MICROWAVE AND RF DESIGN

A Systems Approach

Michael Steer

MICROWAVE AND RF DESIGN

A Systems Approach

MICROWAVE AND RF DESIGN

A Systems Approach

Michael Steer

North Carolina State University

This book is available at special quantity discounts to use as premiums and sales promotions, or for use in corporate training programs. For more information and quotes, please contact the publisher.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To my daughter,

Fiona

Preface

Microwave and RF Design: A Systems Approach is a comprehensive introduction to Radio Frequency (RF) and microwave design with a modern "systems-first" approach. The book has been developed over 10 years of teaching the class to ever-increasing numbers of Master's-level students that is now averaging 90 per class. A strong emphasis on design permeates the book with extensive practical design examples focused on the main technologies encountered in RF engineering today. Design is oriented toward microstrip and planar circuit design so that lessons learned can be applied to real-world design tasks.

Rationale

The central philosophy behind this popular approach is that the student or practicing engineer will complete the book with a full appreciation for RF engineering and gain the practical skills to perform system-level design decisions. Now more than ever, companies need engineers with an ingrained appreciation of systems and armed with the skills to make system decisions. One of the greatest challenges facing RF and microwave engineering, however, is the increasing level of abstraction needed to create innovative microwave and RF systems. This book is organized in such a way that the reader comes to understand the impact that system-level decisions have on component and subsystem design. At the same time, the capabilities of technologies, components, and subsystems impact system design. The book is meticulously crafted to intertwine these themes.

Audience

The book was developed for three courses at North Carolina State University. One is a final-year undergraduate class, another an introductory graduate class, and the third an advanced graduate class. I believe it is completely understandable for senior-level students where a microwave/RF engineering course is offered. The book is a comprehensive RF and microwave text and reference with detailed index, glossary, appendices, and cross-references throughout the text. Practicing engineers will find the book a valuable systems primer, a refresher as needed, and a reference tool in the field. Additionally, it can serve as a valuable, accessible resource for those outside RF circuit engineering who need to understand how they can work with RF hardware engineers.

Organization

The level of sophistication of RF and microwave engineering concepts begins with a broad, easily read treatment of systems in the first three chapters. This develops in the middle of the book with the subtleties of microwave components and immediately useful design knowledge particularly focused on microstrip design. The book finishes with in-depth treatment of network synthesis which embodies the full sophistication of microwave engineering. Detailed design examples are used throughout the book. These are often used to illustrate design decisions that do not lend themselves to formulaic solutions.

The book is organized around five modules that are largely independent but do build on each other and are best taught in sequence. The first three chapters of the book comprise the System Module. The *System Module* (Chapters 1–3) describes system concepts and provides comprehensive knowledge of RF and microwave systems. The emphasis is on understanding how systems are crafted from many different technologies and concepts. For example, the first chapter, "Modulation, Transmitters, and Receivers," considers the system-level trade-offs driven by modulation requirements. The choice of amplifier impacts the microwave requirements that must be designed. Complementary Metal Oxide Semiconductor (CMOS) circuitry in Radio Frequency Integrated Circuits (RFICs) necessarily requires that amplifiers be differential, which in turn drives the design of baluns and other balanced structures. The System Module prepares the reader for subsequent material by providing the rational for design choices. It is written from a hardware perspective with connections between system-level attributes and hardware performance and requirements. The reader therefore gains valuable insight into how different technologies can be traded off in meeting system requirements. I do not believe this systems presentation is available anywhere else in such a compact form.

The *Transmission Line Module* (Chapters 4 and 5) consists of two chapters, beginning with a chapter on transmission line theory, with microstrip lines primarily used as design examples. Design examples illustrate some of the less quantifiable design decisions that must be made, such as choosing manufacturable dimensions and avoiding undesirable modes of operation. This module shows how distributed effects form the basis for a large variety of microwave and RF components that are not available at lower frequencies. The modern treatment emphasizes planar circuit design and the practical aspects of designing around unwanted effects. The text, examples, and problems introduce the often hidden design requirements of designing to mitigate parasitic effects and unwanted modes of operation.

The *Network Module* (Chapters 6–8) presents the descriptions of microwave networks based on S-parameter and ABCD matrices, and the representation of reflection and transmission information on polar plots called Smith charts. Microwave measurement and calibration technology is exam-

ined. A sampling of the wide variety of microwave elements based on transmission lines is presented. It is shown how many of these have lumped element equivalents and how lumped elements and transmission lines can be combined as a compromise between the high performance of transmission line structures and the compactness of lumped elements. The module concludes with an in-depth treatment of matching for maximum power transfer. Both lumped element and distributed element matching are presented.

The *Coupled Line and Filter Module* (Chapters 9 and 10) begins by presenting descriptions of coupling of parallel lines. It is shown that coupled lines can be used to realize a variety of responses. Detailed design of a directional coupler is used to illustrate the use of coupled lines. Network equivalents of coupled lines are introduced as fundamental building blocks that will be used later in filter design and network synthesis. Filter design is presented as a mixture of art and science. This mix, and the thought processes involved, are emphasized through the design of a filter integrated throughout the chapter. This chapter integrates and applies many of the design lessons presented throughout the book.

The final module is the *Active Device and Synthesis Module* (Chapters 11–13) which first considers amplifiers, mixers, oscillators, and switches. Detailed amplifier and voltage-controlled oscillators are presented. Basic RFIC design is presented and the reason for design choices in RFICs made clear. The capstone chapter presents network synthesis through a number of detailed design examples. Network synthesis is the methodical design of microwave networks leading to design insight, novel topologies, and tailored performance. Synthesis is presented as the merger of art, with the choice of topologies with the desired fundamental characteristics, and the mathematical development of successive networks beginning with a lumped element prototype and resulting in the final physical realization.

Course Structures

University courses can be taught around two or three modules, and the book can be used to teach two or three classes. A standard microwave class following the format of earlier microwave texts can be taught around Modules 2 and 3 (i.e., Chapters 4–8). Such a course will benefit from the strong practical design flavor and modern treatment of measurement technology, Smith charts, and matching networks. Transmission line propagation and design is presented in the context of microstrip technology providing an immediately useful skill. The subtleties of multimoding are also presented in the context of microstrip lines. In such a class the easily read chapters in the system module can be assigned for self-learning.

The fundamental philosophy behind the book is that the broader impact of the material must be presented first. Systems should be discussed up front and not left as an afterthought for the final chapter of a textbook or the last class of the semester. A systems first class can be taught beginning

with the systems module and continuing through the transmission line and network module. This class is suited to a final-year class or, with the addition of problems from the extensive set of design-oriented problems at the end of the chapters, for an introductory graduate class. An advanced graduate class can emphasize the final two modules of the book concentrating on amplifiers, oscillators, and network synthesis. This can be supplemented with a laboratory or computer-aided microwave design tools. The book is written so that all electrical engineers can gain an appreciation of RF and microwave hardware engineering. It offers a four-prong flexible approach to RF engineering. For a more traditional approach, the book provides appendices containing detailed reviews of electromagnetic theory, including waveguide theory. The body of the text can be covered without strong reliance on this electromagnetic theory, but it is there for those who desire it for teaching or reader review. The book is rich with detailed information and also serves as a technical reference. The extensive worked examples aid the experienced engineer and the new engineer growing into RF and microwave engineering.

The Systems Engineer

Systems are developed beginning with fuzzy requirements for components and subsystems. Just as system requirements provide impetus to develop new base technologies, the development of new technologies provides new capabilities that drive innovation and new systems. The new capabilities may arise from developments made in support of other systems. Sometimes serendipity leads to the new capabilities. Creating innovative microwave and RF systems that address market needs, or provide for new opportunities, is the most exciting challenge in RF design. The engineers who can conceptualize and architect new RF systems are in great demand. This book began as an effort to train RF systems engineers and as an RF systems resource for practicing engineers. Many RF systems engineers began their careers when systems were simple. Today, appreciating a system requires higher levels of abstraction than in the past, but it also requires detailed knowledge or the ability to access detailed knowledge and expertise. So what makes a systems engineer? There is not a simple answer, but many partial answers. We know that system engineers have great technical confidence and broad appreciation for technologies. They are both broad in their knowledge of a large swath of technologies and also deep in knowledge of a few areas, sometimes called the "T" model. One book or course will not make a systems engineer. It is clear that there must be a diverse set of experiences. This book fulfills the role of fostering both high-level abstraction of RF engineering, and also detailed design skills to realize effective RF and microwave components. My hope is that this book will provide the necessary background for the next generation of RF systems engineers by stressing system principles immediately, followed by

core component technologies. Core technologies are thereby covered within the context of the systems in which they are used.

Supplementary Materials

Supplementary materials presently available include PowerPoint slides and solutions to selected problems, which can be downloaded from www.scitechpub.com/steer. Instructors can also obtain a complete set of solutions by writing editor@scitechpub.com.

Comments and Suggestions

This book will continue to evolve with the experience of instructors, their students, and other interested readers. Please send your comments, questions, and suggestions to RFDesign@scitechpub.com.

Acknowledgements

Executives at various companies and government laboratories helped develop the strategy for teaching RF systems engineering, in particular reemphasizing the systems first approach. In particular, I thank the technical leaders at Northrop Grumman, Harris Corporation, and the U.S. Army Research Laboratory.

Many people helped in producing this book. Ms. Claire Sideri assisted in editing several of the chapters. Ms. Susan Manning provided guidance on layout and developed the cover art. Mr. Robert Lawless was instrumental in seeing this book through the production process. Ms. Laura Bader went through the final manuscript and perfected the writing and grammar. The front cover captures what this book is trying to accomplish, we are looking right into the core of a system. The X-ray image on the cover was taken by Mr. Jonathan Campbell of the Naval Explosive Ordnance Disposal Technology Division. The publisher, task master, and chief coordinator, Mr. Dudley Kay, helped provide focus and the final push to finish. The book was produced using LaTeX and 95% of the graphics were drawn using xfig. So thanks to the many volunteers who developed these packages.

My family, Mary, Cormac, Fiona, and Killian, gracefully put up with my absence for innumerable nights and weekends, many more than I could have ever imagined. Having a professor as a husband and father is not for the mild. I truly thank them.

Michael Steer
North Carolina State University
Raleigh, North Carolina

Contents

Modulation, Transmitters, and Receivers

1.1 Introduction

The front end of a Radio Frequency (RF) communication receiver combines a number of subsystems in cascade to achieve several objectives. Filters and matching networks provide frequency selectivity to eliminate interfering signals. Amplifiers manage noise levels by boosting both received signals and signals to be transmitted. Mixers coupled with oscillators translate the modulated information from one frequency to another.

There are only a few types of receiver and transmitter architectures. In a receiver, the central idea is to take information superimposed on an RF signal or carrier and convert it to a lower frequency form which can be directly applied to a speaker or digitized. In a cellular communication system, the low-frequency signal, often called the baseband signal, could have a bandwidth of 30 kHz to 5 MHz and the carrier frequency could be 500 MHz to 2 GHz. A transmitter takes the baseband signal and superimposes it on an RF carrier which can be more easily radiated into space and propagates easily from one antenna to another. The essential receiver and transmitter architectures are shown in Figure 1-1. In a **receiver**,
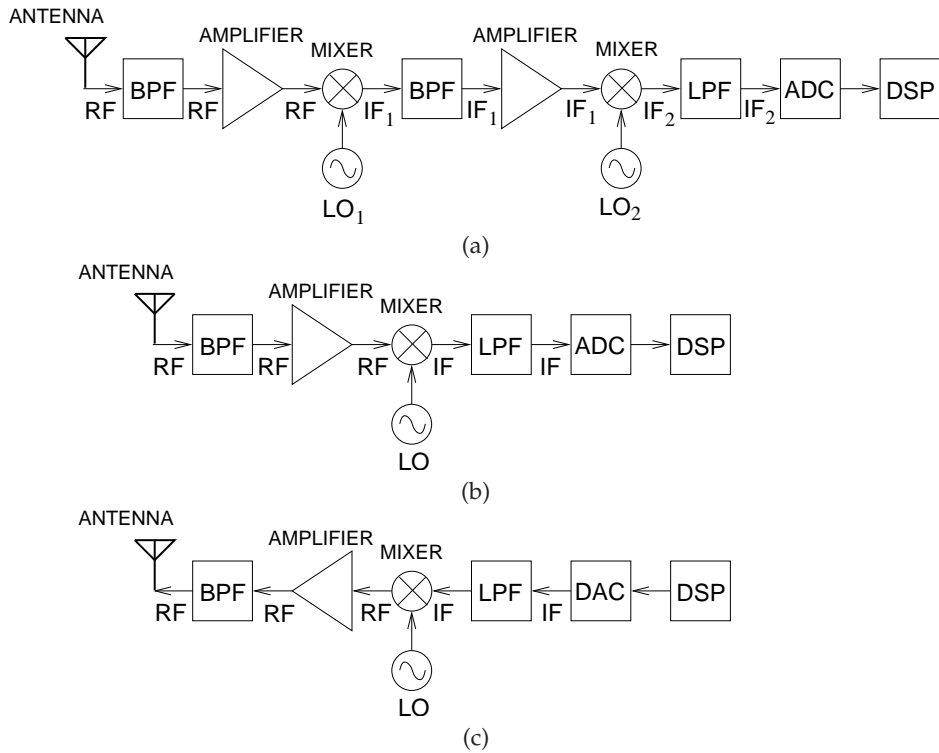
Figure 1-1   Unilateral RF front end: (a) a receiver with two mixing (or heterodyning) stages; (b) a receiver with one heterodyne stage; and (c) a one-stage transmitter.

mixers down-convert information superimposed on an RF carrier to a lower frequency that can be directly connected to speakers or digitized by an **Analog-to-Digital Converter** (**ADC**). With a transmitter, the low-frequency information-bearing signal is translated to a frequency that can be more easily radiated. The most common receiver architecture is shown in Figure 1-1(a). First, an antenna collects a broad portion of the electromagnetic spectrum. Antennas have relatively low frequency selectivity (they have broad bandwidth) and unwanted signal levels can be large, so additional filtering by a **bandpass filter** (**BPF**) is required to reduce the range of voltages presented to the first amplifier. Eventually this signal is digitized by an ADC, but to do this the frequency of the information-carrying part of the signal must be reduced. The stepping down of frequency is accomplished by a mixer stage. With the mixer driven by a large **Local Oscillator** (**LO**) signal, the output at the **Intermediate Frequency** (**IF**) is at the difference frequency of the RF and LO (see Figure 1-2). Thus $f_{\text{IF}} = f_{\text{RF}} - f_{\text{LO}}$ (although sometimes the LO is above the RF so that $f_{\text{IF}} = f_{\text{LO}} - f_{\text{RF}}$). LOs generally have noise close to the operating frequency so that there is a limit
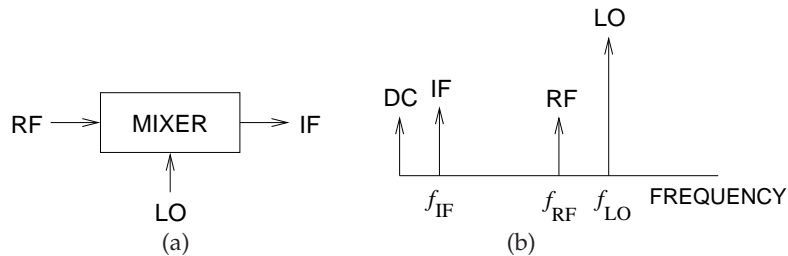
Figure 1-2 Simple mixer circuit: (a) block diagram; and (b) spectrum.

on how close the RF and LO can be in frequency without oscillator noise appearing at the IF. If there is a single mixer, then the IF may still be too high. A solution is to use two stages of mixing. A BPF between the mixing (or heterodyning) stages further blocks unwanted signals. Eventually a **lowpass filter** (LPF) allows only the final IF (here IF_2) to be presented to the ADC. Once digitized, it is possible to further filter the intended signal which originally appeared as modulation at the RF. A one-stage receiver (see Figure 1-1(b)) generally requires a more capable ADC with a higher operating frequency. The elimination of a mixing stage, however, reduces cost and size. The architecture of a transmitter is similar to that of a receiver, with a key difference being the use of a **Digital-to-Analog Converter** (**DAC**) (see Figure 1-1(c)).

The major **active elements** in the RF **front end** of both the transmitter and receiver are the **amplifiers**, **mixers**, and **oscillators**. These subsystems have much in common, using nonlinear devices to convert power at DC to power at RF. In the case of mixers, power at the LO is also converted to power at RF. The front end of a typical cellphone is shown in Figure 1-3. The components here are generally implemented in a module and use different technologies for the various elements, optimizing cost and performance. There are many variants of the architecture shown here. At one extreme, a module is used with all of the components packed in a shielded structure perhaps 1 cm on a side and 2–3 mm thick. Another extreme is a single-chip implementation, usually in **Bipolar with Complementary Metal Oxide Semiconductor** (**BiCMOS**) technology, **Silicon Germanium** (**SiGe**) technology, or high performance CMOS called **RF CMOS**. However, it is necessary to use a gallium arsenide GaAs device to efficiently achieve the hundreds of milliwatts that typically must be transmitted.

Return now to the mixer-based **transceiver** (for transmitter and receiver) architecture shown in a multichip form in Figure 1-3. Here, a single antenna is used, and either a **duplexer**[1] (a combined lowpass and highpass filter)

---

[1] A duplexer separates transmitted and received signals and is generally implemented as a filter and is a type of diplexer. A **diplexer** combines two signals on different frequency bands
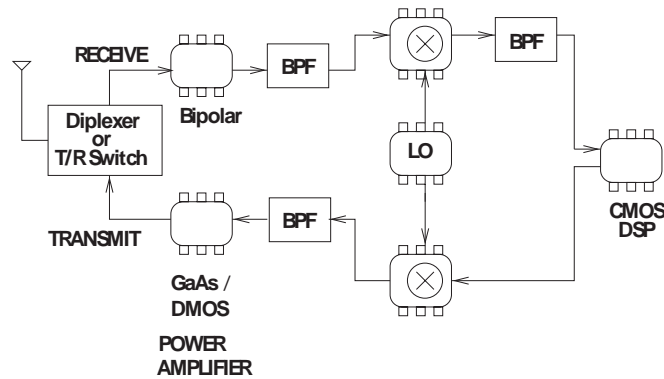
Figure 1-3   RF front end organized as multiple chips.

or a switch is used to separate the (frequency spaced) transmit and receive paths. If the system protocol requires transmit and receive at the same time, a duplexer is required to separate the transmit and receive paths. This filter tends to be large, lossy, or costly (depending on the technology used). Consequently a transistor **switch** is preferred if the transmit and receive signals operate in different time slots. In the receive path, a CMOS or BiCMOS chip initially amplifies the low-level received signal, and so low noise is important. This amplifier is thus called a **Low-Noise Amplifier** (**LNA**). The amplified receive signal is then bandpass filtered and frequency down-converted by a mixer (indicated by a circle with a cross in it) to IF that can be sampled by an ADC to produce a digital signal that is further processed by **Digital Signal Processing** (**DSP**). Variants of this architecture include one that has two down-conversion stages, and another with no mixing that relies instead on direct conversion of the receive signal using a subsampling ADC. In the transmit path, the architecture is reversed, with a DAC driven by the DSP chip that produces an information-bearing signal at the IF which is then frequency up-converted by a mixer, bandpass filtered, and amplified by what is called a power amplifier to generate the hundreds of milliwatts required. An alternative transmitter design is **Direct Digital Synthesis** (**DDS**), which bypasses the conversion stage. Direct conversion and DDS are difficult to implement, but are essential for the highly desired single or few chip solution.

This chapter describes the operation and design strategies for the RF front-end architecture of Figure 1-3, looking at amplifiers, mixers, switches, and oscillators. This architecture is used in most high-performance RF and microwave communication and radar systems. While the subsystems are

---

so that they can be sent over a common path. A diplexer is normally a multiport filter, but diplexing can be implemented in a digital signal processor.

preferably linear at RF, this can only be approximated, as the active devices used are intrinsically nonlinear. Performance is limited fundamentally by distortion, which is related to the characteristics of the RF signal, and this in turn is determined by the modulation scheme that impresses information on an RF carrier.

1.2 RF Signals

RF communication signals are engineered to trade off efficient use of the **electromagnetic** (**EM**) spectrum with the complexity and performance of the required RF hardware. The process of converting baseband (or low-frequency) information to RF is called modulation, of which there are two types: analog and digital modulation. In analog modulation, the RF signal has a continuous range of values; in digital modulation, the output has a number of discrete states. There are just a few modulation schemes that achieve the optimum trade-offs of spectral efficiency and ease of use with hardware complexity. The major modulation schemes include the following:

Analog modulation
AM	Amplitude modulation
FM	Frequency modulation
PM	Phase modulation

Digital modulation
FSK	Frequency shift keying
PSK	Phase shift keying
MSK	Minimum shift keying (a form of FSK)
GMSK	Minimum shift keying using Gaussian filtered data
BFSK	Binary frequency shift keying
BPSK	Binary phase shift keying
QPSK	Quadrature PSK (QPSK is also referred to as quarternary PSK, quadraphase PSK, and quadra PSK)
$\pi/4$-DQPSK	$\pi/4$ Differential encoded QPSK
OQPSK	Offset QPSK
SOQPSK	Shaped Offset QPSK
SBPSK	Shaped BPSK
FOQPSK	Feher Offset QPSK
8PSK	8-state phase shift keying
$3\pi/8$-8PSK	$3\pi/8$, 8-state phase shift keying
16PSK	16-state phase shift keying
QAM	Quadrature amplitude modulation

Frequency modulation, and the similar PM modulation schemes, are used in analog cellular radio. With the addition of legacy AM, the three schemes are the bases of analog radio. The other schemes are used in digital radio, including digital cellular radio. GMSK is used in the **Global System for**
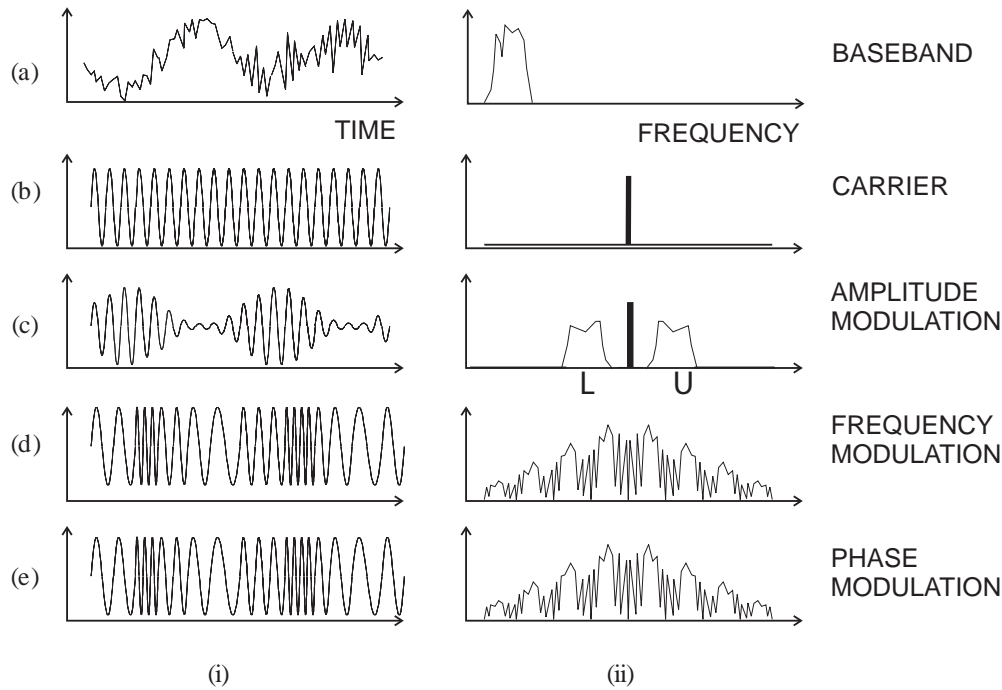
Figure 1-4 Analog modulation showing (i) waveform and (ii) spectrum for (a) baseband signal; (b) carrier; (c) carrier modulated using amplitude modulation; (d) carrier modulated using frequency modulation; and (e) carrier modulated using phase modulation.

Mobile Communications (**GSM**) cellular system and is a form of FSK and produces a constant amplitude modulated signal. The FM, FSK, GMSK, and PM techniques produce constant RF envelopes, thus no information is contained in the amplitude of the signal. Therefore errors introduced into the amplitude of the system are of no significance, so efficient saturating-mode amplifiers such as Class C can be used, thus extending battery life.

So there is a trade-off in the complexity of RF design, choice of modulation format, and battery life. In contrast, the MSK, $\pi/4$-DQPSK, $3\pi/8$-8PSK, and QAM techniques do not result in constant RF envelopes, so information is contained in the amplitude of the RF signal. Thus more sophisticated RF processing hardware is required.

1.3 Analog Modulation

Wireless modulation formats in conventional narrowband radio are based on modifying the properties of a carrier by slowly varying the amplitude and phase of the carrier. The waveforms and spectra of common analog modulation formats are shown in Figure 1-4.

1.3.1 Amplitude Modulation, AM

Amplitude Modulation (**AM**) is the simplest analog modulation scheme to implement. Here a signal is used to slowly vary the amplitude of the carrier according to the level of the modulating signal. The modulating signal is generally referred to as the baseband signal and it contains all of the information to be transmitted or interpreted. The waveforms in Figure 1-4 are stylized, as the variation in the carrier is relatively fast. They are presented this way so that the effects of modulation can be more easily interpreted. The baseband signal (Figure 1-4(a)) is shown as having a period that is not too far away from the period of the carrier (Figure 1-4(b)). In reality, there would be hundreds or thousands of RF cycles for each cycle of the baseband signal so that the frequency of the baseband signal would have frequency components which are a tiny fraction of the frequency of the carrier.

With AM (Figure 1-4(c)), the amplitude of the carrier is modulated, and this results in a broadening of the spectrum of the carrier, as shown in Figure 1-4(c)(ii). This spectrum contains the original carrier component and upper and lower sidebands designated as U and L, respectively. In AM, the two sidebands contain identical information, so all the information would be transmitted if the carrier and one of the sidebands were suppressed. With the carrier present, it is easy to receive a signal by bandpass filtering the incoming modulated signal, rectifying the result, and then lowpass filtering the rectified signal to remove harmonics of the baseband signal.

An AM signal $x(t)$ has the form

$$x(t) = A_c \left[1 + my(t)\right] \cos \omega_c t, \tag{1.1}$$

where m is called the modulation index and $y(t)$ is the baseband information-bearing signal that has frequency components which are below the carrier radian frequency ω_c. Provided that $y(t)$ varies slowly relative to the carrier, that is, the frequency components of $y(t)$ are significantly below the carrier frequency, $x(t)$ looks like a carrier whose amplitude varies slowly. To get an idea of how slowly the amplitude varies in actual systems, consider an AM radio that broadcasts at 1 MHz (which is in the middle of the AM broadcast band). The highest frequency component of the modulating signal corresponding to voice is about 4 kHz. Thus the amplitude of the carrier takes 250 carrier cycles to go through a complete amplitude variation. At all times a cycle of the carrier appears to be periodic, but in fact it is not quite. It is common to refer to the modulated carrier as being quasi-periodic and to the apparent carrier as being the pseudo-carrier.

The concept of the envelope of a modulated RF signal is introduced in Figure 1-5. Figure 1-5(a) is the carrier; the amplitude-modulated carrier is shown in Figure 1-5(b). The outline of the modulated carrier is called the envelope, and for AM this is identical to the modulating signal. Both the envelope and the modulating signal are shown in Figure 1-5(c). At the peak
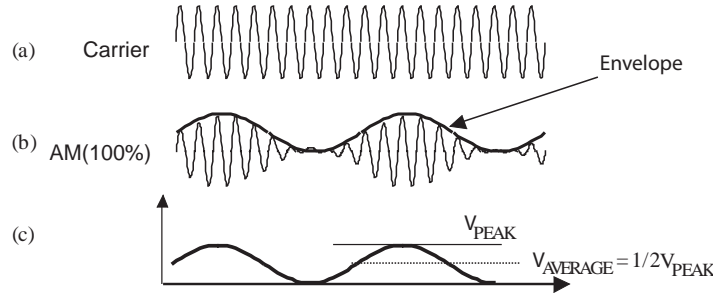
Figure 1-5 AM showing the relationship between the carrier and modulation envelope: (a) carrier; (b) 100% amplitude modulated carrier; and (c) modulating or baseband signal.

of the envelope, the RF signal has maximum power (considering the power of a single RF cycle). Since we are dealing with 100% AM, $m = 1$ in Equation (1.1) and there is no RF power when the envelope is at its minimum.

One of the characteristics of various modulation formats is the ratio of the power of the signal when the carrier is at its peak (i.e., the power in one cycle of the carrier when the envelope is at its maximum) relative to its average value (the power averaged over all time). This is called the **Peak-to-Average Ratio** (**PAR**) and is a good indicator of how sensitive a modulation format is to the effects of nonlinearity of the RF hardware.

It is complex to determine the PAR for a general signal, but a good estimate can be obtained by considering that the modulating signal is a sinewave. Let $y(t)$ $(= \cos \omega_m t)$ be a cosinusoidal modulating signal with radian frequency ω_m. Then (for AM)

$$x(t) = A_c \left[1 + m \cos \omega_m t\right] \cos \omega_c t . \tag{1.2}$$

Thus if just one quasi-period of this signal is considered (i.e., one variation of the modulated signal at the carrier frequency), then the signal has a power that varies with time.

Consider a voltage $v(t)$ across a resistor of conductance G. The power of the signal must be determined by integrating over all time, which is work, and dividing by the time period yields the average power:

$$P_{\text{avg}} = \lim_{\tau \to \infty} \int_{-\tau}^{\tau} \frac{1}{2\tau} G v^2(t) \, dt . \tag{1.3}$$

Now, if $v(t)$ is a cosinusoid, $v(t) = A \cos \omega t$, then

$$P_{\text{avg}} = \lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} A_c^2 G \cos^2 (\omega t) \, dt$$

$$= \lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} A_c^2 G \frac{1}{2} \left[1 + \cos\left(2\omega t\right)\right] dt$$

$$= \frac{1}{2} A_c^2 G \left\{ \lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} 1 \, dt + \lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos\left(2\omega t\right) dt \right\} = \frac{1}{2} A_c^2 G \, . \quad (1.4)$$

In the above equation, a useful equivalence has been employed by observing that the infinite integral of a cosinusoid can be simplified to just integrating over one period, $T = 2\pi/\omega$:

$$\lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos^n\left(\omega t\right) dt = \frac{1}{T} \int_{-T/2}^{T/2} \cos^n\left(\omega t\right) dt \, , \qquad (1.5)$$

where n is a positive integer. In power calculations there are a number of other useful simplifying techniques based on **trigonometric identities**. Some of the ones that will be used are the following:

$$\cos A \cos B \ = \ \frac{1}{2} \left[\cos(A - B) + \cos(A + B)\right]$$

$$\cos^2 A \ = \ \frac{1}{2} \left[1 + \cos(2A)\right] \qquad (1.6)$$

$$\lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos \omega t \, dt \ = \ \frac{1}{T} \int_{-T/2}^{T/2} \cos\left(\omega t\right) dt = 0 \qquad (1.7)$$

$$\frac{1}{T} \int_{-T/2}^{T/2} \cos^2\left(\omega t\right) dt \ = \ \frac{1}{T} \int_{-T/2}^{T/2} \frac{1}{2} \left[\cos\left(2\omega t\right) + \cos(0)\right] \qquad (1.8)$$

$$= \ \frac{1}{2T} \left[\int_{-T/2}^{T/2} \cos\left(2\omega t\right) dt + \int_{-T/2}^{T/2} 1 \, dt\right]$$

$$= \ \frac{1}{2T}(0 + T) = \frac{1}{2} \, . \qquad (1.9)$$

More trigonometric identities are given in Appendix A.3 on Page 828. Also, when cosinusoids $\cos At$ and $\cos Bt$, having different frequencies ($A \neq B$), are multiplied together, then

$$\int_{-\tau}^{\tau} \cos At \cos Bt \, dt = \int_{-\tau}^{\tau} \left[\cos(A + B)t + \cos(A + B)t\right] dt = 0, \qquad (1.10)$$

and, in general, if $A \neq B \neq 0$,

$$\int_{-\infty}^{\infty} \cos At \cos^n Bt \, dt = 0 \, . \qquad (1.11)$$

Now the discussion returns to characterizing an AM signal by considering long-term average power and the short-term power of the signal. The

maximum amplitude of the pseudo-carrier at its peak amplitude is, from Equation (1.2),

$$x_p(t) = A_c \left[1 + m\right] \cos \omega_c t \ . \tag{1.12}$$

Then the power (P_{peak}) contained in the peak pseudo-carrier is obtained by integrating over one period:

$$
\begin{aligned}
P_{\text{peak}} &= \frac{1}{T} \int_{-T/2}^{T/2} G x^2(t) \, dt = \frac{1}{T} \int_{-T/2}^{T/2} A_c^2 G (1 + m)^2 \cos^2 (\omega_c t) \, dt \\
&= A_c^2 G \left(1 + m\right)^2 \frac{1}{T} \int_{-T/2}^{T/2} \cos^2 (\omega_c t) \, dt = \frac{1}{2} A_c^2 G \left(1 + m\right)^2 . \tag{1.13}
\end{aligned}
$$

The **average power** (P_{avg}) of the modulated signal is obtained by integrating over all time, so

$$
\begin{aligned}
P_{\text{avg}} &= \lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} G x^2(t) \, dt \\
&= A_c^2 G \lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \left\{ [1 + m \cos (\omega_m t)] \cos (\omega_c t) \right\}^2 \, dt \\
&= A_c^2 G \lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \left\{ \left[1 + 2m \cos (\omega_m t) + m^2 \cos^2 (\omega_m t) \right] \cos^2 (\omega_c t) \right\} \, dt \\
&= A_c^2 G \lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \left[\cos^2 (\omega_c t) + 2m \cos (\omega_m t) \cos^2 (\omega_c t) \right. \\
&\quad \left. + m^2 \cos^2 (\omega_m t) \cos^2 (\omega_c t) \right] \, dt \\
&= A_c^2 G \left[\lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos^2 (\omega_c t) \, dt \right. \\
&\quad + \lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} 2m \cos (\omega_m t) \cos^2 (\omega_c t) \, dt \\
&\quad \left. + \lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} m^2 \cos^2 (\omega_m t) \cos^2 (\omega_c t) \, dt \right] \\
&= A_c^2 G \left[\frac{1}{2} + 0 + \lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} m^2 \cos^2 (\omega_m t) \cos^2 (\omega_c t) \, dt \right] \\
&= A_c^2 G \left\{ \frac{1}{2} + m^2 \lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \frac{1}{4} [1 + \cos (2\omega_m t)] [1 + \cos (2\omega_c t)] \, dt \right\} \\
&= A_c^2 G \left\{ \frac{1}{2} + \right. \\
&\quad \left. \frac{m^2}{4} \lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} [1 + \cos (2\omega_m t) + \cos (2\omega_c t) + \cos (2\omega_m t) \cos (2\omega_c t)] \, dt \right\}
\end{aligned}
$$

$$P_{\text{avg}} = A_c^2 G \left\{ \frac{1}{2} + \frac{m^2}{4} \left[\lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} 1 \, dt + \lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos\left(2\omega_m t\right) dt \right. \right.$$

$$\left. \left. + \lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos\left(2\omega_c t\right) dt + \lim_{\tau \to \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos\left(2\omega_m t\right) \cos\left(2\omega_c t\right) dt \right] \right\}$$

$$= A_c^2 G \left[1/2 + m^2(1/4 + 0 + 0 + 0) \right]$$

$$= \frac{1}{2} A_c^2 G (1 + m^2/2) \, . \tag{1.14}$$

So the PAR of an AM signal (i.e., PAR_{AM}) is

$$\text{PAR}_{\text{AM}} = \frac{P_{\text{peak}}}{P_{\text{avg}}} = \frac{\frac{1}{2} A_c^2 G \left(1 + m\right)^2}{\frac{1}{2} A_c^2 G (1 + m^2/2)} = \frac{(1 + m)^2}{1 + m^2/2} \, .$$

For 100% AM described by $m = 1$, the PAR is

$$\text{PAR}_{100\%\text{AM}} = \frac{(1 + 1)^2}{1 + 1^2/2} = \frac{4}{1.5} = 2.667 = 4.26 \text{ dB}. \tag{1.15}$$

In expressing the PAR in decibels, the formula $\text{PAR}_{\text{dB}} = 10 \log_{10}(\text{PAR})$ was used, as the PAR is a power ratio. As an example, for 50% AM described by $m = 0.5$, the PAR is

$$\text{PAR}_{50\%\text{AM}} = \frac{(1 + 0.5)^2}{1 + 0.5^2/2} = \frac{2.25}{1.125} = 2 = 3 \text{ dB}. \tag{1.16}$$

The PAR is an important attribute of a modulation format and impacts the types of circuit designs that can be used. It is much more challenging to achieve low levels of distortion when the PAR is high. When comparing two signals it is common to assume that the powers are referenced to the same resistance. In this case the square of the voltage ratios of the signals can be used as the power ratio.

It is tempting to consider if the lengthy integrations can be circumvented. Powers can be added if the signal components (the tones making up the signal) are uncorrelated. If they are **correlated**, then the complete integrations are required.[2] Consider two uncorrelated sinusoids of (average) powers P_1 and P_2, then the average power of the composite signal is $P_{\text{avg}} = P_1 + P_2$. However, in determining peak power, the RF cycle where the two sinusoids align is considered, and here the voltages add to produce a sinewave with a higher amplitude. So peak power applies to just one RF pseudo-cycle. Generally the voltage amplitude of the two sinewaves

---

[2]   For the purposes here, two signals are uncorrelated if the integral of their product over all time and all offsets is zero. That is, $x(t)$ and $y(t)$ are uncorrelated if $C = \int_{-\infty}^{+\infty} x(t) y(t + \tau) \, dt = 0$ for all τ, otherwise they are correlated (or partly correlated). For a more complete definition see Reference [1].

would be added and then the power calculated. If the uncorrelated carriers are modulated and the modulating signals (the baseband signals) are uncorrelated, then the average power can be determined in the same way, but the peak power calculation is much more complicated. The integrations are the only calculations that can always be relied on. They can be used with all signals, including digitally-modulated signals.

1.3.2   *Phase and Frequency Modulation, PM and FM*

The two other analog modulation schemes commonly used are **Phase Modulation** (**PM**) (Figure 1-4(e)) and **Frequency Modulation** (**FM**) (Figure 1-4(d)). The signals produced by the two schemes are identical; the difference is how the signals are generated. In PM, the phase of the carrier depends on the instantaneous level of the baseband signal. In FM, the amplitude of the baseband signal determines the frequency of the carrier. The result in both cases is that the bandwidth of the time-varying signal is spread out, as seen in Figure 1-6. A **receiver** must compress the spread-out information to re-create the original narrowband signal, and this can be thought of as processing gain, as the compression of correlated signals significantly increases the tolerance to noise. As will be seen, processing gain is essential in digital radio, which uses digital modulation. The peak amplitude of the RF **phasor** is equal to the average amplitude and so the PAR is 1 or 0 dB. A summary of the PAR of the primary analog modulated signals is given in Figure 1-7.

Frequency modulation was invented by Edwin H. Armstrong and patented in 1933. FM is virtually static free and clearly superior to AM radio. However, it was not immediately adopted largely because AM radio was established in the 1930s, and the adoption of FM would have resulted in the scrapping of a large installed infrastructure (seen as a commercial catastrophe) and so the introduction of FM was delayed by decades. The best technology does not always win immediately! Commercial interests and the interests of those heavily invested in an alternative technology have a great deal to do with the success of a technology.

Carson's Rule

Frequency- and phase- modulated signals have unlimited bandwidth but the information content of the sidebands drops off rapidly. The bandwidth required to reliably transmit a PM or FM signal is subjective, but the best accepted criterion is called Carson's bandwidth rule or just Carson's rule [2, 3]. It provides an estimate of the bandwidth capturing approximately 98% of the energy when a carrier is frequency or phase modulated by a continuous spectrum baseband signal. An FM signal is shown in Figure 1-6. In particular, Figures 1-6(a) and 1-6(b) show the FM function and then the spectrum that results when a single sinewave modulates the
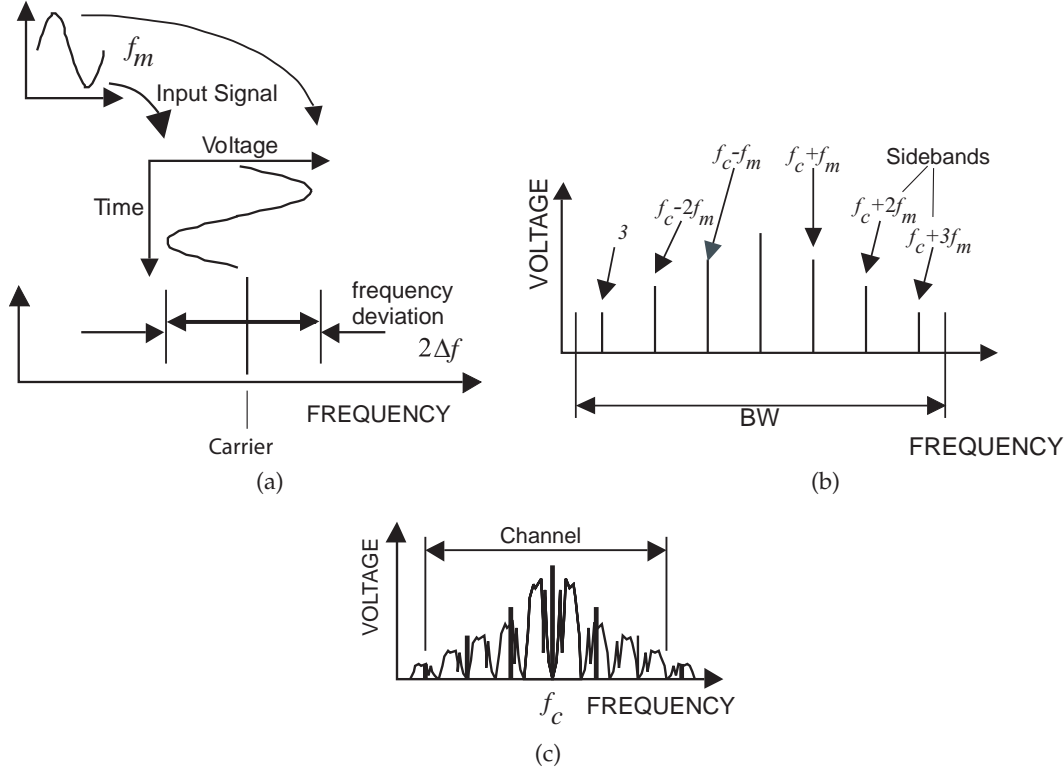
Figure 1-6 Frequency modulation by a sinewave: (a) signal varying the frequency of carrier; (b) spectrum of the resulting waveform; and (c) spectrum when modulated by a continuous baseband signal.
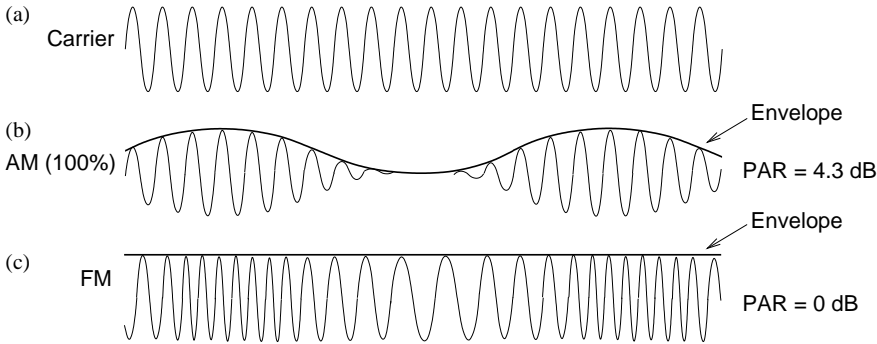
Figure 1-7 Comparison of 100% AM and FM highlighting the envelopes of both: (a) carrier; (b) AM signal with envelope; and (c) FM or PM signal with the envelope being a straight line or constant.

frequency of a carrier. As time passes, the carrier moves up and down in frequency synchronously with the level of the input baseband signal. The level (typically voltage) of the baseband signal determines the frequency deviation of the carrier from its unmodulated value. The frequency shift when the modulating signal is at its maximum amplitude is called the peak frequency deviation, Δf, and the maximum frequency of the modulating frequency is f_m. Figure 1-6(c) shows the spectrum that results when the modulating signal, or baseband signal, is continuous. There are multiple sidebands, with the relative strength of each being dependent on a bessel function of the highest modulation frequency, f_m, and the maximum frequency deviation, Δf. Carson's rule, derived from these considerations, is

$$\text{Bandwidth required} = 2 \times (f_m + \Delta f). \qquad (1.17)$$

Narrowband and Wideband FM

The FM signal, as used in FM broadcast radio, is also called wideband FM, as the maximum frequency deviation is much greater than the highest frequency of the modulating or baseband signal, that is, $\Delta f \gg f_m$. A more spectrally efficient form of FM is called narrowband FM, where $\Delta f \ll f_m$. Narrowband FM was developed as a more bandwidth efficient form of FM, but of course digital radio has passed this now and narrowband FM is no longer an important modulation type. The trade-off is that narrowband FM, as opposed to wideband FM, requires more sophisticated demodulation and hence more complex circuits are required. It should also be noted that FM, as used in conventional FM broadcast radio, is being phased out so that spectrum can be used more efficiently.

1.3.3   Two-Tone Signal

A **two-tone signal** is a signal that is the sum of two cosinusoids. Thus

$$y(t) = X_A \cos(\omega_A t) + X_B \cos(\omega_B t) \qquad (1.18)$$

is a two-tone signal. Generally the frequencies of the two tones are close, with the concept being that the two tones both fit within the passband of a bandpass filter, so it would be reasonable to assume that the individual tones have frequencies that are within 1% of each other. A two-tone signal is not a form of modulation, but is commonly used to characterize the performance of RF systems. The composite signal would then look like a slowly varying pseudo-carrier, not unlike an AM signal. The tones are uncorrelated so that the average power of the composite signal, $y(t)$, is the sum of the powers of each of the individual tones. The peak power of the composite signal is the peak pseudo-carrier, so $y(t)$ has a peak amplitude of $X_A + X_B$. Similar concepts apply to three-tone and n-tone signals.
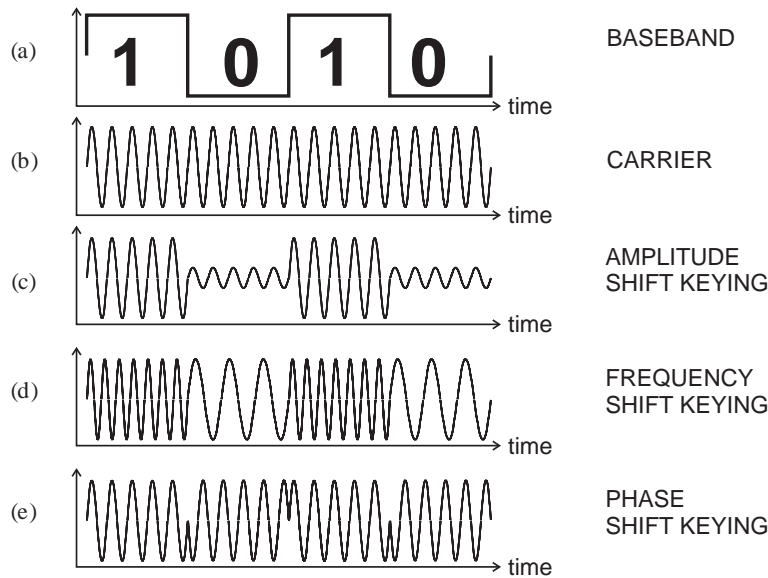
Figure 1-8 Modes of digital modulation: (a) modulating bitstream; (b) **carrier**; (c) carrier modulated using amplitude shift keying (ASK); (d) carrier modulated using frequency shift keying (FSK); and (e) carrier modulated using binary phase shift keying (BPSK).

EXAMPLE 1. 1	PAR of a Two-Tone Signal

What is the PAR of a two-tone signal with both tones having equal amplitude?

Solution:

Let $X_A = X_B = X$, the peak pseudo-carrier has amplitude $2X$, and so the power of the peak RF carrier is proportional to $\frac{1}{2}(2X)^2 = 2X^2$. The average power is proportional to $\frac{1}{2}(X_A^2 + X_B^2) = \frac{1}{2}(X^2 + X^2) = X^2$, as each one is independent of the other, and so the powers can be added.

$$\text{PAR} = \frac{2X^2}{X^2} = 2 = 3 \text{ dB}. \tag{1.19}$$

1.4 Digital Modulation

Digital modulation was first employed in sending telegraph signals wirelessly in which a carrier was switched, or keyed, on and off to create pulses of the carrier signal. This modulation is now known as **Amplitude Shift Keying** (**ASK**), but today this scheme is little used. Several digital modulation formats are shown in Figure 1-8. The fundamental characteristic
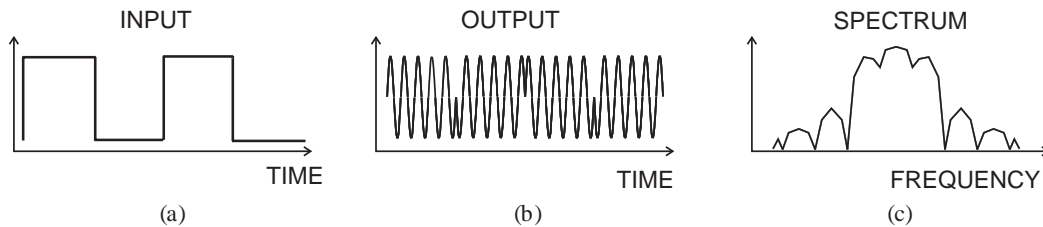
Figure 1-9 Characteristics of phase shift keying (PSK) modulation: (a) modulating bitstream; (b) the waveform of the carrier modulated using PSK, with the phase determined by the 1s and 0s of the modulating bitstream; and (c) the spectrum of the modulated signal.

of **digital modulation** is that there are discrete states, each of which defines a symbol, with a symbol representing one or more bits. In Figure 1-8, there are only two states representing one of two values for a bit (0 or 1). With multiple states, groups of bits can be represented. There are many digital modulation formats that have proved successful and many of these are considered below. In modern communication schemes it is important to be able to recover the original carrier, so it is important that the amplitude of the carrier not be small for an extended period of time as it is in the ASK scheme illustrated in Figure 1-8(c)

1.4.1   Phase Shift Keying, PSK

The waveforms and spectrum of a PSK modulated signal are shown in Figure 1-9. The incoming baseband bitstream (Figure 1-9(a)) is lowpass filtered and used to modulate the phase of a **carrier** (Figure 1-9(b)). The spectrum of this signal is shown in Figure 1-9(c). The PSK modulation scheme is similar to that represented in Figure 1-10, with the FSK modulator replaced by a PSK modulator which shifts the phase of the carrier rather than its frequency. There are many variants of PSK, with the most fundamental characteristics being the number of phase states (e.g., with 2^n phase states, n bits of information can be transmitted) and how the phasor of the RF signal transitions from one phase state to another. Generally PSK schemes shape the spectrum of the modulated signal to fit as much energy as possible within a spectral mask. This results in a modulated carrier whose amplitude varies (and thus a time-varying envelope). Such schemes require highly linear amplifiers to preserve the amplitude variations of the modulated RF signal. Other schemes orchestrate the phase transitions to achieve a constant envelope modulated RF signal but have lower spectral efficiency. Two approaches to achieving this are, first, to slow the transitions down, and, second, to eliminate transitions from a phase state to one which is rotated by 180° and so avoid the RF phasor traversing the origin. The result of both approaches is that relatively simple hardware can be used, as
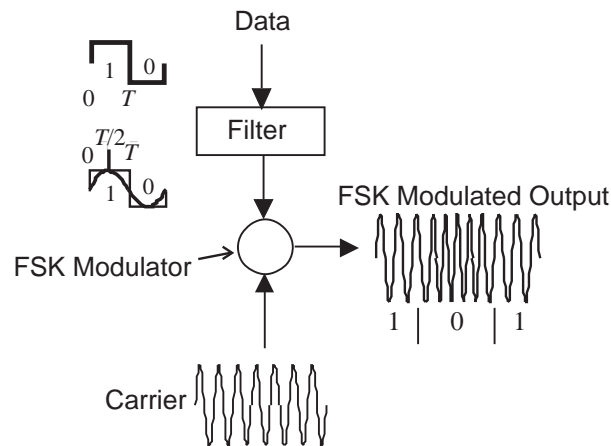
Figure 1-10   The frequency shift keying (FSK) modulation system.

amplitude distortion is not a problem. So system design affects RF hardware complexity, and the sophistication of available and affordable hardware impacts system design. There are only a few variants that achieve optimum properties, and many of these will be considered later in this chapter.

The communication limit of 1 symbol per hertz of bandwidth, the **symbol rate**, comes from the **Nyquist signaling theorem**. Nyquist determined that the number of independent pulses that could be put through a telegraph channel per unit of time is limited to twice the bandwidth of the channel. With a modulated RF carrier, this translates to a pulse of information on the I (or cosine) component, and a pulse of information on the Q (or sine) component, in a unit of time equal to $1/$bandwidth. Combining the I and Q channels, the phasor can move from one value to another in a unit of time equal to $1/$bandwidth. The phasor transition identifies a symbol, and hence one symbol can be sent per hertz of bandwidth.

1.4.2   Binary Phase Shift Keying, BPSK

Phase shift keying demodulation requires more sophisticated signal processing than does FSK. PSK uses prescribed phase shifts to define symbols, each of which can represent one, two, or more bits. **Binary Phase Shift Keying** (**BPSK**), illustrated in Figure 1-8(e), has one bit per symbol and is a relatively spectrally inefficient scheme, with a maximum spectral efficiency of 1 bit/second/hertz (1 b·s^{-1}·Hz^{-1}). Although spectrally inefficient, it is ideally suited to low-power applications and single-chip implementations, perhaps with an off-chip reference resonator. The typical signal flow is from an antenna, through an RF-tuned amplifier, with quadrature mixing to produce I and Q channels which are then lowpass
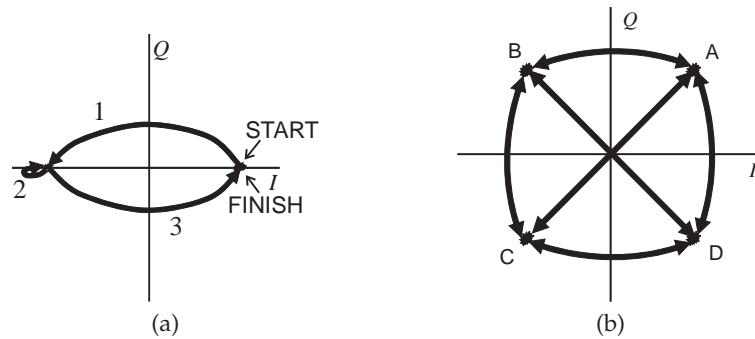
Figure 1-11 Constellation diagrams with possible transitions: (a) a binary modulation scheme; and (b) QPSK, a four-state phase modulation scheme. Each state is a symbol.

filtered. The filtered I and Q channels are then commonly integrated over the duration of a bit. In the most sensitive scheme, the I and Q channels are oversampled (by an ADC) at a multiple of the bit rate and the signal correlated with the expected zerocrossing.

BPSK is commonly used in pagers and is used in **Bluetooth**. The operation of BPSK modulation can be described using the constellation diagram shown in Figure 1-11(a). The BPSK constellation diagram indicates that there are two states. These states can be interpreted as the values of $i(t)$ and $q(t)$ at the sampling points corresponding to the bit rate. The curves in Figure 1-11(a) indicate three transitions. The states are at the ends of the transitions. If a 1, in Figure 1-11(a), is assigned to the positive I value and 0 to a negative I value, then the bit sequence represented in Figure 1-11(a) is "1001."

1.4.3 Quadrature Phase Shift Keying, QPSK

Quadrature Phase Shift Keying (**QPSK**) modulation is usually referred to as quadrature PSK, although it is also referred to as quarternary PSK and quadraphase PSK. In QPSK wireless systems, good spectral efficiency is obtained by sending more than one bit of information per hertz of bandwidth. Information is encoded in four phase states. Thus referring to QPSK as quadraphase shift keying is more precise, but this is not the common usage. The higher-order modulation schemes that achieve more than two states require that the characteristics of the channel be taken into account. The dominant characteristic of the wireless channel are deep fades resulting from destructive interference of multiple reflections. Fades can be viewed as deep amplitude modulation, and so it is difficult to transfer information in the amplitude of a carrier. Consequently phase modulation schemes falling in the class of **M-ary Phase Shift Keying** (**MPSK**) are most appropriate in the mobile context. In mobile environments there are
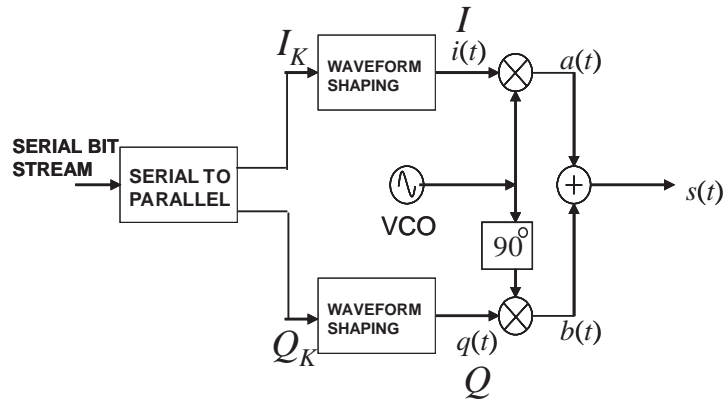
Figure 1-12 Quadrature modulation block diagram indicating the role of pulse shaping.

just a few modulation formats that have been found acceptable. These all fall in the class of either FSK-like schemes or quadraphase shift keying (QPSK) (also called quadrature phase shift keying, as the modulation can be viewed as the superposition of two modulated quadrature carriers). The characteristic of QPSK modulation is that there are four allowable phase states per symbol period, so two bits of information are transmitted per change in the characteristic of the modulated signal. There are many other four-state PSK schemes and there are schemes that have more than four phase states.

Quadrature phase shift keying modulation can be implemented using the quadrature modulator shown in Figure 1-12. The constellation diagram of QPSK is the result of plotting $i(t)$ and $q(t)$ on a rectangular graph (or equivalently $A(t)$ and $\phi(t)$ on a polar plot) for the generalized modulation circuit of Figure 1-12. More commonly these quantities are referred to as I and Q. In Figure 1-12, the input bitstream is first converted into two parallel bitstreams. Thus a two-bit sequence in the serial bitstream becomes one I_K bit and one Q_K bit. The (I_K, Q_K) pair constitutes the Kth symbol. A modulation scheme with four allowable states—A, B, C, D—is shown in Figure 1-11(b). In the absence of wave-shaping circuits, $i(t)$ and $q(t)$ have very sharp transitions, and the paths shown in Figure 1-11 are almost instantaneous. This leads to large spectral spreads in the modulated waveform, $s(t)$. So to limit the spectrum of the RF signal $s(t)$, the shape of $i(t)$ and $q(t)$ is controlled; the waveform is shaped, usually by lowpass filtering. So a pulse-shaping circuit changes binary information into a more smoothly varying signal. Each transition or path in Figure 1-11 represents the transfer of a symbol or minimum piece of information. The best efficiency that can be obtained in point-to-point communication is one

symbol per hertz of bandwidth. In the QPSK modulation scheme shown, there are three possible transitions from each point in the constellation in addition to the possibility of no transition. Thus each symbol contains two bits. So the maximum efficiency of this type of modulation scheme is 2 b·s^{-1}·Hz^{-1} (2 bits/second/hertz of bandwidth). What is actually achieved depends on the pulse-shaping circuits and on the criteria used to establish the bandwidth of $s(t)$. Various modulation schemes have relative merits in terms of spectral efficiency, tolerance to fading (due to destructive interference), carrier recovery, spectral spreading in nonlinear circuitry, and many other issues that are the realm of communication system theorists.

The waveforms corresponding to the state transitions shown in Figure 1-11(b) most immediately affect the bandwidth of $s(t)$ and the ability to demodulate the signal. The constellation diagram is analogous to a phase diagram of the carrier signal.[3] Thus signal trajectories through the origin indicate that the amplitude of the carrier is very small for many RF cycles, and this is particularly troublesome, as it is difficult to track the carrier in the presence of noise. The ability to demodulate signals is equivalent to being able to reconstruct the original constellation diagram of the modulation signal. Also, transitions through the origin indicate that there is significant amplitude variation of the RF signal and so this has high PAR.

EXAMPLE 1. 2	QPSK Modulation and Constellation

The bit sequence 110101001100 is to be transmitted using QPSK modulation. Show the transitions on a constellation diagram.

Solution:

The bit sequence 110101001100 must be converted to a two-bit-wide parallel stream of symbols resulting in the sequence of symbols 11 01 01 00 11 00. The symbol 11 transitions to the symbol 01 and then the symbol 01 and so on. The states (or symbols) and the transitions from one symbol to the next required to send the bitstream 110101001100 are shown in Figure 1-13. QPSK modulation results in the phasor of the carrier transitioning through the origin so that the average power is lower and the PAR is high. A more significant problem is that the phasor will fall below the noise floor, making carrier recovery almost impossible.

1.4.4 Frequency Shift Keying, FSK

Frequency Shift Keying (FSK) is one of the simplest forms of digital modulation, with the frequency of the transmitted signal indicating a

---

[3] This is true here, but not in all cases. The constellation diagram is a way of representing symbols first and is not simply a phasor diagram.
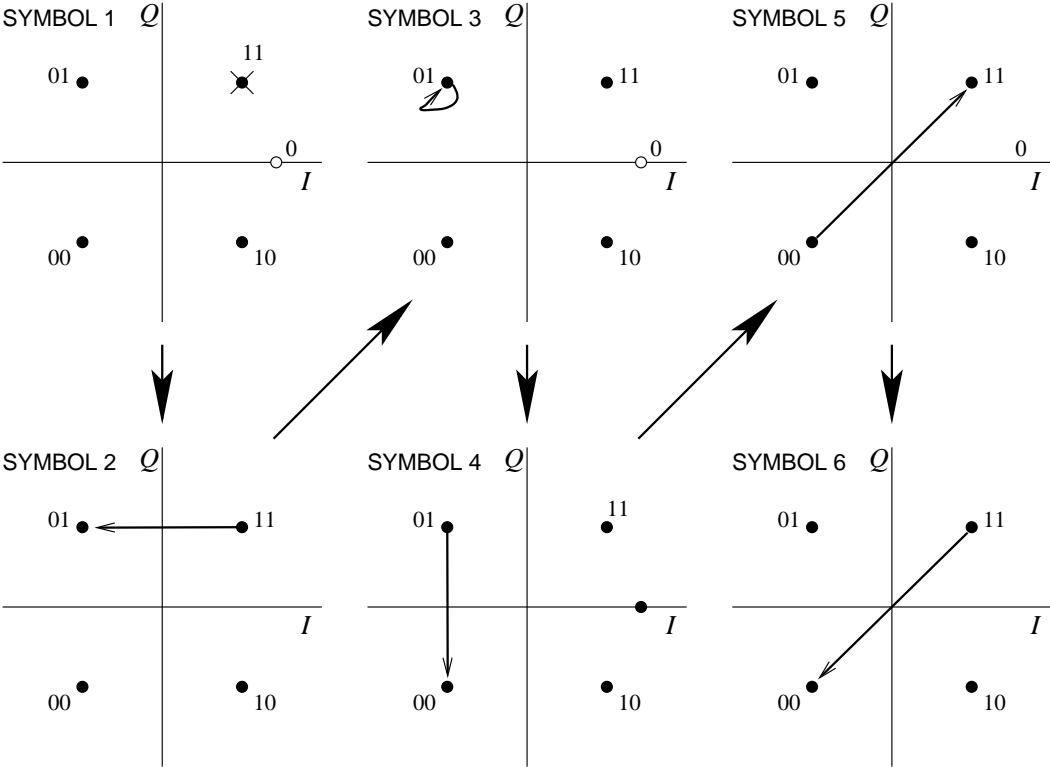
Figure 1-13   Constellation diagram states and transitions for the bit sequence 110101001100 sent as the set of symbols 11 01 01 00 11 00 using QPSK. Note that Symbols 2 and 3 are identical, so there is no transition and this is shown as a self-loop, whereas there will be no transition in going from Symbol 2 to Symbol 3. The SYMBOL numbers indicated reference the symbol at the end of the transition (the end of the arrowhead).

symbol, usually either one or two bits. FSK is illustrated in Figure 1-8(d). The schematic of an FSK modulation system is shown in Figure 1-10. Here, a binary bitstream is lowpass filtered and used to drive an FSK modulator, one implementation of which shifts the frequency of an oscillator according to the voltage of the baseband signal. This function can be achieved using a phase-locked loop (**PLL**) with considerably less sophistication than PSK-based schemes, which require digital signal processing to demodulate a modulated signal. With FSK, an FM demodulator can be used to receive the signal. A characteristic feature of FSK is that the amplitude of the modulated signal is constant, so efficient saturating (and hence nonlinear) amplifiers can be used without introducing much distortion of concern. Not surprisingly, FSK was the first form of digital modulation used in mobile digital radio. Prior to digital radio, FSK was used in analog radio to transmit
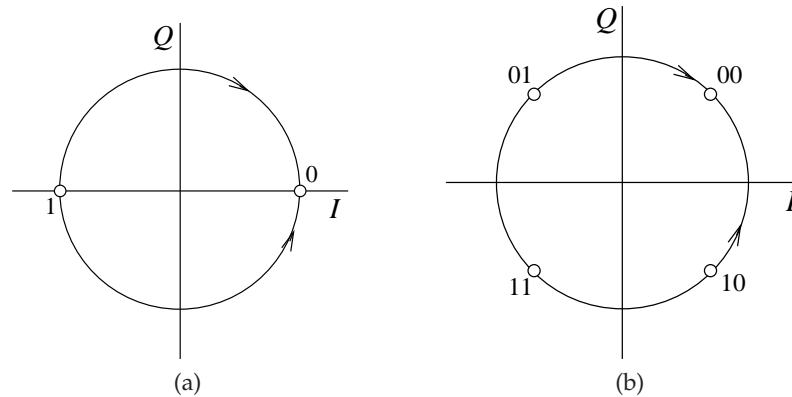
Figure 1-14   Constellation diagrams of FSK modulation: (a) two-state FSK; and (b) four-state FSK.

bits. A particular form of FSK is **Minimum Shift Keying** (**MSK**), which uses a baseband lowpass filter so that the transitions from one state to another are smooth in time and limit the bandwidth of the modulated signal.

In the preceding sections the **constellation diagram** was introduced as a phasor diagram of the (modulated) carrier. However, the equivalence is only approximate and the similarity is most distinct with FSK modulation. Strictly speaking, a phasor diagram describes a phasor that is fixed in frequency. Still, if the phasor is very slowly phase modulated, then this approximation is good. That is, the frequency of the modulated carrier is considered to be fixed and the phase changes over time. FSK modulation cannot be represented on a phasor diagram, as the information is in the frequency transition rather than a phase transition. However, the discrete states must be represented, and the constellation diagram is used to graphically represent them and the transitions. The departure from the phasor diagram can no longer be ignored. In reality, with FSK modulation, the frequency of the modulated carrier changes slowly if the baseband signal is lowpass filtered. For example, consider an FSK modulated signal with a bandwidth of 200 kHz and a carrier at 1 GHz. This is a 0.02% bandwidth, so the phasor changes very slowly. So going from one FSK state to another takes a very long time, about 5000 cycles. In trying to represent FSK modulation on a pseudo-phasor diagram, the frequency is approximated as being fixed and the maximum real frequency shift is arbitrarily taken as being a 180° shift of the phasor.

In FSK, the states are on a circle on the constellation diagram (see Figure 1-14). Note that the constellation diagram indicates that the amplitude of the phasor is constant, as FSK is a form of FM. In four-state FSK modulation (see Figure 1-14(b)), transitions between states take twice as long for states that are on opposite sides of the constellation diagram compared to states

that are only separated by $90°$. Filtering of the baseband modulating signal is required to minimize the bandwidth of the modulated four-state FSK signal. This reduces spectral efficiency to less than the theoretical maximum of 2 bits per hertz. In summary, there are slight inconsistencies and arbitrariness in using a phasor diagram for FSK, but FSK does have a defined constellation diagram which is closely related to a phasor diagram.

1.4.5 Comparison of FSK and QPSK Modulation

The modulation format used impacts the choice of circuitry, battery life, and the tolerance of the system to noise. Figure 1-15 contrasts two types of digital modulation: FSK used in the Global System for Mobile Communications (**GSM**) cellular system, and QPSK used in the **Digital Advanced Mobile Phone System** (**DAMPS**) cellular system. In Figure 1-15(a), f_b is the bit frequency and it is seen that FSK and QPSK have different spectral shapes. Most of the energy is contained within the bandwidth defined by the bit frequency. At multiples of the bit frequency, the power density with FSK is much lower than with QPSK, resulting in less interference (**Adjacent Channel Interference** [**ACI**]) with neighboring radios in adjacent channels. This is an important metric with radios that is captured by the **Adjacent Channel Power Ratio** (**ACPR**), the ratio of the power in the adjacent channel to the power in the main channel. Another important metric is the **Bit Error Rate** (**BER**). Different modulation formats differ in their susceptibility to noise. The level of noise is captured by the ratio of the power in a bit, E_b, to the noise power, N_o, in the time interval of a bit. This ratio, E_b/N_o (often referred to as E B N O), is directly related to the **Signal-to-Noise Ratio** (**SNR**). In particular, consider the plot of the BER against E_b/N_o shown in Figure 1-15(b). QPSK is less susceptible to noise than is FSK.

1.4.6 Implementation Margin

If filters and other hardware in a communication receiver are ideal, E_b/N_o = SNR for BPSK, and $E_b/N_o = \frac{1}{2}$SNR for QPSK. With practical filters there is a performance degradation and a higher SNR is necessary than the theoretical E_b/N_o required for a particular BER. The difference is captured by the implementation margin, k, usually specified in decibels. To achieve a specific BER, SNR is greater than the theoretical E_b/N_o by k. The implementation margin is therefore a measure of the performance of RF hardware.

1.4.7 $\pi/4$ Quadrature Phase Shift Keying, $\pi/4$-QPSK

A major objective in digital modulation is to ensure that the RF trajectory from one phase state to another does not go through the origin. The transition is slow, so that if the trajectory goes through the origin, the
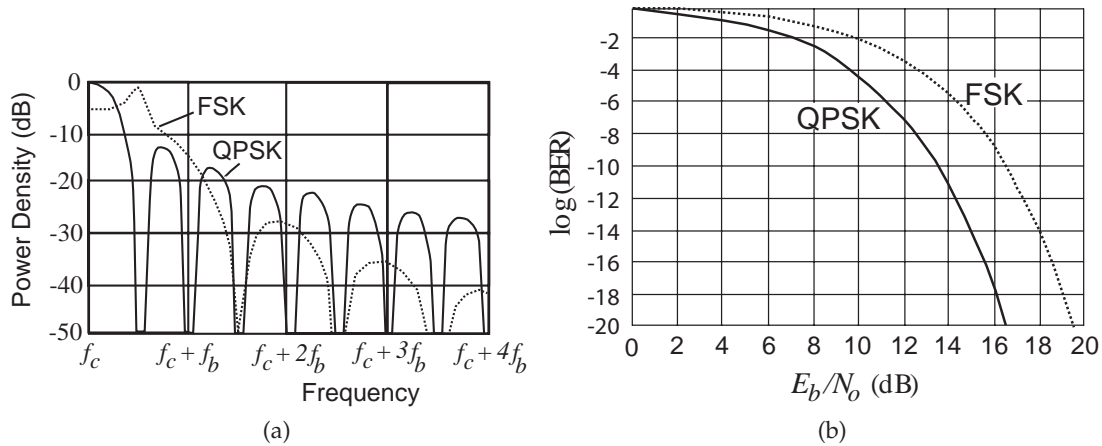
Figure 1-15 Comparison of FSK and QPSK: (a) power spectral density as a function of frequency deviation from the carrier; and (b) BER versus signal-to-noise ratio (SNR) as E_b/N_o (or energy per bit divided by noise per bit).

Table 1-1 Implementation margins for modern communication receivers [4, page 328], [5, 6]. These implementation margins are what can be achieved by the best designs. Not all companies can achieve these. Higher values result from the choice of lower performing technologies, perhaps resulting from a compromise of performance, cost, and design effort.

Modulation	Implementation Margin
BPSK	0.5 dB
QPSK	0.8 dB
8PSK	1–1.6 dB
16QAM	1.5–2.1 dB
CDMAOne	0.5 – 1 dB
WCDMA	2 dB

amplitude of the carrier will be below the noise floor for a considerable time and it will not be possible to recover its frequency. One of the solutions developed to address this problem is the $\pi/4$ quadrature phase shift keying ($\pi/4$-QPSK) modulation scheme. In this scheme the constellation at each symbol is rotated $\pi/4$ radians from the previous symbol, as shown in Figure 1-16.
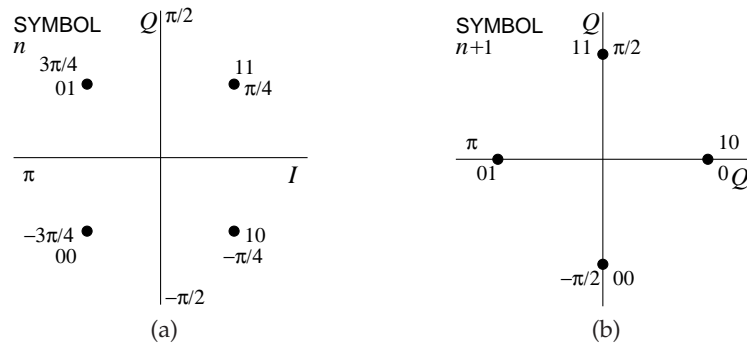
Figure 1-16 Constellation diagram of $\pi/4$-QPSK modulation: (a) constellation diagram at one symbol; and (b) the constellation diagram at the next symbol.

EXAMPLE 1. 3 $\pi/4$-QPSK Modulation and Constellation

The bit sequence 110101001000 is to be transmitted using $\pi/4$-QPSK modulation. Show the transitions on a constellation diagram.

Solution:

The bit sequence 110101001000 must be converted to a two-bit-wide parallel stream of symbols, resulting in the sequence of symbols 11 01 01 00 10 00. The symbol 11 transitions to the symbol 01 and then the symbol 01 and so on. The constellation diagram of $\pi/4$-QPSK modulation really consists of two QPSK constellation diagrams that are shifted by $\pi/4$ radians, as shown in Figure 1-16. At one symbol (or time) the constellation diagram is that shown in Figure 1-16(a) and at the next symbol it is that shown in Figure 1-16(b). The next symbol uses the constellation diagram of Figure 1-16(a) and the process repeats. The states (or symbols) and the transitions from one symbol to the next required to send the bitstream 110101001000 are shown in Figure 1-17.

One of the unique characteristics of $\pi/4$-QPSK modulation is that there is always a change, even if a symbol is repeated. This helps with recovering the carrier frequency, which is an important function in a demodulator. Also, the carrier phasor does not go through the origin and so the PAR is lower than if QPSK modulation were used, as this would result in transitions through the origin. If the binary bitstream itself (with sharp transitions in time) is the modulation signal, then the transition from one symbol to the next occurs instantaneously and hence the modulated signal has a broad spectrum around the carrier frequency. The transition, however, is slower if the bitstream is filtered, and so the bandwidth of the modulated signal will be less. Ideally the transmission of one symbol per hertz would be obtained. However, in $\pi/4$-QPSK modulation the change from one symbol to the next has a variable distance (and so takes different times) so that the ideal spectral efficiency of one symbol per hertz (or 2 bits/Hz) is not obtained. In practice, with realistic filters and allowing for the longer transitions, $\pi/4$-QPSK modulation achieves 1.62 bits/Hz.
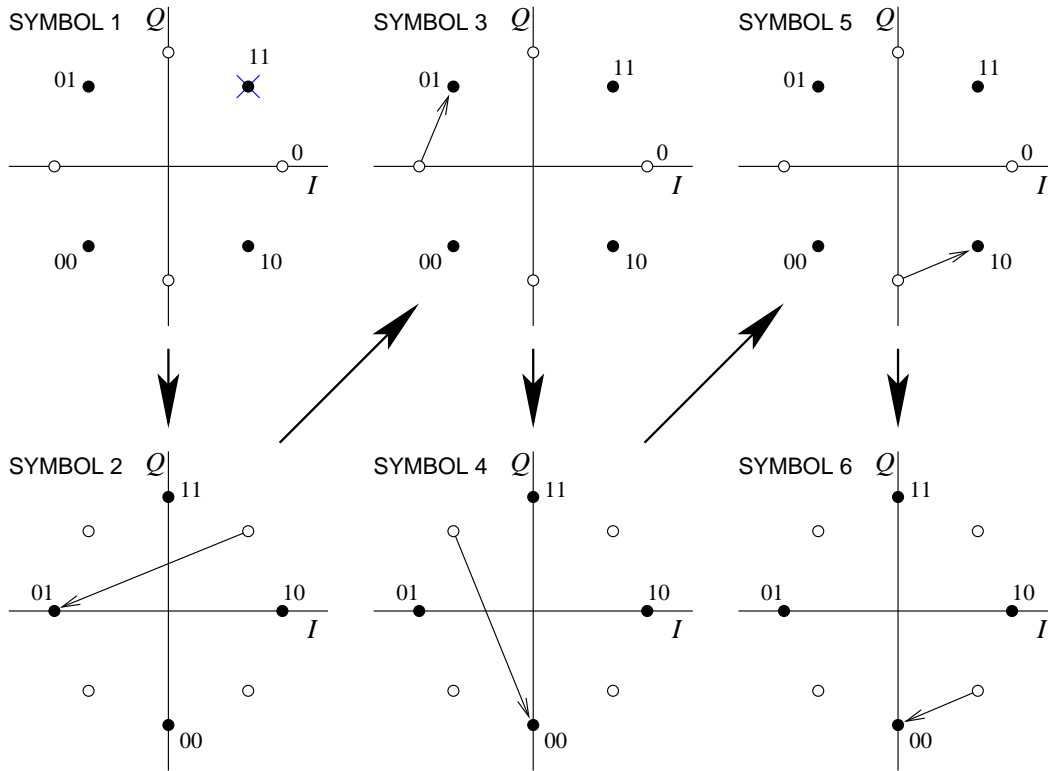
Figure 1-17   Constellation diagram states and transitions for the bit sequence 110101001000 sent as the set of symbols 11 01 01 00 10 00 using $\pi/4$QPSK modulation.

1.4.8   Differential Quadrature Phase Shift Keying, DQPSK

Multiple transmission paths, or multipaths, result in constructive and destructive interference and can result in rapid additional phase rotations. Thus relying on the phase of a phasor at the symbol sample time to determine the symbol transmitted is prone to error. When an error results at one symbol, this error accumulates when subsequent symbols are extracted. The solution is to use encoding, and one of the simplest encoding schemes is differential phase encoding. In this scheme the information of the modulated signal is contained in changes (differences) in phase rather than in the absolute phase.

The $\pi/4$-DQPSK modulation scheme is a differentially encoded form of $\pi/4$-QPSK. The $\pi/4$-DQPSK scheme incorporates the $\pi/4$-QPSK modulator and an encoding scheme, as shown in Figure 1-18(a). The scheme is defined with respect to its constellation diagram, shown in Figure 1-18(b) and repeated in Figure 1-18(c) for clarity. The D indicates **differential coding**,

(a)

(b)   (c)
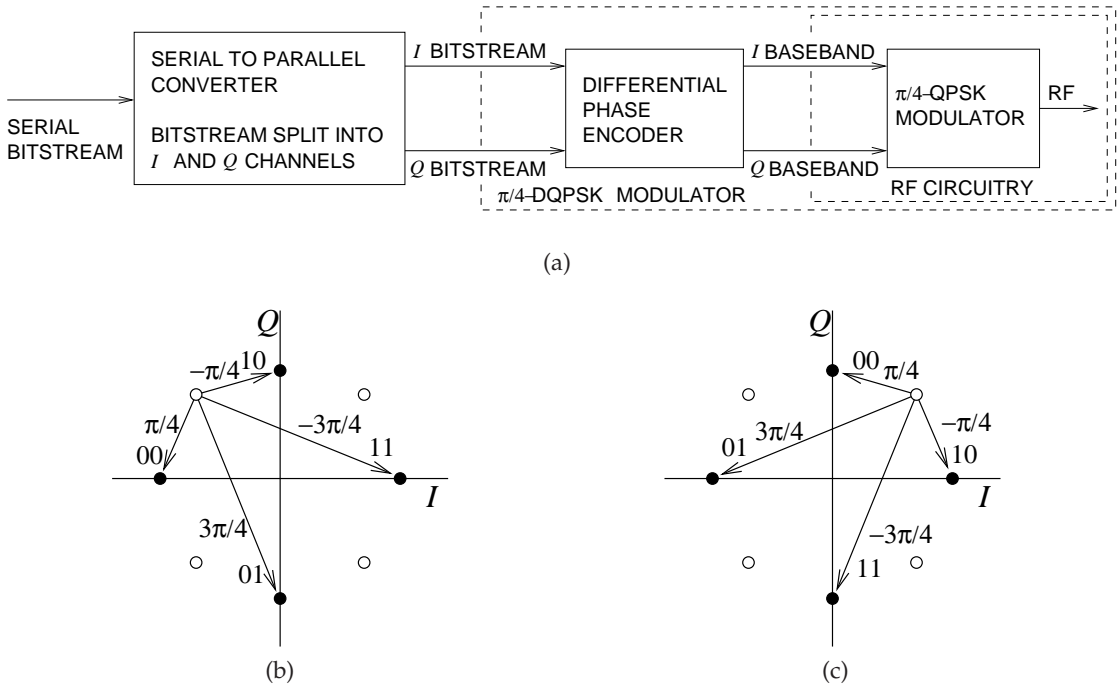
Figure 1-18 A $\pi/4$-DQPSK modulator consisting of (a) a differential phase encoder and a $\pi/4$-QPSK modulator; (b) constellation diagram of $\pi/4$-DQPSK; and (c) a second example clarifying the information is in the phase change rather than the phase state.

Table 1-2 Phase changes in a $\pi/4$-DQPSK modulation scheme.

O_k	E_k	$\Delta\pi$
1	1	$-3\pi/4$
0	1	$3\pi/4$
0	0	$\pi/4$
1	0	$-\pi/4$

while the $\pi/4$ denotes the rotation of the constellation by $\pi/4$ radians or $45°$ from one interval to the next. This can be explained by considering Figure 1-18(a). A four-bit stream is divided into two quadrature **nibbles** of two bits each. These nibbles independently control the I and Q encoding, respectively, so that the allowable transitions rotate according to the last transition. The information or data is in the phase transitions rather than the constellation points themselves. The relationship between the symbol value and the transition is given in Table 1-2. For example, the transitions shown in Figure 1-19 for six successive time intervals describes the input bit sequence
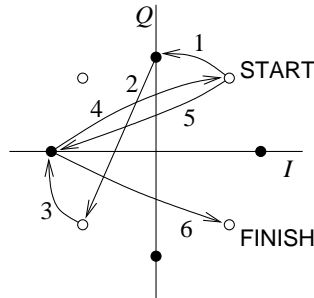
Figure 1-19   Constellation diagram of $\pi/4$-DQPSK modulation showing six symbol intervals coding the bit sequence 000110110101.
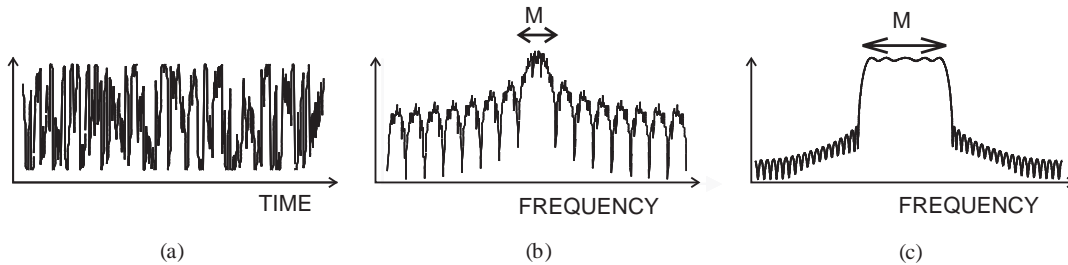
Figure 1-20   Details of digital modulation obtained using differential phase shift keying ($\pi/4$-DQPSK): (a) modulating waveform; (b) spectrum of the modulated carrier, with M denoting the main channel; and (c) details of the spectrum of the modulated carrier focusing on the main channel.

000110110101. Its waveform and spectrum are shown in Figure 1-20. More detail of the spectrum is shown in Figure 1-21. In practice with realistic filters and allowing for the longer transitions, $\pi/4$-DQPSK modulation achieves 1.62 bits/Hz, the same as $\pi/4$-QPSK, but of course with greater resilience to changes in the transmission path.

Sometimes a distinction is made between the transmitted symbols and the encoded symbols. The encoded symbols already have the data represented as transitions from one transmitted symbol to the next. Similar reference is made to received symbols and decoded symbols. The received symbols are the output of the $\pi/4$-QPSK demodulator, while the decoded symbols are the actual data extracted by comparing one received symbol with the previous received symbol. The decoded symbol is extracted in the DSP unit. In a differential scheme, the data transmitted are determined by comparing a symbol with the previously received symbol, so the data are determined from the change in phase of the carrier rather than the actual phase of the carrier. This process of inferring the data actually sent from the received symbols is called decoding. When $\pi/4$-DQPSK encoding was introduced in
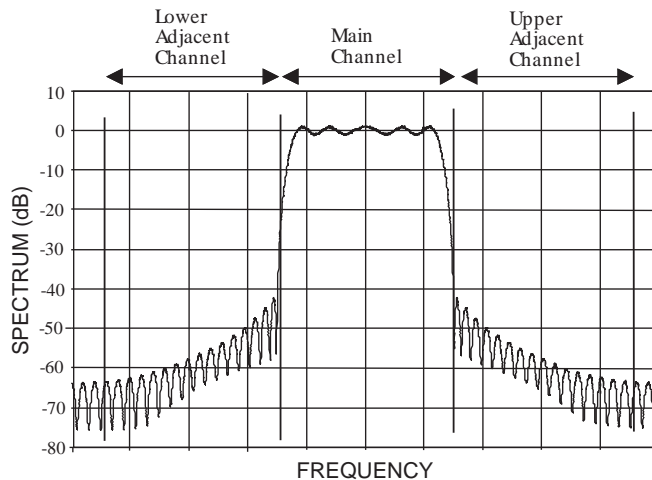
Figure 1-21   Detailed spectrum of a $\pi/4$-DQPSK signal showing the main channel and lower and upper adjacent channels.

the early 1990s the DSP available for a mobile handset had only just reached sufficient complexity. Today, encoding is used with all digital radio systems and is more sophisticated than just the differential scheme of DQPSK. There are new ways to handle carrier phase ambiguity. The sophistication of modern coding schemes is beyond the hardware-centric theme of this book.

1.4.9   Offset Quadrature Shift Keying, OQPSK

The **Offset Quadrature Phase Shift Keying** (**OQPSK**) modulation scheme avoids IQ transitions passing through the origin on the constellation diagram (see Figure 1-23(a)). As in all QPSK schemes, there are two bits per symbol, but now one bit is used to directly modulate the RF signal, whereas the other bit is delayed by half a symbol period, as shown in Figure 1-22. The maximum phase change for a bit transition is $90°$, and as the I_K and Q_K are delayed, a total phase change of approximately $180°$ is possible during one symbol. The constellation diagram is shown in Figure 1-23(a).

The OQPSK modulator can be implemented using relatively simple electronics with a digital delay circuit delaying the Q bit by half a symbol period and lowpass filters shaping the I and Q bits. The OQPSK scheme is also called **Staggered Quadrature Phase Shift Keying** (**SQPSK**). Better performance can be obtained by using DSP to shape the I and Q transitions so that they change smoothly and the phasor trajectory nearly follows a circle. Consequently I and Q change together, but in such a manner that the PAR is maintained close to 0 dB. Two modulation techniques that implement this are the **Shaped Offset QPSK** (**SOQPSK**) and the **Feher QPSK** (**FQPSK**)
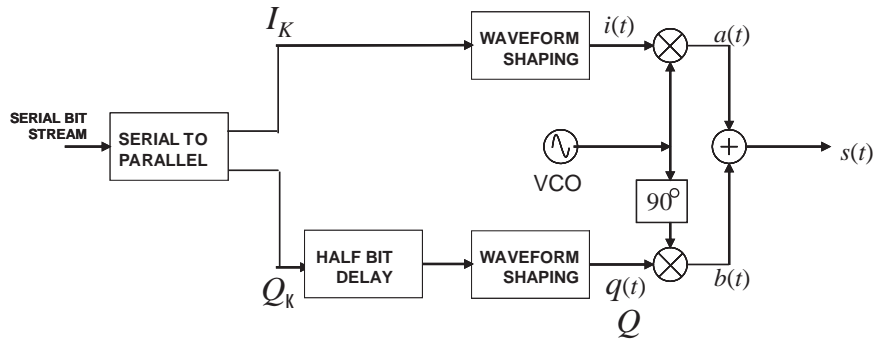
Figure 1-22   Block diagram of an OQPSK modulator.
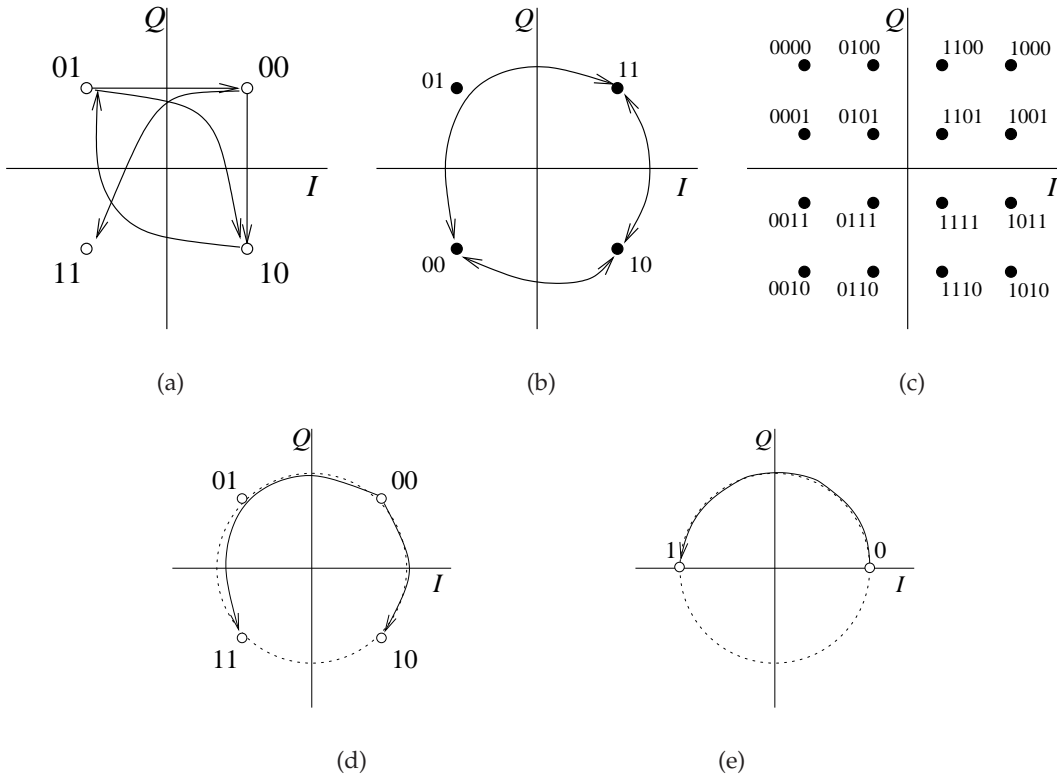
Figure 1-23   Constellation diagrams for various modulation formats: (a) OQPSK; (b) GMSK; (c) 16QAM; (d) SOQPSK (also FOQPSK); and (e) SBPSK.

schemes. The constellation diagrams for SOQPSK and FQPSK are shown in Figure 1-23(d). These are constant envelope digital modulation schemes. As with OQPSK, the Q bit is delayed by one-half of a symbol period. The I and Q baseband signals are shaped by a half-sine filter. The advantage is that saturating amplifier designs can be used and battery life extended. There is a similar modulation format called **Shaped Binary Phase Shift Keying** (**SBPSK**) which, as expected, has two constellation points as shown in Figure 1-23(e). SOQPSK, FOQPSK, and SBPSK are **Continuous Phase Modulation** (CPM) schemes, as the phase never changes abruptly. Instead, the phase changes smoothly, achieving high spectral efficiency and maintaining a constant envelope. Implementation of the receiver, however, is complex. CPM schemes have a good immunity to interference.

EXAMPLE 1. 4   OQPSK Modulation

Draw the constellation diagrams for the bit sequence 010010100110 using OQPSK modulation. With Q delayed by half a bit.

Solution:

The bit sequence is first separated into the parallel stream 01-00-10-10-01-10. The I bit changes first, followed by the Q bit delayed by half of the time of a bit. Five constellation diagrams are required to show the transitions. These are shown in Figure 1-24.

1.4.10 Gaussian Minimum Shift Keying, GMSK

Gaussian Minimum Shift Keying (GMSK) is the modulation scheme used in the **GSM** cellular wireless system and is a variant of **MSK** with waveform shaping coming from a Gaussian lowpass filter. It is a type of FSK modulation and can be implemented with the same hardware, generally using a PLL.

GMSK modulation is also used in the **Digital European Cordless Telephone** (**DECT**) standard. The spectral efficiency of GMSK as implemented in the GSM system (it depends slightly on the Gaussian filter parameters) is 1.35 bits/s/Hz (1.35 b·s^{-1}·Hz^{-1}). Unfiltered MSK has a constant RF envelope, and so the linear amplification requirement is reduced. Filtering is required to limit spectral spreading—in GMSK this results in amplitude variations of about 30%. However, this is still very good, so one of the fundamental advantages of this modulation scheme is that nonlinear, power-efficient amplification can be used. GMSK is essentially a digital implementation of FM with a binary change in the frequency of modulation. The switch from one modulation frequency to the other is timed to occur at zero phase. Put another way, the input bitstream is shaped to form half sinusoids for each
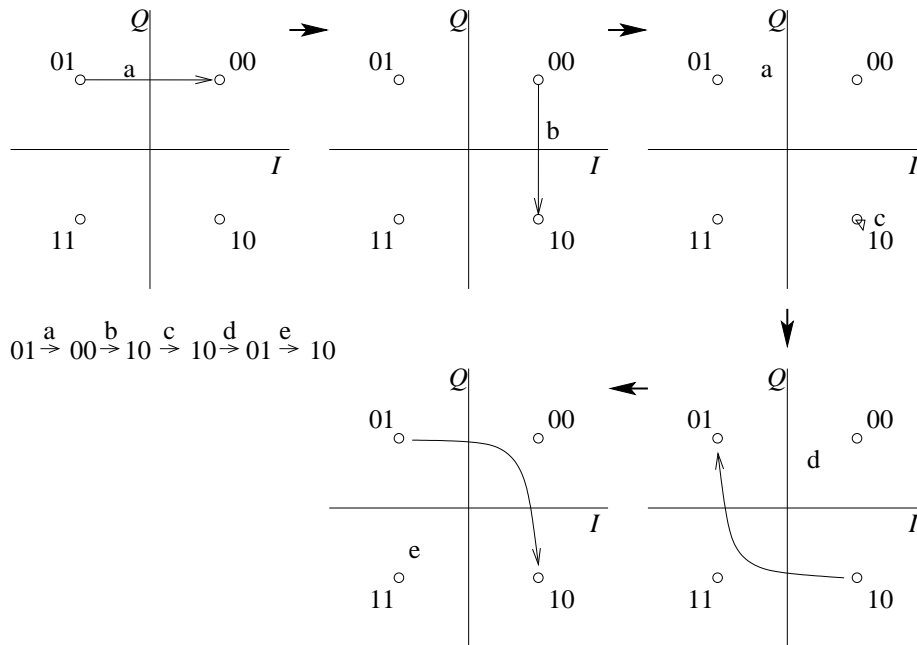
Figure 1-24   Constellation diagram of OQPSK modulation for the bit sequence 010010100110.

bit of the input stream. The phase of the modulating signal is always continuous, but at the zero crossings the half sinusoid continues as a positive or negative half sinusoid depending on the next bit in the input stream. The constellation diagram for GMSK (Figure 1-23(b)) is similar to that for QPSK, but on decoding, the information is not in the phase, but the frequency. So GMSK is an FSK scheme and can be implemented using traditional frequency modulation and demodulation methods.

While QPSK schemes can transmit more data in a given channel bandwidth, GMSK (and other FSK techniques) have the advantage that implementation of the baseband and RF hardware is simpler. A GMSK transmitter can use conventional frequency modulation. On receive, an FM discriminator can be used, avoiding the more complex I and Q demodulation. In GMSK modulation, a data stream is passed through a Gaussian filter and the filtered response drives an FM modulator with the FM deviation set to one-half of the data rate. For example, an 8000 bits/s GMSK data stream is modulated onto an RF carrier with a peak deviation of 4 kHz or ±2 kHz. One type of MSK and GMSK modulator is shown in Figure 1-25.

Most GSM phones input the baseband signal to a PLL to implement frequency modulation. The output of the PLL is input to a power amplifier. This amplifier can be quite efficient, as amplitude distortion is not a concern.
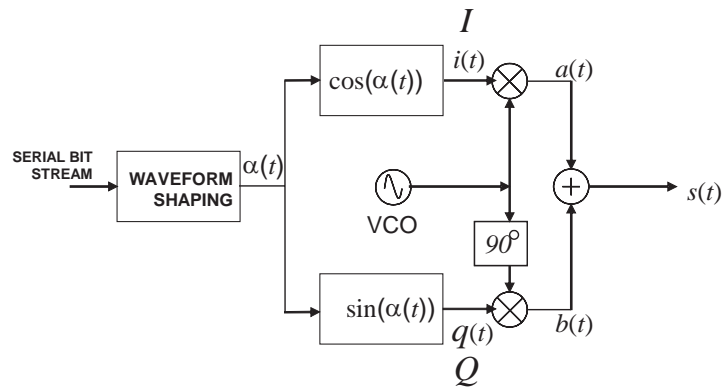
Figure 1-25   Block diagram of an MSK modulator.

1.4.11   3π/8-8PSK, Rotating Eight-State Phase Shift Keying

The $3\pi/8$-8PSK modulation scheme is similar to $\pi/4$-DQPSK in the sense that rotation of the constellation occurs from one time interval to the next. This time, however, the rotation of the constellation from one symbol to the next is $3\pi/8$. This modulation scheme is used in the **Enhanced Data Rates for GSM Evolution** (**EDGE**) system, and provides 3 bits per symbol (ideally) compared to GMSK used in GSM which has 2 bits per symbol. GSM/EDGE provides data transmission of up to 128 kbps, faster than the 48 kbps possible with GSM.

Quadrature modulation schemes with four states, such as QPSK, have two I states and two Q states that can be established by lowpass filtering the I and Q bitstreams. For higher-order modulation schemes such as 8PSK, this approach will not work. Instead, $I(t)$ and $Q(t)$ are established in the DSP unit and then converted using a DAC to generate the analog signals applied to the hardware modulator. Alternatively the modulated signal is created directly in the DSP and a DAC converts this to an IF and a hardware mixer up-converts this to RF. This approach is required in multimode phones supporting multiple standards.

1.4.12   Quadrature Amplitude Modulation, QAM

The digital modulation schemes described so far modulate the phase or frequency of a carrier to convey binary data and the constellation points lie on a circle of constant amplitude. The effect of this is to provide some immunity to amplitude changes to the signal. However, much more information can be transmitted if the amplitude is varied as well as the phase. With sophisticated signal processing it is possible to reliably use **Quadrature Amplitude Modulation** (QAM). In particular, it is necessary to characterize the channel. In wired and **Line-Of-Sight** (**LOS**) systems,

Table 1-3 Spectral efficiencies of various modulation formats. The maximum spectral efficiencies obtained by modulation schemes alone (e.g., BPSK, BFSK, 64-QAM, 256-QAM) result in broad spectra, while actual system implementations achieve less. For example, spectral efficiencies achieved for actual systems are less in an effort to manage bandwidth.

Modulation	bits/s/Hz
BPSK (ideal)	1
BFSK (ideal)	1
QPSK (ideal)	2
GMSK	1.35
$\pi/4$-DQPSK	1.63
$3\pi/8$-8PSK	2.7
64-QAM (ideal)	6
256-QAM (ideal)	8
256-QAM (satellite cable TV)	6.33

the channel changes slowly. However, in non-LOS wireless systems, it is necessary to incorporate a pilot code or pilot signal with the data so that the characteristics of the channel can be continually updated.

A 16-state rectangular QAM constellation is shown in Figure 1-23(c). This constellation can be produced by separately amplitude modulating an I carrier and a Q carrier. Both carriers have the same frequency but are $90°$ out of phase. The two carriers are then combined, with the result that the fixed carrier is suppressed. The most common form of QAM is square QAM, or rectangular QAM with an equal number of I and Q states. The most common forms are 16-QAM, 64-QAM, 128-QAM, and 256-QAM. The constellation points are closer together with high-order QAM and so are more susceptible to noise and other interference. Thus high-order QAM can deliver more data, but less reliably, than can lower-order QAM.

The constellation in QAM can be constructed in many ways, and while rectangular QAM is the most common form, nonrectangular schemes exist; for example, having two PSK schemes at two different amplitude levels. While there are minor advantages to such schemes, square QAM is generally preferred, as it requires simpler modulation and demodulation. The rapid fading in a mobile environment has a bigger impact on amplitude than on phase. As a result, PSK schemes have fewer errors than QAM schemes in mobile use.

1.4.13 Digital Modulation Summary

The spectral efficiencies of various digital modulation schemes are summarized in Table 1-3.

For example, in 1 kHz of bandwidth the $3\pi/8$-8PSK scheme (supported in

third-generation cellular) transmits 2700 bits. Digital transmission requires greater bandwidth than does analog modulation for transmission of the same amount of information that was originally analog (e.g., voice). However, digital modulation is essential for data, and digital modulation is also advantageous for voice. Direct digitization of an audio waveform for high-quality reproduction requires 8 bits of resolution captured at a sample rate of 8000 samples/s for a total 64 kbps (64000 bits/s). The appeal of digital modulation for audio is directly related to the reduction of bit rate accomplished by speech-coding algorithms. Acceptable speech is achieved with bit rates of 3.8 kbps and higher. (The measure of speech quality is purely subjective.) The speech coding algorithms achieve bit rate reduction by utilizing the characteristics of human hearing. There is a lot of redundancy in speech, but this is not used specifically. Human hearing responds to time-varying spectral content, and the unique characteristic is that the statistics of the signal, the autocorrelation functions, and higher-order moments, can be captured at low resolution. The spectrum of the signal can be adequately reconstructed from these statistics.[4] In a typical speech-coding algorithm implemented in a codec, units of 160 samples are characterized by just a few autocorrelation and related parameters. These parameters generally do not require many bits to enable fully intelligible speech to be synthesized, so that a reduction factor of 8 or even 16 in the bit rate is achieved.

The spectral efficiencies shown in Table 1-3 are sometimes less than the ideal. QPSK ideally conveys 2 bits per symbol, but in a communication system, lowpass filtering at baseband is required to constrain the spectrum of the RF-modulated signal. It is clear that an ideal filter cannot be realized, and this, in part, limits the achievable spectral efficiency. A more significant limitation can be understood by considering the constellation diagram of actual schemes, such as those shown in Figure 1-23. Now consider that the constellation diagrams are equivalent to phasor diagrams (which they are in first- and second-generation radio). With the same baseband bandwidth it will take different times for the phasor to make the transition from one symbol to the next; longer transitions require more bandwidth than do shorter transitions. As a result, the spectrum efficiency will be less than the ideal. So in a QPSK-like scheme, 2 bits per symbol are achievable, but in practice, a symbol requires more than the minimum of 1 Hz of bandwidth. Also, with QAM, some of the outlying constellation points at the corners of the constellation cube (high I and Q) are not used.

---

[4] Moments themselves are not transmitted, however. Residuals are transmitted from which the time-varying moments and spectra can be reconstructed. Note that the first moment of a signal is its mean, the second-order moment is the signal's standard deviation, etc.
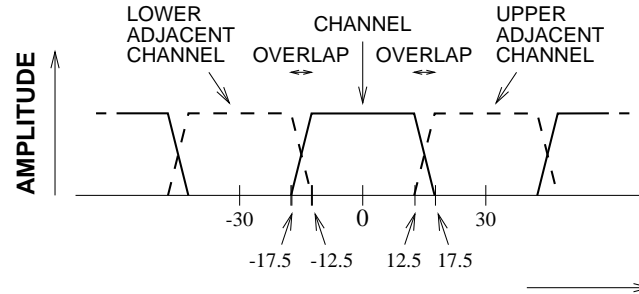
Figure 1-26   Adjacent channels and overlap in the AMPS and DAMPS cellular systems.

1.5   Interference and Distortion

1.5.1   Cochannel Interference

The minimum signal detectable in conventional wireless systems is determined by the **Signal-to-Interference Ratio** (**SIR**) at the input. The noise is due to background noise sources, including galactic noise and thermal noise. In cellular wireless systems, the minimum signal detectable is also determined by the SIR, but now the dominant interference is due to other transmitters in the cell and adjacent cells. The noise that is produced in the signal band from other transmitters operating at the same frequency is called cochannel interference. The level of cochannel interference is dependent on cell placement and frequency reuse patterns. The degree to which cochannel interference can be controlled has a large effect on system capacity.

Control of cochannel interference is largely achieved by controlling the power levels at the base station and at the mobile units. Factors affecting interference are (a) the signal power falling off quickly with distance, and (b) the transmitted power being reduced to the minimum acceptable SIR. Cochannel interference is not a nonlinear affect and is addressed using cell placement.

1.5.2   Adjacent Channel Interference

Adjacent channel interference is the result of several factors. Since ideal filtering cannot be achieved, there is inherent overlap of neighboring channels (Figure 1-26). For this reason, adjacent channels are assigned to different cells. The nonlinear behavior of transmitters also contributes to adjacent channel interference. Thus characterization of nonlinear phenomena is important in RF design. Adjacent channel interference occurs with both digitally-modulated and analog modulated RF signals. It turns out that conventional design approaches can be used to control and predict adjacent channel interference for analog modulated signals, but there is as
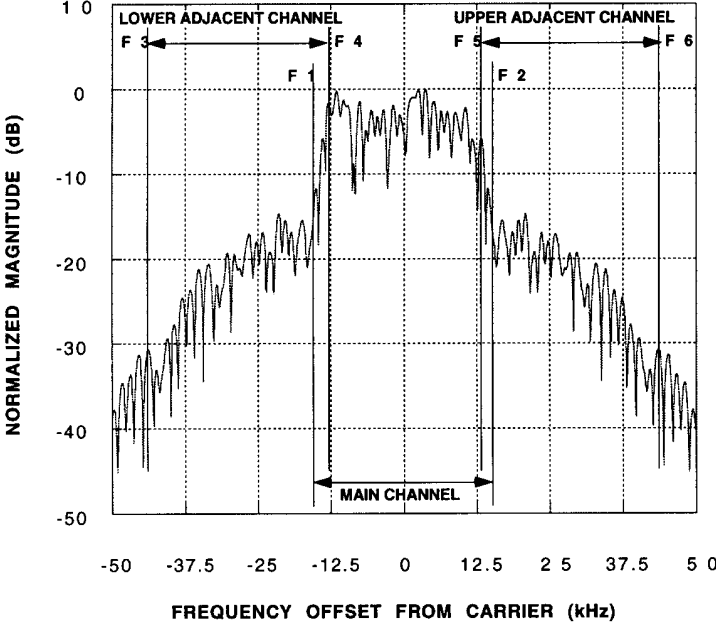
Figure 1-27 Definition of adjacent channel and main channel integration limits using a typical DAMPS spectrum as an example.

yet not a good design practice for digitally-modulated signals.

The spectrum of a DAMPS signal is shown in Figure 1-27. The signal between frequencies f_1 and f_2 is due to the digital modulation scheme and filtering. Most of the signal outside this region is due to nonlinear effects which result in what is called spectral regrowth, a process similar to third- and fifth-order intermodulation in two-tone systems. Using the frequency limits defined in Figure 1-27, the lower channel ACPR is defined as

$$
\begin{aligned}
\text{ACPR}_{\text{ADJ,LOWER}} \quad &= \quad \frac{\text{Power in lower adjacent channel}}{\text{Power in main channel}} \\
&= \quad \frac{\int_{f_3}^{f_4} X(f)df}{\int_{f_1}^{f_2} X(f)df} \; ,
\end{aligned}
\tag{1.20}
$$

where $X(f)$ is the RF signal spectral power density.

1.5.3 Noise, Distortion, and Constellation Diagrams

Noise and nonlinear distortion affect the received constellation diagram and the ability to demodulate signals. Noise is mostly introduced from the environment, particularly from other radios, but the noise introduced by
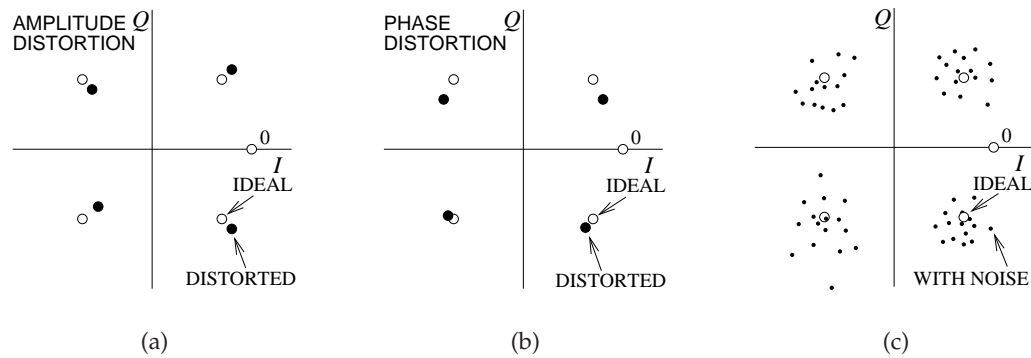
Figure 1-28 Impact of signal impairments on the constellation diagram of QPSK: (a) amplitude distortion; (b) phase distortion; and (c) noise.

the RF hardware itself is significant when the signal received is close to the minimum detectable signal. These distortion effects can be described in part by their effect on constellation diagrams (see Figure 1-28). An additional impact is impairment introduced in adjacent channels. It should be emphasized that the constellation diagram shows the state of the system at the sampling instant that is determined by the recovered clock. Errors in recovering the clock further distort the constellation diagram.

1.5.4 Error Vector Magnitude

The **Error Vector Magnitude** (**EVM**) specifies the accuracy of the waveform at the sampling instances and so is directly related to the bit error rate in a digital modulation scheme. The EVM captures the combined effect of amplifier nonlinearities, amplitude and phase imbalances of separate I and Q signal paths, in-band amplitude ripple (e.g., due to filters), noise, carrier suppression, image rejection, and DAC inaccuracies.

The EVM is a measure of the departure of a sampled phasor from the ideal phasor located at the constellation point (see Figure 1-29(a)). Introducing an error vector, X_{error}, and a reference vector, $X_{\text{reference}}$, which points to the ideal constellation point, the EVM is defined as the ratio of the magnitude of the error vector to the reference vector so that

$$\text{EVM} = \frac{|X_{\text{error}}|}{|X_{\text{reference}}|}. \tag{1.21}$$

Expressing the error and reference in terms of the powers P_{error} and $P_{\text{reference}}$, respectively, enables EVM to be expressed as

$$\text{EVM} = \sqrt{\frac{P_{\text{error}}}{P_{\text{reference}}}} \; ; \tag{1.22}$$
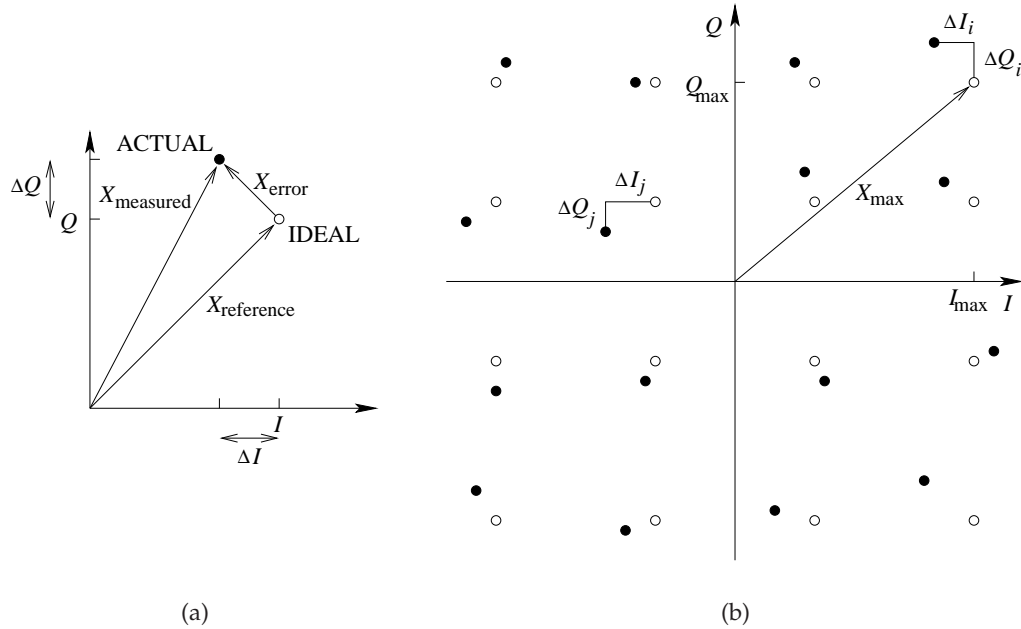
Figure 1-29 Partial constellation diagram showing quantities used in calculating EVM: (a) definition of error and reference signals; and (b) error quantities used when constellation points have different powers.

in decibels,

$$\text{EVM(dB)} = 10\log_{10}\frac{P_{\text{error}}}{P_{\text{reference}}} = 20\log_{10}\frac{|X_{\text{error}}|}{|X_{\text{reference}}|} \; ; \qquad (1.23)$$

or as a percentage,

$$\text{EVM(\%)} = \frac{X_{\text{error}}}{X_{\text{reference}}} \times 100\% \; . \qquad (1.24)$$

If the modulation format results in constellation points having different powers, the constellation point with the highest power is used as the reference and the error at each constellation point is averaged. With reference to Figure 1-29, and with N constellation points,

$$\text{EVM} = \sqrt{\frac{\frac{1}{N}\Sigma_{i=1}^{N}\left(\Delta I_i^2 + \Delta Q_i^2\right)}{X_{\text{max}}^2}} \; , \qquad (1.25)$$

where $|X_{\text{max}}|$ is the magnitude of the reference vector to the most distant constellation point, and ΔI_i and ΔQ_i are the I and Q offsets of the actual constellation point and the ideal constellation point.

A similar measure of signal quality is the **Modulation Error Ratio** (**MER**), a measure of the average signal power to the average error power. In decibels it is defined as

$$\text{MER(dB)} = 10 \log \frac{\frac{1}{N} \Sigma_{i=1}^{N} \left(I_i^2 + Q_i^2 \right)}{\frac{1}{N} \Sigma_{i=1}^{N} \left(\Delta I_i^2 + \Delta Q_i^2 \right)} = 10 \log \frac{\Sigma_{i=1}^{N} \left(I_i^2 + Q_i^2 \right)}{\Sigma_{i=1}^{N} \left(\Delta I_i^2 + \Delta Q_i^2 \right)}.$$

$$(1.26)$$

The advantage of the MER is that it relates directly to the SNR.

Another quantity that is related to both the EVM and MER concepts is the implementation margin, k, described in Section 1.4.5 on Page 23. The implementation margin is a measure of the performance of particular hardware. The required EVM can be estimated from the hardware implementation margin:

$$\text{EVM}_\text{required} = \sqrt{\frac{k}{\text{SNR} \cdot \text{PAR}}} .$$

$$(1.27)$$

All quantities above are absolute. In terms of decibels,

$$\text{EVM}_\text{db, required} = k_\text{dB} - \text{SNR}_\text{dB} - \text{PAR}_\text{dB} .$$

$$(1.28)$$

1.6   Early Receiver Technology

In this section, historical receivers are considered first, in part because the terms associated with the early receivers are still used, but also because the early trade-offs influence the architectures used today. Today receivers use DSP technology, very stable LOs, and sophisticated clock recovery schemes. This was not always so. One of the early problems was using an LO to demodulate a signal when transmitter oscillators drifted by many kilohertz. Radio at first used AM and the carrier was sent with the information-carrying sidebands. With this signal, a simple rectifier circuit connected to a bandpass filter could be used, but the reception was poor. A crystal rectifier consists of a single diode with filters. To improve performance it was necessary to lock an oscillator to the carrier and then amplify the received signal. Here some of the early schemes that addressed some of the problems are discussed. There were many more variants, but the discussion covers the essential ideas.

1.6.1   Heterodyne Receiver

The heterodyning principle mixes a single-tone signal, the LO, with a finite bandwidth signal to produce a lower-frequency version of the information-bearing signal. With the LO frequency set appropriately, the low-frequency signal would be in the audio range. If the information-bearing signal is an AM signal, then the low-frequency version of the signal is the original audio signal, which is the envelope of the AM signal. This type of receiver is
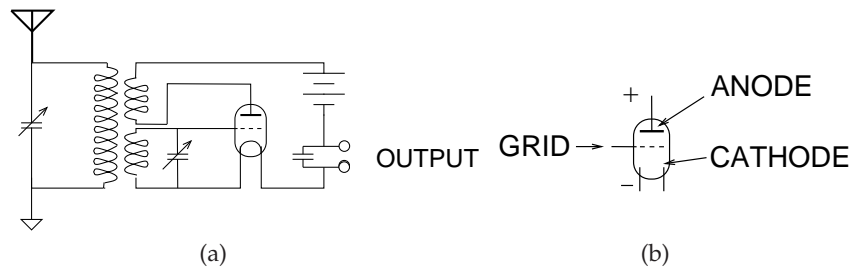
Figure 1-30 The Colebrook's original homodyne receiver: (a) circuit with an antenna, tunable bandpass filter, and triode amplifier; and (b) triode vacuum tube.

called a **tuned radio frequency** (**TRF**) **receiver**, and performance is critically dependent on the stability of the LO and the selectivity of the receive filter. The TRF receiver required the user to adjust a tunable capacitor so that, with a fixed inductor, a tunable bandpass filter was created. Such a filter has a limited Q^5 and a bandwidth that is wider than the bandwidth of the radio channel. Even worse, a user had to adjust both the frequency of the bandpass filter and the frequency of the LO. The initial radios based on this principle were called **audions**, used a triode vacuum tube, and have been in use since 1906. They were an improvement on the **crystal detectors**, but there was a need for something better.

1.6.2 Homodyne Receiver

The homodyne [7], syncrodyne (for synchronous heterodyne) [8], and autodyne (for automatic heterodyne) circuits were the needed improvements on the audion and are based on the regenerative circuit invented by Edwin **Armstrong** in 1912 while he was an electrical engineering student at New York City's Columbia University [9]. Armstrong's circuit fed the input signal into an amplifying circuit and a portion of this signal was coupled back into the input circuit so that the signal was amplified over and over again. This is a positive feedback amplifier. A small input RF signal was amplified to such a large extent that it resulted in the amplifying circuit becoming nonlinear and consequently it rectified the amplitude modulated RF signal. **Colebrook** used this principle and developed the original homodyne receiver shown in Figure 1-30(a). This serves to illustrate the operation of the family of regenerative receivers. The antenna shown on the left-hand side is part of a resonant circuit that is in the feedback path of a triode oscillator. The triode vacuum tube is annotated in Figure 1-30(b). Here the grid coils

---

⁵ Q is the **quality factor** and is the ratio of the energy stored to the energy resistively lost each cycle. Good frequency selectively in a filter requires high-Q components. Tunable components have lower Qs than fixed components.

(which control the flow of carriers between the bottom cathode[6] and top anode) are weakly coupled to the anode circuit. When an AC signal appears at the top anode, the part within the passband of the tuned circuit is fed back to the grid and the signal reinforced. The radio signals of the day were AM and had a relatively large carrier, so the oscillator tended to lock on to the carrier. The AM sidebands were then successfully heterodyned down to the desired audio frequencies.

The **autodyne** worked on a slightly different principle in that the oscillation frequency was tuned to a slightly different frequency from the carrier. Still, the autodyne combined the functions of an oscillator and detector in the same circuit.

1.6.3 *Superheterodyne Receiver*

The superheterodyne receiver was invented by Edwin Armstrong in 1918 [10]. The key concept was to **heterodyne** down in two stages and to use fixed filters and use a tunable LO. The receiving antenna was connected to a bandpass filter that allowed several channels to pass. This relaxed the demands on the receive filter, but also filters with higher selectivity could be constructed if they did not need to be tuned. Today we use high-order filters that are manually or machine tuned, as manufacturing tolerances do not allow high-Q high-order filters to be manufactured unmodified. The filtered received signal is then mixed with an offset LO to produce what is called a **supersonic** signal—a signal above the audio range—and hence the name of this architecture. The performance of the superheterodyne (or super HET) receive architecture has only recently been achieved at cellular frequencies using direct conversion architecture requiring large-scale integrated (and hence silicon) circuits. However, the superheterodyne architecture is still superior above about 6 GHz.

1.7 Modern Transmitter Architectures

Modern transmitters maximize both spectral efficiency and electrical efficiency. Spectral efficiency is achieved by suppressing the carrier on transmit and transmitting a single sideband. The classic technique for achieving this is quadrature modulation, described in the next section. Electrical efficiency must be achieved with tight specifications on allowable distortion and designs must achieve this with minimum manual adjustments. Electrical efficiency has resulted in compound semiconductor transistors, including GaAs HBTs and pHEMTS, mostly preferred for cellular handsets. For base station and point-to-point applications Si LDMOS is the dominant technology below a few gigahertz, with high-

---

[6] The **cathode** is heated (the heater circuit is not shown) and electrons are spontaneously emitted in a process called **thermionic emission**.
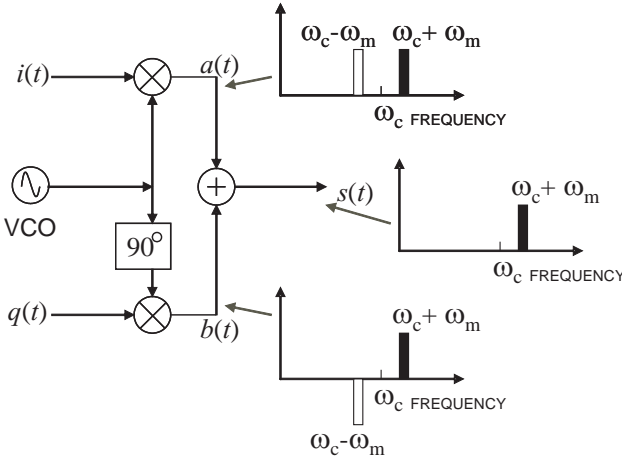
Figure 1-31   Quadrature modulator, showing intermediate spectra.

breakdown gallium nitride (GaN) FETs being introduced. Another trend is the development of universal amplifier concepts so that the same RF front end can be used for a number of different applications. Multifunctional capability is a cost driving transmitter architectures to minimize the RF analog hardware. The discussion here focuses on narrowband communications when the modulated RF carrier can be considered as a slowly varying RF phasor.

1.7.1   Quadrature Modulation

Quadrature modulation describes the frequency conversion process in that the real and imaginary parts of the RF phasor are varied separately. A subsystem that implements quadrature modulation is shown in Figure 1-31. This is quite an ingenious circuit. The operation of this subsystem is described by what is known as the generalized quadrature modulation equation:

$$s(t) = i(t) \cos\left[\omega_c t + \varphi_i(t)\right] + q(t) \sin\left[\omega_c t + \varphi_q(t)\right] . \qquad (1.29)$$

Here, $i(t)$ and $q(t)$ embody the particular modulation rule for amplitude, $\varphi_i(t)$ and $\varphi_q(t)$ embody the particular modulation rule for phase, and ω_c is the carrier radian frequency. In terms of the signals identified in Figure 1-31, the quadrature modulation equation can be written as

$$s(t) = a(t) + b(t) \qquad (1.30)$$
$$a(t) = i(t) \cos\left[\omega_c t + \varphi_i(t)\right] \qquad (1.31)$$
$$b(t) = q(t) \sin\left[\omega_c t + \varphi_q(t)\right] , \qquad (1.32)$$

where $a(t)$ describes the output of the mixer at the top and $b(t)$ describes the output of the mixer on the bottom. The spectrum of $a(t)$ as shown in Figure 1-31 has two bands above and below the frequency of the carrier, ω_c. Similarly the spectrum of $b(t)$ has two bands above and below the frequency of the carrier. However, there is a difference. The LO (here designated as the **voltage-controlled oscillator** (**VCO**) is shifted 90° (perhaps using an RC delay line) so that the frequency components of $b(t)$ have a different phase relationship to the carrier than those of $a(t)$. When $a(t)$ and $b(t)$ are combined, the carrier content is canceled, as is one of the sidebands. This is exactly what is desired: the carrier should not be transmitted, as it contains no information. Also, it is desirable to transmit only one sideband, as it contains all of the information in the modulating signal. This type of modulation is called **suppressed carrier single-sideband** (**SCSS**) modulation. The actual characteristics depend on the particular forms of $i(t)$, $q(t)$, $\varphi_i(t)$, and $\varphi_q(t)$, and these define the modulation schemes. Only a few have the optimum properties of a well-defined spectrum with steep sidewalls so that adjacent channels can be closely packed. In the next section frequency modulation is used to demonstrate SCSS operation. In digital modulation, $i(t)$ and $q(t)$ are each derived from a bitstream, perhaps simply by filtering a binary waveform.

1.7.2 Frequency Modulation

Frequency modulation is considered here to demonstrate SCSS operation. Let $i(t)$ and $q(t)$ be finite bandwidth signals centered at radian frequency ω_m with $(\phi_q(t) - \phi_i(t))$ being 90° on average. This is shown in Figure 1-31, where ω_m represents the frequency components of $i(t)$ and $q(t)$. With reference to Figure 1-31, and setting $\varphi_i(t) = 0 = \varphi_q(t)$,

$$i(t) = \cos(\omega_m t) \quad \text{and} \quad q(t) = -\sin(\omega_m t) \,, \tag{1.33}$$

and the general quadrature modulation equation, Equation (1.29), becomes

$$s(t) = i(t)\cos(\omega_c t) + q(t)\sin(\omega_c t) = a(t) + b(t) \,,$$

where

$$a(t) = \frac{1}{2}\left\{\cos[(\omega_c - \omega_m)t] + \cos[(\omega_c + \omega_m)t]\right\} \tag{1.34}$$

and

$$b(t) = \frac{1}{2}\left\{\cos[(\omega_c + \omega_m)t] - \cos[(\omega_c - \omega_m)t]\right\} \,. \tag{1.35}$$

So the combined frequency modulated signal at the output is

$$s(t) = a(t) + b(t) = \cos[(\omega_c + \omega_m)t] \,, \tag{1.36}$$

and as well as the carrier and lower sideband are suppressed. The lower sideband, $\cos[(\omega_c - \omega_m)t]$, is also referred to as the image. In modulators

$A(t)$

AMPLITUDE
MODULATOR

PA

RF
CARRIER

COMBINER

MODULATED RF
OUTPUT

PHASE
MODULATOR

PA

$\phi(t)$

(a)

ENVELOPE
DETECTOR

$A(t)$ SWITCHING
POWER SUPPLY

MODULATED
RF INPUT

LIMITER

$\phi(t)$

PA

MODULATED RF
OUTPUT

(b)
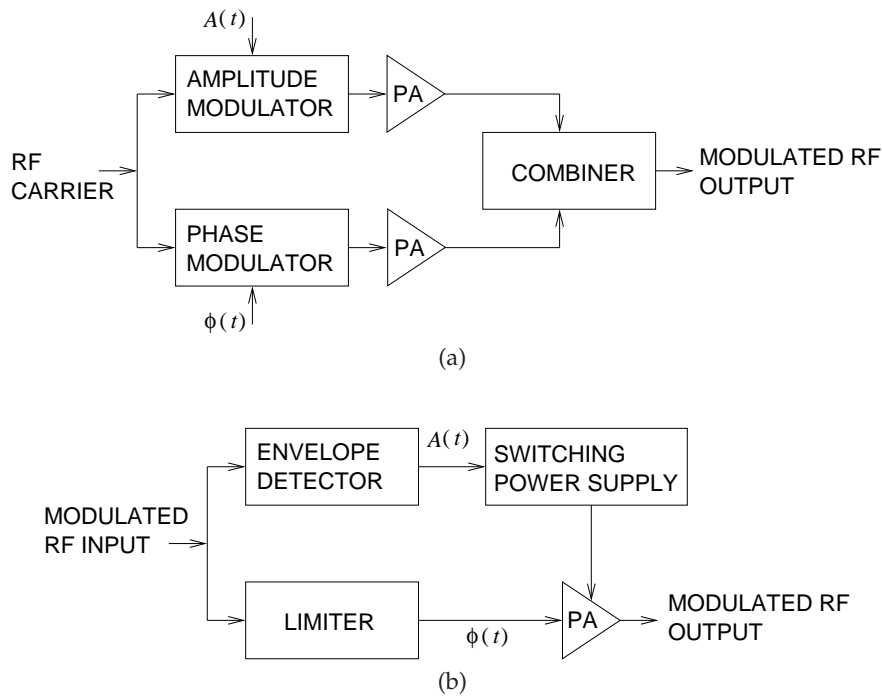
Figure 1-32 Polar modulator architectures: (a) amplitude and phase modulated components amplified separately and combined; and (b) the amplitude used to modulate a power supply driving a saturating amplifier with phase modulated input.

it is important to suppress this image and in demodulators it is important that undesired signals at the image frequency not be converted along with desired signals .

1.7.3 Polar Modulation

Polar modulation is a relatively new modulation scheme for impressing information on a carrier. In polar modulation, the $i(t)$ and $q(t)$ quadrature signals are converted to polar form as amplitude $A(t)$ and phase $\phi(t)$ components. This is either done in the DSP unit or, if a modulated RF carrier is all that is provided, using an envelope detector to extract $A(t)$ and a limiter to extract the phase information corresponding to $\phi(t)$. Two polar modulator architectures are shown in Figure 1-32. In the first architecture, Figure 1-32(a), $A(t)$ and $\phi(t)$ are available and $A(t)$ is used to amplitude modulate the RF carrier, which is then amplified by a **Power Amplifier (PA)**. The phase signal, $\phi(t)$, is the input to a phase modulator implemented as a PLL. The output of the PLL is fed to an efficient (i.e., nonlinear) amplifier
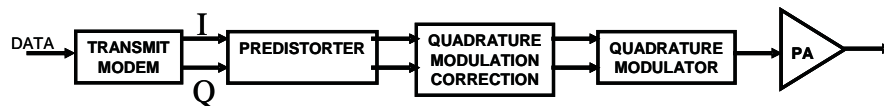
Figure 1-33   Architecture of a direct conversion transmitter.

operating near saturation (also called a **saturating amplifier**). The outputs of the two amplifiers are combined to obtain the large modulated RF signal to be transmitted.

In the second polar modulation architecture, Figure 1-32(b), a low-power modulated RF signal is decomposed into its amplitude and phase modulated components. The phase component, $\phi(t)$, is extracted using a limiter which produces a pulse-like waveform with the same zero crossings as the modulated RF signal. Thus the phase of the RF signal is captured. This is then fed to a saturating amplifier whose gain is controlled by the carrier envelope, or $A(t)$. Specifically, $A(t)$ is extracted using an envelope detector, with a simple implementation being a rectifier followed by a lowpass filter with a corner frequency equal to the bandwidth of the modulation. $A(t)$ then drives a switching (and hence efficient) power supply that drives the **saturating power amplifier**. Polar modulation is finding application in the **Universal Mobile Telecommunications System** (**UMTS**), as the 8PSK modulation used in UMTS results in low efficiency if direct power amplification of a modulated carrier is used.

1.7.4 Direct Conversion Modulation

The transmitted signal has a controlled spectral content, being just a single channel. Issues such as image rejection and interferers are not problems. A direct conversion transmitter generates the RF signal directly without an IF stage using the architecture shown in Figure 1-33. Here, the transmit modem first produces I and Q baseband signals from the data. This is then translated directly to RF via a quadrature modulator and then amplified by a PA. Practically, the nonlinearities of the PA must be linearized using predistortion, and quadrature modulation errors must be accounted for in a quadrature modulation corrector. The transmit modem, the predistorter, and the quadrature modulation corrector can be combined in a DSP, so considerable complexity is shifted to the DSP chip. Errors are associated with mismatches of the ADC and of the analog circuit paths of the separate I and Q paths. Noise-shaping techniques implemented in a DSP have been developed to shift noise outside the bandwidth of the generated signal.

1.8 Modern Receiver Architectures

Communication receivers most commonly use mixing of the RF signal with a fixed signal called an LO to produce a lower-frequency replica of the modulated RF signal. Some receiver architectures use one stage of mixing, while others use two stages of mixing. In cellular systems, the receiver must be sensitive enough to detect signals of 1 pW or less. Some of the architectures used in modern receivers are shown in Figure 1-34. Figure 1-34(a) is the superheterodyne architecture in much the same form that it has been used for almost a century. Key features of this architecture are that there are two stages of mixing, and filtering is required to suppress spurious mixing products. Each mixing stage has its own VCO. The receiver progressively reduces the frequency of the information-bearing signal. The image rejection mixer in the dashed box achieves rejection of the image frequency to produce an IF (or baseband frequency) that can be directly sampled. It is, however, difficult to achieve the required amplitude and phase balance, especially when the image reject filter is realized on an IC. Instead, the architecture shown in Figure 1-34(b) is used. The filter between the two mixers can be quite large. For example, if the incoming signal is 1 GHz, the frequency of the signal after the first mixer could be 100 MHz.

Filters are smaller and have higher performance at higher frequencies. This is exploited in the dual-conversion receiver shown in Figure 1-34(c). This is similar to the traditional superheterodyne architecture except that the IF between the two mixers is high. In the previous example it could be 3 GHz. This architecture also enables broad radio operation with the band selected by choosing the frequencies of the two local oscillators.

The Low-IF or Zero-IF receiver shown in Figure 1-34(d) uses less hardware and is common in less demanding communication applications. In high-performance systems, such as the cellular phone system, this archiecture requires more design time and generally calibration circuitry to trim the I and Q paths so that they are closely matched.

1.8.1 Homodyne Frequency Conversion

Homodyne mixing and detection is one of the earliest wireless receiver technologies and is used in AM radio. Homodyne mixing can be used for detecting modulation formats other than AM, including digitally-modulated signals, and is a particularly attractive architecture for monolithically integrated circuits. In homodyne mixing, the carrier of a modulated signal is regenerated and synchronized in phase with the incoming carrier frequency. Mixing the carrier with the RF signal results in an IF signal centered around zero frequency. The only simple way to ensure that the LO or **pump** is in phase is to transmit the carrier with the RF signal. AM transmission does just this, but at the cost of transmitting large carrier power, as well as the additional prospect of interference that goes along with

(a) SUPERHETERODYNE, HARTLEY IMAGE REJECTION

(b) SUPERHETERODYNE RECEIVER

(c) DUAL CONVERSION RECEIVER

(d) LOW–IF OR ZERO–IF RECEIVER
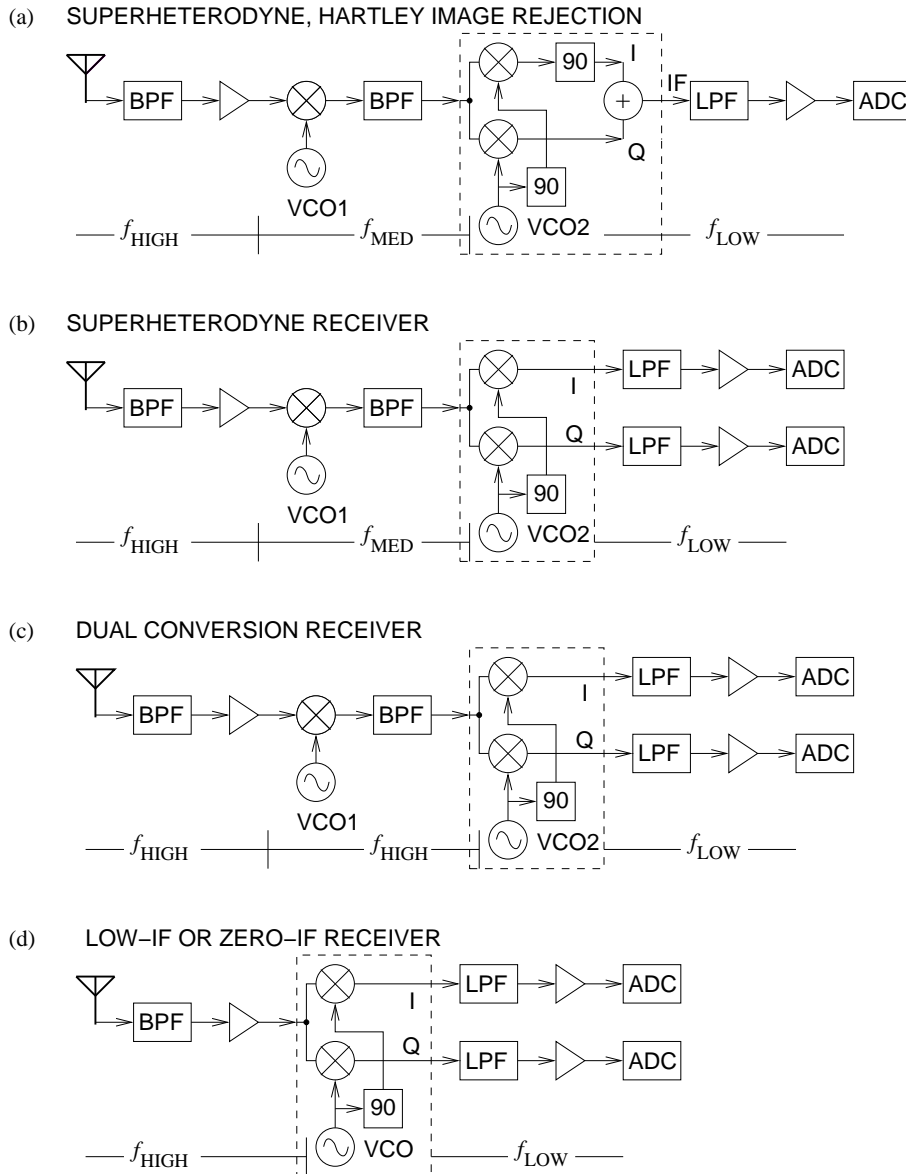
Figure 1-34 Architecture of modern receivers: (a) superheterodyne receiver using the **Hartley** architecture for image rejection; (b) superheterodyne receiver; (c) dual-conversion receiver; low-IF or zero-IF receiver. PBF, bandpass filter; LPF, lowpass filter; ADC, analog-to-digital converter; VCO, voltage-controlled oscillator; 90, 90° phase shifter; I, in-phase component, Q, quadrature component; f_{HIGH}, f_{MED}, and f_{LOW} indicate relatively high, medium, and low frequencies in the corresponding section of the receiver.
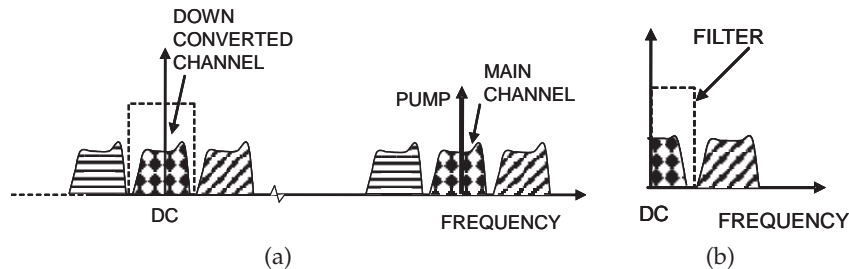
Figure 1-35 Frequency conversion using homodyne mixing: (a) the spectrum with a large LO or pump; and (b) the baseband spectrum showing only positive frequencies.

this. Homodyne mixing can be used with digitally-modulated signals.

Signal spectra that result in homodyne mixing are shown in Figure 1-35. In Figure 1-35(a), the RF signals are shown on the right-hand side and the baseband signals are shown on the left-hand side. It is usual to show both positive and negative frequencies at the lower frequencies so that the conversion process is more easily illustrated. The characteristic of homodyne mixing is that the pump corresponds to the carrier and is in the middle of the desired RF channel. RF signal components mix with the pump, and it appears that the entire RF spectrum is down-shifted around DC. Of course, the actual baseband spectrum (the term for the lowest-frequency signals) is only defined for positive frequencies, so the negative frequency baseband signals and the positive frequency baseband signals add to yield the baseband spectrum shown in Figure 1-35(c). With other modulation schemes, this loss of information is avoided using quadrature demodulation. An amplitude modulated signal has identical modulation sidebands, so the collapsing of positive and negative frequencies at baseband results in no loss of information. Then simple amplitude detection circuitry, such as a rectifier, is used and the rectified signal is passed directly to a speaker.

1.8.2 *Heterodyne Frequency Conversion*

In heterodyne mixing, the pump and the main RF channel are separated in frequency, as shown in Figure 1-36. In this figure the RF signals (shown as three discrete channels on the right-hand side of the spectrum) mix with the pump signal to produce signals at a lower frequency. This lower frequency is usually not the final baseband frequency desired, and so is called the IF. The intermediate frequency of the main channel is at the difference frequency of the RF signal and the pump. The pump must be locally generated, and so is called the LO signal. There are several important refinements to this. The first of these is concerned with limiting the number of signals that can mix
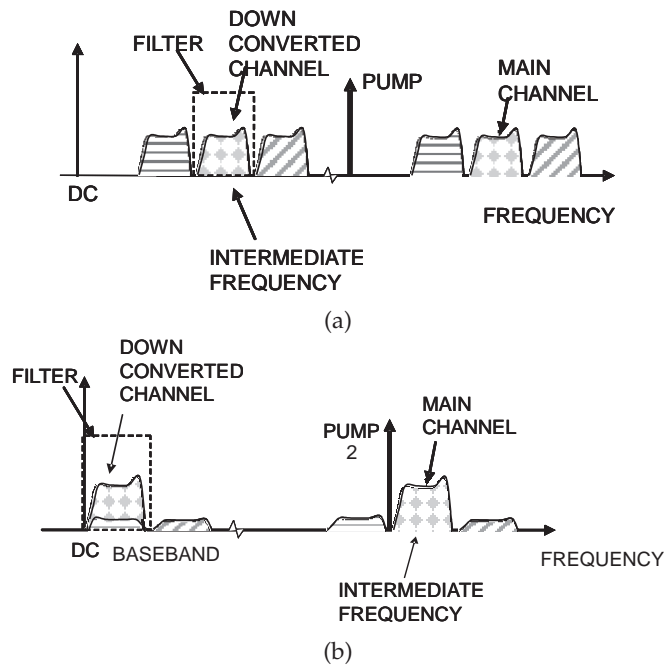
Figure 1-36 Frequency conversion using superheterodyne mixing: (a) the pump is offset from the main channel producing a down-converted channel at the IF; and (b) a second pump down-converts the main channel, now at the IF, to the baseband frequency.

with the pump signal. This is done using an RF preselect filter, resulting in the spectrum shown in Figure 1-37(b). The important characteristic is that the signals at frequencies below the pump have been suppressed, however, the image channel is still of concern. To see the difficulties introduced by the image channel, consider the frequency conversion to baseband process described in Figure 1-38. The RF spectrum after RF **preselect filtering** is shown in Figure 1-38(a) and the baseband (or IF) spectrum is shown in Figure 1-38(b). Again, positive and negative frequencies are used to better illustrate the down-conversion process. Note the down-converted image and the main channel are equidistant from DC. So referring the signals to a positive-frequency-only spectrum (Figure 1-38(c)), it can be seen that the image channel interferes with the main channel. In the worst-case scenario, the IF image could be larger than that of the desired channel. Fortunately there is a circuit fix that compensates for this.

The block diagrams of the circuits that correct this image problem are shown in Figures 1-39(a) and 1-39(b) for both transmit and receive functions. For now, consider the **quadrature receiver** shown in Figure 1-39(a). Shown here is an antenna that takes in the RF signal, and the RF preselect function

(a)

(b)
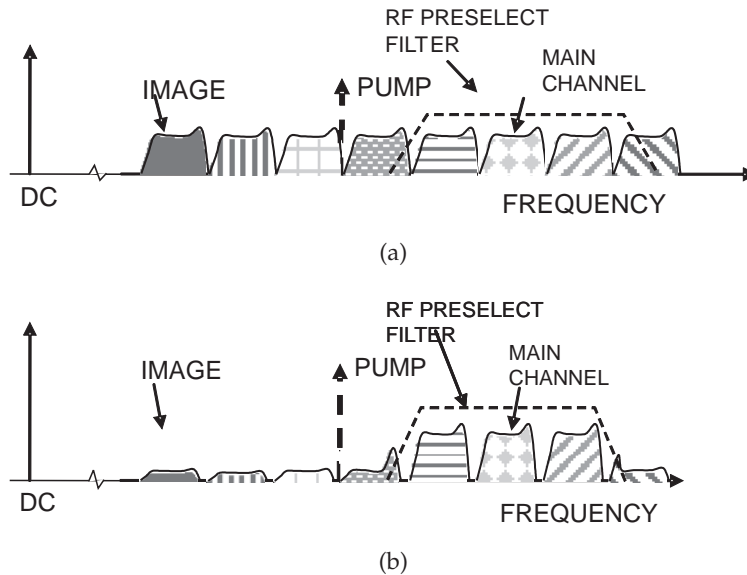
Figure 1-37 Frequency conversion using heterodyne mixing showing the use of an RF preselect filter to reduce the image signal.

is performed by a BPF. This signal is initially amplified (usually), split, and applied equally to two mixers. In an ideal mixer, the two signals are multiplied together. The mixers have pump (LO) signals that are $90°$ out of phase (i.e., in quadrature) with each other and contain different information. The phase relationships at the outputs of the two mixers have very particular relationships with each other such that when I and Q are added the image is eliminated, as shown in Figure 1-39(e). With digitally-modulated signals, however, the I and Q waveforms are individually sampled by ADCs.

Figure 1-39(c) shows the baseband spectrum at the I output of the heterodyne receiver. Figure 1-39(d) shows the baseband spectrum at the Q output of the heterodyne receiver. It also shows the negative frequency components inverted. This is a shorthand way of saying that the negative (image) frequency components are $180°$ out of phase with the down-converted image components of the in-phase signal. However, the down-converted main channel and neighbors are in phase in both spectra. Thus the combination of the I and Q outputs suppresses the image signals from baseband. Figure 1-39(e) illustrates the positive spectrum following the summation of the I and Q channels at the output of the heterodyne receiver. Image rejection is key to all commonly used wireless communications systems today.

The **quadrature transmitter** shown in Figure 1-39(b) works almost

(a)

(b)

(c)
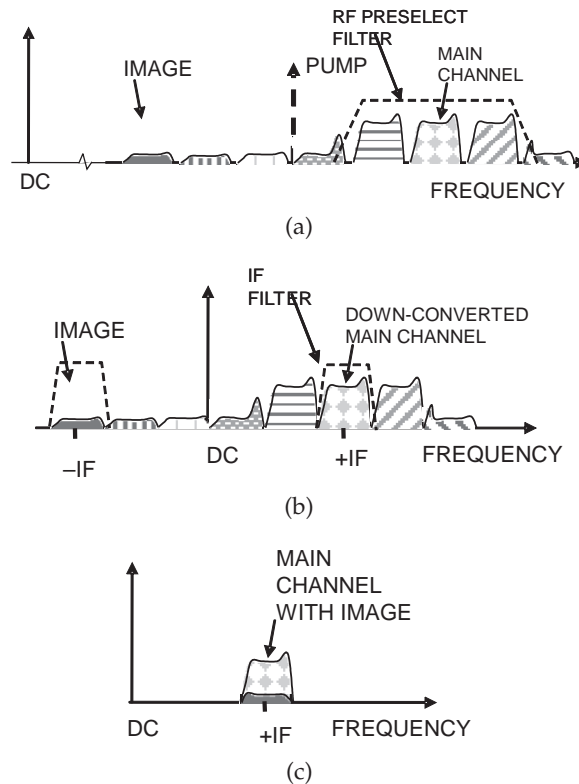
Figure 1-38 Frequency conversion using heterodyne mixing showing the effect of image distortion: (a) the RF spectrum following filtering using an RF preselect filter; (b) the baseband down-converted signal showing positive and negative frequencies; and (c) the single-sided baseband spectrum following IF filtering showing the contamination of the final signal by the image signal.

identically, but in reverse. The I and Q waveforms are each applied to mixers, with one pumped by an LO and the other by the LO shifted by $90°$. The outputs of each mixer contain upper and lower sidebands on either side of the LO (or carrier) frequency. When these are combined at the summing node, one of the sidebands (called the image) is canceled and only a single carrier is presented to the BPF and then radiated by the antenna. In QPSK digitally-modulated systems, the I and Q waveforms are lowpass filtered digital bitstreams.

Heterodyne systems implement signal processing such as filtering, modulation and demodulation, and image rejection at RF and IF using hard-to-integrate discrete components, leading to expense and limitations on size reductions. Heterodyne architectures are regarded as approaching their limit in size, integration, and fabrication cost. The primary issue in mixer
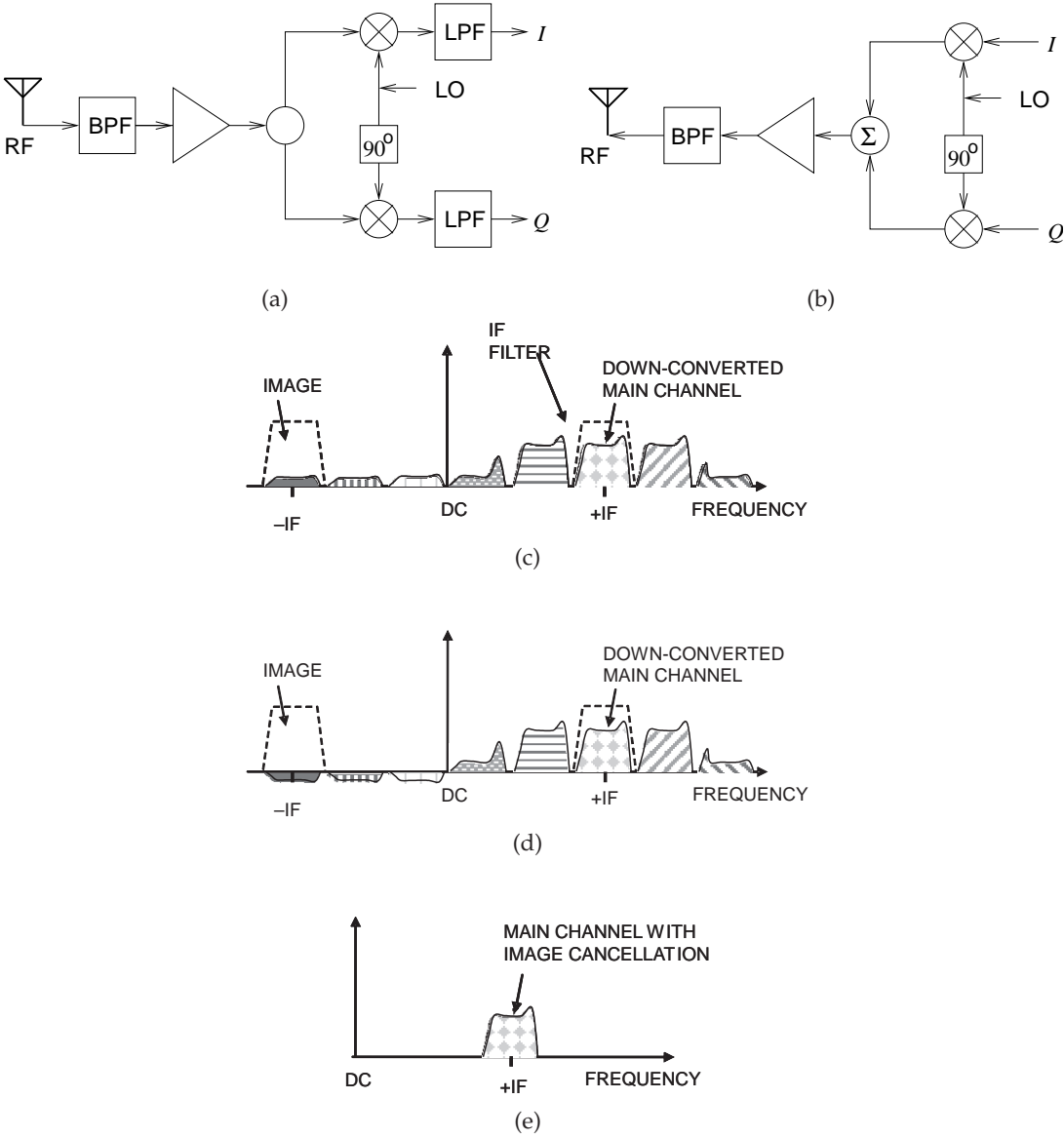
Figure 1-39 Quadrature mixing: (a) receive modulator; and (b) transmit modulator. Frequency conversion using heterodyne mixing and quadrature mixing: (c) the baseband spectrum at the I output of the heterodyne receiver; (d) the baseband spectrum at the Q output; and (e) the positive spectrum following the summation of the I and Q channels at the output of the heterodyne receiver.
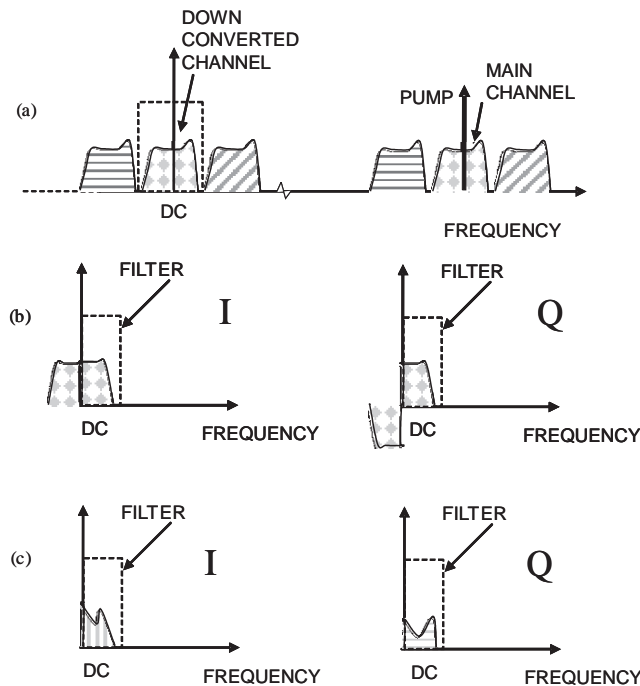
Figure 1-40 Frequency conversion using homodyne mixing and quadrature mixing: (a) the baseband spectrum at the I output of the homodyne receiver; (b) the baseband spectrum at the Q output of the homodyne receiver; and (c) the positive spectrum following the summation of the I and Q channels at the output of the homodyne receiver.

design is limited image rejection resulting from gain and phase mismatches of the I (in-phase) and Q (quadrature) paths.

1.8.3   Direct Conversion Receiver

Zero-IF direct conversion receivers are similar to quadrature homodyne receivers with an LO signal placed at the center of the RF channel. The difference is that in homodyne receivers, the phase of the carrier (i.e., the phase of the LO) is precisely known, as the carrier is transmitted with the signal. In virtually all RF transmission schemes (above a few megahertz) the carrier is not transmitted. Thus in zero-IF direct conversion schemes, the LO signal has inherent phase error with the original carrier. The important characteristic is that there is only one level of mixing. The conversion process is described in Figure 1-40. A particular advantage of direct conversion is that the relatively large IF filters are eliminated. Thus the I and Q mixer outputs are necessary, as the two sides of the RF spectrum contain different

information, and there would be irreversible corruption if a scheme was not available to extract the information in each of the sidebands.

The main nonideality of this design is the DC offset in the down-converted spectrum. DC offset results mostly from self-mixing, or rectification, of the LO. This DC offset can be much larger than the down-converted signal itself, and because of the nonlinearities of baseband amplification stages, either severely limits the dynamic range of the receiver or places limitations on the modulation format that can be used. One way of coping with the DC offset is to highpass filter the down-converted signal, but highpass filtering requires a large passive component (e.g., a series capacitor), at least to avoid dynamic range problems with active filters. Highpass filtering the down-converted signal necessarily throws away information in the signal spectrum, and it is only satisfactory to do this if there is very little information around DC to begin with.

The primary design effort with zero-IF converters is overcoming the DC offset problem, and to a lesser extent coping with jitter of the LO. The primary LO noise of concern is close-in phase noise, which can be at appreciable levels 100 kHz from the carrier. This noise is commonly referred to as **flicker noise**, and increases rapidly as the offset from the carrier is reduced. This is of concern in all conversion processes. However, one of the properties of heterodyne mixing is that the RF signal is considerably offset from the large phase noise region. Consequently the LO phase noise at the frequency of the RF reduces the impact on the resulting offset IF signal. For these reasons, heterodyne mixing provides higher performance than direct conversion. Also, direct conversion receivers are difficult to implement for 8-state (8PSK etc.) and higher-order modulation.

In cellular wireless, the radio signals are spectrally efficient and the spectrum is fairly constant across the channel. So the near-DC signals that result from direct conversion have appreciable information content and cannot be discarded so easily without significant distortion.

The main problems of zero-IF conversion in cellular radio applications include the following:

1. Spurious LO leakage. Retransmission of the LO is possible because the LO is tuned precisely to the RF signal frequency and reverse leakage through the RF path will radiate from the antenna. Spurious LO transmission is severely regulated. The limit on this in-band LO radiation is between -50 and -80 dBm. The problem is reduced by using differential LOs and using multiple RF amplifier stages to increase the reverse isolation between the mixer and the antenna.

2. Interferer leakage. A large RF interferer can leak through the RF amplifier and enter the mixer through both the LO port as well as the RF port. Mixing of these components results in DC offset.

3. Distortion. Direct conversion receivers are more sensitive to undesired

signals than are heterodyne receivers. The nonlinearities of the input mixers will rectify strong spurious RF signals to produce output components around DC. This is the result of second-order nonlinearities, and this effect can be suppressed through the use of balanced RF circuits, which will have only odd-order nonlinearities. This reduces the baseband signals that can result from large RF interferers, but it is still possible to produce baseband distortions if the RF interferer is large enough to produce third harmonics in the mixer. The effect of this distortion can be largely eliminated through the use of highpass filtering at the baseband, but this is not acceptable for cellular radio signals. Heterodyne conversion is susceptible to distortions resulting from odd-order nonlinearities, but in zero-IF converters, second-order distortion is also a problem.

4. LO self-mixing. Mixing of the LO with itself will produce a DC signal in the mixer output. This DC level may be many orders of magnitude larger than the baseband signal itself, so it can desensitize or saturate the baseband amplifier. DC offsets can also result from circuit mismatch problems. The DC offset that results primarily from LO self-mixing is the most significant problem in the use of zero-IF architectures in cellular wireless. The DC offset can be reduced through the use of balanced designs, but circuit mismatch errors still result in very large DC offsets.

5. LO frequency error. A difference between the LO and the carrier will cause the RF signal to be asymmetrically converted around DC.

6. Second-order distortion. Because of second-order distortion, second harmonics of the signal can appear in the baseband. This is a problem if the RF signal is large to begin with. This problem can be circumvented by using designs that utilize differential signals.

7. I/Q mismatches. Mismatches of the I and Q paths also result in DC offsets. These offsets, however, vary negligibly with time, and analog or digital calibration techniques can be used to remove their effect.

The problem of DC offset is made worse because the DC level can vary with time as the amplitude of the interferer varies, or the LO that leaks from the antenna reflects off moving objects and is received as a time-varying interferer itself.

1.8.4 Low-IF Receiver

In a low-IF receiver, single-stage heterodyne mixing is used to down-convert the modulated RF carrier to a frequency just above DC, perhaps to a few hundred kilohertz or a few megahertz, depending on the bandwidth of the

RF channel. In doing this, the DC offset problem of a direct conversion receiver is avoided. The particular advantage of a low-IF receiver is that it can be used with higher-order modulation formats (8PSK and higher). It does, however, require a higher-performance ADC than does a direct conversion receiver.

1.8.5   Subsampling Analog-to-Digital Conversion

Subsampling receivers overcome the DC offset problem typical of other direct conversion receivers. The idea is to sample the modulated RF signal using an exact subharmonic of the carrier of the RF signal to be converted. The sampling rate must be at least twice the bandwidth of the baseband signal and the track-mode bandwidth must be greater than the carrier frequency. Thus the sampling aperture is the critical parameter and must be several times smaller than the period of the carrier. Fortunately the aperture times of CMOS tracking circuits are adequate. It is critical that an RF preselect filter be used to eliminate unwanted interferers and noise outside the communication band. Aliasing of signals outside the **Nyquist** bandwidth onto the baseband signal is a consequence of subsampling. Adjacent channel signals are converted without aliasing, but these will lie outside the bandwidth of the baseband signal. Flicker noise on the sampling clock is multiplied by the subsampling ratio and appears as additional noise in baseband.

1.8.6   First IF-to-Baseband Conversion

In a superheterodyne conversion architecture there are two heterodyne stages, with the IF of the first stage in the range of 20 to 200 MHz. The assignment of frequencies is known as frequency planning, and this is treated as proprietary by the major radio vendors. This IF is then converted to a much lower IF, typically around 100 kHz or higher. This frequency is generally called baseband, but strictly it is not because the signal is still offset in frequency from DC. Some direct conversion architectures leave the first heterodyne mixing stage in place and use direct conversion of the first IF to baseband (true baseband—around DC).

1.8.7   Bilateral Double-Conversion Receiver

The receivers considered so far are suitable for narrowband communications typical of point-to-point and consumer mobile radio. There are many situations were the range of received or transmitted RF signals covers a very wide bandwidth, such as with emergency radios, television, and military communications. Typically, however, the instantaneous bandwidth is small. If narrowband RF front end architectures are used, a switchable filter bank would be required and this would result in an impractically large radio.
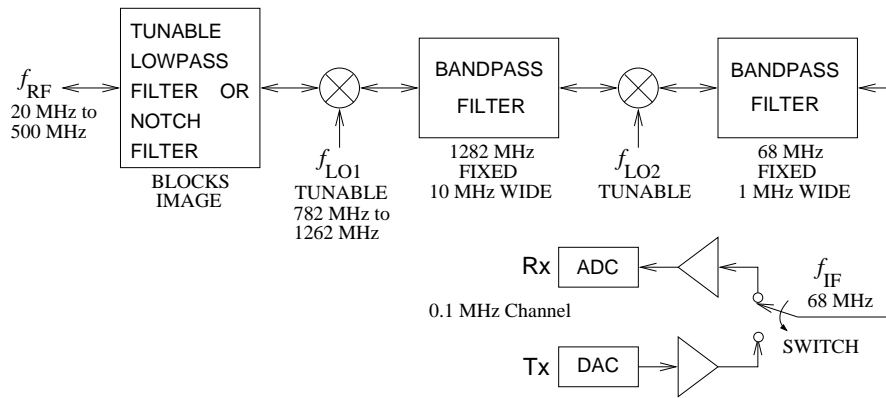
Figure 1-41   Bilateral double-conversion transceiver for wideband operation.

Tunable bandpass filters are one option being explored. One solution to covering wide bandwidths is the double-conversion transceiver architecture shown in Figure 1-41. The frequency plan of a typical radio using 0.1 MHz channels between 20 MHz and 500 MHz is shown. The key feature of this radio is that bidirectional mixers are used, such as the **diode ring mixer** of Figure 12-4 on Page 713. Following the RF chain from left to right, the RF is first mixed up in frequency, bandpass filtered using a high-Q distributed filter, and then down-converted to a lower frequency that can be sampled directly by an ADC. A much higher performance passive (and hence bidirectional) filter can be realized at gigahertz frequencies than at a few tens of megahertz. On transmit, the function is similar, with the mixers and LO source reused. As a receiver, the notch filter or lowpass filter is used to block the image frequency of the first mixer so that only the upper sideband IF is presented to the first bandpass filter. The lowpass or notch filter may be fixed, although, with the plan shown, there must be at least two states of the filters. On transmit, the lowpass or notch filters prevent the image frequency from being radiated.

1.9   Summary

This chapter presented the RF front-end architectures used from the beginnings of wireless communications up to those used in modern systems. Similar architectures are used in the front ends of radar and sensor systems. Wireless systems proliferate, and even in established domains such as cellphones, architectures are evolving to achieve greater efficiency, greater multifunctionality, and lower cost primarily by monolithically integrating and digitizing as much as possible of the RF front end. Size drives the replacement of superheterodyne architecture by eliminating large intermediate filters.

1.10 Exercises

1. An amplifier has a power gain of 1200. What is the power gain in decibels?

2. Short answer questions on gain calculations:

   (a) An amplifier with 50 Ω input impedance and 50 Ω load impedance has a voltage gain of 100. What is the gain in decibels?

   (b) An attenuator reduces the power level of a signal by 75%. What is the gain of the attenuator in decibels?

   (c) What is the wavelength in free space of a signal at 4.5 GHz?

3. The PAR of a signal is an important parameter in determining the efficiency that can be achieved by an amplifier with an allowable amount of distortion. The following questions are about determining the PAR of various signals. Note that the average power level is not necessarily the average of the minimum and maximum power levels. A full power calculation and integral should be performed.

   (a) Write down a formula for the average power of a signal $x(t)$. You can consider $x(t)$ to be a voltage across a 1 Ω resistor.

   (b) What is the PAR of an FM signal at 1 GHz with a maximum modulated frequency deviation of ± 10 kHz?

   (c) What is the PAR of a two-tone signal (consisting of two sinewaves at different frequencies that are, say, 1% apart)? First, use a symbolic expression, then consider the special case when the two amplitudes are equal. Consider that the two tones are close in frequency.

   (d) What is the PAR of a three-tone signal (consisting of three sinewaves, say, 1% apart in frequency) when the amplitude of each sinewave is the same?

   (e) What is the PAR of an AM signal with 75% amplitude modulation?

4. An FM signal has a maximum frequency deviation of 20 kHz and a modulating signal between 300 Hz and 5 kHz. What is the bandwidth required to transmit the modulated RF signal when the carrier is 200 MHz? Is this considered to be narrowband FM or wideband FM?

5. Consider two uncorrelated analog signals combined together. One signal is denoted $x(t)$ and the other $y(t)$, where $x(t) = 0.1 \sin (10^9 t)$ and $y(t) = 0.05 \sin (1.01 \cdot 10^9 t)$. What is the PAR of this combined signal? Express PAR in decibels.

6. A high-fidelity stereo audio signal has frequency content ranging from 50 Hz to 20 kHz. If the signal is to be modulated on an FM carrier at 100 MHz, what is the bandwidth required for the modulated RF signal? The maximum frequency deviation is 5 kHz when the modulating signal is at its peak value.

7. What is the PAR of a ten-tone signal when the amplitude of each tone is the same?

8. Consider FM signals.

   (a) What is the PAR of just one FM modulated signal? Express your answer in decibels.

   (b) What is the PAR of a signal comprised of two different FM signals with the same average power.

9. The following sequence of bits 0100110111 is to be transmitted using QPSK modulation. Take these data in pairs, that is, as 01 00 11 01 11. These pairs, one bit at a time, drive the I and Q channels. Show the transitions on a constellation diagram.

10. The following sequence of bits 0100110111 is to be transmitted using OQPSK modulation. Take these data in pairs, that is, as 01 00 11 01 11. These pairs, one bit at a time, drive the I and Q channels. Show the transitions on a constellation diagram.

11. The following sequence of bits 0100110111 is to be transmitted using $\pi/4$-DQPSK modulation. Take these data in pairs, that is, as 01 00 11 01 11. These pairs, one bit at a time, drive the I and Q channels. Use five constellation diagrams, with each diagram showing one transition or symbol.

12. A 16QAM modulated signal has a a maximum RF phasor amplitude of 4 V. If the noise on the signal has an RMS value of 0.1 V, what is the EVM of the modulated signal?

13. A 16QAM modulated signal has a a maximum RF phasor amplitude of 4 V. If the noise on the signal has an RMS value of 0.1 V, what is the modulation error ratio of the modulated signal in decibels?

14. A superheterodyne receiver has, in order, an antenna, a low-noise amplifier, a bandpass filter, a mixer, a second bandpass filter, a second mixer, a lowpass filter, an ADC, and a DSP which will implement quadrature demodulation. Develop the frequency plan of the receiver if the input RF signal is at 2 GHz and has a 200 kHz single-channel bandwidth. The final signal applied to the ADC must be between DC and 400 kHz so that I/Q demodulation can be done in the DSP unit. Noise considerations mandate that the LO of the first mixer must be more than 10 MHz away from the input RF. Also, for a bandpass filter to have minimum physical size , the center frequency of the filter should be as high as possible. It has been determined that the appro-priate trade-off of physical size and cost is to have a 100 MHz bandpass filter between the two mixers. (Note: a 100 MHz bandpass filter has a center frequency of 100 MHz.)

 (a) Draw a block diagram of the receiver and annotate it with symbols for the frequencies of the LOs and the RF and IF signals.

 (b) What is the LO frequency f_{LO1} of the first mixer?

 (c) What is the LO frequency f_{LO2} of the second mixer?

 (d) Specify the cutoff frequency of the lowpass filter following the second mixer.

 (e) Briefly discuss in less than one-half page other design considerations as they relate to the frequency plan, filter size, and filter specification.

15. Short answer questions. Each part requires a short paragraph of about five lines and a figure where appropriate to illustrate your understanding.

 (a) Explain the operation of a superheterodyne receiver.

 (b) Compare zero-IF and low-IF receivers.

Antennas and RF Link

2.1   Introduction

The **RF link** is the path between the output of the transmitter and the input of the receiver (see Figure 2-1). In many communication and radar systems, this includes the cable from the transmitter to the transmit antenna, the transmit antenna itself, the propagation path, the receive antenna, and the transmission line connecting the receive antenna to the receiver. Of these, the overwhelming majority of the loss is from the propagation path. Generally not just one path is taken, as reflections from the ground, buildings, and other objects lead to what is called a multipath situation where, commonly, in urban areas, 10 or 20 paths have significant power in them and each combines at the receive antenna.

   The first half of this chapter is concerned with the properties of antennas. One of the characteristics of antennas is that the energy can be focused in a particular direction, a phenomenon captured by the concept of antenna gain, which partially compensates for path loss. The second half of this chapter considers modeling the RF link and the geographical arrangement of antennas to manage the SIR while providing support for as many users as possible.
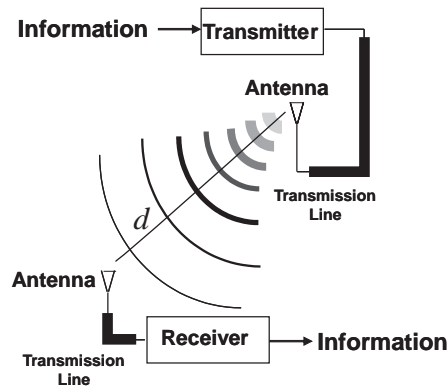
Figure 2-1   RF link.

2.2   RF Antennas

There is what seems to be an inexhaustible number of antenna structures, some of which are shown in Figure 2-2. Antennas are of two fundamental types. Resonant antennas establish a standing wave of current on one or more resonant conductor sections. Resonance occurs when the antenna section is either a quarter- or half-wavelength long—which one depends on the coupling mechanism of the transmit cable connected to the antenna. Resonant antennas are inherently narrowband. Figures 2-2(b), 2-2(c), 2-2(f), 2-2(h), and 2-2(i) show resonant antennas. The other family of antennas, shown in Figures 2-2(a), 2-2(d), 2-2(g), 2-2(j), and 2-2(k), is traveling-wave antennas, which act as extended delay lines that gradually flare out so that a traveling wave on the original transmission line transitions into free space. Traveling-wave antennas tend to be two or more wavelengths long at the lowest frequency of operation. While relatively long, they are broadband, many 3 or more octaves wide (e.g., 2 to 18 GHz for the double-ridge antenna of Figure 2-2(j)). While all **electromagnetic** (**EM**) phenomena are described by the same physics, it is easiest to describe resonant antennas as producing fields from a current distribution. With these antennas, a standing wave of current is established and the current form can be reasonably assumed and is little affected by the EM fields surrounding the antenna.

EXAMPLE 2. 1   Interference

In the figure below there are two transmitters, Tx_1 and Tx_2, operating at the same power level, and one receiver, Rx. Tx_1 is an intentional transmitter and its signal is intended to be received at Rx. Tx_2 uses the same frequency channel as Tx_1, but it transmits an interfering signal. Assume that antennas are omnidirectional and that the transmitted power density drops off as $1/d^2$, where d is the distance from the transmitter. Calculate the SIR at Rx.
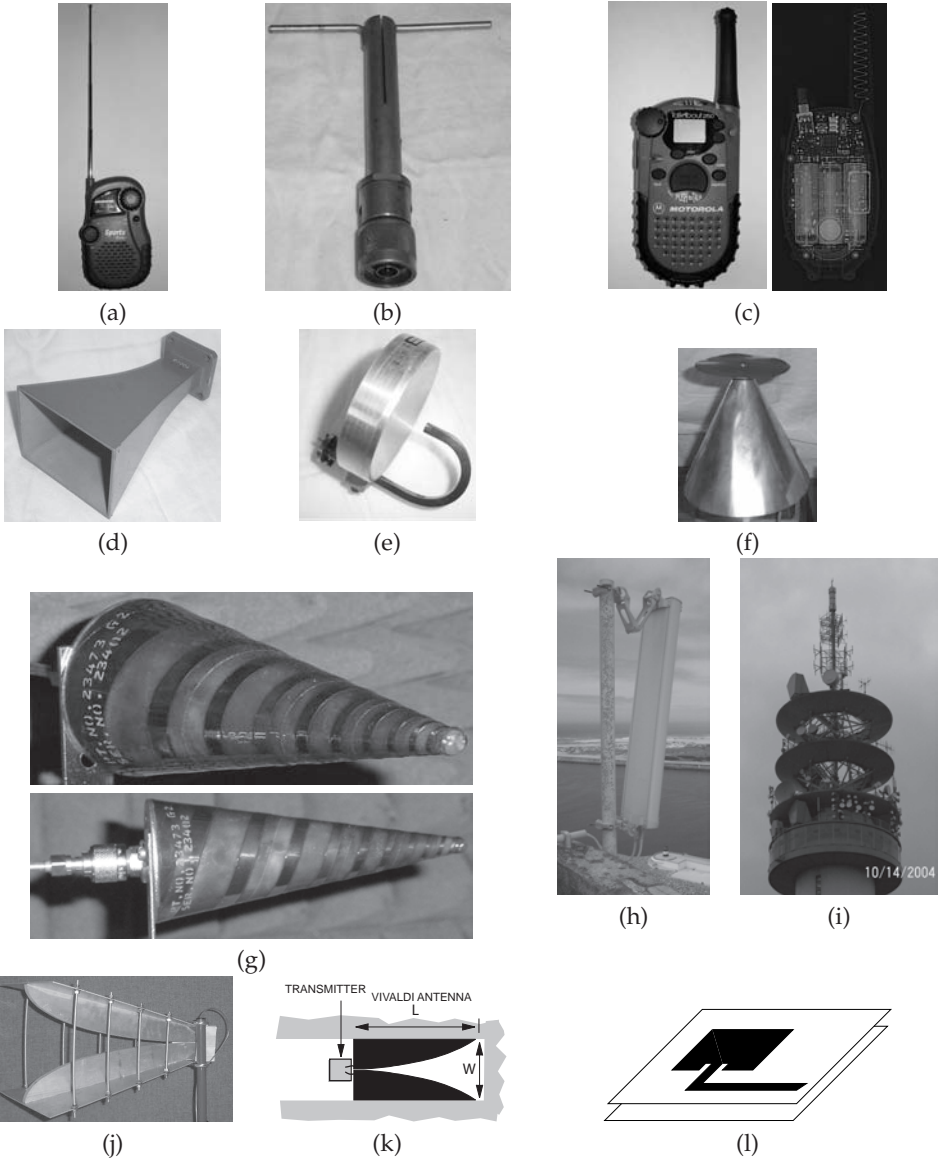
Figure 2-2 Various antennas: (a) radio with a monopole antenna; (b) dipole antenna (with coaxial balun); (c) X-ray image of a Family Radio Service (FRS) handset showing helical antenna; (d) rectangular waveguide horn antenna; (e) reflector antenna; (f) disc-cone antenna; (g) conical spiral antenna; (h) tilted patch array antenna; (i) major communication tower; (j) double-ridge horn antenna; (k) Vivaldi antenna; and (l) a microstrip patch antenna (used in the patch array of (h)).

Solution:

Let the transmitted power be P_T.
$D_1 = 2$ km and $D_2 = 4$ km.
P_1 is the signal.
P_2 is the interferer.
So

$$\text{SIR} = \frac{P_1}{P_2} = \left(\frac{D_2}{D_1}\right)^2 = 4 = 6.02 \text{ dB}.$$

2.3   Radiation from a Current Filament

The fields radiated by a resonant antenna are most conveniently calculated by considering the distribution of current on the antenna. This current distribution can generally be intuitively determined and assumed to be little affected by the radiating EM fields. The analysis begins by considering a short filament of current which is also known as a Hertzian dipole (see Figure 2-3(a)). Considering the sinusoidal steady state, the current on the filament is $I(t) = I_0 \cos(\omega t + \chi)$, so that $\mathbf{I}_0 = I_0 e^{-\jmath \chi}$ is the phasor of the current on the filament. To support this current there must be charges of opposite polarity at either end of the filament, and this is where the name Hertzian dipole comes from. The length of the dipole is h, but it has no other dimensions, that is, it is considered to be infinitely thin. Resonant antennas are conveniently modeled as being made up of an array of current filaments with the spacing of the dipoles and their lengths being a tiny fraction of a wavelength. Wire antennas are even simpler and can be considered to be line of current filaments. Ramo, Whinnery, and Van Duzer [11] calculate the spherical EM fields at the point P with the spherical coordinates (ϕ, θ, r) generated by the z-directed current filament centered at the origin in Figure 2-3. The total EM fields are

$$H_\phi = \frac{\mathbf{I}_0 h}{4\pi} e^{-\jmath k r} \left(\frac{\jmath k}{r} + \frac{1}{r^2} \right) \sin \theta \qquad (2.1)$$
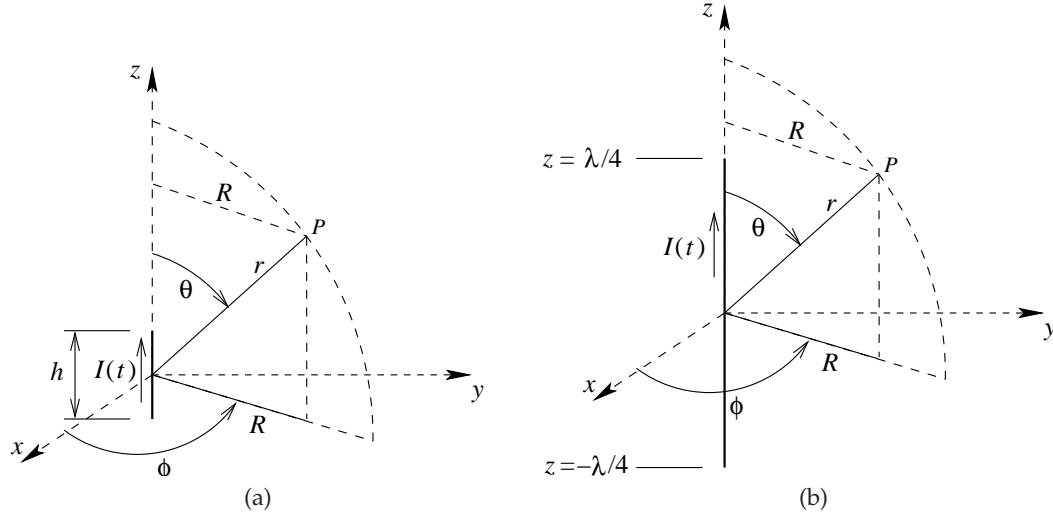
Figure 2-3 A wire antenna: (a) a current filament known as a Hertzian dipole; and (b) a half-wavelength long wire. $R = r \sin \theta$

$$E_r = \frac{\mathbf{I}_0 h}{4\pi} e^{-\jmath k r} \left(\frac{2\eta}{r^2} + \frac{2}{\jmath \omega \varepsilon_0 r^3} \right) \cos \theta \qquad (2.2)$$

$$E_\theta = \frac{\mathbf{I}_0 h}{4\pi} e^{-\jmath k r} \left(\frac{\jmath \omega \mu_0}{r} + \frac{1}{\jmath \omega \varepsilon_0 r^3} + \frac{\eta}{r^2} \right) \sin \theta \,, \qquad (2.3)$$

where η is the free-space characteristic impedance, ε_0 is the permittivity of free space, and μ_0 is the permeability of free space. The variable k is called the **wavenumber** and $k = 2\pi/\lambda = \omega\sqrt{\mu_0\varepsilon_0}$. Equations (2.1)–(2.3) are the complete fields with the $1/r^2$ and $1/r^3$ dependence describing the near-field components. In the far field, the components with $1/r^2$ and $1/r^3$ dependence become negligible and the only field components are the propagating components H_ϕ and E_θ:

$$H_\phi = \frac{\mathbf{I}_0 h}{4\pi} e^{-\jmath k r} \left(\frac{\jmath k}{r} \right) \sin \theta \qquad (2.4)$$

$$E_r = 0 \qquad (2.5)$$

$$E_\theta = \frac{\mathbf{I}_0 h}{4\pi} e^{-\jmath k r} \left(\frac{\jmath \omega \mu_0}{r} \right) \sin \theta \,. \qquad (2.6)$$

As a check, consider the fields in the plane normal to the filament, that is, with $\theta = \pi/2$ radians. Now $\sin \theta = 1$ and the fields are

$$H_\phi = \frac{\mathbf{I}_0 h}{4\pi} e^{-\jmath k r} \left(\frac{\jmath k}{r} \right) \qquad (2.7)$$

$$E_\theta = \frac{\mathbf{I}_0 h}{4\pi} e^{-\jmath k r} \left(\frac{\jmath \omega \mu_0}{r} \right) \qquad (2.8)$$

and the wave impedance is

$$\eta = \frac{E_\theta}{H_\phi} = \frac{\mathbf{I}_0 h}{4\pi} e^{-\jmath k r} \frac{\jmath \omega \mu_0}{r} \cdot \left(\frac{\mathbf{I}_0 h}{4\pi} e^{-\jmath k r} \frac{\jmath k}{r} \right)^{-1} = \frac{\omega \mu_0}{k} \ . \tag{2.9}$$

Now $k = \omega \sqrt{\mu_0 \epsilon_0}$, so

$$\eta = \frac{\omega \mu_0}{\omega \sqrt{\mu_0 \epsilon_0}} = \sqrt{\frac{\mu_0}{\epsilon_0}} = 377 \ \Omega \ , \tag{2.10}$$

as expected. Further comments can be made about the propagating fields (Equations (2.4)–(2.6)). The EM field propagates in all directions except not directly in line with the filament. The strength of the propagating field increases sinusoidally until they are maximum in the direction normal to the filament. So the power in the radiated field is concentrated in particular directions and there is effectively a gain of the maximum power density compared to the situation where the power is evenly distributed over a sphere. This concept is revisited in Section 2.5.2, where this effect is captured in the concept of antenna gain.

The power radiated is obtained from the **Poynting vector** which is the cross-product of the propagating electric and magnetic fields. The time-average propagating power density is

$$P_R = \frac{1}{2} \Re \left(E_\theta H_\phi^* \right) = \frac{\eta k^2 I_0^2 h^2}{32\pi^2 r^2} \sin^2 \theta \ \mathrm{W/m^2} \ , \tag{2.11}$$

so as expected the power density is proportional to $1/r^2$.

2.3.1 Finite Length Wire Antennas

The EM wave propagated by a wire of finite length is obtained by considering the wire as being made up of many filaments and the field is then the superposition of the fields from each filament. As an example, consider Figure 2-3(b), where the wire is half a wavelength long. The current on the wire will be a standing wave with all of the current along the wire in phase so that

$$I(t) = I_0 \sin(\beta z) \ . \tag{2.12}$$

From Equations (2.4) and (2.6) and referring to Figure 2-3(b), the fields in the far field are

$$H_\phi = \int_{-\lambda/4}^{\lambda/4} \frac{I_0 \sin(kz)}{4\pi} e^{-\jmath k r'} \left(\frac{\jmath k}{r'} \right) \sin \theta' \ dz \tag{2.13}$$

$$E_\theta = \int_{-\lambda/4}^{\lambda/4} \frac{I_0 \sin(kz)}{4\pi} e^{-\jmath k r'} \left(\frac{\jmath \omega \mu_0}{r} \right) \sin \theta' \ dz \ , \tag{2.14}$$
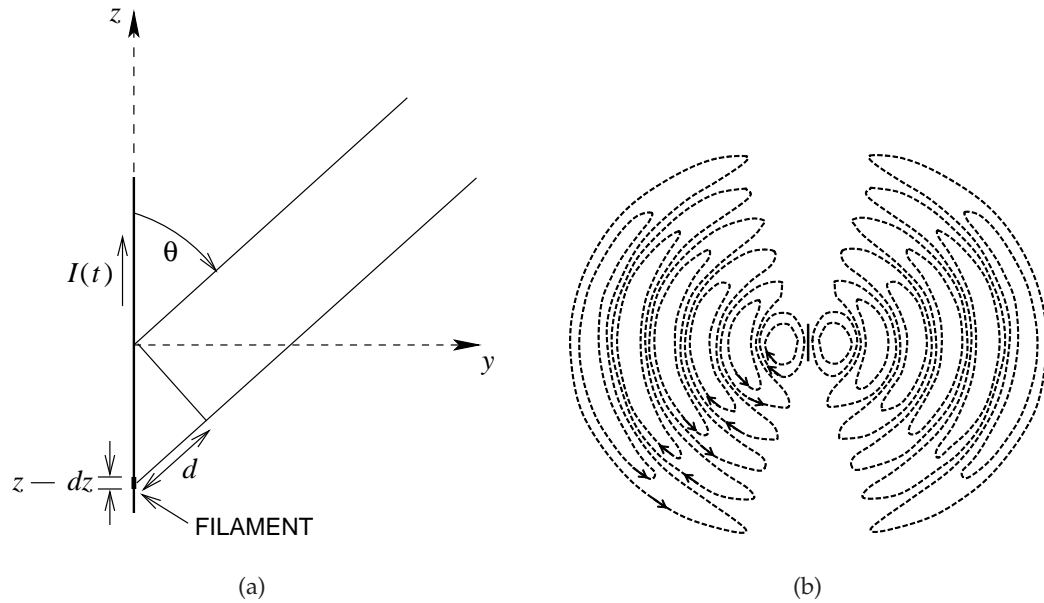
Figure 2-4 Wire antenna calculations: (a) geometry for calculating phase contributions from current filaments with $d = -z \sin \theta$, where z is the coordinate of the filament; and (b) EM field ($|H_\phi|$ and $|H_\theta|$) in the y-z plane due to a half-wavelength-long current element.

where θ' is the angle from the filament to the point P. Solving these equations is involved and will not be done here. It can, however, be conveniently done using MATLAB. The net result is that the fields are further concentrated in the direction in the plane normal to the wire. At large r, at least several wavelengths distant from an antenna, only the field components decreasing as $1/r$ are significant. At large r, the magnitude of the field contribution from each of the current filaments is approximately the same, however, the phase contribution difference is significant and results in shaping of the fields. The geometry to be used in calculating the far field is shown in Figure 2-4(a). The phase contribution of each element relative to that at $z = 0$ is $(z \sin \theta)/\lambda$, where z is the coordinate of the filament. So Equations (2.13) and (2.14) become

$$H_\phi \quad = \quad I_0 \left(\frac{\jmath k}{4\pi r} \right) \sin \theta e^{-\jmath k r} \int_{-\lambda/4}^{\lambda/4} \frac{\sin(kz)}{4\pi} \sin(z \sin(\theta)).dz \quad (2.15)$$

$$E_\theta \quad = \quad I_0 \left(\frac{\jmath \omega \mu_0}{4\pi r} \right) \sin \theta e^{-\jmath k r} \int_{-\lambda/4}^{\lambda/4} \sin(kz) \sin(z \sin(\theta)).dz \quad (2.16)$$

Figure 2-4(b) is the result of a calculation of the above fields in the x-z plane. In the extreme for an infinitely long wire with a current that oscillates

in phase along the wire (something that is actually impossible to realize), all of the propagating fields are in the $\theta = \pi/2$ plane.

A summary of the implications of the above equations are, first, that the strength of the radiated electric and magnetic fields is proportional to the current on the wire antenna, so establishing a standing wave of current is important to an efficient wire antenna. A second result is that, in the absence of obstructions to propagation, the power density of propagating EM fields is proportional to $1/r^2$, where r is the distance from the antenna. A third interpretation is that the longer the antenna, the flatter the radiated transmission profile; that is, the more tightly confined the radiated energy. In the case of the wire antenna, the peak radiated field is in the plane normal to the antenna, and thus the wire antenna is generally oriented vertically so that transmission is in the plane of the earth and power is not radiated unnecessarily into the ground or into the sky. The problem, of course, is that the current along the wire antenna should have the same phase relationship. A standing wave has the property that the current along the wire is in phase as long as it is no more than a half-wavelength long; any longer and the current would change direction. An ingenious solution to this problem is the stacked dipole antenna (Figure 2-5). In this antenna, the radiating element is hollow and a coaxial cable is passed through the antenna elements and the half-wavelength sections are feed separately to create a "wire" antenna that is one or more wavelengths long. Most cellular antennas using wire antennas are stacked dipole antennas.

Standing waves of current can be realized by structures other than wires. Microstrip patch antennas are an example, but the underlying principle is that an array of current filaments generates EM components that combine through superposition to generate a propagating field. Resonant antennas are inherently narrowband because of the reliance on the establishment of a standing wave. A relative bandwidth of 5–10% is typical. One solution to this is to use an antenna with many radiating elements of different lengths. The **Yagi-Uda antenna** is the most famous of these achieving wide bandwidth. Readers are referred to **Balanis** [12] and **Fusco** [13] for a detailed derivation of the properties of antennas. The discussion of antennas can now return to a more qualitative basis.

2.4   Resonant Antennas

On a coaxial line, voltage and current waves are guided and confined by the inner and outer conductors. If the outer conductor is suddenly removed, the current wave continues traveling on the center conductor, or wire, as shown in Figure 2-6(a). At this point it makes little sense to talk about voltage, as there is no convenient path over which to integrate the electric field. Without the outer shield of a coaxial cable, the EM fields begin to spread out, but the current in particular is guided by the single conductor, as shown. So it is
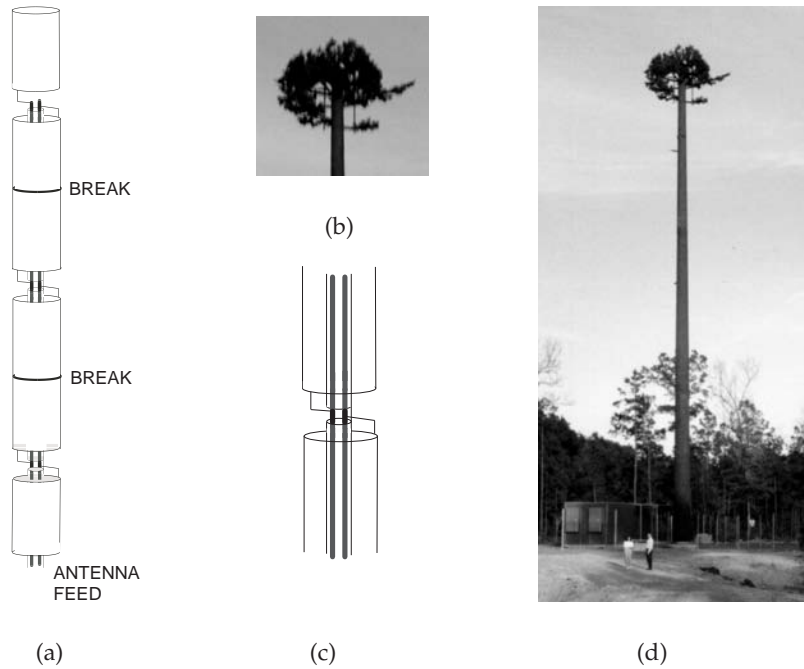
Figure 2-5 Stacked dipole antenna: (a) schematic representation; (b) in a base station antenna camouflaged as a pine tree; (c) detail showing a trisectional stacked dipole antenna; and (d) detail of a stacked dipole connection.
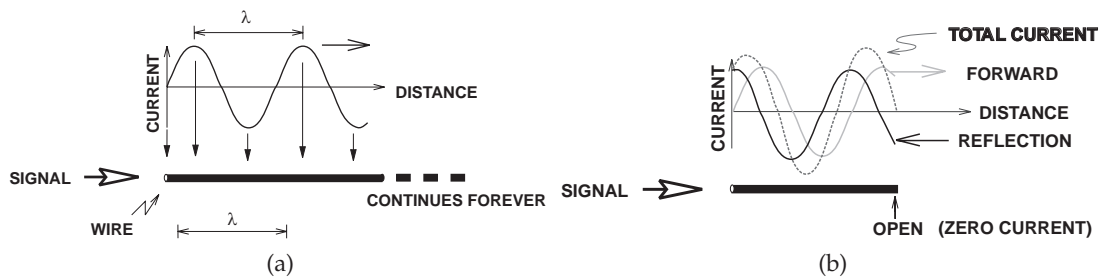
Figure 2-6 Currents on a wire: (a) section of an infinitely long wire; and (b) a truncated section of wire.

reasonable to model the single conductor as a lossy transmission line. If the wire is infinitely long, the current wave continues to travel down the line as a sinusoidal current whose amplitude decreases exponentially with distance. Such a wire antenna (as it is called) would not be an efficient antenna. The electric field that is radiated is directly proportional to the current. In particular, if the current along the wire is divided into short segments, or what are called current filaments, the electric field phasor $\mathbf{E}$ at a distance r

due to just one filament $\mathbf{I}$ at a position x is

$$\mathbf{E} = \frac{\mathbf{I}(x)}{r} \ . \tag{2.17}$$

The electric field is in the same direction as the current filament and the magnitude of the field is inversely proportional to distance. The total electric field is the integral of the electric field from all of the current filaments:

$$\mathbf{E} = \int \frac{\mathbf{I}(x)}{r} \, dx \ . \tag{2.18}$$

This is exactly the way an EM simulation tool evaluates the far field; that is, the field many wavelengths distant from the antenna. The far field is usually at several wavelengths and at least several times the length of the antenna away. Closer to the antenna the fields are more complex,[1] but the interest here is in the far-field characteristic.

At a distance, the current in the wire averages out as the wire antenna begins to subtend a smaller and smaller angle. Returning again to the infinitely long wire of Figure 2-6(a), the effective current seen at a distance is virtually zero, as the current changes direction every half wavelength along the wire. To obtain an efficient antenna, all of the current should be pointed in the same direction at a particular time. One way of achieving this is to establish a standing wave, as shown in Figure 2-6(b), where the wire is of finite length. At the open-circuited end, the current reflects with a reflection coefficient of -1 so that the total current at the end of the wire is zero. The forward- and backward-traveling current waves combine to create a standing wave. Provided that the antenna is sufficiently short, all of the total current—the standing wave—is pointed in the same direction. The length, though, should be as long as possible, subject to this constraint, so that the field radiated (see Equation (2.18)) is maximum. The optimum length is a half wavelength. When the wire is longer, the contributions to the field from the oppositely directed current segments cancel (see Figure 2-7(b)).

A wire above a ground plane, as in Figure 2-8(a), is called a monopole. A coaxial cable is attached below the ground plane to the antenna. Often a series capacitor provides a low level of coupling leading to the transmission-line equivalent circuit of Figure 2-8(b). In effect, a resonant structure is established with open circuits at the two ends so that resonance occurs at $\lambda/2$. The wave along the antenna rapidly decays, with the effect that the wave stretches out so that the actual resonant condition occurs when the length is closer to $5/8\lambda$. Sometimes a series capacitor is not used and instead the discontinuity between the cable and the wire establishes one end of the resonant structure. A matching network is used either adjacent to the

---

[1]   A complete description is given by Fusco [13].
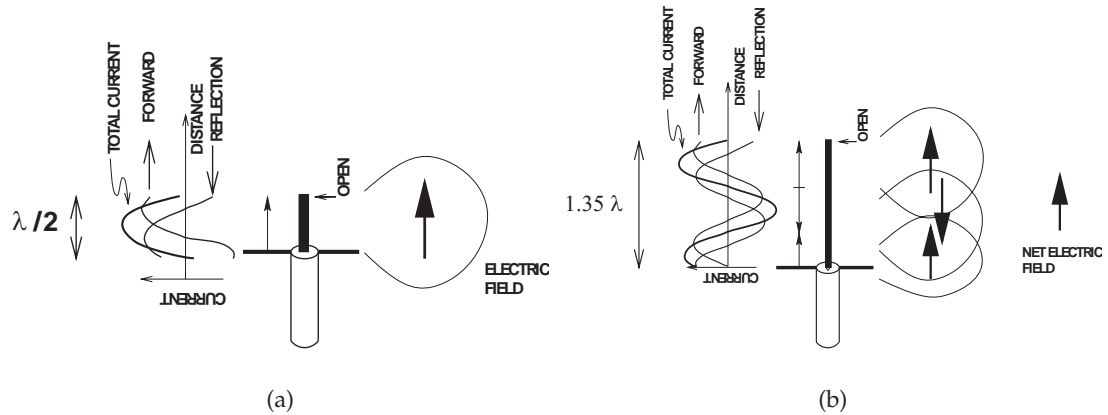
Figure 2-7 A monopole antenna showing forward- and backward-traveling currents as well as the total current: (a) a half-wavelength-long antenna; and (b) a relatively long antenna.

monopole or back at the transmitter to ensure maximum power transfer to the antenna.

The monopole above is nominally one-half wavelength long. If the length of the monopole is reduced to one-quarter wavelength long, as in Figure 2-8(d), it is again resonant, and now the input impedance, Z_{in}, is found to be 36 Ω. Thus a 50 Ω cable can be directly connected to the antenna and nearly all the power is transferred to the antenna and from there radiated. Another variation on the monopole is shown in Figure 2-9, where the key component is the phasing coil. The phasing coil rotates the electrical angle of the current phasor on the line so that the current in the $\lambda/4$ segment is in the same direction as the $\lambda/2$ segment. The result is that the two straight segments of the loaded monopole radiate a more tightly confined EM field. The phasing coil itself does not effectively radiate.

A monopole above a ground plane results in current flow in the ground plane. The effect is that the monopole above-ground structure is equivalent to a monopole plus an image having an image current in the same direction as the current on the monopole. This situation is shown in Figure 2-10(a). In effect, the antenna is twice as long with the current in the same direction. The field generated is compressed into a "squashed-balloon shape," focusing the radiated energy perpendicular to the monopole. In reality, there are no fields below the ground plane and there is only energy in the top half of the radiated pattern shown, so the image is not doubling the energy, but acting solely to further focus the fields. Thus the radiated energy density in one direction is higher than it would otherwise be.
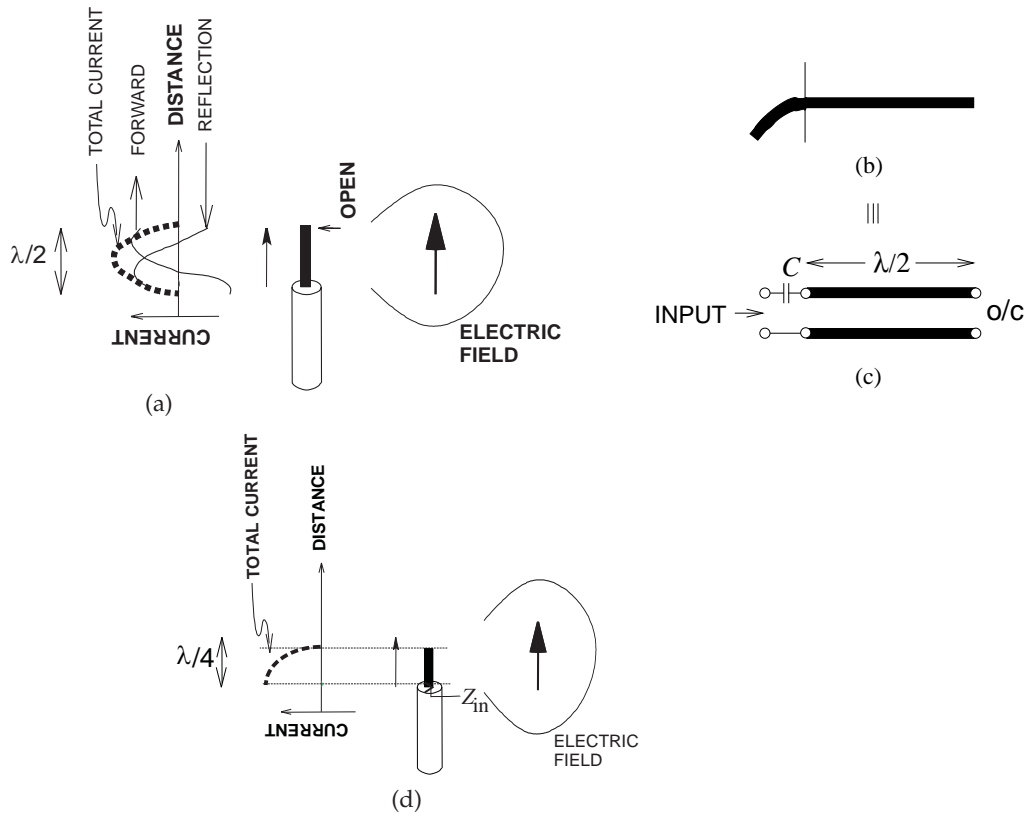
Figure 2-8   Half-wavelength-long monopole: (a) vertically mounted showing orientation of the radiated electric field; (b) rotated so that it can be cast into a conventional circuit form; (c) transmission line-based equivalent circuit of monopole with small coupling capacitor C; and (d) quarter-wavelength-long monopole.

2.4.1   Resonant Scattering

Propagation is rarely from point to point as the path is often obstructed. One type of event that interferes with transmission is scattering. The effect of scattering depends on the size of the objects causing scattering. Here the effect of the pine needles of Figure 2-11 will be considered. The pine needles (as most objects in the environment) conduct electricity, especially when wet. So when an EM field is incident, the individual needles act as wire antennas, with the current on them maximum when the wavelength of the signal is one-half wavelength long. At this length, the "needle" antenna supports a standing wave and will reradiate the signal in all directions. This is scattering, and there is a considerable loss in the direction of propagation of the original fields. As well, there is loss due to a needle not being a very
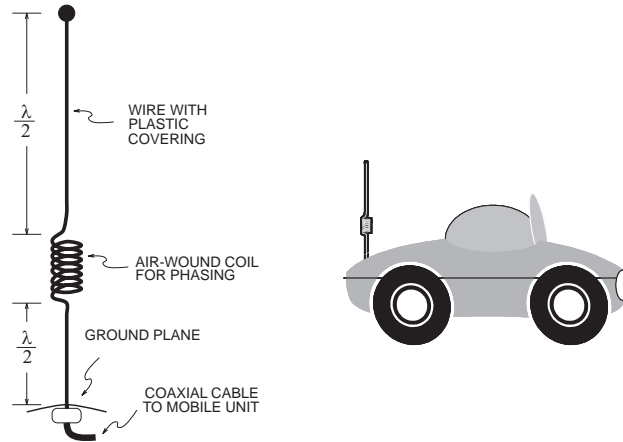
Figure 2-9 Mobile antenna with phasing coil extending effective length of antenna.
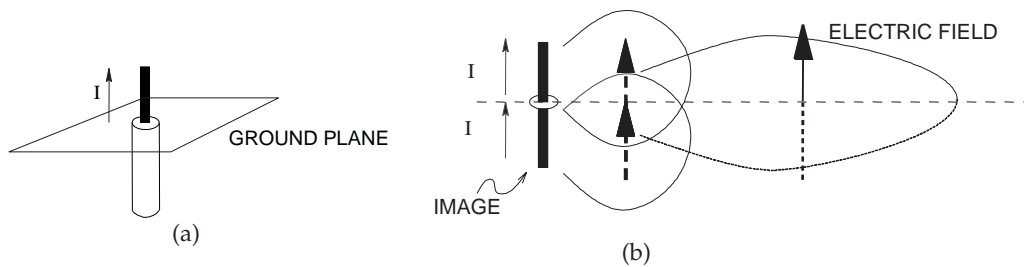
Figure 2-10 Monopole above a ground plane: (a) actual structure; and (b) equivalent image structure.
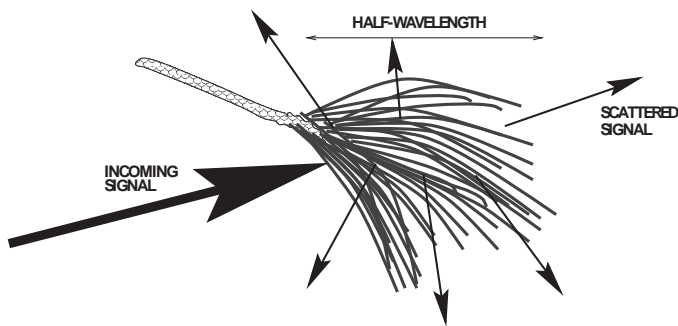
Figure 2-11 Pine needles illustrating scattering.

good conductor, so energy is lost as heat. As was mentioned, the effect of scattering is signal dependent. A typical pine needle is 15 cm long, which is exactly $\lambda/2$ at 1 GHz, and so pine trees have an extraordinary impact on
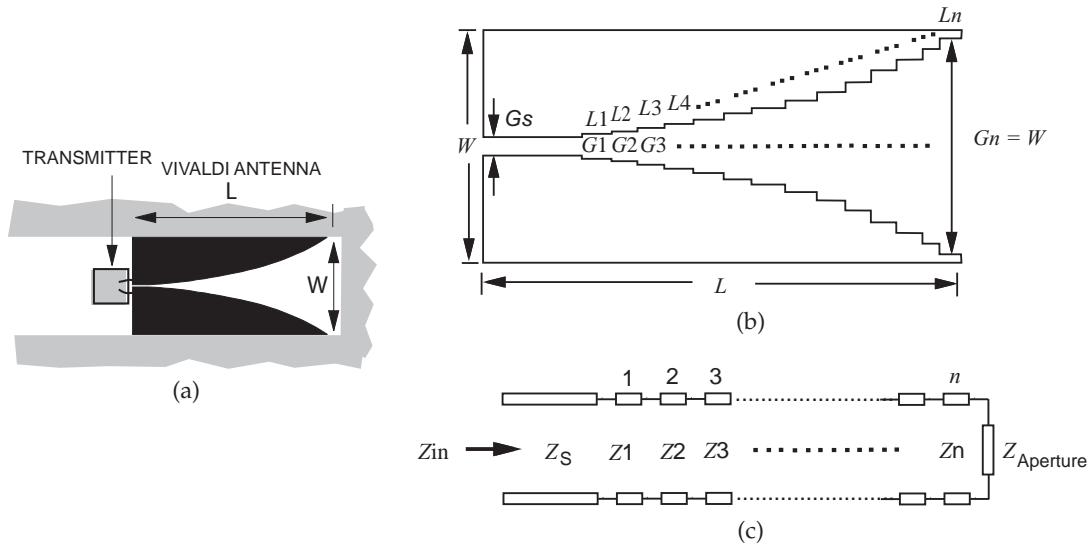
Figure 2-12   Vivaldi antenna showing design procedure: (a) the antenna; (b) a stepped approximation; and (c) transmission line approximation.

cellular communications at 1 GHz. As a rough guide, 20 dB of a signal is lost when passing through a stand of pine trees. At 2 GHz, a similar impact comes from other leaves, and they do not need to look like wire antennas. Consider an oak leaf having a dimension of 7.5 cm. This is $\lambda/2$ at 2 GHz, the other dominant cellular frequency. So scattering results, but now the loss is season dependent, with the loss due to scattering being considerably smaller in winter (when trees such as oaks do not have leaves) than in summer. Consequently the size of a cell expands and contracts during the year.

2.5 Traveling-Wave Antennas

Traveling-wave antennas share the characteristic of broad bandwidth. These antennas begin as a transmission line structure that flares out slowly, providing a low reflection transition from a transmission line or waveguide to free space. The bandwidth can be very large and is primarily dependent on how gradual the transition is. For example, the double ridge horn antenna of Figure 2-2(j) has a bandwidth of 500 MHz to 12 GHz, but is quite large, having a volume of 30 cm $\times$ 30 cm $\times$ 50 cm.

2.5.1   Vivaldi Antenna

One of the more interesting traveling-wave antennas is the **Vivaldi antenna** of Figure 2-12. This is a good example of both the operation and design of traveling-wave antennas. The Vivaldi antenna is an extension of a slotline
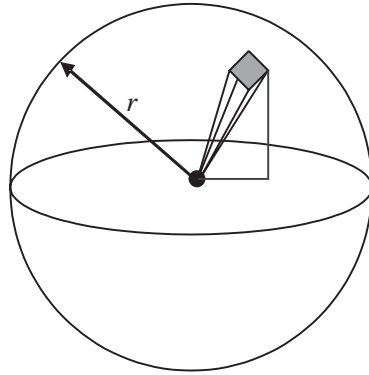
Figure 2-13 Free-space spreading loss. The energy intercepted by the red square is proportional to $1/r^2$.

in which the fields are confined in the space between two metal sheets in the same plane. The slotline spacing increases gradually in an exponential manner, much like that of a Vivaldi violin shape, over a distance of a wavelength or more. The forward-traveling wave on the antenna continues to propagate with negligible reflected field. Eventually the slot opens sufficiently that the effective impedance of the slot is that of free space and the traveling wave continues to propagate in air, being guided by the conductors.

The other traveling-wave antennas work similarly and all are at least a wavelength long, with the central concept being a gradual taper from the characteristic impedance of the originating transmission line to free space. As well, the final aperture is at least one-half wavelength across so that the fields can curl on themselves and are self-supporting.

2.5.2 Antenna Gain

Antennas do not radiate uniformly in all directions and instead concentrate power in particular directions. So that this focusing effect can be quantified, an isotropic radiator that radiates equally in all directions is introduced as a reference.

The fictitious **isotropic antenna** radiates uniformly in all directions, with the radiated power density at a distance r uniform over a sphere of radius r and falling off as $1/r^2$. If the total radiated power is P_r, then, for the isotropic antenna, the power radiated per unit solid angle (i.e., power radiated per steradian), P_D, is

$$P_D|_{\text{ISOTROPIC}} = \frac{dP_r}{d\Omega} = \frac{P_r}{4\pi} \, , \qquad (2.19)$$

since there are 4π steradians in a sphere. The symbol for steradians is sr so that P_D has the units of W/sr (watts per steradian). For an isotropic antenna

$$\left(\frac{dP_r}{d\Omega}\right)_{\text{ISOTROPIC}} = \frac{P_r}{4\pi} \; . \tag{2.20}$$

One metric used with antennas is **antenna efficiency**. Antenna efficiency takes into account losses in the antenna principally due to resistive (I^2R) losses. Antennas are of two types: resonant antennas and traveling-wave antennas. Most antennas are resonant antennas and work by creating a large current that is maximized through the generation of a standing wave at resonance. There is a lot of current, and even just a little resistance results in substantial resistive loss. There is one other form of loss and this is the power that is reflected from the input of the antenna. This is usually small. It occurs when the input impedance of the antenna is not perfectly matched to the driving transmission line. The total radiated power (in all directions) is the power input to the antenna less losses. The efficiency of the antenna is therefore defined as

$$\eta = P_r/P_{\text{IN}} \; , \tag{2.21}$$

where P_{IN} is the power input to the antenna. Antenna efficiency is very close to one for many antennas, but can be 50% for patch antennas (the antenna in Figure 2-2(l)).

Antenna gain is a quantity that can only be measured. Antenna gain incorporates the resistive losses in the antenna as well as dielectric losses. The input power to an antenna can be measured and the region of maximum power density from the antenna found. The peak radiated power density is compared to the calculated density from an ideal (lossless) isotropic antenna at the same distance with the same input power. It is impossible to measure the resistive and dielectric losses of an antenna directly and impossible to simulate it accurately. The way antenna efficiency is obtained is to use theoretical calculations of the directivity of an antenna; that is, calculate theoretically the expected antenna gain assuming no losses in the antenna itself. This is compared to the measured antenna gain. The difference yields the antenna efficiency.

Antenna gain is not a gain in the same sense as the gain of an amplifier, however, it can be used in calculations of power as though it were. Antennas concentrate the radiated power in one or more directions so that the density of the power radiated in the direction of the peak field is higher than the power density from an isotropic antenna that radiates equally in all directions. Power radiated from a base station antenna, such as that shown in Figure 2-14, is concentrated in a region that looks like a toroid or, more closely, a balloon squashed at its north and south poles. This is exactly the antenna pattern desired because when the antenna is mounted vertically, it will not radiate much power into space and will concentrate power in a region skimming the surface of the earth. Antenna gain is a measure of the effectiveness of an antenna to concentrate power in one direction, and, to be fair, we need to consider the efficiency of the antenna. So antenna gain

is a quantity that can only be measured, as it is difficult to model losses. In this measurement, the region of maximum radiated power density is found at several wavelengths from the antenna. Then, imagining an isotropic antenna, the power density on the surface of a sphere with a radius equal to the distance from the antenna is calculated, assuming that all of the power available at the input of the antenna is radiated. The ratio of the maximum power density of the actual antenna to the power density of the isotropic antenna is taken as the antenna gain. The result is that the efficiency of the antenna is incorporated in antenna gain. Antenna gain is defined as

$$
\begin{aligned}
G_A &= 4\pi \frac{\text{Radiated power per unit solid angle}}{\text{Total input power to the antenna}} \\
&= 4\pi \frac{(dP_r/d\Omega)}{P_{\text{IN}}} \\
&= \eta \frac{(dP_r/d\Omega)}{(dP_r/d\Omega)_{\text{ISOTROPIC}}} \; .
\end{aligned}
\tag{2.22}
$$

Thus, in free space where power spreads out by $1/d^2$, the maximum power density at a distance d is

$$
P_D = \frac{G_A P_{\text{IN}}}{4\pi d^2} \; ,
\tag{2.23}
$$

where $4\pi d^2$ is the area of a sphere of radius d and P_T is the total power radiated by the antenna. So the propagation loss, sometimes called the path loss, is $4\pi d^2$. There are two definitions of path loss and this is one. This will be revisited in Section 2.6.1 on Page 83.

Antenna gain is expressed in units of decibels or more commonly as dBi to indicate that antenna gain is with respect to an isotropic antenna. The gains of common resonant and traveling-wave antennas are given in Table 2-1.

Antenna loss refers to the same mechanism that gives rise to antenna efficiency. So an antenna with an antenna efficiency of 50% has an antenna loss of 3 dB. Generally losses are resistive loss due to I^2R loss and mismatch loss of the antenna that occurs when the input impedance is not matched to the impedance cable connected to the antenna. There is a source of possible confusion here, as antenna loss is not directly related to antenna gain (they are not the inverses of each other). Recall that antenna gain captures the focusing power or directivity of the antenna as well as antenna loss. Therefore, when discussing these antenna matters it is important that the well-defined terms **antenna gain**, **antenna loss**, and **antenna efficiency** are used.

Table 2-1 Antenna gains of several resonant and traveling-wave antenna systems. The traveling-wave antennas have a bandwidth of 3 to 4 octaves except as indicated.

Antenna	Type	Figure	Gain (dBi)	Notes
Lossless isotropic antenna			0	
$\lambda/2$ dipole	Resonant	2-2(b)	2	$R_{in} = 73\ \Omega$
3λ diameter parabolic	Traveling	2-2(e)	38	$R_{in} = $ match
Patch	Resonant	2-2(l)	9	$R_{in} = $ match
Vivaldi	Traveling	2-2(k)	10	$R_{in} = $ match
Double ridge horn	Traveling	2-2(j)	9	$R_{in} = $ match
Waveguide horn	Traveling	2-2(d)	12–25	$\frac{1}{2}$ octave
Conical spiral	Traveling	2-2(g)	9	$R_{in} = $ match
Conical	Traveling	2-2(f)	9	$R_{in} = $ match
Loaded monopole over ground	Resonant	2-9	8	$R_{in} = 50\ \Omega$
Helical monopole	Resonant	2-2(c)	5	$R_{in} = 35\ \Omega$
$\lambda/4$ monopole over ground	Resonant	2-10	2	$R_{in} = 36\ \Omega$
$5/8\lambda$ monopole over ground	Resonant	2-8(a)	3	Matching required

---

EXAMPLE 2. 2   Antenna Loss

An antenna has an antenna gain of 13 dB and an antenna efficiency of 50%.

(a) What is the antenna loss?

From $\eta = 50\%$, antenna loss = 3 dB.

(b) If the antenna is cooled to near absolute zero so that it is lossless, what would the antenna gain be?

The antenna gain would increase by 3 dB and antenna gain incorporates both directivity and antenna losses. So the gain of the cooled antenna is 16 dB.

EXAMPLE 2. 3   Antenna Calculations

In a resonant antenna, a large current is established by creating a standing wave. The current peaking that thus results establishes a strong electric field (and hence magnetic field) that radiates away from the antenna. A typical microstrip patch antenna loses 50% of the power input to it as resistive (I^2R) losses and has an antenna gain of 11 dB. Consider a base station patch antenna that has a 40 W input. Also consider that the transmitted power density falls off with distance d as $1/d^2$.

(a) What is the input power in dBm?

P_{in} = 40 W = 46.02 dBm

(b) What is the power transmitted in dBm?

$$P_{\text{RADIATED}} = 50\% \text{ of } P_{\text{IN}} = 20 \text{ W or } 43.01 \text{ dBm}$$
$$\text{Alternatively, } P_{\text{RADIATED}} = 46.02 \text{ dBm} - 3\text{dB} = 43.02 \text{ dBm}$$

Note $\frac{1}{2} = -3.01$ dB, but this is approximated as $\frac{1}{2}$ just as 3 dB is approximated as a power ratio of 2.

(c) What is the power density at 5 km?

A sphere of radius 5 km has an area $A = 4\pi r^2 = 3.142 \cdot 10^8 \text{ m}^2$
11 dB = 12.6
In the direction of peak radiated power, the power density is

$$P_D = \frac{P_{\text{in}} \cdot (\text{ANTENNA GAIN})}{A} = \frac{40 \cdot 12.6 \text{ W}}{3.142 \cdot 10^8 \text{m}^2} = 1.603 \ \mu\text{W/m}^2$$

(d) What is the power captured by a receive antenna (at 5 km) that has an effective antenna aperture of 6 cm^2?

The power received
$$P_{Rx} = P_D \cdot \left(6 \text{ cm}^2\right) = P_D \cdot 6 \cdot 10^{-4} \text{m}^2 = (1.603 \ \mu\text{W} \cdot \text{m}^{-2}) \cdot (6 \cdot 10^{-4} \text{m}^2) = 962 \text{ pW}$$

(e) If the background noise level captured by the antenna is 100 fW, what is the SNR in dB. Ignore interference that comes from other transmitters.

$$\text{SNR} = (962 \text{ pW})/(100 \text{ fW}) = 9\,618 = 39.8 \text{ dB}$$

2.5.3 Effective Isotropic Radiated Power

A base station does not radiate power equally in all directions, as it is inefficient to do so. Instead, power is concentrated horizontally, as shown in Figure 2-14. The peak radiated power density is incorporated in another quantity called the **effective isotropic radiated power** (**EIRP**):

$$\text{EIRP} = P_{\text{IN}}G_{\text{MAX}} . \tag{2.24}$$

This uses antenna gain to derive the power density that would come from an isotropic antenna. Many limits on the power levels of transmitters are in terms of the EIRP rather than the total radiated power. Sometimes **Equivalent Radiated Power** (**ERP**) is used instead of EIRP: they refer to the same thing.

2.5.4 Effective Aperture Size

Another concept used with antennas is effective aperture size. An antenna has an effective size which is more than its actual physical size. This is because of its influence on the fields around it. This concept is captured
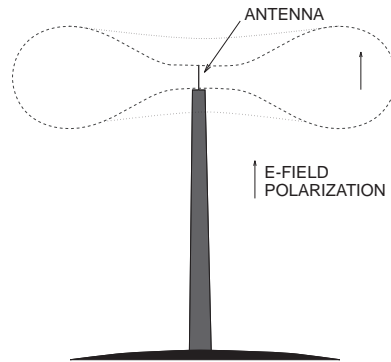
Figure 2-14   Base station transmitter pattern.

by the concept of effective aperture size whereby the power captured by
an antenna is the effective aperture size (or area) multiplied by the power
density. That is, the effective aperture size of an antenna is the area of
a circle which captures all of the power passing through it and delivers
this power to the output terminals of the antenna. Antennas are reciprocal
networks, and there is a direct relationship between effective aperture area
(the usage preferred with receivers) and antenna gain (preferentially used
with transmitters). The effective aperture area of a receive antenna, A_R, is
related to the receive antenna gain, G_R, as follows:

$$A_R = \frac{G_R \lambda^2}{4\pi} \, , \tag{2.25}$$

where λ is wavelength. Both antenna gain and effective aperture account for
losses in the antenna. So if P_D is the power density at the antenna, the power
received is

$$P_R = P_D A_R = P_D \frac{G_R \lambda^2}{4\pi} \, . \tag{2.26}$$

P_R is the power delivered at the output connector of the receive antenna,
as loss in the receive antenna is incorporated in G_R and A_R. The total
power radiated by the transmit antenna in the direction of maximum power
density is given by multiplying the power input to a transmit antenna, P_T,[2]
by the antenna gain of the transmit antenna, G_T. This can be converted to
the power density at a distance d (ignoring multipath effects),

$$P_D = \frac{P_T G_T}{4\pi d^2} \, , \tag{2.27}$$

---

[2]   Note that this was P_{in} previously and here the subscript T is used to indicate the power
input to the transmit antenna.

and the power received by the receive antenna is

$$P_R = P_D A_R = \frac{P_T G_T}{4\pi d^2} \frac{G_R \lambda^2}{4\pi} = P_T G_T G_R \left(\frac{\lambda}{4\pi d}\right)^2 . \qquad (2.28)$$

The effective aperture area of a high-frequency parabolic or dish antenna is very nearly equal to their physical area. Wire antennas such as monopole and dipole antennas have effective aperture sizes that greatly exceed their physical areas because of their influence on the fields near the antenna. Multipath effects and nonidealities such as soil, being a lossy conductor, result in excess losses between the antennas. This is considered further in Section 2.6.1 on Page 83.

EXAMPLE 2. 4   Point-to-Point Communication

In a point-to-point communication system, a parabolic receive antenna has an antenna gain of 60 dB. If the signal is 60 GHz and the power density at the receive antenna is 1 pW/cm^2, what is the power at the output of the antenna connected to the RF electronics?

Solution:

The first step in solving this problem is determining the effective aperture area of the parabolic antenna using Equation (2.25). The frequency is 60 GHz and so the wavelength (λ) is 5 mm. Also note that $G_R = 60$ dB $= 10^6$. From Equation (2.25), the effective aperture area of the antenna is

$$A_R = \frac{G_R \lambda^2}{4\pi} = \frac{10^6 \cdot 0.005^2}{4\pi} = 1.989 \text{ m}^2. \qquad (2.29)$$

Note that A_R is the effective aperture area and is not the physical size of the antenna. It also includes loss in the antenna. The antenna affects the fields in a volume that is larger than the physical volume of the antenna. Equation (2.26) can be used to obtain the power received by the antenna. Now $P_D = 1$ pW/cm^2 = 10 nW/m^2. Thus the total power delivered to the RF electronics is

$$P_R = P_D A_R = 10 \text{ nW} \cdot \text{m}^{-2} \cdot 1.989 \text{ m}^2 = 19.89 \text{ nW}. \qquad (2.30)$$

2.6 The RF link

The RF link is between a transmit antenna, and a receive antenna and the term is used in the context of establishing loss. Sometimes the RF link includes the antenna and sometimes it does not. Whether or not the receive and transmit antenna gains need to be considered will be clear from the context. The principle source of loss is the spreading out of the fields as they propagate. This was described for the isotropic antenna on Page 75. In the absence of any other effects the power density reduces as $1/d^2$, where d is
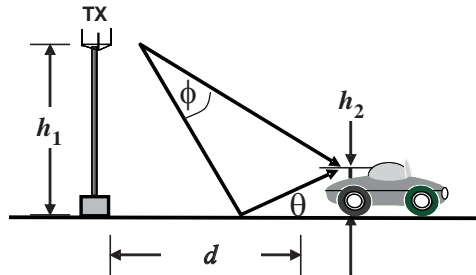
Figure 2-15   Ground reflection.
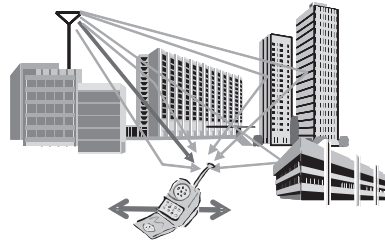
Figure 2-16   Multipaths in an urban environment.

distance. With **Line Of Sight** (LOS), the transmission loss is calculated from the antenna gain of the transmit antenna, G_T, the antenna gain of the receive antenna, G_R, and the free-space path loss, L. The power density at a distance d from the transmit antenna is

$$P_D = \frac{P_T G_T}{4\pi d^2} \,, \tag{2.31}$$

where P_T is the power input to the transmit antenna. The power captured by the receive antenna is calculated from the power density at the receive antenna and the effective aperture area of the antenna.

The antennas of point-to-point links are high gain and are mounted high on masts, and the propagating energy avoids the ground. With handheld devices there will be reflections from the ground and structures. Considering just LOS and a single ground reflection, the situation is as shown in Figure 2-15. This is a special situation, as the ground changes the phase of the signal upon reflection, generally by $180°$. When the car, in this case, is a long way from the base station, the lengths of the two paths are almost identical and the level of the signals in the two paths are almost the same. The net result is that these two signals tend to cancel, and so instead of the power falling off by $1/d^2$, it falls of by $1/d^3$. In an urban area such as that shown in Figure 2-16, there are many significant multipaths and the power falls of by $1/d^4$. The multipath situations are so varied that measurements are the only reliable indicator of how power reduces with distance.
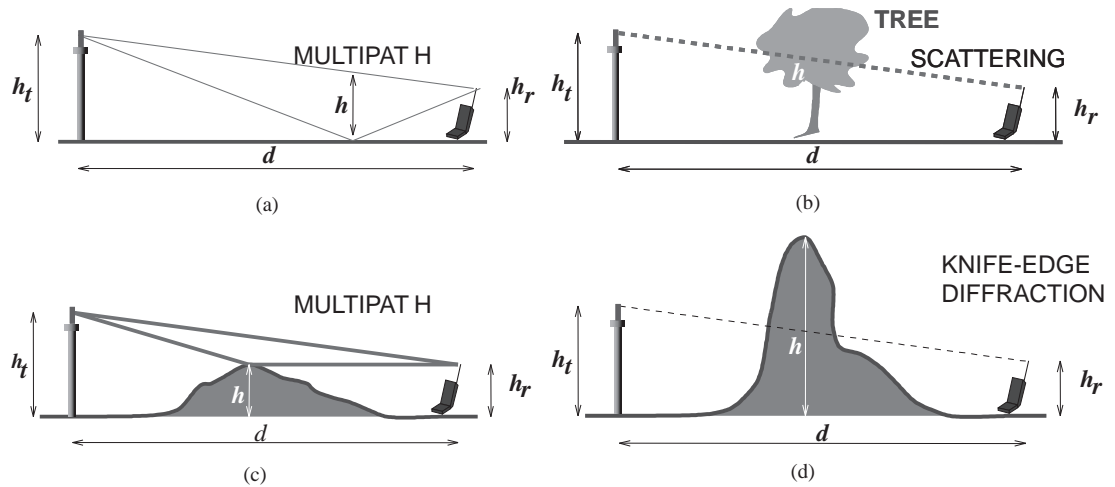
Figure 2-17 Common multipath scenarios in cellular communications: (a) ground effect; (b) transmission through a scattering such as a tree; (c) ground bounce from a hill; and (d) knife-edge diffraction.

Common multipaths encountered in cellular radio are shown in Figure 2-17. As a very rough guide, each diffraction and scattering event reduces the signal received by 20 dB. This is based on an average of measurements, but it is a good rough guide. The knife-edge diffraction scenario is shown in more detail in Figure 2-18. This case is fairly easy to analyze and can be used as a manual tool to estimate the effects of individual obstructions. The diffraction model is derived from the theory of half-infinite screen diffraction [14]. First, calculate the parameter ν from the geometry of the path using

$$\nu = -H\sqrt{\frac{2}{\lambda}\left(\frac{1}{d_1} + \frac{1}{d_1}\right)}.$$ (2.32)

Next, consult the table in Figure 2-18(b) to obtain the obstruction loss in decibels. This loss should be added (using decibels) to the otherwise-determined path loss to obtain the total path loss. Other losses such as reflection cancelation still apply, but are computed independently for the path sections before and after the obstruction.

2.6.1 Multipath and Path Loss

Path loss is affected by the antenna gain, atmospheric conditions, and multipath effects that increase the roll-off of signals above the free-space roll-off which is inversely proportional to the square of distance [15, 16]. The first and last factors in the RF link are the antennas, and their impact is captured by antenna gain. Antenna gain is generally measured in an anechoic chamber, which is a room that has negligible reflections, and so

(a)

(b)
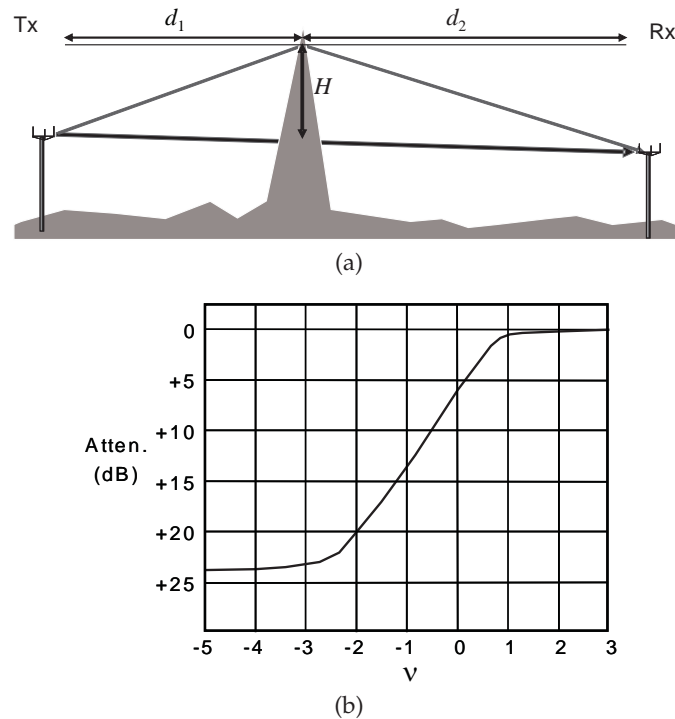
Figure 2-18 Knife-edge diffraction: (a) definition of variables used in calculations; and (b) chart for determining attenuation using the parameter derived in Equation (2.32).

the fall off of power is $1/d^2$, where d is the distance from the antenna. An alternative is to conduct the measurements in an open-field site generally using a very large flat metal ground plane. It is straightforward to calculate the power density at any distance using Equation (2.23).

The **propagation loss**, or **path loss**, of a LOS path is inversely proportional to the increase in the area of a region on the surface sphere subtending a constant solid angle as the radius d of the sphere increases. This loss is captured in Equation (2.23). There are several formulas used for path loss. The one that comes directly from Equation (2.23) is the LOS path loss of the first kind:

$$^1L_{\mathrm{PATH,LOS}} = 4\pi d^2 \ . \tag{2.33}$$

More commonly an equivalent form that includes wavelength is used and is derived from the equation for LOS link loss in Equation (2.28). Link loss is defined as the ratio of the power input to the transmit antenna (P_T) to the power delivered by the receive antenna (P_R). Rearranging Equation (2.28) and taking logarithms yields the total link loss, L_{LINK}, between the power input to the transmit antenna and the power output from the receive

antenna is (in decibels):

$$L_{\text{LINK,LOS}} = 10\log\left(\frac{P_T}{P_R}\right) = 10\log\left(\frac{P_T}{P_D A_R}\right) \tag{2.34}$$

$$= 10\log\left[P_T\left(\frac{4\pi d^2}{P_T G_T}\right)\left(\frac{4\pi}{\lambda^2 G_R}\right)\right] \tag{2.35}$$

$$= 10\log\left[\left(\frac{1}{G_T G_R}\right)\left(\frac{4\pi d}{\lambda}\right)^2\right] \tag{2.36}$$

$$= -10\log G_T - 10\log G_R + 20\log\left(\frac{4\pi d}{\lambda}\right) , \tag{2.37}$$

where the path loss in decibels is

$$L_{\text{PATH,LOS}} = 20\log\left(\frac{4\pi d}{\lambda}\right) . \tag{2.38}$$

This is the preferred form of the expression for path loss, as it can be used directly in calculating link loss using the antenna gains of the transmit and receive antennas without the exercise of calculating the **effective aperture size** of the receive antenna.

Multipath effects result in losses that are proportional to d^n so that the general path loss, including multipath effects, is (in decibels):

$$L_{\text{PATH}} = L_{\text{PATH,LOS}} + \text{excess loss} \tag{2.39}$$

$$= 20\log\left(\frac{4\pi d}{\lambda}\right) + 10(n-2)\log\left(\frac{d}{1\,\text{m}}\right) , \tag{2.40}$$

where the distance d is normalized to 1 m. In decibels, Equation (2.40) becomes

$$L_{\text{PATH}} = 10n\log[d/(1\,\text{m})] + C , \tag{2.41}$$

where C is a constant that captures the effect of wavelength. Here,

$$C = 20\log(4\pi/\lambda) . \tag{2.42}$$

Combining this with Equation (2.37) yields the link loss between the input to the transmit antenna and the output of the receive antenna:

$$L_{\text{LINK}}(\text{dB}) = -G_T - G_R + 10n\log[d/(1\,\text{m})] + C . \tag{2.43}$$

As you can imagine, a few constants, here n and C, cannot capture the full complexity of the propagation environment. Many models have been developed to capture particular environments better and incorporate mast height, experimental correction factors, and statistical limits. One of these models is the **Okumura-Hata model**, discussed in Section 2.6.5 on Page 91.

Transmitter	+43	dBm TX output
Trans. Line	- 3 = +40	dB line efficiency dBm to antenna
Antenna	+13 = +53	dB antenna gain dBm ERP
	-158	dB path attenuation
	= -105	dBm if intercepted by dipole antenna
Antenna	+13 = - 92	dB antenna gain dBm into line
Trans. Line	- 3 = - 95	dB line efficiency dBm to receiver
Receiver		
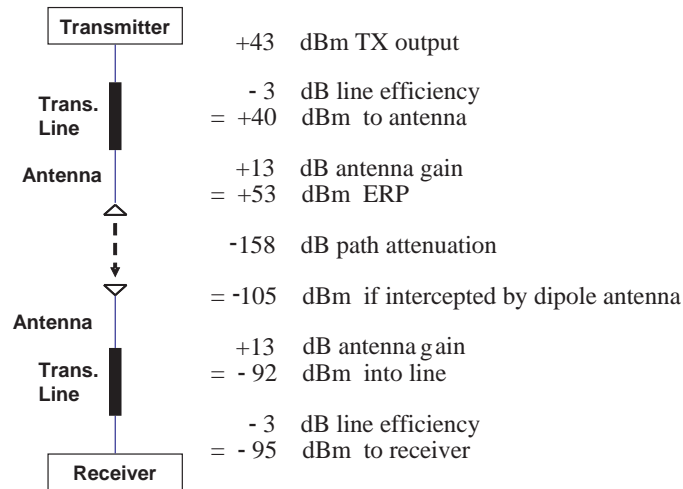
Figure 2-19   RF link example.

The RF link calculation can also include losses in cables connecting the RF electronics to the antennas. An example of such an RF link calculation is given in Figure 2-19.

EXAMPLE 2. 5   Power Density

A communication system operating in a dense urban environment has a power density roll-off of $1/d^{3.5}$ between the base station transmit antenna and the mobile receive antenna. At 10 m from the transmit antenna, the power density is 0.3167 W/m². What is the power density at the receive antenna located at 1 km from the base station.

Solution: $P_D(10\text{ m}) = 0.3167\text{ W} \cdot \text{m}^{-2}$ and let $d_c = 10$ m, so at $d = 1$ km, the power density $P_D(1\text{ km})$ is obtained from

$$\frac{P_D(1\text{ km})}{P_D(10\text{ m})} = \frac{d_c^{3.5}}{d^{3.5}} = \frac{10^{3.5}}{1000^{3.5}} = 10^{-7} \,, \tag{2.44}$$

so

$$P_D(1\text{ km}) = P_D(10\text{ m}) \cdot 10^{-7} = 31.7\text{ nW/m}^2 \,. \tag{2.45}$$

EXAMPLE 2. 6   Link Loss

A 5.6 GHz communication system uses a transmit antenna with an antenna gain G_T of 35 dB and a receive antenna with an antenna gain G_R of 6 dB. If the distance between the antennas is 200 m. What is the link loss if multipath effects result in the power density reducing as $1/d^3$. The link loss here is between the input to the transmit antenna and the output from the receive antenna.
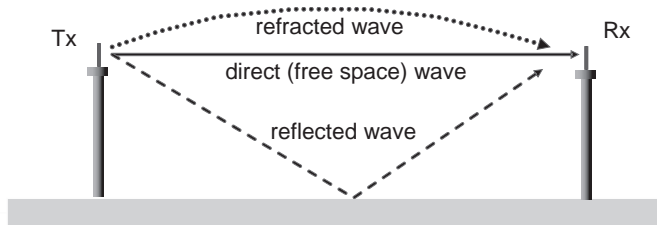
Figure 2-20 Three propagation characteristic paths: line of sight, reflection, and refraction.

Solution:

The link loss is provided by Equation (2.43),

$$L_{\mathrm{LINK}}(\mathrm{dB}) = -G_T - G_R + 10n\log[d/(1\ \mathrm{m})] + C\ ,$$

and C comes from Equation (2.42), where λ = 5.36 cm. So

$$C = 20\log\left(\frac{4\pi}{\lambda}\right) = 20\log\left(\frac{4\pi}{0.0536}\right) = 47.4\ \mathrm{dB}.$$

With n = 3 and d = 200 m,

$$L_{\mathrm{LINK}}(\mathrm{dB}) = -35 - 6 + 10 \cdot 3 \cdot \log(200) + 47.4\ \mathrm{dB} = 75.4\ \mathrm{dB}.$$

2.6.2 Fresnel Zones

The dominant propagation paths are shown in Figure 2-20. The refracted wave path arises because of density variations in the air producing a permittivity profile that varies with height. The reflected wave from the ground can be important if the ground is too close to the propagation path. Both of these effects will be considered in this section.

   The density variation of air resulting primarily from temperature variation produces a variation in the index of refraction with height, as shown in Figure 2-21(a). This results in **beam bending** so that the dominant propagation path is the refracted wave path shown in Figure 2-20. The most convenient way of accommodating beam bending is to use a curved-earth model, as shown in Figure 2-21(b), so that subsequent calculations can use LOS considerations.

   As radio waves propagate they spread out, and in a plane perpendicular to the direction of propagation, the power density of the radio waves reduces with distance from the centerline. The spreading out is understood from Huygens principle that every point of a propagating EM wave reradiates in every direction. One of the consequences of this is that an obstruction that is not in the LOS path can still interfere with signal

(a)   (b)
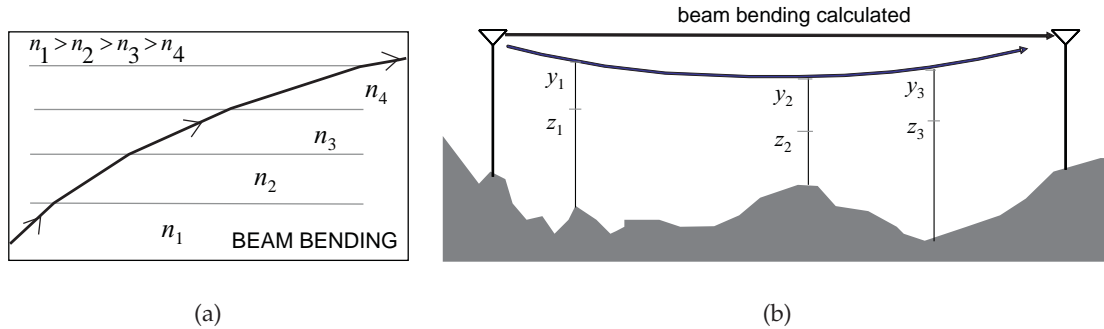
Figure 2-21   Beam bending from density variation in air: (a) profile of refraction index in air with the density of air reducing with height; (b) incorporating beam bending by using a curved-earth model.
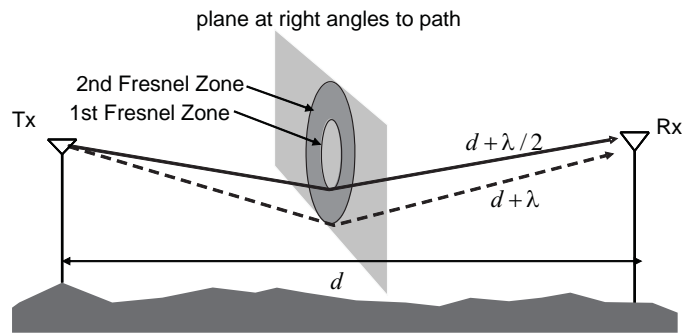
Figure 2-22   Fresnel zones in a plane perpendicular to a direct propagation path which is a distance d long.

propagation. The appropriate clearance is determined from the Fresnel zones, which are shown in Figure 2-22. The direct LOS path between the antennas has a length d. If there is a reflecting object near the LOS path, then there can be a second path between the transmit and receive antennas. The path from the first antenna to the circle defined as the first Fresnel zone and then to the second antenna has a path length $d + \lambda/2$, and so at the receive antenna this signal is $180°$ out of phase with the LOS signal and there will be cancelation. The radius of the nth Fresnel zone at point P is

$$F_n = \sqrt{\frac{n\lambda d_1 d_2}{d_1 + d_2}} \, , \qquad (2.46)$$

where d_1 is the distance from the first antenna to P, d_2 is the distance from the first antenna to P, and λ is the wavelength of the propagating signal. Ninety percent of the energy in the wave is in the first Fresnel zone. As an example, if $d_1 = d_2 = 5$ km, $F_1 = 17.66$ m at 2.4 GHz. A guideline is that an obstacle should be at a distance from the LOS path that is at least 60% of the
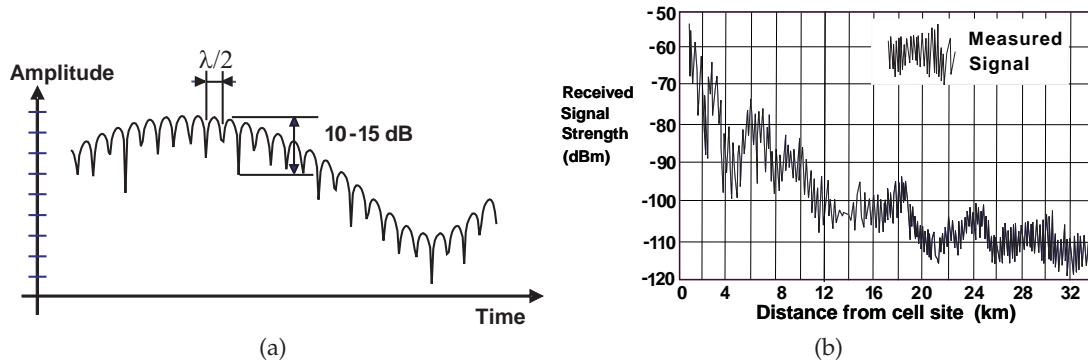
Figure 2-23   Fast fading: (a) in time as a radio and obstructions move; and (b) in distance.

first Fresnel zone radius distance. A more conservative guideline is that it should be at least one Fresnel zone distant.

Both beam bending and Fresnel zone clearance must be accounted for. In Figure 2-21, z_1, z_2, and z_3 are calculated Fresnel zone clearances and y_1, y_2, and y_3 are heights of obstacles adjusted to account for beam bending (i.e., straight-line path between antennas for a curved earth). In a point-to-point communication system, antenna heights are adjusted so that the beam just touches one or more Fresnel zone clearances.

2.6.3   Rayleigh Fading

Rayleigh fading, or **fast fading**, results from destructive cancelation as individual paths drift in and out of phase as the receiver and sources of multiple reflections, diffractions, and refractions move. Rayleigh fading is named after the statistical model that describes it. The result is the amplitude response shown in Figure 2-23(a). These are rapid fades and so are also called fast fades. Similar fading occurs with distance, as shown in Figure 2-23(b). The fades occur over distances roughly $\frac{1}{2}\lambda$ apart.

In a fixed wireless system, Rayleigh fading is due to rapidly changing atmospheric conditions, with the refractive index of small regions varying. These fades occur over a few seconds. In a mobile wireless system, the fast fades are principally due to movement of the mobile terminal unit or moving reflection and diffraction objects. Slow variations come from blockage and shadowing by large objects such as hills and buildings, as for fixed wireless systems. The time interval between fades depends on the speed of the mobile radio, but can be in the millisecond range. The main strategy for overcoming the effects of the deep fades is to boost the transmit power level by 10 to 15 dB. In the frequency domain, the fades are about $\frac{1}{2}$ MHz wide and almost independent of frequency for radios operating from hundreds of begahertz to 100 GHz. Signals that are spread
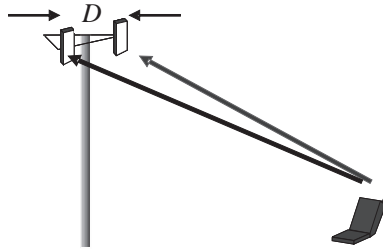
Figure 2-24 Two receive antennas used to achieve space diversity and overcome the effects of Rayleigh fading.

over a few megahertz, such as **Code Division Multiple Access** (**CDMA**) (typically 1.23 MHz) and **Wideband CDMA** (**WCDMA**) (around 5 MHz) are relatively immune to deep fades and so can operate at lower average transmit power levels than communication systems with channels narrower than a megahertz or so.

Fortunately Rayleigh fades are very short, last a small percentage of the time, and slight changes to the propagation environment can circumvent their effects. One strategy for overcoming fades is to use two receive antennas as shown in Figure 2-24. Here the signals received by two antennas separated by several wavelengths will rarely fade concurrently. One useful strategy is to switch instant by instant to whichever has the best signal. In practice, the required separation for good decorrelation is found to be 10 to 20λ or 4 to 7 m at 800 MHz or 1.5 to 3 m at 1900 MHz. This space diversity can be applied only on the receiving end of a link, so fading of the downlink with a mobile handset cannot be easily overcome.

EXAMPLE 2. 7	Radiated Power Density

In free space, radiated power density drops off with distance d as $1/d^2$. However, in a terrestrial environment there are multiple paths between a transmitter and a receiver, with the dominant paths being the direct LOS path and the path involving reflection off the ground. Reflection from the ground partially cancels the signal in the direct path, and in a semiurban environment results in an attenuation loss of 40 dB per decade of distance (instead of the 20 dB per decade of distance roll-off in free space). Consider a transmitter that has a power density of 1 W·m^{-2} at a distance of 1 m from the transmitter.

(a) The power density falls off as $1/d^n$, where d is distance and n is an index. What is n?

(b) At what distance from the transmit antenna will the power density reach 1 μW·m^{-2}?

(a) Power drops off by 40 dB per decade of distance. 40 dB corresponds to a factor of 10,000 (= 10^4). So, at distance d, $P_d = k/d^n$ (k is a constant).

Solution:

At a decade of distance, $10d$, $P_{10d} = k/(10d)^n = P_d/10000$, i.e., $\frac{k}{10^n d^n} = \frac{1}{10,000}\frac{k}{d^n}$; $10^n = 10,000 \Rightarrow n = 4$.

Note: P_d is power density.

(a) Power drops off by 40 dB per decade of distance. 40 dB corresponds to a factor of 10,000 (= 10^4). So, at distance d, $P_d = k/d^n$ (k is a constant).

At a decade of distance, $10d$, $P_{10d} = k/(10d)^n = P_d/10000$,

i.e., $\frac{k}{10^n d^n} = \frac{1}{10,000}\frac{k}{d^n}$; $10^n = 10,000 \Rightarrow n = 4$.

Note: P_d is power density.

(b) At d = 1 m, $P_d = 1\ \text{W}\cdot\text{m}^{-2}$.

At a distance x,

$$P_x = 1\mu\text{W}\cdot\text{m}^{-2} = \frac{k}{x^4}m^2 = \frac{k}{x^4}$$

$$x^4 = \frac{1}{10^{-6}} \quad \text{and so} \quad x = 31.6\ \text{m}\,.$$

2.6.4 Radio Link Reciprocity

The path loss between two antennas is exactly the same in both directions when the frequency of the signal in each direction is the same. This is radio link reciprocity. However, nearly all communication systems use different frequencies in the two directions, and so the links are not reciprocal. Also, in a mobile environment, the link will change with time, generally significantly over a few milliseconds, as the source of obstructions moves. The link characteristics in point-to-point links change much more slowly, with localized pockets of high-density air, humidity, and rainfall causing the main changes. In a cellular environment, the more significant impact is the variation in background noise level. Most of the noise is generated by other radios, and while this does not impact link loss, it does require signal power levels to be boosted to maintain the signal level relative to the level of noise and interference.

2.6.5 Propagation Model

Radio frequency propagation in the mobile environment can only be described statistically. One of the models is the Okumura-Hata model [17], which calculates the path loss as

$$L_{\text{PATH}}(\text{dB}) = 69.55 + 26.16\log f\, 13.82\log h + (44.96.55\log h)\cdot\log d + c\,,\quad (2.47)$$

where f is the frequency (in MHz), d is the distance between the base station and terminal (in km), H is the effective height of the base station antenna (in m), and c is an environment correction factor ($c = 0$ dB in a dense urban area, $c = -5$ dB in an urban area, $c = -10$ dB in a suburban area, and $c = -17$ dB
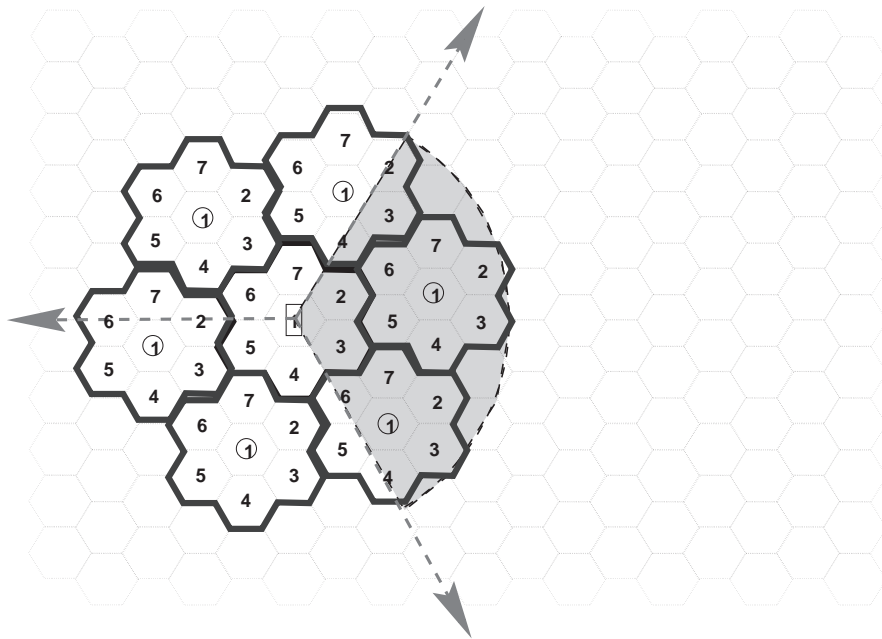
Figure 2-25   A cellular radio system with a trisector antenna, such as in Figure 2-26, showing the area of possible interference as a shaded region.

in a rural area).

More sophisticated characterization of the propagation environment uses ray-tracing models to follow individual propagation paths. The ray-tracing models are based on deterministic methods using terrain data path profile calculations accounting for obstruction and reflection analyses. Each refraction and reflection event is characterized either experimentally or through detailed EM simulations. Appropriate algorithms are applied for best compliance with radio physics. Commonly required inputs to these models include frequency, distance from the transmitter to the receiver; effective base station height, obstacle height and geometry, radius of the first Fresnel zone, forest height/roof height, distance between buildings, arbitrary loss allowances based on land use (forest, water, etc.), and loss allowance for penetration of buildings and vehicles.

2.7   Radio Link Interference

Radio systems using the same channel are geographically spaced to control interference. In a cellular system, the coverage areas are arranged in cells represented by the hexagons in Figure 2-25. The actual shape of the cells is influenced by obstructions such as hills and buildings, but the hexagonal
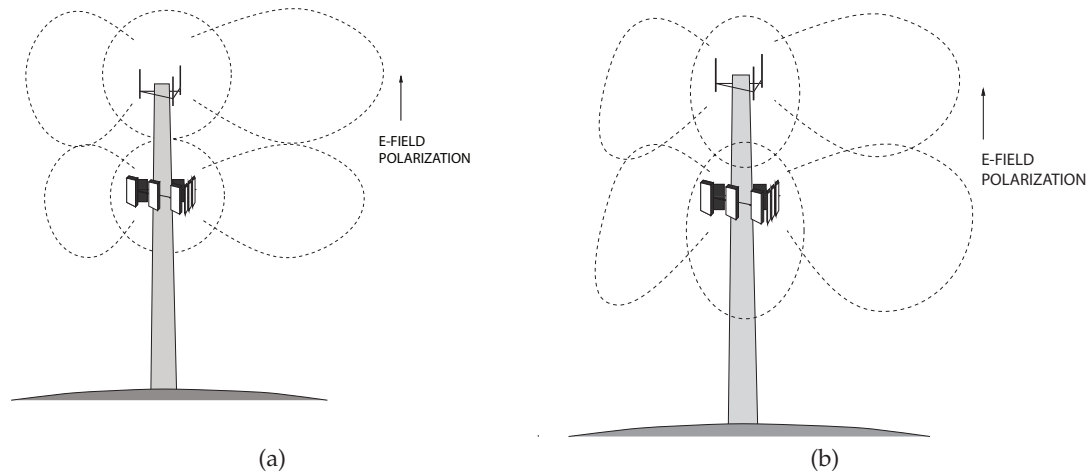
Figure 2-26 Base station tri-sector antenna with separate transmit and receive in: (a) normal configuration; and (b) look down mode. antennas.

shape is used to convey geographic partitioning. In a cellular system the cells are arranged in clusters such as the 3-, 7-, and 12-cell clusters shown, and the total number of channels available is divided among the cells in a cluster with the full set of frequency channels repeated in each cluster. The size of the clusters is the major component of what is called the **frequency reuse** plan. So a three-cell cluster has a spacing of approximately one cell diameter to the next cell using the same frequency channels. The signal level transmitted from the original cell will interfere with the signal in its corresponding cell in adjacent clusters. The level of interference is reduced with 7-cell and then 12-cell clusters. There is also background noise coming from cosmic sources as well as artificial sources, but in a cellular system, interference from other radios operating in the same system tends to dominate.

The use of directional antennas at the base station increases the SIR. The interference pattern obtained using a **trisector antenna** (each segment of the antenna providing 120° of coverage) is shown in Figure 2-25. The trisectorial antenna can be arranged so that the transmit and receive antennas are separated (see Figure 2-26(a)), and can also be arrange to tilt the coverage toward the ground, as shown in Figure 2-26(b). It is also possible to have smaller cells (achieved possibly by relatively low-power transmissions) to provide higher levels of coverage at critical regions such as intersections of roads, as shown in Figure 2-27. These smaller cells are referred to as microcells or picocells.

Antennas with higher-order sectoring are also used. Figure 2-28 is a four-section (or quadrasectional) antenna providing better coverage at a road
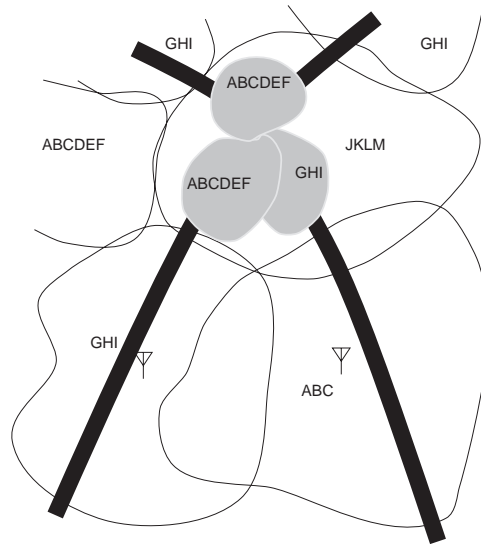
Figure 2-27   Nested cells using sectored antennas in combination with conventional antennas to provide a microcell boosting capacity.
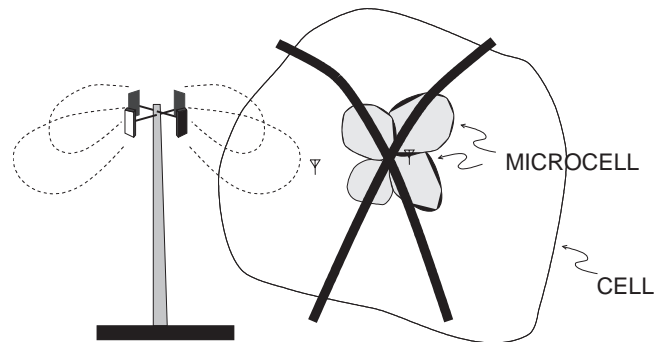
Figure 2-28   Quadrasectional antenna and additional coverage at an intersection.

intersection, which naturally lends itself to division by four. With such an antenna, more users can be supported, but more importantly there would be much lower handoff from one sector to the next. An extreme situation occurs with **Wide Wireless Area Networks** (**WWANs**) where 12 sectors or more are being used, but in this case the terminals are usually not mobile and sector-to-sector handoff is relatively low.
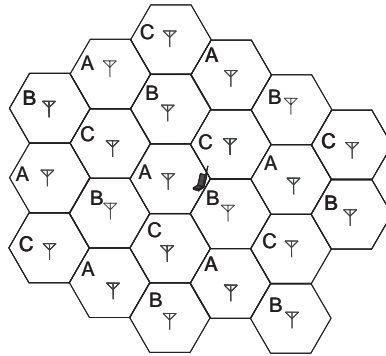
Figure 2-29   A three-cell cluster with a mobile unit at the edge of a cell.

EXAMPLE 2. 8   Cellular Interference

In a cellular system, a signal is intentionally transmitted from a base station nominally located in the center of a cell to a mobile unit in the same cell. However, nearby transmitters using the same channel cause interference. In Figure 2-29, a mobile unit is located at the edge of a cell and uses frequency channel A. Many nearby transmitters also operate using channel A and the six nearest transmitters can be considered as causing significant interference. Consider that the mobile unit is a distance r from its cell's transmitter along the line connecting two channel A base stations; that the transmitters all operate at the same power level; and that the distance between base stations operating using channel A is $3r$. This three-cell cluster operates in a suburban area and the power density drops off with distance d as $1/d^3$ due to multipath effects. What is the SIR at the mobile unit? Express your answer as both a power ratio and in decibels.

Solution:

There are seven close towers transmitting signals on the same channel. Call these A_1–A_7 . A_1 is the desired signal and A_2–A_7 are interferers. The distances from the transmitters to the mobile units are d_1–d_7, and the powers from the transmitters received at the mobile unit are P_1–P_7.
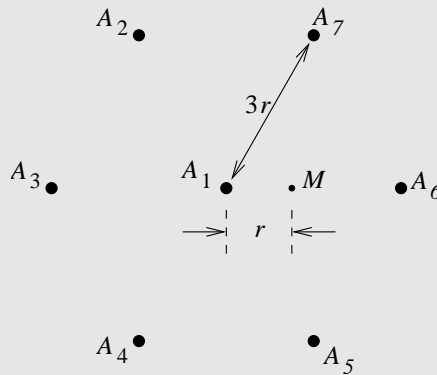
The SIR is

$$\text{SIR} = \frac{P_1}{P_2 + P_3 + P_4 + P_6 + P_7} \; .$$

This problem requires us to determine d_2–d_7. We know that $d_1 = r$. Also, the distance from A_2 to A_7 is $3r$. Now $d_3 = 3r + r = 4r, \quad d_6 = 3r - r = 2r$.
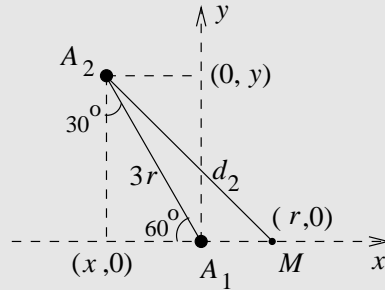
Consider d_2:

$$d_2 = \left[(x+r)^2 + Y^2\right]^{1/2} = \left(2.5^2 + 2.6^2\right) r = 3.607r = d_4$$
$$X = 3r \sin 30° = 1.5r \quad \text{and} \quad Y = 3r \cos 30° = 2.6r \; .$$

Similarly, $d_7 = d_5 = \left[(1.5-1)^2 + 2.6^2\right]^{1/2} = 2.648r$

$$P_2/P_1 = \left(\frac{d_1}{d_3}\right)^3 = 1/3.607^3 = 0.0213 = P_4/P_1$$

$$P_7/P_1 = \left(\frac{d_1}{d_7}\right)^3 = 1/2.648^3 = 0.0539 = P_5/P_1$$

$$P_3/P_1 = \left(\frac{d_1}{d_3}\right)^3 = 1/4^3 = 0.0156$$

$$P_6/P_1 = \left(\frac{d_1}{d_6}\right)^3 = 1/2^3 = 0.1250$$

$$\text{SIR} = (0.0213 + 0.0156 + 0.0213 + 0.0539 + 0.1250 + 0.0539)^{-1} = 3.44 = 5.36 \text{ dB.}$$

2.8 Summary

This chapter provided an overview of radio signal propagation. The aim was to provide a broad perspective of the signaling environment and the impact on signal level and SIR levels. This is part of what drives the design of communication systems. The environment impacts RF design by determining the standards to which they must be designed.

2.9 Exercises

1. The output stage of an RF front end consists of an amplifier followed by a filter and then an antenna. The amplifier has a gain of 27 dB, the filter has a loss of 1.9 dB, and of the power input to the antenna, 35% is lost as heat due to resistive losses. If the power input to the amplifier is 30 dBm, calculate the following: [Parallels the example on Page 838.]

   (a) What is the power input to the amplifier in watts?
   (b) Express the loss of the antenna in dB.
   (c) Express the gain of the antenna in dB.
   (d) What is the total gain of the RF front end (amplifier + filter + antenna)?
   (e) What is the total power radiated by the antenna in dBm?
   (f) What is the total power radiated by the antenna in mW?

2. The output stage of an RF front end consists of an amplifier followed by a filter and then an antenna. The amplifier has a gain of 27 dB, the filter has a loss of 1.9 dB, and of the power input to the antenna, 45% is lost as heat due to resistive losses. If the power input to the amplifier is 30 dBm, calculate the following:

   (a) What is the power input to the amplifier in watts.
   (b) Express the loss of the antenna in decibels?
   (c) What is the total gain of the RF front end (amplifier + filter + antenna)?
   (d) What is the total power radiated by the antenna in dBm?
   (e) What is the total power radiated by the antenna in milliwatts?

3. Consider a point-to-point communication system. Parabolic antennas are mounted high on a mast so that ground effects do not exist, thus power falls off as $1/d^2$. The gain of the transmit antenna is 20 dB and the gain of the receive antenna is 15 dB. The distance between the antennas is 10 km. The effective area of the receive antenna is 3 cm^2. If the power input to the transmit antenna is 600 mW, what is the power delivered at the output of the receive antenna?

4. Consider a 28 GHz point-to-point communication system. Parabolic antennas are mounted high on a mast so that ground effects do not exist, thus power falls off as $1/d^2$. The gain of the transmit antenna is 20 dB and the gain of the receive antenna is 15 dB. The distance between the antennas is 10 km. If the power output from the receive antenna is 10 pW, what is the power input to the transmit antenna?

5. In the figure below there are two transmitters, Tx$_1$ and Tx$_2$, operating at the same power level, and one receiver, Rx. Tx$_2$ is an intentional transmitter and its signal is intended to be received at Rx. Tx$_1$ uses the same frequency channel as Tx$_2$, but it transmits an interfering signal. [Parallels Example 2.1 on Page 62.]
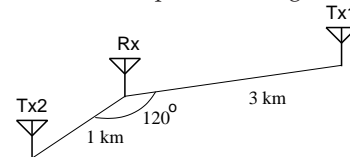
   (a) Assume that antennas are omnidirectional (they receive and transmit a signal equally from and to all directions) and, being in a semiurban area, that the transmitted power density drops off as $1/d^3$, where d is the distance from the transmitter. Calculate the SIR at Rx. Express your answer in decibels.
   (b) Now consider a directional antenna at Rx while the transmit antennas remain omnidirectional antennas. The antenna at Rx is directed toward the transmitter Tx$_2$ and the antenna gain is 6 dB. In the direction of Tx$_1$ the effective antenna gain is -3 dB. Now recalculate the SIR. Express your answer in decibels.

6. On a resonant antenna, a large current is established by creating a standing wave. The current peaking that thus results establishes a strong electric field (and hence magnetic field) that radiates away from the antenna. A typical dipole loses 15% of the power input to it as resistive (I^2R) losses and has an antenna gain of 10 dB measured at 50 m. Consider a base station dipole antenna that has 100 W input to it. Also consider that the transmitted power density falls off with distance d as $1/d^3$. Hint, calculate the power density at 50 m. [Parallels Example 2.3 on Page 78.]

   (a) What is the input power in dBm?
   (b) What is the power transmitted in dBm?
   (c) What is the power density at 1 km? Express your answer as W/m^2.
   (d) What is the power captured by a receive antenna (at 1 km) that has an effective antenna aperture of 6 cm^2? Express your answer in first dBm and then watts.
   (e) If the background noise level captured by the antenna is 1 pW, what is the SNR in decibels? Ignore interference that comes from other transmitters.

7. Stacked dipole antennas are often found at the top of cellphone masts, particularly for large cells and operating frequencies below 1 GHz. These antennas have an efficiency that is close to 90%. Consider an antenna that has 40 W of input power, an antenna gain of 10 dB, and transmits a signal at 900 MHz.

   (a) What is the EIRP in watts?
   (b) If the power density drops as $1/d^3$, where d is the distance from the transmit tower, what is the power density at 1 km if the power density is 100 mW/m^2 at 10 m?

8. Consider an 18 GHz point-to-point communication system. Parabolic antennas are mounted on masts and the LOS between the antennas is just above the tree line. As a result, power falls off as $1/d^3$, where d is the distance between the antennas. The gain of the transmit antenna is 20 dB and the gain of the receive antenna is 15 dB. The antennas are aligned so that they are in each other's

main beam. The distance between the antennas is 1 km. The transmit antenna is driven by a power amplifier with an output power of 100 W. The amplifier drives a coaxial cable that is connected between the amplifier and the transmit antenna. The cable loses 75% of its power due to resistive losses. On the receive side, the receive antenna is directly connected to a masthead amplifier with a gain of 10 dB and then a short cable with a loss of 3 dB before entering the receive base station.

   (a) Draw the signal path.
   (b) What is the loss and the gain of the transmitter coaxial cable in decibels?
   (c) What fraction of the power input to the receive coaxial cable is lost in the receive cable? Express your answer as a percentage.
   (d) Express the power of the transmit amplifier in dBW and dBm.
   (e) What is the propagation loss in decibels?
   (f) Determine the total power delivered to the receive base station. Express you answer in watts.

9. Consider a point-to-point communication system. Parabolic antennas are mounted high on a mast so that ground effects are minimal. Thus power density falls of as $1/d^{2.3}$ where d is the distance from the transmitter. The gain of the transmit antenna is 15 dB and the gain of the receive antenna is 12 dB. These gains are normalized to 1 m. These antenna gains are normalized to a distance of 1 m. The distance between the antennas is 15 km. The output power or the receive antenna must be 1 pW? The RF frequency is 2 GHz; treat the receive and transmit antennas as lossless.

   (a) What is the received power in dBm?
   (b) What is the path loss in decibels?
   (c) What is the link loss in decibels?
   (d) Using the link loss, calculate the input power, P_T, of the transmitter. Express the answer in dBm.
   (e) What is the aperture area of the receiver in square meters?

(f) Determine the radiated power density at the receiver in terms of the transmitter input power. That is, if P_T is the power input to the transmit antenna, determine the power density, P_D, at the receive antenna where $P_D = xP_T$. What is x in units of m^{-2}?

(g) Using the power density calculation and the aperture area, calculate P_T in watts. This should be the same as the answer you calculated in part (c).

(h) What is P_T in dBm?

(i) What is the total power radiated by the transmit antenna in dBm?

10. Describe the following concepts and, if applicable, the reason behind their usage.

(a) Rayleigh fading.

(b) Wireless LAN systems are operating at 2.4 GHz and 5.6 GHz and systems are being developed at 40 GHz and 60 GHz. Contrast the performance of these schemes inside a building in terms of range and explain your answer.

(c) At 60 GHz the atmosphere strongly attenuates a signal. Discuss the origin of this attenuation and indicate an advantage and a disadvantage of this property.

11. A transmitter and receiver operating at 2 GHz are at the same level, but the direct path between them is blocked by a building and the signal must diffract over the building for a communication link to be established. This is a classic knife-edge diffraction situation. The transmit and receive antennas are separated from the building by 4 km and the building is 20 m higher than the antennas (which are at the same height). Consider that the building is very thin. It has been found that the path loss can be determined by considering loss due to free-space propagation and loss due to diffraction over the knife edge.

(a) What is the additional attenuation (in decibels) due to diffraction?

(b) If the operating frequency is 100 MHz, what is the attenuation (in decibels) due to diffraction?

(c) If the operating frequency is 10 GHz, what is the attenuation (in decibels) due to diffraction?

12. A hill is 1 km from a transmit antenna and 2 km from a receive antenna. The receive and transmit antennas are at the same hight and the hill is 200 m above the height of the antennas. What is the additional loss caused by diffraction over the top of the hill? Treat the hill as causing knife-edge diffraction.

13. A transmit antenna and a receive antenna are separated by 10 km.

(a) What is the radius of the first Fresnel zone?

(b) What is the radius of the second Fresnel zone?

(c) To ensure LOS propagation, what should the clearance be from the direct line between the antennas and obstructions such as hills and vegetation?

CHAPTER 3

Radio Frequency Systems

3.1 Introduction

Radio frequency systems drive the requirements of microwave and RF circuits and the capabilities of RF and microwave circuits fuel the evolution of RF systems. This relationship has become more intertwined, requiring communication engineers and circuit designers to have a broad appreciation of technology, communication principles, and circuit attributes. In this chapter we explore the development of RF systems, addressing principally the relationship of circuits and communications. Similar relationships have evolved for radar used in detection and ranging. Other radio systems are used in navigation, astronomy (radio telescopes), and heating (e.g., microwave ovens). The history of RF communications has led to the current

mode of operation, allocating a narrow slice of EM spectrum to one or a few users. This has dictated the need for stable oscillators and high-rejection filters, for example.

3.2   Communication Over Distance

Communicating using EM signals has been an integral part of society since the transmission of the first telegraph signals over wires in the early 19th century. This development derived from an understanding of magnetic induction based on the experiments of Faraday[1] in 1831 [18] in which he investigated the relationship of current and magnetic field.

3.2.1   Telegraph

While traveling by ship back to the United States from Europe in 1832, Samuel Morse learned of Faraday's experiments and conceived of an EM telegraph[2]. He sought out partners in Leonard Gale, a professor of science at New York University, and Alfred Vail, "skilled in the mechanical arts," who constructed the telegraph models used in their experiments. In 1835 this collaboration led to an experimental version transmitting a signal over 16 km of wire. Morse was not alone in imagining an EM telegraph, and in 1837 Charles Wheatstone opened the first commercial telegraph line between London and Camden Town, England, a distance of 2.4 km. Subsequently, in 1844, Morse designed and developed a line to connect Washington, DC, and Baltimore, Maryland. This culminated in the first public transmission on May 24, 1844, when Morse sent a telegraph message from the Capitol in Washington to Baltimore. This event is recognized as the birth of communication over distance using wires. This rapid transition from basic research (Faraday's experiment) into electromagnetism to a fielded transmission system has been repeated many times in the evolution of wireless technology.

   The early telegraph systems used EM induction and multicell batteries that were switched in and out of circuit with the long telegraph wire to create pulses of current. We now know that these current pulses created magnetic fields that propagated along the wire together with an accompanying electric field. In 1840 Morse applied for a U.S. patent for "Improvement in the Mode of Communicating Information by Signals by the Application of Electro-Magnetism Telegraph" which described "lightning wires" and "Morse code." By 1854, 37,000 km of wire crossed the United States, and had a profound effect on the development of the country. Railroads made early extensive use of telegraph and a new industry was created. In the United States the telegraph industry was dominated

---

[1]   Faraday's name is immortalized in the Farad, the unit of capacitance.
[2]   Morse had been studying art; he taught art at New York University.

Table 3-1 International Morse code.

Symbol	Code	Symbol	Code
1	. - - - -	I	. .
2	. . - - -	J	. - - -
3	. . . - -	K	- . -
4	. . . . -	L	. - . .
5	. . . . .	M	- -
6	-	N	- .
7	- - . . .	O	- - -
8	- - - . .	P	. - - .
9	- - - - .	Q	- - . -
0	- - - - -	R	. - .
A	. -	S	. . .
B	- . . .	T	-
C	- . - .	U	. . -
D	- . .	V	. . . -
E	.	W	. - -
F	. . - .	X	- . . -
G	- - .	Y	- . - -
H	. . . .	Z	- - . .

by Western Union (the name coming from uniting the western and eastern telegraphs), which became one of the largest companies in the world. Just as with the telegraph, the history of wired and wireless communication has been shaped by politics, business interests, market risk, entrepreneurship, and patent ownership as much by the technology itself.

The first telegraph signals were just short bursts and slightly longer bursts of noise using what is known as Morse code in which sequences of dots, dashes, and pauses represent numbers and letters (Table 3-1).[3] The speed of transmission was determined by an operator's ability to key and recognize the codes. Information transfer using EM signals in the late 19th century was therefore about 5 **bits per second** (5 **bps**). Morse himself achieved 10 words per minute. It was not long before telegraph signals were sent wirelessly. By the end of the 19th century, wireless telegraph signals

---

[3] Morse code uses sequences of dots, dashes, and spaces to represent 26 letters and 10 numbers. The time of a dash (or "dah") is three times the length of a dot (or "dit"). An "L" is a long dash equal in length to one dot-one dash. For zero, either the code for "O" or for 'T" is used. Between letters there is a small gap. For example, the Morse code for PI is ".–. .." . Between words there is a slightly longer pause and between sentences an even longer pause. Table 3-1 lists the International Morse code adopted in 1848. The original Morse code developed in the 1830s is now known as "American Morse code" or "Railroad code." The 'Modern International Morse Code' extends the International Morse code with dot-dash sequences for non-English letters and special symbols.

were being sent over considerable distances, but there were a handful of transmitters and receivers in just a few countries in the world. Much has happened since, with now more than four billion individuals[4] on the planet regularly transmitting information wirelessly at **bit rates** of 10 thousand bits per second (**kbps**), and many much higher (up to 2 million bits per second (**Mbps**) or so), over long distances.

Individuals use transmitters and receivers that can be smaller than an infants hand, contain hundreds of millions of transistors, and provide wireless connectivity in nearly any location. Many hundreds of millions of people also transmit data over **Wireless Local Area Networks** (**WLANs**) and **Wireless Metropolitan Area Networks** (**WMANs**) at speeds exceeding 50 Mbps. The history of radio communications has been remarkable and nearly every aspect of electrical engineering has been involved. The most important factor of all, at least in the last few decades which have seen explosive growth, is that consumers are prepared to part with a large portion of their income to have untethered connectivity. This overwhelming desire for ubiquitous communication surprised even the most optimistic proponents of personal wireless communication. The growth of wireless has also led to clear productivity gains in industry. Wireless communication is now a significant part of every country's economy and governments are very involved in setting standards and protecting the competitiveness of their own industries.

3.2.2 The Origins of Radio

Morse discovered that he could send pulses along a pair of wires by connecting and disconnecting a battery coupled to a coil to send pulses that were detected at a distance by a galvanometer completing the wired circuit. In the 1850s Morse began to experiment with wireless transmission, but this was still based on the principle of conduction. He used a flowing river, which as we now know, is a medium rich with ions, to carry the charge. He set up a series connection of a metal plate, a battery, a Morse key, and a second metal plate. This formed the transmitter circuit. The metal plates were inserted into the water and separated by a distance considerably greater than the width of the river. On the other side of the river, metal plates were placed directly opposite the transmitter plates and this second set of plates was connected by a wire to a galvanometer in series. This formed the receive circuit, and electric pulses established by the transmitter resulted in the charge being "transmitted" across the river by conduction and the pulses subsequently detected by the galvanometer. This was the first wireless transmission using electromagnetism.

Morse relied entirely on conduction to achieve wireless transmission and

---

[4] The number of cellular phone subscribers passed the 4 billion mark in 2008.

we now know that we need alternating electric and magnetic fields to propagate information over distance without change carriers. The next steps in the progress to radio were experiments in induction. These culminated in an experiment by Loomis described on the historical marker at Beorse Deer Mountain, Virginia, US: "Forerunner of Wireless Telegraphy - From nearby Bear's Den Mountain to the Catoctin Ridge, a distance of fourteen miles, Dr. Mahlon Loomis, dentist, sent the first aerial wireless signals, 1866-73, using kites flown by copper wires. Loomis received a patent in [July] 1872 [For Improvements in Telegraphing]." The transmitter kite had a Morse key at the ground end and an electric potential would have been developed between the ground and the kite itself. Closing the key resulted in current flow along the wire and this created a magnetic field circulating around the wire. The magnetic fields spread out and induced a current in the receive kite and this was detected by a galvanometer.

Loomis, in 1865, is believed to have been the first to achieve nonconducting EM wireless communication. But this was not radio. Loomis mistakenly believed that there was a conduction layer in the upper atmosphere and that this completed his circuit. This was the effect that he was trying to exploit, but we know that the kite acquired a high potential from the surrounding clouds. When the wire to ground was closed, current flowed; that is, the current in one wire was producing a large magnetic field that induced current in the receive wire. However, there is not much electric field produced and an EM wave is not transmitted. As such, the range of this system is very limited. Practical wireless communication requires something more than induction, it needs an EM wave at a high-enough frequency that it can be efficiently generated by short wires; such a frequency is known as RF.

About the same time as Loomis' induction experiments, in 1864, James Maxwell [19] laid the foundations of modern EM theory in his work, *A Treatise on Electricity and Magnetism* [20]. Maxwell theorized that electric and magnetic fields are different manifestations of the same phenomenon. The revolutionary conclusion is that if they are time varying then they would travel through space in a wave.[5] This insight was accepted almost immediately by many people and initiated a large number of endeavors in several related fields: in induction, in the generation of an oscillating signal, and in electromagnetism in general. The period of 1875 to 1900 was a period of tremendous innovation in wireless communication. On November 22, 1875, Edison observed EM sparks. Previously sparks were considered to be an induction phenomenon, but Edison thought that he was producing a new kind of force, which he called the etheric force. He believed that this would enable communication without wires. To put this in context, the telegraph

---

[5]   A copy of a biography of Maxwell, written in 1882 by Lewis Cambell, can be downloaded from the Sonnet Web site: www.sonnetusa.com under "products."

was invented in the 1830s and the telephone was invented in 1876.

The next stage leading to radio was orchestrated by **D.E. Hughes** beginning in 1879. Hughes experimented with a spark gap and reasoned that in the gap there was a rapidly alternating current and not a single, unidirectional constant current as others of his time believed. The electric oscillator was born. The spark gap transmitter was augmented with a clockwork mechanism to interrupt the transmitter circuit and produce pulsed radio signals. He used a telephone as a receiver and walked around London and detected the transmitted signals over distance. Hughes noted that he had good reception at 180 feet. Hughes publicly demonstrated his "radio" in 1870 to the Royal Society, but the eminent scientists of the society determined that the effect was simply due to induction. This discouraged Hughes from going any further. However, Hughes has a legitimate claim to having invented radio, mobile digital radio at that, and probably was transmitting pulses on a 100 kHz or slightly higher carrier. In Hughes' radio the RF carrier was produced by the spark gap oscillator and the information was coded as pulses. It was a small leap to move from the clockwork RF pulse generator to a Morse key-based pulsing system.

The invention of practical radio can be attributed to many people, beginning with **Heinrich Hertz**, who in the period from 1885 to 1889 successfully verified the essential prediction of Maxwell's equations that EM energy could propagate through the atmosphere. Hertz was much more thorough than Hughes and his results were widely accepted. In 1891 Tesla developed what we now call the Tesla coil, which is a transformer with a primary and a secondary coil, one inside the other. When one of the coils is charged, a large voltage is produced across the terminals of the other coil. Tesla pursued the application of his coils to radio and realized that the coils could be tuned so that the resulting resonance greatly amplified a radio signal.

The next milestone was the establishment of the first practical radio system by Marconi, with experiments beginning in 1894. Oscillations were produced in a spark gap which were amplified by a Tesla coil. The work culminated in the transmission of telegraph signals across the Atlantic (from Ireland to Canada) by Marconi in 1901. In 1904, crystal radio kits to detect wireless telegraph signals could be readily purchased.

Spark gap transmitters could only send pulses of noise and not voice. One generator that could be amplitude modulated, the simplest way to modulate with a time-varying signal, was an alternator. At the end of the 19th century, readily available alternators produced a 60 Hz signal. Reginald Fessenden attempted to make a higher-frequency alternator and the best he achieved operated at 1 kHz. Fessenden realized that Maxwell's equations indicated that radiation increased dramatically with frequency and so he needed a much higher frequency signal source. Under contract, General Electric developed a 2 kW, 100 kHz alternator designed by Ernst
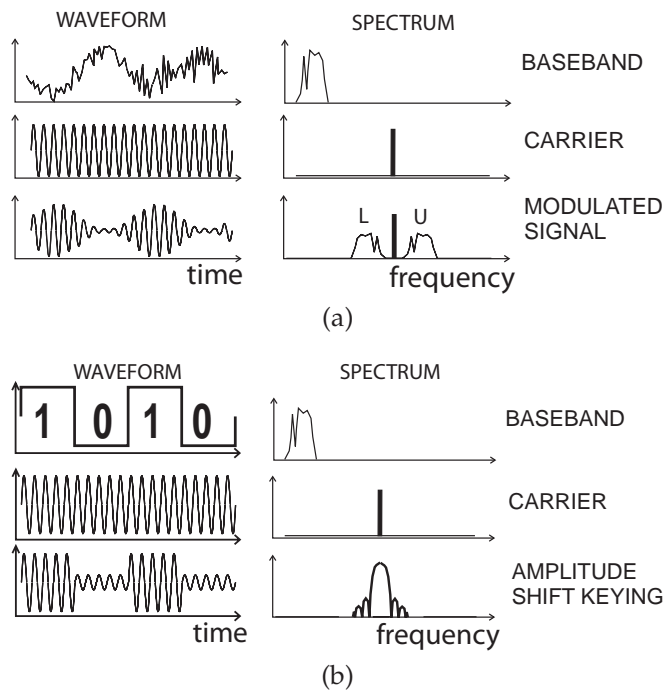
Figure 3-1 Waveform and spectra of simple modulation schemes: (a) Amplitude Modulation (AM); and (b) Amplitude Shift Keying (ASK) modulation. The modulating signal is called the baseband signal.

Alexanderson. With this alternator, the first radio communication of voice occurred on December 23, 1900, in a transmission by Fessenden from an Island in the Potomac River, near Washington, DC. Then on December 24, 1906, Fessenden transmitted voice from Massachusetts to ships hundreds of miles away in the Atlantic Ocean. This milestone is regarded as the beginning of the radio era.

Marconi subsequently purchased 50 and 200 kW Alexanderson alternators for his trans-atlantic transmissions. Marconi was a great integrator of ideas, with particular achievements being the design of transmitting and receiving antennas that could be tuned to a particular frequency and the development of a coherer to improve detection of a signal.

3.2.3   Early Radio

Early radio systems were based on modulating an oscillating carrier either using pulsing in the case of Morse code—this modulation scheme is called Amplitude Shift Keying (ASK)—or Amplitude Modulation (AM), in the case of voice transmission. The waveforms and spectra of these modulation schemes are shown in Figure 3-1. The information is contained in the

baseband signal, which is also called the modulating signal. The spectrum of this signal extends to DC or perhaps down to a roll-off at a low frequency. The carrier is a single sinewave and contains no information. The carrier is varied by the baseband signal to produce the modulated signal. In general, there are many cycles of the carrier relative to variations of the baseband signal so that the bandwidth of the modulated signal is relatively small compared to the bandwidth of the carrier.

AM and ASK radios are narrowband communication systems (they use a small portion of the EM spectrum), so to avoid interference with other radios it is necessary to search for an open part of the spectrum to place the carrier frequency. At first, there was little organization and a listener needed to search to find the desired transmission. The technology of the day necessitated this anyway, as the carrier would drift around since it was then not possible to build a stable oscillator. It was not until the Titanic sinking in 1912 that regulation was imposed on the wireless industry. Investigations of the Titanic sinking concluded that most of the lives lost would have been saved if a nearby ship had been monitoring its radio channels and if the frequency of the emergency channel was fixed and not subject to interference by nonemergency radios. However, a second ship, but not close enough, did respond to Titanic's "SOS" signal.[6] A result of the investigations was the Service Regulations of the 1912 London International Radiotelegraph Convention. In the United States this led to the Act to Regulate Radio Communication, passed on August 13, 1912. Similar regulations were adopted by governments around the world. These early regulations were fairly liberal and radio stations were allowed to use radio wavelengths of their own choosing, but restricted to four broad bands: a single band at 1500 kHz for amateurs; 187.5 to 500 kHz, appropriated primarily for government use; below 187.5 kHz for commercial use, mainly for long-distance communication because of the excellent ground wave effect;[7] and 500 kHz to 1500 kHz, also a commercial band. Subsequent years saw more stringent assignment of narrow spectral bands and assignment of channels. The standards and regulatory environment for radio were set—there would be assigned bands for particular purposes. Very quickly strong government and commercial interests struggled for exclusive use of particular bands and thus the EM spectrum developed considerable value. Entities "owned" portions of the spectrum either through a license or through government allocation. It did not matter how efficiently the bands were used, and now, overall, the spectrum utilization is quite low even though spectrum allocation is almost 100% up to the tens of gigahertz range.

---

[6]   SOS is used in an emergency as the Morse code for SOS is easily recognizable.
[7]   The ground wave effect is when propagating EM signals above ground have a portion of their energy in the skin of the earth. The ground wave effect is more pronounced at lower frequencies where the propagating EM signals bend to follow the earth's curvature.

RF hardware design, and the competitive positioning of RF companies, is intricately entwined with the regulatory environment.

While most of the spectrum is allocated, there are several open bands where licenses are not required. The **instrumentation, scientific, and medical (ISM)** bands at 2.4 and 5.8 GHz are examples. Since these bands are loosely regulated, radios must cope with potentially high levels of interference. Many modulation and spectrum utilization schemes are being used with the aim of maximizing the use of spectrum and coping with interference. This is driving the evolution of RF hardware, as well as different bands in various countries, the need to support multiple standards, and the need to efficiently support various data rates ranging from approximately 10 kbps (for voice) to 2 Mbps for video (albeit on a small screen). Service providers[8] project increasing demand for ever higher data rates.

3.2.4 Radio Frequency Design

A radio device, be it a radar, sensor, or communicator, is comprised of several key reasonably well-defined units. By frequency we have baseband, IF, and RF partitions. In a typical device, the information—either transmitted or received, bits or analog waveforms—is fully contained in the baseband unit. In the case of digital radios, the digital information originating in the baseband DSP is converted to an analog waveform typically using a DAC. When the basic information is analog, say a voice signal in analog broadcast radios, the information is already a baseband analog waveform. This analog baseband signal can have frequency components that range from DC to hundreds of kilohertz or megahertz. However, the baseband signal can range from DC to gigahertz in the case of some radars and point-to-point links that operate at tens of gigahertz. The RF hardware interfaces the external EM environment with the rest of the phone. The information that is represented at baseband is translated to a higher-frequency signal which can more easily propagate and for which antennas can be more easily built with manageable sizes. Thus the information content is generally contained in a narrow band of frequencies centered at the RF. The information content generally occupies a relatively small slice of the EM spectrum. We use the term generally, as it is not strictly necessary that communication be confined to a narrow band: that is, narrow in percentage terms relative to the RF. The trade-offs in choice of RF are that lower-frequency EM signals require large antennas, typically one-quarter to one-half wavelength long, but propagate over long distances and tend to follow the curvature of the earth. AM broadcast radio stations operate around 1 MHz (where the wavelength, λ, is 300 m) using transmit antennas

---

[8]   Service provider is the term given to operators of a cellular phone system, for example.

that are 100 m high or more, but good reception is possible at hundreds of kilometers from the transmitter. At higher frequencies, antennas can be smaller, a much larger amount of information can be transmitted with a fixed fractional bandwidth, and there is less congestion. An antenna at 2 GHz (where λ = 15 cm) is around 7 cm long (and smaller when folded or coiled), which is a very convenient size for a hand-held communicator. The concept of an IF is related to the almost universal architecture of transmitters in the 20th century when baseband signals were first translated, or heterodyned, to a band around an IF before a second translation to a higher RF frequency. Initially the IF was just above the audible range and was known as the supersonic frequency (hence this conversion scheme is known as superheterodyne conversion or just superhet). The same progression applies in reverse in a receiver where information carried at RF is first translated to IF before finally being converted to baseband. This architecture resulted in near optimum noise performance and relatively simple hardware, particularly at RF, where components are much more expensive than at lower frequencies.

The above discussion is a broad description of how radios work. There are many qualifications, as there are many evolving architectures and significant rethinking of the way radios can operate. Architectures and basic properties of radios are trade-offs of the capabilities of technologies, signal processing capability, cost, market dynamics, and politics.

3.3   Wireless Standards

The worldwide demand for wireless systems has led to a plethora of standards and different types of systems. Contrary to what would be expected, evolving standards are increasing the diversity of RF systems rather than leading to the adoption of a few communication formats. In addition to the enormous pressure resulting from the expansion of communications, wireless networks, RF-based sensor systems, and radar are pushing the limits of RF and microwave design. This is partly a result of the maturity of high-volume RF systems and the opportunities provided by **Very-Large Scale Integration VLSI**, mixed-signal circuits, algorithmic advances, and advances in digital signal processing. New mobile phones are becoming universally ubiquitous communicators supporting multiple wireless communications standards, including WiFi, Bluetooth, and WiMax, as well as multiple cellular bands and standards.

One trend is that systems are designed reflecting the capabilities of RF hardware. Similarly an understanding of systems is required to understand the specifications for RF hardware. This is particularly important, as the actual performance required is often not directly related to the specifications developed by system designers. For example, one of the most important characteristics of digital radio systems is the **Bit Error Rate** (**BER**). The

BER is a quantity that cannot be determined until the components of a system are integrated. Thus in the design of subsystems, indirect measures such as **intermodulation distortion IMD** (referring to the generation of spurious signals when discrete tones[9] are applied to the subsystem) are specified. The relationship between (**IMD**) and BER is weak. Clearly higher IMD tends to indicate a higher BER for the same technology, but the relative performance of different technologies cannot be evaluated this way. One way of addressing levels of IMD that are too high is to use additional filtering. This however can result in additional in-band distortion and increased BER. Thus the essential system design problem is developing sufficient and tractable criteria that enable subsystems to be locally designed.

3.4   Conventional Wireless Communications

Up until the mid-1970s most wireless communications were based on centralized high-power transmitters, often operating in a wide-area broadcast mode, and reception (e.g., by a television or radio unit) was expected until the signal level fell below the noise threshold or several multiples of power above it.[10] Many systems still operate in this mode. These systems are particularly sensitive to interference, therefore systems transmitting at the same frequency are geographically separated so that the signals fall below the background noise threshold before there is a chance to interfere with a neighboring system operating at the same frequency. This situation is illustrated in Figure 3-2. Here there are a number of base stations (BSs), each operating at a frequency (or set of frequencies) designated by numerals referring to the frequency of operation which we will correspondingly designate as f_1, f_2, etc. The coverage by two base stations both operating at the frequency f_1 are highlighted, indicating the geographical region over which the signals are above the minimum detectable signal threshold. The regions of acceptable signal reception are small, as the signals must be several factors above the noise level for reasonable reception. The frequency reuse factor of these types of systems is low, as there is a large geographical area where there is no reception at a particular frequency. The coverage area will not be circular or constant because terrain is not flat, signals are blocked by and reflected from buildings, and background noise levels vary during the day and from season to season as vegetation coverage changes. Allowances must be made in the allocation of broadcast areas to account for the changing coverage

---

[9] A tone is a **continuous wave** (**CW**) signal or sinewave.

[10] This discussion refers to conventional analog radio and television broadcasts using amplitude and frequency modulation. The situation is different with digital radio and television broadcasts, as error correction codes enable stations to be placed much closer together.
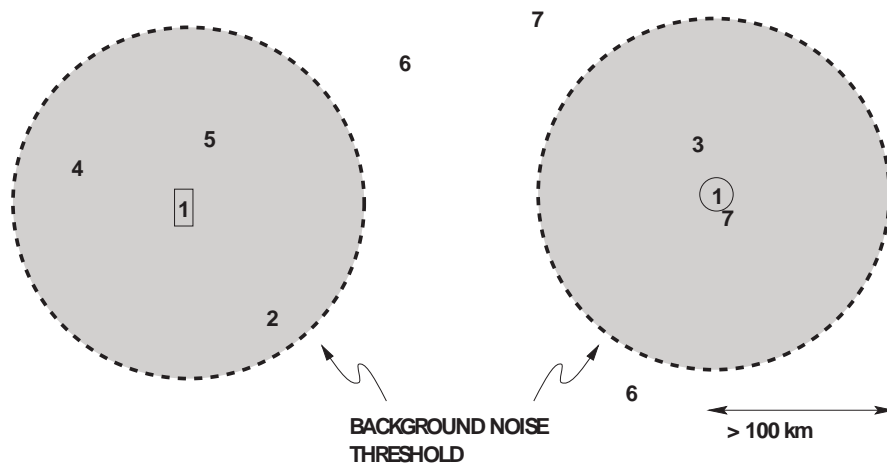
Figure 3-2 Interference in a conventional radio system. The two transmitters, 1, are at the centers of the coverage circles.

level. At the same time, it is necessary, in conventional radio, for the coverage area to be large so that reception, particularly for mobile devices, is continuous over a region corresponding to metropolitan-size areas.

Conventional consumer-oriented wireless systems have been available for some time with wireless telephone service.[11] In large metropolitan areas the wireless telephone service could only simultaneously support 10 or so two-way conversations at a time in each of the available bands. Mobile telephone service expanded to operate at the 40 MHz, 150 MHz, and 450 MHz bands, thus tripling the number of available channels. As a consequence, in a region the size of New York City only 30 conversations could be supported. The demand was much greater than the capacity of the wireless service and was only controlled by the frustration people had in using the system and the long waiting periods in obtaining a subscription to the service. The demand was believed to be many orders of magnitude greater than could possibly be satisfied using the mobile telephone system architecture. It was clear that only a radically different system could satisfy demand. The solutions arrived at are responsible for the current "wireless" proliferation.

The mobile service described above is called **0G** for zero-generation radio. The broadcast concept of widely separating sites operating at the same frequency is behind conventional broadcast radio and television.

---

[11] Common forms of nonbroadcast wireless communications are paging services operating in the 40 MHz, 150 MHz, 450 MHz, and 900 MHz bands, and cordless telephones in the 800–900 MHz and 1800–1900 MHz bands.
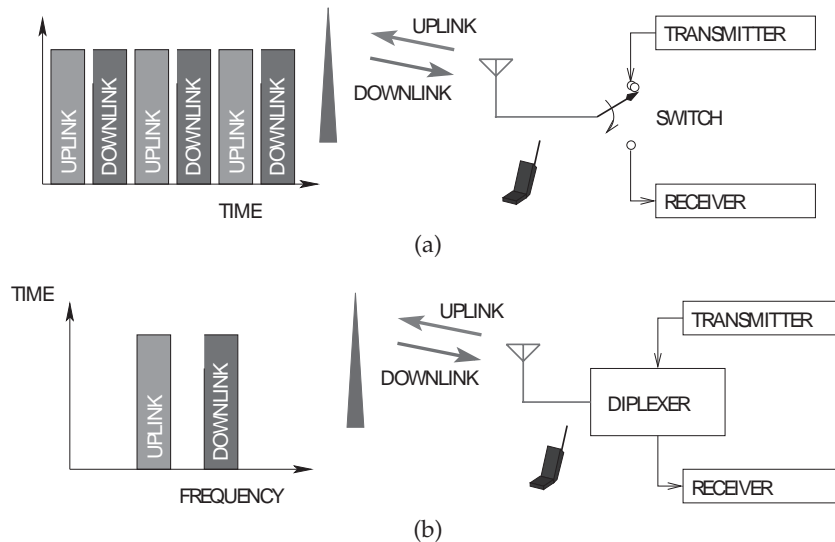
Figure 3-3 Duplex schemes: (a) time-division duplex (TDD); and (b) frequency-division duplex (FDD).

3.5 Simplex and Duplex Operation

Simplex operation refers to one-way communication in which there is only one transmitter and at least one, and perhaps more, receivers. So broadcast radio and television use simplex communication. Two-way communication uses a duplex scheme. Duplex operation requires management of the way in which information passes first from one radio to a second radio and then transmitting information in reverse. The management scheme is called duplex operation. The two predominant duplex schemes are **Frequency-Division Duplex** (**FDD**) and **Time-Division Duplex** (**TDD**), as illustrated in Figure 3-3. In these diagrams the mobile unit is shown communicating with a base station, in which case transmission from the mobile unit to the base station is called the **uplink** (**UL**) or **reverse link** (**RL**), and communication from the base station and received by the mobile unit is called the **downlink**, (**DL**), **forward link** (**FL**), or **forward path**. In FDD it is necessary to use a filter to separate the uplink and downlink signals, as shown in Figure 3-4, as the two links are in use simultaneously.

In TDD (Figure 3-3(a)), the uplink and downlink are separated in time and a switch connects the antenna to the transmitter, to send the uplink, and then to the receiver to receive the downlink. The sequence is then repeated. The two units in a communication link do not transmit simultaneously. This operation is called **half-duplex**. Generally the uplink and downlink are at different frequencies, although in "walkie-talkie" or **Family Radio Service** (**FRS**) systems the same narrow frequency band is used for uplink and downlink. Early radios operating using simplex or duplex operation
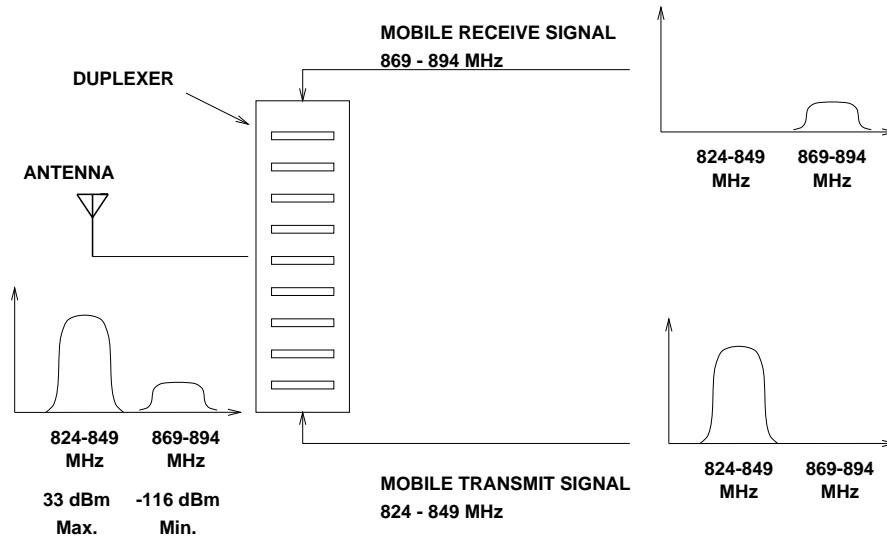
Figure 3-4   A duplexer which separates receive and transmit signals.

are shown in Figure 3-5. In FDD the uplink and downlink are at different frequencies and the transmitter and receiver are connected permanently to the antenna. This communication scheme is called **full-duplex** operation,

| (a) | (b) | (c) |
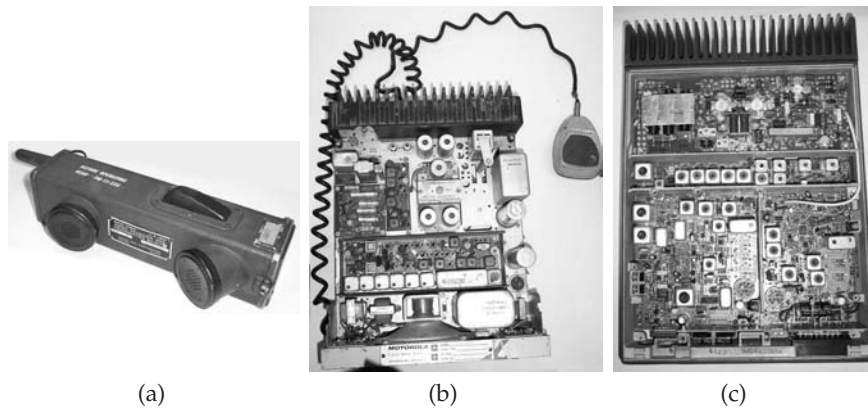
Figure 3-5   Early radios: (a) walkie-talkie with push-to-talk (PTT) switch on top; (b) Motorola business dispatcher two-way radio operating at 33.220 MHz designed in the 1960s as a dash mount unit; (c) the Mitrek two-way radio designed by Motorola in 1977. The radio has two PC boards and was crystal controlled with a channel scanning control head. The radio was trunk mounted, with the control head, microphone, and speaker mounted under the dash board. Each unit is approximately 20 cm long.

although sometimes just the term duplex is used. FDD is implemented by the duplex filter, which is a special filter with three ports that looks like a lowpass filter for the uplink when the transmitter is connected to the antenna (the uplink is generally at a lower frequency than the downlink in first- and second-generation radio systems) and a highpass filter for the downlink when the receiver is connected to the antenna.

3.6 Cellular Communications

Cellular communications, as the name implies, is based on the concept of cells in which a terminal unit communicates with a base station at the center of a cell. Each cell can be relatively small and a terminal unit travels smoothly from cell to cell with calls transferring using what is called a hand-off process. For communication in closely spaced cells to work, interference from other radios must be managed. This is facilitated using the processing gain available with different modulation and coding schemes and the use of antenna beam-forming technology. There are a large number of cellular systems that are being used now or have been in the past. Cellular systems are examined in Section 3.9 on Page 132, but before that the underlying technology and concepts must be explored.

3.6.1 Cellular Concept

The cellular concept was outlined in a 1947 Bell Laboratories technical memorandum [21]. It described a system of frequency reuse with small geographical cells and this remains the key concept of cellular radio. This was elaborated on in two articles published in 1957 and 1960 [22, 23]. The first widespread cellular radio system was called the **Advanced Mobile Phone System** (**AMPS**) and was fully described by Bell Laboratories in a submission to the U.S. **Federal Communications Commission** (**FCC**) and in a patent filed on December 21, 1970 [24]. Bell Labs petitioned the FCC in 1958 for a frequency band around 800 MHz for a cellular system. The FCC, believing that it was better to allocate spectrum for the public good, including radio, television, and emergency services, was reluctant to act on the petition. In 1968 pressure on the FCC became too great and an agreement was reached in principle to make frequencies available. Thus began the research and development of cellular systems in the United States. In 1961 Ericsson reorganized to address mobile radio, including cellular radio systems. Nokia, formed in 1865, did not begin developing cellular systems until the 1970s. NTT was working away as well and began developing cellular radio systems in 1967 [25]. Meanwhile, in January 1969, the Bell System launched an experimental cellular radio system employing frequency reuse for the first time to achieve optimum use of a limited number of RF channels. The first commercial cellular system was launched by the Bahrain Telephone Company in May 1978 using Matsushita

equipment. This was followed by the launch of AMPS by Illinois Bell and AT&T in the United States in July 1978.

In 1979 the **World Administrative Radio Conference** (**WARC**), allocated the 862–960 MHz band for mobile radio, leading to the FCC releasing, in 1981, 40 MHz in the 800–900 MHz band for "cellular land-mobile phone service." The service, as defined in the original documents, is (and this is still the best definition of cellular radio)

> A high capacity land mobile system in which assigned spectrum is divided into discrete channels which are assigned in groups to geographic cells covering a cellular geographic area. The discrete channels are capable of being reused within the service area.

The key attributes here are

- High capacity. Prior to the availability of the cellular system other radio systems were used. Users were not always able to gain access to the radio network and frequently access required multiple attempts.

- The concept of cells. The idea is to divide a large geographic area into cells, shown as the hexagons in Figure 3-6. The actual shape of the cells is influenced by obstructions such as hills and buildings, but the hexagonal shape is used to convey the concept of cells. The cells are arranged in clusters and the total number of channels available is divided among the cells in a cluster and the full set is repeated in each cluster. In Figure 3-6, 3-, 7- and 12-cell clusters are shown. As will be explained later, the number of cells in a cluster affects both capacity (the fewer cells the better) and interference (the more cells per cluster the better). The size of a cell can be reduced, thus increasing system capacity, in a process called cell splitting. In cell splitting, additional base stations are introduced and the base stations and mobile units operate at lower power levels to reduce the geographical coverage or cell size.

- Frequency reuse. Frequencies used in one cell are reused in the corresponding cell in another cluster. As the cells are relatively close, it is important to dynamically control the power radiated by each radio, as radios in one cell will produce interference in other clusters.

Cells are often represented as having a hexagonal shape. The shape, however, can take nearly any form. In a flat desert the coverage area of each base station would be circular, so that with a cluster of cells there would be overlapping circles of coverage. (Power levels are adjusted to minimize the overlap of these circles.) Buildings, hills, lakes, etc. affect cell size. In a city, what is called the **urban canyon effect**, or **urban waveguide effect** greatly distorts cells and creates havoc in managing cellular systems. In the urban
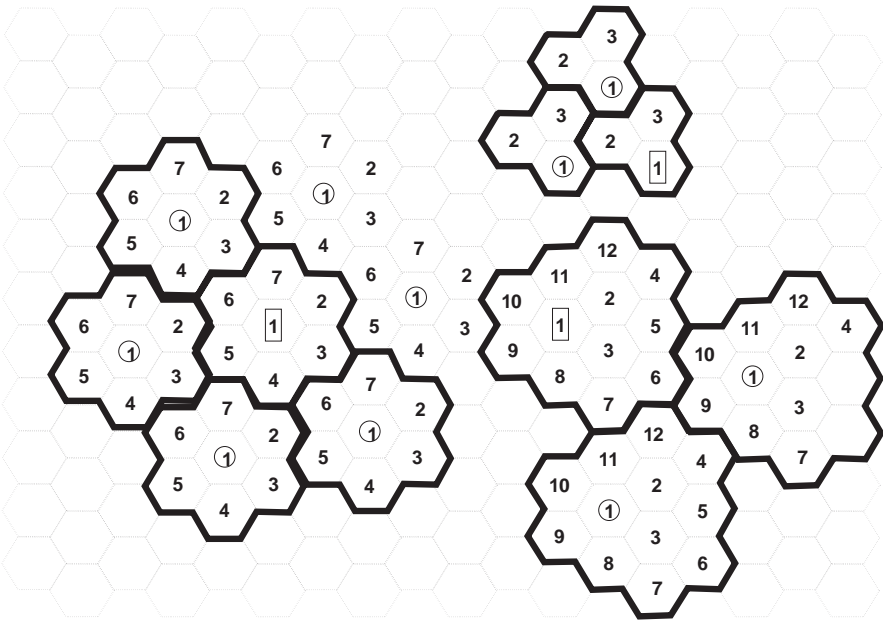
Figure 3-6   Cells arranged in clusters.

canyon effect good coverage extends for large distances down a street, so a cell may look like a four-armed starfish if the cell site is placed near an intersection of two streets.

Achieving maximum frequency reuse is essential in achieving high capacity. In a conventional wireless system, be it broadcast or mobile telephone service, base stations are separated by sufficient distance that the signal levels fall below a noise threshold before the same frequencies are reused, as shown in Figure 3-2. There is clearly poor geographical use of the spectrum here. The geographic areas could be pushed closer to each other at the expense of introducing what is called cochannel interference—a receiver could pick up transmissions from two or more base stations operating at the same frequency. Interference in conventional systems results solely from background noise, as the concept is to eliminate cochannel interference entirely. In a cellular system, there is a radical departure in concept from this. Consider the interference in a cellular system as shown in Figure 3-7. The signals in corresponding cells in different clusters interfere with each other and generally the interference is much larger than that of the background noise. Generally only in rural areas and when mobile units are near the boundaries of cells will the background noise level be significant. Interference can also be controlled by dynamically adjusting the base station and mobile transmit powers to the minimum acceptable level. Operating with interference from neighboring clusters is a key concept in cellular
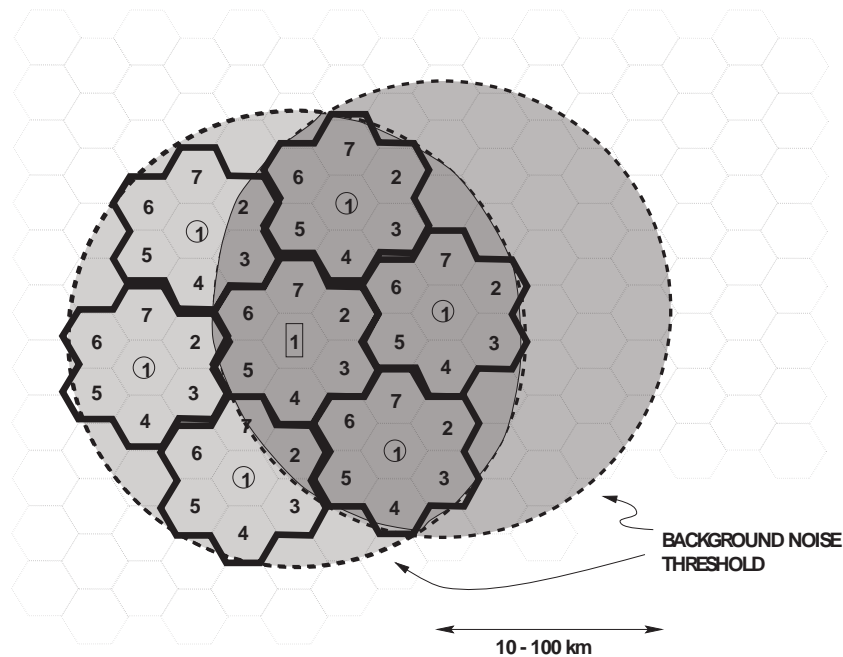
Figure 3-7   Interference in a cellular radio system.

radio. The interference levels that can be tolerated vary among the various cellular systems.

The AMPS system, which uses frequency modulation, has a qualitative minimum SIR of 17 dB (about a factor of 50) that was determined via subjective tests with a criterion that 75% of listeners ranked the voice quality as good or excellent, the top two levels on a five-level scale. The seven-cell clustering shown in Figure 3-7 does not achieve this required minimum SIR.

Traffic engineering tells us that using smaller clusters results in greater spectrum efficiency. With fewer cells per cluster, more frequency channels, or trunks, are available. Since calling patterns are statistical, increasing the number of trunks greatly increases the number of subscribers that can be supported. Thus digital systems that use error correction coding tolerate high levels of interference and can reuse frequency channels more efficiently. This increases the number of trunks available and greatly increases capacity. Indeed, in the CDMA system the tolerance to interference is so high that the concept of clustering is not required and every frequency channel is available in each cell. The increased user capacity of digital radio has compelled all new radio systems to use digital modulation and adopt CDMA-like schemes. Along with greater capacity, digital systems allow for a whole range of advanced functionality, including paging, texting and

messaging, computer communications, Web browsing, and picture transfer.

3.6.2 Personal Communication Services

Personal Communication Services (**PCS**), implemented in the early 1990s, was a development in the thinking of wireless communications. PCS is an evolution of digital cellular radio with the key conceptual difference being that communication is from person to person, whereas in cellular radio communication, as originally conceived, it was from terminal-to-terminal. The idea is that when a call is placed, an individual is being contacted rather than a piece of hardware. One way this is achieved is by using a credit card-size (or smaller) smart card, a **Subscriber Identification Module** (**SIM**) card, to identify the user. The user can insert his or her smart card into any (appropriate) handset and be contacted. The particular hardware or handset a user is using does not matter. So in PCS, each person has a unique phone number. The term PCS is not commonly used now, as the concept has now been incorporated in all evolved cellular systems.

3.6.3 Processing Gain

Processing gain is an essential reason why the cellular concept works, as the interference coming from radios in corresponding cells in adjacent clusters can be significant. In the discussion of FM in Section 1.3.2 on Page 12, the concept of processing gain was introduced. In FM, processing gain is obtained by spreading the information-carrying signal over a broader bandwidth. There are some inefficiencies: in FM the spectrum of the spread signal tapers off and does not fill the channel with an even distribution of power (and so the edges are more susceptible to interference). If the power is uniformly distributed over the channel of bandwidth B_m, then the processing gain is

$$G_P = B_m/B_b \, . \tag{3.1}$$

where B_b is the bandwidth of the baseband signal. Demodulation collapses the correlated information in the bandwidth B_m to the smaller bandwidth B_b, but the uncorrelated noise and interference are suppressed. That is, if interference is uncorrelated to the signal, then the interference can be treated as noise. So the recovered SIR of the signal is increased by G_P; that is, with SIR_b being the SIR of the recovered baseband signal and SIR_m being the SIR of the received modulated signal, the two are ideally related by the processing gain:

$$\text{SIR}_b = G_P \text{SIR}_m \, . \tag{3.2}$$

The actual processing gain of FM is less than that indicated by Equation (3.1), as the energy is not uniformly distributed across the channel. It is not possible to precisely and analytically calculate the processing gain of an FM system (or of any modulation scheme).
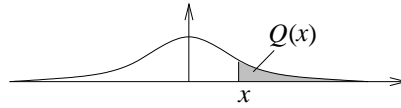
Figure 3-8 Guassian distribution function showing the Q function as the area of the shaded region.

The essential concept of processing gain is that by spreading a communication signal over a greater bandwidth than that of the information bearing signal, correlation of the transmitted signal can be used to discriminate between the signal and noise. With a digitally-modulated signal processing gain is, roughly, the ratio of the bit rate of what is transmitted to the bit rate of the information bearing signal.

3.6.4 Symbol Error Rate and Bit Error Rate

In this section, SIR is related to the bit error rate, BER. The discussion begins with the SIR of the incoming RF signal. Through sampling of the RF signal at appropriate times discrete symbols are obtained. The digital form of SIR relates the energy of a symbol, E_s, to the noise and interference energy corresponding to the symbol (generally the noise and interference in the duration of the symbol). The height of the double-sided noise spectral density is conventionally taken as $N_o/2$, so the noise power corresponding to a symbol is N_o. Thus [30–32]

$$\frac{E_s}{N_o} = \text{SIR} \; . \tag{3.3}$$

Error in a digitally-modulated communication systems is first manifested as a symbol error which occurs when a symbol selected in a receiver is not the symbol transmitted. The probability of a symbol error is a function of E_s/N_o. It is not always possible to develop a closed form expression relating the two however. For a BPSK system the probability of a symbol error, the symbol error rate, is [30, page 187]

$$\Pr[\text{symbol error}] = \Pr_s^{\text{BPSK}} = Q\left(\sqrt{\frac{2E_s}{N_o}}\right) = Q\left(\sqrt{2 \cdot \text{SIR}}\right), \tag{3.4}$$

where $Q(x)$ is known as the Q function [30] and is the integral of the tail of the Gaussian density function (see Figure 3-8). It can be expressed in terms of the error function $\text{erf}(x)$ [30, page 63]:

$$Q(x) = \frac{1}{2}\left[1 - \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right]. \tag{3.5}$$

For M-PSK [30, page 191]

$$\mathrm{Pr}_s^{M-\mathrm{PSK}} \approx 2Q\left[\sqrt{\frac{2E_s}{N_o}}\sin\left(\frac{\pi}{M}\right)\right] \tag{3.6}$$

and for M-QAM [30, page 190]

$$\mathrm{Pr}_s^{M-\mathrm{QAM}} = 4\left(1 - \frac{1}{\sqrt{M}}\right)Q\left(\sqrt{\frac{3E_s/N_o}{M-1}}\right)$$
$$- 4\left(1 - \frac{1}{\sqrt{M}}\right)Q^2\left(\sqrt{\frac{3E_s/N_o}{M-1}}\right) \tag{3.7}$$

where M is the number of modulation levels (e.g. for QPSK $M = 4$, and for 64-QAM $M = 64$).

The per-bit SNR is obtained after noting that each symbol can represent several bits. With a uniform constellation and the same number of bits per symbol, b, the signal energy received per bit is

$$E_b = E_s/b, \qquad b = \log_2 M, \tag{3.8}$$

and

$$M = 2^b \tag{3.9}$$

For high SIR and using **gray mapping**, the probability of a bit error, the bit error rate, is

$$\mathrm{Pr}[\text{bit error}] = \mathrm{Pr}_b = \frac{1}{b}\mathrm{Pr}_s. \tag{3.10}$$

Gray mapping is coding that ensures that nearest-neighbor symbols differ by at most one-bit. The high SIR requirement results in a symbol error being the erroneous selection of a nearest neighbor symbol by the receiver. The final results are the bit-error probablity for BPSK, M-QAM and M-PSK [30, page 193]:

$$\mathrm{Pr}_b^{\mathrm{BPSK}} = Q\left(\sqrt{2E_s/N_o}\right) = Q\left(\sqrt{2\cdot\mathrm{SIR}}\right) \tag{3.11}$$

$$\mathrm{Pr}_b^{M-\mathrm{PSK}} = \frac{2}{b}Q\left[\sqrt{2E_s/N_o}\sin\left(\frac{\pi}{M}\right)\right] = \frac{2}{b}Q\left[\sqrt{2\cdot\mathrm{SIR}}\sin\left(\frac{\pi}{M}\right)\right] \tag{3.12}$$

$$\mathrm{Pr}_b^{M-\mathrm{QAM}} = \frac{4}{b}\left(1 - 2^{\frac{-b}{2}}\right)Q\left(\sqrt{\frac{3E_s/N_o}{2^b-1}}\right) = \frac{4}{b}\left(1 - 2^{\frac{-b}{2}}\right)Q\left(\sqrt{\frac{3\cdot\mathrm{SIR}}{2^b-1}}\right) \tag{3.13}$$

Thus for the same SIR the probablity of bit errors drops with higher-order modulation, i.e. larger M. This is the major driving factor behind the move to higher-order modulation.

Because of coding, the actual transmitted code rate, R_c, in a digital system is higher than the information bit rate, R_i, so that the processing gain

$$G_P = \frac{R_c}{R_i}. \tag{3.14}$$

(This corresponds to the bandwidth-based definition of processing gain in Equation (3.1) if the signal is digitally-modulated with maximum efficiency to fully and uniformly occupy the available bandwidth.) Now, however, more bits must be transmitted for the same transmit power. The growth in the bit error rate, see Equations (3.11)–(3.13), is slower than the growth rate of G_P. With optimal coding, the transmitted code rate is proportional to the bandwidth of the transmitted signal so the processing gain is

$$G_P = \frac{\text{Transmitted bandwidth}}{\text{Information bit rate}}. \tag{3.15}$$

If $E_{b,\text{eff}}$ is the energy of a bit of data (after despreading and error correction), then

$$\text{Effective SIR} = \frac{E_{b,\text{eff}}}{N_o} = G_P \frac{E_b}{N_o} \approx \frac{G_P}{b}\frac{E_s}{N_o} = \frac{G_P}{b}\text{SIR}_{\text{RF}} \tag{3.16}$$

where E_b is the energy of a received bit, E_s is the energy of a received symbol, and SIR_{RF} is the SIR of the RF signal. With spreading, as used in CDMA, the number of transmitted bits is much greater than the number of data bits, typically 2 to 4 orders of magnitude greater. Forward error correcting bits further increases the number of bits transmitted. The spread and error coded signal has considerable correlation while noise is uncorrelated. So on despreading and error correction the effective SIR is greatly increased, an affect captured by the processing gain.

EXAMPLE 3. 1   Processing Gain

A new communication system is being investigated for sending data to a printer. The system will use GMSK modulation and a channel with 10 MHz bandwidth and the bit rate will be 1 Mbps. The modulation format will result in a spectrum that distributes power almost uniformly over the 10 MHz bandwidth.
  (a) What is the processing gain?

$$\text{Processing gain} = \frac{\text{Bandwidth}}{\text{Bit rate}} = \frac{10\text{ MHz}}{1\text{ Mbps}} = 10 = 10\text{ dB}$$

  (b) If the received RF SIR is 6 dB. What is the effective system SIR (or E_b/N_o) after the digital signal processor? Express your answer in decibels.

For GMSK, $b = 2$ and so

$$\text{Effective SIR} = \frac{E_{b,\text{eff}}}{N_o} = \frac{1}{b}(\text{RF SIR}) \cdot \text{Processing Gain}$$

$$= -3\text{ dB} + 6\text{ dB} + 10\text{ dB} = 13\text{ dB}. \tag{3.17}$$

EXAMPLE 3. 2   Signal-to-Interference Ratio

At the output of a receiver antenna, the level of interfering signals is 1 pW, the level of background noise is 500 fW, and the level of the desired signal is 4 pW.

(a) What is the SIR? Note that SIR includes the effect of the signal, interference, and noise.

(b) If the processing gain is 20 dB and 16-QAM is the modulation scheme used, what is the effective system SIR, that is, what is the signal energy in a bit versus the noise energy in the duration of the bit, E_b/N_o?

Solution:

(a) Interfering signal $P_I = 1$ pW

Noise signal $P_N = 500$ fW

Signal $P_S = 4$ pW

$$\text{SIR} = P_S/(P_I + P_N)$$
$$= 4\,\text{pW}/(1\,\text{pW} + 0.5\,\text{pW}) = 4/1.5 = 2.667 = 4.26\,\text{dB}$$

(b) Processing gain, $G_P = 20$ dB and there are 4 bits ($b = 4$), so

$$\frac{E_{b,\text{eff}}}{N_o} = \frac{1}{4}\text{SIR} \cdot G_P$$
$$= -6\,\text{dB} + 4.26\,\text{dB} + 20\,\text{dB} = 18.26\,\text{dB} .$$

3.6.5 Call Flow and Handoff

With small cells, mobile users can be expected to move frequently between cells and thus call handoff procedures for transferring calls and users from one cell to the next are necessary. The triggering events that enable call handoff are shown in Figure 3-9. The main aspect is monitoring of the signal strength, the **Received Signal Strength Indicator** (**RSSI**), both in the handset for the signal received from a base station and in the base station for the signal received from a handset. If either of these falls below a threshold, computers in the base station initiate a handoff procedure by polling nearby base stations for the RSSI they have for the user. If a suitable RSSI is found handoff proceeds.

Call flow from mobile to mobile and from mobile to **Public Switched Telephone Network** (**PSTN**) must also be orchestrated and a **Mobile Switching Center** (**MSC**) becomes the key component in orchestrating call connection. In digital radio the MSC is always aware of the location of a mobile handset, but in analog radio, increasingly large geographic areas have to be polled to find a mobile unit. The basic call-flow architecture of cellular systems is shown in Figure 3-10.

(a)

(b)

(c)
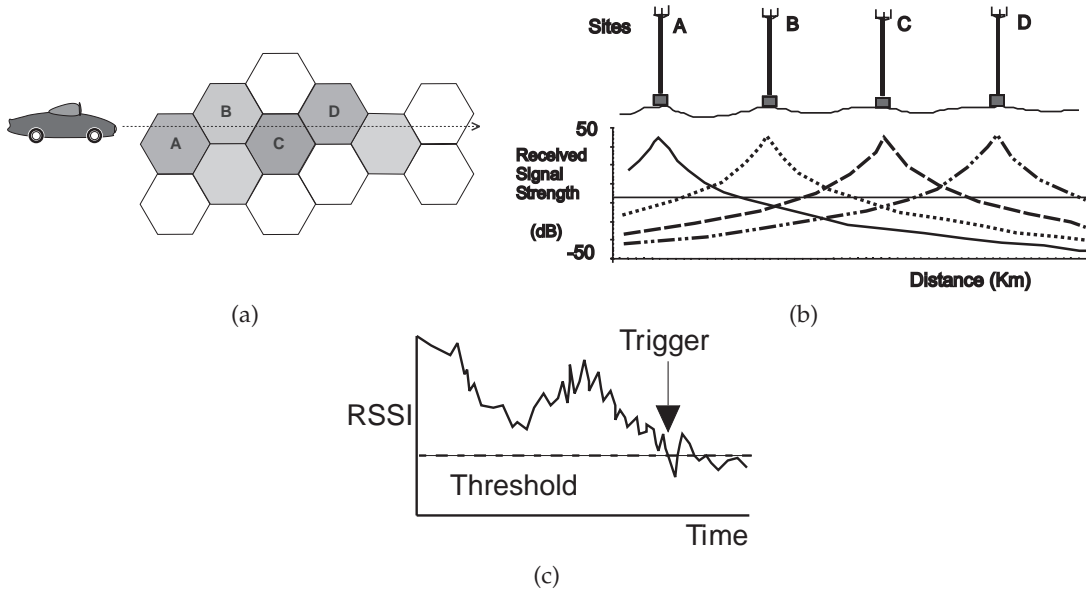
Figure 3-9 Process of handoff: (a) movement of a mobile unit through cells; (b) received signal strength indicator (RSSI) during movement of a mobile unit through cells; and (c) RSSI of a mobile unit showing the handoff triggering event.

(a)   (b)   (c)
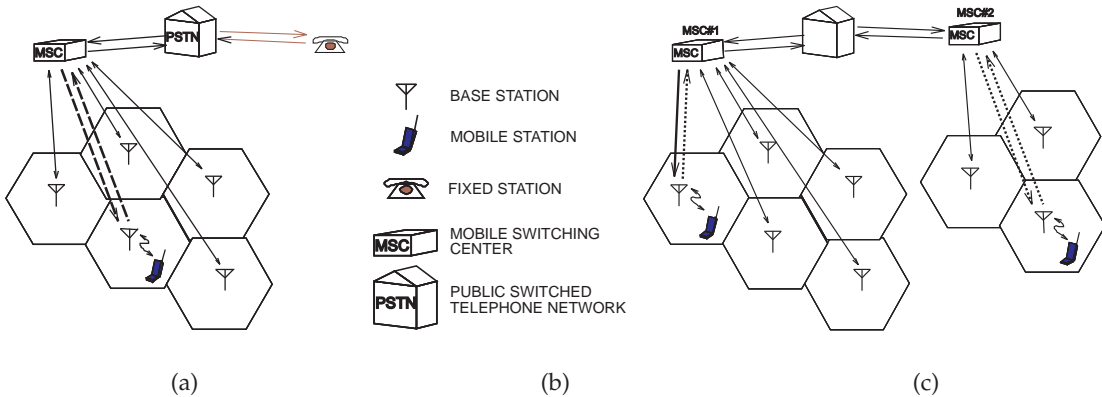
Figure 3-10 Call flow: (a) during initiation of a call; (b) legend; and (c) call flow for mobile to mobile call through a telephone network.

3.6.6 Cochannel Interference

The minimum signal detectable in conventional wireless systems is determined by the SIR at the input. The noise is due to background noise sources, including cosmic noise and thermal noise. In cellular wireless

systems the **Minimum Detectable (or Discernible) Signal** (**MDS**) is also determined by the SIR but now the dominant interference is due to other transmitters in the cell and adjacent cells. The noise that is produced in the signal band from other transmitters operating at the same frequency is called cochannel interference. The level of cochannel interference is dependent on cell placement and the frequency reuse pattern. The degree to which cochannel interference can be controlled has a large effect on system capacity. Control of cochannel interference is largely achieved by controlling the power levels at the base station and at the mobile units.

3.7 Multiple Access Schemes

Many schemes have been developed to enable multiple users to share a frequency band. The simplest scheme requiring the least sophistication in channel management is the **Frequency Division Multiple Access** (**FDMA**) scheme, shown in Figure 3-11(a), where the numerals indicate a particular user. In FDMA the available spectrum is divided by frequency, with each user assigned a narrow frequency band that is kept for the duration of a call. This can be conveniently implemented in analog radio. In duplex operation a frequency channel would be assigned for the uplink as well as the downlink. An example of where FDMA is used is AMPS, where 30 kHz is assigned to each channel.[12] Clearly this is wasteful of spectrum, as not all of the band is used continuously. There are really only three ways to increase the spectrum efficiency of this system. One is to make the channels closer, but this is difficult to do. Another is to use digital techniques, and in effect fill one physical channel with a bitstream. Different users use different segments of the bitstream. The third approach is called spread spectrum, which spreads the signal over a relatively large bandwidth using a unique spreading code whose rate is much faster than the bit rate of the information being sent.

In a radio system, a base station simultaneously transmits to multiple users. The process of combining user communications is called **multiplexing**, often referred to as **MUX** or MUXing. The reverse operation of separating a signal with multiple users into their component users is called **demultiplexing**, and is referred to as **DEMUX** or DEMUXing. The hardware that implements multiplexing is called a multiplexer and demultiplexing is implemented by a demultiplexer.

In the first digital access technique, **Time Division Multiple Access** (**TDMA**), shown in Figure 3-11(b), the bitstream is divided among a few users using the same physical channel. In the **GSM**[13] mobile phone system,

---

[12] In the subsequent **Narrowband Advanced Mobile Phone Service** system (**NAMPS**), the channels were 10 kHz wide. The **Total Access Communications System** (**TACS**) was a European analog cellular telephone standard derived from the AMPS standard.

[13] GSM stands for Global System for Mobile Communications and was formerly known as
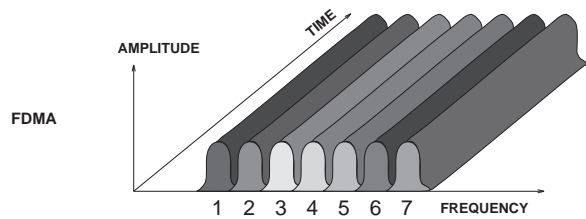
(a) Frequency Division Multiple Access
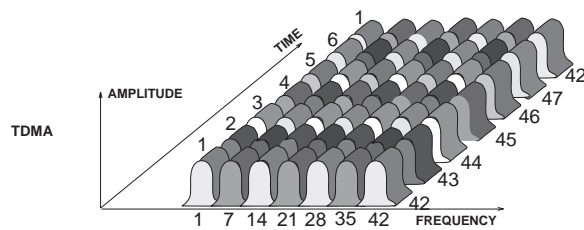
(b) Time Division Multiple Access
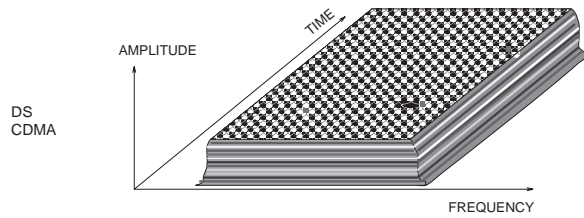
(c) Direct Sequence Code Division Multiple Access
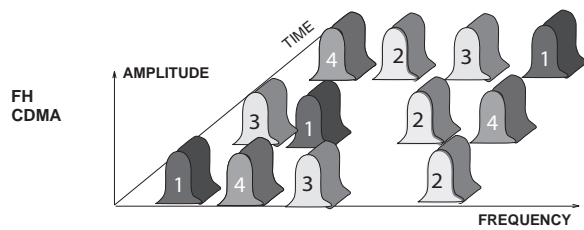
(d) Frequency Hopping Code Division Multiple Access
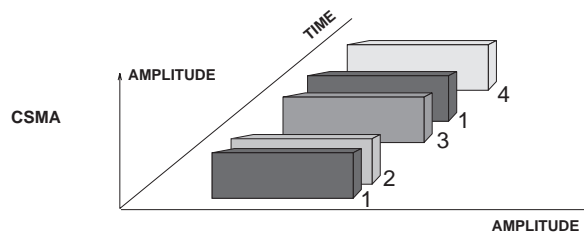
(e) Carrier Sense Multiple Access

Figure 3-11   Multiple access schemes. The numerals indicate the user number.

a physical channel is divided into eight time slots and a user is allocated to each so that eight users can be supported in each physical channel. In other TDMA schemes, a different number of time slots are used. Thus the logical channels are divided in both frequency and time. In a TDMA system, the base station transmits a continuous stream of data containing frames of time slots for multiple users. The mobile unit listens to this continuous stream, extracting and processing only the time slots assigned to it. On the reverse transmission, the mobile unit transmits to the base station in bursts only in its assigned time slots. This is yet another complication to the RF design of TDMA systems. As well as dealing with a nonconstant RF envelope, and hence the requirement for linear amplifiers, the RF circuitry operates in a burst mode with additional constraints on settling time and frequency spreading.[14] One of the advantages of TDMA is that multiple slots can be assigned to the same user to support high data rates and a time slot can be skipped if it is not used by a user (e.g., when the mobile user is not talking) and can be used for other revenue-generating purposes such as text messaging.

Another technique that can be used in digital radio is **Spread Spectrum** (**SS**) the invention of which is attributed to Hedwig Kiesler Markey[15]. Most of the development of spread spectrum was secret up until the mid-1970s. There are two types of spread-spectrum techniques: direct-sequence, and frequency-hoping. It is both an access technique and way to secure communications. **Direct Sequence Code Division Multiple Access** (**DS-CDMA**)[16] is shown in Figure 3-11(c). DS-CDMA mixes the baseband signal with a broadband spreading coding signal to produce a broadband signal which is then used to modulate an RF signal. This process is illustrated in Figure 3-12. The rate of the spreading code is referred to as the **chip rate** rather than the bit rate, which is reserved for the rate of the information-bearing code. A unique code is used to "scramble" the

Groupe Spécial Mobile. The system was deployed worldwide in 1991 and is now used by more than 2 billion people in more than 210 countries. An interesting footnote is that the GSM group began discussions leading to the system in 1982. By the mid-1980s, there were many different versions of the GSM system implemented in many European countries. The European Union (EU) intervened and all 15 countries (the membership of the EU at the time) decided to choose a single standard with eight tested in a competition held in Paris. The system developed by Torleiv Maseng and Odd Trandem at the Norwegian University of Science and Technology (without industrial backing) was chosen. The key feature of this system was its ability to handle multipath reflections.

[14] These RF bursts can sometimes be heard when a GSM phone is near other electronics, such as a wired telephone.

[15] The 1940 invention is described in "Secret Communications System," Secret U.S. patent 2,292,387 submitted by H. Markey and G. Antheil [33]. The patent described a frequency-hopping scheme to render radio-guided torpedoes immune from jamming using a piano roll to hop between 88 carrier frequencies. H. Markey was the actress Hedy Lamarr (born Hedwig Eva Maria Kiesler).

[16] Sometimes DS-CDMA is referred to as just CDMA.
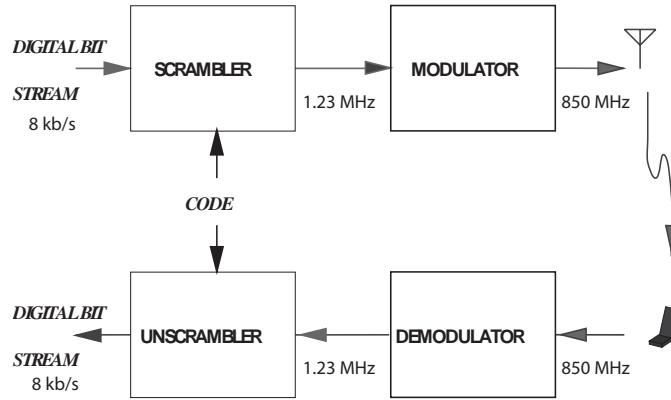
Figure 3-12   Block schematic of the operation of a CDMAOne spread spectrum radio.
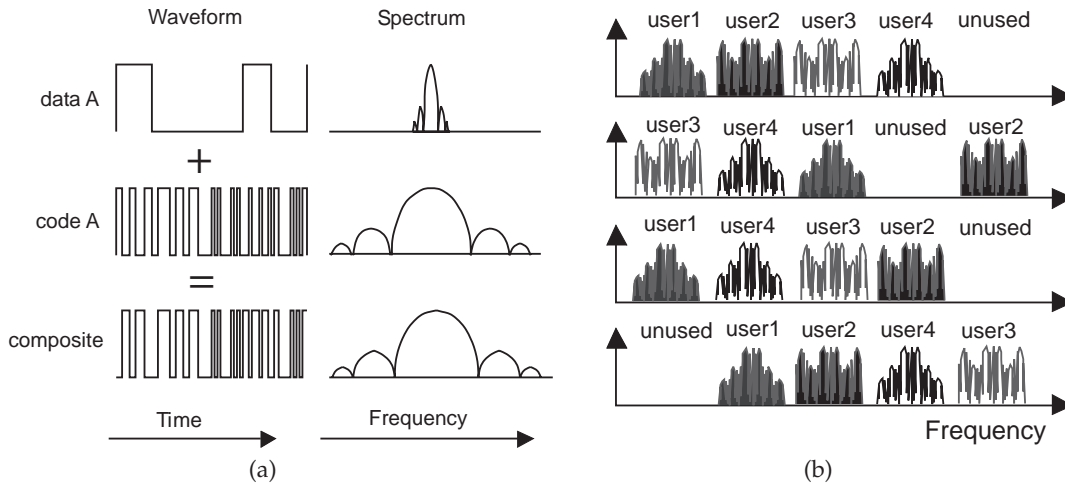
Figure 3-13   Basis of spread spectrum communication: (a) direct sequence; and (b) frequency hopping.

original baseband signal, and the original signal can only be recovered using that particular code. The system can support many users, with each user assigned a unique code, hence the term code division multiple access (CDMA). Another interpretation of the DS-CDMA process is shown in Figure 3-13(a). One of the unique characteristics is that different users are using the same frequency band at the same time. CDMA is used in the **CDMAOne** (described in the IS-95 standard and by which it is often known), **CDMA2000**, and **Wideband CDMA** (**WCDMA**) systems.

Another form of CDMA access is the **Frequency-Hopping CDMA**

scheme (**FH-CDMA**) shown in Figure 3-11(d). This scheme uses **Frequency-Hoping Spread Spectrum** (**FHSS**) to separate users. For one particular user, a code is used to determine the hopping sequence and hops can occur faster than the data rate. The signal can only be reassembled at the receiver if the hopping code is known. FH-CDMA is useful in unregulated environments, such as an ad hoc radio network, where it is not possible to coordinate multiple users. In this case, two users can transmit information in the same frequency band at the same time but the error tolerance of digital communication can still enable the original signals to be extracted. So frequency hopping is preferred in hostile environments, where interference with other radios cannot be coordinated. FH-CDMA is commonly used in radios for emergency workers and by the military.

The final multiple access scheme to be described here is **carrier-sense multiple access** (**CSMA**), which is used in the WiFi system (IEEE 802.11). The access scheme is illustrated in Figure 3-11(e). This scheme is also used in hostile environment where various radios cannot be well coordinated. In CSMA, users transmit at different times, but without coordination. A terminal unit listens to the channel and transmits a data packet when it is free. Collisions are unavoidable and data are lost, requiring retransmission of data, but the terminals use a random delay before retransmitting data. In the case of WiFi there are several channels and if collisions are excessive another channel can be selected either as the preferred startup channel or in evolved systems under central unit control.

Orthogonal Frequency Division Multiple Access (**OFDMA**) is used in **Worldwide Interoperability for Microwave Access** (**WiMax**) and overcomes many of the performance limitations encountered with CSMA as assignment of effective channels is used. As **Orthogonal Frequency Division Multiplexing** (**OFDM**) must first be defined to discuss OFDMA, the OFDMA access scheme will be considered in Section 3.11.2 on Page 145.

3.8 Spectrum Efficiency

The concept of spectral efficiency is important in contrasting different radio and modulation systems. Spectral efficiency has its origins in **Shannon's theorem**, which expresses the information-carrying capacity of a channel as [30–32]

$$\hat{C} = B_c \ln(1 + \text{SNR}) , \tag{3.18}$$

where $\hat{C}$ is the capacity in units of bits/second (b.s^{-1}); B_c is the channel bandwidth in hertz; and SNR is the signal-to-noise ratio. N is assumed to be Gaussian noise, so interference that can be approximated as Gaussian can be incorporated by adding the noise and interference powers, and then it is more appropriate to use the SIR. Thus Shannon's capacity theorem, Equation (3.18), becomes

$$\hat{C} = B_c \ln[1 + S/(N + I)] = B_c \ln(1 + \text{SIR}). \tag{3.19}$$

Shannon's theorem is widely accepted as the upper limit on the information-carrying capacity of a channel. So the stronger the signal, or the lower the interfering signal, the greater a channel's information-carrying capacity. Indeed, if there is no noise and no interference, the information-carrying capacity is infinite. Shannon's capacity formula, Equation (3.19), tells us that increasing the interference level (lower SIR) has a weakened effect on the decrease in capacity than may initially be expected; that is, doubling the interference level does not halve $\hat{C}$. This is the conceptual insight that supports the use of closely packed cells and frequency reuse, as the resulting increase in interference, and its moderated effect on capacity, is offset by having more cells and supporting more users.

Shannon's carrying capacity limit has not been reached, but today's radio systems are very close. Current systems operate with SNRs approximately 8 dB away from the limit [30, page 193].[17] Different modulation and radio schemes come closer to the limit, and two quantities will be introduced to describe the performance of different schemes. From the capacity formula, a useful metric for the performance of modulation schemes can be defined. This is the channel efficiency (or channel spectrum efficiency, sometimes called channel spectral efficiency),

$$\eta_c = R_c/B_c \,, \tag{3.20}$$

where R_c (in bits/s) is the bit rate transmitted on the channel, so η_c has the units of bits per second per hertz (b·s^{-1}·Hz^{-1}). The unit b·s^{-1}·Hz^{-1} is dimensionless, as hertz has the units of s^{-1}, but using b·s^{-1}·Hz^{-1} is a reminder of the meaning of a quantity. (Similarly decibels is dimensionless but is an important qualifier.)

In a cellular system, the number of cells in a cluster must also be incorporated to obtain a system metric [34]. The available channels are divided among the cells in a cluster, and a channel in one cell appears as interference to a corresponding cell in another cluster. Thus the SIR is increased and the capacity of the channel drops. System throughput increases, however, because of closely packed cells. So the system throughput is a function of the frequency reuse pattern. The appropriate system-level metric is the radio spectrum efficiency, η_r, which incorporates the number of cells, K, in a cluster:

$$\eta_r = \frac{R_b}{B_c K} = \frac{\eta_c}{K}\frac{R_b}{R_c}. \tag{3.21}$$

Here R_b is the bit rate of useful information and discounts the **channel bit rate**, R_c, (R_c is higher than R_b because of coding). Coding is used to

---

[17] In **Multiple Input, Multiple Output** (**MIMO**) systems, the limit is apparently exceeded as the effective number of channels increases by using multiple transmit and receive antennas. More on this in Section 3.11.3 on Page 146

enable error correction, assist in identifying the start and end of a packet, and also to provide orthogonality of users in some systems that overlap users as with CDMA. The units of η_r are bits per second per hertz per cell (b·s^{-1}·Hz^{-1}cell^{-1}). The decrease in channel capacity resulting from the increased SIR associated with fewer cells in a cluster (i.e., lower K) is more than offset by the increased system throughput.

So there are two definitions of spectral efficiency: the **channel spectrum efficiency**, η_c, which characterizes the efficiency of a modulation scheme; and the **radio spectrum efficiency**, η_r, which incorporates the added interference that comes from frequency reuse. Indeed, the frequency reuse interference dominates noise in a cellular system, and background noise is often ignored in evaluating performance. Commonly both measures of efficiency are referred to as spectral efficiency, and then only the units distinguish which is being referred to. In summary:

- $\hat{C}$ is a theoretical maximum channel bit rate for a given set of conditions.

- R_c is the actual channel bit rate for a given set of conditions.

- R_b is the actual channel bit rate for a given set of conditions minus overhead associated with coding, etc.

One may well ask why efficiency is not expressed as a ratio of actual bit rate to Shannon's limit for a given set of conditions. There are several reasons:

1. The historical use of bits per hertz to characterize a modulation scheme was used long before cellular systems became popular.

2. Shannon's capacity limit is so high that back in the 1950s people would have been talking about extremely low efficiencies if performance was referred to the capacity limit.

3. Only additive white noise is considered, but this does not capture all types of interference, which can be multiplicative or partially correlated with the signal. Shannon's limit is not really a theoretical limit, there is no proof. In the digital communications world, much has been published about how close Shannon's capacity limit can be approached. In a direct LOS system, such as a point-to-point microwave link, the limit is approached within a few percent. In MIMO systems, (see Section 3.11.3 on Page 146) the limit has been exceeded, prompting a redefinition of the limit when multiple transmit and receive antennas are used.

3.9   Cellular Phone Systems

The evolution of cellular communications is described by generations of radio. The major mobile communication systems are outlined in Table 3-2. Few first-generation (1G) systems remain, most services are now second generation, (2G) dominated by GSM, but also with widespread development of CDMA. Third generation (3G) offers a significant increase in capacity and is optimized for broadband data access.

From an RF point of view, the major distinction among the various phone systems is whether they use analog or digital modulation, how the channels are partitioned, and the type of duplex operation. Systems using analog modulation typically use FM (e.g., the AMPS system), but phase modulation is also common. The main attributes of FM and PM analog modulation schemes is that the signal amplitude is constant, only the frequency or phase varies. This means that RF transmitters can use transistor amplifiers operating in high-efficiency saturating modes (such as classes AB, B, C, and higher), as the introduced distortion, being mostly amplitude distortion, does not have as much impact as with other modulation formats. Newer systems, however, use some form of digital modulation, which generally combines frequency, phase, and amplitude variations as well as a modulation scheme that uses spectrum more efficiently by maintaining a near-constant spectral density in the channel. In digital modulation, the output has a number of prescribed discrete states. In some systems, information is carried in amplitude variations so that circuit nonlinearities that result in amplitude distortion (the most common kind) can significantly affect signal integrity. This presents enormous challenges for the design of efficient transistor amplifiers.

3.10   Generations of Radio

3.10.1   First Generation: Analog Radio

The initial cellular radio system was analog, with the dominant system being theAMPS system, the attributes of which are given in Table 3-3. This is a relatively simple system, but appropriate for the low levels of integration of the 1970s, as most of the functionality could be realized using analog circuits. The first-generation systems handled analog 3 kHz voice transmissions with very limited ability to transmit digital information.

3.10.2   Second Generation: Digital Radio

The second generation of radio is characterized by digitization.Many different types of digital systems have been installed around the world. The systems can transmit data and voice at rates of 8 to 14.4 kbps. This can be contrasted to the wireline system where, once signals reach the exchange, the 3 kHz analog signals are sampled at 64 kbps to achieve an undistorted

Table 3-2 Major mobile communication systems with the year of first widespread use.

System	Year	Description
0G		Broadcast, no cells, few users, analog modulation
MTS	1946	Mobile Telephone Service, half-duplex, operator assist to establish call, push to talk
AMTS	1965	Advanced Mobile Telephone System, Japan, full-duplex, 900 MHz
IMTS	1969	Improved Mobile Telephone Service, full-duplex, up to 13 channels, 60–100 km (40–60 mile) radius, direct dial using dual-tone multifrequency (DTMF) keypad
0.5G		FDMA, analog modulation
PALM	1971	(also Autotel) Public Automated Land Mobile radiotelephone service, used digital signaling for supervisory messages, technology link between IMTS and AMPS
ARP	1971	AutoRadioPuhelin (Car Radio Phone), obsolete in 2000, used cells (30 km radius) but not hand-off, 80 channels at 150 MHz, half-duplex and later full-duplex
1G		Analog modulation, FSK for signaling, cellular, FDMA
NMT	1981	Nordic Mobile Telephone, 12.5 kHz channel, 450 MHz, 900 MHz
AMPS	1983	Advanced Mobile Phone System, 30 kHz channel
TACS	1985	Total Access Communication Systems, 25 kHz channel, widely used until 1990s, similar to AMPS
Hicap	1988	NTT's mobile radiotelephone service in Japan
Mobitex	1990	National public access wireless data network, first public access wireless data communication services including two-way paging network services, 12.5 kHz channel, GMSK
DataTac	1990	Point-to-point wireless data communications standard (like Mobitex), wireless wide area network, 25 kHz channels, maximum bandwidth 19.2 kbps (used by the original Blackberry device)
2G		Digital modulation
PHS	1990	Personal Handyphone System, originally a cordless phone, now functions as both a cordless phone and as a mobile phone elsewhere
GSM	1991	Global System for Mobile Communications (formerly Groupe Spécial Mobile), TDMA, GMSK, constant envelope, 200 kHz channel, maximum 13.4 kb per time slot (at 1900 MHz), 2 billion customers in 210 countries
DAMPS	1991	Digital AMPS (formerly NADC [North American Digital Cellular] and prior to that as U.S. Digital Cellular [USDC]), narrowband, $\pi/4$DQPSK, 30 kHz channel
PDC	1992	Personal Digital Cellular, Japan, 25 kHz channel
CDMAOne	1995	Brand name of first CDMA system known as IS-95, spread spectrum, CDMA, 1.25 MHz channel, QPSK
CSD	1997	Circuit Switched Data, original data transmission format developed for GSM, maximum bandwidth 9.6 kbps, uses a single time slot

Table 3-2 continued.

System	Year	Description
2.5G		Higher data rates
WiDEN	1996	Wideband Integrated Dispatch Enhanced Network, combines four 25 kHz channels, maximum bandwidth 100 kbps
GPRS	2000	General Packet Radio System, compatible with GSM network, used GSM time slot and higher-order modulation to send 60 kb per time slot, 200 kHz channel, maximum bandwidth 171.2 kbps
HSCSD	2000	High-Speed Circuit-Switched Data, compatible with GSM network, maximum bandwidth 57.6 kbps based on CSD, higher quality of service than GPRS
2.75G		Medium bandwidth data—1 Mbps
CDMA2000	2000	CDMA, upgraded CDMAOne, double data rate, 1.25 MHz channel
EDGE	2003	Enhanced data rate for GSM Evolution, compatible with GSM network, 8PSK, TDMA, maximum bandwidth 384 kbps, 200 kHz channel
3G		Spread Spectrum
FOMA	2001	Freedom of Mobile Multimedia Access, first 3G service, NTT's implementation of WCDMA
UMTS		Universal Mobile Telephone Service, 5 MHz channel, data up to 2 Mbps
WCDMA	2004	Main 3G outside China
OFDMA	2007	Evolution to 4G (downlink high bandwidth data)
1xEV-DO		(IS-856) Evolution of CDMA2000, maximum downlink bandwidth 307 kbps, maximum uplink bandwidth 153 kbps
TD-SCDMA	2006	Time Division Synchronous CDMA, China, uses the same band for transmit and receive, base stations and mobiles use different time slots to communicate, 1.6 MHz channel
GAN/UMA	2006	Generic Access Network, formerly known as Unlicensed Mobile Access, provides GSM and GPRS mobile services over unlicensed spectrum technologies (e.g., Bluetooth and WiFi)
3.5G		
UMTS/HSDPA	2006	High-Speed Downlink Packet Access, download speeds to 7.2 Mbit/s
EV-DO Rev A	2006	CDMA2000 EV-DO Revision A, downlink speeds to 3.1 Mbps, uplink speeds to 1.8 Mbps
3.75G		
UMTS/HSUPA	2007	High-Speed Uplink Packet Access, upload speeds to 5.76 Mbps
EV-DO Rev B	2008	CDMA2000 EV-DO Revision B, Downlink speeds to 73 Mbps, uplink speeds to 27 Mbps
UMTS/HSPA	2009	High-Speed Packet Access, downlink speeds to 40 Mbps, upload speeds to 10 Mbps
4G	2011	Long-term evolution (LTE); Beyond 3G Low latency (e.g., for VoIP) + MIMO + OFDM + wireless broadband (WBB, > 100 Mbps) + software defined radio

signal representation. The cellular systems sacrifice some voice quality but use reasonably sophisticated algorithms that use the characteristics of speech to achieve greater than a factor of 4 compression.

Table 3-3 Attributes of the AMPS system.

Property	Attribute
Number of physical channels	832, 2 groups of 416 channels, each group has 21 signaling channels and 395 traffic or voice channels.
Bandwidth per channel	30 kHz
Cell radius	2–20 km
Base-to-mobile frequency	869–894 MHz (downlink)
Mobile-to-base frequency	824–849 MHz (uplink)
	45 MHz between transmit and receive channels.
Channel spacing	30 kHz
Modulation	FM with peak frequency deviation of ±12 kHz
	Signaling channel uses FSK
	Can send data at 10 kbps
Access method	FDMA
Base station ERP	100 W per channel (maximum)
Channel coding	None
RF specifications of mobile unit	
Transmit RF power	3 W maximum (33 dBm) (600 mW for hand-held)
Transmit power control	10 steps of 4 dB attenuation each, minimum power is -4 dBm
Receive sensitivity	-116 dBm from a 50 W source applied at antenna terminals
Receive noise figure	6 dB measured at antenna terminals
Receive spurious response	-60 dB from center of the passband
Number of synthesizer channels	832

In North America the first digital system introduced was the **digital advanced mobile phone system** (**DAMPS**) (originally known as **North American Digital Cellular** [**NADC**] and as the EIA/TIA interim standard **IS-54**). The system was designed to provide a transition from the then current analog system to a fully digital system by reusing existing spectrum. The idea is that system providers can allocate a few of their channels for digital radio out of the total available. As analog radio is phased out, more of the channels can be committed to digital radio. The main motivation behind this system is that it provides three to five times the capacity of the analog system. The European-origin GSM system provides a similar increase in capacity, is compatible with **Integrated Services Digital Network** (**ISDN**) and has largely replaced DAMPS. The GSM system was initially (early 1990s) dominant in Europe and had the advantage that it did not need to be upward compatible with uncoordinated analog phone systems. The attributes of the GSM system are shown in Tables 3-4 and 3-5.

From an RF design perspective, the main differences between analog and digital standards are

Table 3-4 Attributes of the GSM system. Uplink and downlink frequencies of the various GSM implementations are given in Table 3-5.

Property	Attribute
Number of physical channels	125 (GSM-900, 1 channel is special so sometimes 124 is quoted)
Bandwidth per channel	200 kHz
Channel spacing	200 kHz
Cell radius	2–20 km
Base-to-mobile frequency	935–960 MHz (GSM-900)
Mobile-to-base frequency	890–915 MHz (GSM-900)
	55 MHz between transmit and receive channels.-
Modulation	GMSK
	Slow frequency hopping (217 hops/s)
Access method	TDMA
	8 slots per frame, user has one slot,
	each frame is 4.615 ms and each slot is 557 μs.
Speech code	8 kbps, VSELP code
Channel coding	Block and convolutional coding
Transmit rate	270.833 kbps
Equalizer	16 μs
Synchronization	Up to 233 μs absolute delay

1. The RF envelope. In AMPS, FSK is used which produces a constant envelope RF signal. Consequently high-efficiency saturation mode amplifiers (such as class C) can be used. In most digital modulation schemes, but not GSM, the modulation results in a nonconstant envelope. The information contained in the amplitude of the RF signal is just as important as the information contained in the phase or frequency of the signal. Consequently saturation mode amplifiers which severely distort the amplitude characteristic must be avoided.

2. Bursty RF. In an analog system, RF power is continually being transmitted. In a digital system, transmission is intermittent and the RF signal is bursty. Therefore an RF designer must be concerned about turn-on transients and thermal stability of the power amplifier.

3.10.3 Generation 2.5: Direct Sequence Code Division Multiple Access, CDMA

CDMA, or more specifically CDMAOne, was initially promoted as being third generation, but the definition now is that data rates of at least 2 Mbps must be supported. Thus CDMAOne is now referred to as 2.5G. A depiction of spread spectrum is shown in Figure 3-14, in which a very fast code is superimposed on a slower data sequence and the combined code is used

Table 3-5 GSM frequency bands. GSM channels have a bandwidth of 200 kHz. The base-to-mobile transmission is the downlink and the mobile-to-base station transmission is the uplink.

Band	Uplink (MHz)	Downlink (MHz)	Duplex Spacing (MHz)
GSM-900 Most of world	890–915	935–960	45
GSM-900 extended	876–915	921–960	45
GSM-1800 (Americas)	1710–1785	1805–1880	95
GSM-400 GSM-400 alt. (Nordic, Eastern Europe, Russia)	450.4–457.6 478.8–486	460.4–467.6 488.8–496	10 10
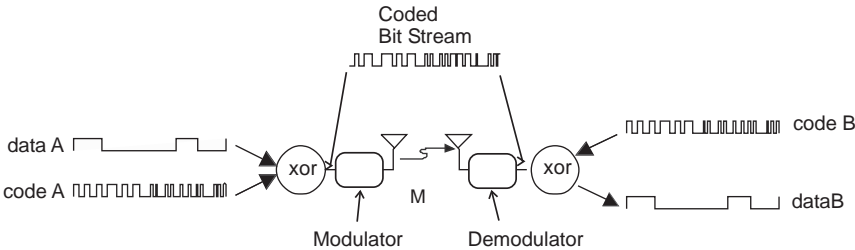
Figure 3-14 Depiction of direct sequence CDMA transmission.

to modulate a carrier. The same code is used to extract the baseband signal from the received bitstream. The effect of the code is to greatly spread out the baseband signal, transforming perhaps a 12 kbps bitstream into an RF signal with a bandwidth of 1.23 MHz.

The key feature of the DS-CDMA system is the use of lengthy codes to spread the spectrum of the signal that is to be transmitted. In the case of voice an 8 kbps second code (for example) is mixed with a much faster code which is unique to a particular user. Thus the 8 kbps bitstream becomes a 1.23 MHz wide analog or baseband signal. This signal is then modulated up to RF and transmitted. On the receiver side, the demodulated RF signal can only be decoded using the original code. Use of the original code to decode the signal rejects virtually all interference and noise in the received signal. Since orthogonal codes are used many radio channels can be supported on the same radio link. CDMA can support approximately 120 radio channels on the same physical channel. Another important feature is that the same 120 frequencies can be reused in adjacent cells, as the radio channels can still be extracted. Thus there is no need for clustering as in the TDMA and FDMA systems. The attributes of the cellular CDMAOne system are given

Table 3-6   Attributes of the cellular CDMAOne system.

Property	Attribute
Bandwidth per channel	1.23 MHz
Channel spacing	1.25 MHz (20 kHz guard band)
Cell radius	2–20 km
Base-to-mobile frequency	869–894 MHz
Mobile-to-base frequency	824–849 MHz
	45 MHz between transmit and receive channels.
Modulation	OPSK
Access method	CDMA
	64 radio channels per physical channel
	Forward:
	55 traffic channels, 7 paging channels
	1 pilot channel 1 sync channel
	Reverse:
	55 traffic channels, 9 access channels
RF specifications of mobile unit	
Transmit power control	1 dB power control

in Table 3-6.

3.10.4   Third Generation: High Speed Data

Third-Generation Radio,[18] (**3G**) includes wideband mobile multimedia networks and broadband mixed wireless systems. The mobile systems support variable data rates depending on demand and the level of mobility. At least 144 kbps is supported for full vehicular mobility and higher bit rates for pedestrian levels of mobility. Switched packet radio techniques and wideband CDMA-like systems are required to support this bandwidth-on-demand environment. Here the physical channel is shared (i.e., packet-switched) rather than the user being assigned a physical channel for exclusive use (referred to as circuit-switched).

   The drive for 3G systems was partly fueled by the saturation of 2G systems in many places and a desire to increase revenues by supporting high-speed data. Prior to the rollout of 3G systems, the increased demand primarily resulted from an increased consumer base rather than the emergence of significant data traffic. The increased subscriber base was addressed by 2.5G systems, which have some of the 3G concepts, but

---

[18]   Third-generation radio is coordinated by the **3rd Generation Partnership Project** (**3GPP**). This is a collaborative agreement of standards development organizations and other related bodies for the production of a complete set of globally applicable technical specifications and reports for a 3G system, see http://www.3gpp.org.

only partially implemented. However, the belief of service providers is that the data communications will significantly outpace voice and full implementation of 3G systems is essential. Third-generation systems are about more than increased capacity. The driving concepts are the development of a global standard that supports global roaming and supports advanced features including two-way motion video and Internet browsing.

Third-generation cellular radio is defined by the **International Telecommunications Union** (ITU) [35] and is formally called International Mobile Telecommunications 2000 (IMT-2000). The basic requirements are for a system that supports data rates up to 2 Mbps in fixed environments ranging down to 144 kbps in wide area mobile environments. In 1999 the ITU adopted five radio interfaces for IMT-2000:

1. IMT-DS Direct-Sequence CDMA, more commonly known as wideband CDMA (WCDMA)

2. IMT-MC Multi-Carrier CDMA, more commonly known as CDMA2000, the successor to CDMAOne (specifically international standard IS-95)

3. IMT-SC Time-Division CDMA, which includes Time Division CDMA (TD-CDMA) and Time Division Synchronous CDMA (TD-SCDMA)

4. IMT-SC Single Carrier, more commonly known as EDGE

5. IMT-FT Frequency Time, more commonly known as DECT

The dominant choices for 3G are WCDMA and CDMA2000. In October 2007 the ITU Radiocommunication Assembly included WiMax-derived technology, specifically **Orthogonal Frequency Division Multiple Access** (**OFDMA**) and MIMO, in the set of IMT-2000 standards as the sixth radio interface. The **3rd Generation Partnership Project** (**3GPP**) [5] unites telecommunications standards related to 3G, and establishes a migration path from 2G to 3G and eventually beyond 3G or **Long-Term Evolution** (**LTE**). The ITU family of 3G standards are collectively known as IMT-2000 standards.

The term "3GPP specification" covers GSM (including GPRS and EDGE), W-CDMA, and LTE specifications. EDGE has intermediate data speeds between those of GSM and WCDMA. The following terms are also used to describe networks using the 3G WCDMA specification: Universal Mobile Telecommunication System (UMTS, in Europe), UMTS Terrestrial Radio Access Network (UTRAN), and **Freedom of Mobile Multimedia Access** (**FOMA**, in Japan). UMTS is the 3G successor of the GSM standard with the air interface now using WCDMA. The terminology used in UMTS, listed in part in Table 3-7, is based on the terminology used in GSM, with subtle differences. UMTS was first deployed in Japan in 2001 (by NTT DoCoMo).

Table 3-7 UMTS terminology.

Term	Description
AuC	Authentication Center
GGSN	Gateway GPRS Support Node
GMSC	Gateway MSC
HLR	Home Location Register
ISDN	Integrated Services Digital Network
MSC	Mobile Switching Center
Node B	Base station
PSTN	Public Switched Telephone Network
RNC	Radio Network Controller
SGSN	Serving GPRS Support Node
UE	User Equipment
USIM	Universal Subscriber Identity Module
VLR	Visitor Location Register

Table 3-8 Spectrum assignments for FDD 3GPP [5, Release 8].

Band	Uplink (MHz)	Downlink (MHz)	Available spectrum					
			NA	LA	EMEA	ASIA	Oceania	Japan
1	1920–1980	2110–2170	■	■	■	■	■	■
2	1850–1910	1930–1990	■	■		■	■	■
3	1710–1785	1805–1880		■	■			
4	1710–1755	2110–2155	■		■			
5	824–849	869–894	■	■	■	■	■	■
6	830–840	875–885						■
7	2500–2570	2620–2690			■			
8	880–915	925–960	■	■	■	■	■	■
9	1749.9–1784.9	1844.9–1879.9						■
10	1710–1770	2110–2170						

UMTS is now deployed in most regions of the world, but coverage is limited to major cities. UMTS combines the WCDMA air interface with the speech codecs and application support of GSM. The term WCDMA describes the physical interface and protocols that support it, while UMTS refers to the whole network. In most countries UMTS uses 1850–1910 MHz and 1920–1980, for the mobile-to-base station link (the uplink), and 1930–1990 MHz and 2110–2170 MHz for the base station-to-mobile link (the downlink). However, there are a large number of frequencies designated for FDD 3GPP with multiple band variants based on existing spectrum allocations in different geographic regions (see Table 3-8).

The 3G timeline is summarized in Figure 3-15. The CDMA2000 and WCDMA paths become the single LTE path beyond 3G. The CDMA2000 (the IS-2000 standard) path builds on the original CDMA system defined
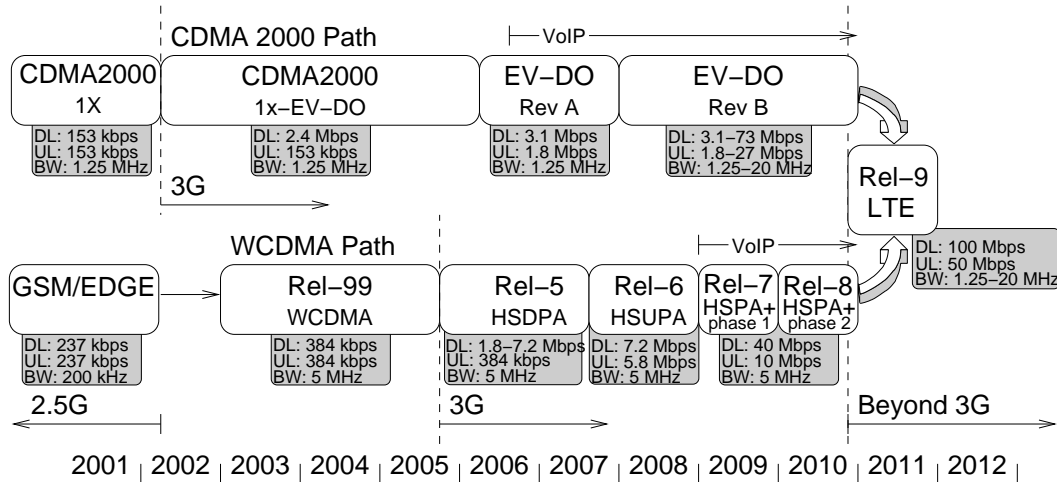
Figure 3-15 Timeline for implementation of 3G and Long-Term Evolution (LTE). DL indicates the downlink data rate; UL indicates the uplink data rate; BW indicates the channel bandwidth. Both paths support Internet Protocol (IP) and Voice over IP (VoIP) during their evolution. The two 3G paths combine into a single LTE path.

by the IS-95 standard and commonly known as CDMAOne. CDMA2000 1xEV-DO is the first evolution of CDMA2000 that meets the ITU basic specification for 3G. **Evolution-Data Optimized** (**EV-DO**), combines CDMA and TDMA for higher data throughput. Subsequent revisions (Rev-A and Rev-B) provide higher data rates and can use more than one 1.25 MHz frequency channel. Rev-A and Rev-B are designed to operate end-to-end as an Internet Protocol (IP)-based network. It can support any application which can operate on such a network. For voice communications it uses Voice over IP (VoIP).

The WCDMA and LTE evolution is defined by releases beginning with an initial release known as Release 99 (Rel-99) [36]. The releases are designed to protect the installed investment in GSM-based 2G systems while providing a migration path to 3G and beyond. With Rel-9 and the definition of LTE, the CMDA2000 and WCDMA paths become a single LTE path. The timeline indicates the earliest adoption times, but widespread adoption is expected to be many years out, with the deployment of LTE expected to be significant after 2015.

So far the 3G schemes discussed have used FDD. China is promoting a TDD form of 3GPP, TD-CDMA. TD-CDMA has its own bands. These are shown in Table 3-9 and Bands 38 and 40 are dominant.

Table 3-9 Spectrum assignments for TDD 3GPP [5, Release 8].

Band	Uplink (MHz)	Downlink (MHz)
33	1900–1920	1900-1920
34	2010–2025	2010–2025
35	1850–1910	1850–1920
36	1930–1990	1930–1990
37	1910–1930	1910–1930
38	2570–2620	2570–2620
39	1880–1920	1880–1920
40	2300–2400	2300–2400

3.11 Long-Term Evolution: Beyond 3G

"Beyond 3G" is a common designation for what many call 4G, long term evolution (LTE), or NextGen[19]. There are two essential concepts beyond 3G. One of these is providing data transmission at rates of 100 Mbps while mobile and 1 Gbps while stationery. The other concept is that of pervasive networks where a handset supports many access technologies (e.g., cellular, UMTS, WiFi, etc.), perhaps simultaneously, and smoothly transitions between them. For example, to support high-speed data it is more efficient to use an available WiFi network than a UMTS network.

Two technological advances that enable the high data rate concept are the bases of the evolution of cellular communications and wireless networks. One advance is **Orthogonal Frequency Division Multiplexing** (OFDM), which sends multiple relatively slow bitstreams with one of the bitstreams on each of a large number of carriers. OFDM reduces the impact of fading as symbols are spread out over relatively long periods of time. The other advance is MIMO, which relies on multipath to send multiple versions of several bitstreams transmitted from several antennas. These schemes are discussed below, and with them it is possible to greatly increase spectral efficiency.

LTE cannot co-exist with 3G systems and separate bands have been allocated. The bands designated by the ITU World Radiocommunication Conference 2007 (WRC-2007) are listed in Table 3-10.

---

[19] This is a political statement, as many wireless providers do not want 4G to be hyped while 3G is not fully adopted. There is no official definition of 4G. Both LTE and WiMax have been called 4G.

Table 3-10 Frequency bands allocated by the ITU World Radiocommunication Conference 2007 (WRC-2007) for LTE or 4G.

Frequency band	Bandwidth
450–470 MHz	20 MHz
790–862 MHz ITU Reg. 1	72 MHz
698–806 MHz ITU Reg. 2,3	108 MHz
2300–2400 MHz	100 MHz
3400–3600 MHz	200 MHz
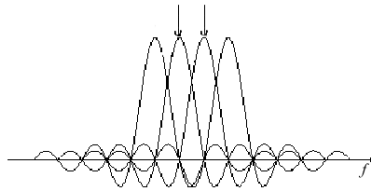
Figure 3-16 OFDM spectrum with four subcarriers showing orthogonality.

3.11.1 Orthogonal Frequency Division Multiplexing, OFDM

In OFDM, data are simultaneously sent over multiple channels or **subcarriers** with the special property that the subcarriers are orthogonal and precisely spaced in frequency and time. Each subcarrier is transmitted simultaneously however. Each subcarrier is modulated so each becomes a subchannel carrying its own datastream. This is shown in the spectrum of Figure 3-16, where the arrows at the top indicate sampling points for two subcarriers. The subcarriers are ideally orthogonal. This can be seen in part by noticing that the peak of one subcarrier is at the zeros of the other subcarriers. When one subcarrier is sampled, the contribution from all other carriers is zero; they are orthogonal. The spectra of the subcarriers overlap, but this does not matter. This scheme enables high-speed data transmission over possibly hostile channels of an unregulated band. OFDM is a spread spectrum technique in that data are spread over a large number of subcarriers. Thus the impact of multipath is mitigated, as each subcarrier has a relatively narrow bandwidth, or in time, a long duration. Signal strength and interference, and hence SIR, can differ for each channel and this is compensated for by having different bit rates in each subchannel, and also by adjusting the power level in each subchannel.

OFDM can be implemented by using separate modulators and demodulators for each subcarrier, however, this is rarely done. It is also possible and more practical to replace the separate modulators and demodulators by a **Fast Fourier Transform** (**FFT**) and an **Inverse FFT** (**IFFT**), respectively, implemented in DSP. The signal comprising all of the subcarriers is then modulated onto a single RF carrier. This FFT/IFFT implementation of OFDM is

Table 3-11 Modulation formats of the OFDM WLAN system denoted by IEEE Standard 802.11a, generally known as WiFi [37–39].

Data rate (Mbps)	Modulation	Coding rate	Coded bits per subcarrier	Code bits per OFDM symbol	Data bits per OFDM symbol
6	BPSK	1/2	1	48	24
9	BPSK	3/4	1	48	36
12	QPSK	1/2	2	64	48
18	QPSK	3/4	2	96	72
24	16QAM	1/2	4	192	96
36	16QAM	1/2	4	192	144
48	16QAM	3/4	4	288	192
54	64QAM	2/3	6	288	216

called **Discrete Multi-Tone** (**DMT**) or **OFDM/DMT**. Here the subcarriers share a common carrier and the frequency outputs of the FFT of the datastream are the subcarriers. Generally each subcarrier has its own bitstream and the total bitstream is carried on perhaps 256 subcarriers. With **Forward Error Correction** (**FEC**) coding, a large fraction of the bits could be lost (because individual subcarriers collide with other signals), but the datastream can still be recovered.

Unfortunately the **Peak-to-Average power Ratio** (**PAR**) of OFDM is large. An approximation of the problem is to view an OFDM signal as the composite of a large number of tones. In practice, the PAR is reduced through clipping of the signal in DSP (and hence some data are lost). However, using error correction coding, it is possible to correct and recover from these errors and recover the missing bits. In the ideal situation the subcarriers are orthogonal, but timing and frequency offsets cause the subchannels to interfere with each other. This interference can be reduced by limiting the number of subchannels and using special pulse shapes that are more robust to timing errors.

OFDM is the transmission method for digital radio, digital TV (in Europe), high-speed wireless local and metropolitan area networks (WLAN and WMAN respectively; e.g., WiFi[20] and WiMax [21]), and broadband Internet over phone lines in Digital Subscriber Line (DSL). OFDM achieves close to maximum spectral efficiency. The OFDM system implemented in the IEEE Standard 802.11a, one form of WiFi, uses several modulation formats with different throughputs (see Table 3-11). The difference between the number of data bits per symbol and the number of coded bits per symbol is the coding added in the DSP for error correction and to provide additional information

[20] Specifically IEEE 802.11a WLAN standard achieving 54 Mbps [37–39].
[21] Specifically IEEE 802.16 WMAN standard achieving 18.36 Mbps [40–42].

Table 3-12 Modulation formats of the OFDM WLAN system, denoted IEEE Standard 802.16d, with a 5 MHz bandwidth, generally known as WiMax [40–42].

Data rate (Mbps)	Modulation	Coding Rate	Coded bits per subcarrier	Code bits per OFDM symbol	Data bits per OFDM symbol
1.89	BPSK	1/2	0.5	176	88
3.95	QPSK	1/2	1	368	184
6.00	QPSK	3/4	1.5	512	280
8.06	16QAM	1/2	2	752	376
12.18	16QAM	3/4	3	752	568
16.30	64QAM	2/3	4	1140	760
18.36	64QAM	3/4	4.5	1136	856

about the channel. The ratio of the number of data bits to the total number of bits (useful data bits plus forward error correction code bits) is the **coding rate** or **code rate**.

The system described in IEEE Standard 802.16, generally known as **WiMax**, also uses OFDM. It is designed to cover a wide geographical area and hence is designated as WMAN. It is a cellular system implementing universal frequency reuse with cells that are typically a few kilometers in diameter. The downlink throughput typically averages 3 Mbps over a 5 MHz bandwidth when there is a single antenna at the receiver and with a trisector transmit antenna. In the same 5 MHz bandwidth 7 Mbps can be achieved with two receive antennas and with six-sector cells (and hence lower interference from other cells). WiMax uses several modulation formats, as shown in Table 3-12. Higher-order modulation such as 64QAM can only be achieved when interference is low.

WiMax has been specified to operate in frequency bands from 2 to 66 GHz, with few bands requiring licenses. Most activity is below 6 GHz with most systems expected to operate in the 2.3 GHz, 2.5 GHz, or 5 GHz bands. With mesh networking capability, WiMax nodes will simultaneously operate in "subscriber station" and "base station" modes. WiMax handles multiple users efficiently. In WiFi subscriber units compete to connect to an **Access Point** (**AP**), creating collisions when there are multiple users and a rapid degradation in the **quality of service** (**QoS**) In WiMax the AP assigns a time slot that a terminal unit uses indefinitely. The time slot can change in length, but collisions are avoided.

3.11.2 Orthogonal Frequency Division Multiple Access—OFDMA

OFDMA is the multi-user version of OFDM that supports simultaneous communication by multiple users and so is an access technology. In OFDM, multiple orthogonal subcarriers are used with each subcarrier carrying an
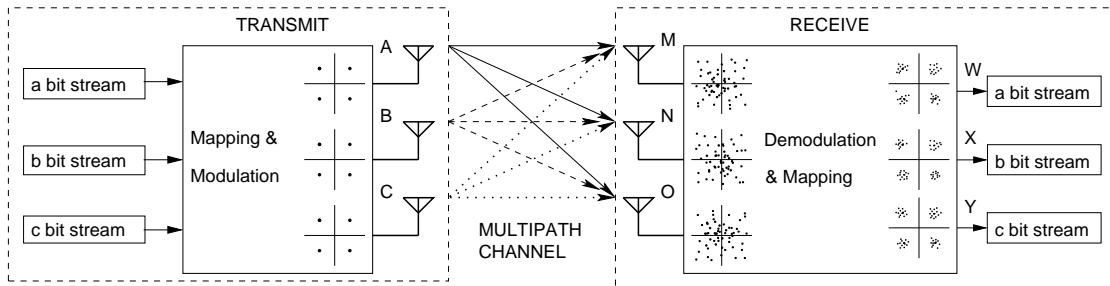
Figure 3-17   A MIMO system showing multiple paths between each transmit antenna and each receive antenna.

independent bitstream. In OFDMA, generally one or more subcarriers is assigned to a particular user and this are orthogonal to the subcarriers used by other users. (It is possible, however, for users to share subcarriers if other schemes are used to separate users.) OFDMA assigns subcarriers to a particular user for a specific time period. OFDMA is also known as multiuser-OFDM and is used in the WiMax **Wireless MAN** air interface standard.

3.11.3   Multiple Input, Multiple Output, MIMO

Multiple input, multiple output (MIMO, pronounced my-moe) technology uses multiple antennas to transmit and receive signals. The MIMO concept was developed in the 1990s [43, 44] and implemented in a variety of WLAN systems and evolving cellular communications standards. There are several aspects to MIMO. First, each transmit antenna sends different datastreams simultaneously on the same frequency channel as other transmit antennas. The most interesting feature is that MIMO relies on signals traveling on multiple paths between an array of transmit antennas and an array of receive antennas. In a conventional communications system the various paths result in interference and fading, but in MIMO these paths are used to carry more information. In a MIMO system, each path propagates an image of one transmitted signal (from one antenna) that differs in both amplitude and phase from the images following other paths. Each image arrives at one of the receive antennas at slightly different times and the phase differences are used to differentiate between them. Effectively there are multiple connections between each transmit antenna and each receive antenna (see Figure 3-17). For simplicity, three transmit antennas and three receive antennas are shown. MIMO, however, can work with as few as two transmit antennas and one receive antenna, but the capacity is much higher with more. The high-speed datastream is split into several slower datastreams, shown in Figure 3-17 as the a, b, and c datastreams.

The bitstreams are mapped so different versions of the datastreams are combined, modulated, and sent from each transmit antenna, with the constellation diagrams labeled A, B, and C. The signal from each transmit antenna reaches all of the receive antennas by following different paths.

The output of each receive antenna is a linear combination of the multiple transmitted datastreams, with the sampled RF phasor diagrams labeled M, N, and O. (It is not really appropriate to call these constellation diagrams.) That is, each receive antenna has a different linear combination of the multiple images. In effect, the output from each receive antenna can be thought of as the solution of a linear equation, with each path corresponding to an equation. Continuing the analogy, the signal from each transmit antenna represents a variable. So the set of simultaneous equations can be solved to obtain the original bitstreams. This is accomplished by demodulation and mapping with some knowledge of the channel characteristics, yielding the original transmitted signals modified by interference. The result is that the constellation diagrams W, X, and Y are obtained. The composite channel can be characterized using test signals. Special coding called **space-time** (or **spatiotemporal**) coding embedded in the transmitted datastream also enables estimation of the communication matrix. Space-time coding encodes each transmitted datastream with information that can be used to assist in reconstructing the signals on the others. This is more robust than characterizing the channel with test signals sent at a different time.

The capacity of a MIMO system with high SIR scales approximately linearly with the minimum of M and N, $\min(M, N)$, where M is the number of transmit antennas and N is the number of receive antennas (provided that there is a rich set of paths) [45, 46]. So a system with $M = N = 4$ will have four times the capacity of a system with just one transmit antenna and one receive antenna. Table 3-13 presents the capacity of a MIMO system with ideal PSK modulation (i.e., without modifications to control PAR) and two transmit and two receive antennas. This is compared to the capacity of a conventional (non-MIMO) system. The capacity is presented in bits per second per hertz and it is seen that significant increases in throughput are obtained when SIR is high. MIMO is incorporated in the WiMax (IEEE 802.16d) and WiFi (IEEE 802.11n) standards, where spectral efficiencies of $6.35\,\mathrm{b \cdot s^{-1} \cdot Hz^{-1}}$ have been achieved in commercial systems.

In summary, MIMO systems achieve throughput and range improvements through four gains achieved simultaneously:

1. Array gain resulting from increased average received SNR obtained by coherently combining the incoming and outgoing signals. To exploit this the channel must be characterized. This increases coverage and QoS.

2. Diversity gain obtained by presenting the receiver with multiple

Table 3-13 Capacity of MIMO schemes with PSK modulation for different received SIRs compared to the maximum capacity of a conventional non-MIMO scheme. M is the number of transmit antennas, N is the number of receive antennas. Data are from He and Georghiades [47].

Modulation scheme	Capacity b·s^{-1}Hz·$^{-1}$ (bits per second per hertz)				
	Non-MIMO $M = 1, N = 1$ maximum	MIMO with $M = 2, N = 2$			
		SIR 0 dB	SIR 10 dB	SIR 20 dB	SIR 30 dB
BPSK	1	1.2	2	2	2
QPSK	2	1.6	3.7	4	4
8PSK	3	1.6	4.8	6	6
16PSK	4	1.6	4.9	7.5	8

identical copies of a given signal. This combats fading. This also increases coverage and QoS.

3. Multiplexing gain by transmitting independent data signals from different antennas to increase throughput. This increases spectral efficiency.

4. Cochannel interference reduction. This increases cellular capacity.

MIMO can be combined with spreading to obtain a scheme denoted as MIMO-CDMA. MIMO-CDMA achieves greater capacity than MIMO-OFDM when SIR is low, generally below 10 dB [48]. At high SIR, however, MIMO-OFDM achieves higher capacity than MIMO-CDMA.

3.11.4   Spectrum Utilization

Allocation of the EM spectrum is controlled by governments with a reasonable degree of coordination through the ITU. The spectrum is divided into nonoverlapping blocks of frequencies with restrictions to specific geographic areas, technical characteristics, output power levels, and applications. Generally, rigid specifications are required to avoid interference with other users. The structure of frequency assignment in the EM spectrum and the control of spectrum usage is essential to radio technology development which utilizes fixed frequency, narrowband circuits, and architectures. It is also much easier to develop RF hardware for fixed narrowband systems. Another important factor is that spectrum assignment has enabled auctions of spectrum bands and a significant source of revenue for governments; no government will relinquish this revenue easily. Owners of spectrum are also quite pleased with the current system, as they can be reasonably sure of being free of interference. It is important to have a predictable medium after significant investments in

capital expenditure (CapEx) and license costs. However, a big price has been paid for the current situation.

For many years there have been claims that the spectrum is crowded and that the spectrum is a limited resource, resulting in what is called "spectrum crowding." This is a consequence of the assignment of specific bands for specific purposes and dedicated, except for a few ISM bands, to licensed operators.[22] The latest paradigm shift in spectrum assignments has been the adoption of cellular radio technology, which, in part, addresses this problem of limited spectrum. In radio systems, the concept of avoiding interference is central to the assignment of spectrum utilization to geographic areas. This promotes the move to higher frequencies of operation, higher licensing fees (through competitive bidding), and domination of the industry often by very large enterprises. However, there are significant changes under way. WiFi is having a tremendous impact, with significant erosion of the price of wireless connectivity. WiFi has been quickly adopted by consumers and many stores and cities are installing open WiFi systems to attract customers. WiFi operates in two bands around 2.4 GHz and 5.6 GHz, but other than that it is fairly loose, with individual radio units determining the best conditions under which to operate. The shared spectrum movement enables different vendors to share the same spectrum and to rent spectrum as needed per use. Commercial exploitation of shared spectrum began in 2008 using **cognitive radio**. This is a new paradigm where a radio senses its environment and decides how to make best use of it. Many different kinds of radio units have been considered—sophisticated signal processing is enabling RF hardware to be multifunctional, operate in a number of bands, and implement a number of functions (GPS, cellular, etc.). In the case of cellular radio handsets some units have a common transmit amplifier covering the 1 GHz and 2 GHz bands. Antenna systems are being built that can operate from 800 MHz to 2.5 GHz. All this leads to a radical idea for the use of the EM spectrum. The idea is to develop wideband hardware that adaptively utilizes the spectrum. One approach is to sniff out the EM environment and use spectrum that is not currently being utilized, even if it is in a licensed band. The promise is a dramatic increase in spectrum utilization. Implementing these ideas will require significant changes to the regulatory environment. Concepts of renting spectrum are being thought through. Much more likely in the long term is the idea of transmitting signals at power levels significantly below the noise floor—so called **Ultra-Wideband Systems** (**UWB**).

---

[22] This has its roots in the Titanic shipwreck aftermath. A decision that was necessary at the time, but which had a significant impact on the development of radio.
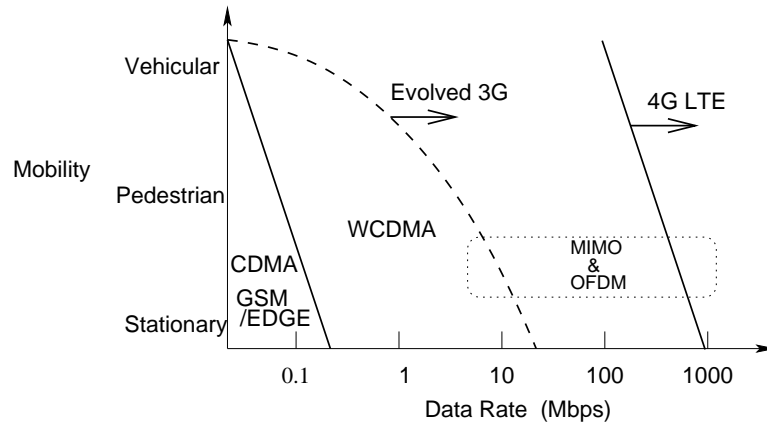
Figure 3-18 Data rate capacity of evolved 3G and beyond cellular communications. LTE is long term evolution which is also called 4G.

3.11.5 Summary of Evolved 3G and 4G

With the advent of MIMO it is necessary to modify Shannon's capacity limit as MIMO systems can exceed the limit defined for a single channel. Shannon's capacity limit for a MIMO system becomes

$$\hat{C} = B_c \ln(1 + \text{SIR} \cdot H) \qquad (3.22)$$

where H is a MIMO capacity factor that depends on $\min(M, N)$ [45]. There seems little doubt that 4G cellular communication technologies will be based on a combination of OFDM and MIMO (MIMO-OFDM) or a combination of MIMO and CDMA (MIMO-CDMA). There is a tremendous increase in channel carrying capacity, especially when the SIR is high. The data rate capacity of 2G systems (GSM and CDMA) is contrasted with the capacity of 3G (WCDMA) and what are called evolved 3G and 4G systems in Figure 3-18. Evolved 3G incorporates many of the features of 4G. Fourth generation (4G), including LTE, is the implementation of MIMO and OFDM or CDMA technologies with space-time coding. achieving packet-switched low-latency data transmission at mobile rates up to 100 Mbps and rates while stationary of 1 Gbps. Some 4G systems will use multiple channels, including multiple standards simultaneously. The defining characteristic of 4G will be universal pervasiveness.

3.12 Radar Systems

Radar uses EM signals to determine the range, altitude, direction, and speed of objects called targets by looking at the signals received from transmitted signals called radar waveforms. The earliest use of EM signals to detect targets was demonstrated in 1904 by Christian Hülsmeyer using a spark gap

Table 3-14   IEEE radar bands and applications.

Band	Frequency	Wavelength	Application
HF	3–30 MHz	10–100 m	Over-the-horizon radar, oceanographic mapping
VHF	30–300 MHz	1–10 m	Oceanographic mapping, atmospheric monitoring, long-range search
UHF	0.3–1 GHz	1 m–30 cm	Long-range surveillance, foliage penetration, ground penetration, atmospheric monitoring
L	1–2 GHz	15–30 cm	Satellite imagery, mapping, long-range surveillance, environmental monitoring
S	2–4 GHz	7.5–15 cm	Weather radar, air traffic control, surveillance, search, IFF (identify, friend or foe)
C	4–8 GHz	3.75–7.5 cm	Hydrologic radar, topography, fire control, weather
X	8–12 GHz	2.5–3.75 cm	Cloud radar, air-to-air missile seeker, maritime, air turbulence, police radar, high-resolution imaging, perimeter surveillance
Ku	12–18 GHz	1.7–2.5 cm	Remote sensing, short-range fire control, perimeter surveillance; pronounced "kay-you"
Ka	27–40 GHz	7.5–12 mm	Police radar, weapon guidance, remote sensing, perimeter surveillance, weapon guidance; pronounced "kay-a"
V	40–75 GHz	4–7.5 mm	Perimeter surveillance, remote sensing, weapon guidance
W	75–110 GHz	2.7–4 mm	Perimeter surveillance, remote sensing, weapon guidance

generator [49]. This system was promoted as a system to avoid collisions of ships and detected the direction of targets only. Research contributed to further developments, with a significant acceleration during World War II. Radar is now a word in its own right, but in 1941 the term RADAR was created as an acronym for radio detection and ranging.

In a radar system, typically a high-gain antenna such as a parabolic antenna is used to transmit a radar signal, but always a high-gain antenna is used to receive the signal. If the same antenna is used for transmit and receive (but possibly two similar antennas at the same site) the system is called a **monostatic radar** (see Figure 3-19(a)). Radar with transmit and receive antennas at different sites is called **bistatic radar** (shown in Figure 3-19(b)).

In a monostatic radar using the same antenna for transmit and receive, the space is painted with a radar signal and the received signal captured after the propagation delay from the antenna to the target and back again. A radar image can then be developed. It is only necessary, however, that the receive antenna move. In many radars the receive antenna is mechanically steered and often a regular rotation is used. With so-called synthetic aperture radars, a platform such as an aircraft moves the radar in one direction and a one-dimensional mechanical or electrical scan enables a two-dimensional image to be developed.
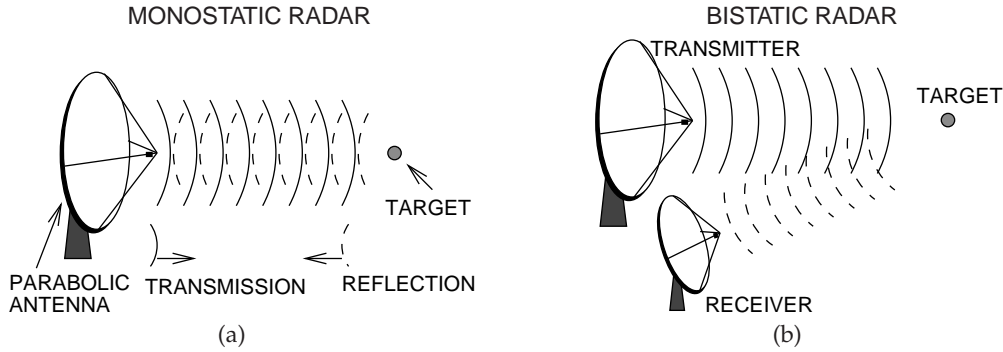
Figure 3-19 Radar system: (a) monostatic radar with the same site used for transmission of the radar signal and receipt of the reflection from the target; and (b) bistatic radar with different transmit and receive sites.
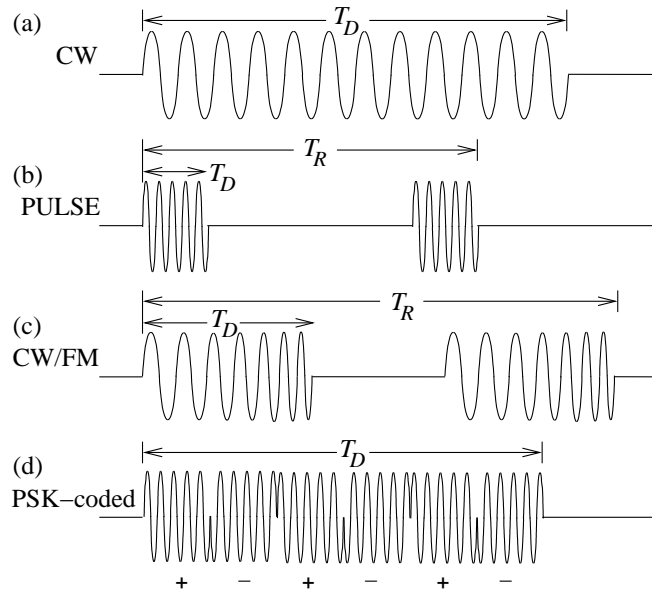
Figure 3-20 Radar waveforms: (a) continuous wave; (b) pulsed wave; (c) frequency modulated continuous wave; and (d) phase-encoded (PSK-coded) waveform.

The categories of radar waveforms are shown in Figure 3-20. The continuous wave (CW) waveform shown in Figure 3-20(a) is on all or most of the time and is used to detect a reflection from a target. This reflected signal is much smaller than the transmitted signal and it can be difficult to separate the transmitted and received signals. A monostatic CW radar architecture is shown Figure 3-21(a), where a circulator[23] is used to separate
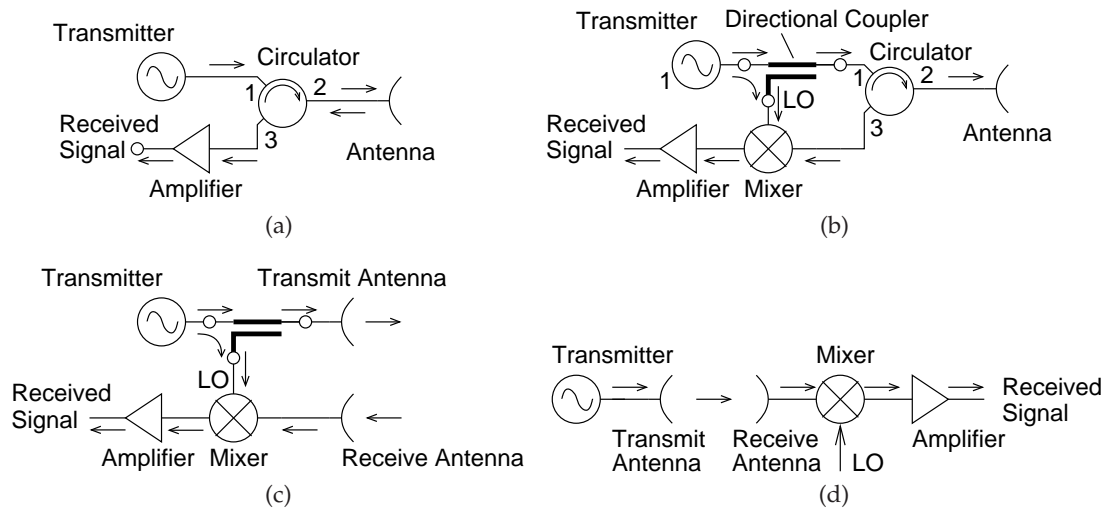
Figure 3-21 Radar architectures: (a) CW radar; (b) CW Doppler radar; (c) CW Doppler radar with separate transmit and receive antennas; and (d) bistatic CW Doppler radar.

the transmitted and received signals. The received signal is converted to digital form using an ADC and the bandwidth of the ADC with the required dynamic range determines the limit on the bandwidth of the radar signal. Generally the broader the bandwidth, the better the radar system is at identifying objects. A CW signal can be used to develop an image, but is not good at determining the range of a target. For this, a pulsed radar waveform, as shown in Figure 3-20(b), is better. The repetition period, T_R, is more than the round-trip time to the target, thus the time interval between the transmitted signal and the returned signal can be used to estimate range. Direction is determined by the orientation of the antenna.

The CW architecture can also be used with **pulsed radar**. In pulsed radar, the signal received contains the desired target signal, multipath and echoes, and clutter, as shown in Figure 3-22(a). These effects also appear in the signal received in CW radar, but it is much easier to see in pulsed radar. Identifying clutter and multipath effects is a major topic in radar processing. Alternative waveforms, especially digitally modulated waveforms, aid in extracting the desired information. The frequency modulated waveform, or chirp waveform, in Figure 3-20(c) will have a reflection that will also be chirped, and the difference between the frequency being transmitted and that received indicates the range of the target, provided that the target is not

---

[23] Circulators are considered in Section 7.15 on Page 406. For the circulator shown, power entering Port 1 of a circulator leaves at Port 2 and power entering Port 2 is delivered to Port 3. So Ports 1 and 3 are isolated.
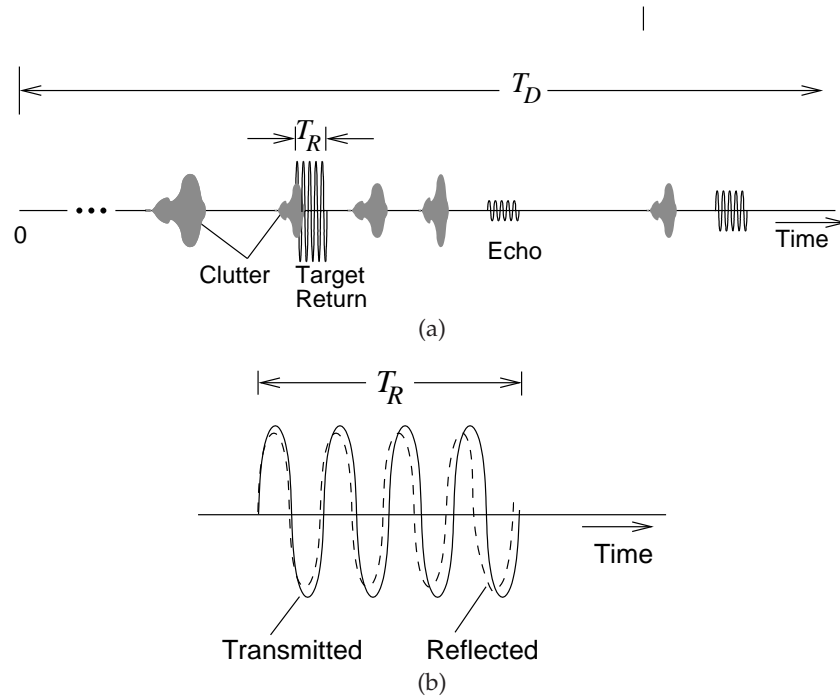
Figure 3-22 Radar returns: (a) pulsed radar; and (b) CW Doppler radar.

moving. The radar architecture that can be used to extract this information is shown in Figure 3-21(b). The directional couplers[24] tap off a small part of the transmit signal and use it as the LO of a mixer with the received signal as input. A similar architecture is shown in Figure 3-21(c), but now separate transmit and receive antennas are used to separate the transmitted and received signals rather than using a circulator. Better separation of the transmitted and received signals can be obtained this way. The IF that results is proportional to the range of the target.

The power of the signal reflected by the target and captured by the receive antenna is given by the **radar equation**:

$$P_R = \frac{P_T G_T A_R \sigma F^4}{(4\pi)^2 R_T^2 R_R^2} \ , \tag{3.23}$$

where P_T is the transmit power delivered to the transmit antenna, G_T is the antenna gain of the transmitter, A_R is the affective aperture area of the receive antenna, σ is the radar cross section of the target, F is the pattern propagation factor, R_T is the distance from the transmitter to the target, and

---

[24] Directional couplers are discussed in Section 9.7 on Page 483.

Table 3-15 Doppler frequency shifts for targets moving toward a radar at speed v_R.

Radar frequency, f_T	Relative speed, v_R		
	1 m/s	100 km/hr	1000 km/hr
500 MHz	3.3 Hz	92.6 Hz	925.9 kHz
2 GHz	13.3 Hz	370.4 Hz	3.704 kHz
10 GHz	66.7 Hz	1.852 kHz	18.519 kHz
40 GHz	266.7 Hz	7.407 kHz	74.074 kHz

R_R is the distance from the target to the receiver. F captures the effective loss due to multipath, which was captured in communications by introducing a signal dependence of $1/d^n$, where n ranges from 2 to 4 depending on the environment. In free space, $F = 1$. If the same antenna is used for transmit and receive, and if multipath is not important (so $F = 1$), then Equation (3.23) becomes

$$P_R = \frac{P_T G_T A_R \sigma}{(4\pi)^2 R^4} \,, \tag{3.24}$$

where $R = R_T = R_R$. In Section 2.5.4 on Page 79 the antenna effective area is related to the antenna gain. From Equation (2.25),

$$A_R = \frac{G_R \lambda^2}{4\pi} \,, \tag{3.25}$$

and so, with $G_R = G_T$, Equation (3.24) becomes

$$P_R = \frac{P_T G_T^2 \lambda^2 \sigma}{(4\pi)^3 R^4} \,. \tag{3.26}$$

If a signal of frequency f_T is transmitted, then the received signal will be at frequency

$$f_R = f_T + f_D \quad \text{(target moving toward the radar)} \tag{3.27}$$

$$f_R = f_T - f_D \quad \text{(target moving away from the radar)}, \tag{3.28}$$

where f_D is the **Doppler shift**:

$$f_D = \frac{2v_R f_T}{c} \,, \tag{3.29}$$

where v_R is the radial component of the speed of the object relative to the radar and c is the speed of light. Typical Doppler shifts are shown in Table 3-15.

If the target is moving, then there will be a Doppler shift. If the target is moving toward a CW radar, then the frequency of the returned signal will be higher, as shown in Figure 3-22(b). A similar architecture to that

used with chirp radar can be used (see Figures 3-21(b) and 3-21(c)). The concept can be extended to bistatic radars, but now the LO reference must be generated. As was seen with digitally modulated signals, the frequency of the transmit carrier can be recovered for digitally modulated signals such the PSK-encoded signals shown in Figure 3-20(d). Advanced high-end radars use digital modulation and CDMA-like waveforms and exploit space-time coding. Typically the radar waveforms to be transmitted are chirped, which is a technique that takes the desired transmit waveform and stretches it out in time so that it can be more efficiently amplified and more power can be transmitted. At the receiver, the radar signal is compressed in time so that it corresponds to the original transmit signal prior to chirping.

It should be apparent that radars and radar waveforms can be optimized for imaging or exploiting Doppler shifts to track moving targets. Imaging is suitable when there is little clutter, such as looking into the air. However, it is difficult to detect targets such as cars that are moving on the ground. So-called **ground moving target indication radar** (**GMTI**) relies on Doppler shifts to discriminate moving targets, and then the ability to image accurately is compromised. Considerable effort is devoted to developing waveforms that are difficult to detect (stealthy) and are optimized for either imaging or GMTI.

Another important radar variant is **synthetic aperture radar** (**SAR**), which utilizes the movement of the radar platform, generally an aircraft, and is used in remote sensing and mapping. SAR yields static imagery and DSP produces a very narrow effective beam derived from the movement of the platform.

3.13   Radiometer System

A radiometer measures the power in EM radiation predominantly at microwave frequencies and most commonly measures noise. Radiometers are used in remote sensing and radio astronomy, especially by satellites and aircraft. Understanding the physical process that creates uncorrelated radiation enables vegetation, air and sea temperatures, ice coverage, ocean salinity, and other surface and atmospheric sources to be identified from the spectrum captured by a radiometer. Radiometers monitor discrete windows of the spectrum, particularly at frequencies corresponding to molecular resonances. A radiometer includes a mechanism for rapid calibration, quickly switching between the object being observed and another object serving as a calibrated noise source. In aircraft and on land, the calibration source is a resistor, often held at low temperatures, and switching uses what is called a **Dicke switch**. A satellite-based radiometer is shown in Figure 3-23, and instead of a Dicke switch, the calibration is obtained by switching the antenna beam between the observed region (beam A) and an empty area of space (beam B). With a Dicke switch or antenna beam switching,
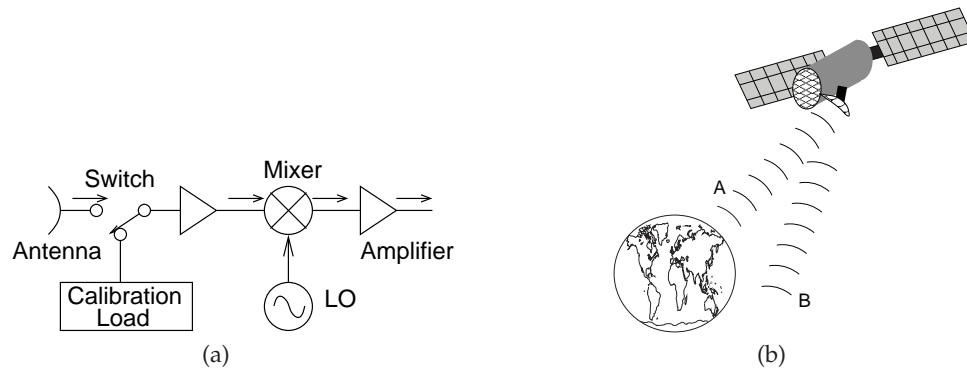
Figure 3-23 Radiometer: (a) heterodyne architecture showing calibration switch; and (b) satellite radiometer that switches between two beams, one oriented at the region being monitored and the other the much colder space which serves as calibration.

it is possible to achieve 0.1 K of stability.

3.14   Summary

Radio frequency and microwave design continues to rapidly evolve, responding to new communication, radar, and sensor architectures and exploiting opportunities made available with monolithic integration. An advanced integrated comunications system incorporates several important trends and desired attributes:

1. Direct digital synthesis (DDS):   This requires the use of fast digital-to-analog and analog-to-digital converters to directly produce and sample RF signals, with minimum mixers eliminated entirely. For systems operating in the low gigahertz region, it can be expected that mixers are eliminated entirely. For millimeter wave communications, the number of mixers would be reduced to one.

2. Digitization: The concentration of functionality in the digital domain. The availability of enormous computing complexity with low DC power requirements facilitates novel approaches to low-power wireless design. Improved performance relative to the traditional analog circuit implementation can be accomplished using digital signal processing techniques.

3. Reprogrammable transceiver module: The combination of the above functions leads to a reprogrammable universal transmit/receive (T/R) module. This is a wireless solution that can adapt to changes in communication standards and frequency bands through changes in the DSP program.

4. Linearization: Digital predistortion technique based in DSP. An optimum predistortion function will be developed by macromodeling the nonlinear transmit path.

5. Adaptivity: Adaptable, switchable, and tunable RF and microwave components are becoming available. Components becoming available are tunable capacitors based on micromachining technology (MEMS), ferroelectrics, and MEMS switches. Tunable subsystem components include tunable filters, delay lines, and phase shifters.

Radio frequency systems will continue their rapid pace of evolution.

3.15 Exercises

1. Write your name in Morse code (see Table 3-1).

2. A new communication system is being investigated for sending data to a printer. The system will use GMSK modulation and a channel with 25 MHz bandwidth and a bit rate of 10 Mbps. The modulation format will result in a spectrum that distributes power almost uniformly over the 25 MHz bandwidth. [Parallels Example 3.1 on Page 122]

   (a) What is the processing gain?
   (b) If the received RF SIR is 6 dB, what is the effective system SIR (or E_b/N_o) after DSP. Express your answer in decibels.
   (c) What is the system BER prior to the use of error correction coding? Use Figure 1-15(b) noting that GMSK is a type of FSK modulation.

3. Research at Bell Labs in the 1960s showed that the minimum acceptable SIR that was comfortable in voice communications is 17 dB. This applies to analog modulated signals, but not digitally modulated signals, where BER is important. Consider a seven-cell cluster. If the power falls off as $1/d^3$, where d is distance (this corresponds to a suburban environment), determine the worst possible SIR considering only interference from other radios. The worst situation will be when a mobile handset is at the edge of its cell. To do this you need to estimate the distance from the handset to the other base stations (in neighboring clusters) that are operating at the same power levels.

Consider the cells to be hexagons. Develop a symbolic expression for the total interference signal level at the handset, assuming that all base stations are radiating at the same power level, P. You can be reasonably approximate in determining distances. For example, each distance can be expressed in terms of integer multiples of cell radii, R. Is the 17 dB SIR achieved with analog radio?

4. Short answer questions. Each part requires a short paragraph of about five lines and a figure, where appropriate, to illustrate your understanding.

   (a) Cellular communications systems use two frequency bands to communicate between the base station and the mobile unit. The bands are generally separated by 50 MHz or so. Which band (higher or lower) is used for the downlink from the base station to the mobile unit and what are the reasons behind this choice?

   (b) Describe at least two types of interference in a cellular system from the perspective of a mobile handset.

   (c) The three main cellular communication bands are centered around 450 MHz, 900 MHz, and 2 GHz. Compare these three bands in terms of multipath effects, diffraction around buildings, object (such as a wall) penetration, scattering from trees and parts of trees, and the ability to follow the curvature of hills. You can use a table and indicate the relative attributes: HIGH, MEDIUM, and LOW.

5. An antenna with a gain of 10 dB presents a signal with a power of 5 dBm to a low-noise amplifier along with noise of 1 mW and an interfering signal of 2 mW.

   (a) What is the SIR? Include both noise and the interfering signal in your calculation. Express your answer in decibels.

   (b) The modulation format and coding scheme used have a processing gain, G_P, of 7 dB. The modulation scheme has four states. What is the ratio of the energy per bit to the noise per bit, that is, what is the effective E_b/N_o?

6. Describe the following concepts and, if applicable, the reason behind their usage.

   (a) Clusters in a cellular phone system.

   (b) Multipath effects in a central city area compared to multipath effects in a desert.

   (c) There are three major cell phone frequency bands centered at 450 MHz, 900 MHz, and 1900 MHz. Contrast operation of a cell phone system at these different frequency bands in terms of performance, number of users supported, cell size, and signal propagation.

7. The power of a received signal is 1 pW and the received noise power is 200 fW. In addition the level of the interfering signals is 100 fW. What is the signal to noise ratio? Treat interference as if it was an additional noise signal.

8. A 4 kHz bandwidth voise signal is coded as an 8 kbps data stream. Coding increases the data stream to 64 kbps. What is the processing gain that can be achieved at the receiver?

9. If a received signal as an SIR of -5 dB and the processing gain that can be achieved for the

moudlation and coding used is 15 dB, what is the E_b/N_0 after processing?

10. Short answer questions on modulation and spectral efficiency.

   (a) What is the PAR of a phase modulated signal?

   (b) In less than five lines explain your understanding of spectral efficiency as it relates to bits per hertz. That is, how can you have a spectral efficiency of n bits/Hz where n is more than 1? [Note that sometimes this is expressed as bits/sec/Hz as well as bits/Hz.]

   (c) What is the spectral efficiency of a QPSK modulated signal? Ignore the impact of the number of cells in a cluster.

   (d) A proposed modulation format has a spectral efficiency of 3.5 bits/sec/Hz. Antenna sectoring and required SNR lead to a system with seven cells per cluster. You can ignore the impact of coding so you can assume that $R_b = R_c$. What is the spectral efficiency in terms of bits/sec/Hz/cell modulated signal?

11. Short answer questions. Each part requires a short paragraph of about five lines and a figure, where appropriate, to illustrate your understanding.

   (a) Explain how OFDM reduces the impact of multiple paths in a wireless communication system.

   (b) Explain how MIMO exploits multipath to enhance the capacity of a digital communication system.

Transmission Lines

4.1 Introduction

A transmission line stores electric and magnetic energy distributed in space and alternating between the two forms in time. That is, at any position along the line the energy is stored in a combination of electric and magnetic forms and, for an alternating signal at any position on the line, converted from one form to the other as time progresses. As such, a transmission line has a circuit form that combines inductors, Ls (for the magnetic energy), capacitors, Cs (for the electric energy), and resistors, Rs (modeling losses), whose values are dependent on the geometry of the line and the properties of the materials comprising the line. Thus transmission lines of various lengths and crosssections mimic circuits. **Distributed structures**,
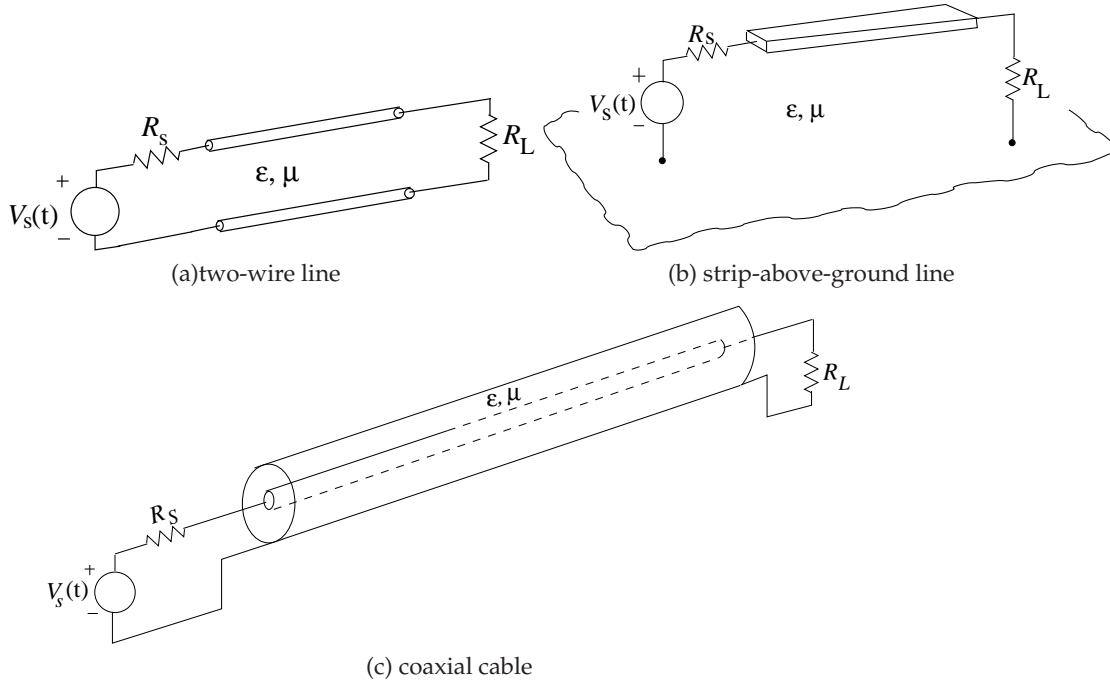
Figure 4-1   Several homogeneous transmission line structures: (a) two-wire line; (b) strip-above-ground plane; and (c) coaxial cable.

of which transmission lines are the most fundamental members, are what distinguishes RF, microwave, packaging, and high-speed digital design from lumped-element (R, L, and C) circuit design. In this chapter the properties of transmission lines are considered. It will be seen how they can be modeled using lumped elements and it will be seen how simple lumped-element circuits can be realized using combinations of transmission lines, and how transmission lines can be used to achieve surprising functionality beyond that which can be achieved with lumped-element circuits. The discussion of transmission lines introduces concepts that apply to all distributed structures.

   The transmission lines considered here are systems of two or more closely spaced parallel conductors.[1] For now, the discussion is restricted to considering just two parallel conductors, as shown in Figure 4-1(a), with the distance between the two wires being substantially smaller than the wavelengths of the signals on the line. Then the structure may be satisfactorily analyzed on the basis of voltages and currents. As the frequency increases, and therefore the wavelength becomes smaller, and the

---

[1]   Rectangular waveguide is a transmission line that has just one conductor. Dielectric lines can have no conductor.

crosssectional dimensions become electrically larger, it becomes necessary to set up a complete EM field solution in order to analyze the structures. By considering only transmission lines where electrically small transverse line dimensions[2] prevail, a number of useful results are obtained on a voltage and current basis. The theory thus developed is called distributed circuit theory.

Metallic conductors forming a transmission line are **crystals** with positively charged **ions** locked into position in a regular lattice. Each ion comprises a nucleus of an atom which is positively charged and a complement of electrons local to each atom which almost, but not quite, balances the positive charge locally. In a metallic crystal there are free **electrons** shared by several ions, with the overall effect that the positive and negative charges are balanced and the free electrons can wander around the lattice. The movement of charge results in a **magnetic field**, and hence **magnetic energy** storage. The ability of a structure to store magnetic energy is described by its inductance, L. Similarly the rearrangement of charge to produce localized net positive or negative charge on one conductor of the pair is matched by an opposing charge on the adjacent conductor. This separation of charge results in an **electric field**, and thus **electric energy** storage, with the capacitance, C, indicating the amount of energy that can be stored. The ratio of the energy stored in the magnetic and electric forms is proportional to L/C and the rate at which the energy can be moved is proportional to $1/\sqrt{LC}$; together these determine the characteristics of the transmission line. A lossless transmission line is generally characterized by its **characteristic impedance**:

$$Z_0 = \sqrt{L/C}\,, \tag{4.1}$$

with the units of ohms (Ω) and its propagation constant

$$\gamma = \sqrt{LC}\,, \tag{4.2}$$

which in the SI system has the units of inverse meters (m^{-1}). Another way of looking at transmission lines is that they confine and guide an EM field between them. In fact, transmission lines are also, but less commonly, called waveguides. Many structures can successfully guide waves, and two additional ones are shown in Figures 4-1(b) and 4-1(c). The "strip-above-ground" transmission line shown in Figure 4-1(b) confines the EM field mostly between the flat metal strip and the metallic ground plane below it. However, some energy is distributed above and to the sides of the lines so that the EM energy is not completely confined. As long as the distance between the strip and the ground plane is less than half a wavelength of the signal on the line, the energy will follow the strip along its length and

---

[2]   Say, less than 1/20 of a wavelength.

around bends, etc. When the separation is more than one-half wavelength EM energy will radiate away from the line. The coaxial line shown in Figure 4-1(c) completely confines the field between the inner conductor and the outer conductor. Now, if the spacing between the conductors is less than one-half wavelength the fields will, in general, be coherently guided in what is called a single mode. With some structures, such as the strip-above-ground line of Figure 4-1(b), there is an additional criterion that the width of the strip be less than $\lambda/2$.

In low-frequency analog and digital circuits, transmission lines are often referred to as **interconnects** and can be viewed simply as wires, and provided that the wire has sufficiently low resistance, the interconnect can be largely ignored. However, if transmission must be over a nonnegligible distance compared to a **wavelength** (λ), then the interconnect must be considered as part of the circuit.

The earliest fundamental understanding of signal transmission led to telegraphy over distances. The critical theoretical step that enabled transmission over more than short distances to be achieved was the development of an understanding of signal transmission on lines using what is now called **phasor analysis**.[3] The mathematician and engineer Oliver Heaviside [50] developed the frequency-domain-based treatment of signal propagation on transmission lines. Frequency-domain analysis (i.e., using phasors) is still the best way to develop a fundamental understanding of transmission lines, even if they are used for purely digital signals.

The key determinant of whether a transmission line can be considered as an invisible connection is whether the signal anywhere along the interconnect has the same value at a particular instant. If the value of the signal (say, voltage) varies along the line (at an instant), then it may be necessary to consider transmission line effects. A typical criterion used is that if the length of the interconnect is less than $1/20$ of the wavelength of the highest-frequency component of a signal, then transmission line effects can be safely ignored and the circuit can be modeled as a single RLC circuit [51]. The actual threshold used—$\lambda/20$, $\lambda/10$, or $\lambda/5$—is based on experience and the particular application. For example, an interconnect on a silicon chip clocking at 4 GHz has an appreciable frequency component at 20 GHz. Then the interconnect reaches the $\lambda/10$ threshold when it is 4.5 mm long. This is less than the dimensions of most chips, which can be up to 2 cm on a side. Thus it takes a finite time for the variation of a voltage at one end of an interconnect to impact the voltage at the other end. The ultimate limit

---

[3] A phasor is a complex number that combines the amplitude and phase of a sinewave. With transmission line equations, the introduction of phasors eliminates time dependence from the equations and the dimensionality of the equations is reduced by one. Also, in circuit analysis, when phasors are introduced, differential equations in time become algebraic equations, which are much easier to handle. When phasors are used, no information on frequency is retained.

is determined by the speed of light, c, but this is reduced by the relative permittivity and permeability of the material in which the fields exist. The relative permittivity and permeability describe the effect of excess potential energy storage in the material in a manner that is analogous to storing mechanical energy in a spring.

4.1.1 Movement of a Signal on a Transmission Line

A **coaxial line** (Figure 4-2(a)) is the quintessential transmission line, as it is one of the few transmission line structures that can be described exactly from first principles[4] when there is no loss. Here a realistic coaxial line is considered with conductors having a small amount of loss, a structure that does not have an exact solution. When a positive voltage pulse is applied to the center conductor of the coaxial line, an electric field results that is essentially directed from the center conductor to the outer conductor. A much smaller component of the electric field will also be directed along the line. The direction of the electric field is the direction in which a positive charge would move if it was released into the field. The component of the field that is directed along the shortest path from the center conductor to the outer conductor (in what is called the **transverse plane**) is denoted E_T, and the component directed along the line is denoted E_L. (The subscripts T and L denote transverse and longitudinal components, respectively.) Thus, while $E_L \ll E_T$, it is necessary to accelerate electrons on the conductors and so give rise to current flow, and hence the movement of the pulse along the line. Figure 4-2(b) shows the fields in the structure after the pulse has started moving along the line. This is shown in another view in Figure 4-3. The transverse voltage, V_T, is given by E_T integrated along a path between the inner and outer conductors: $V_T \approx E_T(a - b)$. This is a good measure, provided that the transverse dimensions are sufficiently small compared to a wavelength (otherwise the integral does not come out so simply, as it is path dependent).

The pulse moves down the lossless line at the **phase velocity**, v_p.[5] It is determined by the physical properties of the region between the conductors. The permittivity, ε, describes the energy storage capability associated with the electric field, E, and the energy storage associated with magnetic field, H, is described by the permeability, μ. (Both ε and μ are properties of the

---

[4] Applying Maxwell's equations and assuming perfect conductors and using cylindrical symmetry and phasors reduces the dimensionality of the problem from four dimensions (the three spatial dimensions plus time) to two spatial dimensions. Maxwell's equations are discussed in Appendix D.

[5] Specifically, the phase velocity is the apparent velocity of a point of constant phase on a sinewave. While v_p can vary with frequency, it is almost frequency independent for a low-loss coaxial line of relatively small transverse dimensions (less than $\lambda/10$).

(a)　　　　　　　　　　　　　　　　(b)
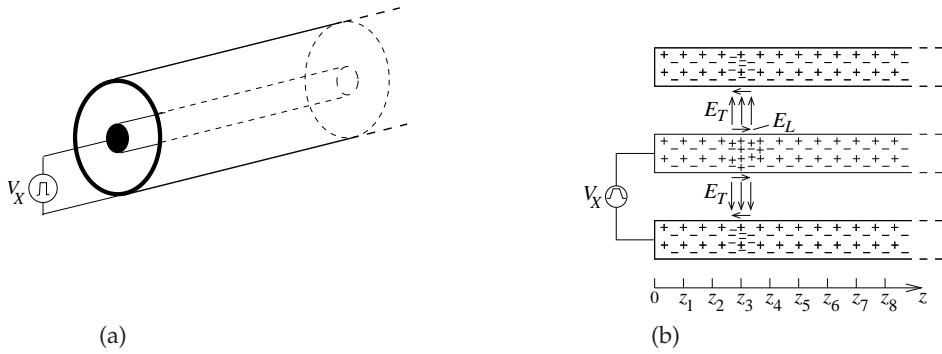
Figure 4-2　A coaxial transmission line: (a) three-dimensional view; (b) the line with pulsed voltage source showing the electric fields at an instant in time as a voltage pulse travels down the line.

Q_{CENTER} is the net free charge on the center conductor.
I_{CENTER} is the current on the center conductor.
$T_4 > T_3 > t_2 > t_1$
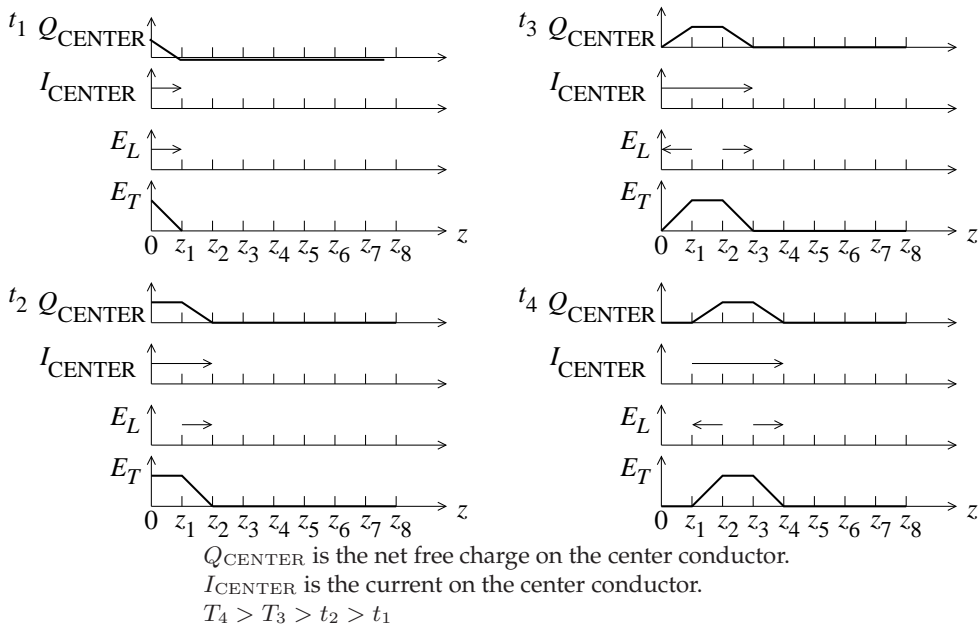
Figure 4-3　Fields, currents, and charges on the coaxial transmission line of Figure 4-2.

medium—the material.) It has been determined that[6]

$$v_p = 1/\sqrt{\mu\varepsilon} \ . \tag{4.3}$$

In a vacuum $\varepsilon = \varepsilon_0$, the free-space permittivity, and $\mu = \mu_0$, the free-space

---

[6]　This is derived from Maxwell's equations; there is no underlying theory for these equations, but they have been verified by numerous experiments.

permeability. These are physical constants and have the values:

Permittivity of free space $\quad \varepsilon_0 = 8.854 \times 10^{-12}$ F·m^{-1}
Permeability of free space $\quad \mu_0 = 4\pi \times 10^{-7}$ H·m^{-1}

One conclusion here is that EM energy can be stored in a **vacuum** (or **free space**). So in free space, or on an air-filled coaxial line, $v_p = c = 1/\sqrt{\mu_0 \varepsilon_0} = 3 \times 10^8$ m·s^{-1}. The wavelengths, $\lambda_0 = c/f$, at several different frequencies are

f	100 MHz	1 GHz	10 GHz
λ_0	3 m	30 cm	3.0 cm

These are dimensions comparable to the sizes of many circuits. Commonly λ_0 is used to indicate the wavelength in free space and λ_g, the guide wavelength, is used to denote the wavelength of a transmission line (or waveguide). A dielectric cannot have a dielectric constant less than ε_0, and so it is convenient to use the relative permittivity (or the less commonly used term dielectric constant), ε_r, defined as

$$\varepsilon_r = \varepsilon/\varepsilon_0 . \tag{4.4}$$

Similarly the relative permeability is

$$\mu_r = \mu/\mu_0 , \tag{4.5}$$

and most materials have $\mu_r = 1$. The permittivity, permeability, and conductivity of materials used in RF and microwave circuits are given in Appendix C.

EXAMPLE 4. 1   Transmission Line Wavelength

A length of coaxial line is filled with a dielectric having a relative dielectric constant of 20 and is designed to be one-quarter wavelength long at frequency, f, of 1.850 GHz.

(a) What is the free-space wavelength at 1850 MHz?

(b) What is the wavelength of the signal in the dielectric-filled coaxial line?

(c) How long is the line?

Solution:

(a) $\lambda_0 = c/f = 3 \times 10^8/1.85 \times 10^9 = 0.162$ m $= 16.2$ cm

(b) $\lambda_g = \lambda_0/\sqrt{\varepsilon_r} = 16.2$ cm$/\sqrt{20} = 3.62$ cm

(c) $\lambda_g/4 = 3.62$ cm$/4 = 9.05$ mm

4.1.2 Current and Voltage on Transmission Lines

The majority of transmission lines used in design are planar, as these can be defined using masks, photoresist, and etching of metal sheets. Such lines are called planar interconnect. A common planar interconnect is the microstrip line shown in crosssection in Figure 4-4[7]. This crosssection is typical of what would be found on a semiconductor or **printed wiring board** (**PWB**), which is also called a **printed circuit board** (**PCB**). Current flows in both the top and bottom conductor, but in opposite directions. The physics is such that if there is a signal current on the top conductor, there must be a return signal current, which will tend to be as close to the signal current as possible to minimize stored energy. The provision of a signal return path close to the signal path is important in maintaining the integrity (i.e., predictable signal waveform) of an interconnect.

In the microstrip line, electric field lines start on one of the conductors and finish on the other and are located almost entirely in the plane transverse to the long length of the line. The magnetic field is also mostly confined to the transverse plane, and so this line is referred to as a **transverse electromagnetic** (**TEM**) line. Integrating the electric field along a path gives the voltage. Since the voltage between the top and bottom conductors is more or less the same everywhere, longer E field lines correspond to lower levels of E field. The strength of the E field is also indicated by the density of the E field lines. This is a drawing convention for electric and magnetic fields. A further comment is warranted for this line. This line is more commonly called a **quasi-TEM** line, as the longitudinal fields are not as negligible as with the coaxial line considered previously. On a microstrip line, the relative level of the longitudinal fields increases with frequency, but below about 10 GHz and for typical dimensions used, the line is still essentially TEM. Figure 4-4 illustrates an important point: current flows in the strip and a return current flows in what is normally regarded as the grounded conductor. Both the signal and return currents induce a magnetic field and the closed path integral of the magnetic field is equal to the current contained within the path.

Various schematic representations of a transmission line are used. To illustrate this, consider the various representations in Figure 4-5 of a length of microstrip line shorted by a via at the end denoted by "2" (specifically the "2" refers to Port 2). The representations shown in Figures 4-5(d), 4-5(e), and 4-5(f) are commonly used in circuit diagrams. The preference depends on the number of transmission lines in a circuit diagram. If a circuit diagram has many transmission line elements, such as the filter circuits of Chapter 10, then the simple representation of Figure 4-5(f) is most common. If there are

---

[7] An entertaining account of the early application of microstrip transmission lines can be found in Barrett [52]. The original analysis of microstrip was based on the unfolding of a coaxial line.
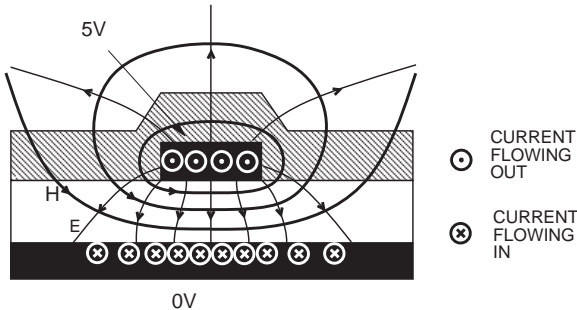
Figure 4-4 Crosssectional view of a microstrip line showing the electric and magnetic field lines and current flow for a microstrip interconnect. The electric and magnetic fields are in two media—the dielectric and air.
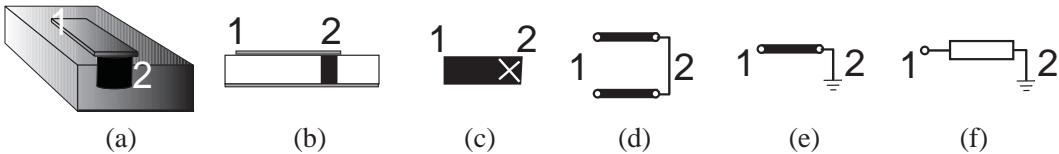
Figure 4-5 Representations of a shorted microstrip line with a short (or via) at Port 2: (a) three-dimensional (3D) view indicating via; (b) side view; (c) top view with via indicated by X; (d) schematic representation of transmission line; (e) alternative schematic representation; and (f) representation as a circuit element.

few transmission line elements, then the representation of Figure 4-5(d) is most common.

4.1.3 Forward- and Backward-Traveling Pulses

Forward- and backward-traveling pulses are shown in Figure 4-6(a), for the situation where the resistance at the end of the line is lower than the characteristic impedance of the line ($Z_L < Z_0$). The voltage source is a step voltage which is zero for time $t < 0$. At time $t = 0$, the step is applied to the line and it begins traveling down the line, as shown at time $t = 1$. This voltage step moving from left to right is called the forward-traveling voltage wave. At time $t = 2$, the leading edge of the step reaches the load, and as the load has lower resistance than the characteristic impedance of the line, the total voltage across the load drops below the level of the forward-traveling voltage step. The reflected wave is called the backward-traveling wave and it must be negative, as it adds to the forward-traveling wave to yield the total voltage. Thus the voltage reflection coefficient, Γ, is negative and the total voltage on the line, which is all we can directly observe, drops. A reflected, smaller, and opposite step signal travels in the backward direction and adds to the forward-traveling step to produce the waveform shown at $t = 3$. The
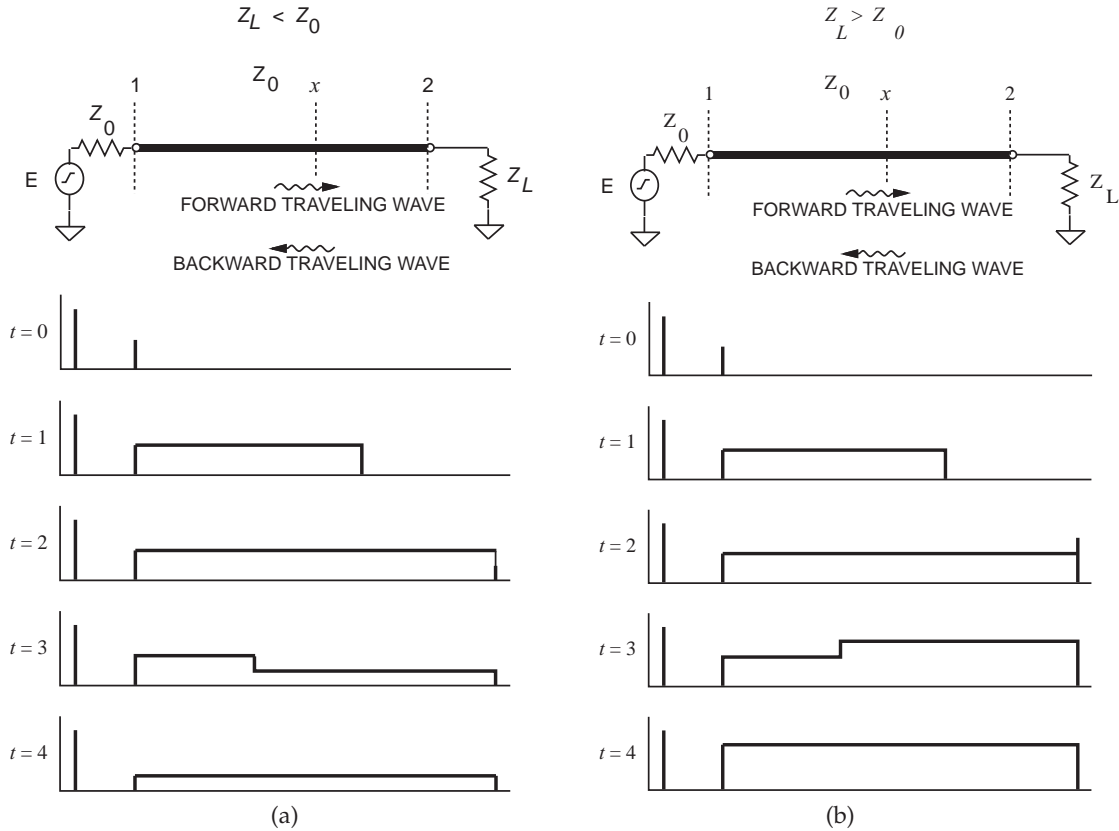
Figure 4-6   Reflection of a pulse from a resistive load: (a) when the resistance of the load is lower than the characteristic impedance of the line; and (b) when the resistance of the load is higher than the characteristic impedance of the line.

impedance of the source is matched to the transmission line impedance so that the reflection at the source is zero. The signal on the line at time $t = 4$, the round-trip propagation time of the line, therefore remains at the lower value. The easiest way to remember the polarity of the reflected pulse is to consider the situation with a short-circuit at the load. Then the total voltage on the line at the load end must be zero. The only way this can occur when a signal is incident is if the reflected signal is equal in magnitude but *opposite* in sign, in this case $\Gamma = -1$. So whenever $|Z_L| < |Z_0|$, the reflected pulse will tend to subtract from the incident pulse. You will note that in Figure 4-6(a) a schematic symbol is used for the line. Even though it appears that just one conductor is shown, this symbol represents the full transmission line with the signal path and the signal return path.

The opposite situation occurs when the resistance at the load end is higher

than the characteristic impedance of the line (Figure 4-6(b)). In this case the reflected pulse has the same polarity as the incident signal. Again, to remember this, think of the open-circuited case. The voltage across the load does not need to be zero, and indeed doubles, as the reflected pulse has the same sign as well as magnitude as that of the incident signal, in this case $\Gamma = +1$.

A more illustrative situation is shown in Figure 4-7, where a more complicated signal is incident on a load that has a resistance higher than that of the characteristic impedance of the line. The peaking of the voltage that results at the load is typically the design objective in many long digital interconnects, as less overall signal energy needs to be transmitted down the line, or equivalently a lower current drive capability of the source is required to achieve first incidence switching. This is at the price of having reflected signals on the interconnects, but these can be dissipated through a combination of the interconnect loss and absorption of the reflected signal at the driver.

4.2 Media

Design involves choosing the transmission line structure to use and the substrate. In this section, the electrical properties of materials will be discussed and then substrates commonly used with planar interconnects will be considered.

4.2.1 Dielectric Effect

The presence of material between the conductors alters the electrical characteristics of the interconnect. With a dielectric, the application of an electric field moves the centers of positive and negative charge at the atomic and molecular level. Moving the charge centers changes the amount of energy stored in the electric field—a process akin to storing energy in a stretched spring. The extra energy storage property is described by the **relative permittivity**, ϵ_r, which is the ratio of the permittivity of the material to that of free space:

$$\varepsilon = \varepsilon_r \varepsilon_0 \ . \tag{4.6}$$

The relative permittivities of materials commonly used with interconnects range from 2.08 for Teflon$^{\text{TM}}$, used in high-performance PCBs and coaxial cables, to 11.9 for silicon (Si), to 3.8–4.2 for silicon dioxide (SiO$_2$), and 12.4 for gallium arsenide (GaAs). Values of permittivity for other materials are given in Table 4-1.

When the fields are in more than one medium (a **nonhomogeneous transmission line**), as for the microstrip line shown in Figure 4-4, the effective relative permittivity, ε_{eff}, is used. The characteristics of the nonhomogeneous line are then more or less the same as for the same
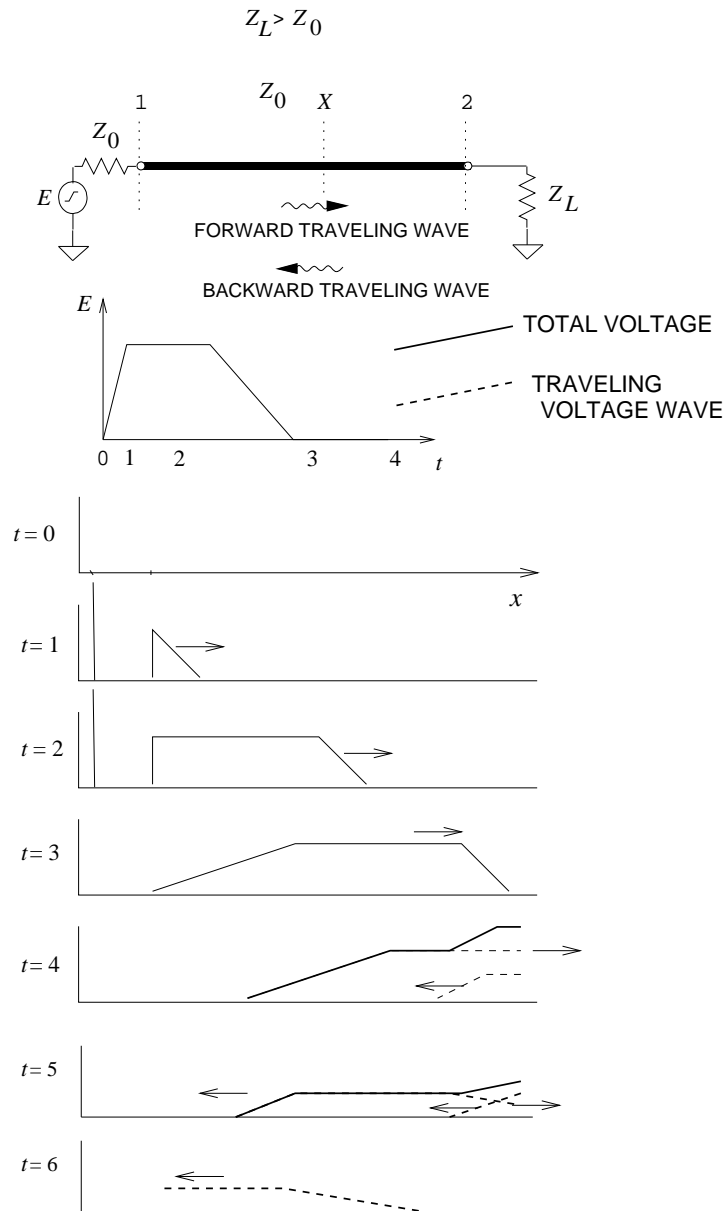
Figure 4-7 Reflection of a pulse on an interconnect showing forward- and backward-traveling pulses.

structure with a uniform dielectric of permittivity, ε_{eff}. The ε_{eff} changes with frequency as the proportion of energy stored in the different regions changes. This effect is called **dispersion** and causes a pulse to spread out as the different frequency components of the pulse travel at different speeds. Dispersion is elaborated on more in Section 5.2.5 on Page 252.

4.2.2 Dielectric Loss Tangent $\tan \delta$

Loss in a dielectric comes from two sources: (a) **dielectric damping**, and (b) conduction losses in the dielectric. Dielectric damping originates from the movement of charge centers resulting in mechanical distortion of the dielectric lattice. With an alternating electric field this results in vibrational (or phonon) energy in the dielectric, thus energy is lost from the electric field. It is easy to see that this loss will increase linearly with frequency and be zero at DC. Because of this, loss is described by incorporating an imaginary term in the **permittivity**. So now the permittivity becomes

$$\varepsilon = \varepsilon' - \jmath \varepsilon'' = \varepsilon_0 \left(\varepsilon'_r - \jmath \varepsilon''_r \right). \tag{4.7}$$

If there is no dielectric damping loss, $\varepsilon'' = 0$. The other type of loss that occurs in the dielectric is due to the movement of charge carriers in the dielectric. The ability to move charges is described by the conductivity, σ, and the loss due to current conduction is independent of frequency. So the energy lost in the dielectric is proportional to $\omega \varepsilon'' + \sigma$ and the energy stored in the electric field is proportional to $\omega \varepsilon'$. Thus a loss tangent is introduced:

$$\tan \delta = \frac{\omega \varepsilon'' + \sigma}{\omega \varepsilon'} \ . \tag{4.8}$$

Alternatively an effective relative permittivity, $\varepsilon_{r,e}$, can be defined which is the complex permittivity that would be measured:

$$\varepsilon_{r,e} = \varepsilon'_{r,e} - \jmath \varepsilon''_{r,e} = \varepsilon'_r - \jmath \left(\varepsilon''_r + \frac{\sigma}{\omega} \right). \tag{4.9}$$

The loss tangent is very small for dielectrics that are useful at RF and microwave frequencies and so

$$|\varepsilon| \approx \varepsilon' \ , \tag{4.10}$$

and this is what is quoted as the permittivity of materials. Thus

$$\begin{aligned} \varepsilon &= \varepsilon' - \jmath \varepsilon'' \\ &\approx \varepsilon' \left(1 - \jmath \tan \delta \right). \end{aligned} \tag{4.11}$$

4.2.3 Magnetic Material Effect

A similar effect on energy storage in the magnetic field occurs for a few materials. The magnetic properties of materials are due to the magnetic

dipole moments resulting from alignment of electron spins—an intrinsic property of electrons. In most materials the electron spins occur in pairs with opposite signs, with the result that there is no net **magnetic moment**. However, in magnetic materials, some of the electron spins are not canceled by balancing spins and there is a net magnetic moment. This net magnetic moment aligns itself with an applied H field and so provides a mechanism for additional storage of **magnetic energy**. The relative permeability, μ_r, describes this effect and

$$\mu = \mu_r \mu_0 . \tag{4.12}$$

Most materials have $\mu_r = 1$. One notable exception is nickel, which has a high permeability, is a very convenient processing material, and is often used in electronic packaging for its desirable processing properties.

As with dielectrics, the effect of loss in a magnetic material can be described by its complex permeability:

$$\mu = \mu' - \jmath\mu''. \tag{4.13}$$

Lossy magnetic effects are due to the movement of magnetic dipoles which again creates vibrations in the material and loss.

4.3 Substrates

The properties of a number of typical substrate materials are given in Table 4-1. We will consider just one family of substrates, the ceramics, to illustrate some of the considerations involved in substrate choice. Alumina is a typical ceramic well suited for production of circuits functioning at frequencies up to about 60 GHz.

4.4 Transmission Line Structures

There are two major categories of transmission lines sorted according to the uniformity of the medium surrounding the transmission line conductors. When the embedding medium is uniform the transmission line structures are referred to as homogeneous. Several important homogeneous transmission line structures were introduced in Figure 4-1 and more are shown in Figures 4-8(a) and 4-8(b).

In these homogeneous lines virtually all of the fields are in the plane transverse to the direction of propagation (i.e., the longitudinal direction) along the line, that is, $E_T \gg E_L$ and $H_T \gg H_L$, where the subscripts T and L refer to the transverse and longitudinal directions respectively. We refer to transmission lines where the longitudinal fields are almost insignificant as supporting a TEM mode, and they are called TEM lines.

The most important inhomogeneous lines are shown in Figures 4-8(c), 4-8(d), and 4-8(e). The main difference between the two sets of configurations (homogeneous and inhomogeneous) is the frequency-dependent variation of the EM field distributions with inhomogeneous

Table 4-1 Properties of common substrate materials. The dielectric loss tangent is scaled. For example, for glass, $\tan \delta$ is typically 0.002.

Material	$10^4 \tan \delta$ (at 10 GHz)	ε_r
Air (dry)	≈ 0	1
Alumina, 99.5%	1–2	10.1
Sapphire	0.4–0.7	9.4, 11.6
Glass, typical	20	5
Polyimide	50	3.2
Quartz (fused)	1	3.8
FR4 circuit board	100	4.3–4.5
RT-duroid 5880	5–15	2.16–2.24
RT-duroid 6010	10–60	10.2–10.7
AT-1000	20	10.0–13.0
Si (high resistivity)	10–100	11.9
GaAs	6	12.85
InP	10	12.4
SiO$_2$ (on-chip)	—	4.0–4.2
LTCC (typical, green tape(TM) 951)	15	7.8

lines. With inhomogeneous lines, the EM fields are not confined entirely to the transverse plane even if the conductors are perfect. Even so, $E_T \gg E_L$, and so these lines are called **quasi-TEM lines**. The inhomogeneous lines are simpler to make. Each transmission line structure comprises a combination of metal, shown as dense black, and dielectric, indicated by the shaded region and having permittivity ε. It is common not to separately designate the permeability, μ, of the materials as, except for magnetic materials, $\mu = \mu_0$, the free-space permeability. The region with permittivity ε_0 is free space or air. In most cases the dielectric principally supports the metal pattern, acting as a substrate, and clearly influences the wave propagation. The actual choice of structure depends on several factors, including operating frequency and the type of substrate and metallization system available.

4.4.1 Stripline

Stripline is a symmetrical structure somewhat like a coaxial line completely flattened out so that the center conductor is a rectangular metal strip and the outer grounded metal is simply a rectangular box. The entire structure is 100% filled with dielectric, and therefore transmission is TEM and dependent upon the relative permittivity, ε_r, explicitly. Therefore the wavelength is simply the free-space value divided by the square root of ε_r.

HOMOGENEOUS   **INHOMOGENEOUS**

(a) Stripline   (c) Microstrip

(b) Embedded differential line   (d) Coplanar waveguide (CPW)

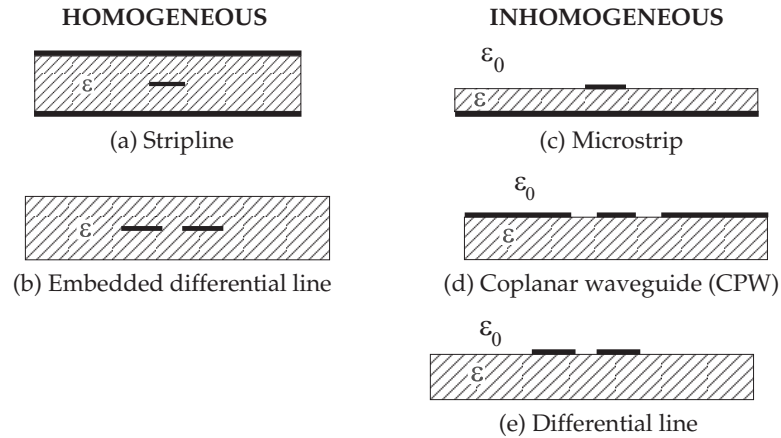   (e) Differential line

Figure 4-8 Crosssections of several homogeneous and inhomogeneous transmission line structures. Homogeneous lines: (a) stripline; and (b) embedded differential line. Inhomogeneous lines: (c) microstrip; (d) CPW or coplanar waveguide; and (e) differential line.

4.4.2 Embedded Differential Line

This simple transmission structure is formed by having just two conductors embedded in a substrate with no specific ground plane. In this structure the possible existence of ground planes is incidental, and ideally these should not influence the field pattern. Essentially the substrate merely acts as a mechanical supporting element and a quasi-TEM mode forms the main propagating field distribution.

4.4.3 Microstrip

Although microstrip (Figure 4-8(d)) has a very simple geometric structure the EM fields involved are actually complex. However, simple approaches to the quasi-TEM mode calculations combined with frequency-dependent expressions yield quite acceptable design accuracy for many applications. **Microwave Integrated Circuits**[8] (**MICs**) using microstrip can be designed for frequencies ranging from a few gigahertz, or even lower, up to at least many tens of gigahertz. At higher frequencies, particularly into the millimeter wavelength ranges (the wavelength at 30 GHz is 1 mm), losses (including radiation) increase greatly, higher-order modes[9] become

---

[8] Circuits integrated on compound semiconductor substrates are referred to as **Microwave Monolithic Integrated Circuits** (**MMICs**). For greater distinction, MICs are sometimes referred to as hybrid MICs.

[9] Avoiding higher-order modes is a key aspect of transmission line design. Higher order modes and design strategies to avoid exciting them are discussed in Section 5.6 on Page

a considerable problem, and fabrication tolerances become exceedingly difficult to meet using MICs. It is probable that the frequency limit for the extensive use of microstrip is in the region of 110 GHz. With monolithic ICs, fabrication tolerances are much finer than with hybrid MICs and the options available for both microstrip and other transmission structures are extended considerably.

4.4.4 Coplanar Waveguide (CPW)

Coplanar Waveguide (**CPW**)[10] supports a quasi-TEM mode of propagation with the active metallization and the ground planes on the same side of the substrate. Each 'side-plane" conductor is grounded and the center strip carries the signal; thus much less field enters the substrate when compared with microstrip. In conventional CPW the ground planes extend indefinitely, but in **finite ground CPW** (**FGCPW**), the extent of the grounds is limited and this results in reduced coupling of adjacent and crossing CPW lines. It is important to connect the ground strips every tenth of a wavelength or so. This is done using wire bonds, via structures, or in integrated circuit form using air bridges. CPW does a good job of suppressing radiation and is preferred to microstrip at frequencies above 10 GHz or so. It does have drawbacks, including the increased area required (compared to microstrip) and the need to use ground straps.

4.4.5 Coplanar Strip (CPS) and Differential Line

This simple transmission structure is formed by two conductors in the same plane. As with the embedded differential line, the possible existence of ground planes is incidental and ideally these should not influence the field pattern. This structure is formed on the surface of a dielectric substrate, as shown in Figure 4-8(e). In one realization, one of the conductors is grounded, and this form is called coplanar stripline or **coplanar strip** (**CPS**) [54–64]. In this configuration, CPS is used as an area-efficient variation of CPW. When neither of the conductors is grounded and the line is driven differentially, the interconnect is called a differential line. A differential line is used extensively with RFICs and in critical nets in high-speed digital ICs. The two forms have essentially identical electrical characteristics, with differences resulting from interaction with other metallic structures such as ground planes.

Silicon-based RFIC chips generally use differential signaling for analog signal to overcome the limitations of metal-oxide semiconductor (MOS) devices and the finite conductivity of the silicon substrate that results in high levels of common-mode noise. The currents on each of the differential signal

---

267.

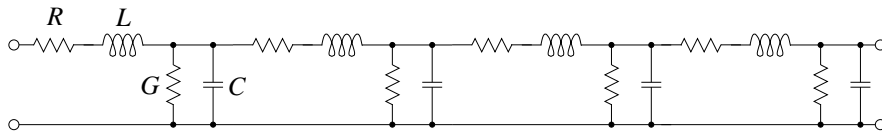[10] CPW was invented in 1969 by Cheng P. Wen [53].

Figure 4-9 $RLGC$ model of an interconnect.

paths balance each other and so each provides the signal return path for the other. This design practice effectively eliminates RF currents that would occur on ground conductors. The off-chip RF interconnects interfacing the chip to the outside world also require differential signaling, but now, because of the larger electrical dimensions, differential transmission lines are required.

4.5   Modeling of Transmission Lines

Describing the signal on a line in terms of E and H requires a description of the E and H field distributions in the transverse plane. It is fortunate that current and voltage descriptions can be successfully used to describe the state of a circuit at a particular position along a TEM or quasi-TEM line. This is an approximation and the designer needs to be aware of situations where this breaks down. Such extraordinary effects are left to the next chapter. Once the problem of transmission line descriptions has been simplified to current and voltage, R, L, and C models of a transmission line can be developed. A range of models are used for transmission lines depending on the accuracy required and the frequency of operation.

Uniform interconnects (with regular crosssection) can be modeled by determining the characteristics of the transmission line (e.g., Z_0 and γ versus frequency) or arriving at a distributed lumped-element circuit, as shown in Figure 4-9. Typically EM modeling software models planar interconnects as having zero thickness, as shown in Figure 4-10. This is reasonable for microwave interconnects as the thickness of a planar strip is usually much less than the width of the interconnect. Many analytic formulas have also been derived for the characteristics of uniform interconnects. These formulas are important in arriving at synthesis formulas that can be used in design (i.e., arriving at the physical dimensions of an interconnect structure from its required electrical specifications). Just as importantly, they provide insight into the effects of materials and geometry.

Simplification of the geometry of the type illustrated in Figure 4-10 for microstrip can lead to appreciable errors in some situations. More elaborate computer programs that capture the true geometry must still simplify the real situation. An example is that it is not possible to account for density variations of the dielectric. Consequently characterization of many RF and microwave structures requires measurements to "calibrate" simulations.
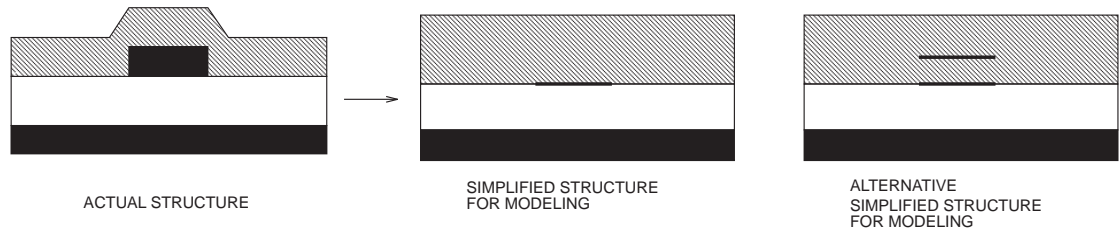
Figure 4-10 A microstrip line modeled as an idealized zero-thickness microstrip line. Also shown is an alternative simplified structure that can be used with some EM analysis programs as a more accurate approximation. (Crosssectional view.)

Unfortunately it is also difficult to make measurements at microwave frequencies. Thus one of the paradigms in RF circuit engineering is to require measurements and simulations to develop self-consistent models of transmission lines and distributed elements.

4.6 Transmission Line Theory

Regardless of the actual structure, a segment of uniform transmission line (i.e., a transmission line with constant crosssection along its length) can be modeled by the circuit shown in Figure 4-11(b). The primary constants can be defined as follows:

Resistance along the line	$= R$	
Inductance along the line	$= L$	all specified
Conductance shunting the line	$= G$	per unit length.
Capacitance shunting the line	$= C$	

Thus R, L, G, and C are also referred to as resistance, inductance, conductance, and capacitance per unit length. (Sometimes p.u.l. is used as shorthand for **per unit length**.) In the metric system we use ohms per meter (Ω/m), henries per meter (H/m), siemens per meter (S/m) and farads per meter (F/m), respectively. The values of R, L, G, and C are affected by the geometry of the transmission line and by the electrical properties of the dielectrics and conductors. G and C are almost entirely due to the properties of the dielectric and R is due to loss in the metal more than anything else. L is mostly a function of geometry, as most materials used with transmission lines have $\mu_r = 1$.

In most transmission lines the effects due to L and C tend to dominate because of the relatively low series resistance and shunt conductance. The propagation characteristics of the line are described by its loss-free, or lossless, equivalent line, although in practice some information about R or G is necessary to determine actual power losses. The lossless concept is just a useful and good approximation. The lossless approximation is not valid
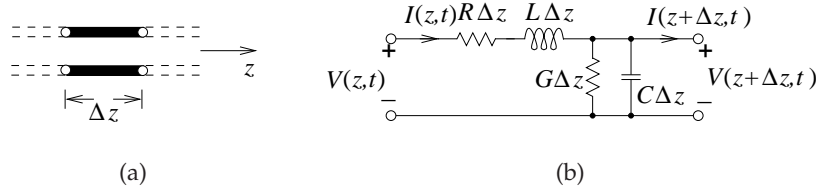
Figure 4-11 The uniform transmission line: (a) transmission line segment of length Δz; and (b) primary constants assigned to a lumped-element model of a transmission line.

for narrow on-chip interconnections, as their resistance is very large.

4.6.1 Derivation of Transmission Line Properties

In this section the differential equations governing the propagation of signals on a transmission line are derived. Solution of the differential equations describes how signals propagate, and leads to the extraction of a few parameters that describe transmission line properties.

From **Kirchoff's laws** applied to the model of Figure 4-11(b) and taking the limit as $\Delta x \to 0$ the transmission line or telegraphist's equations are

$$\frac{\partial v(z,t)}{\partial z} = -Ri(z,t) - L\frac{\partial i(z,t)}{\partial t} \tag{4.14}$$

$$\frac{\partial i(z,t)}{\partial z} = -Gv(z,t) - C\frac{\partial v(z,t)}{\partial t} . \tag{4.15}$$

For the sinusoidal steady-state condition with cosine-based phasors[11] these become

$$\frac{\mathrm{d}V(z)}{\mathrm{d}z} = -(R + \jmath\omega L)I(z) \tag{4.16}$$

$$\frac{\mathrm{d}I(z)}{\mathrm{d}z} = -(G + \jmath\omega C)V(z) . \tag{4.17}$$

Eliminating $I(z)$ in Equations (4.16) and (4.17), yields the wave equation for $V(z)$:

$$\frac{\mathrm{d}^2V(z)}{\mathrm{d}z^2} - \gamma^2 V(z) = 0 . \tag{4.18}$$

Similarly,

$$\frac{\mathrm{d}^2I(z)}{\mathrm{d}z^2} - \gamma^2 I(z) = 0 , \tag{4.19}$$

---

[11] $V(z)$ and $I(z)$ are phasors and $v(z,t) = \Re\{V(z)e^{\jmath\omega t}\}$, $i(z,t) = \Re\{I(z)e^{\jmath\omega t}\}$. $\Re\{w\}$ denotes the real part of a complex number w.

where the propagation constant is

$$\gamma = \alpha + j\beta = \sqrt{(R + j\omega L)(G + j\omega C)} \,. \tag{4.20}$$

In Equation (4.20) α is called the attenuation coefficient and has units of Nepers per meter; and β is called the phase-change coefficient, or phase constant, and has units of radians per meter (expressed as rad/m or radians/m). Nepers and radians are dimensionless units, but serve as prompts for what is being referred to.

Equations (4.18) and (4.19) are second-order differential equations that have solutions of the form

$$\begin{align} V(z) &= V_0^+ e^{-\gamma z} + V_0^- e^{\gamma z} \tag{4.21} \\ I(z) &= I_0^+ e^{-\gamma z} + I_0^- e^{\gamma z} \,. \tag{4.22} \end{align}$$

The physical interpretation of these solutions is that $V_0^+ e^{-\gamma z}$ and $I_0^+ e^{-\gamma z}$ are forward-traveling waves (moving in the $+z$ direction) and $V_0^- e^{\gamma z}$ and $I_0^- e^{\gamma z}$ are backward-traveling waves (moving in the $-z$ direction). Substituting Equation (4.21) in Equation (4.16) results in

$$I(z) = \frac{\gamma}{R + j\omega l}\left[V_0^+ e^{-\gamma z} - V_0^- e^{\gamma z}\right] . \tag{4.23}$$

Then from Equations (4.23) and (4.22) we have

$$I_0^+ = \frac{\gamma}{R + j\omega L} V_0^+ ; I_0^- = \frac{\gamma}{r + j\omega L}(-V_o^-) \,. \tag{4.24}$$

Defining what is called the characteristic impedance as

$$Z_0 = \frac{V_0^+}{I_0^+} = \frac{-V_0^-}{I_0^-} = \frac{R + j\omega L}{\gamma} = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \,. \tag{4.25}$$

Now Equations (4.21) and (4.22) can be rewritten as

$$\begin{align} V(z) &= V_0^+ e^{-\gamma z} + V_0^- e^{\gamma z} \tag{4.26} \\ I(z) &= \frac{V_0^+}{Z_0} e^{-\gamma z} - \frac{V_0^-}{Z_0} e^{\gamma z} \,. \tag{4.27} \end{align}$$

Converting back to the time domain:

$$\begin{align} V(z,t) &= \left|V_0^+\right| \cos(\omega t - \beta z + \varphi^+) e^{-\alpha z} \tag{4.28} \\ &\quad + \left|V_0^-\right| \cos(\omega t + \beta z + \varphi^-) e^{\alpha z} \,, \tag{4.29} \end{align}$$

where

$$V_0^+ = \left|V_0^+\right| e^{j\varphi^+}, \quad \text{and} \quad V_0^- = \left|V_0^-\right| e^{j\varphi^-} \,, \tag{4.30}$$

and so the following quantities are defined:

$$\text{Propagation constant:} \quad \gamma \quad = \sqrt{(R + \jmath\omega L)(G + \jmath\omega C)} \qquad (4.31)$$

$$\text{Attenuation constant:} \quad \alpha \quad = \Re\{\gamma\} \qquad (4.32)$$

$$\text{Phase constant:} \quad \beta \quad = \Im\{\gamma\} \qquad (4.33)$$

$$\text{Wavenumber:} \quad k \quad = -\jmath\gamma \qquad (4.34)$$

$$\text{Phase velocity:} \quad v_p \quad = \frac{\omega}{\beta} \qquad (4.35)$$

$$\text{Wavelength:} \quad \lambda \quad = \frac{2\pi}{|\gamma|} = \frac{2\pi}{|k|} \ , \qquad (4.36)$$

where $\omega = 2\pi f$ is the radian frequency and f is the frequency in hertz. The wavenumber k as defined here is used in electromagnetics and where wave propagation is concerned.[12]

For low-loss materials (and for all of the substrate materials that are useful for transmission lines), $\alpha \ll \beta$ and so $\beta \approx |k|$, then the following approximates are valid:

$$\text{Wavenumber:} \quad k \quad \approx -\jmath\beta \qquad (4.37)$$

$$\text{Phase velocity:} \quad v_p \quad \approx \frac{\omega}{\beta} \qquad (4.38)$$

$$\text{Wavelength:} \quad \lambda \quad \approx \frac{2\pi}{\beta} = v_p/f \ . \qquad (4.39)$$

The important result here is that a voltage wave (and a current wave) can be defined on a transmission line. One more parameter needs to be introduced: the group velocity,

$$v_g = \frac{\partial\omega}{\partial\beta} \ . \qquad (4.40)$$

The group velocity is the velocity of a modulated waveform's envelope and describes how fast information propagates. It is the velocity at which the energy (or information) in the waveform moves. Thus group velocity, can never be more than the speed of light in a vacuum, c. Phase velocity, however, can be more than c. For a lossless, dispersionless line, the group and phase velocity are the same. If the phase velocity is frequency independent, then β is linearly proportional to ω and the group velocity is the same as the phase velocity ($v_g = v_p$).

Electrical length is often used in working with transmission line designs prior to establishing the physical length of a line. The electrical length of a transmission line is expressed either as a fraction of a wavelength or

---

[12] There is an alternative definition for wavenumber, $\nu = 1/\lambda$, which is used by physicists and engineers dealing with particles. When ν is used as the wavenumber, k is referred to as the circular wavenumber or angular wavenumber.

in degrees (or radians), where a wavelength corresponds to $360°$ (or 2π radians). So if β is the phase constant of a signal on a transmission line and ℓ is its physical length, the electrical length of the line in radians is $\beta\ell$.

EXAMPLE 4. 2   Physical and Electrical Length

A transmission line is 10 cm long and at the operating frequency the phase constant β is 30 m^{-1} and the wavelength is 40 cm. What is the electrical length of the line?

Solution: Let the physical length of the line be ℓ = 10 cm = 0.1 m. Then the electrical length of the line is $\ell_e = \beta\ell = (30 \text{ m}^{-1}) \times 0.1 \text{ m} = 3$ radians. The electrical length can also be expressed as $\ell_e = (3 \text{ radians}) = 3 \times 180/\pi = 171.9°$ or as $\ell_e = 3/(2\pi) \, \lambda = 0.477 \, \lambda$.

EXAMPLE 4. 3   Forward- and Backward-Traveling Waves

A transmission line ends (i.e., is terminated) in an open circuit. What is the relationship between the forward-traveling and backward-traveling voltage waves at the end of the line.

Solution: At the end of the line the total current is zero, so that $I^+ + I^- = 0$ and so

$$I^- = -I^+ . \tag{4.41}$$

Also, the forward-traveling voltage and forward-traveling current are related by the characteristic impedance:

$$Z_0 = V^+/I^+ . \tag{4.42}$$

Similarly the backward-traveling voltage and backward-traveling current are related by the characteristic impedance:

$$Z_0 = -V^-/I^- , \tag{4.43}$$

however, there is a change in sign, as there is a change in the direction of propagation. Combining Equations (4.41)–(4.43),

$$Z_0 = V^+/I^+ = -V^-/I^- \tag{4.44}$$

and so substituting for I^-,

$$V^+ = -V^- I^+/I^- = -V^- I^+/(-I^+) = V^- . \tag{4.45}$$

So the total voltage at the end of the line, V_{TOTAL} is $V^+ + V^- = 2V^+$—the total voltage at the end of the line is double the incident (forward-traveling) voltage.

EXAMPLE 4. 4   *RLGC* Parameters

A transmission line has the following $RLCG$ parameters: $R = 100 \ \Omega\text{m}^{-1}$, $L = 80 \ \text{nH·m}^{-1}$, $G = 1.6 \ \text{S·m}^{-1}$, and $C = 200 \ \text{pF·m}^{-1}$. Consider a traveling wave on the transmission line with a frequency of 2 GHz.

(a) What is the attenuation constant?

(b) What is the phase constant?

(c) What is the phase velocity?

(d) What is the characteristic impedance of the line?

(e) What is the group velocity?

Solution:

(a) α: $\gamma = \alpha + \jmath\beta = \sqrt{(R + \jmath\omega L)(G + \jmath\omega C)}; \quad \omega = 12.57 \cdot 10^9 \ \text{rad} \cdot \text{s}^{-1}$

$$\gamma = \sqrt{(100 + \jmath\omega \cdot 80 \cdot 10^{-9})\left(1.6 + \jmath\omega 200\text{x}10^{-12}\right)} = 17.94 + \jmath 51.85 \ \text{m}^{-1}$$

$$\alpha = \Re\{\gamma\} = 17.94 \ \text{Np} \cdot \text{m}^{-1}$$

(b) Phase constant: $\beta = \Im\{\gamma\} = 51.85 \ \text{rad} \cdot \text{m}^{-1}$

(c) Phase velocity:

$$v_p = \frac{\omega}{\beta} = \frac{2\pi f}{\beta} = \frac{12.57 \times 10^9 \text{rad} \cdot \text{s}^{-1}}{51.85 \ \text{rad} \cdot \text{m}^{-1}} = 2.42 \times 10^8 \ \text{m} \cdot \text{s}^{-1}$$

(d) $Z_0 = (R + \jmath\omega L)/\gamma = (100 + \jmath\omega \cdot 80 \cdot 10^{-9})/(17.94 + \jmath 51.85)$

$$Z_0 = 17.9 + j4.3 \ \Omega$$

Note also that

$$Z_0 = \sqrt{\frac{(R + \jmath\omega L)}{(G + j\omega C)}} \ ,$$

which yields the same answer.

(e) Group velocity:

$$v_g = \left.\frac{\partial\omega}{\partial\beta}\right|_{f=2 \ \text{GHz}}$$

Numerical derivatives will be used, thus

$$v_g = \frac{\Delta\omega}{\Delta\beta} \ .$$

Now β is already known at 2 GHz. At 1.9 GHz $\gamma = 17.884 + \jmath 49.397 \ \text{m}^{-1}$ and so $\beta = 49.397 \ \text{rad/m}$.

$$v_g = \frac{2\pi(2 - 1.9)}{51.85 - 49.397} = 2.563 \times 10^8 \ \text{m/s}$$

4.6.2   Relationship to Signal Transmission in a Medium

In the previous section the telegraphists equations for a transmission modeled as subsections of *RLGC* elements was derived. In this section these are related to signal transmission described by the physical parameters of permittivity and permeability. The development does not go into much detail as the derivation of the wave equations for a particular physical transmission line are involved and can only be derived for a few regular structures. If you are curious, the development is done for a parallel plate and rectangular waveguide in Appendix E on Page 869. The main parameters of describing propagation on a transmission line are Z_0 and γ, and these depend on the permeability and permittivity of the medium containing the EM fields, but also on the spatial variation of the E and H fields. As a result, Z_0 must be numerically calculated or derived analytically. The propagation constant is derived from the field configurations as well with

$$\gamma^2 = - \left(k^2 - k_c^2 \right) \tag{4.46}$$

where the wavenumber is

$$k = \jmath\omega\sqrt{\mu\varepsilon} \tag{4.47}$$

and k_c is called the cutoff wavenumber. For TEM modes, $k_c = 0$. For non-TEM modes, k_c requires detailed evaluation.

The other propagation parameters are unchanged:

$$\begin{aligned} \text{Attenuation constant:} \quad & \alpha &&= \Re\{\gamma\} & (4.48) \\ \text{Phase constant:} \quad & \beta &&= \Im\{\gamma\} & (4.49) \\ \text{Phase velocity:} \quad & v_p &&= \frac{\omega}{\beta} & (4.50) \\ \text{Wavelength:} \quad & \lambda &&= \frac{v_p}{f} \, , & (4.51) \end{aligned}$$

where loss is incorporated in the imaginary parts of ε and μ. When $k_c = 0$ (as it is with coaxial lines, microstrip, and many other two-conductor transmission lines),

$$\gamma = \jmath\omega\sqrt{\mu\varepsilon} \tag{4.52}$$

Comparing γ in Equation (4.20) and Equation (4.52), an equivalence can be developed between the lumped-element form of transmission line propagation and the propagation of an EM wave in a medium. Specifically,

$$-\omega^2\mu\varepsilon = (R + \jmath\omega L)(G + \jmath\omega C) \, . \tag{4.53}$$

If the medium is lossless (μ and ε are real and $R = 0 = G$), then

$$\mu\varepsilon = -\omega^2 LC \, . \tag{4.54}$$

When the medium is free space (a vacuum), then a subscript zero is generally used. Free space is also lossless, so the following results hold:

$$\alpha_0 = 0 \quad \text{and} \quad \beta_0 = -\jmath\gamma = \omega\sqrt{\mu_0\epsilon_0} \, . \tag{4.55}$$

If frequency is specified in gigahertz (indicated by f_{GHz})

$$\beta_0 = 20.958 f_{\mathrm{GHz}} \, . \tag{4.56}$$

So at 1 GHz, $\beta_0 = 20.958 \ \mathrm{rad \cdot m^{-1}}$. In a lossless medium with effective relative permeability $\mu_r = 1$ and effective relative permittivity ε_r,

$$\beta = \sqrt{\varepsilon_r}\beta_0 \, . \tag{4.57}$$

Z_0 depends strongly on the spatial variation of the fields. When there is no variation in the plane transverse to the direction of propagation

$$Z_0 = \sqrt{\frac{\mu}{\varepsilon}} \, . \tag{4.58}$$

However, if there is variation of the fields

$$Z_0 = \kappa\sqrt{\frac{\mu}{\varepsilon}} \, , \tag{4.59}$$

where κ captures the geometric variation of the fields.

If the boundary conditions on a transmission line are such that a required spatial variation of the fields cannot be supported then the signal cannot propagate. The critical frequency at which $k = \jmath\omega\sqrt{\mu\epsilon} = k_c$ is called the cutoff frequency, f_c. Signals cannot propagate on the line if the frequency is below f_c.

4.6.3 Dimensions of γ, α, and β

In the above expressions the **propagation constant**, γ, is multiplied by length in determining impedance and signal levels. It is not surprising then that the SI units of γ are inverse meters ($\mathrm{m^{-1}}$). The **attenuation constant**, α, and the **phase constant**, β, have, strictly speaking, the same units. However, the convention is to introduce the dimensionless quantities **Neper** and radian to convey additional information. Thus the attenuation constant α has the units of Nepers per meter (Np/m), and the phase constant, β, has the units radians per meter (rad/m). The unit Neper comes from the name of the e (= 2.7182818284590452354...) symbol, which is called the Neper[13] [65] or Napier's constant. The number e is sometimes called Euler's

---

[13] The name is derived from John **Napier**, who developed the theory of logarithms described in his treatise *Mirifici Logarithmorum Canonis Descriptio*, 1614, translated as *A Description of the Admirable Table of Logarithms* (see http://www.johnnapier.com/table_of_logarithms_001.htm).

constant after the Swiss mathematician Leonhard Euler. The Neper is used in calculating transmission line signal levels, as in Equations (4.21) and (4.22). The attenuation and phase constants are often separated and then the attenuation constant, or more specifically $e^{-\alpha}$, describes the decrease in signal amplitude per unit length as the signal travels down a transmission line. So when $\alpha = 1$ Np, the signal has decreased to $1/e$ of its original value, and power drops to $1/e^2$ of its original value. The decrease in signal level represents loss and, as with other forms of loss, it is common to describe this loss using the units of decibels per meter (dB/m). Thus 1 Np $= 20 \log_{10} e$ = 8.6858896381 dB. So expressing α as 1 Np/m is the same as saying that the attenuation loss is 8.6859 dB/m. To convert from dB to Np multiply by 0.1151. Thus $\alpha = x$ dB/m $= x \times 0.1151$ Np/m.

EXAMPLE 4. 5   Transmission Line Characteristics

A transmission line has an attenuation of 10 dB·m^{-1} and a phase constant of 50 radians·m^{-1} at 2 GHz.

(a) What is the complex propagation constant of the transmission line?

(b) If the capacitance of the line is 100 pF·m^{-1} and the conductive loss is zero (i.e., $G = 0$), what is the characteristic impedance of the line?

Solution:

(a) $\alpha|_{\mathrm{Np}} = 0.1151 \times \alpha|_{\mathrm{dB}} = 0.1151\times (10 \text{ dB/m}) = 1.151$ Np/m

$\beta = 50$ rad/m

Propagation constant, $\gamma = \alpha + \jmath\beta = 1.151 + \jmath 50$ m^{-1}

(b) $$\gamma = \sqrt{(R + \jmath\omega L)(G + \jmath\omega C)}$$
$$Z_0 = \sqrt{(R + \jmath\omega L)/(G + \jmath\omega C)},$$
therefore $Z_0 = \gamma/(G + j\omega C) . \omega = 2\pi \cdot 2 \times 10^9 \ s^{-1}; \ G = 0; \ C = 100 \times 10^9$ F,
so $Z_0 = 39.8 - \jmath 0.916 \ \Omega$.

4.6.4   Lossless Transmission Line

If the conductor and dielectric are ideal (i.e., lossless), then $R = 0 = G$ and the equations for the transmission line characteristics simplify. The transmission line parameters from Equations (4.25) and (4.31)–(4.36) are then

$$Z_0 = \sqrt{\frac{L}{C}} \tag{4.60}$$

$$\alpha = 0 \tag{4.61}$$

$$\beta = \omega\sqrt{LC} \tag{4.62}$$

$$v_p = \frac{1}{\sqrt{LC}} \tag{4.63}$$

$$\lambda_g = \frac{2\pi}{\omega\sqrt{LC}} = \frac{v_p}{f} \ . \tag{4.64}$$

Note that there is a distinction between a transmission line and an RLC circuit. When referring to a transmission line having an impedance of 50 Ω, this is not the same as saying that the transmission line can be replaced by a 50 Ω resistor. The 50 Ω resistance is the characteristic impedance of the line. That is, the ratio of the forward-traveling voltage wave and the forward-traveling current wave is 50 Ω. It is not correct to call a lossless line reactive. Instead, the input impedance of a lossless line would be reactive if the line is terminated in a reactance. If the line is terminated in a resistance then the input impedance of the line would, in general, be complex, having a real part and a reactive part.

A transmission line cannot be replaced by a lumped element except as follows:

1. When calculating the forward voltage wave of a line which is infinitely long (or there are no reflections from the load). Then the line can be replaced by an impedance equal to the characteristic impedance of the line. The total voltage is then only the forward-traveling component.

2. The characteristic impedance and the load impedance can be plugged into the telegraphists equation (or transmission line equation) to calculate the input impedance of the terminated line.

4.6.5   Coaxial Line

The characteristic impedance of a transmission line is the ratio of the strength of the electric field to the strength of the magnetic field. The calculation of the impedance from the geometry of the line is not always possible except for a few regular geometries. For a coaxial line, the electric fields extend in a radial direction from the center conductor to the outer conductor. So it is possible to calculate the voltage by integrating this E field from the center to the outer conductor. The magnetic field is circular, centered on the center conductor, so the current on the conductor can be calculated as the closed integral of the magnetic field. Solving for the fields in the region between the center and outer conductors yields the following formula for the characteristic impedance of a coaxial line:

$$Z_0 = \frac{138}{\sqrt{\varepsilon_r}} \log \frac{b}{a} \ \Omega \ , \tag{4.65}$$

where ε_r is the relative permittivity of the medium between the center and outer conductors, b is the inner diameter of the outer conductor, and a is
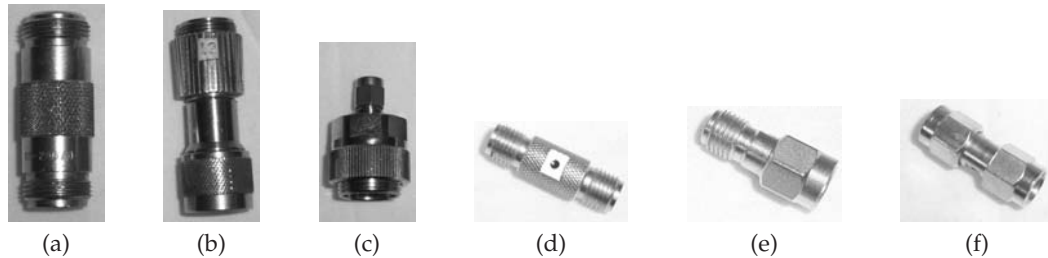
Figure 4-12 Various coaxial transmission line adaptors: (a) N-type female-to-female (N(f)-to-N(f)); (b) APC-7 to N-type male (APC-7-to-N(m)); (c) APC-7 to SMA-type male (APC-7-to-SMA(m)); SMA adapters: (d) SMA-type female-to-female (SMA(f)-to-SMA(f)); (e) SMA-type male-to-female (SMA(m)-to-SMA(f)); and (f) SMA-type male-to-male (SMA(m)-to-SMA(m)).

the outer diameter of the inner conductor. With a higher ε, more energy is stored in the electric field and the capacitance per unit length of the line C increases. As the relative permittivity of the line increases, the characteristic impedance of the line reduces. Equation (4.65) is an exact formulation for the characteristic impedance of a coaxial line. Such a formula can only be approximated for nearly every other line.

Most coaxial cables have a Z_0 of 50 Ω, but different ratios of b and a yield special properties of the coaxial line. When the ratio is 1.65, corresponding to an impedance of 30 Ω, the line has maximum power-carrying capability. The ratio for maximum voltage breakdown is 2.7, corresponding to $Z_0 = 60$ Ω. The characteristic impedance for minimum attenuation is 77 Ω, with a diameter ratio of 3.6. A 50 Ω line is a reasonable compromise. Also the dimensions required for a 50 Ω line filled with polyethylene with a relative permittivity of 2.3 has dimensions that are most easily machined.

The velocity of propagation in a lossless coaxial line of uniform medium is the same as that for a plane wave in the medium. There is one caveat. This is true for all transmission line structures supporting the minimum variation of the fields corresponding to a TEM mode. Higher-order modes, with spatial variations of the fields, will be considered in Chapter 5. The diameter of the outer conductor and the type of internal supports for the internal conductor determine the frequency range of coaxial components. Various transmission line adapters are shown in Figure 4-12. It is necessary to convert between series and also to convert between the sexes (plug and jack) of connectors. The different construction of connectors can be seen more clearly in Figure 4-13. The APC-7 connector is shown in Figure 4-13(c). With this connector, the inside diameter of the outer conductor is 7 mm. The unique feature of this connector is that it is sexless with the interface plate being spring-loaded. These are precision connectors used in microwave measurements. The N-type connectors in Figures 4-13(a) and 4-13(b) are more common day-to-day connectors. There are a large number of
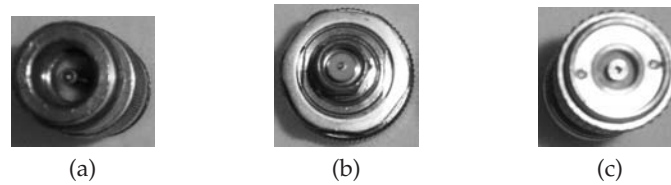
(a)   (b)   (c)

Figure 4-13   Various coaxial transmission line connectors: (a) female N-type (N(f)), coaxial connector; (b) male N-type (N(m)), coaxial connector; and (c) APC-7 coaxial connector.
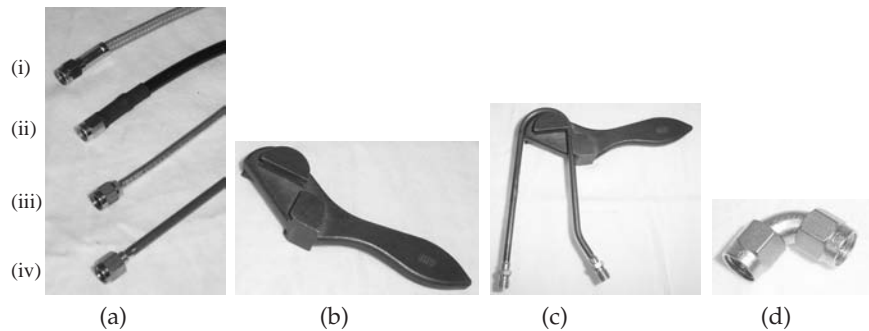

(i)

(ii)

(iii)

(iv)

(a)   (b)   (c)   (d)

Figure 4-14   Coaxial transmission line sections and tools: (a) SMA cables (from the top): flexible cable type I, type II, type III, semirigid cable; (b) semirigid coaxial line bender; (c) semirigid coaxial line bender with line; and (d) SMA elbow.

different types or series f connectors for high-power applications, different frequency ranges, low distortion, and low cost. There are also many types of coaxial cables, as shown in Figure 4-14(a). These are cables for use with SMA connectors (with 3.5 mm outer conductor diameter). These cables range in cost, flexibility, and the number of times they can be reliably flexed or bent. The semirigid cable shown at the bottom of Figure 4-14(a) must be bent using a bending tool, as shown in Figure 4-14(b) and in use in Figure 4-14(c). The controlled bending radius ensures minimal change in the characteristic impedance and propagation constant of the cable. Semirigid cables can only be bent once however. The highest precision bend is realized using an elbow bend, shown in Figure 4-14(d). Various flexible cables have different responses to bending, with higher precision (and more expensive) cables having the least impact on characteristic impedance and phase variations as cables are flexed. The highest-precision flexible cables are used in measurement systems.


EXAMPLE 4. 6   Transmission Line Resonator

Communication filters are often constructed using several shorted transmission line resonators that are coupled to each other. Consider a coaxial line that is short-circuited at one end. The permittivity filling the coaxial line has a relative dielectric constant of 20 and the resonator is to be designed to resonate at a center frequency, f_0, of 1850 MHz when it is one-quarter wavelength long.

(a) What is the wavelength in the dielectric-filled coaxial line?

(b) What is the form of the equivalent circuit (in terms of inductors and capacitors) of the one-quarter wavelength long resonator if the coaxial line is lossless?

(c) What is the length of the resonator?

Solution: The first thing to realize with this example is that the first resonance will occur when the length of the resonator is one-quarter wavelength ($\lambda/4$) long. Resonance generally means that the impedance is either an open or a short circuit and there is energy stored. When the shorted line is $\lambda/4$ long, the input impedance will be an open circuit and energy will be stored.

(a) $\lambda_g = \lambda_0/\sqrt{\varepsilon_r} = 16.2 \text{ cm}/\sqrt{20} = 3.62$ cm.

(b)
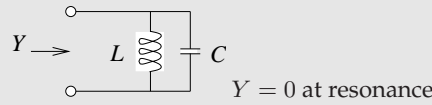
$$Y = Y_L + Y_C = \frac{1}{j\omega L} + j\omega C = \frac{\omega^{-1}}{jL} + j\omega C \,.$$

(c) $\ell = (0.0362 \text{ m})/4 = 9.05$ mm.

4.6.6 Attenuation on a Low-Loss Line

Recall that γ, the propagation constant, is given by

$$\gamma = \sqrt{(R + \jmath\omega L)(G + \jmath\omega C)} \,. \qquad (4.66)$$

This can be written as

$$\gamma = \jmath\omega\sqrt{LC}\sqrt{\left(1 + \frac{R}{\jmath\omega L}\right)\left(1 + \frac{G}{\jmath\omega C}\right)} \,. \qquad (4.67)$$

With a low-loss line, $R \ll \omega L$ and $G \ll \omega C$, and so using a Taylor series approximation

$$\left(1 + \frac{R}{\jmath\omega L}\right)^{1/2} \approx 1 + \frac{1}{2}\frac{R}{\jmath\omega L} \qquad (4.68)$$

$$\left(1 + \frac{G}{\jmath\omega C}\right)^{1/2} \approx 1 + \frac{1}{2}\frac{G}{\jmath\omega C} \, , \tag{4.69}$$

thus

$$\gamma \approx \frac{1}{2}\left(R\sqrt{\frac{C}{L}} + G\sqrt{\frac{L}{C}}\right) + \jmath\omega\sqrt{LC} \, . \tag{4.70}$$

Hence for low-loss lines,

$$\alpha \approx \frac{1}{2}\left(\frac{R}{Z_0} + GZ_0\right) \tag{4.71}$$

$$\beta \approx \omega\sqrt{LC} \, . \tag{4.72}$$

What Equation (4.72) indicates is that for low-loss lines the attenuation constant, α, consists of dielectric- and conductor-related parts; that is,

$$\alpha = \alpha_d + \alpha_c \, , \tag{4.73}$$

where

$$\alpha_d = GZ_0/2 \tag{4.74}$$

is the loss contributed by the dielectric, called the dielectric loss, and

$$\alpha_c = R/2Z_0 \tag{4.75}$$

is the loss contributed by the conductor, called the ohmic or conductor loss.

For a microstrip line, an estimate of G is [51]

$$G = \frac{\epsilon_e - 1}{\epsilon_r - 1}\omega\tan\delta_\ell\,\epsilon_r C_{\text{air}} \, , \tag{4.76}$$

where $\tan\delta$ is the loss tangent of the microstrip substrate. So from Equations (4.173) and (4.74)

$$\alpha_d = \frac{GZ_0}{2} = \frac{1}{2}\frac{\epsilon_e - 1}{\epsilon_r - 1}\omega\tan\delta_\ell\,\epsilon_r C_{\text{air}}\frac{1}{c\sqrt{CC_{\text{air}}}} \, . \tag{4.77}$$

Or, using Equation (4.176), this can be written as

$$\alpha_d = \frac{\omega}{c}\,\tan\delta_\ell\,\frac{\epsilon_r(\epsilon_e - 1)}{2\sqrt{\epsilon_e}(\epsilon_r - 1)}\,\text{Np}\cdot\text{m}^{-1} \, . \tag{4.78}$$

4.6.7 Lossy Transmission Line Dispersion

On a lossy line, both phase velocity and attenuation constant are, in general, frequency dependent and so a lossy line is, in general, dispersive. That is, different frequency components of a signal travel at different speeds, and the phase velocity, v_p, is a function of frequency. As a result and the signal

will spread out in time and, if the line is long enough, it will be difficult to extract the original information.

In the previous section it was seen, in Equation (4.72), that $\beta/\omega = v_p$ is approximately frequency independent for a low-loss line. Also, the conductor component of the attenuation constant, α_c in Equation (4.75), is approximately frequency independent. However, the dielectric component, α_d in Equation (4.78), is frequency dependent even for a low-loss line. If the transmission line has moderate loss, as with microstrip lines, all of the propagation parameters will be frequency dependent and the line is dispersive.

4.6.8 Design of a Dispersionless Lossy Line

The parameters that are important in describing the signal propagation properties of a transmission line are the propagation constant, γ, and the characteristic impedance, Z_0. Instead of γ it is more appropriate to examine α and $v_p = \beta/\omega$ as these are the parameters that are ideally frequency independent if a signal, such as a pulse or modulated carrier, are to travel down the line and not be distorted. As was seen in the previous section these are generally frequency dependent for a lossy line. However, it is possible to design a line that is lossy but dispersionless, that is, α, β/ω, and Z_0 are independent of frequency. In this section a transmission line design is presented for a dispersionless line.

For any line the propagation constant is

$$\gamma = \sqrt{(R + \jmath\omega L)(G + \jmath\omega C)} = \jmath\omega\sqrt{LC}\left[\left(1 + \frac{R}{\jmath\omega L}\right)\left(1 + \frac{G}{\jmath\omega C}\right)\right]^{1/2}.$$
(4.79)

If R, L, C, and G are selected so that

$$\frac{R}{L} = \frac{G}{C},$$
(4.80)

then for this case

$$\gamma = \alpha + \jmath\beta = \jmath\omega\sqrt{LC}\left(1 + \frac{R}{\jmath\omega L}\right) = R\sqrt{\frac{C}{L}} + \jmath\omega\sqrt{LC}.$$

From this the attenuation constant, α, and phase constant, β, are given by

$$\alpha = R\sqrt{\frac{C}{L}}, \quad \beta = \omega\sqrt{LC},$$
(4.81)

and the phase velocity is
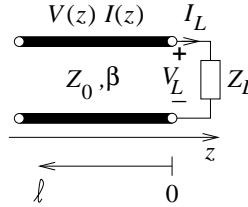
$$v_p = \frac{1}{\sqrt{LC}}.$$
(4.82)

Figure 4-15 A terminated transmission line.

To complete the analysis consider the characteristic impedance

$$Z_0 = \sqrt{\frac{R + \jmath\omega L}{G + \jmath\omega C}} = \sqrt{\frac{L}{C}}\sqrt{\frac{R/L + \jmath\omega}{G/C + \jmath\omega}} \tag{4.83}$$

and referring to Equation (4.80) it is seen that the second square root is just 1, so

$$Z_0 = \sqrt{\frac{L}{C}} \ , \tag{4.84}$$

which is frequency independent. So the important characteristics describing signal propagation are independent of frequency and so the transmission line is dispersionless.

4.7 The Terminated Lossless Line

4.7.1 Total Voltage and Current

Consider the terminated line shown in Figure 4-15. Assume an incident or forward-traveling wave, with traveling voltage $V_0^+ e^{-\jmath\beta z}$ and current $I_0^+ e^{-\jmath\beta z}$, respectively, propagating toward the load Z_L at $z = 0$. The characteristic impedance of the transmission line is the ratio of the voltage and current traveling waves so that

$$\frac{V_0^+ e^{-\jmath\beta z}}{I_0^+ e^{-\jmath\beta z}} = \frac{V_0^+}{I_0^+} = Z_0 \ . \tag{4.85}$$

The reflected wave has a similar relationship (but watch the sign change):

$$\frac{V_0^- e^{\jmath\beta z}}{-I_0^- e^{\jmath\beta z}} = \frac{V_0^-}{-I_0^-} = Z_0 \ . \tag{4.86}$$

The load Z_L imposes an additional constraint on the relationship of the total voltage and current at $z = 0$:

$$\frac{V_L}{I_L} = \frac{V(z = 0)}{I(z = 0)} = Z_L \ . \tag{4.87}$$

When $Z_L \neq Z_0$ there must be a reflected wave with appropriate amplitude to satisfy the above equations. Now the total voltage, $V(z)$, is related to the traveling voltage waves by

$$V(z) = V_0^+ e^{-\jmath\beta z} + V_0^- e^{\jmath\beta z} , \qquad (4.88)$$

and the total current, $I(z)$, is related to the traveling current waves by

$$I(z) = \frac{V_0^+}{Z_0} e^{-\jmath\beta z} - \frac{V_0^-}{Z_0} e^{\jmath\beta z} = I_0^+ e^{-\jmath\beta z} + I_0^- e^{\jmath\beta z} . \qquad (4.89)$$

Thus at the termination of the line ($z = 0$),

$$\frac{V(0)}{I(0)} = Z_L = Z_0 \frac{V_0^+ + V_0^-}{V_0^+ - V_0^-} .$$

This can be rearranged as the ratio of the reflected voltage to the incident voltage:

$$\frac{V_0^-}{V_0^+} = \frac{Z_L - Z_0}{Z_L + Z_0} .$$

This ratio is defined as the voltage reflection coefficient

$$\Gamma^V = \frac{V_0^-}{V_0^+} = \frac{Z_L - Z_0}{Z_L + Z_0} . \qquad (4.90)$$

This relationship can be rewritten so that the input load impedance can be obtained from the reflection coefficient:

$$Z_L = Z_0 \frac{1 + \Gamma^V}{1 - \Gamma^V} . \qquad (4.91)$$

Similarly the current reflection coefficient can be written as

$$\Gamma^I = \frac{I_0^-}{I_0^+} = \frac{-Z_L + Z_0}{Z_L + Z_0} = -\Gamma^V. \qquad (4.92)$$

The voltage reflection coefficient is used most of the time, so the reflection coefficient, Γ, on its own refers to the voltage reflection coefficient, $\Gamma^V = \Gamma$.

The total voltage and current waves on the line can be written as

$$V(z) = V_0^+ [e^{-\jmath\beta z} + \Gamma e^{\jmath\beta z}] \qquad (4.93)$$

$$I(z) = \frac{V_0^+}{Z_0} [e^{-\jmath\beta z} - \Gamma e^{\jmath\beta z}] . \qquad (4.94)$$

From the above equations, it can be seen that the voltage and current on the line consist of a superposition of incident and reflected waves. This

superposition is also a wave, but not a propagating wave, and is referred to as a **standing wave**. There are several special cases that are noteworthy. The most important of these is the case when there is no reflected wave and $\Gamma = 0$. To obtain $\Gamma = 0$ the value of load impedance, Z_L, is equal to Z_0, the characteristic impedance of the transmission line as seen from Equation (4.90). Such a load is then said to be matched to the line (it is a **matched termination**) since there is no reflection of the incident wave.

EXAMPLE 4. 7	Reflection Coefficient

A load consists of a shunt connection of a capacitor of 10 pF and a resistor of 60 Ω. The load terminates a lossless 50 Ω transmission line. The operating frequency is 5 GHz.

(a) What is the impedance of the load?

(b) What is the normalized impedance of the load (normalized to Z_0 of the line)?

(c) What is the reflection coefficient of the load?

(d) What is the current reflection coefficient of the load? (When the term reflection coefficient is used without a qualifier it is assumed to be the voltage reflection coefficient.)

Solution:

(a) $C = 10 \cdot 10^{-12}\text{F} \ ; \ R = 60 \ \Omega \ ; \ f = 5.10^9 ; \omega = 2\pi f \ ; \ Z_0 = 50 \ \Omega$

$$ Z_L = R||C = (1/R + \jmath\omega C)^{-1} = 0.168 - \jmath 3.174 \ . $$

(b) $z_L = Z_L/Z_o = 3.368 \cdot 10^{-3} - \jmath 0.063$.

(c) $\Gamma_L = (z_L - 1)/(z_L + 1) = -0.985 - \jmath 0.126 = 0.993 \angle 187.3° $.

(d) $\Gamma_L^i = -\Gamma_L = 0.985 - \jmath 0.126 = 0.993 \angle (187.3 - 180)° = 0.993 \angle 7.3° $.

4.7.2 *Power Flow and Return Loss*

Now consider the power flow on the line. The incident (forward-traveling) wave carries the power

$$ P^+ = \frac{1}{2}\Re\left\{ \left(V_0^+ e^{-\jmath\beta z}\right)\left(\frac{V_0^+ e^{-\jmath\beta z}}{Z_0}\right)^* \right\} = \frac{1}{2}\frac{|V_0^z|^2}{Z_0} \tag{4.95} $$

and the reflected wave carries the power

$$ P^- = \frac{1}{2}\frac{\left|V_0^-\right|^2}{Z_0} = \frac{|\Gamma|^2}{2}\frac{\left|V_0^+\right|^2}{Z_0} \ , $$

where Equations (4.93) and (4.94) are used. Considering the conservation of power, the power delivered to the load is

$$P_L = \frac{1}{2}\Re\{V_L I_L^*\} = P^+ - P^- = P^+\left(1 - |\Gamma_L|^2\right) . \qquad (4.96)$$

Thus the total power on the line is equal to the incident power minus the reflected power. The other noteworthy cases are when there is an open circuit, a short circuit, or a purely reactive load at the end of a transmission line. These have $|\Gamma| = 1$. Thus all power is reflected back to the source and $P_L = 0$.

The power that is absorbed by the load appears as a loss as far as the incident and reflected waves are concerned. To describe this, the concept of return loss (RL) is introduced and defined as

$$\mathrm{RL} = -20 \log |\Gamma| \; \mathrm{dB} . \qquad (4.97)$$

RL is a measure of how much of the available power is not delivered to the load. A matched load ($\Gamma = 0$) has a return loss of ∞ dB, whereas a total reflection ($|\Gamma| = 1$) has a return loss of 0 dB.

EXAMPLE 4. 8   Transmission Line Model at the Load

A transmission line with a characteristic impedance of 75 Ω supports a forward-traveling wave with a power of 1 μW. The line is terminated in a resistance of 100 Ω. Draw the lumped-element equivalent circuit at the interface between the transmission line and the load.

Solution:

The equivalent circuit will have the form
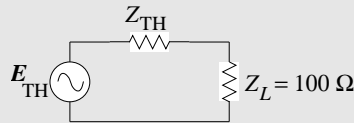
where E_{TH} is the Thevenin equivalent generator and Z_{TH} is the Thevenin equivalent generator impedance. The amplitude of the forward-traveling voltage wave is obtained by calculating the power in the forward-traveling wave:

$$
\begin{aligned}
P^+ &= \frac{1}{2}(V^+)^2/Z_0 \\
&= (V^+)^2/150 \\
&= 1 \, \mu\mathrm{W} = 10^{-6} .
\end{aligned}
$$
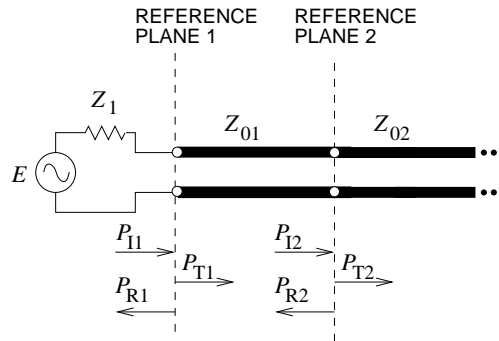
Figure 4-16 Transmission line network with a finite-length line of characteristic impedance Z_{01} and an infinitely long transmission line of characteristic impedance Z_{02}.

So $V^+ = \sqrt{150.10^{-6}} = 12.25$ mV. Note that V^+ is not E_{TH}. To calculate E_{TH}, consider the following circuit which will result in maximum power transfer:
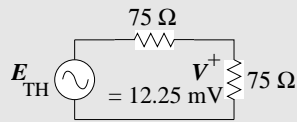
So $E_{\text{TH}} = 2V^+ = 24.5$ mV. Since the line has a characteristic impedance of 75 Ω, then $Z_{\text{TH}} = 75$ Ω. So the lumped-element equivalent circuit at the load is

---
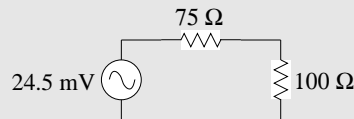
EXAMPLE 4. 9   Transmission Line Power Calculations

The transmission line shown in Figure 4-16 consists of a source with Thevenin impedance $Z_1 = 40$ Ω and source $E = 5$ V (peak), connected to a quarter-wavelength long line of characteristic impedance $Z_{01} = 40$ Ω, which in turn is connected to an infinitely long line of characteristic impedance $Z_{02} = 100$ Ω. The transmission lines are lossless. Two reference planes are shown in Figure 4-16. At reference plane 1 the incident power is P_{I1} (the maximum available power from the source), the reflected power is P_{R1}, and the transmitted power is P_{T1}. P_{I2} (the maximum available power from Z_{01}), P_{R2}, and P_{T2} are similar quantities at reference plane 2. P_{I1}, P_{R1}, P_{T1}, P_{I2}, P_{R2}, and P_{T2} are steady-state quantities.

(a) What is P_{I1}?

(b) What is P_{T2}?

Solution:

(a) P_{I1} is the available power from the generator. Since the Thevenin impedance of the generator is 40 Ω, P_{I1} is the power that would be delivered to a matched load (the maximum available power). An equivalent problem is

$Z_1 = 40\ \Omega$

$Z_L = 40\ \Omega$
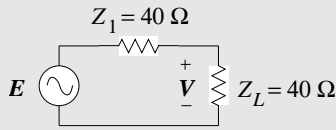
where $V = \frac{1}{2}E = 2.5$ V$_{\text{peak}}$.

$$P_{I1} = \text{Power in } Z_L$$
$$= \frac{1}{2}(\mathbf{V})^2 \frac{1}{Z_L} = \frac{1}{2}(2.5)^2 \frac{1}{40}$$
$$= 0.07813 \text{ W} = 78.13 \text{ mW}. \tag{4.98}$$

Note that the $\frac{1}{2}$ is used because peak voltage is used.

(b) Now the problem becomes interesting and there are many ways to solve it. One of the key observations is that the first transmission line has the same characteristic impedance as the Thevenin equivalent impedance of the generator, $Z_{01} = Z_1$, and so can be ignored where appropriate. This observation will be used in this example. One way to proceed is to directly calculate P_{T2}, and a second approach is to calculate the incident and reflected powers at reference plane 2 and then to determine P_{T2}.

(i) First approach: Looking to the left from Port 2 the circuit can be modeled as an equivalent circuit having a Thevenin equivalent resistance of 40 Ω and a Thevenin equivalent voltage that produces the right amount of power. So consider the following circuit:

$Z_{01} = 40\ \Omega$

$Z_L = 100\ \Omega$
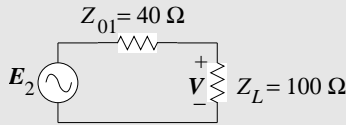
Here, E_2 is the voltage required to deliver the maximum possible power to the load and we know that this power is P_{I1} (= 78 mW), as calculated above. The load is 100 Ω as the second transmission line is infinitely long. So based on the previous problem we know that E_2 is just E or 5 V. A reasonable question would be why is it not 2.5 V, as this would be the voltage across $Z_L = 40\ \Omega$ in Part (a). However, 2.5 V is the voltage of the forward-traveling voltage wave on the first transmission line with characteristic impedance $Z_{01} = 40\ \Omega$. So the voltage across the load is

$$V = E\frac{100}{40 + 100} = E\frac{100}{140} = 3.57 \text{ V}_{\text{peak}}. \tag{4.99}$$

So

$$P_{T2} = P_L = \frac{1}{2}(V)^2 \frac{1}{Z_L}$$
$$= \frac{1}{2}(3.57)^2 \frac{1}{100}$$
$$= 0.0638 \text{ W} = 63.8 \text{ mW}. \tag{4.100}$$

A quick check is that this is less than P_{I1}, as it should be.

(ii) Second Approach: This time P_{R2}, the reflected power at reference plane 2, will be calculated. The incident power at plane 2, P_{I2}, is just P_{I1}, so $P_{I2} = P_{I1} = 78.13$ mW.

P_{R2} can be calculated from the voltage reflection coefficient at plane 2:

$$\Gamma_2 = \frac{Z_L - Z_{01}}{Z_L + Z_{01}} = \frac{100 - 40}{100 + 40} = 0.429$$
$$P_{R2} = \Gamma_2^2 P_{I2} = 0.429^2 \times 78 \text{ mW} = 14.36 \text{ mW} .$$

So $P_{T2} = P_{I1} - P_{R2} = 78.13 \text{ mW} - 14.36 \text{ mW} = 63.7 \text{ mW}$, which is the same transmitted power as in the first approach, allowing for rounding error.

EXAMPLE 4. 10   Advanced Power Calculations

This example is similar to Example 4.9. Again, the transmission line network of Figure 4-16 is considered, but now the characteristic impedance of the first transmission line is not the same as the generator impedance so the simplification used in the previous example can no longer be used. Now the generator has a Thevenin impedance $Z_1 = 40$ Ω and source $E = 5$ V_{peak}, connected to a one-quarter wavelength long line of characteristic impedance $Z_{01} = 30$ Ω which in turn is connected to an infinitely long line of characteristic impedance $Z_{02} = 100$ Ω. The transmission lines are lossless. Two reference planes are shown in Figure 4-16. At reference plane 1 the incident power is P_{I1} (the maximum available power from the source), the reflected power is P_{R1} and the transmitted power is P_{T1}. P_{I2} (the maximum available power from Z_{01}), P_{R2}, and P_{T2} are similar quantities at reference plane 2. P_{I1}, P_{R1}, P_{T1}, P_{I2}, P_{R2}, and P_{T2} are steady-state quantities.

(a) What is P_{I1}?
(b) What is P_{T2}?
(c) Determine P_{T1}, P_{I2}, P_{R1}, and P_{R2}.

Solution:

One of the first things to note is that the infinitely long 100 Ω transmission line is indistinguishable from a 100 Ω resistor, so the reduced form of the problem is as shown below:
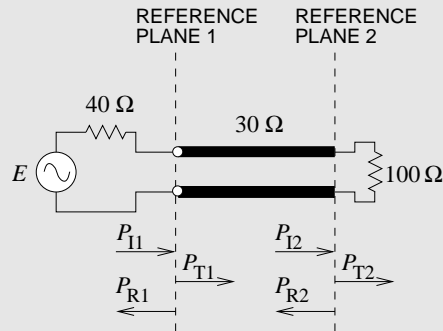
(a) P_{I1} was calculated in Example 4.9 in Equation (4.98):

$$P_{I1} = 78.13 \text{ mW} . \tag{4.101}$$

(b) The problem here is finding P_{T2}. Recall that the powers here are steady state quantities so that we are not considering multiple reflections of, say, a pulse. Since the system is lossless, the power delivered by the generator must be the power delivered to the infinitely long transmission line Z_{02} (i.e., P_{T2}). The telegraphist's equation can be used to calculate the input impedance, Z_{in}, of the two transmission line system; that is, the input impedance of Z_{01} from the generator end. However, a simpler way to find this impedance is to realize that the Z_{01} line is a one-quarter wave transformer so that

$$Z_{01} = 30 = \sqrt{Z_{\text{in}} Z_{02}} = \sqrt{100 Z_{\text{in}}} , \tag{4.102}$$

and so

$$Z_{\text{in}} = 9 \ \Omega . \tag{4.103}$$

The equivalent circuit is thus
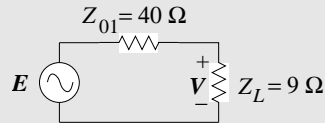
where E is the original generator voltage of 5 V and

$$V = \frac{9}{40 + 9} 5 = 0.9184 \text{ V} . \tag{4.104}$$

The power delivered by the generator to the 9 Ω load is

$$P_{T2} = \frac{1}{2} V^2 / 9 = 0.04686 \text{ W} = 46.86 \text{ mW} . \tag{4.105}$$

(c) The power transmitted into the system at reference plane 1, P_{T1}, is the same as the power transmitted to the 100 Ω load, as the first transmission line is lossless; that is,

$$P_{T1} = P_{T2} = 46.86 \text{ mW} . \tag{4.106}$$

Also
$$P_{R1} = P_{I1} - P_{T2} = (78.13 - 46.86) \text{ mW} = 31.27 \text{ mW}. \tag{4.107}$$

The two remaining quantities to determine are P_{I2} and P_{R2}. There can be a number of interpretations of what these should be, but one thing that is certain is that

$$P_{T2} = P_{I2} - P_{R2} = 46.86 \text{ mW}. \tag{4.108}$$

One interpretation that will be followed here is based on the equivalent circuit at reference plane 2. Let Z_{out} be the impedance looking to the left from reference plane 2. Again, using the property of a one-quarter wavelength long transmission line,

$$Z_{01} = 30 = \sqrt{Z_{\text{out}} \times 40} \tag{4.109}$$

and so
$$Z_{\text{out}} = Z_{01}^2/40 = 30^2/40 = 22.5 \ \Omega. \tag{4.110}$$

The equivalent circuit at reference plane 2 is then
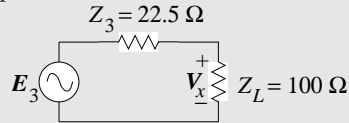
Now determine V_x that results in a power P_{T2} being delivered to the 100 Ω load, so

$$P_{T2} = \frac{1}{2}V_x^2/100 = 0.04686 \text{ W}, \tag{4.111}$$

and
$$V_x = 3.06 \text{ V}. \tag{4.112}$$

From the circuit above,
$$V_x = \frac{100}{100 + 22.5}E_3, \tag{4.113}$$

and so
$$E_3 = 3.75 \text{ V}. \tag{4.114}$$

The available power from this source is obtained by considering
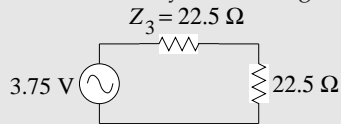
The available power at reference plane 2 is

$$P_{I2} = \frac{1}{2}\frac{(E_3/2)^2}{22.5} = \frac{1.875^2}{2 \times 22.5} = 0.07813 \text{ W} = 78.13 \text{ mW}. \tag{4.115}$$

From Equation (4.108),

$$P_{R2} = P_{I2} - P_{T1} = (78.13 - 46.86) \text{ mW} = 31.27 \text{ mW}. \tag{4.116}$$

Note that $P_{R2} = P_{R1}$, as expected.
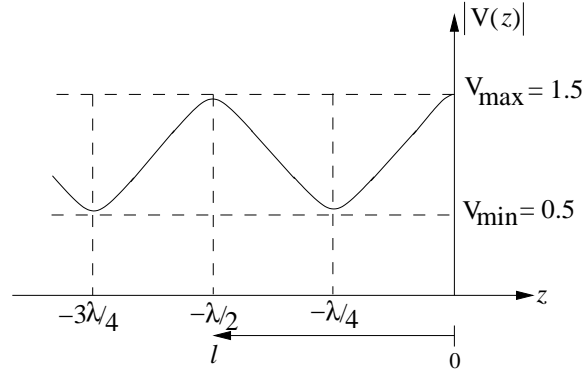
Figure 4-17   Standing wave pattern on a line with $Z_{\text{in}} = 3Z_0$.

4.7.3   Standing Waves and Voltage Standing Wave Ratio

If the load is matched to the line, $\Gamma = 0$, then the magnitude of the total voltage on the line, $|V(z)|$, is equal to $|V_0^+|$, which is a constant. For this reason, such a line is said to be "flat." If the load is mismatched, the presence of reflected waves leads to standing waves where the magnitude of the voltage on the line is not a constant (see Figure 4-17). Thus from Equations (4.93) and (4.94):

$$|V(z)| = |V_0^+|\,|1 + \Gamma e^{2\jmath\beta z}| = |V_0^+|\,|1 + \Gamma e^{2\jmath\beta l}|,$$

where $z = -l$ is the positive distance measured from the load at $z = 0$ toward the generator. Or, setting $\Gamma = |\Gamma|e^{\jmath\Theta}$,

$$|V(z)| = |V_0^+|\,\left|1 + |\Gamma|e^{\jmath(\Theta - 2\beta\ell)}\right|, \tag{4.117}$$

where Θ is the phase of the reflection coefficient ($\Gamma = |\Gamma|e^{\jmath\Theta}$). This result shows that the voltage magnitude oscillates with position z along the line. The maximum value occurs when $e^{\jmath(\Theta - 2\beta\ell)} = 1$ and is given by

$$V_{\text{max}} = |V_0^+|(1 + |\Gamma|). \tag{4.118}$$

Similarly the minimum value of the voltage magnitude occurs when the phase term $e^{\jmath(\Theta - 2\beta l)} = -1$, and is given by

$$V_{\text{min}} = |V_0^+|(1 - |\Gamma|). \tag{4.119}$$

As $|\Gamma|$ increases, the ratio of V_{max} to V_{min} increases and the mismatch defined by the **Voltage Standing Wave Ratio** (**VSWR**) is given by

$$\text{VSWR} = \frac{V_{\text{max}}}{V_{\text{min}}} = \frac{(1 + |\Gamma|)}{(1 - |\Gamma|)}. \tag{4.120}$$

Notice that, in general, Γ may be complex, but VSWR is necessarily always real and $1 \leq \text{VSWR} \leq \infty$. For the matched condition, $\Gamma = 0$ and $\text{VSWR} = 1$, and the closer VSWR is to 1, the better the load is matched to the line and the more power is delivered to the load. The reflection coefficients and standing-wave ratios for short-circuit and open-circuit terminated conditions are -1 and $+1$, respectively, and in both cases the VSWR is infinite.

To determine the position of $V_{\max}$, the voltage maxima, using Equation (4.117) we have

$$\Theta - 2\beta\ell_{\max} = 2n\pi \, , n = 0, 1, 2, \ldots .$$

This can be rewritten in the form

$$\Theta - 2n\pi = 2\frac{2\pi}{\lambda_g}\ell_{\max} \, .$$

Thus the position of voltage maxima, $l_{\max}$, normalized to the wavelength is given by

$$\frac{\ell_{\max}}{\lambda_g} = \frac{1}{2}\left(\frac{\Theta}{2\pi} - n\right), n = 0, -1, -2, \ldots . \tag{4.121}$$

Similarly the position of the voltage minima can be found using Equation (4.117):

$$\Theta - 2\beta\ell_{\min} = 2n + 1\pi \, .$$

Rearranging the terms, we get $l_{\min}$ normalized to the wavelength:

$$\frac{\ell_{\min}}{\lambda_g} = \frac{1}{2}\left(\frac{\Theta}{2\pi} - n + \frac{1}{2}\right), n = 0, -1, -2, \ldots . \tag{4.122}$$

Summarizing from Equations (4.121) and (4.122):

1. The distance between two successive maxima is $\lambda_g/2$.

2. The distance between two successive minima is $\lambda_g/2$.

3. The distance between a maximum and an adjacent minimum is $\lambda/4$.

In a similar manner to that employed above, the magnitude of the total current on the line is found to be

$$|I(\ell)| = \frac{\left|V_0^+\right|}{Z_0}\left|1 - |\Gamma|e^{\jmath(\Theta - 2\beta\ell)}\right| \, . \tag{4.123}$$

Hence the magnitude of the current has its maxima where the voltage is minimum and has its magnitude minima where the voltage magnitude is maximum.

EXAMPLE 4. 11   Standing Wave Ratio

In Example 4.7 a load consisted of a shunt connection of a capacitor of 10 pF and a resistor of 60 Ω. The load terminates a lossless 50 Ω transmission line. The operating frequency is 5 GHz.

(a) What is the standing wave ratio, SWR?

(b) What is the current standing wave ratio (ISWR)? (When SWR is used on its own it is assumed to refer to VSWR.)

Solution:

(a) From Example 4.7 on Page 196, $\Gamma_L = 0.993\angle 187.3°$ and so

$$\text{VSWR} = \frac{1 + |\Gamma_L|}{1 - |\Gamma_L|} = \frac{1 + 0.993}{1 - 0.993} = 285 \ .$$

(b) ISWR = VSWR = 285 .

EXAMPLE 4. 12   Standing Waves

A load has an impedance $Z_L = 45 + j75$ and the system reference impedance, Z_0, is 100 Ω.

(a) What is the reflection coefficient?

(b) What is the current reflection coefficient?

(c) What is the SWR?

(d) What is the ISWR?

(e) The power available from a source with a 100 Ω Thevenin equivalent impedance is 1 mW. The source is connected directly to the load, Z_L. Use the reflection coefficient to calculate the power delivered to Z_L.

(f) What is the total power absorbed by the Thevenin equivalent source impedance?

(g) Discuss the effect of inserting a lossless 100 Ω transmission line between the source and the load.

Solution:

(a) The voltage reflection coefficient:

$$\begin{aligned}
\Gamma_L &= (Z_L - Z_0)/(Z_L + Z_0) = (45 + j75 - 100)/(45 + j75 + 100) \\
&= (93.0\angle 2.204 \text{ rads})/(163.2\angle 0.4773 \text{ rads}) \\
&= 0.570\angle 1.726 \text{ rads} = 0.570\angle 98.9° \\
&= -0.0881 + j0.563 = \Gamma_V \ .
\end{aligned} \tag{4.124}$$

(b) The current reflection coefficient:

$$\Gamma_I = -\Gamma_V = 0.0881 - \jmath 0.563$$
$$= 0.570\angle(98.9° - 180°) = 0.570\angle 81.1° \ . \tag{4.125}$$

(c) The SWR is the VSWR so

$$\mathrm{SWR} = \mathrm{VSWR} = V_{\mathrm{max}}/V_{\mathrm{min}} = (1 + |\Gamma_V|)/(1 - |\Gamma_V|)$$
$$= \frac{1 + 0.570}{1 - 0.570} = 3.65 \ . \tag{4.126}$$

(d) The current SWR is ISWR = VSWR.

(e) To determine the reflection coefficient of the load we begin by developing the Thevenin equivalent circuit of the load. The power available from the source is $P_A = 1$ mW, so the Thevenin equivalent circuit is
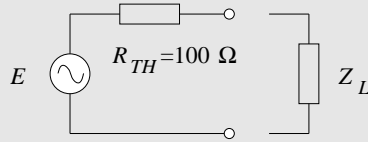
The power reflected by the load is

$$P_R = P_A \left(\Gamma_L^2\right) = 1 \text{ mW} \cdot (0.570)^2 = 0.325 \text{ mW}$$

and so the power delivered to load is

$$P_D = P_A \left(1 - \Gamma_L^2\right) = 0.675 \text{ mW}.$$

(f) It is tempting to think that the power dissipated in R_{TH} is just P_R. However, this is not correct. Instead, the current in R_{TH} must be determined and then the power dissipated in R_{TH} found. Let the current through R_{TH} be I, and this is composed on forward- and backward-traveling components:

$$I = I^+ + I^- = (1 + \Gamma_I)I^+,$$

where I^+ is the forward-traveling current wave. Thus

$$P_A = \frac{1}{2}|I^+|^2 R_{\mathrm{TH}} = \frac{1}{2}|I^+|^2 \times 100 = 1 \text{ mW} = 10^{-3},$$

so

$$I^+ = 4.47 \text{ mA},$$

and

$$I = (1 + \Gamma_I)I^+ = (1 - 0.881 + \jmath 0.563) \times 4.47 \times 10^{-3}$$
$$|I| = 2.57 \text{ mA}.$$
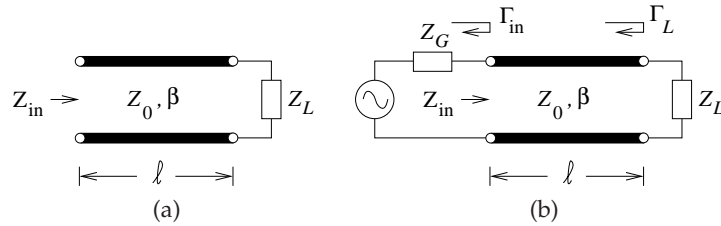
Figure 4-18 Terminated transmission line: (a) a transmission line terminated in a load impedance, Z_L, with an input impedance of Z_{in}; and (b) a transmission line with source impedance Z_G and load Z_L.

Power dissipated in R_{TH} is

$$P_{\text{TH}} = \frac{1}{2}|I|^2 R_{\text{TH}} = \frac{1}{2}(2.57 \times 10^{-3})^2 R_{\text{TH}} = 331 \ \mu\text{W}.$$

Recall that the $\frac{1}{2}$ comes about because in electromagnetics the voltage and currents are peak values and not RMS values.

(g) Inserting a transmission line with the same characteristic impedance as the Thevenin equivalent impedance will have no effect on power flow.

4.7.4 Input Impedance of a Lossless Line

The impedance looking into the line varies with position, as the forward- and backward-traveling waves combine. At a distance $z = -\ell$, the input impedance seen looking toward the load is given by

$$Z_{\text{in}}|_{z=-\ell} = \frac{V(z = -\ell)}{I(z = -\ell)} = Z_0 \frac{1 + |\Gamma|\, e^{\jmath(\Theta - 2\beta\ell)}}{1 - |\Gamma|\, e^{\jmath(\Theta - 2\beta\ell)}} \, , \qquad (4.127)$$

and so

$$Z_{\text{in}} = Z_0 \frac{Z_L + \jmath Z_0 \tan \beta\ell}{Z_0 + \jmath Z_L \tan \beta\ell} \, . \qquad (4.128)$$

This equation is also known as the lossless telegraphist's equation. Also $\beta\ell$ is called the electrical length and is expressed in degrees or fractions of a wavelength, but in calculations it must be converted to radians with 1 wavelength being 2π radians.

Let us note here that Z_{in} is a periodic function of ℓ with period $\lambda/2$ and it varies between the Z_{max} and Z_{min}, where

$$Z_{\text{max}} = \frac{V_{\text{max}}}{I_{\text{min}}} = Z_0(\text{VSWR}) \qquad (4.129)$$

and

$$Z_{\text{min}} = \frac{V_{\text{min}}}{I_{\text{max}}} = \frac{Z_0}{\text{VSWR}} \, . \qquad (4.130)$$

| EXAMPLE 4. 13 | Forward-Traveling Wave Calculations |

A transmission line is driven by a generator with a maximum available power of 20 dBm and a Thevenin equivalent impedance of 50 Ω. The transmission line has a characteristic impedance of 50 Ω.

   (a) What is the Thevenin equivalent generator voltage?

   (b) What is the magnitude of the forward-traveling voltage wave on the line? Assume the line is infinitely long.

   (c) What is the power of the forward-traveling voltage wave?

Solution:

   (a) Maximum available power is delivered to the load when it is conjugately matched to the generator impedance (see right diagram). Then $P_{\text{LOAD}} = \frac{1}{2} V^2/R$ (V is peak voltage) (i.e., $P_{\text{LOAD}} = 20$ dBm $= 0.1$ W) and

$$V = \sqrt{2P_{\text{LOAD}} \cdot R} = \sqrt{2 \times 0.1 \times 50} \ \text{V}_{\text{peak}}$$

$$V = 3.16 \text{ V}$$

$$E_G = 2 \text{ V} = 6.32 \text{ V}.$$

   (b) This is just V above, so $V^+ = 3.16$ V.

   (c) $P^+ = 20$ dBm.

| EXAMPLE 4. 14 | Coaxial Line Resonator |

A shorted coaxial line is used as a resonator. The first resonance is determined to be a parallel (or shunt) resonance and is at 1 GHz.

   (a) Draw the lumped-element equivalent circuit of the resonator.

   (b) What is the electrical length of the resonator?

   (c) What is the impedance looking into the line at resonance?

   (d) If the resonator is $\lambda_g/4$ longer, what is the input impedance of the resonator now?

Solution:

   (a)
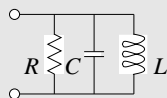
LC is an open circuit at resonance. If the line is lossless, $R = 0$.

(b) $\lambda_g/4 = 90° = \pi/2$.

$\lambda_g/4$
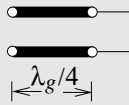
(c) $Z_{\text{in}} = \infty$ (for a lossless line).

(d) $Z_{\text{in}} = 0 \ \Omega$.

EXAMPLE 4. 15   Resonator Model

This example builds on Example 4.6. Communication filters are often constructed using several shorted transmission line resonators that are coupled to each other. Consider a coaxial line that is short-circuited at one end. The dielectric filling the coaxial line has a relative permittivity of 20 and the resonator is to be designed to resonate at a center frequency, f_0, of 1850 MHz.

(a) What is the wavelength in the dielectric-filled coaxial line?

(b) What is the form of the equivalent circuit (in terms of inductors and capacitors) of a one-quarter wavelength long resonator if the coaxial line is lossless?

(c) What is the length of the resonator?

(d) Determine the derivative with respect to frequency of the admittance of the LC equivalent circuit developed in (b). Determine an analytic expression of the derivative at the resonant frequency.

(e) If the diameter of the inner conductor of the coaxial line is 2 mm and the inside diameter of the outer conductor is 5 mm, what is the characteristic impedance of the coaxial line?

(f) Calculate the input admittance of the dielectric-filled coaxial line at $0.99 f_0$, f_0, and $1.01 f_0$. Determine the numerical derivative of the line admittance at f_0.

(g) Derive the numeric values of the equivalent circuit of the resonator at the resonant frequency. Hint: Match the derivative expression derived in (d) with the actual derivative calculated in (f).

Solution:

The first thing to realize with this example is that the first resonance will occur when the length of the resonator is one-quarter wavelength, $\lambda/4$, long. Resonance generally means that the impedance is either an open or a short circuit and there is energy stored. When the shorted line is $\lambda/4$ long, the input impedance will be an open circuit. When the line is of zero length, one of the resonance conditions is met (the input impedance is zero) but energy is not stored, so a zero-length line is not a resonator.

(a) $\lambda_g = \lambda_0/\sqrt{\varepsilon_r} = 16.2 \ \text{cm}/\sqrt{20} = 3.62 \ \text{cm}$.

(b)
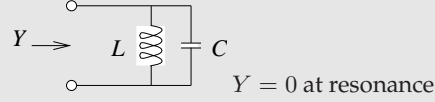
$$Y = Y_L + Y_C = \frac{1}{\jmath \omega L} + \jmath \omega C = \frac{\omega^{-1}}{\jmath L} + \jmath \omega C.$$

(c) $\ell = (0.0362 \text{ m})/4 = 9.05$ mm.

(d) From (b),

$$\frac{\partial Y}{\partial \omega} = -\frac{\omega^{-2}}{\jmath L} + \jmath C = \frac{\jmath}{\omega^2 L} + \jmath C = \jmath \left(\frac{1}{\omega^2 L} + C \right). \tag{4.131}$$

(e) From Equation (4.65), $Z_0 = 12.28 \ \Omega$.

(f) The input impedance of the line can be calculated from Equation (4.128), repeated below:

$$Z_{\text{in}} = Z_0 \frac{Z_L + \jmath Z_0 \tan \beta \ell}{Z_0 + \jmath Z_L \tan \beta \ell}. \tag{4.132}$$

In this problem $Z_L = 0 \ \Omega$, $\ell = (0.0362 \text{ m})/4 = 9.05$ mm, $\beta = \beta_0 \sqrt{\varepsilon_r}$, where β_0 is the phase constant of free space. From Equations (4.56) and (4.57),

$$\beta = 20.958 \times f_0|_{(GHz)} \times \sqrt{20} = 173.4 \text{ rad/m}$$

At $0.99 f_0$, $Z_{\text{in}} = \jmath 781.7 \ \Omega$; $Y_{\text{in}} = -\jmath 0.001279$ S.
At f_0, $Z_{\text{in}} = -\jmath \infty \ \Omega$; $Y_{\text{in}} = 0$ S.
At $1.01 f_0$, $Z_{\text{in}} = -\jmath 781.7 \ \Omega$; $Y_{\text{in}} = \jmath 0.001279$ S.
So the derivative of the input admittance is

$$\frac{\partial Y}{\partial \omega} \approx \frac{\Delta Y_{\text{in}}}{\Delta \omega} = \frac{\jmath (0.001279 + 0.001279)}{2\pi f_0 (1.01 - 0.99)} = \jmath 1.101 \cdot 10^{-11}. \tag{4.133}$$

(g) The equivalent circuit resonates at $f_0 = 1.85$ GHz when $1/(\jmath \omega_0 L) = -\jmath \omega_0 C$; $\omega_0 = 2\pi f_0$. That is,

$$LC = \omega_0^{-2} = 0.7401 \times 10^{-20} \text{ s}^{-2}. \tag{4.134}$$

Equating Equations (4.131) and (4.133),

$$\frac{1}{\omega_0^2 L} + C = 1.101 \times 10^{-11} \text{ F}$$

$$C = 1.101 \times 10^{-11} - \frac{1}{\omega_0^2 L} \text{ F}. \tag{4.135}$$

Substituting this in Equation (4.134)

$$L \left(1.101 \times 10^{-11} - \frac{1}{\omega_0^2 L} \right) = 0.7401 \times 10^{-20} \text{ H}$$

$$1.101 \times 10^{-11} L = 1.4802 \times 10^{-20} \text{ H}$$

$$L = 1.345 \times 10^{-9} \text{ H} = 1.345 \text{ nH}. \tag{4.136}$$

SHORT–CIRCUITED STUB

OPEN–CIRCUITED STUB
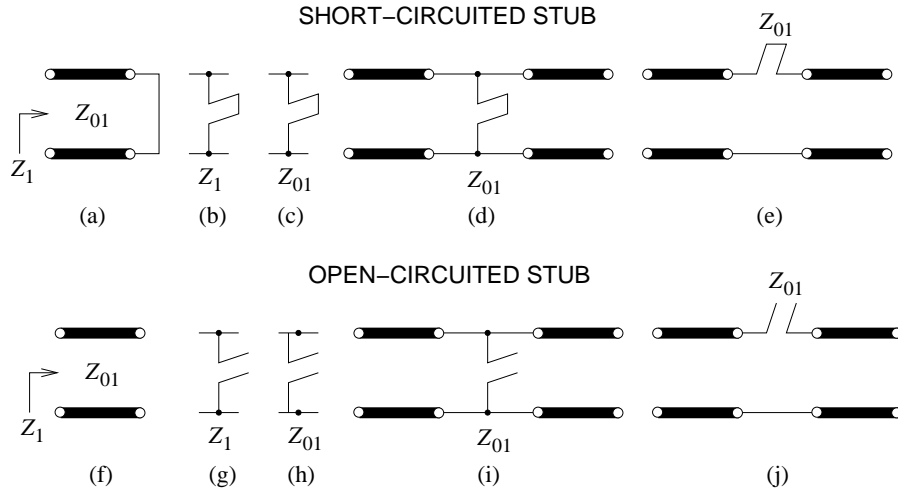
Figure 4-19   Transmission line stubs: (a)–(e) short-circuited stubs; and (f)–(j) open-circuited stubs.

From Equation (4.134),

$$C = 0.7401 \times 10^{-20}/L = 5.503 \times 10^{-12} \text{ F} = 5.503 \text{ pF}. \tag{4.137}$$

So the equivalent circuit of the resonator is
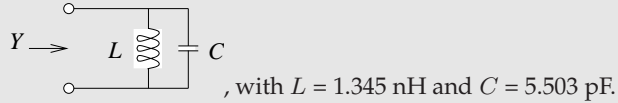
, with $L = 1.345$ nH and $C = 5.503$ pF.
This equivalent circuit is valid for a narrow range (say 5%) of frequencies around 1.85 GHz.

4.7.5   Short-Circuited Stub

A short-circuited transmission line is commonly used as a circuit element that is called a stub. Figures 4-19(a)–(e) show short-circuited stubs. In Figure 4-19(a), and noting that in Equation (4.128) $Z_L = 0$, the transmission line has an input impedance

$$Z_1 = \jmath Z_{01} \tan \beta \ell . \tag{4.138}$$

With the stub one-quarter wavelength long at a frequency f_r, the input impedance is an open circuit and the stub is said to be resonant at a frequency f_r. Then, at a frequency f, the input impedance of the stub is

$$Z_1 = \jmath Z_{01} \tan \left(\frac{\pi}{2} \frac{f}{f_r} \right) . \tag{4.139}$$

When $f = \frac{1}{2}f_r$, the stub is one-eighth of a wavelength long and the argument of the $\tan$ function in Equation (4.139) is $\pi/4$ and Z_1 becomes

$$Z_1 = \jmath Z_{01} \tan\left(\frac{\pi}{4}\right) = Z_1 = \jmath Z_{01} \; . \tag{4.140}$$

Figure 4-19(b) is a shunt short-circuited stub with input impedance Z_1. The representation of the stub in Figure 4-19(c) indicates the characteristic impedance of the stub, Z_{01}, and in this representation it is necessary to know the resonant frequency (at which the stub is one-quarter wavelength long) of the stub. If Z_1 or Z_{01} are replaced by numbers, a real number (such as 10 Ω) would indicate the characteristic impedance of the stub, but an imaginary number (such as $\jmath 10$ Ω) would indicate the input impedance of the stub. If the characteristic impedance of the stub is given, then the length of the stub must be indicated by the context, typically by its resonant frequency, f_r, the frequency at which the stub is one-quarter wavelength long. Figure 4-19(d) shows a short-circuited stub inserted between two transmission lines, while Figure 4-19(e) shows the stub in series connection.

4.7.6 Open-Circuited Stub

An open-circuited transmission line is commonly used as a circuit element that is called an open stub. Figure 4-19(f–j) show open-circuited stubs. In Figure 4-19(f), and noting that in Equation (4.128) $Z_L = \infty$, the transmission line has an input impedance of

$$Z_1 = -\jmath Z_{01} \frac{1}{\tan \beta \ell} \; . \tag{4.141}$$

With the stub one-quarter wavelength long at a frequency f_r, the input impedance is a short circuit and the stub is said to be resonant at a frequency f_r. Then at a frequency f, the input impedance of the stub is

$$Z_1 = -\jmath Z_{01} \frac{1}{\tan\left(\frac{\pi}{2}\frac{f}{f_r}\right)} \; . \tag{4.142}$$

Figure 4-19(g) is a representation of a shunt open-circuited stub with input impedance Z_1. The representation of the stub in Figure 4-19(h) indicates the characteristic impedance of the stub, Z_{01}, and in this representation it is necessary to know the resonant frequency (at which the stub is one-quarter wavelength long) of the stub. Figure 4-19(i) shows an open-circuited stub inserted between two transmission lines, while Figure 4-19(j) shows the stub in series connection.

The stub introduces an impedance in shunt across a transmission line. This can be used in tuning. A common microstrip trick is to allow for optimization of a design using the layout shown in Figure 4-20, where bonding to different pads enables a variable length stub to be realized.

Figure 4-20   Open-circuited stub with variable length realized using wire bonding from the fixed stub to one of the bond pads.
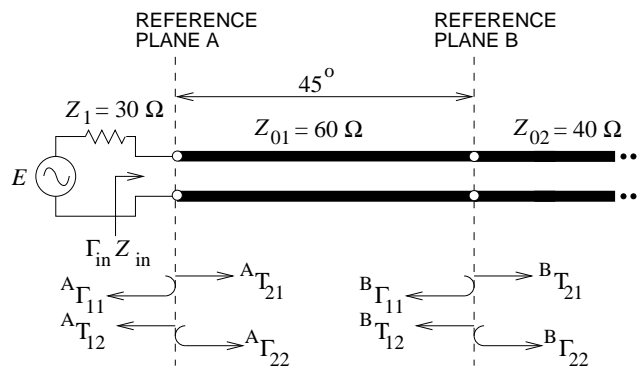

Figure 4-21   Transmission line network with a finite length line of characteristic impedance Z_{01} and an infinitely long transmission line of characteristic impedance Z_{02}.

4.7.7   Bounce Diagram

The bounce diagram follows reflections at the interfaces between networks. The utility of bounce diagrams will be addressed using an example.

EXAMPLE 4. 16   Bounce Diagram

In the transmission line system in Figure 4-21 there are two planes at which reflection and transmission (or scattering) occur. The electrical length of the Z_{01} line is $45°$ and the Z_{02} line is infinitely long. The aim here is to find the input impedance, Z_{in}. This will be arrived at in two ways. The first technique uses a bounce diagram approach and emphasizes reflection and transmission at the interface planes. The second approach uses the telegraphist's equation. It is more subtle to use the bounce diagram technique, but it yields the same answer as using the telegraphist's equation.

(a) What are the scattering parameters at reference plane A?

$^{A}\Gamma_{11}$ is the reflection coefficient of signals incident on reference plane A from the left. $^{A}T_{21}$ is the transmission coefficient at the plane for signals from the left. $^{A}\Gamma_{22}$ and $^{A}T_{12}$ are the corresponding parameters for scattering from signals coming from the right to the plane. So, normalizing to Z_1,

$$^A\Gamma_{11} = \frac{Z_{01} - Z_1}{Z_{01} + Z_1} = 0.333; \qquad\qquad ^A\Gamma_{22} = -^A\Gamma_{11} = -0.333 \qquad (4.143)$$

$$^AT_{21} = 1 - {^A\Gamma_{11}} = 0.667; \qquad\qquad ^AT_{12} = 1 - {^A\Gamma_{22}} = 1.333 \qquad (4.144)$$

While $^AT_{12}$ is greater than one, this does not indicate power gain because of the differences in impedance levels.

(b) What are the scattering parameters Γ_2 and T_2 at reference plane B? Now the reference is Z_0, and so

$$^B\Gamma_{11} = \frac{Z_{02} - Z_0}{Z_{02} + Z_0} . \qquad (4.145)$$

The question now is what should the system reference impedance be—should it be Z_1, Z_{01}, or even Z_{02}? The problem could be solved using any of these, but the simplest procedure is to use the same reference impedance throughout, and since the eventual aim is to calculate the overall input reflection coefficient, the appropriate choice is $Z_0 = Z_1$. Note, however, that the actual voltage levels on the lines are not being calculated (which would need to be referenced to the characteristic impedance of the lines they are on), but instead to a traveling wave referenced to a universal system impedance. If it were required that the actual voltages be calculated, then a system reference impedance change would be necessary. So the scattering parameters at reference plane B referenced to the impedance Z_1 are

$$^B\Gamma_{11} = \frac{Z_{02} - Z_1}{Z_{02} + Z_1} = 0.143; \qquad\qquad ^B\Gamma_{22} = -^B\Gamma_{11} = -0.143 \qquad (4.146)$$

$$^BT_{21} = 1 - {^A\Gamma_{11}} = 0.857; \qquad\qquad ^BT_{12} = 1 - {^B\Gamma_{22}} = 1.143 . \qquad (4.147)$$

The second transmission line is infinitely long and so no signal from the line will be incident at reference plane B from the right.

(c) What is the transmission coefficient of the Z_{01} transmission line?

Using a reference impedance of Z_{01}, the transmission coefficient is one and rotates by $45°$ or 0.785 radians:

$$T_{01} = \exp(\jmath 0.785) = 0.707 + \jmath 0.707 .$$

(d) Draw the bounce diagram of the transmission line network.

The bounce diagram is shown in Figure 4-22.

(e) What is Γ_{in} and hence what is Z_{in}? Γ_{in} is the steady-state input reflection coefficient and is obtained by adding all of the reflections leaving to the left from plane A in Figure 4-22. So

$$\begin{aligned}
\Gamma_{\text{in}} &= {^A\Gamma_{11}} + {^AT_{12}}{^AT_{21}}T_{01}^2{}^B\Gamma_{11} \\
&\quad + {^AT_{12}}{^AT_{21}}T_{01}^4{}^B\Gamma_{11}^2{}^A\Gamma_{22} + {^AT_{12}}{^AT_{21}}T_{01}^6{}^B\Gamma_{11}^2{}^A\Gamma_{22}^2 \cdots \\
&= {^A\Gamma_{11}} + {^AT_{12}}{^AT_{21}}T_{01}^2{}^B\Gamma_{11}\left[1 + x + x^2 + \cdots\right], \qquad (4.148)
\end{aligned}$$

where $x = T_{01}^2{}^B\Gamma_{11}{}^A\Gamma_{22}$. Now $1/(1-x) = 1 + x + x^2 + \cdots$, and so

$$\begin{aligned}
\Gamma_{\text{in}} &= {}^A\Gamma_{11} + \frac{{}^A T_{12}\,{}^A T_{21}\,T_{01}^2\,{}^B\Gamma_{11}}{1 - T_{01}{}^{2\,B}\Gamma_{11}\,{}^A\Gamma_{22}} \\
&= 0.333 + \frac{1.333 \times 0.667 \times (0.707 - \jmath 0.707)^2 \times (-0.2)}{1 - (0.707 - \jmath 0.707)^2 \times (-0.2) \times (-0.333)} \\
&= 0.345 - \jmath 0.177 \,. \hspace{4cm} \text{(4.149)}
\end{aligned}$$

Γ_{in} is the reflection at reference plane A and is referenced to $Z_1 = 30\,\Omega$. So the input impedance is

$$Z_{\text{in}} = Z_1 \left(\frac{1 + \Gamma_{\text{in}}}{1 - \Gamma_{\text{in}}}\right) = 30\left(\frac{1 + 0.345 - \jmath 0.177}{1 - 0.345 - \jmath 0.177}\right) = 55.39 + \jmath 23.08\,\Omega. \quad \text{(4.150)}$$

(f) Use the lossless telegraphist's equation, Equation (4.128), to find Z_{in}.

The infinitely long line presents an impedance Z_{02} to the 60 Ω transmission line. So the input impedance looking into the 60 Ω line at reference plane A is, using the lossless telegraphist's equation,

$$Z_{\text{in}} = Z_{01}\left(\frac{Z_{02} + \jmath Z_{01}\tan\beta l}{Z_{01} + \jmath Z_{02}\tan\beta l}\right),$$

where the electrical length βl is 45° or $\pi/8$ radians. So

$$Z_{\text{in}} = 60\left(\frac{40 + \jmath 60\tan(\pi/8)}{60 + \jmath 40\tan(\pi/8)}\right) = 55.39 + \jmath 23.08\,\Omega,$$

which is the same result obtained using the bounce diagram method.

In the above, two techniques were used to evaluate the input impedance of a transmission line system. The bounce diagram technique required a subtle choice of several different reference impedances which were required to keep track of real power flow. It is quite difficult to make these choices. It can be viewed that the bounce diagram considers instantaneous reflections of pulses. These are summed to yield the steady-state input reflection coefficient, and thus the input impedance. The telegraphist's equation directly captures the steady-state response. So while the bounce diagram technique aids in physical understanding, using the telegraphist's equation is a much less error prone approach to solving transmission line problems.

4.8 Special Cases of Lossless Terminated Lines

A number of special cases of lossless terminated transmission lines are useful circuit elements in RF design.
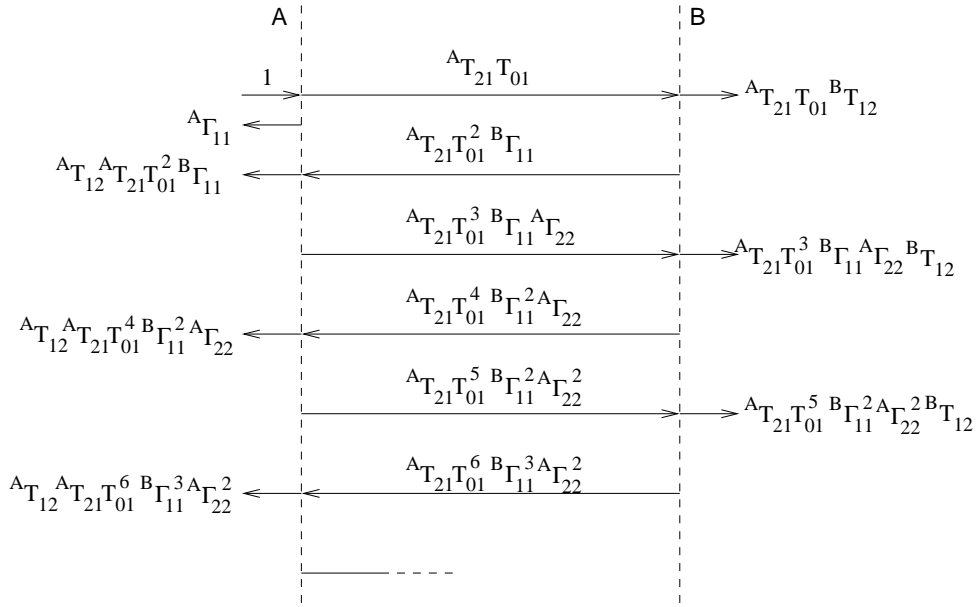
Figure 4-22   Bounce diagram.

4.8.1   *Electrically Short Lossless Line*

Now consider the case of electrically short lines, that is, $\beta\ell$ goes to zero (i.e., $2\pi\ell/\lambda$ goes to zero), then Equation (4.128) becomes

$$Z_{\text{in}} \approx \frac{Z_L + \jmath Z_0(\beta\ell)}{1 + \jmath \frac{Z_L}{Z_0}(\beta\ell)} \approx [Z_L + \jmath Z_0(\beta\ell)]\left[1 - \jmath\frac{Z_L}{Z_0}(\beta\ell)\right].$$

This can be written as

$$Z_{\text{in}} \approx Z_L\left[1 + (\beta\ell)^2\right] + \jmath\left[\omega(L\ell) - Z_L^2\omega(C\ell)\right].$$

Thus the input impedance of an electrically short line terminated in impedance Z_L is

$$Z_{\text{in}} \approx Z_L + \jmath\left[\omega(L\ell) - Z_L^2\,\omega(C\ell)\right]. \tag{4.151}$$

Some special cases of this result will be considered next.

4.8.2   *Short Length of Short-Circuited Line*

A transmission line terminated in a short circuit ($Z_L = 0$) has the input impedance

$$Z_{\text{in}} = \jmath Z_0 \tan(\beta\ell). \tag{4.152}$$

So a short length of line, $\ell < \lambda_g/4$, looks like an inductor with inductance L_s,

$$Z_0 \tan(\beta\ell) = \omega L_s, \tag{4.153}$$

and so

$$L_s = \frac{Z_0}{\omega} \tan \frac{2\pi\ell}{\lambda_g} . \tag{4.154}$$

From Equation (4.154) it is seen that for a given ℓ, L_s is proportional to Z_0. Hence, for larger values of L_s, sections of transmission line of high characteristic impedance are needed. So microstrip lines with narrow strips can be used to realize inductors in planar microstrip circuits.

EXAMPLE 4. 17   Inductive Transmission Line

This example demonstrates that one does not need an ideal short to effect a (predominantly) inductive behavior from a segment of transmission line. Consider the transmission line system shown below with lines having two different characteristic impedances, Z_{01} and Z_{02}.
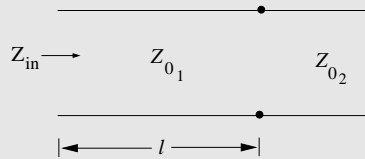
The value of Z_{in} is

$$Z_{\text{in}} = Z_{01} \frac{Z_{02} + \jmath Z_{01} \tan \beta\ell}{Z_{02} + \jmath Z_{01} \tan \beta\ell} ,$$

which for a short line can be expressed as

$$Z_{\text{in}} \approx Z_{02} \left(1 + \tan^2(\beta l)\right) + \jmath Z_{01} \tan(\beta\ell) .$$

Note that $\jmath Z_{01} \tan(\beta\ell)$ is the dominant part for $l < \lambda/8$ and $Z_{02} \ll Z_{01}$.

From this example it can be concluded that a microstrip realization of a series inductor can be a high-impedance line embedded between two low-impedance lines. A top view of such a configuration in microstrip is shown in the figure below. Recall that a narrow microstrip line has high characteristic impedance.

(a)   (b)
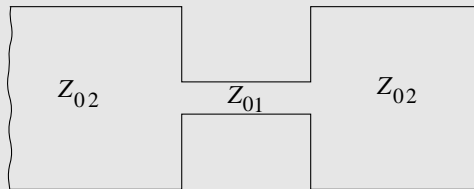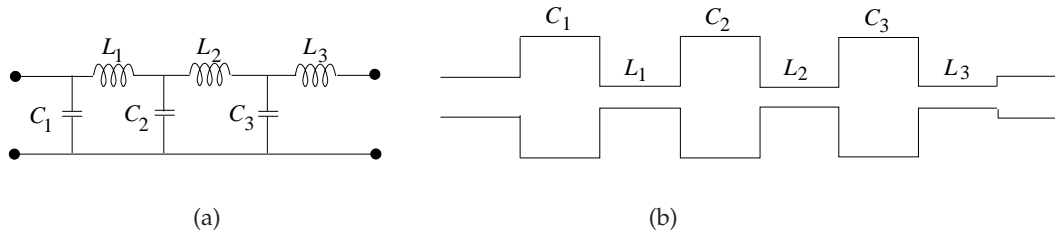
Figure 4-23   A lowpass filter: (a) in the form of an LC ladder network; and (b) realized using microstrip lines.

4.8.3   Short Length of Open-Circuited Line

An open-circuited line has $Z_L = \infty$ and so

$$Z_{\text{in}} = -\jmath \frac{Z_0}{\tan \beta \ell} \ . \tag{4.155}$$

For lengths ℓ such that $\ell < \lambda/4$, an open-circuited segment of line realizes a capacitor C_0 for which

$$\frac{1}{\omega C_0} = \frac{Z_0}{\tan \beta \ell} \tag{4.156}$$

and so

$$C_0 = \frac{1}{Z_0} \frac{\tan \beta \ell}{\omega} \ . \tag{4.157}$$

From the above relationship, we see that for a given length of line ℓ, C_0 is inversely proportional to Z_0. Hence for larger values of C_0 we need to use sections of transmission line of low characteristic impedance.

EXAMPLE 4. 18   Transmission Line Filter

A microstrip realization of a lowpass filter consists of Ls and Cs as shown in Figure 4-23(a). This lowpass filter section can be realized using wide and narrow microstrip lines, as shown in Figure 4-23(b).

4.8.4   Quarter-Wave Transformer

The next important case of a transmission line section is the quarter-wave transformer. Figure 4-24(a) shows a resistive load R_L and a section of transmission line whose length ℓ is $\lambda_g/4$ (hence the name quarter-wave transformer). The input impedance of the line is

$$Z_{\text{in}} = Z_1 \frac{R_L + \jmath Z_1 \tan(\beta \ell)}{R_L + \jmath Z_1 \tan(\beta \ell)} = Z_1 \frac{R_L + \jmath Z_1 \infty}{R_L + \jmath Z_1 \infty} = \frac{Z_1^2}{R_L} \ . \tag{4.158}$$
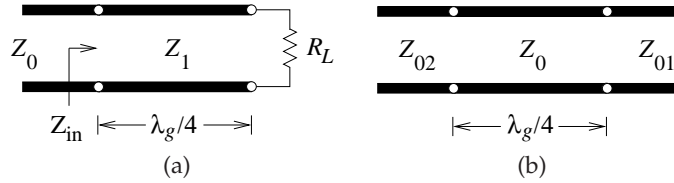
Figure 4-24   The quarter-wave transformer line: (a) interfacing two lines; and (b) transforming a load.

So the input impedance is matched to the transmission line Z_0 if

$$Z_{\text{in}} = Z_0; \tag{4.159}$$

that is,

$$Z_1 = \sqrt{Z_0 R_L} \, . \tag{4.160}$$

Thus the quarter-wavelength long line acts as an ideal impedance transformer.

Another example of the quarter-wave transformer is shown in Figure 4-24(b). It is clear from the figure that the input impedance looking into the quarter-wave transformer is given by

$$Z_{\text{in}} = Z_0 \frac{Z_{01} + \jmath Z_0 \tan(\beta\ell)}{Z_0 + \jmath Z_{01} \tan(\beta\ell)} = Z_0 \frac{Z_{01} + \jmath Z_0 \infty}{Z_0 + \jmath Z_{01} \infty} = \frac{Z_0^2}{Z_{01}} \, . \tag{4.161}$$

Hence a section of the transmission line of length $\ell = \lambda_g/4 + n\lambda_g/2$, where $n = 0, 1, 2, \ldots$, can be used to match two different impedances Z_{01}, Z_{02} by constructing the line such that its characteristic impedance Z_0 is

$$Z_0 = \sqrt{Z_{01} Z_{02}} \, . \tag{4.162}$$

Note that for a design center frequency f_0, the electrical length of the matching section is $\lambda_g/4$, but at different frequencies the electrical length is different and a perfect match is no longer achieved. In general, a perfect match is obtained only at the frequencies at which $\ell = \lambda_g/4 + n\lambda_g/4$.

There is an interesting property of a quarter-wave transformer that is widely used in filters. Examine the final result in Equation (4.161) which is repeated here:

$$Z_{\text{in}} = \frac{Z_0^2}{Z_{01}} \, . \tag{4.163}$$

This result applies to complex impedances as well. Equation (4.163) indicates that a quarter-wavelength long line is an impedance inverter, presenting, at Port 1 the inverse of the impedance presented at Port 2, Z_{01}. This inversion is normalized by the square of the characteristic impedance
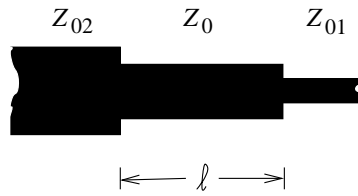
Figure 4-25   Layout of a microstrip quarter-wave transformer.
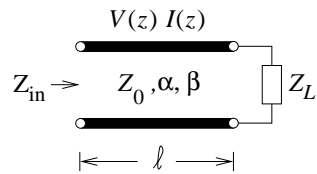
Figure 4-26   A low-loss transmission line.

of the line. This inversion holds in the reverse direction as well. The quarter-wave transformer is also called an **impedance inverter** (or equivalently an admittance inverter).

The layout of a microstrip quarter-wave transformer is shown in Figure 4-25, where $\ell = \lambda_g/4$ and the characteristic impedance of the transformer, Z_0, is the geometric mean of the impedances on either side, that is $Z_0 = \sqrt{Z_{0,1}Z_{0,2}}$.

4.9   Input Impedance of a Terminated Lossy Line

In this section a lossy transmission line with low-loss is considered so that $R \ll \omega L$ and $G \ll \omega C$, and the characteristic impedance is $Z_0 \approx \sqrt{L/C}$. Figure 4-26 is a lossy transmission line whose attenuation constant is α and the phase constant is β. The propagation constant, γ, is a complex function of both α and β. The total voltage and current at any point on the line are given by

$$V(z) = V_0^+ \left[e^{-\gamma z} + \Gamma e^{\gamma z} \right] \tag{4.164}$$

and

$$I(z) = \frac{V_0^+}{Z_0} \left[e^{-\gamma z} - \Gamma e^{\gamma z} \right] . \tag{4.165}$$

In a similar fashion, the impedance seen looking into the line for this case is given by the lossy telegraphist's equation:

$$Z_{\text{in}} = Z_0 \frac{Z_L + Z_0 \tanh \gamma \ell}{Z_0 + Z_L \tanh \gamma \ell} . \tag{4.166}$$

For a lossless line this becomes

$$Z_{\text{in}} = Z_0 \frac{Z_L + \jmath Z_0 \tan \beta \ell}{Z_0 + \jmath Z_L \tan \beta \ell} \qquad (4.167)$$

as $\gamma = \alpha + \jmath \beta$ and with $\alpha = 0$, $\tanh \gamma \ell$ becomes $\tanh \jmath \beta \ell = \jmath \tan \beta \ell$.

For a lossy transmission line it is clear that not all of the power applied at the input will be delivered to the load. There will be a certain amount of power lost on the line due to the finite attenuation constant that results from loss on the line. Equations (4.168)–(4.170) are the equations of power delivered to the load, the power incident on the line, and the power lost on the line, respectively. So the power delivered to the load is

$$P_L = \frac{1}{2} \mathcal{R} \left\{ V(0) I^*(0) \right\} = \frac{\left| V_0^+ \right|^2}{2 Z_0} (1 - |\Gamma|^2). \qquad (4.168)$$

Similarly the power at the input of the line is

$$P_{\text{in}} = \frac{1}{2} Re \left\{ V(-\ell) I^*(-l) \right\} = \frac{\left| V_0^+ \right|^2}{2 Z_0} \left[1 - |\Gamma|^2 e^{-4\alpha\ell} \right] e^{2\alpha\ell} \qquad (4.169)$$

and the power lost in the line is

$$P_{\text{loss}} = P_{\text{in}} - P_L = \frac{\left| V_0^+ \right|^2}{2 Z_0} \left[(e^{2\alpha\ell} - 1) + |\Gamma|^2 (1 - e^{-2\alpha\ell}) \right]. \qquad (4.170)$$

4.10 Microstrip Transmission Lines

Transmission lines with conductors embedded in an inhomogeneous dielectric medium cannot support a pure TEM mode. This is the case even if the conductors are lossless. The most important member of this class is the microstrip transmission line (Figure 4-8(c)). Part of the field is in the air and part in the dielectric between the strip conductor and ground. In most practical cases, the dielectric substrate is *electrically thin*, that is, $h \ll \lambda$. It can be shown that the transverse field is dominant, $|E_L| \ll |E_T|$, and the fields are called quasi-TEM.

4.10.1 Microstrip Line in the Quasi-TEM Approximation

In this section it is shown that the static solutions for the transverse electric field alone can be used to calculate the characteristics of a transmission line. The procedure described is used in many EM computer programs to calculate the characteristics of transmission lines.

As a first step, the potential of the conductor strip is set to V_0 and Laplace's equation is solved for the electrostatic potential everywhere in the dielectric. Then, the p.u.l. electric charge, Q, on the conductor is determined. Using this

in the following relation gives the line capacitance:

$$C = \frac{Q}{V_0}.$$

In the next step, the process described in the step above is repeated with $\epsilon_r = 1$. This is done to calculate C_{air} (the capacitance of the line without a dielectric). Since the transmission line is now a TEM structure,

$$LC_{\text{air}} = \mu_0\epsilon_0,$$

and so the velocity on a vacuum-filled line with a TEM field distribution is identical to that in free space, where $c = v_p = 1/\sqrt{\mu_0\varepsilon_0}$. Thus

$$L = \frac{1}{c^2 C_{\text{air}}} . \tag{4.171}$$

L is not affected by the dielectric properties of the medium.[14] L calculated above is the desired p.u.l inductance of the line with the dielectric as well as in free space. Once L and C have been found, the characteristic impedance can be found using

$$Z_0 = \sqrt{\frac{L}{C}} , \tag{4.172}$$

which can be written as

$$Z_0 = \frac{1}{c}\frac{1}{\sqrt{C\,C_{\text{air}}}} , \tag{4.173}$$

and the phase velocity is given by

$$v_p = \frac{1}{\sqrt{LC}} = c\sqrt{\frac{C_{\text{air}}}{C}} . \tag{4.174}$$

It can be seen that the field is distributed in the inhomogeneous medium and in free space, as shown in Figure 4-27. So the effective permittivity, ϵ_e, of the microstrip line is defined by

$$\sqrt{\epsilon_e} = \frac{c}{v_p} . \tag{4.175}$$

Combining Equations (4.174) and (4.175), the effective relative permittivity is obtained:

$$\epsilon_e = \frac{C}{C_{\text{air}}} . \tag{4.176}$$

---

[14] The assumption that L is not affected by the dielectric is in general a good approximation. However, there is a small discrepancy due to a change in the electric field orientation which will also affect the distribution of the magnetic field, but there is negligible additional magnetic energy storage.
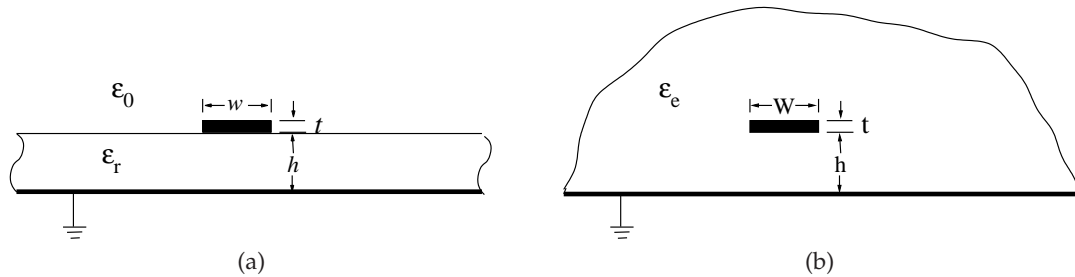
Figure 4-27   Microstrip line: (a) actual crosssection; and (b) approximate equivalent structure where the strip is embedded in a dielectric of semi-infinite extent with effective relative permittivity ε_e.

The effective dielectric constant can be interpreted as the dielectric constant of a homogeneous medium that replaces the air and the dielectric regions of the microstrip, as shown in Figure 4-27. Since some of the field is in the dielectric and some is in air, the effective dielectric constant satisfies the relation

$$1 < \epsilon_e < \epsilon_r. \tag{4.177}$$

4.10.2   Effective Dielectric Constant and Characteristic Impedance

This section presents formulas for the effective dielectric constant and characteristic impedance of the microstrip. These formulas are fits to the results of detailed EM simulations. Also, the form of the equations is based on good physical understanding. First, assume that the thickness, t, is zero. This is not a bad approximation, as in practice $t \ll w, h$ for most microwave circuits (this, however, is not true with digital ICs). It is not possible to develop useful formulas for the situation where the lines are not very thin.

Hammerstad and others provided well-accepted formulas for calculating the effective permittivity and characteristic impedance of microstrip lines [66–68]. Given ε_r, w, and h, the effective relative permittivity is

$$\epsilon_e = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2}\left(1 + \frac{10h}{w}\right)^{-a \cdot b}, \tag{4.178}$$

where

$$a(u)|_{u=w/h} = 1 + \frac{1}{49}\ln\left[\frac{u^4 + \{u/52\}^2}{u^4 + 0.432}\right] + \frac{1}{18.7}\ln\left[1 + \left(\frac{u}{18.1}\right)^3\right] \tag{4.179}$$

and

$$b(\epsilon_r) = 0.564\left[\frac{\epsilon_r - 0.9}{\epsilon_r + 3}\right]^{0.053}. \tag{4.180}$$

Take some time to interpret Equation (4.178) for effective relative permittivity. First, assume for simplicity that the dimensions of the line are such that $a \cdot b = 1$. If $\varepsilon_r = 1$, then $\varepsilon_e = (1+1)/2 + 0 = 1$, as expected. If ε_r is not that of air then ε_e will be between 1 and ε_r, dependent on the geometry of the line, or more specifically, the ratio w/h. For a very wide line, $w/h \gg 1$, $\varepsilon_e = (\varepsilon_r + 1)/2 + (\varepsilon_r - 1)/2 = \varepsilon_r$, corresponding to the EM energy being confined to the dielectric. For a thin line, $w/h \ll 1$, so $\varepsilon_e = (\varepsilon_r + 1)/2$, the average of the dielectric and air regions. Commonly the shorter-term effective permittivity is used, but nearly always the effective relative permittivity is meant.

The characteristic impedance is given by

$$Z_0 = \frac{Z_{01}}{\sqrt{\epsilon_e}} \, , \tag{4.181}$$

where the characteristic impedance of the micostrip line in free space is

$$Z_{01} = Z_0|_{(\epsilon_r = 1)} = 60 \ln \left[\frac{F_1 h}{w} + \sqrt{1 + \left(\frac{2h}{w} \right)^2} \right] , \tag{4.182}$$

and

$$F_1 = 6 + (2\pi - 6) \exp \left\{ -(30.666 h/w)^{0.7528} \right\} . \tag{4.183}$$

The accuracy of Equation (4.178) is better than 0.2% for $0.01 \leq w/h \leq 100$ and $1 \leq \epsilon_r \leq 128$. Also, the accuracy of Equation (4.182) is better than 0.1% for $w/h < 1000$. $Z_0(\epsilon_r = 1)$ has a maximum value when w is vanishingly small. This is not seen in Equation (4.182), as this is only valid over a range of w.

Now consider the special case where w is vanishingly small. Then $Z_0(\epsilon_r = 1)$ is that of free space; that is, $Z_0(\epsilon_r = 1) = \sqrt{\mu_0 \varepsilon_0} = 377 \, \Omega$. As well, ϵ_e has its minimum value when w is vanishingly small, then

$$\epsilon_e = \frac{\epsilon_r + 1}{2} \, . \tag{4.184}$$

This leads to an approximate (and more convenient) form of Equation (4.178):

$$\epsilon_e = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2} \frac{1}{\sqrt{1 + 12h/w}} \, . \tag{4.185}$$

This approximation has its greatest error for low and high ϵ_r and narrow lines, $w/h \ll 1$, where the maximum error is 1%. Again, Equation (4.181) would be used to calculate the characteristic impedance. A comparison of the line characteristics using the more precise formula for ϵ_e (Equation (4.178)) and the slightly less accurate fit (Equation (4.185)) is given in Table 4-2.

Table 4-2 Comparison of two fitted equations for the effective permittivity and characteristic impedance of a microstrip line. Equation (4.178) has an accuracy of better than 0.2% and Equation (4.185) has an accuracy of better than 1%.

ϵ_r	w/h	ϵ_e	Z_0	ϵ_e	Z_0
		Eq. (4.178), 0.2%		Eq. (4.185), 1%	
1	0.01	1	401.1	1	401.1
1	0.1	1	262.9	1	262.9
1	1	1	126.5	1	126.5
1	10	1	29.04	1	29.04
1	100	1	3.61	1	3.61
2	0.01	1.525	322.5	1.514	325.9
2	0.1	1.565	210.0	1.545	211.5
2	1	1.645	98.64	1.639	98.83
2	10	1.848	21.36	1.837	21.43
2	100	1.969	2.58	1.972	2.58
10	0.01	5.683	165.8	5.630	169.0
10	0.1	6.016	107.0	5.909	108.2
10	1	6.705	48.86	6.748	48.70
10	10	8.556	9.93	8.534	9.94
10	100	9.707	1.16	9.752	1.16
20	0.01	10.88	119.6	10.77	122.2
20	0.1	11.57	77.14	11.36	78.00
20	1	13.01	35.07	13.13	34.91
20	10	16.93	7.06	16.91	7.06
20	100	19.38	0.821	19.48	0.819
128	0.01	66.90	48.15	66.30	49.25
128	0.1	71.51	31.02	70.27	31.37
128	1	81.12	14.05	82.11	13.96
128	10	71.51	2.80	70.27	2.80
128	100	123.8	0.325	124.5	0.324

The more exact analysis, represented by Equation (4.178), was used to develop Table 4-3 which can be used in the design of microstrip on a silicon dioxide (SiO_2) substrate and on FR-4 printed circuit board which both have a relative permittivity of 4; on alumina with a permittivity of 10; on a silicon (Si) substrate with a relative permittivity of 11.9; and on a gallium arsenide (GaAs) substrate with a relative permittivity of 12.85.

EXAMPLE 4. 19   Microstrip Calculations

The strip of a microstrip has a width of 600 μm and is fabricated on a lossless substrate that is 635 μm thick and has a relative permittivity of 4.1.

(a) What is the effective relative permittivity?

(b) What is the characteristic impedance?

(c) What is the propagation constant at 5 GHz ignoring any losses?

Solution:
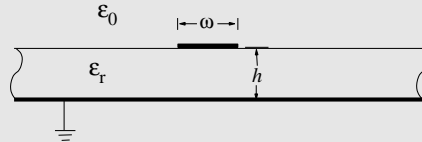
Use the formulas for effective dielectric constant, characteristic impedance, and attenuation constant from Section 4.10.2 with $w = 600\ \mu\text{m}$; $h = 635\ \mu\text{m}$; $\varepsilon_r = 4.1$; $w/h = 600/635 = 0.945$.

(a)
$$\varepsilon_e = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2}\left(1 + \frac{10h}{w}\right)^{-a \cdot b}$$

From Equation (4.179),

$$a = 1 + \frac{1}{49}\ln\left[\frac{(w/h)^4 + \{w/(52h)\}^2}{(w/h)^4 + 0.432}\right] + \frac{1}{18.7}\ln\left[1 + \left(\frac{w}{18.1h}\right)^3\right] = 0.991,$$

and from Equation (4.180),

$$b = 0.564\left[\frac{\varepsilon_r - 0.9}{\varepsilon_r + 3}\right]^{0.053} = 0.541\ .$$

From Equation (4.178),
$$\varepsilon_e = 2.967\ .$$

(b) In free space,
$$Z_0\,(\varepsilon_r = 1) = 60\ln\left[\frac{F_1 \cdot h}{w} + \sqrt{1 + \left(\frac{2h}{w}\right)^2}\right],$$

where

$$F_1 = 6 + (2\pi - 6)\exp\left\{-(30.666h/w)^{0.7528}\right\}, \quad Z_0 = \frac{Z_0\,(\varepsilon_r = 1)}{\sqrt{\varepsilon_e}}$$

$$Z_0\,(\text{air}) = 129.7\ \Omega, \quad Z_0 = Z_0\,(\text{air})/\sqrt{\varepsilon_e} = 75.3\ \Omega$$

(c) $f = 5\ \text{GHz}$, $\omega = 2\pi f$, $\gamma = \jmath\omega\sqrt{\mu_0\varepsilon_0\varepsilon_e}$:

$$\gamma = \jmath 180.5\ \text{m}^{-1}\ .$$

Table 4-3 Microstrip line normalized width u $(= w/h)$, and effective permittivity, ε_e, for specified characteristic impedance Z_0. Data derived from the analysis in Section 4.10.2 on Page 223.

Z_0 (Ω)	$\varepsilon_r = 4$ (SiO$_2$, FR-4)		$\varepsilon_r = 10$ (Alumina)		$\varepsilon_r = 11.9$ (Si)		Z_0 (Ω)	$\varepsilon_r = 4$ (SiO$_2$, FR-4)		$\varepsilon_r = 10$ (Alumina)		$\varepsilon_r = 11.9$ (Si)	
	u	ε_e	u	ε_e	u	ε_e		u	ε_e	u	ε_e	u	ε_e
140	0.171	2.718	0.028	5.914	0.017	6.907	110	0.378	2.781	0.089	6.025	0.062	7.032
139	0.176	2.720	0.029	5.917	0.018	6.910	109	0.389	2.783	0.093	6.031	0.065	7.038
138	0.181	2.722	0.030	5.919	0.019	6.914	108	0.399	2.786	0.097	6.036	0.068	7.044
137	0.185	2.723	0.031	5.922	0.020	6.919	107	0.410	2.789	0.100	6.040	0.071	7.050
136	0.190	2.725	0.032	5.924	0.021	6.923	106	0.421	2.791	0.104	6.046	0.074	7.055
135	0.195	2.727	0.033	5.927	0.022	6.925	105	0.432	2.794	0.109	6.052	0.077	7.061
134	0.201	2.729	0.035	5.931	0.022	6.927	104	0.444	2.797	0.113	6.057	0.080	7.066
133	0.206	2.731	0.036	5.933	0.023	6.930	103	0.456	2.800	0.117	6.062	0.084	7.073
132	0.212	2.733	0.037	5.936	0.024	6.934	102	0.468	2.803	0.122	6.069	0.087	7.079
131	0.217	2.734	0.038	5.939	0.025	6.937	101	0.481	2.806	0.127	6.075	0.091	7.085
130	0.223	2.736	0.040	5.942	0.026	6.941	100	0.494	2.809	0.132	6.081	0.095	7.092
129	0.229	2.738	0.043	5.949	0.028	6.948	99	0.507	2.812	0.137	6.087	0.099	7.099
128	0.235	2.740	0.044	5.951	0.029	6.951	98	0.521	2.815	0.143	6.094	0.103	7.105
127	0.241	2.742	0.046	5.955	0.030	6.954	97	0.535	2.819	0.148	6.100	0.108	7.113
126	0.248	2.744	0.048	5.958	0.031	6.957	96	0.550	2.822	0.154	6.106	0.112	7.120
125	0.254	2.746	0.050	5.962	0.033	6.963	95	0.565	2.825	0.160	6.113	0.117	7.127
124	0.261	2.748	0.052	5.966	0.034	6.966	94	0.580	2.829	0.167	6.121	0.122	7.135
123	0.268	2.750	0.054	5.970	0.035	6.969	93	0.596	2.832	0.173	6.127	0.128	7.144
122	0.275	2.752	0.056	5.973	0.038	6.977	92	0.612	2.836	0.180	6.134	0.133	7.151
121	0.283	2.755	0.058	5.977	0.039	6.980	91	0.629	2.839	0.187	6.142	0.139	7.159
120	0.290	2.757	0.061	5.982	0.041	6.985	90	0.646	2.843	0.195	6.150	0.145	7.168
119	0.298	2.759	0.063	5.985	0.043	6.990	89	0.664	2.847	0.202	6.157	0.151	7.176
118	0.306	2.761	0.066	5.990	0.045	6.995	87	0.701	2.855	0.219	6.173	0.164	7.193
117	0.314	2.763	0.068	5.993	0.047	6.999	86	0.721	2.859	0.228	6.182	0.171	7.203
116	0.323	2.766	0.071	5.998	0.049	7.004	85	0.740	2.863	0.237	6.190	0.179	7.213
115	0.331	2.768	0.074	6.003	0.051	7.008	84	0.761	2.867	0.246	6.198	0.187	7.223
114	0.340	2.771	0.077	6.007	0.053	7.013	83	0.782	2.872	0.256	6.208	0.195	7.233
113	0.349	2.773	0.080	6.012	0.055	7.017	82	0.804	2.876	0.266	6.216	0.203	7.242
112	0.359	2.776	0.083	6.016	0.057	7.022	81	0.826	2.881	0.277	6.226	0.212	7.253
111	0.368	2.778	0.086	6.021	0.060	7.028	80	0.849	2.885	0.288	6.235	0.221	7.263

Table 4-3 continued.

Z_0 (Ω)	$\varepsilon_r = 4$ (SiO$_2$, FR-4)		$\varepsilon_r = 10$ (Alumina)		$\varepsilon_r = 11.9$ (Si)		Z_0 (Ω)	$\varepsilon_r = 4$ (SiO$_2$, FR-4)		$\varepsilon_r = 10$ (Alumina)		$\varepsilon_r = 11.9$ (Si)	
	u	ε_e	u	ε_e	u	ε_e		u	ε_e	u	ε_e	u	ε_e
79	0.873	2.890	0.299	6.245	0.230	7.274	44	2.518	3.131	1.229	6.828	1.047	7.936
78	0.898	2.895	0.311	6.255	0.240	7.285	43	2.609	3.140	1.283	6.856	1.096	7.968
77	0.923	2.900	0.324	6.265	0.251	7.297	42	2.703	3.150	1.340	6.884	1.147	8.002
76	0.949	2.905	0.337	6.276	0.262	7.309	41	2.803	3.160	1.400	6.913	1.201	8.036
75	0.976	2.910	0.350	6.286	0.273	7.321	40	2.908	3.171	1.464	6.944	1.259	8.072
74	1.003	2.915	0.364	6.297	0.285	7.333	39	3.019	3.181	1.531	6.974	1.319	8.108
73	1.032	2.921	0.379	6.309	0.297	7.345	38	3.136	3.192	1.602	7.006	1.384	8.147
72	1.062	2.926	0.394	6.320	0.310	7.359	37	3.259	3.203	1.677	7.039	1.452	8.186
71	1.092	2.932	0.410	6.332	0.323	7.371	36	3.390	3.214	1.757	7.073	1.524	8.226
70	1.123	2.937	0.426	6.344	0.338	7.386	35	3.528	3.226	1.841	7.108	1.600	8.268
69	1.156	2.943	0.444	6.357	0.352	7.399	34	3.675	3.237	1.931	7.143	1.682	8.311
68	1.190	2.949	0.462	6.369	0.368	7.414	33	3.831	3.250	2.027	7.180	1.769	8.355
67	1.224	2.955	0.480	6.382	0.384	7.429	32	3.997	3.262	2.129	7.218	1.862	8.402
66	1.260	2.961	0.500	6.396	0.400	7.444	31	4.174	3.275	2.238	7.258	1.961	8.449
65	1.298	2.968	0.520	6.410	0.418	7.460	30	4.364	3.288	2.355	7.298	2.067	8.498
64	1.336	2.974	0.541	6.424	0.436	7.476	29	4.567	3.301	2.480	7.340	2.181	8.549
63	1.376	2.980	0.563	6.439	0.455	7.492	28	4.785	3.315	2.615	7.384	2.304	8.601
62	1.417	2.987	0.586	6.454	0.475	7.509	27	5.020	3.329	2.760	7.428	2.436	8.655
61	1.460	2.994	0.610	6.470	0.496	7.527	26	5.273	3.344	2.917	7.475	2.579	8.712
60	1.504	3.001	0.635	6.486	0.518	7.545	25	5.547	3.359	3.087	7.523	2.734	8.770
59	1.551	3.008	0.661	6.502	0.541	7.564	24	5.845	3.374	3.272	7.573	2.902	8.831
58	1.598	3.015	0.688	6.519	0.564	7.583	23	6.169	3.390	3.474	7.625	3.086	8.894
57	1.648	3.022	0.717	6.538	0.589	7.603	22	6.523	3.407	3.694	7.679	3.287	8.960
56	1.700	3.030	0.746	6.556	0.616	7.624	21	6.912	3.424	3.936	7.734	3.508	9.028
55	1.753	3.037	0.777	6.575	0.643	7.645	20	7.341	3.441	4.203	7.793	3.752	9.100
54	1.809	3.045	0.809	6.594	0.672	7.667	19	7.815	3.459	4.499	7.854	4.022	9.174
53	1.867	3.053	0.843	6.614	0.702	7.690	18	8.344	3.478	4.829	7.917	4.323	9.252
52	1.927	3.061	0.878	6.635	0.733	7.713	17	8.936	3.497	5.199	7.983	4.661	9.334
51	1.991	3.069	0.915	6.657	0.766	7.738	16	9.603	3.517	5.616	8.053	5.043	9.419
50	2.056	3.077	0.954	6.679	0.800	7.763	15	10.361	3.538	6.090	8.126	5.476	9.509
49	2.125	3.086	0.995	6.702	0.837	7.790	14	11.229	3.559	6.633	8.202	5.972	9.604
48	2.197	3.094	1.037	6.726	0.875	7.817	13	12.233	3.581	7.262	8.282	6.547	9.704
47	2.272	3.103	1.081	6.750	0.914	7.845	12	13.407	3.604	7.997	8.367	7.219	9.809
46	2.350	3.112	1.128	6.775	0.956	7.874	11	14.798	3.628	8.868	8.456	8.016	9.920
45	2.432	3.121	1.177	6.801	1.000	7.904	10	16.471	3.652	9.916	8.550	8.975	10.038

4.10.3 Microstrip Resistance

In calculating R for a microstrip line, the conductor thickness must be taken into account. The total resistance is the sum of the resistance of the strip and the resistance of the ground plane:

$$R = R_{\text{strip}} + R_{\text{ground}} , \tag{4.186}$$

where [51]

$$R_{\text{strip}} = \frac{R_s}{w} \left(\frac{1}{\pi} + \frac{1}{\pi^2} \ln \frac{4\pi w}{t} \right) q , \tag{4.187}$$

R_s is the sheet resistance, t is the thickness of the strip, and q is the filling factor of the line:

$$q = \begin{cases} 1, & \text{for } w/h \leq 0.5 \\ 0.94 + 0.132w/h - 0.0062 \left(w/h \right)^2 , & \text{for } 0.5 \leq w/h \leq 10 . \end{cases} \tag{4.188}$$

The resistance of the ground plane is

$$R_{\text{ground}} = \frac{R_s}{w} \frac{w/h}{w/h + 5.8 + 0.03h/w}, \text{ for } 0.1 \leq w/h \leq 10 . \tag{4.189}$$

Finally, the conductor loss, α_c, from Equation (4.75), is $R/2Z_0$. In summary, the conductor resistance is comprised of a strip component and a component due to the ground plane.

EXAMPLE 4. 20   Microstrip Attenuation

If the strip in Example 4.19 has a resistance of $1\ \Omega \cdot \text{cm}^{-1}$ and the ground plane resistance can be ignored, what is the attenuation constant at 5 GHz?

Solution:
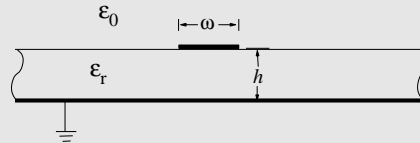
For a low-loss line, $\alpha = R/2Z_0$ (since there is no dielectric loss), $R = 1\ \Omega{\cdot}\text{cm}^{-1}$, $Z_0 = 86.1\ \Omega$, and so

$$\alpha = 0.581\ m\text{Np} \cdot \text{m}^{-1}.$$

4.11 Microstrip Design Formulas

The formulas developed in Section 4.10.2 on Page 223 enable the electrical characteristics to be determined given the material properties and the

physical dimensions of a microstrip line. In design, the physical dimensions must be determined given the desired electrical properties. Several people have developed procedures that can be used to synthesize microstrip lines [67, 69–72]. This subject is considered in much more depth by Edwards and Steer [51], and here just one approach is reported which is applicable to alumina-type substrates where $8 \ll \varepsilon_r \ll 12$. The formulas are useful outside this range, but with reduced accuracy. Again, these formulas are the result of curve fits, but starting with physically based equation forms.

4.11.1 High Impedance

For narrow strips, that is, when $Z_0 > (44 - \varepsilon_r) \ \Omega$:

$$\frac{w}{h} = \left(\frac{\exp H'}{8} - \frac{1}{4 \exp H'} \right)^{-1}, \tag{4.190}$$

where

$$H' = \frac{Z_0 \sqrt{2(\varepsilon_r + 1)}}{119.9} + \frac{1}{2} \left(\frac{\varepsilon_r - 1}{\varepsilon_r + 1} \right) \left(\ln \frac{\pi}{2} + \frac{1}{\varepsilon_r} \ln \frac{4}{\pi} \right) \tag{4.191}$$

and [72]

$$\varepsilon_e = \frac{\varepsilon_r + 1}{2} \left[1 + \frac{29.98}{Z_0} \left[\frac{2}{\varepsilon_r + 1} \right] \left(\frac{\varepsilon_r - 1}{\varepsilon_r + 1} \right) \left(\ln \frac{\pi}{2} + \frac{1}{\varepsilon_r} \ln \frac{4}{\pi} \right) \right]^{-2}. \tag{4.192}$$

4.11.2 Low Impedance

Strips with low Z_0 are relatively wide and the formulas below can be used when $Z_0 < (44 - \varepsilon_r) \ \Omega$. The crosssectional geometry is given by

$$\frac{w}{h} = \frac{2}{\pi} \left[(d_{\varepsilon_r} - 1) - \ln (2d_{\varepsilon_1} - 1) \right] + \frac{(\varepsilon_r - 1)}{\pi \varepsilon_r} \left[\ln (d_{\varepsilon_r} - 1) + 0.293 - \frac{0.517}{\varepsilon_r} \right], \tag{4.193}$$

where

$$d_{\varepsilon_r} = \frac{59.95 \pi^2}{Z_0 \sqrt{\varepsilon_r}}, \ \ d_{\varepsilon_1} = \frac{59.95 \pi^2}{Z_0}, \tag{4.194}$$

and

$$\varepsilon_e = \frac{\varepsilon_r}{0.96 + \varepsilon_r (0.109 - 0.004 \varepsilon_r)[\log (10 + Z_0) - 1]}. \tag{4.195}$$

For microstrip lines on alumina, which has $\varepsilon_r \approx 10$, the expressions above are accurate to $\pm 0.2\%$ for

$$8 \leq Z_0 \leq 45 \ \ \Omega. \tag{4.196}$$

EXAMPLE 4. 21	Microstrip Design

Design a microstrip line to have a characteristic impedance of 75 Ω at 10 GHz. The microstrip is to be constructed on a substrate that is 500 μm thick with a relative dielectric constant of 5.6. What is the width of the line? Ignore the thickness of the strip. What is the effective permittivity of the line?

Solution:

(a) Two design formulas were introduced for microstrip: one for high impedance and one for low. The high-impedance (or narrow-strip) formula (Equation (4.190)) is to be used for $Z_o > (44 - \varepsilon_r)$ $[= (44 - 5.6) = 48.4]$ Ω.

With $\varepsilon_r = 5.6$ and $Z_0 = 75$ Ω, Equation (4.191) yields $H' = 2.445$. From Equation (4.190) $w/h = 0.704$, thus

$$ w = \frac{w}{h} \times h = 0.704 \times 0.5 = 352 \ \mu\text{m}. $$

(b) The effective permittivity formula is Equation (4.192) and so $\varepsilon_e = 3.04$.

4.11.3   Comment on Formulas for Effective Permittivity

Two formulas have been presented that enable the effective permittivity of a microstrip line to be calculated. Equation (4.178) provides the effective permittivity from the physical dimensions, the width and height of the line, and the relative permittivity of the medium. For a high-impedance line, Formula (4.192) (and Formula (4.195) for a low-impedance line) provides the effective permittivity using an electrical characteristic, the characteristic impedance, and the relative permittivity of the medium. Both formulas for effective permittivity are curve fit equations, although they are based on physical insight. So the two formulas are unlikely to provide answers that are within only a rounding error. So which one do you really want? Most likely, once you have set the width you would really like to know the effective permittivity as accurately as possible. Equation (4.178) is known to be quite accurate, better than 0.2%, compared to detailed computer simulations. Can you really believe this? No, there are variations of the permittivity from place to place as the density of the material changes. This is particularly true of composite materials such as an FR-4 circuit board substrate, where the lower permittivity resin moves around during manufacture while the glass fiber stays fixed. With silicon ICs the density of the silicon dioxide varies. Another factor is that the thickness of the strip affects the field distribution and hence the effective permittivity.
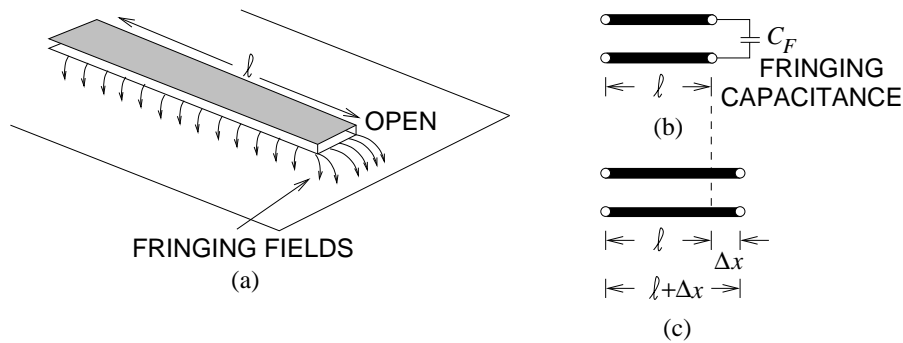
Figure 4-28 An open on a microstrip transmisison line: (a) microstrip line showing fringing fields at the open; (b) fringing capacitance model of the open; and (c) an extended line model of the open with Δx being the extra transmission line length that captures the open.

4.12 Transmission Line Components

The simplest microwave circuit element is a uniform section of transmission line which can be used to introduce a time delay or a frequency-dependent phase shift. More commonly it is used to interconnect other components. Other line segments are used for interconnections, including bends, corners, twists, and transitions between lines of different dimensions. The dimensions and shapes are designed to minimize reflections and thus maximize return loss and minimize insertion loss.

4.12.1 Open

Many transmission line discontinuities arise from fringing fields. One element is the microstrip open shown in Figure 4-28. The fringing fields at the end of the transmission line in Figure 4-28(a) store energy in the electric field, and this can be modeled by the fringing capacitance, C_F, shown in Figure 4-28(b). This effect can also be modeled by an extended transmission line, as shown in Figure 4-28(c). For a typical microstrip line with $\varepsilon_r = 9.6$, $h = 600$ μm, and $w/h = 1$, C_F is approximately 36 fF. However C_F varies with frequency, and the extended length is a much better approximation to the effect of fringing, as it has very little frequency dependence [73]. For the same dimensions, the length section is approximately $0.35h$ and almost independent of frequency. A full treatment is provided in Edwards and Steer [51]. As with many fringing effects, a capacitance or inductance is used to describe the effect of fringing, but generally a distributed model is better.
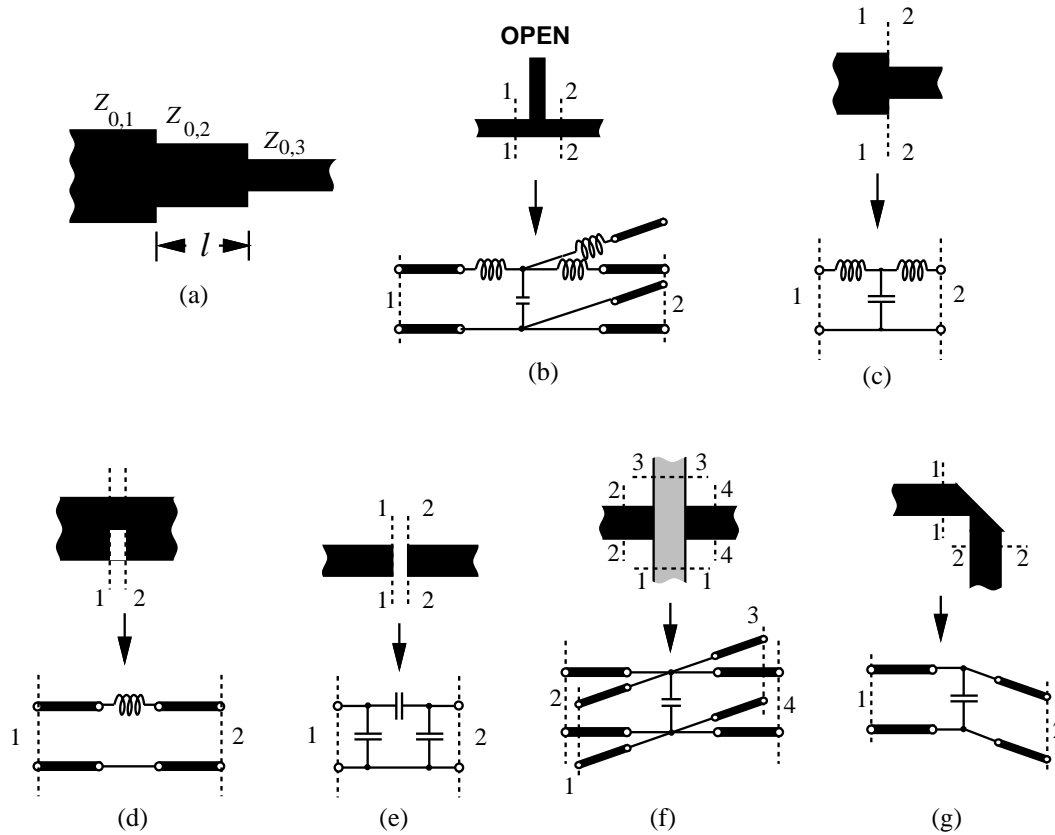
Figure 4-29 Microstrip discontinuities: (a) quarter-wave impedance transformer; (b) open microstrip stub; (c) step; (d) notch; (e) gap; (f) crossover; and (g) bend .

4.12.2 Discontinuities

The equivalent circuits of microstrip discontinuities (Figure 4-29(b–g)) are modeled by capacitive elements if the E field is interrupted and by inductive elements if the H field (or current) is disturbed. The stub shown in Figure 4-29(b) presents a short circuit to the through transmission line when the length of the stub is $\lambda_g/4$. When the stub is electrically short ($<< \lambda_g/4$) it introduces a shunt capacitance in the through transmission line.

4.12.3 Impedance Transformer

Impedance transformers interface two sections of line of different characteristic impedance. The smoothest transition and the one with the broadest bandwidth is a tapered line This element tends to be very long, as $\ell > \lambda_g$, and so step terminations called quarter-wave impedance transform-
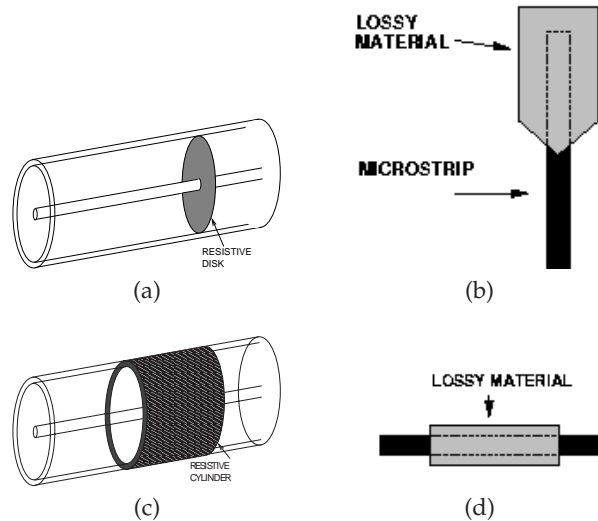
Figure 4-30   Terminations and attenuators: (a) coaxial line resistive termination; (b) microstrip matched load; (c) coaxial attenuator; and (d) microstrip attenuator.

ers (see Figure 4-29(a)) are sometimes used, although their bandwidth is relatively small and centered on the frequency at which $l = \lambda_g/4$. Ideally $Z_{0,2} = \sqrt{Z_{0,1}Z_{0,3}}$.

4.12.4 Termination

In a termination, power is absorbed by a length of lossy material at the end of a shorted piece of transmission line (Figure 4-30(a)). This type of termination is called a matched load, as power is absorbed and reflections are small irrespective of the characteristic impedance of the transmission line. This is generally preferred to a lumped resistor at high frequencies. If size is critical, as the characteristic impedance of transmission lines varies with frequency, a simpler and smaller termination can be realized by placing a resistor to ground (Figure 4-30(b)).

4.12.5 Attenuator

Attenuators reduce the level of a signal traveling along a transmission line. The basic design is to make the line lossy, but with characteristic impedance approximating that of the connecting lines so as to reduce reflections. In the case of wireless circuits, a microstrip line is made lossy by covering the line with resistive material (Figures 4-30(c) and 4-30(d)).

(a)   (b)   (c)
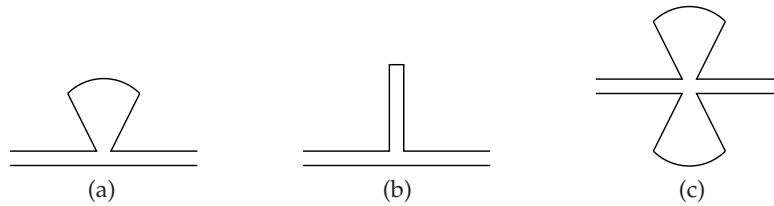
Figure 4-31 Microstrip stubs: (a) radial shunt-connected stub; (b) conventional shunt stub; and (c) butterfly radial stub.

4.12.6 Planar Radial Stub

The use of a radial stub (Figure 4-31(a)), as opposed to the conventional microstrip stub (Figure 4-31(b)), can improve the bandwidth of many microstrip circuits. A major advantage of a radial stub is that the input impedance presented at the through line generally has broader bandwidth than that obtained with the conventional stub. When two shunt-connected radial stubs are introduced in parallel (i.e., one on each side of the microstrip feeder line) the resulting configuration is termed a "butterfly" structure (see Figure 4-31(c)). Critical design parameters include the radius, r, and the angle of the stub.

EXAMPLE 4. 22   Rat-Race Hybrid

In this example the "rat-race" circuit shown in Figure 4-32(a) is considered. One of the functions of this circuit is that with an input at Port 1, the power of this signal is split between Ports 2 and 3. At the same time, no signal appears at Port 4. This example is an exercise in exploiting the impedance transformation properties of the transmission line.

From Figure 4-32(a) it is seen that each port is separated from the other port by a specific electrical length. Intuitively one can realize that there will be various possible outputs for excitation from different ports. Each case will be studied.
When Port 1 of the hybrid is excited or driven, the outputs at Ports 3 and 4 are in phase, as both are distanced from Port 1 by an electrical length of $\lambda_g/4$, while Port 2 remains isolated, as the electrical distance from Port 1 to Port 2 is an even multiple of $\lambda_g/2$. Thus Port 2 will be an electrical short circuit to the signal at Port 1.

In a similar way, a signal excited at Port 2 will result in outputs at Ports 3 and 4, though with a phase difference of $180°$ between the two output ports and Port 1 remains isolated, which is directly from the same analysis done in the earlier case.

Finally, a signal excited at Ports 3 and 4 will result in the sum of the two signals at Port 1 and the difference of two signals at Port 2. This combination of output is again due to varying electrical length between every port and every other port in the rat-race hybrid. The equivalent transmission line model and equivalent circuit of the rat-race hybrid are shown in Figures 4-32(b) and 4-32(c), respectively.
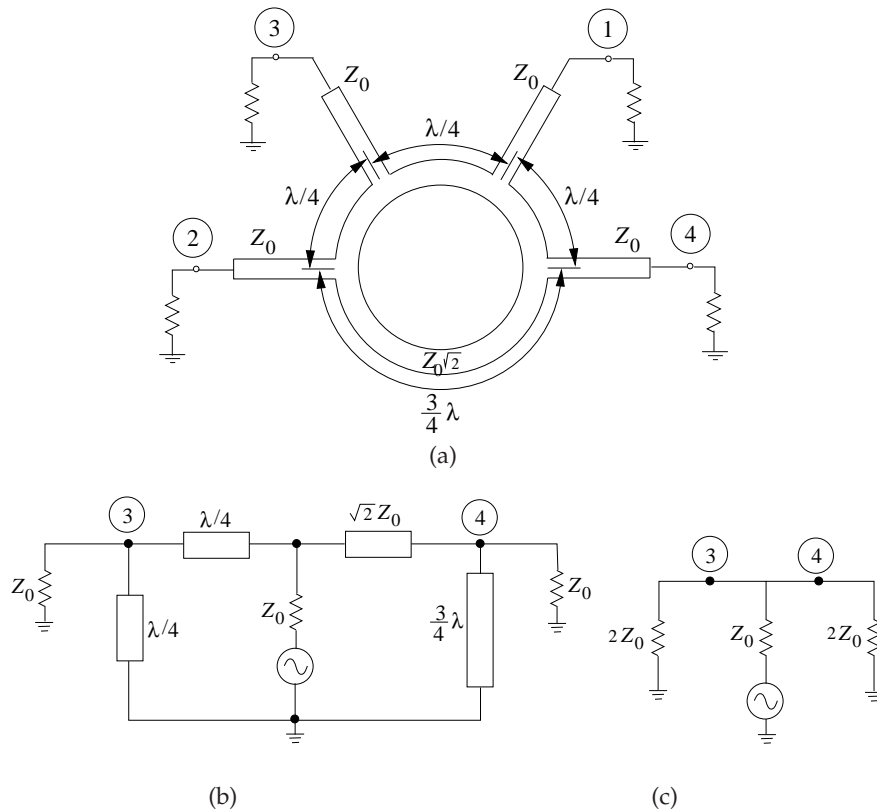
Figure 4-32 Rat-race hybrid with input at Port 1, outputs at Ports 3 and 4, and virtual ground at Port 2: (a) implementation as a planar circuit; (b) transmission-line model; and (c) equivalent circuit model.

4.13 Resonators

In a lumped-element resonant circuit, stored energy is transferred between an inductor which stores magnetic energy and a capacitor which stores electric energy, and back again every period. Microwave resonators function the same way, exchanging energy stored in electric and magnetic forms, but with the energy stored spatially. Resonators are described in terms of their quality factor,

$$Q = 2\pi f_0 \left(\frac{\text{average energy stored in the resonator at } f_0}{\text{power lost in the resonator}} \right), \qquad (4.197)$$

where f_0 is the resonant frequency. The Q is reduced and thus the resonator bandwidth is increased by the power lost to the external circuit so that the
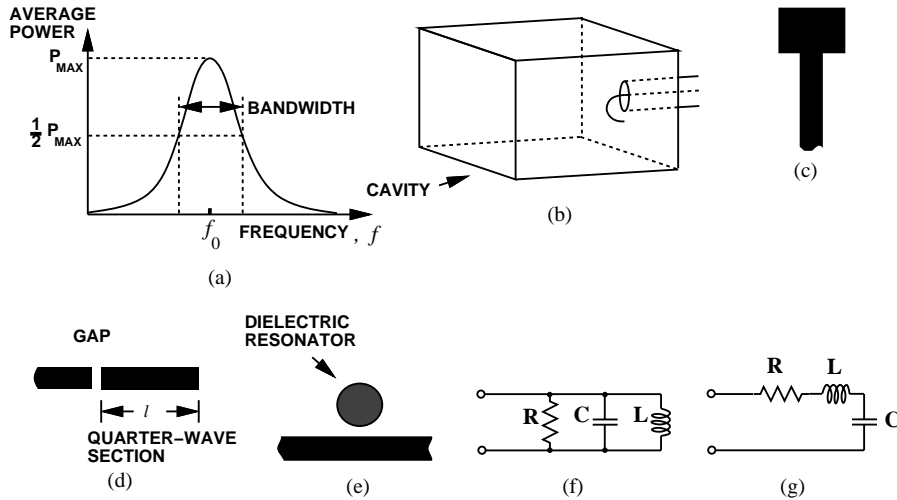
Figure 4-33 Microwave resonators: (a) resonator response, (b) rectangular cavity resonator, (c) microstrip patch resonator (d) microstrip gap-coupled reflection resonator, (e) transmission dielectric transmission resonator in microstrip, (f) parallel equivalent circuits, and (g) series equivalent circuits.

loaded Q is

$$Q_L = 2\pi f_0 \left(\frac{\text{average energy stored in the resonator at} f_0}{\text{power lost in the resonator and to the external circuit}} \right)$$

$$= \frac{1}{1/Q + 1/Q_X} , \tag{4.198}$$

where Q_X is called the external Q. Q_L accounts for the power extracted from the resonant circuit. For the simple response shown in Figure 4-33(a), the half power (3 dB) bandwidth is f_0/Q_L. Near resonance the response of a microwave resonator is very similar to the resonance response of a parallel or series LC resonant circuit, shown in Figures 4-33(f) and 4-33(g). These equivalent circuits can be used over a narrow frequency range.

Several types of resonators are shown in Figure 4-33(b). Figure 4-33(b) is a rectangular cavity resonator coupled to an external coaxial line by a small coupling loop. Figure 4-33(c) is a microstrip patch reflection resonator. This resonator has large coupling to the external circuit. The coupling can be reduced and photolithographically controlled by introducing a gap, as shown in Figure 4-33(d), for a microstrip gap-coupled transmission line reflection resonator. The Q of a resonator can be dramatically increased by using a low-loss, high dielectric constant material, as shown in Figure 4-33(e), for a dielectric transmission resonator in microstrip. Here the resonant frequency of a rectangular cavity is varied by changing the physical dimensions of the cavity, with the null of the detector indicating the cavity resonant frequency.

4.14 Summary

In this chapter a classical treatment of transmission lines was presented. Transmission lines are distributed elements and form the basis of microwave circuits. A distinguishing feature is they support forward- and backward-traveling waves. This chapter provided an understanding of signals on transmission lines. The next chapter points out problems that can occur and provides intuitive understanding and guidelines for the design of transmission lines and interconnects.

4.15 Exercises

1. A coaxial line is short-circuited at one end and is filled with a dielectric with a relative dielectric constant of 64. It is used at a frequency, f_0, of 18 GHz. [Parallels Example 4.1 on Page 167]

    (a) What is the free-space wavelength at 18 GHz?
    (b) What is the wavelength in the dielectric-filled coaxial line at 18 GHz?
    (c) The first resonance of the coaxial resonator occurs when it is one-quarter wavelength long. How long is the resonator at this first resonance?

2. Design a microstrip line having a 50 Ω characteristic impedance. The substrate has a dielectric constant of 2.3 and is 250 μm thick. The operating frequency is 18 GHz. You need to determine the width of the microstrip line.

3. A transmission line has the per unit length parameters $R = 2\ \Omega$/cm, $L = 100\ \text{nH}\cdot\text{m}^{-1}$, $G = 1\ \text{mS}\cdot\text{m}^{-1}$, $C = 200\ \text{pF}\cdot\text{m}^{-1}$.

    (a) What is the propagation constant of the line at 5 GHz?
    (b) What is the characteristic impedance of the line at 5 GHz?
    (c) Plot the magnitude of the characteristic impedance versus frequency from 100 MHz to 10 GHz.

4. A transmission line has an attenuation of 2 $\text{dB}\cdot\text{m}^{-1}$ and a phase constant of 25 $\text{radians}\cdot\text{m}^{-1}$ at 2 GHz. [Parallels Example 4.5 on Page 187]

    (a) What is the complex propagation constant of the transmission line?

(b) If the capacitance of the line is 50 $\text{pF}\cdot\text{m}^{-1}$ and the conductive loss is zero (i.e., $G = 0$), what is the characteristic impedance of the line?

5. A transmission line has the following $RLCG$ parameters: $R = 100\ \Omega\cdot\text{m}^{-1}$, $L = 85\ \text{nH}\cdot\text{m}^{-1}$, $G = 1\ \text{S}\cdot\text{m}^{-1}$, and $C = 150\ \text{pF}\cdot\text{m}^{-1}$. Consider a traveling wave on the transmission line with a frequency of 1 GHz. [Parallels Example 4.4 on Page 184]

    (a) What is the attenuation constant?
    (b) What is the phase constant?
    (c) What is the phase velocity?
    (d) What is the characteristic impedance of the line?
    (e) What is the group velocity?
    (f) If the line resistance is $R = 0\ \Omega\cdot\text{m}^{-1}$, what is the phase velocity?
    (g) If the line resistance is $R = 0\ \Omega\cdot\text{m}^{-1}$, what is the group velocity?
    (h) If the line resistance is $R = 10\ \text{k}\Omega\cdot\text{m}^{-1}$, what is the phase velocity?
    (i) If the line resistance is $R = 10\ \text{k}\Omega\cdot\text{m}^{-1}$, what is the group velocity?

6. A very low-loss microstrip transmission line has the following per unit length parameters: $R = 2\ \Omega\cdot\text{m}^{-1}$, $L = 80\ \text{nH}\cdot\text{m}^{-1}$, $C = 200\ \text{pF}\cdot\text{m}^{-1}$, and $G = 1\ \mu\text{S}\cdot\text{m}^{-1}$.

    (a) What is the characteristic impedance of the line if loss is ignored?
    (b) What is the attenuation constant due to conductor loss?
    (c) What is the attenuation constant due to dielectric loss?

7. A lossless transmission line has the following per unit length parameters: $L = 80$ nH·m^{-1}, $C = 200$ pF·m^{-1}. Consider a traveling wave on the transmission line with a frequency of 1 GHz.

   (a) What is the attenuation constant?

   (b) What is the phase constant?

   (c) What is the phase velocity?

   (d) What is the characteristic impedance of the line?

   (e) Now consider that the dielectric is replaced by a dielectric with $\varepsilon_r = 1$ (or air). The capacitance per unit length of the line is now C(air) = 50 pF·m^{-1}. What is the effective relative dielectric constant of the line?

8. The resonator below is constructed from a 3.0 cm length of 100 Ω air-filled coaxial line, shorted at one end and terminated with a capacitor at the other end.
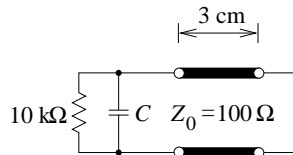
   (a) What is the lowest resonant frequency of this circuit without the capacitor (ignore the 10 kΩ resistor)?

   (b) What is the capacitor value to achieve the lowest-order resonance at 6.0 GHz (ignore the 10 kΩ resistor)?

   (c) Assume that loss is introduced by placing a 10 kΩ resistor in parallel with the capacitor. What is the Q of the circuit?

   (d) Approximately what is the bandwidth of the circuit?

9. A transmission line has an attenuation of 0.2 dB/cm and a phase constant of 50 radians·m^{-1} at 1 GHz.

   (a) What is the complex propagation constant of the transmission line?

   (b) If the capacitance of the line is 100 pF·m^{-1} and the conductive loss is zero (i.e., $G = 0$), what is the complex characteristic impedance of the line?

   (c) If the line is driven by a source modeled as an ideal voltage and a series impedance, what is the impedance of the source for maximum transfer of power to the transmission line?

   (d) If 1 W is delivered to the transmission line by the generator, what is the power in the forward-traveling wave on the line at 2 m from the generator?

10. A transmission line is driven by a 1 GHz generator having a Thevenin equivalent impedance of 50 Ω. The transmission line is lossless, has a characteristic impedance of 75 Ω, and is infinitely long. The maximum power that can be delivered to a load attached to the generator is 2 W.

   (a) What is the total (phasor) voltage at the input to the transmission line? Calculate the forward-traveling voltage wave (at the generator end of the transmission line)?

   (b) What is the magnitude of the forward-traveling voltage wave at the generator side of the line?

   (c) What is the magnitude of the forward-traveling current wave at the generator side of the line?

11. A transmission line has a characteristic impedance Z_0 and is terminated in a load with a reflection coefficient of $0.8\angle 45°$. A forward-traveling voltage wave on the line has a power of 1 dBm.

   (a) How much power is reflected by the load?

   (b) What is the power delivered to the load?

12. A 50 Ω transmission line is terminated in a load that results in a reflection coefficient of $0.5 + \jmath 0.5$.

   (a) What is the load impedance?

   (b) What is the VSWR on the line?

   (c) What is the input impedance if the line is one-half wavelength long?

13. A transmission line has a characteristic impedance Z_0 and is terminated in a load with a reflection coefficient of 0.8. A forward-traveling voltage wave on the line has a power of 1 W.

    (a) How much power is reflected by the load?
    (b) What is the power delivered to the load?

14. Communication filters are often constructed using several shorted transmission line resonators that are coupled by passive elements such as capacitors. Consider a coaxial line that is short-circuited at one end. The dielectric constant filling the coaxial line has a relative dielectric constant of 64 and the resonator is to be designed to resonate at a center frequency, f_0, of 800 MHz. [Parallels Example 4.15 on Page 209]

    (a) What is the wavelength in the dielectric-filled coaxial line?
    (b) What is the form of the equivalent circuit (in terms of inductors and capacitors) of the quarter-wavelength long resonator if the coaxial line is lossless?
    (c) What is the length of the resonator?
    (d) If the diameter of the inner conductor of the coaxial line is 2 mm and the inside diameter of the outer conductor is 5 mm, what is the characteristic impedance of the coaxial line?
    (e) Calculate the input admittance of the dielectric-filled coaxial line at $0.99f_0$, f_0, and $1.01f_0$. Determine the numerical derivative of the line admittance at f_0.
    (f) Derive the numeric values of the equivalent circuit of the resonator at the resonant frequency and derive the equivalent circuit of the resonator. Hint: Match the derivative expression derived in (e) with the actual derivative derived in Example 4.15.

15. A load consists of a shunt connection of a capacitor of 10 pF and a resistor of 25 Ω. The load terminates a lossless 50 Ω transmission line. The operating frequency is 1 GHz. [Parallels Example 4.7 on Page 196]

    (a) What is the impedance of the load?
    (b) What is the normalized impedance of the load (normalized to the characteristic impedance of the line)?
    (c) What is the reflection coefficient of the load?
    (d) What is the current reflection coefficient of the load? (When the term reflection coefficient is used without a qualifier it is assumed to be the voltage reflection coefficient.)
    (e) What is the standing wave ratio (SWR)?
    (f) What is the current standing wave ratio (ISWR)? (When SWR is used on its own it is assumed to refer to the voltage standing wave ratio [VSWR].)

16. The transmission line shown in the Figure 4-16 consists of a source with Thevenin impedance $Z_1 = 40$ Ω and source $E = 5$ V (peak) connected to a quarter-wavelength long line of characteristic impedance $Z_{01} = 50$ Ω, which in turn is connected to an infinitely long line of characteristic impedance $Z_{02} = 100$ Ω. The transmission lines are lossless. Two reference planes are shown in Figure 4-16. At reference plane 1 the incident power is P_{I1}, the reflected power is P_{R1}, and the transmitted power is P_{T1}. P_{I2}, P_{R2}, and (P_{T2}) are similar quantities at reference plane 2. [Parallels Example 4.9 on Page 198]

    (a) What is P_{I1}?
    (b) What is P_{T2}?

17. A transmission line is driven by a 1 GHz generator with a Thevenin equivalent impedance of 50 Ω. The maximum power that can be delivered to a load attached to the generator is 2 W. The generator is connected to a 10 cm long lossless transmission line with a characteristic impedance of 75 Ω. Finally, the line is terminated in a load that has a complex reflection coefficient (referred to 50 Ω) of $0.65 + \jmath0.65$. The effective relative permittivity, ϵ_{eff}, of the transmission line medium is 2.0, and the effective relative permeability of the line is that of free space.

(a) Calculate the forward-traveling voltage wave (at the generator end of the transmission line).

(b) What is the load impedance?

(c) What is the wavelength of the forward-traveling voltage wave?

(d) What is the VSWR on the line?

(e) What is the propagation constant of the transmission line?

(f) What is the input reflection coefficient (at the generator end) of the line?

(g) What is the power delivered to the load?

18. A shorted coaxial line is used as a resonator. The first resonance is determined to be a parallel resonance and is at 1.4 GHz, and in a standard resonator test fixture the unloaded Q is determined to be 520.

(a) Draw the lumped-element equivalent circuit of the resonator.

(b) How long is the resonator in terms of wavelength?

(c) Briefly describe how energy is stored in the resonator.

(d) Briefly describe the sources of loss in the resonator.

(e) Now the resonator is used in an actual application and the 3 dB bandwidth is found to be 0.4%. What is the external (sometimes called extrinsic) Q, Q_e, of the resonator in this application? (First you will need to determine the loaded Q, i.e., Q_L.)

19. A transmission line is driven by a generator with a maximum available power of 23 dBm and a Thevenin equivalent impedance of 60 Ω. The transmission line has a characteristic impedance of 25 Ω. [Parallels Example 4.13 on Page 208]

(a) What is the Thevenin equivalent generator voltage?

(b) What is the magnitude of the forward-traveling voltage wave on the line? Assume the line is infinitely long.

(c) What is the power of the forward-traveling voltage wave?

20. A 50 Ω air-filled transmission line is connected between a 40 GHz source with a Thevenin equivalent impedance of 50 Ω and a load. The SWR on the line is 3.5.

(a) What is the magnitude of the reflection coefficient, Γ_L, at the load.

(b) What is the phase constant, β, of the line?

(c) If the first minimum of the standing wave voltage on the transmission line is at a distance 2 mm from the load, determine the electrical distance (in degrees) of the SWR minimum from the load angle of the Γ_L at the load.

(d) Determine the angle of Γ_L at the load.

(e) What is Γ_L in magnitude-phase form?

(f) What is Γ_L in complex (rectangular) form?

(g) Determine the load impedance, Z_L.

21. A load has an impedance $Z_L = 45 + \jmath 75$. The reference system impedance is 100 Ω.

(a) What is the reflection coefficient?

(b) What is the current reflection coefficient?

(c) What is the SWR?

(d) What is the ISWR?

(e) The power available from a source with a 100 Ω Thevenin equivalent impedance is 1 mW. The source is connected directly to the load Z_L. Calculate the power delivered to Z_L using the reflection coefficient?

(f) What is the total power absorbed by the Thevenin equivalent source impedance?

(g) Discuss the effect of inserting a lossless 100 Ω transmission line between the source and the load.

22. A load consists of a resistor of 100 Ω in parallel with a 5 pF capacitor with an electrical signal at 2 GHz.

(a) What is the load impedance?

(b) What is the reflection coefficient in a 50 Ω reference system?

(c) What is the SWR on a 50 Ω transmission line connected to the load?

(d) Develop an analytic formula relating the reflection coefficient (Γ_1) in one reference system (Z_{01}) to the reflection coefficient (Γ_2) in another reference system (Z_{02}).

(e) Develop an analytic formula relating the SWR in one reference system (SWR$_1$) to the SWR (SWR$_2$) in another reference system.

(f) Calculate the SWR on a 100 Ω line.

23. An amplifier is connected to a load by a transmission line matched to the amplifier. If the SWR on the line is 1.5, what percentage of the available amplifier power is absorbed by the load?

24. The output amplifier of a cellular phone can tolerate a mismatch characterized by a maximum SWR of 2.0. The amplifier is characterized by a Thevenin equivalent circuit with an impedance of 50 Ω and is connected directly to an antenna characterized by a load impedance, Z_L. Determine the tolerance limits on Z_L so that the amplifier does not self-destruct.

25. A source is connected to a load through a length of transmission line having a loss of 1.5 dB. The source reflection coefficient (referred to the transmission line) is 0.2 and the load reflection coefficient is 0.5.

(a) What is the transmission coefficient?

(b) Draw the bounce diagram using the transmission and reflection coefficients. Determine the overall effective transmission coefficient from the source to the load. Calculate the power delivered to the load from a source with an available power of 600 mW.

26. The resonator below is constructed from a 3.0 cm length of 100 Ω air-filled coaxial line, shorted at one end and terminated with a capacitor at the other end, as shown:

27. A load of 100 Ω is to be matched to a transmission line with a characteristic impedance of 50 Ω. Use a quarter-wave transformer. What is the characteristic impedance of the quarter-wave transformer?

28. A shorted coaxial line is used as a resonator. The first resonator is determined to be a parallel resonance and is at 1 GHz.

(a) Draw the lumped-element equivalent circuit of the resonator.

(b) What is the electrical length of the resonator?

(c) What is the impedance looking into the line at resonance?

(d) If the resonator is $\lambda/4$ longer, what is the impedance of the resonator now?

29. A quarter-wave transformer is to be used to match a load of 50 Ω to a generator with a Thevenin equivalent impedance of 75 Ω. What is the characteristic impedance of the quarter-wave transformer?

30. A coaxial transmission line is filled with lossy material with a relative dielectric constant of $5 - \jmath 0.2$. If the line was airfilled it would have a characteristic impedance of 100 Ω.

(a) What is the characteristic impedance of the dielectric-filled line?

(b) What is the propagation constant at 500 MHz?

(c) What is the input impedance of the line if it has an electrical length of 280° and is terminated in a 35 Ω resistor?

(d) What is the input impedance of the line if it has an electrical length of 280° and is terminated in an inductor of impedance $\jmath 35\ \Omega$?

The following belongs to problem 26 and 27 area (left column top):

(a) What is the lowest resonant frequency of this circuit without the capacitor (ignore the resistor)?

(b) What is the capacitor value to achieve resonance at 6.0 GHz?

(c) Assume that loss is introduced by placing a 10 kΩ resistor in parallel with the capacitor. What is the Q of the circuit?

(d) What is the bandwidth of the circuit?

(e) What is the input impedance of the line if it is 1 km long? Use reasonable approximations. [Hint: Does it matter what the termination is?]

31. A coaxial line is filled with a very slightly lossy material with a relative dielectric constant of 5. The line would have a characteristic impedance of 100 Ω if it was airfilled.

    (a) What is the characteristic impedance of the dielectric-filled line?

    (b) What is the propagation constant at 500 MHz? Use the fact that the velocity of an EM wave in a lossless air-filled line is the same as that of free-space propagation in air. That is, the propagation constant is the same.

    (c) What is the input impedance of the line if it has an electrical length of $90°$ and it is terminated in a 35 Ω resistor?

    (d) What is the input impedance of the line if it has an electrical length of $180°$ and is terminated in an inductor of impedance $j35\ \Omega$?

    (e) What is the input impedance of the line if it is 1 km long? Use reasonable approximations, remembering that the line is slightly lossy.

32. The strip of a microstrip has a width of 250 μm and is fabricated on a lossless substrate that is 500 μm thick and has a relative permittivity of 2.3. [Parallels Example 4.19 on Page 225]

    (a) What is the effective relative permittivity?

    (b) What is the characteristic impedance?

    (c) What is the propagation constant at 3 GHz ignoring any losses?

    (d) If the strip has resistance of 0.5 Ω/cm and the ground plane resistance can be ignored, what is the attenuation constant at 3 GHz?

33. The strip of a symmetrical stripline has a width of 200 μm and is embedded in a lossless medium that is 400 μm thick and has a relative permittivity of 13, thus the separation, h, from the strip to each of the ground planes is 200 μm.

    (a) Draw the effective waveguide model of stripline with magnetic walls and an effective strip width, w_{eff}, which will be approximately the same as with microstrip.

    (b) What is the effective relative permittivity of the stripline waveguide model?

    (c) What is w_{eff}?

    (d) At what frequency will the first transverse resonance occur?

    (e) At what frequency will the first higher-order stripline mode occur?

    (f) At what frequency will the first parallel-plate waveguide mode occur? Do not consider the mode with no field variation, as this cannot be excited.

    (g) Identify the useful operating frequency range of the stripline.

34. Design a microstrip line to have a characteristic impedance of 65 Ω at 5 GHz. The microstrip is to be constructed on a substrate that is 635 μm thick with a relative dielectric constant of 9.8. Ignore the thickness of the strip. [Parallels Example 4.21 on Page 231]

    (a) What is the width of the line?

    (b) What is the effective permittivity of the line?

35. A load has a reflection coefficient of 0.5 when referred to 50 Ω. If the load is placed at the end of a transmission line with a 100 Ω characteristic impedance.

    (a) What is the complex ratio of the forward-traveling wave to the backward-traveling wave on the 100 Ω line at the load end of the line?

    (b) What is the VSWR on the 100 Ω line?

    (c) Now consider that the line has a characteristic impedance of 50 Ω. If the line has an electrical length of $45°$, what is the reflection coefficient calculated at the input of the line?

    (d) What is the VSWR on the 50 Ω line?

36. Design a microstrip shorted stub at 10 GHz with the following characteristics:

    • Characteristic impedance of 60 Ω.

    • A substrate with a permittivity of 9.6 and thickness of 500 μm.

- Input impedance that is a reactance of $\jmath 60$ Ω.

(a) What is the width of the microstrip line?

(b) What is the length of the line in centimeters?

(c) What is the effective permittivity of the line?

(d) If the line is a one-quarter wavelength longer than that calculated in (b), what will the input reactance be?

(e) Regardless of your calculations above, what is the input admittance of a one-quarter wavelength long shorted stub?

37. A load has an impedance $Z = 75 + \jmath 15$ Ω.

(a) What is the load reflection coefficient, Γ_L, if the system impedance is 75 Ω?

(b) Design a stub at the load that will make the impedance of the load plus the stub, call this Z_1, purely real; that is, the reflection coefficient of the effective load, Γ_1, has zero phase. Choose a stub impedance of 75 Ω. (Design specifications require complete electrical information such as whether the stub is open- or short-circuited, and the electrical length of the stub.)

(c) Design a quarter-wave transformer that will present a matched termination to a source with a system reference impedance of 50 Ω. (Again the design must include full electrical specifications such as the characteristic impedance of the transmission line and its electrical length.)

(d) Now convert the electrical specifications of the design into a physical specification. Assuming that the transmission line technology to be used is a microstrip line and the substrate medium is fixed with the following parameters: frequency $f = 1$ GHz, substrate thickness $h = 0.5$ mm, substrate relative dielectric constant $\epsilon_r = 10$. You must design the widths and lengths of the stub and the quarter-wave transformers.

38. Design a microstrip line to have a characteristic impedance of 20 Ω. The microstrip is to be constructed on a substrate that is 1 mm thick with a relative dielectric constant of 12. [Parallels Example 4.21 on Page 231]

(a) What is the width of the line? Ignore the thickness of the strip and frequency effects.

(b) What is the effective permittivity of the line?

Extraordinary Transmission Line Effects

5.1 Introduction

In the previous chapter, the basic operation of transmission lines was discussed with a concentration on designing and understanding the operation of microstrip lines. This chapter focuses on the frequency-dependent behavior of microstrip lines and on designing the lines so that undesirable behavior is avoided. The major limitation on the dimensions of transmission lines is determined by considering when higher-order modes (field orientations) can exist. Different modes on a transmission line travel at different velocities. Thus the problem is that if a signal on a line is split between two modes, then the information sent from one end of the line will reach the other end in two packets arriving at different times. A combination of the two modes will be detected and in general the information will be lost as the two modes combine in an incoherent manner. This multimoding must be avoided at all costs. The purpose of this chapter is to help you gain an understanding of frequency-dependent behavior and multimoding on transmission lines. You will also be able to design the dimensions of lines to avoid the excitation of higher-order modes, and will be able to debug RF and microwave circuits that fail to work correctly because of multimoding.
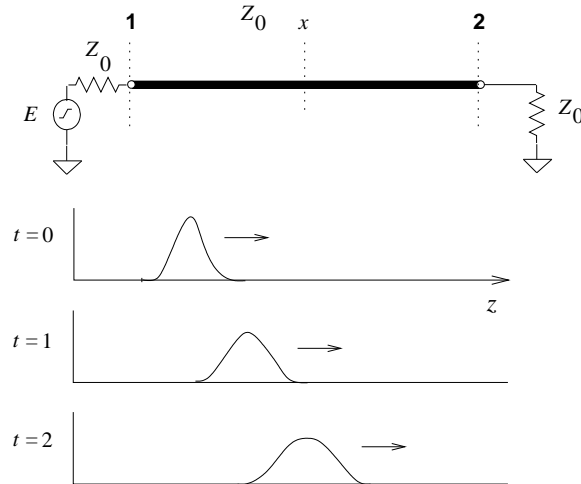
Figure 5-1 Dispersion of a pulse along a line.

5.2 Frequency-Dependent Characteristics

All interconnects have frequency-dependent behavior. In this section the origins of frequency-dependent behavior of the microstrip line are examined because the microstrip line has the most significant frequency-dependence among interconnects of general interest. Frequency-dependent behavior other than multimoding often results in dispersion. The effect can be seen in Figure 5-1 for a pulse traveling along a line. The pulse spreads out as the different frequency components travel at different speeds. You can imagine a long line with this problem—successive pulses would start merging and the signal would become unintelligible.

The most important frequency-dependent effects are

- Changes of material properties (permittivity, permeability, and conductivity) with frequency (Section 5.2.1)

- Current bunching (Section 5.2.3)

- Skin effect (Section 5.2.4)

- Internal conductor inductance variation (Section 5.2.4)

- Dielectric dispersion (Section 5.2.5)

- Multimoding (Section 5.4).

While the discussion focuses on microstrip lines, the effects occur with other planar and nonplanar transmission lines.

5.2.1 Material Dependency

Changes of permittivity, permeability, and conductivity with frequency are properties of the materials used. Fortunately the materials of interest in

microwave technology have characteristics that are almost independent of frequency, at least up to 300 GHz or so. Skin effect, charge bunching, and internal conductor inductance variation are due to the same physical phenomenon but have different effects on frequency dependence. The effects are due to the finite time it takes to transfer information from one part of a distributed structure to another. No information can travel faster than the speed of light in the medium. Dielectric dispersion is also related to the finite time required to transfer information, and the effect results from changes in the distribution of energy in the different dielectrics of an inhomogeneous structure. So dielectric dispersion relates to the differences of the speed of light in different media. As frequency increases it is possible for field distributions involving looping or variations of the fields to exist. This is called multimoding.

5.2.2   Frequency-Dependent Charge Distribution

Skin effect, current bunching, and internal conductor inductance are all due to the necessary delay in transferring EM information from one location to another. This information cannot travel faster than the speed of light in the medium. In a dielectric material, the speed of light will be slower than that in free space by a factor of $\sqrt{\varepsilon_r}$, where ε_r is the relative permittivity of the material. The speed of light in the dielectric is reduced from that in free space typically by a factor ranging from just over 1 to 300. However, the velocity in a conductor is extremely low because of high conductivity. In brief, current bunching is due to changes related to the finite velocity of information transfer through the dielectric, and skin effect and internal conductor inductance variation are due to the very slow speed of information transfer inside a conductor. As frequency increases, only limited information to rearrange charges can be sent before the polarity of the signal reverses and information is sent to reverse the charge changes. The charge distribution on the conductors depends on how fast the signal changes. What occurs is best described by considering a sinusoidal signal. Both the skin effect and charge bunching effect on a microstrip line are illustrated in Figure 5-2.

5.2.3   Current Bunching

Consider the charge distribution for a microstrip line shown in the crosssectional views in Figure 5-2. The microstrip crosssections shown here are typical of an interconnect on a printed circuit board with the top layer being solder resist. (For a monolithically integrated circuit, the top cover is a **passivation** layer.) The thickness of microstrip is often a significant fraction of its width, although this is exaggerated in Figure 5-2. The charge distribution shown in Figure 5-2(a) applies when there is a positive DC voltage on the strip (the top conductor). In this case there are positive charges on the top conductor arranged with a fairly uniform distribution. The individual positive charges (caused by the absence of some balancing
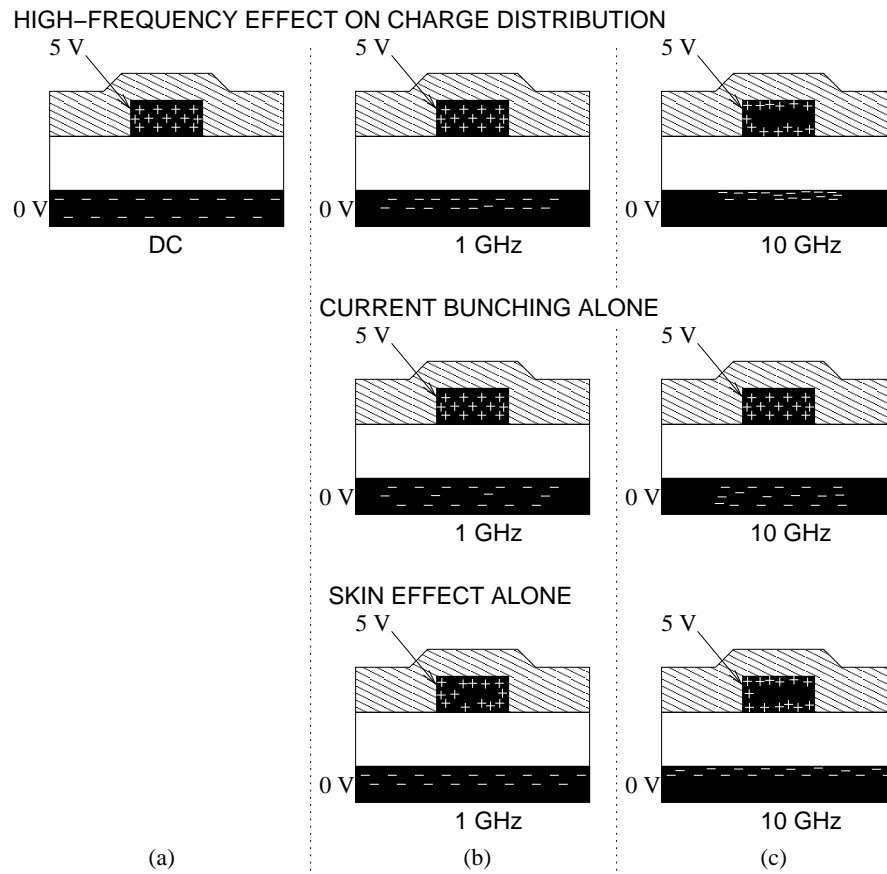
Figure 5-2   Crosssectional view of the charge distribution on an interconnect at different frequencies. The + and − indicate charge concentrations of different polarity and corresponding current densities. There is no current bunching or skin effect at DC.

electrons exposing positively charged ions) tend to repel each other, but this has little effect on the charge distribution for practical conductors with finite conductivity. (If the conductor had zero resistance then these charges would be confined to the surface of the conductor.) The bottom conductor is known as the ground plane and there are balancing negative charges, or a surplus of electrons, so that electric field lines begin on the positive charges and terminate on the negative charges. The negative charges on the ground plane are uniformly distributed across the whole of the ground plane. An important point is that where there are unbalanced, or net, charges there can be current flow. So the charge distribution at DC, shown in Figure 5-2(a), indicates that for the top conductor, current would flow uniformly
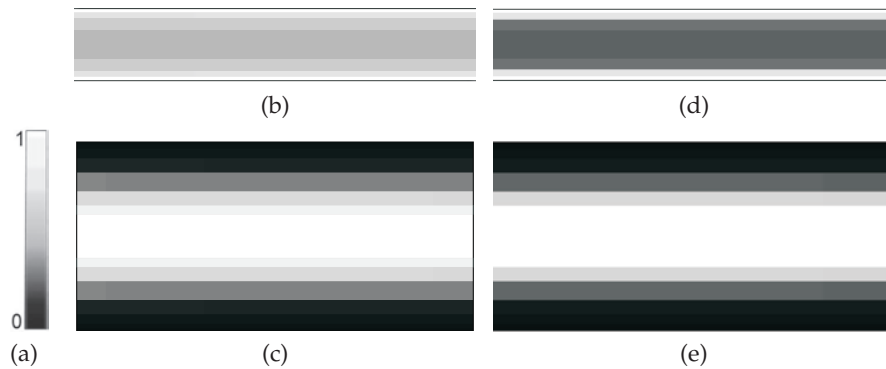
Figure 5-3 Normalized magnitudes of current and charge on an alumina microstrip line at 1 GHz: (a) normalized scale; (b) longitudinal current, i_z, on the strip (10–26 A/m); (c) on the ground plane (0–3.2 A/m); (d) the charge on the strip (80–400 nC/m^2); and (e) on the ground plane (0–33 nC/m^2). (For alumina, $\varepsilon = 10.0$ and $\mu_r = 1$.)

distributed throughout the conductors and the return current in the ground plane would be distributed over the whole of the ground plane.

   The charge distribution becomes less uniform as frequency increases and eventually the signal changes so quickly that information to rearrange charges on the ground plane is soon (half a period latter) countered by reverse instructions. Thus the charge distribution depends on how fast the signal changes. The effects are seen in the higher-frequency views shown in Figures 5-2(b) and 5-2(c). (The concentration of charges near the surface of the metal is a separate effect known as the skin effect.) The longitudinal impact of current or charge bunching alone is illustrated in Figures 5-3, 5-4, and 5-5. These figures present amplitudes of the current and charge phasors at various frequencies and were calculated using the **SONNET** EM simulator. In interpreting these figures, please take into account the magnitudes of the current and charge distributions as identified in the captions, as the scales are normalized. An alternative view (or time-domain view) is the instantaneous snapshot of current and charge shown in Figure 5-6. This situation is not just confined to the transverse plane, and regions further along the interconnect also send instructions. The net effect is bunching of charges and hence of current on both the ground plane and the strip.

5.2.4   *Skin Effect and Internal Conductor Inductance*

From the previous discussion it was seen that at low frequencies currents are distributed uniformly throughout a conductor. Thus there are magnetic fields inside the conductor and hence magnetic energy storage. As a result, there is internal conductor inductance. Transferring charge to the interior of
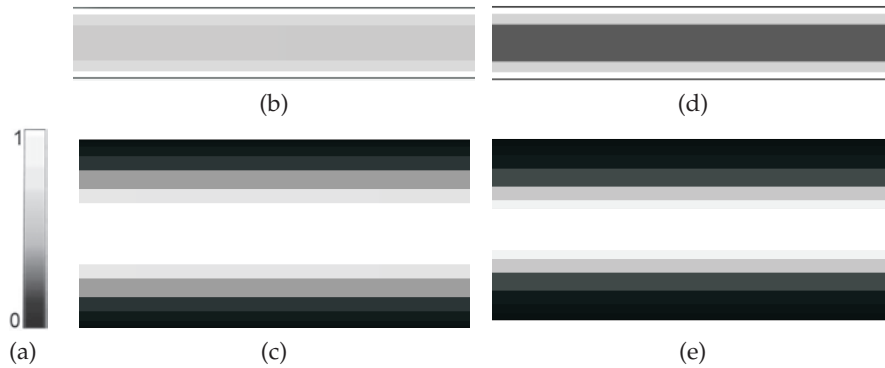
Figure 5-4 Normalized magnitudes of current and charge on an alumina microstrip line at 10 GHz: (a) normalized scale; (b) longitudinal current, i_z, on the strip (10–28 A/m); (c) on the ground plane (0–4.1 A/m); (d) the charge on the strip (114–512 nC/m^2); and (e) on the ground plane (0–39 nC/m^2).
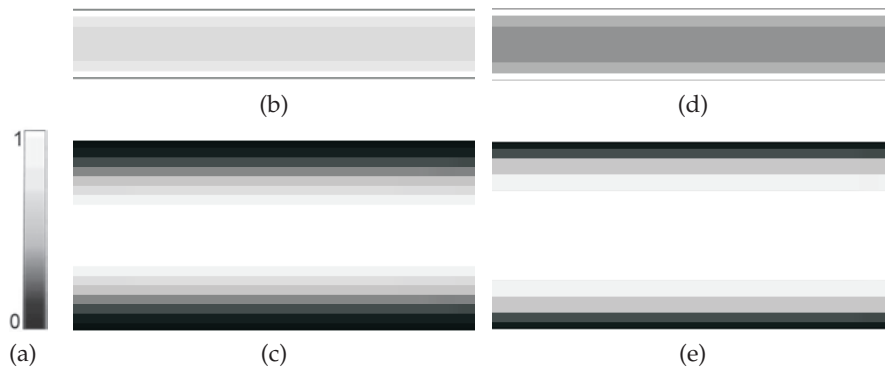
Figure 5-5 Normalized magnitudes of current and charge on an alumina microstrip line at 30 GHz: (a) normalized scale; (b) longitudinal current, i_z, on the strip (10–31 A/m); (c) on the ground plane (0–6 A/m); (d) the charge on the strip (200–575 nC/m^2); and (e) on the ground plane (0–68 nC/m^2).

conductors is particularly slow, and as the frequency of the signal increases charges are confined close to the surface of the metal. Another equally valid interpretation is that time-varying EM fields are not able to penetrate the conductors as much when the frequency increases. This phenomenon is known as the skin effect. With fewer internal currents, the internal conductor inductance reduces and the total inductance of the line drops, thus the redistribution of the current results in a change of the inductance with frequency. This is principally because, for the same current, magnetic energy is stored inside as well as outside the conductors at low frequencies. As frequency increases, the magnetic field becomes confined almost entirely to the region outside the conductors and the line inductance asymptotically reduces to a constant as the internal conductor inductance goes to zero. Only
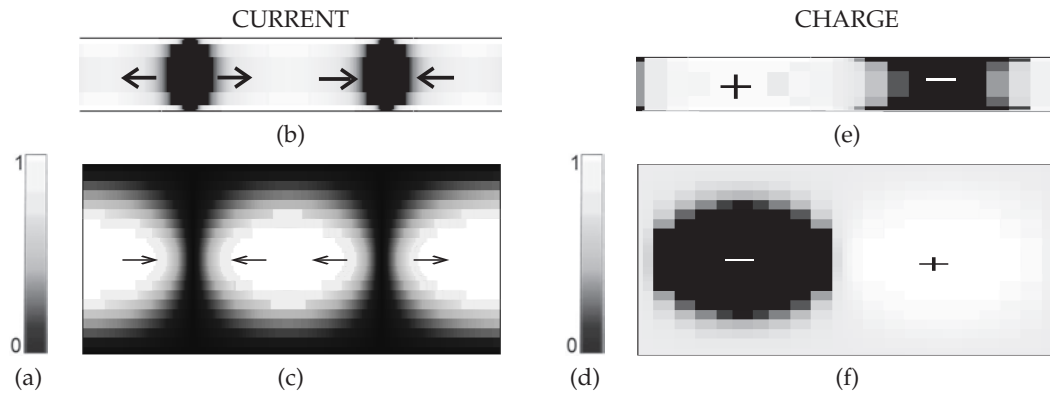
CURRENT CHARGE

Figure 5-6 Normalized instantaneous current and charge distribution on an alumina microstrip line and ground plane at 30 GHz. Current: (a) normalized scale; (b) magnitude of the longitudinal current, i_x, on the microstrip; and (c) on the ground plane. Arrows indicate the direction of current flow. Charge: (d) scale; (e) microstrip; and (f) ground plane. Signs indicate the polarity of charge.

above a few gigahertz or so can the line inductance be approximated as a constant for the transverse interconnect dimensions of a micron to several hundred microns.

Note that the line inductance is not the inductance of the conductor, but the inductance (per unit length) of the entire transmission line. Inductance is a measure of the energy stored in a magnetic field. Calculation of inductance requires a volumetric integral of energy stored in a magnetic field. If there is no magnetic field in a region then there is no inductance associated with the region. So as frequency increases and the magnetic field inside the conductor reduces further from the conductor surface, the internal inductance of the conductor goes to zero.

The skin effect is characterized by the skin depth, δ, which is the distance at which the electric field, or equivalently the charge density, reduces to $1/e$ of its value at the surface. By considering the attenuation constant of fields in a conductive medium, the skin depth is determined to be

$$\delta = 1/\sqrt{\pi f \mu_0 \sigma_2} \,. \tag{5.1}$$

Here f is frequency and σ_2 is the conductivity of the conductor. The permeability of metals typically used for interconnects (e.g., gold, silver, copper, and aluminum) is that of free space, μ_0.

The skin effect is illustrated in Figure 5-2(b) at 1 GHz. (The reader should distinguish between skin effect and current bunching.) The situation is more extreme as the frequency continues to increase (e.g., to 10 GHz) as in Figure 5-2(c). There are several important consequences of this. On the top conductor the negative charges are not uniformly distributed with respect to the depth of penetration into the conductor. Consequently, as

frequency increases, current flow is mostly concentrated near the surface of the conductors and the effective crosssectional area of the conductor, as far as the current is concerned, is less. Thus the resistance of the top conductor increases. A more dramatic situation exists for the charge distribution in the ground plane. From the previous discussion of current bunching it was seen that charge is not uniformly distributed over the whole of the ground plane, but instead becomes more concentrated under the strip. In addition to this, charges and current are confined to the skin of the ground conductor so the frequency-dependent relative change of the resistance of the ground plane with increasing frequency is greater than for the strip.

The skin effect, and to a lesser extent current bunching, results in frequency dependence of line resistance, R, with

$$R(f) = R_{\mathrm{DC}} + R_s(f). \qquad (5.2)$$

R_{DC} is the resistance of the line at DC and R_s is the skin resistance:

$$R_s(f) = R_{\mathrm{DC}} k \sqrt{f}. \qquad (5.3)$$

Here k is a constant, and while Equation (5.3) indicates proportionality to $\sqrt{f}$, this is an approximation and the actual frequency dependence may be different, but R_s always increases more slowly than frequency.

5.2.5 Dielectric Dispersion

Dispersion is principally the result of the velocity of the various frequency components of a signal being different. The electric field lines shift as a result of the different distributions of charge, with more of the electric energy being in the dielectric. Thus the effective permittivity of the line increases with increasing frequency. At high frequencies, the fundamental result of the field rearrangement is that the capacitance of the line increases, but this change can be quite small—typically less than 10% over the range of DC to 100 GHz. (This effect is described by the frequency dependence of the effective permittivity of the transmission line.) To a lesser extent, dispersion is also the result of other parameters changing with frequency, such as an interconnect's resistance. For an IC where the interconnects can have very small transverse dimensions (e.g., microns) of digital interconnects, the line resistance is the most significant source of dispersion. The qualitative effect of dispersion is the same whether it is related to the resistance (resistance-induced dispersion) or change in the effective permittivity (dielectric-inhomogeneity-induced dispersion).

Different interconnect technologies have different dispersion characteristics. For example, with a microstrip line the effective permittivity changes with frequency as the proportion of the EM energy in the air region to that in the dielectric region changes. Dispersion is reduced if the fields are localized

and cannot change orientation with frequency. This is the case with coplanar interconnects—in particular, coplanar waveguide (CPW) and coplanar strip (CPS) lines have lower dispersion characteristics than does microstrip (for small geometries). The stripline of Figure 4-8(a) also has low dispersion, as the fields are confined in one medium and the effective permittivity is just the permittivity of the medium. Thus interconnect choices can have a significant effect on the integrity of a signal being transmitted.

As discussed earlier, as the frequency is increased, the fields become more concentrated in the region beneath the strip—where the substrate permittivity has already resulted in a relatively large electric field displacement. Since the fields are forced into the dielectric substrate to an increasing extent as the frequency rises, a frequency-dependent effective microstrip permittivity, $\varepsilon_e(f)$, can be defined. This quantity clearly increases with frequency and the wave is progressively slowed down. The effective microstrip permittivity is now

$$\varepsilon_e(f) = \{c/\left[v_p(f)\right]\}^2 . \tag{5.4}$$

Fundamentally the dispersion problem then consists of solving the transmission line fields for the phase velocity, $v_p(f)$. The limits of $\varepsilon_e(f)$ are readily established; at the low-frequency extreme it reduces to the static-TEM value ε_e (or $\varepsilon_e(0)$ or $\varepsilon_e(DC)$), while as frequency is increased indefinitely $\varepsilon_e(f)$ approaches the substrate permittivity itself, ε_r. This is summarized as follows:

$$\varepsilon_e(f) \rightarrow \begin{cases} \varepsilon_e & \text{as } f \rightarrow 0 \\ \varepsilon_r & \text{as } f \rightarrow \infty \end{cases} . \tag{5.5}$$

Between these limits $\varepsilon_e(f)$ changes smoothly.

5.3 High-Frequency Properties of Microstrip Lines

Here the high-frequency properties of microstrip lines are discussed and formulas are presented for effective permittivity, characteristic impedance, and attenuation loss which incorporate frequency dependence. In the previous chapter, frequency-dependent dispersion was not incorporated; those results (or formulas) are called the quasi-static approximation. The effective dielectric constant at DC (as calculated in the previous chapter) is denoted $\varepsilon_e(0)$ and the characteristic impedance at DC is $Z_0(0)$. These are also called the quasi-static effective dielectric constant and quasi-static characteristic impedance. Detailed analysis [51] yields the following formula for the frequency-dependent effective permittivity of a microstrip line:

$$\varepsilon_e(f) = \varepsilon_r - \frac{\varepsilon_r - \varepsilon_e(0)}{1 + (f/f_a)^m} , \tag{5.6}$$

where the corner frequency

$$f_a = \frac{f_b}{0.75 + \left(0.75 - 0.332\, \varepsilon_r^{-1.73}\right)(w/h)} \tag{5.7}$$

$$f_b = \frac{47.746 \times 10^6}{h\sqrt{\varepsilon_r - \varepsilon_e(0)}} \tan^{-1}\left\{\varepsilon_r\sqrt{\frac{\varepsilon_e(0) - 1}{\varepsilon_r - \varepsilon_e(0)}}\right\} \tag{5.8}$$

$$m = \left\{ \begin{array}{ll} m_0 m_c & m_0 m_c \leq 2.32 \\ 2.32 & m_0 m_c > 2.32 \end{array} \right. \tag{5.9}$$

$$m_0 = 1 + \frac{1}{1 + \sqrt{w/h}} + 0.32\left(1 + \sqrt{w/h}\right)^{-3} \tag{5.10}$$

$$m_c = \left\{ \begin{array}{ll} 1 + \dfrac{1.4}{1 + w/h}\left\{0.15 - 0.235 e^{-0.45 f/f a}\right\}, & \text{for } w/h \leq 0.7 \\ 1, & \text{for } w/h > 0.7 \end{array} \right. \tag{5.11}$$

In all the equations given above, SI units are used. The accuracy of the equations above is within 0.6% for $0.1 \leq w/h \leq 10$, $1 \leq \varepsilon_r \leq 128$, and for any value of h/λ provided that $h < \lambda/10$.[1]

5.3.1 Frequency-Dependent Loss

The effect of loss on signal transmission is captured by the attenuation constant α. There are two primary sources of loss resulting from the dielectric, captured by the dielectric attenuation constant, α_d, and from the conductor loss, captured in the conductor attenuation constant, α_c. Thus

$$\alpha = \alpha_d + \alpha_c. \tag{5.12}$$

For dielectric loss, Equation (4.78) provides a good estimate for the attenuation when ε_e is replaced by the frequency-dependent effective dielectric constant, $\varepsilon_e(f)$.

Frequency-dependent conductor loss results from the concentration of current as frequency increases. For a wide strip of thickness three times greater than the skin depth,

$$\alpha_c(f) = \frac{R_s(f)}{w\, Z_0}. \tag{5.13}$$

For a narrow strip, say $w/h < 1$, the quasi-static loss equations (Equations (4.186) to (4.189)) can be used with acceptable accuracy.

---

[1] Note that the free-space wavelength is λ_0, the wavelength in the medium is λ, and the guide wavelength is λ_g.
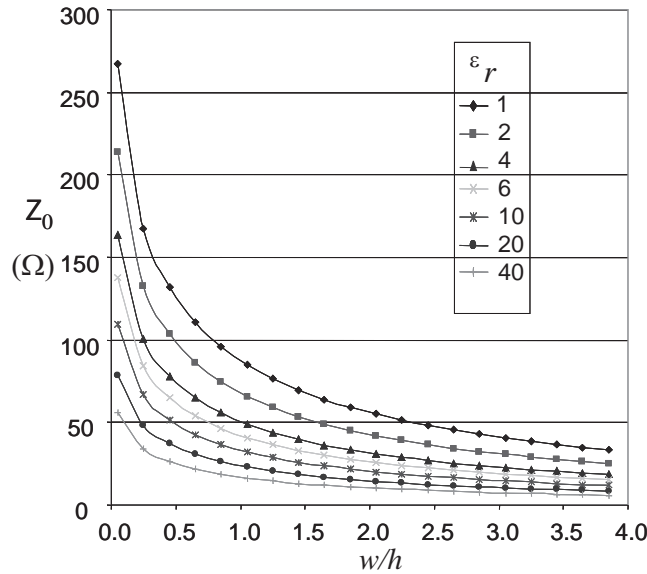
Figure 5-7 Dependence of Z_0 of a microstrip line at 1 GHz for various dielectric constants and aspect (w/h) ratios.

A third source of loss is radiation loss, α_r, and at the frequencies at which a transmission line is generally used, is usually smaller than dielectric and conductor losses. So, in full,

$$\alpha(f) = \alpha_d(f) + \alpha_c(f) + \alpha_r(f). \tag{5.14}$$

5.3.2 Field Simulations

In this section results are presented for EM simulations of microstrip lines with a variety of parameters. These simulations were performed using the SONNET EM simulator. Figure 5-7 presents calculations of Z_0 for various aspect ratios (w/h) and substrate permittivities (ε_r) when there is no loss. The key information here is that narrow strips and low-permittivity substrates have high Z_0. Conversely, wide strips and high-permittivity substrates have low Z_0. The dependence of permittivity on aspect ratio is shown in Figure 5-8, where it can be seen that the effective permittivity, ε_e, increases for wide strips. This is because more of the EM field is in the substrate.

When loss is incorporated, ε becomes complex and the imaginary components indicate loss. Figures 5-9, 5-10, and 5-11 present the frequency dependence of three microstrip lines with different substrates and aspect ratios. These simulations took into account finite loss in the conductors and in the dielectric. In Figure 5-9(a) it can be seen that the effective permittivity,
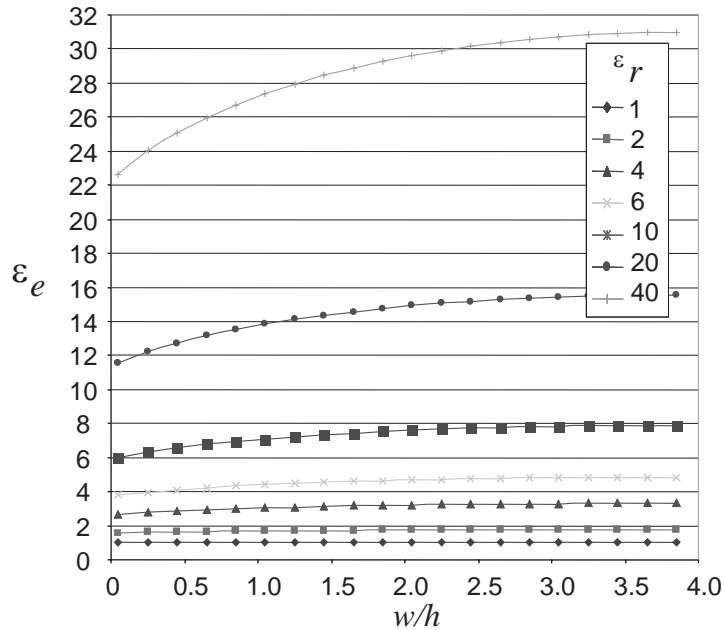
Figure 5-8   Dependence of effective relative permittivity ε_e of a microstrip line at 1 GHz for various dielectric constants and aspect ratios (w/h).

ε_e, increases with frequency as the fields become confined more to the substrate. Also, the real part of the characteristic impedance is plotted with respect to frequency. A drop-off in Z_0 is observed at the low end of the frequencies as frequency increases. This is due to both reduction of internal conductor inductances as charges move to the skin of the conductor and also to greater confinement (the same phenomenon as charge or current bunching) of the EM fields in the dielectric as frequency increases. It is not long before the characteristic impedance increases. This effect is not due to the skin effect and current bunching that were previously described. Rather it is due to other EM effects that are only captured in EM simulation. It is a result of spatial variations being developed in the fields (related to the fact that not all parts of the fields are in instantaneous contact). Figure 5-9(b) shows the imaginary parts of ε_e and Z_0. These imaginary parts are a result of loss, primarily loss in the conductors.

5.3.3   Filling Factor, q

Defining a filling factor, q, provides useful insight into the distribution of energy in an inhomogeneous transmission line. The effective microstrip permittivity is

$$\varepsilon_e = 1 + q(\varepsilon_r - 1), \tag{5.15}$$
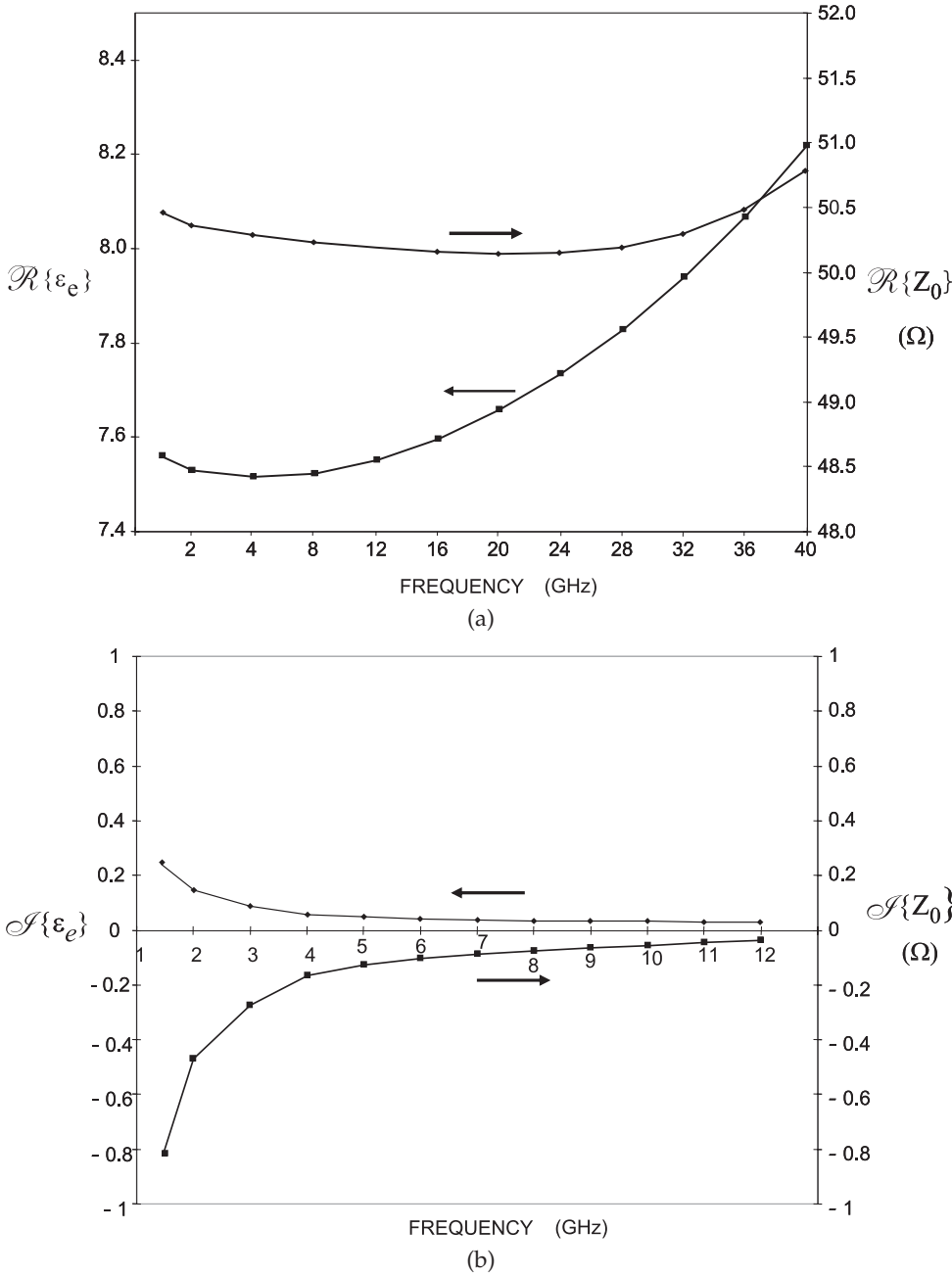
Figure 5-9 Frequency dependence of the real and imaginary parts of the effective permittivity, ε_e, and characteristic impedance Z_0 of a microstrip line on alumina with $\varepsilon_r(DC) = 9.9$, $w = 70\ \mu$m, $h = 500\ \mu$m: (a) real parts $\Re()$; and (b) imaginary parts $\Im()$.
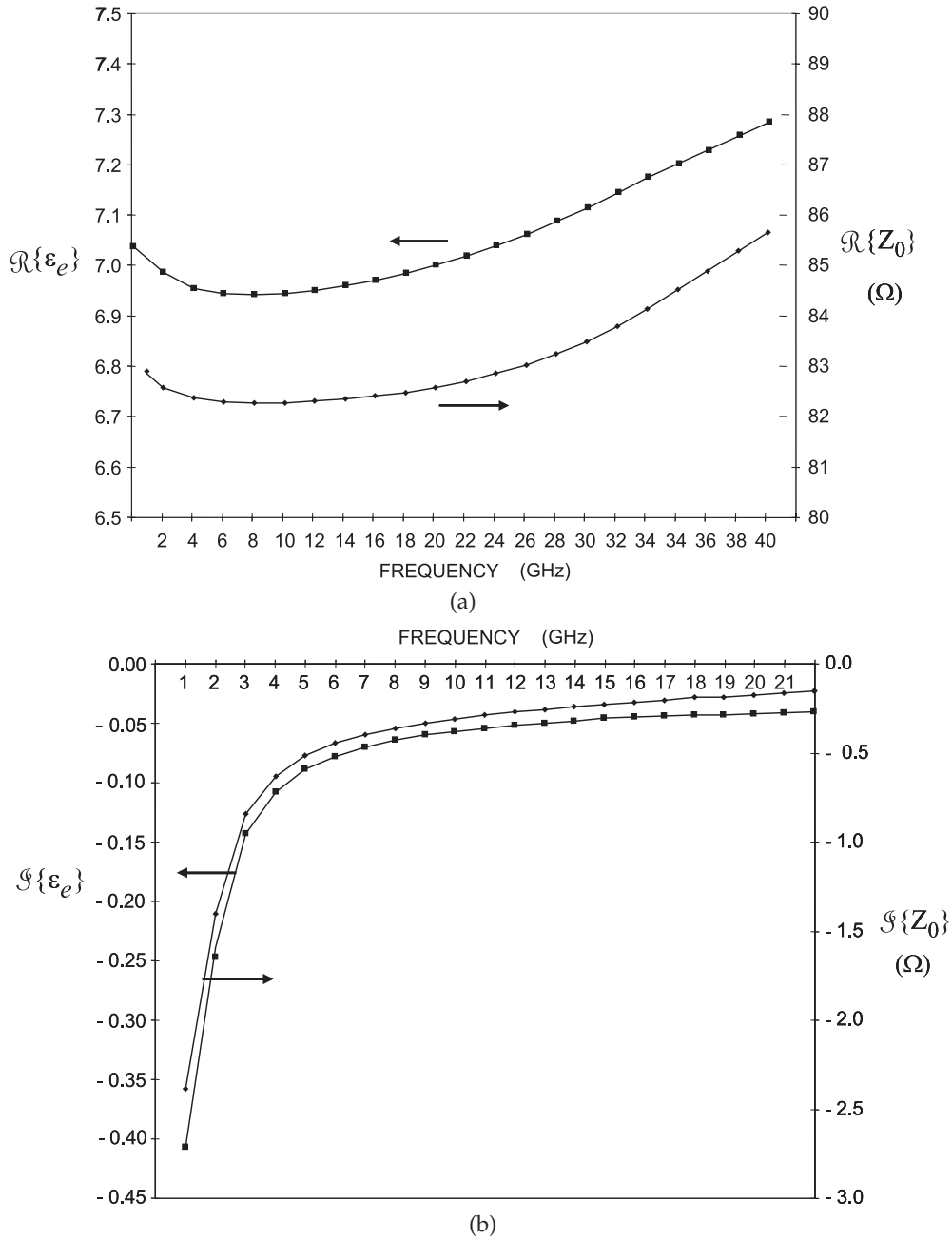
Figure 5-10 Frequency dependence of the effective permittivity, ε_e, and characteristic impedance Z_0 of a microstrip line with $\varepsilon_r(DC) = 9.9$, $w = 240$ μm, $h = 635$ μm: (a) real parts; and (b) imaginary parts.
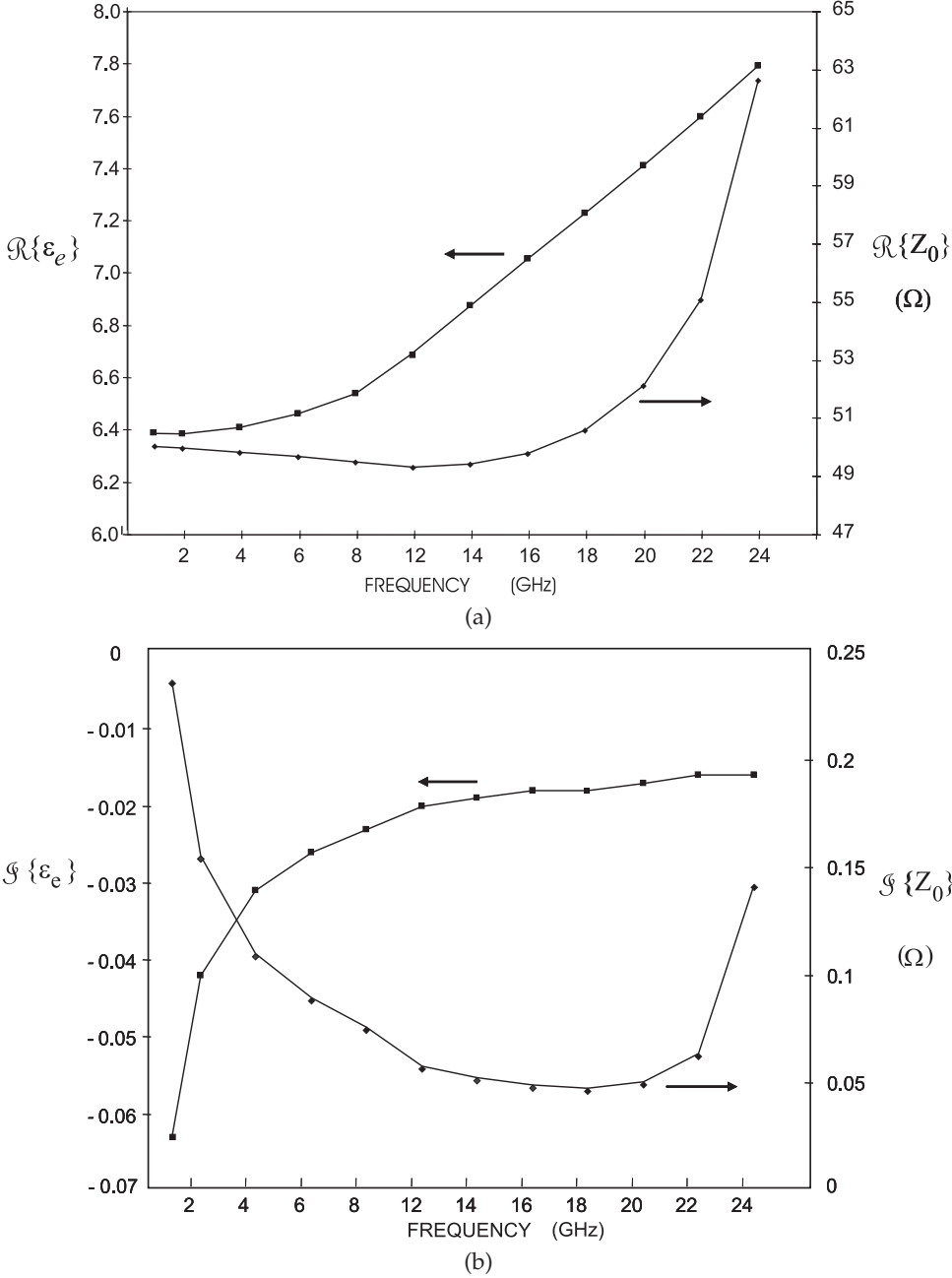
Figure 5-11 Frequency dependence of the effective permittivity, ε_e, and characteristic impedance Z_0 of a microstrip line with $\varepsilon_r(DC) = 9.9$, $w = h = 635$ μm: (a) real parts; and (b) imaginary parts.
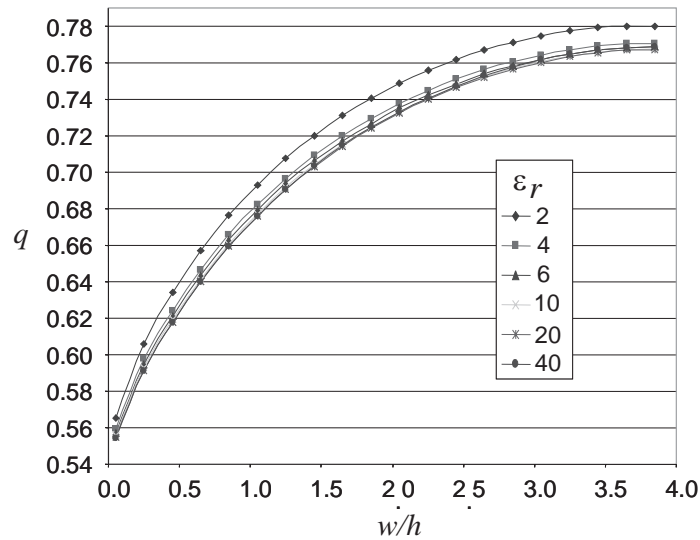
Figure 5-12   Dependence of the q factor of a microstrip line at 1 GHz for various dielectric constants and aspect (w/h) ratios. (Data obtained using SONNET.)

where q has the bounds

$$\frac{1}{2} \leq q \leq 1. \tag{5.16}$$

The useful aspect of q is that it is almost independent of ε_r. A q factor of 1 would indicate that all of the fields are in the dielectric region. The dependence of the q of a microstrip line at 1 GHz for various dielectric constants and aspect (w/h) ratios is shown in Figure 5-12, thus q is almost independent of the permittivity of the line. The properties of a microstrip line, and uniform transmission lines in general, can be described very well by considering the geometric filling factor, q, and the dielectric permittivity separately.

5.4   Multimoding on Transmission Lines

Multimoding is a phenomenon that affects the integrity of a signal as it travels on a transmission line. For transmission lines, multimoding occurs when there are two or more EM field configurations that can support a propagating wave. Different field configurations travel at different speeds so that the information traveling in the two modes will combine incoherently and, if the energy in the two modes is comparable, it will be impossible to discern the intended information being sent. It is critical that transmission line structures be designed to avoid multimoding. The most common mode on a transmission line is when there is no, or minimum, variation of the fields in the transverse direction (perpendicular to the direction

of propagation). There must of course be a variation in the longitudinal direction or else the wave will not propagate. The transmission structures of interest here, are those that have conductors to establish boundary conditions to guide a wave along an intended path. The conductors are electric walls that confine the fields. The lowest-order mode with minimum transverse field variations is called the transverse EM mode (TEM). Higher-order modes occur when the fields can vary. From here the discussion necessarily invokes EM theory. If you need to do this, see Appendix D on Page 847, where EM theory is reviewed specifically with respect to multimoding. One of the important concepts is that electric and magnetic walls impose boundary conditions to the fields. Electric walls are conductors, whereas a magnetic wall is formed approximately at the interface of two regions with different permittivity.

It is the property of EM fields that spatial variations of the fields cannot occur too quickly. This comes directly from Maxwell's equations which relate the spatial derivative (the derivative with respect to distance) of the electric field to the time derivative of the magnetic field. The same is true for spatial variation of the magnetic field and time variation of the electric field. How fast a field varies with time depends on frequency. How fast an EM field changes spatially, its curl, depends on wavelength relative to geometry and on boundary conditions. Without electric and magnetic walls establishing boundary conditions, as in free space, a full wavelength is required to obtain the lowest-order variation of the fields. With electric or magnetic walls, a smaller distance will be sufficient. Between two electric walls one-half wavelength of distance is required. The same is true for magnetic walls. With one electric wall and one magnetic wall a quarter-wavelength separation of the walls will support a higher-order mode. A general rule for avoiding multimoding is that critical transverse geometries must be kept to under a fraction of a wavelength (say, $< \lambda/2$ or $< \lambda/4$). Identification of exactly what the critical geometries are requires some understanding of EM fields, of Maxwell's equations, and of boundary conditions. Appendix D on Page 847 provides a review of Maxwell's equations directed at understanding multimoding and the distribution of fields on transmission lines.

One type of multimoding has already been described. In the previous chapter it was seen that the signals on a regular transmission line have two simple solutions that are interpreted as the forward-traveling and backward-traveling modes. Each **mode** is a possible solution of the differential equations describing the signals. The boundary conditions in the longitudinal direction are established by the source and load impedances, and so the variation can be any fraction of a wavelength. This section is concerned with other solutions to the equations describing the fields on a transmission line structure. In general, the other solutions arise when the transverse dimensions, such as the distance between the two

Table 5-1 Properties of the EM fields at electric and magnetic walls.

	E field	H field
Electric wall	Normal	Parallel
Magnetic wall	Parallel	Normal

conductors of a two-conductor transmission line, permits a variation of the fields. An EM treatment cannot be avoided if **spatial modes** on a transmission line are to be described. The dimensions of structures can be designed to avoid multimoding. Also, multimoding must be understood in resolving signal integrity problems that manifest themselves when circuits do not function correctly. The concern is that with multimoding various components of a signal travel at different velocities and generally combine at a load incoherently. Multimoding can be easily described mathematically for transmission line structures with uniform geometries.

The aim in this section is to understand moding and to develop an intuitive understanding of transmission line design and of microwave circuits in general. The boundary conditions established at electric and magnetic walls were derived in Section D.5 on Page 858. The properties of the EM walls are summarized in Table 5-1. These rules provide a quick way of understanding multimoding. Circuit structures such as transmission lines, substrate thicknesses, and related geometries are nearly always chosen so that only one solution of Maxwell's equations are possible. In particular, if the crosssectional dimensions of a transmission line are much less than a wavelength then it will be impossible for the fields to curl up on themselves and so perhaps there will be only one or, in some cases, no solutions to Maxwell's equations. The simplest illustration of this phenomena, which also happens to be particularly relevant to planar transmission lines, is signal propagation on the parallel-plate waveguide, shown in Figure 5-13. These rules are used in the next section to describe multimoding in parallel-plate waveguides.

5.5 Parallel-Plate Waveguide

The parallel-plate waveguide shown in Figure 5-13 is the closest regular structure to planar transmission lines such as microstrip. While the profile is not the same as a microstrip line, it is as close a structure as is available that still has a reasonably straightforward field solution that can be developed analytically. The aim here is to develop an understanding of the origins of multimoding and develop design guidelines that will enable transmission line structures to be designed to avoid multimoding. Intuition is essential in debugging circuits that are not working properly because of multimoding.

The development begins with **Maxwell's equations** (Equations (D.1)–
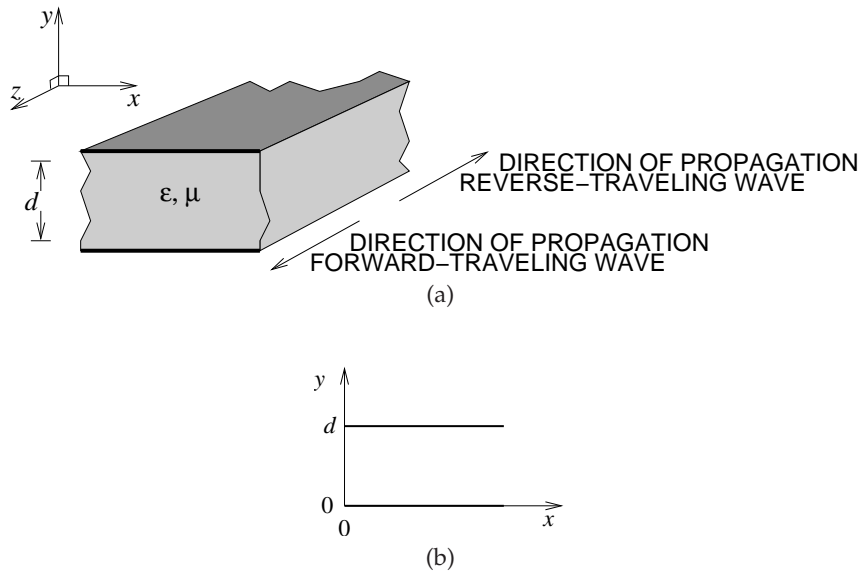
Figure 5-13 Parallel-plate waveguide: (a) three-dimensional view; and (b) crosssectional (transverse) view.

(D.4) on Page 848). A further simplification to the equations is to assume a linear, isotropic, and homogeneous medium, a uniform dielectric, so that ε and μ are independent of signal level and are independent of the field direction and of position. Thus

$$\nabla \times \mathbf{E} = \frac{\partial \mathbf{B}}{\partial t} \qquad (5.17)$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \qquad (5.18)$$

$$\nabla \cdot \mathbf{D} = \rho_V \qquad (5.19)$$

$$\nabla \cdot \mathbf{B} = 0, \qquad (5.20)$$

where ρ_V is the charge density and J is the current density. J and ρ_V will be zero except at an electric wall. The above equations do not include magnetic charge or magnetic current density. They do not actually exist and so a modified form of Maxwell's equations incorporating these is not necessary to solve the fields on a structure. They are invoked in Appendix D to draw an analogy between an electric wall at a conductor and a magnetic wall at the interface between two regions of different permittivity.

In the lateral direction, the parallel-plate waveguide in Figure 5-13 extends indefinitely, and this is indicated by the broken continuation of the top and bottom metal planes in Figure 5-13(b). For this regular
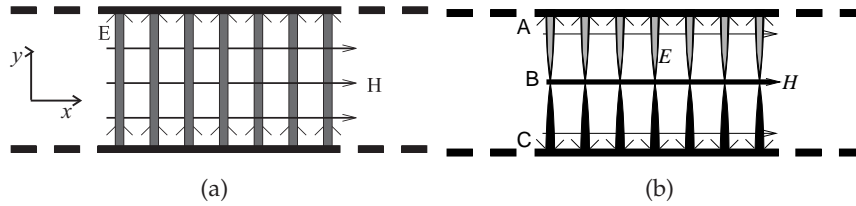
Figure 5-14 Parallel-plate waveguide showing electric, **E**, and magnetic, **H**, field lines: (a) the TEM mode; and (b) the TE mode. The electric (E) and magnetic (H) field lines are shown with the line thickness indicating field strength. The shading of the E field lines indicates polarity (+ or −) and the arrows indicate direction (and also polarity). The z direction is into the page.

structure, and with a few assumptions, the form of Maxwell's equations with multidimensional spatial derivatives can be simplified. One approach to solving differential equations, one that works very well, is to assume a form of the solution and then test to see if it is a valid solution. The first solution to be considered is called the TEM mode and corresponds to the minimum possible variation of the fields and will be the only possible solution when the interconnect transverse dimensions are small (relative to a wavelength). Also, it is assumed that the variation in the z direction is described by the traveling-wave equations. So the only fields of interest here are **E** and **H** in the transverse plane; all that is seen in Figure 5-13(b). If all the fields are in the xy plane, then it is sufficient to apply just the boundary conditions that come from the top and bottom ground planes. At first it appears that there are many possible solutions to the differential equations. This is simplified by assuming certain variational properties of the E_x, E_y, H_x, and H_y fields and then seeing if these solutions can be supported. The simplest solution is when there is no variation in the fields and then the only possible outcome is that $E_x = 0 = H_y$. This is the TEM mode indicated in Figure 5-14(a). At the boundaries, the top and bottom metal planes, there is a divergence of the electric field, as immediately inside the (ideal) conductor there is no electric field and immediately outside there is. This divergence is supported by the surface charge on the ground planes (see Equation (D.3)).

In Figure 5-14(a), the thickness of the lines indicates relative field strength and there is no variation in field strength in either the electric or magnetic fields shown. The coefficients of the field components are determined by the boundary conditions. The trial solution used here has $E_x = 0 = H_y$ and $E_z = 0 = H_z$; that is, the trial solution has the electric field in the y direction and the magnetic field is in the x direction. Maxwell's equations (Equations (5.17)–(5.20)) become

$$\nabla \times E_y \hat{\mathbf{y}} = \frac{\partial B_x \hat{x}}{\partial t} \tag{5.21}$$

$$\nabla \times H_x \hat{\mathbf{x}} = \frac{\partial D_y \hat{x}}{\partial t} + \mathbf{J} \tag{5.22}$$

$$\nabla \cdot D_y \hat{\mathbf{y}} = \rho \qquad (5.23)$$

$$\nabla \cdot B_x \hat{\mathbf{x}} = 0. \qquad (5.24)$$

Expanding the curl, $\nabla \times$, and div, $\nabla \cdot$, operators using Equations (D.15) and (D.16) these become (with $D = \varepsilon E$ and $B = \mu H$)

$$-\hat{x}\frac{\partial E_y}{\partial z} = \frac{\partial B_x \hat{x}}{\partial t} = \frac{\partial \mu H_x \hat{x}}{\partial t} \qquad (5.25)$$

$$\hat{y}\frac{\partial H_x}{\partial z} = \frac{\partial D_y \hat{y}}{\partial t} + \mathbf{J} = \frac{\partial \varepsilon E_y \hat{y}}{\partial t} + \mathbf{J} \qquad (5.26)$$

$$\frac{\partial D_y}{\partial y} = \frac{\partial \varepsilon E_y}{\partial y} = \rho \qquad (5.27)$$

$$\frac{\partial B_x}{\partial x} = \frac{\partial \mu H_x}{\partial x} = 0. \qquad (5.28)$$

These equations describe what happens at each point in space. Equation (5.25) indicates that if there is a time-varying component of the x-directed B field then there must be a z-varying component of the E_y field component. This is just part of the wave equation describing a field propagating in the z direction. Equation (5.26) indicates the same thing, but now the roles of the electric and magnetic fields are reversed. Equation (5.25) shows that the y component of the electric field can be constant between the plates, but at the plates there must be charge on the surface of the conductors (see Equation (5.27)),to terminate the electric field (as indicated by Equation (5.27)), as there is no electric field inside the conductors (assuming that they are perfect conductors for the moment). Since magnetic charges do not exist, captured by the zero on the right-hand side of Equation (5.28), Equation (5.28) indicates that the x component of the magnetic field cannot vary in the x direction (i.e., B_x and H_x are constant). Thus the assumption behind the trial solution about the form of this mode is correct. The electric and magnetic field are uniform in the transverse plane, the xy plane, and the only variation is in the direction of propagation, the z direction. Thus the test solution satisfies Maxwell's equations. This is the TEM mode, as all field components are in the x and y directions and none are in the z direction. The TEM mode can be supported at all frequencies in the parallel-plate waveguide.

Putting Equations (5.25)–(5.28) in phasor form, and considering the source free region between the plates (so $J = 0$ and $\rho = 0$), these become

$$-\frac{\partial E_y}{\partial z} = \jmath\omega\mu H_x \qquad (5.29)$$

$$\frac{\partial H_x}{\partial z} = \jmath\omega\varepsilon E_y \qquad (5.30)$$

$$\frac{\partial E_y}{\partial y} = 0 \qquad (5.31)$$

$$\frac{\partial H_x}{\partial x} = 0 \tag{5.32}$$

and the solution becomes

$$\frac{\partial^2 E_y}{\partial z^2} = \gamma^2 E_y \tag{5.33}$$

$$\frac{\partial^2 H_x}{\partial z^2} = \gamma^2 H_x\,, \tag{5.34}$$

where $\gamma = \jmath\omega\sqrt{\mu\varepsilon}$ and there is no variation of the E_y component of the electric field in the y direction and no variation of the H_x component of the magnetic field in the x direction.

Another possible set of modes occurs when the electric field is only in the y direction, but then there must be a variation of the field strength, as shown in Figure 5-14(b). The full EM development of the fields is given in Appendix E and here we summarize the results qualitatively. The simplest variation is when there is a half-sinusoidal spatial variation of the E_y field component. Applying the methodology described above, it is found that there must be a component of the magnetic field in the z direction to support this mode. Hence these modes are called **Transverse Electric** (TE) modes. (Interchanging the roles of the electric and magnetic fields yields the **Transverse Magnetic** (TM) modes where there is an electric field component in the z direction.) The half-sinusoidal variation still enables the charge to support the existence of an E_y electric field. A key result from our previous discussion is that there must be enough distance for the field to curl by the half-sinusoidal spatial variation and this is related to wavelength, λ. This transverse electric mode can only exist when $h \geq \lambda/2$. When h is smaller than one-half wavelength, this mode cannot be supported, and is said to be cut off. Only the TEM mode can be supported all the way down to DC, so modes other than TEM have a cutoff frequency, f_c, and a cutoff wavelength, λ_c. The concept of wavenumber $k(= 2\pi/\lambda = \omega\sqrt{\mu\varepsilon})$ is also used, and for the lowest TE mode, $k_c = 2\pi/\lambda_c$ with $\lambda_c = 2h$. The cutoff wavelength, λ_c, and the cutoff wavenumber, k_c, are both related to the dimension below in which a mode cannot "curl" sufficiently to be self-supporting. In general, for TE modes, there can be n variations of the electric field, and so we talk about the nth TE mode, denoted as TE$_n$, for which $k_{c,n} = n\pi/h$ and $\lambda_c = 2h/n$. Also note that $k_c = 2\pi/\lambda_c$. The propagation constant of any mode in a uniform lossless medium (not just in a parallel-plate waveguide mode) is

$$\beta = \sqrt{k^2 - k_c^2}. \tag{5.35}$$

For the TEM mode we have $k_c = 0$. High-order modes are described by their cutoff wavenumber, k_c.

Transverse magnetic (TM) modes are similarly described, and again $k_c = n\pi/h$ for the TM$_n$ mode. A mode can be supported at any frequency above

the cutoff frequency for the mode. It just cannot be supported at frequencies below the cutoff frequency, as it is not possible for the fields to vary (or curl) below cutoff.

5.5.1 Multimoding and Electric and Magnetic Walls

In the above discussion, the parallel-plate waveguide had two electric walls—the top and bottom metal walls. Without going into much detail, results will be presented when magnetic walls are introduced. A magnetic wall can only be approximated, as ideal magnetic conductors do not exist (since magnetic charges do not exist). Whereas an electric wall appears as a short circuit, a magnetic wall is an open circuit, so an open-circuited coaxial line appears to have a transverse magnetic wall at the open circuit. Maxwell's equations impose the following boundary conditions:

Electric wall	Perpendicular electric field
	Tangential magnetic field
Magnetic wall	Perpendicular magnetic field
	Tangential electric field

The lowest-order modes that can be supported by combinations of electric and magnetic walls are shown in Figure 5-15. With two electric or two magnetic walls, a TEM mode (having no field variations in the transverse plane) can be supported. Of course, there will be variations in the field components in the direction of propagation. The modes with the simplest geometric variations in the plane transverse to the direction of propagation establish the critical wavelength. In Figure 5-15 the distance between the walls is d. For the case of two like walls (Figures 5-15(a) and 5-15(c)), $\lambda_c = 2h$, as one-half sinusoidal variation is required. With unlike walls (see Figure 5-15(b)), the varying modes are supported with just one-quarter sinusoidal variation, and so $\lambda_c = 4h$.

5.6 Microstrip Operating Frequency Limitations

Different types of higher-order modes can exist with microstrip and the two maximum operating frequencies of microstrip lines are the (a) the lowest-order TM mode, and (b) the lowest-order transverse microstrip resonance mode. In practice, **multimoding** is a problem when two conditions are met. First it must be possible for higher-order field variations to exist, and second, that energy can be effectively coupled into the higher-order mode. Generally this requires significant discontinuity on the line or that the phase velocities of two modes coincide. An early discussion said that the phase velocities of two modes would be different and this would be true when the dielectric is uniform. However, with a nonhomogeneous line like microstrip there
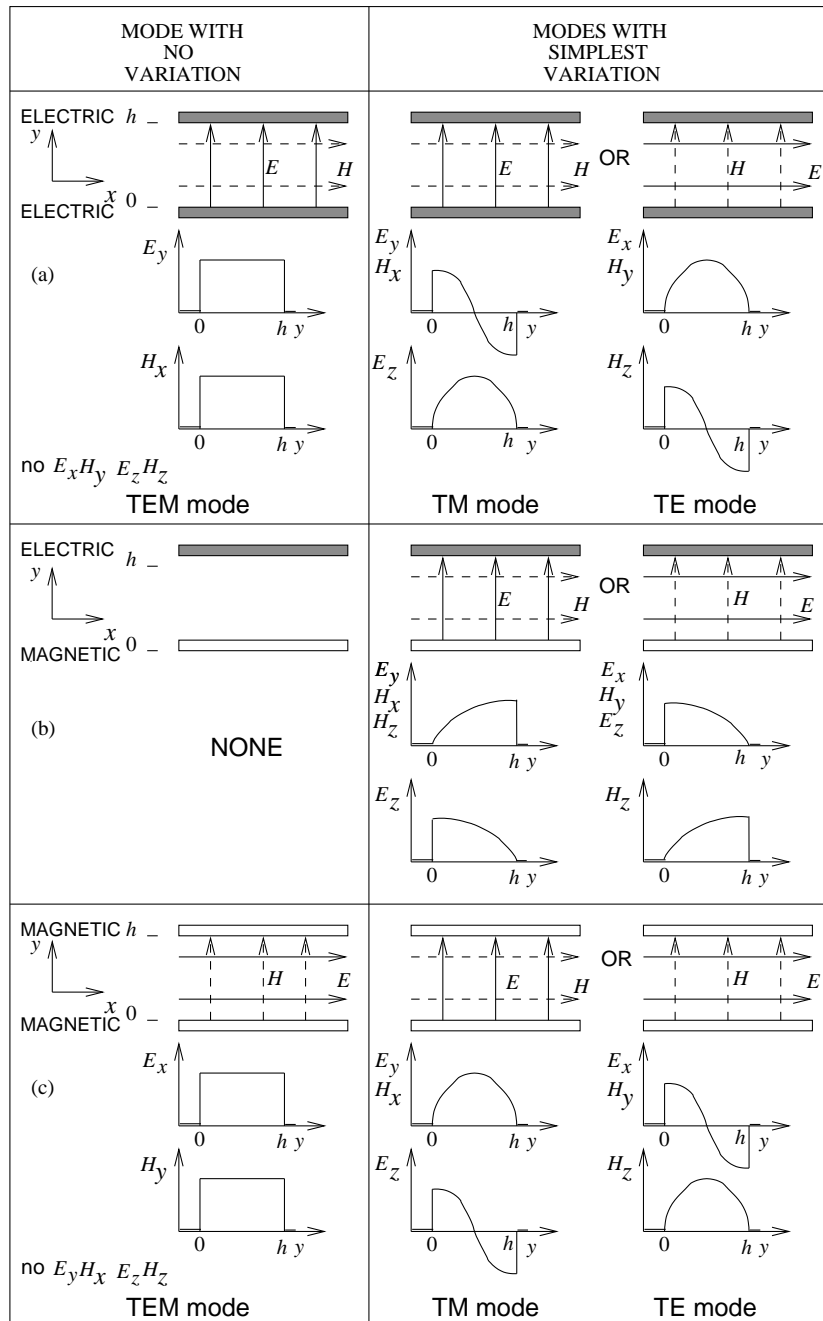
Figure 5-15 Lowest-order modes supported with combinations of electric and magnetic walls.

can be frequencies where the phase velocities of two modes can coincide. Since discontinuities are inevitable it is always a good idea to use the first consideration. As was noted previously, the modal analysis that was possible with the parallel-plate waveguide cannot be repeated easily for the microstrip line because of the irregular crosssection, but the phenomena is very similar. Instead of a TEM mode, there is a quasi-TEM mode and there are TE and TM modes.

5.6.1 Microstrip Dielectric Mode

A dielectric on a ground plane with an air region (of a wavelength or more above it) can support a TM mode, generally called a microstrip dielectric mode, substrate mode, or slab mode. The microstrip dielectric mode is a problem for narrow microstrip lines. However, only at high frequencies does it become important. Whether this mode exists in a microstrip environment depends on whether energy can be coupled from the quasi-TEM mode (which is always generated) of the microstrip line into the TM dielectric mode. The critical frequency at which the TM mode becomes important is when there is significant coupling. Coupling is a problem with a microstrip line having a narrow strip, as the field orientations of the quasi-TEM mode and the dielectric mode align. Also, coupling occurs when the phase velocities of the two modes coincide. A detailed analysis reported in Edwards and Steer [51, page 143], and Vendelin [74], shows that this occurs at the first critical frequency,

$$f_{c1} = \frac{c \tan^{-1}(\varepsilon_r)}{\sqrt{2}\pi h \sqrt{\varepsilon_r - 1}} \, . \tag{5.36}$$

So at f_{c1}, in Equation (5.36), the dielectric mode will be generated even if there is not a discontinuity. If there is a discontinuity, say a split of one microstrip line into two microstrip lines, multimoding will occur when the dielectric mode can exist. From Figure 5-15(b), the dielectric slab mode can be supported when $h > \lambda_g/4$, where λ_g is the wavelength in the dielectric. Now $\lambda_g = \lambda_0/\sqrt{\varepsilon_r} = c/(f\sqrt{\varepsilon_r})$, so the second critical frequency is

$$f_{c2} = \frac{c}{4h\sqrt{\varepsilon_r}} \, . \tag{5.37}$$

The development here assumes that the interface between the dielectric and air forms a good magnetic wall. With a dielectric having a permittivity of 10, typical for microwave circuits, the effective value of h would be increased by up to 10%. However, it is difficult to place an exact value on this.

In summary, f_{c2} is the lowest frequency at which the dielectric mode will exist if there is a discontinuity, and f_{c1} is the lowest frequency at which the dielectric mode will exist if there is not a discontinuity.

EXAMPLE 5. 1   Dielectric Mode

The strip of a microstrip has a width of 1 mm and is fabricated on a lossless substrate that is 2.5 mm thick and has a relative permittivity of 9. At what frequency does the substrate (or slab) mode first occur?

$|\!\leftarrow w \rightarrow\!|$

h

$\varepsilon_r = 9$

$w + 0.4h$
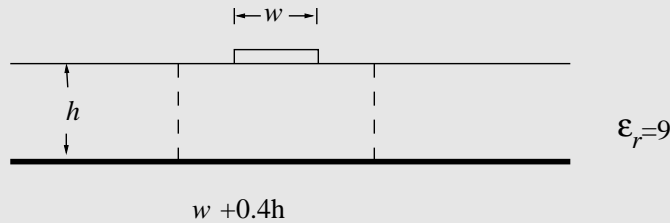
Solution:

There are two frequencies that must be considered. One that comes from the dimensions of the dielectric slab and the other from considerations of matching phase velocities. From phase velocity consideration, the development behind Equation (5.36), the first critical frequency is

$$f_{c1} = \frac{c \tan^{-1}(\varepsilon_r)}{\sqrt{2}\pi h \sqrt{\varepsilon_r - 1}} = 13.9 \text{ GHz}. \tag{5.38}$$

The first slab mode occurs when a variation of the magnetic or electric field can be supported between the ground plane and the approximate magnetic wall supported by the dielectric/free-space interface; that is, when $h = \frac{\lambda}{4} = \lambda_0/(4\sqrt{9}) = 2.5$ mm $\Rightarrow \lambda_0 = 3$ cm. Thus the second slab mode critical frequency is

$$f_{c2} = 10 \text{ GHz}. \tag{5.39}$$

Since discontinuities cannot be avoided (you cannot build an interesting circuit with just a transmission line), f_{c2} is the critical frequency to use.

5.6.2 *Higher-Order Microstrip Mode*

If the crosssectional dimensions of a microstrip line are smaller than a fraction of a wavelength, then the electric and magnetic field lines will be as shown in Figure 4-4 on Page 169. These field lines have the minimum possible spatial variation and the fields are almost entirely confined to the transverse plane; this mode is called the quasi-TEM microstrip mode. However, as the frequency of the signal on the line increases it is possible for these fields to have one-quarter or one-half sinusoidal variation. Determining the frequency at which this higher-order mode can be supported requires more involved intuition than is appropriate to develop here.

The following is a summary of a more complete discussion on Operating Frequency Limitations in Edwards and Steer [51]. Some variations of the fields or modes do not look anything like the field orientations shown in

Figure 4-4. However, the variations that are closest to the quasi-TEM mode are called higher-order microstrip modes and the one that occurs at the lowest frequency corresponds to a half-sinusoidal variation of the electric field between the edge of the strip and the ground plane. This path is a little longer than the path directly from the strip to the ground plane. However, for a wide strip, most of the EM energy is between the strip and the ground plane (both of which are electric walls) with approximate magnetic walls on the side of the strip. The modes are then similar to the parallel-plate waveguide modes described in Appendix D on Page 847. The next highest microstrip mode (or parallel-plate TE mode) occurs when there can be a half-sinusoidal variation of the electric field between the strip and the ground plane. This corresponds to Figure 5-15(a). However, for finite-width strips the first higher-order microstrip mode occurs at a lower frequency than implied by the parallel-plate waveguide model. In reality, multimoding occurs at a slightly lower frequency than would be implied by this model. This is because the microstrip fields are not solely confined to the dielectric region, and in fact the electric field lines do not follow the shortest distance between the strip and the ground plane. Thus the fields along the longer paths to the sides of the strip can vary at a lower frequency than if we considered only the direct path. With detailed EM modeling and with experimental support it has been established that the first higher-order microstrip mode can exist at frequencies greater than [51, page 143]:

$$f_{\text{Higher}-\text{Microstrip}} = \frac{c}{4h\sqrt{\varepsilon_r - 1}} \, . \qquad (5.40)$$

This is only an approximate guide, and it is best to use the thinnest substrate possible. Most EM software programs used to model microstrip and other planar transmission lines report when higher-order moding can occur.

EXAMPLE 5. 2   Higher-Order Microstrip Mode

The strip of a microstrip has a width of 1 mm and is fabricated on a lossless substrate that is 2.5 mm thick and has a relative permittivity of 9. At what frequency does the first higher microstrip mode first propagate?

$w + 0.4h$
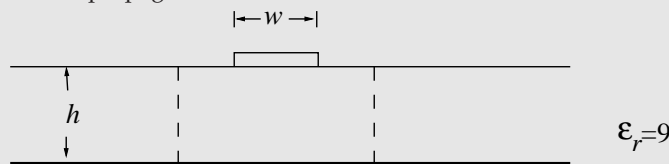
Solution:

The higher-order microstrip mode occurs when a half-wavelength variation of the electric field between the strip and the ground plane can be supported. When $h = \lambda/2 = \lambda_0/(3 \cdot 2) = 2.5$ mm; that is, the mode will occur when $\lambda_0 = 15$ mm. So
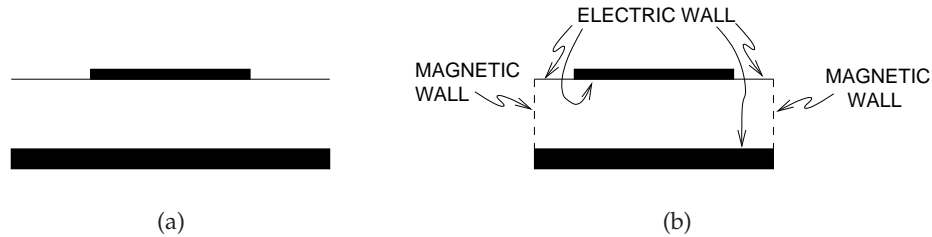
Figure 5-16 Approximation of a microstrip line as a waveguide: (a) crosssection of microstrip; and (b) model with magnetic and electric walls.

$$f'_{\text{Higher}-\text{Microstrip}} = 20 \text{ GHz}. \tag{5.41}$$

A better estimate of the frequency where the higher-order microstrip mode becomes a problem is given by Equation (5.40):

$$f_{\text{Higher}-\text{Microstrip}} = c/(4h\sqrt{\varepsilon_r - 1}) = 10.6 \text{ GHz}. \tag{5.42}$$

So two estimates have been calculated for the frequency at which the first higher-order microstrip mode can first exist. The estimate in Equation (5.41) is approximate and is based on a half-wavelength variation of the electric field confined to the direct path between the strip and the ground plane. Equation (5.42) is more accurate as it considers that on the edge of the strip the fields follow a longer path to the ground plane. It is the half-wavelength variation on this longer path that determines if the higher-order microstrip mode will exist. Thus the more precise determination yields a lower critical frequency.
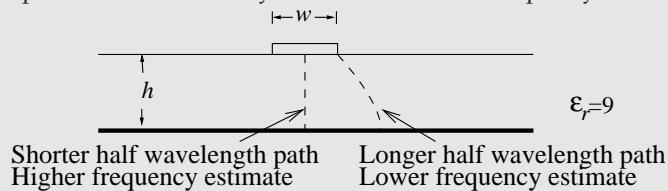
5.6.3 Transverse Microstrip Resonance

For a wide microstrip line, a transverse resonance mode can exist. This is the mode that occurs when EM energy bounces between the edges of the strip with the discontinuity at the strip edges forming a weak boundary. This is illustrated in Figure 5-16, where the microstrip shown in crosssection in Figure 5-16(a) is approximated as a rectangular waveguide in Figure 5-16(b) with magnetic walls on the sides and an extended electrical wall on the sides of the strip. The transverse resonance mode corresponds to the lowest-order H field variation between the magnetic walls. At the cutoff frequency for this transverse-resonant mode the equivalent circuit is a resonant transmission line of length $w + 2d$, as shown in Figure 5-17, where $d = 0.2h$ accounts for the microstrip side fringing. A half-wavelength must
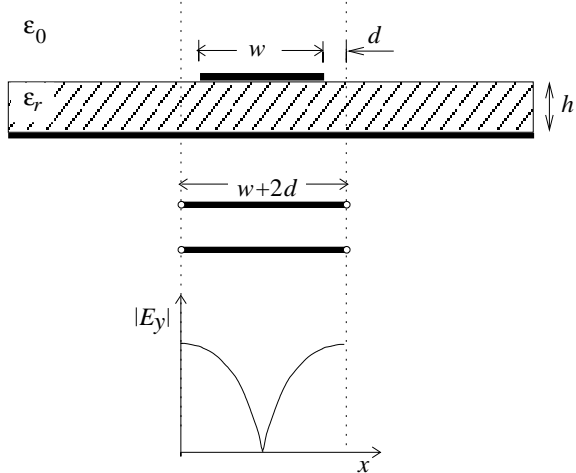
Figure 5-17 Transverse resonance: standing wave ($|E_y|$) and equivalent transmission line of length $w + 2d$ where $d = 0.2h$.

be supported by the length $w + 2d$. Therefore the cutoff half-wavelength is

$$\frac{\lambda_c}{2} = w + 2d = w + 0.4h \tag{5.43}$$

or

$$\frac{c}{2f_c\sqrt{\varepsilon_r}} = w + 0.4h. \tag{5.44}$$

Hence

$$f_c = \frac{c}{\sqrt{\varepsilon_r}\,(2w + 0.8h)}. \tag{5.45}$$

EXAMPLE 5. 3   Transverse Resonance Mode

The strip of a microstrip has a width of 1 mm and is fabricated on a lossless substrate that is 2.5 mm thick and has a relative permittivity of 9.
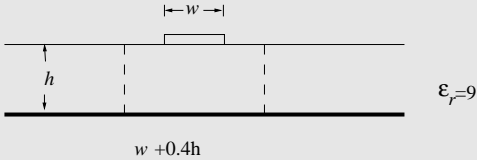
(a) At what frequency does the transverse resonance first occur?

(b) What is the operating frequency range of the microstrip line?

Solution: $h = 2.5$ mm, $w = 1$ mm , $\lambda = \lambda_0/\sqrt{\varepsilon_r} = \lambda_0/3$

(a) The magnetic waveguide model of Figure 5-16 can be used in estimating the frequency at which this occurs. The frequency at which the first transverse resonance mode occurs is when there is a full half-wavelength variation of the magnetic field between the magnetic walls, that is, the first transverse resonance when $\omega + 0.4h = \lambda/2 = 2$ mm:

$$\frac{\lambda_0}{3 \cdot 2} = 2 \text{ mm} \Rightarrow \lambda_0 = 12 \text{ mm},$$

and so

$$f'_{\text{TRAN}} = 25 \text{ GHz} . \qquad (5.46)$$

(b) All of the critical multimoding frequencies must be considered here and the minimum taken. The slab mode critical frequencies are f_{c1} (Equation (5.38)) and f_{c2} (Equation (5.39)); the higher-order mode critical frequency is $f_{\text{High−Microstrip}}$ (Equation (5.42)); and the transverse resonance frequency (Equation (5.46)). So the operating frequency range is DC to 10 GHz.

5.6.4 Summary of Multimoding on Microstrip Transmission Lines

There are four principal higher-order modes that need to be considered with microstrip transmission lines.

Mode	Critical frequency
Dielectric (or substrate) mode with no discontinuity	Equation (5.36)
Dielectric (or substrate) mode with discontinuity	Equation (5.39)
Higher Order Microstrip Mode	Equation (5.42)
Transverse Resonance Mode	Equation (5.45)

5.7 Power Losses and Parasitic Effects

Four separate mechanisms lead to power losses in microstrip lines:

(a) Conductor losses

(b) Dissipation in the dielectric of the substrate

(c) Radiation losses

(d) Surface-wave propagation.

The first two items are dissipative effects, whereas radiation losses and surface-wave propagation are essentially parasitic phenomena. The reader is directed to Edwards and Steer [51] for an extensive treatment. Here, summary results are presented.

Conductor losses greatly exceed dielectric losses for most microstrip lines fabricated on low-loss substrates. Lines fabricated on low-resistivity silicon wafers, however, can have high dielectric loss. These wafers are most commonly used for digital circuits and the interconnect transverse dimensions are generally very small so that line resistance is very high, and again, resistive losses dominate.

Radiation from a microstrip line results from asymmetric structures. In particular, discontinuities such as abruptly open-circuited microstrip (i.e., open ends), steps, and bends will all radiate to a certain extent. Such discontinuities form essential features of a microwave IC and therefore radiation cannot be avoided. Efforts must be made to reduce such radiation and its undesirable effects. In circuits such as filters, amplifiers, etc., this radiation is an acknowledged nuisance. In most cases, radiation can be represented as shunt admittances.

Surface waves, trapped just beneath the surface of the substrate dielectric, will propagate away from microstrip discontinuities as TE and TM modes. The effect of surface waves can also be treated as shunt conductance.

Various techniques can be used to suppress radiation and surface waves:

(a) Metallic shielding or "screening."

(b) The introduction of lossy (i.e., absorbent) material near any radiative discontinuity.

(c) The utilization of compact, planar, inherently enclosed circuits such as inverted microstrip and stripline.

(d) Reducing the current densities flowing in the outer edges of any metal sections and concentrate currents toward the center and in the middle of the microstrip.

(e) Possibly shape the discontinuity in some way to reduce the radiative efficiency.

5.8 Lines on Semiconductor Substrates

Propagation on transmission line structures fabricated on semiconductor substrates can have peculiar behavior. The interest in such lines is in relation to **Monolithic Microwave Integrated Circuits** (**MMICs**) using both silicon (Si) and compound semiconductor technologies (especially GaAs). More specifically, interconnects on **Metal-Oxide-Semiconductor** (**MOS**) or **Metal-Insulator-Semiconductor** (**MIS**) systems, where an insulating layer exists between the conductor and the semiconductor wafer (see Figure 5-18(a)), are of particular interest due to their ability to support a "**slow-wave**." These structures find major use in distributed elements on chip at RF. In particular, a silicon substrate can have a significant impact on microstrip
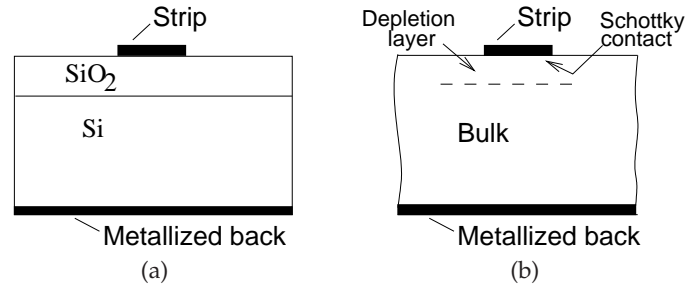
Figure 5-18 Transmission lines on silicon semiconductor: (a) silicon-silicon dioxide sandwich; and (b) bulk view.
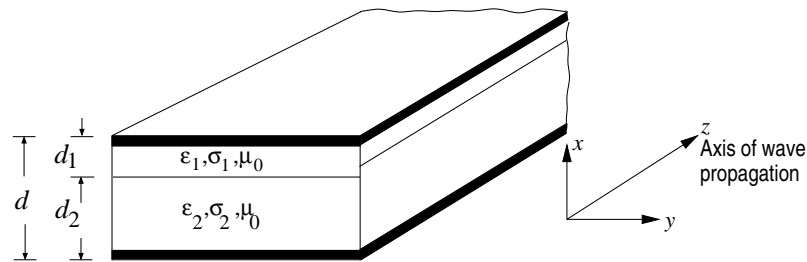
Figure 5-19 Parallel-plate transmission line structure.

propagation that derives from the charge layer formed at the silicon-silicon dioxide interface. The slow-wave effect is utilized in delay lines, couplers, and filters. With Schottky contact lines, the effect is used in variable-phase shifters, voltage-tunable filters, and various other applications.

Now, an intuitive explanation of the propagation characteristics of microstrip lines of this type can be based on the parallel-plate structure shown in Figure 5-19. For an exact EM analysis of the slow-wave effect with silicon substrates see Hasegawa et al. [75]. As well as developing a very useful approximation for the important situation of transmission lines on a semiconductor, the treatment below indicates the type of approach that can be used to analyze unusual structures. In this structure, assume a quasi-TEM mode of propagation. In other words, the wave propagation parameters α, β, and Z_0 can be deduced from electrostatic and magnetostatic solutions for the per unit length parameters C, G, L, and R.

The analysis begins with a treatment of the classic **Maxwell-Wagner capacitor**. Figure 5-20(a) shows the structure of such a capacitor where there are two different materials between the parallel plates of the capacitor with different dielectric constants and conductivity. The equivalent circuit is shown in Figure 5-20(b) with the elements given by

$$C_1 = \varepsilon_1 \frac{A}{d_1}, \ R_1 = \frac{1}{\sigma_1} \frac{d_1}{A} \tag{5.47}$$
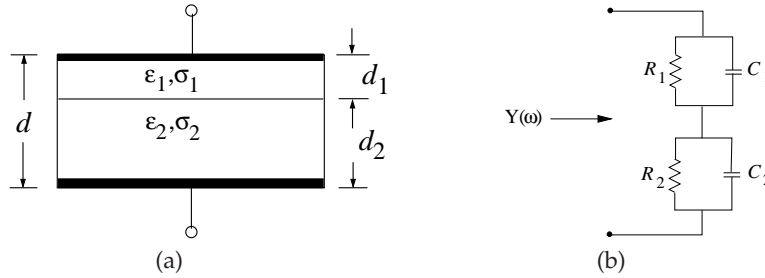
Figure 5-20 Structure of the Maxwell-Wagner capacitor: (a) structure; and (b) its equivalent circuit.

$$C_2 = \varepsilon_2 \frac{A}{d_2}, \ R_2 = \frac{1}{\sigma_2} \frac{d_2}{A} \tag{5.48}$$

$$\tau_1 = R_1 C_1 = \frac{\varepsilon_1}{\sigma_1} \tag{5.49}$$

$$\tau_2 = R_2 C_2 = \frac{\varepsilon_2}{\sigma_2}. \tag{5.50}$$

The admittance of the entire structure at radian frequency ω is

$$Y(\omega) = \frac{1}{R_1 + R_2} \frac{(1 + \jmath\omega\tau_1)(1 + \jmath\omega\tau_2)}{1 + \jmath\omega\tau}, \tag{5.51}$$

where

$$\tau = \frac{R_1\tau_2 + R_2\tau_1}{R_1 + R_2}. \tag{5.52}$$

Introducing

$$Y(\omega) = \jmath\omega \left(\varepsilon_e^* \frac{A}{d} \right), \tag{5.53}$$

the effective complex permittivity, $\varepsilon_e^* = \varepsilon_e' - \jmath\varepsilon_e''$, can be defined in terms of Equation (5.51). Using Equations (5.51) and (5.53) yields

$$\varepsilon_e' = \frac{\tau_1 + \tau_2 - \tau + \tau\tau_1\tau_2\omega^2}{(R_1 + R_2)\left[1 + (\omega\tau)^2\right]} \frac{d}{A}. \tag{5.54}$$

Consider now the case when R_1 goes to infinity (i.e., layer 1 in Figure 5-20 is an insulator). For this case, Equation (5.54) becomes

$$\varepsilon_e' = \frac{1 + \left(1 + \frac{d_2}{d_1}\frac{\varepsilon_1}{\varepsilon_2}\right)\left(\frac{\omega\varepsilon_2}{\sigma_2}\right)^2}{1 + \left(1 + \frac{d_2}{d_1}\frac{\varepsilon_1}{\varepsilon_2}\right)^2 \left(\frac{\omega\varepsilon_2}{\sigma_2}\right)^2} \left(\varepsilon_1 \frac{d}{d_1}\right). \tag{5.55}$$

It is clear from Equations (5.54) and (5.55) that the effective complex permittivity has a frequency-dependent component. Consider how this

varies with a few cases of ω. For $\omega = 0$ (static value),

$$\varepsilon'_{e,0} = \varepsilon_1 \frac{d}{d_1}. \tag{5.56}$$

For the case where ω goes to infinity (the optical value), the real part of the effective permittivity is

$$\varepsilon'_{e,\infty} = \frac{\varepsilon_1 \varepsilon_2 (d_1 + d_2)}{\varepsilon_2 d_1 + \varepsilon_1 d_2}. \tag{5.57}$$

Note also that the value of $\varepsilon'_{e,\infty}$ can be approximately achieved for a large value of $\omega \varepsilon_2 / \sigma_2$ (i.e., low-conductivity substrates can be used to ensure that the displacement currents dominate). In a similar way, it is clear that $\varepsilon'_{e,0}$ can be achieved by having a small value of $\omega \varepsilon_2 / \sigma_2$. Also note that $\varepsilon'_{e,0}$ can be made very large by making d_1 much smaller than d.

EXAMPLE 5. 4   Two-Layer Substrate

Consider the structure in Figure 5-19. Determine the guide wavelength, λ_g, and the wavelength in the insulator, λ_1, at a frequency of 1 GHz. SiO_2 is the dielectric, with permittivity $\varepsilon_1 = 4\varepsilon_0$. The depths d_2 and d_1 of the two dielectrics are $d_2 = 250$ μm and $d_1 = 0.1$ μm.

Solution:

$$\lambda_1 \quad = \quad \frac{3 \times 10^8}{10^9 \sqrt{4}} = 0.15 \text{ m} = 15 \text{ cm}, \quad \lambda = 15 \sqrt{\frac{0.1}{250.1}} \text{cm} \approx 3 \text{ mm}. \tag{5.58}$$

5.8.1 Modes on the MIS (MOS) Line

The previous description of the properties of a Maxwell-Wagner capacitor leads to a discussion of the possible modes on the MIS (MOS) line. To make the problem tractable, the transmission line shown in Figure 5-18(a) will be approximated as having the crosssection shown in Figure 5-21(a).

Dielectric Quasi-TEM Mode

The first possible mode is the dielectric quasi-TEM mode, for which the sectional equivalent circuit model of Figure 5-21(b) is applicable. In this mode $\sigma_2 \ll \omega \varepsilon_2$. This implies from our earlier discussion that $\varepsilon'_e = \varepsilon'_{e,\infty}$ and $\mu'_e = \mu_0$. Thus the per unit length parameters are

$$L = \mu_0 \frac{d_1}{W} \tag{5.59}$$

$$C_1 = \varepsilon_1 \frac{W}{d_1} \tag{5.60}$$
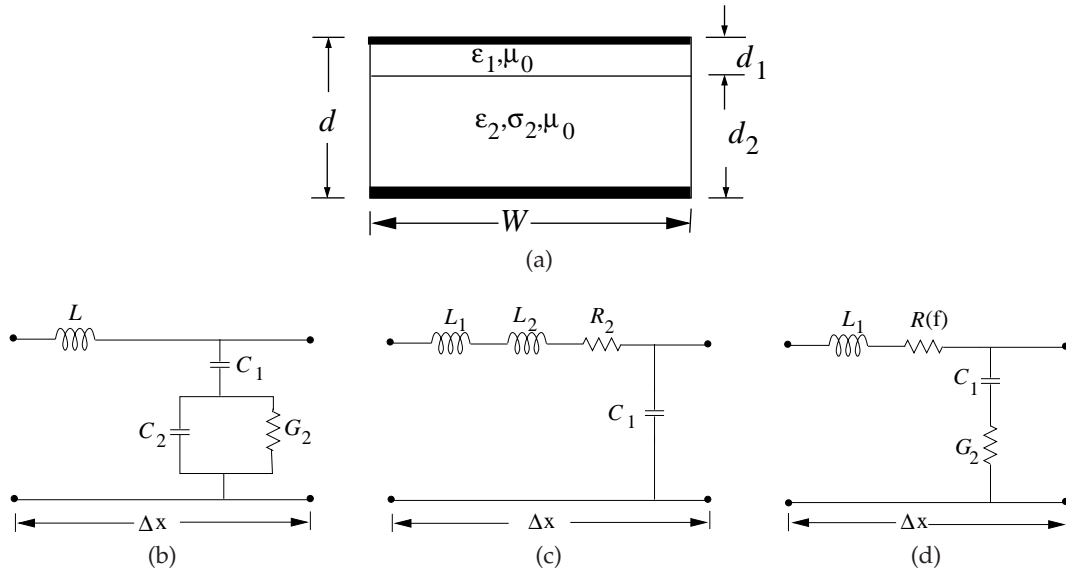
Figure 5-21 Modes on an MOS transmission line: (a) equivalent structure of the MOS line of Figure 5-18; (b) quasi-TEM mode; (c) skin effect mode; and (d) slow-wave mode.

$$C_2 = \varepsilon_2 \frac{W}{d_2} \tag{5.61}$$

$$G_2 = \sigma_2 \frac{W}{d_2} . \tag{5.62}$$

In SI units these have the units: H/m for L, F/m for C_1, F/m for C_2, and S/m for G_2.

Skin-Effect Mode

The second possible mode is the **skin-effect mode**, for which the sectional equivalent circuit model of Figure 5-21(c) is applicable. Here, $\sigma_2 \ll \omega \varepsilon_2$ is such that the skin depth $\delta = 1/\sqrt{\pi f \mu_0 \sigma_2}$ in the semiconductor is much smaller than d_2 and

$$\varepsilon'_e = \varepsilon'_{e,0} , \; \mu'_e = \frac{\mu_0}{d}\left(d_1 + \frac{\delta}{2}\right),$$

Thus

$$L_1 = \mu_0 \frac{d_1}{W} , \qquad C_1 = \varepsilon_1 \frac{W}{d_1} \tag{5.63}$$

$$L_2 = \mu_0 \frac{1}{W}\left(\frac{\delta}{2}\right) , \qquad R_2 = 2\pi f \mu_0 \frac{1}{W}\left(\frac{\delta}{2}\right) . \tag{5.64}$$

These have the SI units: H/m for L_1 and L_2, F/m for C_1, and Ω/m for R_2.

Slow-Wave Mode

The third possible mode of propagation is the slow-wave mode [76,77], for which the sectional equivalent circuit model of Figure 5-21(d) is applicable. This mode occurs when f is not so large and the resistivity is moderate so that the skin depth, δ, is larger than (or on the order of) d_2. Thus $\varepsilon'_e = \varepsilon'_{e,0}$, but $\mu'_e = \mu_0$. Therefore

$$v_p = \frac{1}{\sqrt{\varepsilon'_e \mu'_e}} = \frac{1}{\sqrt{\mu_0 \varepsilon_0}} \frac{1}{\sqrt{\varepsilon_1}} \sqrt{\frac{d_1}{d}} \qquad (5.65)$$

(with SI units of m/s) and $\lambda_g = \lambda_1 \sqrt{d_1/d}$, where λ_1 is the wavelength in the insulator.

5.9 Summary

In this chapter, extraordinary transmission line effects were considered. Multimoding can occur in many forms, but whenever crosssectional dimensions are large enough compared to a wavelength multimoding occurs. This sets the upper bound on the frequency of operation of most transmission line structures. There are also significant frequency-dependent effects with all transmission lines and arise from a number of sources. However, most of these effects in transmission lines of two or more conductors can be understood as resulting from skin effect and charge bunching. Different transmission line structures have varying degrees of frequency-dependent properties.

5.10 Exercises

1. A medium has a relative permittivity of 13 and supports a 5.6 GHz EM signal. By default, if not specified otherwise, a medium is loss-less and will have a relative permeability of 1. [Based on Appendix D.]

    (a) Calculate the characteristic impedance of an EM plane wave.

    (b) Calculate the propagation constant of the medium.

2. A medium has a dielectric constant of 20. What is the index of refraction of the medium? [Based on Appendix D.]

3. A plane wave in free space is normally incident on a lossless medium occupying a half space with a dielectric constant of 12. [Based on Appendix D. Parallels examples on Pages 866 and 866.]

    (a) Calculate the electric field reflection coefficient referred to the interface medium.

    (b) What is the magnetic field reflection coefficient?

4. Water, or more specifically tap water or sea water, has a complex dielectric constant resulting from two effects: conductivity resulting from dissolved ions in the water leading to charge carriers that can conduct current under the influence of an electric field, and dielectric loss resulting from rotation or bending of the water molecules themselves under the influence of an electric field. The rotation or bending of the water molecules results in motion of the water molecules and thus heat. The relative permittivity is

$$\varepsilon_{wr} = \varepsilon'_{wr} - \jmath \varepsilon''_{wr} \qquad (5.66)$$

and the real and imaginary components are

$$\varepsilon_r' = \varepsilon_{w00} + \frac{\varepsilon_{w0} - \varepsilon_{w00}}{1 + (2\pi f \tau_w)^2} \qquad (5.67)$$

$$\varepsilon_{wr}'' = \frac{2\pi f \tau_w (\varepsilon_{w0} - \varepsilon_{w00})}{1 + (2\pi f \tau_w)^2} + \frac{\sigma_i}{2\pi \varepsilon_0 f} (5.68)$$

where f is frequency, $\varepsilon_0 = 8.854 \times 10^{-12}$ F/m is the absolute permittivity, and $\varepsilon_{w00} = 4.9$ is the high-frequency limit of ε_r and is known to be independent of salinity and is also that of pure deionized water. The other quantities are τ_w, the relaxation time of water; ε_{w0}, the static relative dielectric constant of water (which for pure water at 0°C is equal to 87.13); and σ_i, the ionic conductivity of the water. At 5°C and for tap water with a salinity of 0.03 (in parts per thousand by weight) [78,79]

$$\varepsilon_{w0} = 85.9, \qquad \tau_w = 14.6 \text{ ps}, \qquad \sigma_i = 0.00548.$$

[Based on Appendix D.]

(a) Calculate and then plot the real and imaginary dielectric constant of water over the frequency range from DC to 40 GHz for tap water at 5°C. You should see two distinct regions. Identify the part of the imaginary dielectric constant graph that results from ionic conductivity and the part of the graph that relates to dielectric relaxation.

(b) Consider a plane wave propagating in water. Determine the attenuation constant and the propagation constant of the wave at 500 MHz.

(c) Consider a plane wave propagating in water. Determine the attenuation constant and the propagation constant of the wave at 2.45 GHz.

(d) Consider a plane wave propagating in water. Determine the attenuation constant and the propagation constant of the wave at 22 GHz.

(e) Again considering the plane wave in water, calculate the loss factor over a distance of 1 cm at 2.45 GHz. Express your answer in terms of decibels.

(f) Consider a plane wave making normal incidence from above into a bucket of tap water that is 1 m deep so that you do not need to consider the reflection at the bottom of the bucket. Calculate the reflection coefficient of the incident plane wave at 2.45 GHz referred to the air immediately above the surface of the water.

(g) Again considering the situation in (f), but this time with the plane wave incident on the air-water interface from the water side, calculate the reflection coefficient of the incident plane wave at 2.45 GHz referred to the water immediately below the surface of the water.

(h) If the bucket in (f) and (g) is 0.5 cm deep you will need to consider the reflection at the bottom of the bucket. Consider that the bottom of the bucket is a perfect conductor. Draw the bounce diagram for calculating the reflection of the plane wave in air. Determine the total reflection coefficient of the plane wave (referred to the air immediately above the surface of the water) by using the bounce diagram.

(i) Repeat (h) using the formula for multiple reflections.

5. The plane wave electric field of a transmitted radar pulse propagating in free space in the z direction is given by

$$\mathbf{E}(t) = \begin{cases} A \cos\left(\frac{\pi t}{2\tau}\right) \cos\left(\omega_0 t\right)\hat{x} & -\tau \le t \le \tau \\ \\ 0 & \text{otherwise,} \end{cases}$$
$$(5.69)$$

with $\omega_0 \gg \pi/(2\tau)$. [Based on Appendix D.]

(a) Sketch $\mathbf{E}(t)$ over the interval $-1.5\tau \le t \le 1.5\tau$.

(b) Derive the magnetic field expression, that is, $\mathbf{H}(t)$.

(c) Derive the instantaneous Poynting vector, $\mathcal{S}(t)$.

(d) Determine the total energy density in the pulse (in units of joules per square meter, J/m^2).

6. Consider a plane EM wave propagating a medium with a permittivity ε and permeability μ. The complex permittivity measured at

two frequencies is characterized by the relative permittivity ($\varepsilon/\varepsilon_0$) of the real $\Re\{\varepsilon/\varepsilon_0\}$ and imaginary parts $\Im\{\varepsilon/\varepsilon_0\}$ as follows:

Measured relative permittivity:

Frequency	Real part	Imaginary part
1 GHz	3.8	−0.05
10 GHz	4.0	−0.03

Measured relative permeability:

Frequency	Real Part	Imaginary Part
1 GHz	0.999	−0.001
10 GHz	0.998	−0.001

Since there is an imaginary part of the dielectric constant there could be either dielectric damping or material conductivity, or both. [Based on Appendix D. Parallels example on Page 862.]

(a) Determine the dielectric loss tangent at 10 GHz.

(b) Determine the relative dielectric damping factor at 10 GHz (the part of the permittivity due to dielectric damping).

(c) What is the conductivity of the dielectric at 10 GHz?

7. Consider a material with a relative permittivity of 72, a relative permeability of 1, and a static electric field (E) of 1 kV/m. How much energy is stored in the E field in a 10 cm^3 volume of the material? [Based on Appendix D. Parallels example on Page 853.]

8. A time-varying electric field in the x direction has a strength of 1 kV/m and a frequency of 1 GHz. The medium has a relative permittivity of 70. What is the polarization vector? Express this vector in the time domain? [Based on Appendix D. Parallels example on Page 853.]

9. A 4 GHz time-varying EM field is traveling in the $+z$ direction in Region 1 and is incident on another material in Region 2, as shown in the Figure D-8. The permittivity of Region 1 is $\varepsilon_1 = \varepsilon_0$ and that of Region 2 is $\varepsilon_2 = (4 - \jmath 0.04)\varepsilon_0$. For both regions $\mu_1 = \mu_2 = \mu_0$.

[Based on Appendix D. Parallels example on Page 867.]

(a) What is the characteristic impedance (or wave impedance) in Region 2?

(b) What is the propagation constant in Region 1?

10. A 4 GHz time-varying EM field is traveling in the $+z$ direction in Medium 1 and is normally incident on another material in Region 2, as shown in Figure **??**. The boundary between the two regions is in the $z = 0$ plane. The permittivity of Region 1, $\varepsilon_1 = \varepsilon_0$, and that of Region 2 is $\varepsilon_2 = 4\varepsilon_0$. For both regions, $\mu_1 = \mu_2 = \mu_0$. The phasor of the forward-traveling electric field (i.e., the incident field) is $\mathbf{E}^+ = 100\ \hat{y}$ and the phase is normalized with respect to $z = 0$. [Based on Appendix D. Parallels example on Page 867.]

(a) What is the wave impedance of Region 1?

(b) What is the wave impedance of Region 2?

(c) What is the electric field reflection coefficient at the boundary?

(d) What is the magnetic field reflection coefficient at the boundary?

(e) What is the electric field transmission coefficient at the boundary?

(f) What is the power transmitted into Region 2?

(g) What is the power reflected from the boundary back into Region 1?

11. What is the skin depth on a copper microstrip at 10 GHz? Assume that the conductivity of the deposited copper forming the strip is half that of bulk single crystal copper. Use the data in the table on Page 844.

12. What is the skin depth on a silver at 1 GHz? Assume that the conductivity of the fabricated silver conductor is 75% that of bulk single crystal silver. Use the data in the table on Page 844.

13. An inhomogeneous transmission line is fabricated using a medium with a relative permittivity of 10 and has an effective permittivity of 7. What is the fill factor q?

14. Using Table 5-9 on Page 257, determine the complex characteristic impedance and complex effective permittivity of a microstrip line at 12 GHz. The line is fabricated on alumina with $\varepsilon_r(DC) = 9.9$, $w = 70$ μm, $h = 500$ μm.

15. A magnetic wall and an electric wall are 2 cm apart and are separeted by a lossless material having an effective permittivity of 10 and an effective permeability of 23. What is the cut-off frequency of the lowest-order mode in this system.

16. The strip of a microstrip has a width of 600 μm and is fabricated on a lossless substrate that is 1 mm thick and has a relative permittivity of 10.

    (a) Draw the magnetic waveguide model of the microstrip line. Put dimensions on your drawing.

    (b) Sketch the electric field distribution of the first transverse resonance mode and calculate the frequency at which the transverse resonance mode occurs.

    (c) Sketch the electric field distribution of the first higher-order microstrip mode and calculate the frequency at which it occurs.

    (d) Sketch the electric field distribution of the slab mode and calculate the frequency at which it occurs.

17. A microstrip line has a width of 352 μm and is constructed on a substrate that is 500 μm thick with a relative dielectric constant of 5.6.

    (a) Determine the frequency at which transverse resonance would first occur.

    (b) When the dielectric is slightly less than one quarter wavelength in thickness the dielectric slab mode can be supported. Some of the fields will appear in the air region as well as in the dielectric, extending the effective thickness of the dielectric. Ignoring the fields in the air (so that we are considering a quarter-wavelength criterion), at what frequency will the dielectric slab mode first occur?

18. The strip of a microstrip has a width of 600 μm and is fabricated on a lossless substrate that is

635 μm thick and has a relative permittivity of 4.1.

    (a) At what frequency will the first transverse resonance occur?

    (b) At what frequency will the first higher-order microstrip mode occur?

    (c) At what frequency will the slab mode occur?

    (d) Identify the useful operating frequency range of the microstrip.

19. The electrical design space for transmission lines in microstrip designs is determined by the design project leader to be 20 to 100 Ω. This is based on experience with the types of loads and sources that need to be matched. The preferred substrate has a thickness of 500 μm and a relative permittivity of 8. What is the maximum operating frequency range of designs supported by this technology choice. Ignore frequency dispersion effects (i.e., frequency dependence of effective permittivity and losses). [Hint: First determine the dimensions of the strip that will provide the required impedance range and then determine the frequency at which lowest-order multimoding will occur.]

20. Consider the design of transmission lines in microstrip technology using a lossless substrate with relative permittivity of 10 and thickness of 400 μm. You will want to use the formulas in Section 4.10.2 on Page 223.

    (a) What is the maximum characteristic impedance that can be achieved for a transmission line fabricated in this technology?

    (b) Plot the characteristic impedance versus strip width.

    (c) From manufacturing tolerance considerations, the minimum strip width that can be manufactured is 20 μm. What is the maximum characteristic impedance that can be achieved in practice?

    (d) If the operating frequency range is 1 to 10 GHz, determine the maximum width of the strip from higher-order mode considerations. You must consider the trans-

verse resonance mode as well as higher-order microstrip modes.

(e) Identify the electrical design space (i.e., the achievable characteristic impedance range).

(f) Identify the physical design space (i.e., the range of acceptable strip widths).

(g) If the electrical design space requires that transmission line impedances be achieved within ± 2 Ω. What tolerance must be achieved in the manufacturing process if the substrate thickness can be achieved exactly? [Hint: First identify the critical physical process corner and thus the critical strip width that is most susceptible to width variations. Then determine the tolerance on the strip width to achieve the allowable characteristic impedance variation. That is, characteristic impedance is a function of strip width and height. If the substrate is perfect (no height variation), then how much can the strip width vary to keep the impedance within ± 2 Ω of the desired value? You can solve this graphically using a plot of Z_0 versus width or you can iteratively arrive at the answer by recalculating Z_0.]

(h) If the electrical design space requires that transmission line impedances be achieved within ± 2 Ω. What tolerance must be achieved in the manufacturing process if the substrate thickness tolerance is ± 2 μm? (Assume that the perfectly symmetrical stripline property will be achieved accurately.) If the substrate is not perfect (the height variation is ± 2 μm), then how much can the strip width vary to keep the impedance within ± 2 Ω of the desired value? This problem is directly applicable to real-world process/design trade-offs.

21. The strip of a microstrip has a width of 500 μm and is fabricated on a lossless substrate that is 635 μm thick and has a relative permittivity of 12. [Parallels Examples 5.1, 5.2, and 5.3 on Pages 270, 271 and 273.]

(a) At what frequency does the transverse resonance first occur?

(b) At what frequency does the first higher-order microstrip mode first propagate?

(c) At what frequency does the substrate (or slab) mode first occur?

22. The strip of a microstrip has a width of 250 μm and is fabricated on a lossless substrate that is 300 μm thick and has a relative permittivity of 15.

(a) At what frequency does the transverse resonance first occur?

(b) At what frequency does the first higher-order microstrip mode first propagate?

(c) At what frequency does the substrate (or slab) mode first occur?

23. A microstrip line has a strip width of 100 μm is fabricated on a lossless substrate that is 150 μm thick and has a relative permittivity of 9.

(a) Draw the microstrip waveguide model and indicate and calculate the dimensions of the model.

(b) Based only on the microstrip waveguide model, determine the frequency at which the first transverse resonance occurs?

(c) Based only on the microstrip waveguide model, determine the frequency at which the first higher-order microstrip mode occurs?

(d) At what frequency will the slab mode occur? For this you cannot use the microstrip waveguide model.

24. A microstrip line has a strip width of 100 μm and is fabricated on a lossless substrate that is 150 μm thick and has a relative permittivity of 9.

(a) Define the properties of a magnetic wall.

(b) Identify two situations where a magnetic wall can be used in the analysis of a microstrip line; that is, give two locations where a magnetic wall approximation can be used.

(c) Draw the microstrip waveguide model and indicate and calculate the dimensions of the model.

25. Consider the design of transmission lines in symmetrical stripline technology using a lossless substrate with relative permittivity of 20 and thickness of 200 μm.

    (a) Using qualitative arguments, show that the maximum characteristic impedance that can be achieved for a transmission line fabricated in this technology is 116 Ω. The maximum is not actually 116 Ω, but a simple argument will bring you to this conclusion. Hint: This will occur at the minimum width possible.

    (b) Plot the characteristic impedance versus substrate width.

    (c) From manufacturing tolerance considerations, the minimum strip width that can be manufactured is 20 μm. What is the maximum characteristic impedance that can be achieved in practice?

    (d) If the operating frequency range is 2 to 18 GHz, determine the maximum width of the strip from higher-order mode considerations. You must consider the transverse resonance mode, the higher-order stripline mode, and the slab mode.

    (e) Identify the electrical design space (i.e., the achievable characteristic impedance range).

    (f) Identify the physical design space (i.e., the range of acceptable strip widths).

    (g) If the electrical design space requires that transmission line impedances be achieved within ± 2 Ω. What tolerance must be achieved in the manufacturing process if the substrate thickness can be achieved exactly? [Hint: First identify the critical physical process corner and thus the critical strip width that is most susceptible to width variations. Then determine the tolerance on the strip width to achieve the allowable characteristic impedance variation.]

    (h) If the electrical design space requires that transmission line impedances be achieved within ± 2 Ω. What tolerance must be achieved in the manufacturing process if the substrate thickness tolerance is ± 2 μm. (Assume that the perfectly symmetrical stripline property will be achieved accurately.)

26. Consider the structure in Figure 5-19. Determine the guide wavelength, λ_g, and the wavelength in the insulator, λ_1, at a frequency of 20 GHz. SiO$_2$ is the dielectric, with permittivity $\varepsilon_1 = 3.9\varepsilon_0$. The depths d_2 and d_1 of the two dielectrics are $d_2 = 100$ μm and $d_1 = 1.0$ μm. [Parallels Examples 5.4, on Page 278.]

27. Consider a metal-oxide-semiconductor transmission medium as examined in Section 5.8. The structure in the form of the Maxwell-Wagner capacitor is shown in Figure 5-20(a) with $d_1 = 100$ μm, $d_1 = 500$ μm, and relative permittivities $\varepsilon_1 = 3.9$ and $\varepsilon_1 = 13$. Ignore the finite conductivity. What is the capacitance model of this structure (see Figure 5-20(b)) and what are the values of the capacitance?

Microwave Network Analysis

6.1 Introduction

Analog circuits at frequencies up to a few tens of megahertz are characterized by admittances, impedances, voltages, and currents. Above these frequencies it is not possible to measure voltage, current, or impedance directly. One reason for this is that measurement equipment is separated from the device by lengths of transmission line that are electrically long (i.e., at least an appreciable fraction of a wavelength long). It is better to use quantities such as voltage reflection and transmission coefficients that can be quite readily measured and are related to power flow. As well, in RF and microwave circuit design the power of signals and of noise is always of interest. Thus there is a predisposition to focus on measurement parameters that are related to the reflection and transmission of power.

   Scattering parameters, S parameters, embody the effects of reflection
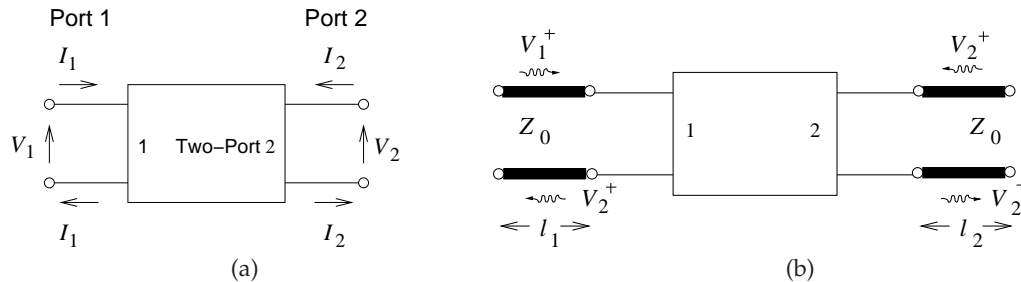
Figure 6-1   A two-port network: (a) port voltages; and (b) with transmission lines at the ports.

and transmission. As will be seen, it is easy to convert these to more familiar network parameters such as admittance and impedance parameters. In this chapter S parameters will be defined and related to impedance and admittance parameters, then it will be demonstrated that the use of S parameters helps in the design and interpretation of RF circuits. S parameters have become the most important parameters for RF and microwave engineers and many design methodologies have been developed around them.

A graphical technique called signal flow graph analysis is introduced for manipulating S parameter-based models of elements. Finally, polar plots of S parameters are introduced as well as an annotated form called a Smith chart. Both are used for displaying and interpreting power flow in circuits.

6.2   Two-Port Networks

Many of the techniques employed in analyzing circuits require that the voltage at each terminal of a circuit be referenced to a common point such as ground. In microwave circuits it is generally difficult to do this. Recall that with transmission lines it is not possible to establish a common ground point. However, with transmission lines (and circuit elements that utilize distributed effects) it was seen that for each signal current there is a signal return current. Thus at radio frequencies, and for circuits that are distributed, ports are used, as shown in Figure 6-1, which define the voltages and currents for what is known as a two-port network, or just two-port.[1] The network in Figure 6-1(a) has four terminals and two ports. A port voltage is defined as the voltage difference between a pair of terminals with one of the terminals in the pair becoming the reference terminal. Port 1 is on the left of the diagram, where port voltage V_1 is defined. The current entering the network at the top terminal of Port 1 is I_1 and there is an equal current leaving the reference terminal. This arrangement clearly makes sense when

---

[1]   Even when the term "two-port" is used on its own the hyphen is used, as it is referring to a two-port network.

transmission lines are attached to Ports 1 and 2, as in Figure 6-1(b). With transmission lines at Ports 1 and 2 there will be traveling-wave voltages, and at the ports the traveling-wave components add to give the total port voltage. In dealing with nondistributed circuits it is preferable to use the total port voltages and currents—V_1, I_1, V_2, and I_2, shown in Figure 6-1(a). However, with distributed elements it is preferable to deal with traveling voltages and currents—V_1^+, V_1^-, V_2^+, and V_2^-, shown in Figure 6-15(b). RF and microwave design necessarily requires switching between the two forms.

6.2.1   Reciprocity, Symmetry, Passivity, and Linearity

Reciprocity, symmetry, passivity, and linearity are fundamental properties of networks. A network is linear if the response, voltages, and currents are linearly dependent on the drive level, and superposition also applies. So if the two-port shown in Figure 6-1 is linear, the currents I_1 and I_2 are linear functions of V_1 and V_2. An example of a linear network would be one with resistors and capacitors. A network with a diode would be an example of a nonlinear network.

A passive network has no internal sources of power and so a network with an embedded battery is not a passive network. A more complicated situation to consider is a class A transistor amplifier where the transistor is biased so that the response to small input signals is linear. In network analysis, the network parameters at the bias point are used. These network parameters are linear, but they are not passive, as the transistor amplifier is adding power to the small signals. The amplifier converts DC power to small signal power. If the whole amplifier is considered and the network parameters include multiple ports (a port is required for the applied DC bias) then the network parameters would be passive but they would still be nonlinear. A symmetrical two-port has the same characteristics at each of the ports. An example of a symmetrical network is a transmission line.

Finally, a reciprocal two-port has a response at Port 2 from an excitation at Port 1 that is the same as the response at Port 1 to the same excitation at Port 2. As an example, consider the two-port in Figure 6-1 with $V_2 = 0$. If the network is reciprocal, the ratio I_2/V_1 will be the same as the ratio I_1/V_2 with $V_1 = 0$. Most networks with resistors, capacitors, and transmission lines, for example, are reciprocal. A transistor amplifier is not reciprocal, as gain, analogous to the ratio I_2/V_1, is just in one direction (or unidirectional).

6.2.2   Parameters Based on Total Voltage and Current

First, port-based impedance parameters will be considered based on total port voltages and currents as defined for the two-port in Figure 6-1. These parameters are also referred to as port impedance parameters or just impedance parameters when the context is understood to be ports.

(a)   (b)
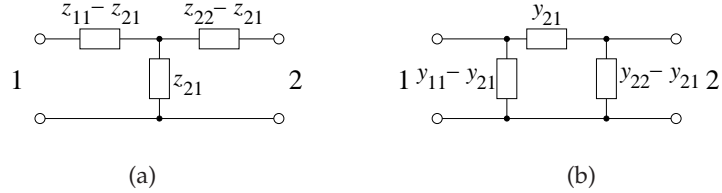
Figure 6-2 Circuit equivalence of the z and y parameters for a reciprocal network: (a) z parameters; and (b) y parameters

Impedance parameters, or z parameters, are defined as

$$V_1 = z_{11}I_1 + z_{12}I_2 \tag{6.1}$$
$$V_2 = z_{21}I_1 + z_{22}I_2, \tag{6.2}$$

or in matrix form as

$$\mathbf{V} = \mathbf{ZI}. \tag{6.3}$$

The double subscript on a parameter is ordered so that the first refers to the output and the second refers to the input, so z_{ij} relates the voltage output at Port i to the current input at Port j. If the network is reciprocal then $z_{12} = z_{21}$, but this simple type of relationship does not apply to all network parameters. The reciprocal circuit equivalence of the z parameters is shown in Figure 6-2(a). It will be seen that the z parameters are convenient parameters to use when an element is in series with one of the ports, as then the operation required in developing the z parameters of the larger network is just addition.

Figure 6-3(a) shows the series connection of a two-terminal element with a two-port designated as network A. The z parameters of network A are

$$\mathbf{Z}_A = \begin{bmatrix} z_{11}^{(A)} & z_{12}^{(A)} \\ z_{21}^{(A)} & z_{22}^{(A)} \end{bmatrix}, \tag{6.4}$$

so that

$$V_1^{(A)} = z_{11}^{(A)}I_1 + z_{12}^{(A)}I_2 \tag{6.5}$$
$$V_2 = z_{21}^{(A)}I_1 + z_{22}^{(A)}I_2. \tag{6.6}$$

Now

$$V_1 = zI_1 + V_1^{(A)} \tag{6.7}$$
$$= zI_1 + z_{11}^{(A)}I_1 + z_{12}^{(A)}I_2, \tag{6.8}$$

so the z parameters of the whole network can be written as

$$\mathbf{Z} = \begin{bmatrix} z & 0 \\ 0 & 0 \end{bmatrix} + \mathbf{Z}_A. \tag{6.9}$$

(a)

(b)
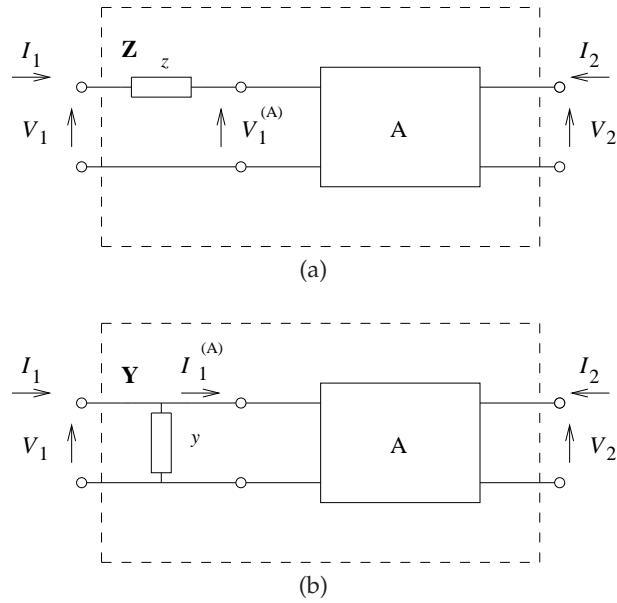
Figure 6-3 Connection of a two-terminal to a two-port: (a) series connection; and (b) parallel or shunt connection.

When an element is in shunt with a two-port, admittance parameters, or y parameters, are the most convenient to use. These are defined as

$$I_1 = y_{11}V_1 + y_{12}V_2 \tag{6.10}$$
$$I_2 = y_{21}V_1 + y_{22}V_2 \,, \tag{6.11}$$

or in matrix form as

$$\mathbf{I} = \mathbf{YV} \,. \tag{6.12}$$

Now, for reciprocity, $y_{12} = y_{21}$ and the circuit equivalence of the y parameters is shown in Figure 6-2(b). Consider the shunt connection of an element shown in Figure 6-3(b) where the y parameters of network A are

$$\mathbf{Y}_A = \begin{bmatrix} y_{11}^{(A)} & y_{12}^{(A)} \\ y_{21}^{(A)} & y_{22}^{(A)} \end{bmatrix} = \mathbf{Z}_A^{-1} \,. \tag{6.13}$$

The port voltages and currents are related as follows

$$I_1^{(A)} = y_{11}^{(A)}V_1 + y_{12}^{(A)}V_2 \tag{6.14}$$

$$I_2 = y_{21}^{(A)}V_1 + y_{22}^{(A)}V_2 \tag{6.15}$$

$$I_1 = yV_1 + I_1^{(A)} \tag{6.16}$$

$$= (y + y_{11}^{(A)})V_1 + y_{12}^{(A)}V_2. \tag{6.17}$$
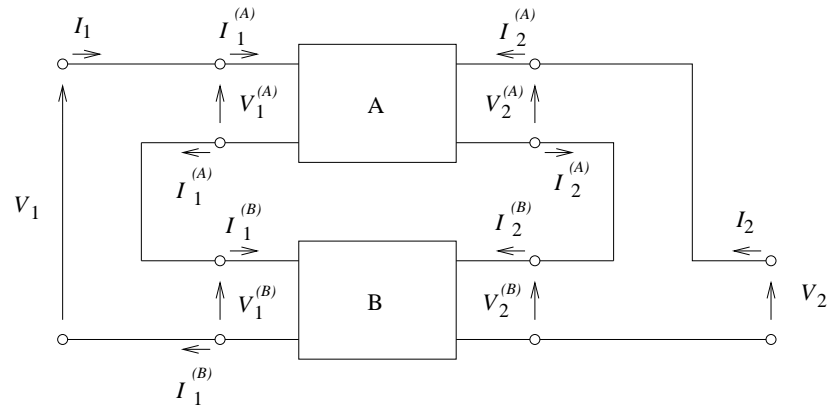
Figure 6-4   Series connection of 2 two-port networks.

Thus the y parameters of the whole network are

$$\mathbf{Y} \;=\; \left[\begin{array}{cc} y & 0 \\ 0 & 0 \end{array}\right] + \mathbf{Y}_A. \tag{6.18}$$

Sometimes it is more convenient to use hybrid parameters, or h parameters defined by

$$V_1 \;=\; h_{11}I_1 + h_{12}V_2 \tag{6.19}$$

$$I_2 \;=\; h_{21}I_1 + h_{22}V_2, \tag{6.20}$$

or in matrix form as

$$\left[\begin{array}{c} V_1 \\ I_2 \end{array}\right] = \mathbf{H} \left[\begin{array}{c} I_1 \\ V_2 \end{array}\right]. \tag{6.21}$$

These parameters are convenient to use with transistor circuits as they conveniently describe a voltage-controlled current source which is a simple model of a transistor.

The choice of which network parameters to use depends on convenience, but as will be seen throughout this text, some parameters naturally describe a particular characteristic desired in a design. For example, with Port 1 being the input port of an amplifier and Port 2 being the output port, h_{21} describes the current gain of the amplifier.

6.2.3   Series Connection of Two-Port Networks

A series connection of 2 two-ports is shown in Figure 6-4. An example of when the series connection occurs is shown in Figure 6-5, which is the schematic of a transistor amplifier configuration with an inductor in the source leg. The transistor and the inductor can each be represented as two-ports so that the circuit of Figure 6-5(a) can be represented as the series
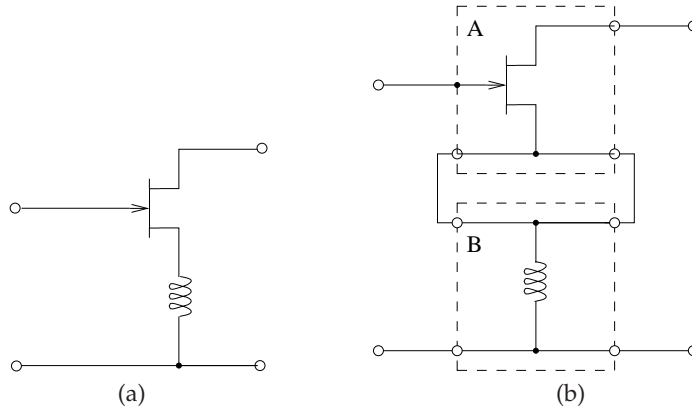
Figure 6-5 Example of a series connection of two-port networks: (a) an amplifier; and (b) its components represented as two-port networks.

connection of 2 two-ports, as shown in Figure 6-5(b). In the following, two-port parameters of the complete circuit are developed using the two-port parameter descriptions of the component two-ports. The procedure with any interconnection is to write down the relationships between the voltages and currents of the constituent networks. Which network parameters to use requires identification of the arithmetic path requiring the fewest operations.

Simple algebra relates the various total voltage and current parameters. From Figure 6-4,

$$I_1^{(A)} = I_1^{(B)} = I_1 \tag{6.22}$$

$$I_2^{(A)} = I_2^{(B)} = I_2 \tag{6.23}$$

$$V_1 = V_1^{(A)} + V_1^{(B)} \tag{6.24}$$

$$V_2 = V_2^{(A)} + V_2^{(B)}, \tag{6.25}$$

and in matrix form this becomes

$$\begin{bmatrix} V_1^{(A)} \\ V_2^{(A)} \end{bmatrix} = \begin{bmatrix} z_{11}^{(A)} & z_{12}^{(A)} \\ z_{21}^{(A)} & z_{22}^{(A)} \end{bmatrix} \begin{bmatrix} I_1^{(A)} \\ I_2^{(A)} \end{bmatrix}; \begin{bmatrix} V_1^{(B)} \\ V_2^{(B)} \end{bmatrix} = \begin{bmatrix} z_{11}^{(B)} & z_{12}^{(B)} \\ z_{21}^{(B)} & z_{22}^{(B)} \end{bmatrix} \begin{bmatrix} I_1^{(B)} \\ I_2^{(B)} \end{bmatrix} \tag{6.26}$$

or

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = (\mathbf{Z}^A + \mathbf{Z}^B) \begin{bmatrix} I_1 \\ I_2 \end{bmatrix}. \tag{6.27}$$

Thus the impedance matrix of the series connection of two-ports is

$$\mathbf{Z} = \mathbf{Z_A} + \mathbf{Z_B}. \tag{6.28}$$
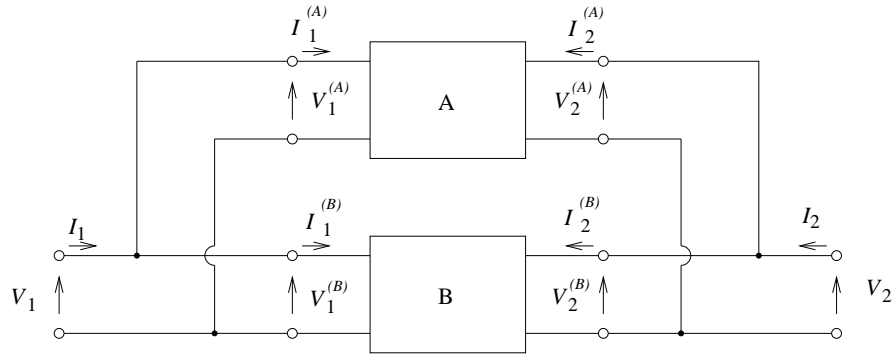
Figure 6-6 Parallel connection of two-port networks.

6.2.4 Parallel Connection of Two-Port Networks

Admittance parameters are most conveniently combined to obtain the overall parameters of the two-port parallel connection of Figure 6-6. The total voltage and current are related by

$$
\begin{align}
V_1^{(A)} &= V_1^{(B)} = V_1 \tag{6.29}\\
V_2^{(A)} &= V_2^{(B)} = V_2 \tag{6.30}\\
I_1 &= I_1^{(A)} + I_1^{(B)} \tag{6.31}\\
I_2 &= I_2^{(A)} + I_2^{(B)} \tag{6.32}
\end{align}
$$

and

$$
\left[\begin{array}{c} I_1^{(A)} \\ I_2^{(A)} \end{array}\right] = \left[\begin{array}{cc} y_{11}^{(A)} & y_{12}^{(A)} \\ y_{21}^{(A)} & y_{22}^{(A)} \end{array}\right] \left[\begin{array}{c} V_1^{(A)} \\ V_2^{(A)} \end{array}\right] \tag{6.33}
$$

$$
\left[\begin{array}{c} I_1^{(B)} \\ I_2^{(B)} \end{array}\right] = \left[\begin{array}{cc} y_{11}^{(B)} & y_{12}^{(B)} \\ y_{21}^{(B)} & y_{22}^{(B)} \end{array}\right] \left[\begin{array}{c} V_1^{(B)} \\ V_2^{(B)} \end{array}\right]. \tag{6.34}
$$

So the overall y parameter relation is

$$
\left[\begin{array}{c} I_1 \\ I_2 \end{array}\right] = (\mathbf{Y}_A + \mathbf{Y}_B) \left[\begin{array}{c} V_1 \\ V_2 \end{array}\right]. \tag{6.35}
$$

Figure 6-7 is an example of subcircuits that can be represented as two-ports that are then connected in parallel.

6.2.5 Series-Parallel Connection of Two-Port Networks

A similar approach to that in the preceding subsection is followed in developing the overall network parameters of the series-parallel connection of two-ports shown in Figure 6-8. Now,
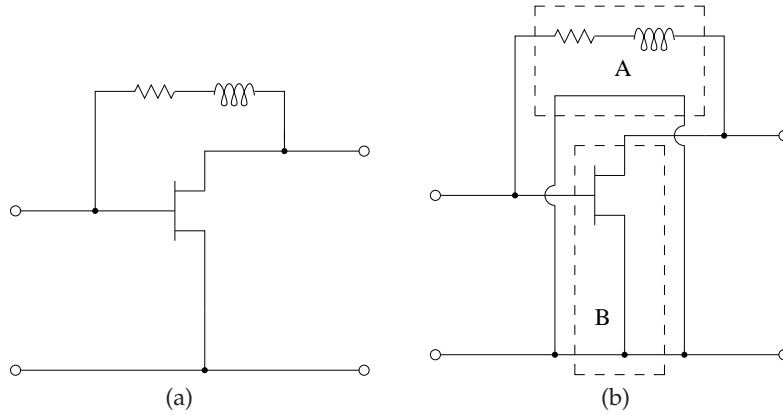
Figure 6-7 Example of the parallel connection of 2 two-ports .
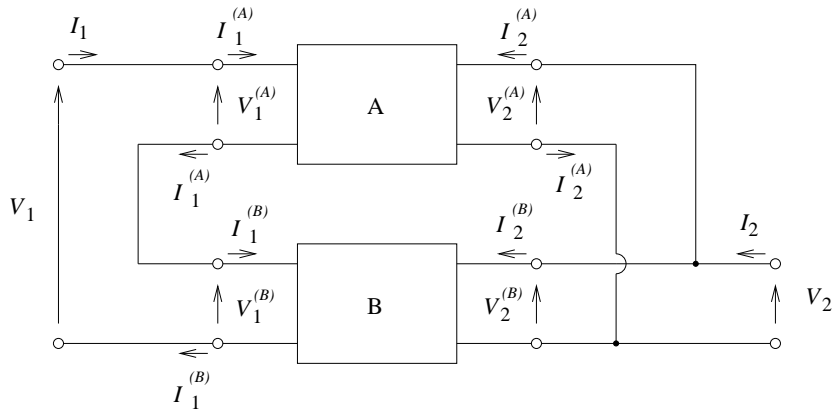
Figure 6-8 Series-parallel connection of two-ports.

$$I_1^{(A)} = I_1^{(B)} = I_1 \qquad (6.36)$$

$$V_2^{(A)} = V_2^{(B)} = V_2 \qquad (6.37)$$

$$V_1 = V_1^{(A)} + V_1^{(B)} \qquad (6.38)$$

$$I_2 = I_2^{(A)} + I_2^{(B)}, \qquad (6.39)$$

and so, using hybrid parameters,

$$
\begin{bmatrix} V_1^{(A)} \\ I_2^{(A)} \end{bmatrix} = \begin{bmatrix} h_{11}^{(A)} & h_{12}^{(A)} \\ h_{21}^{(A)} & h_{22}^{(A)} \end{bmatrix} \begin{bmatrix} I_1^{(A)} \\ V_2^{(A)} \end{bmatrix} \qquad (6.40)
$$

$$
\begin{bmatrix} V_1^{(B)} \\ I_2^{(B)} \end{bmatrix} = \begin{bmatrix} h_{11}^{(B)} & h_{12}^{(B)} \\ h_{21}^{(B)} & h_{22}^{(B)} \end{bmatrix} \begin{bmatrix} I_1^{(B)} \\ V_2^{(B)} \end{bmatrix}. \qquad (6.41)
$$
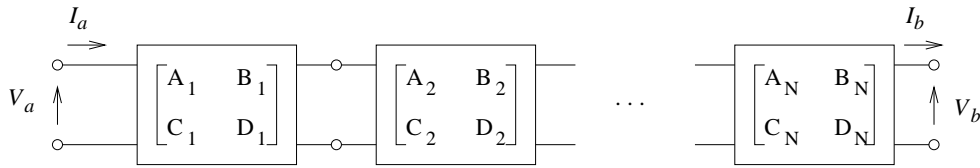
Figure 6-9   Cascade of two-port networks.
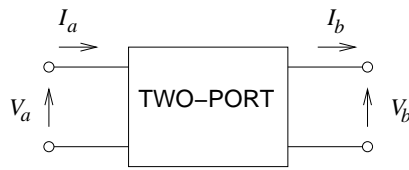
Figure 6-10   Two-port network with cascadable voltage and current definitions.

Putting this in compact form:

$$\left[\begin{array}{c} V_1 \\ I_2 \end{array} \right] = (\mathbf{H}_A + \mathbf{H}_B) \left[\begin{array}{c} I_1 \\ V_2 \end{array} \right]. \tag{6.42}$$

6.2.6 $ABCD$ Matrix Characterization of Two-Port Networks

$ABCD$ parameters are the best parameters to use when cascading two-ports, as in Figure 6-9, and when total voltage and current relationships are required. First, consider Figure 6-10, which puts the voltages and currents in cascadable form with the input variables in terms of the output variables:

$$\left[\begin{array}{c} V_a \\ I_a \end{array} \right] = \left[\begin{array}{cc} A & B \\ C & D \end{array} \right] \left[\begin{array}{c} V_b \\ I_b \end{array} \right]. \tag{6.43}$$

The reciprocity relationship of the $ABCD$ parameters is not as simple as for the z and y parameters. To see this, first express $ABCD$ parameters in terms of z parameters. Note that the current at Port 2 is in the opposite direction to the usual definition of the two-port current shown in Figure 6-1. So

$$
\begin{align}
V_a &= V_1 \tag{6.44} \\
I_a &= I_1 \tag{6.45} \\
V_b &= V_2 \tag{6.46} \\
I_b &= -I_2 \tag{6.47} \\
V_a &= z_{11}I_a - z_{12}I_b \tag{6.48} \\
V_b &= z_{21}I_a - z_{22}I_b. \tag{6.49}
\end{align}
$$

From Equation 6.49,

$$I_a = \frac{z_{22}}{z_{21}}I_b + \frac{1}{z_{21}}V_b, \tag{6.50}$$

and substituting this into Equation (6.48) yields

$$V_a = \frac{z_{11}}{z_{21}}V_b + \left(\frac{z_{11}z_{22}}{z_{21}} - z_{12}\right)I_b . \qquad (6.51)$$

Comparing Equations (6.50) and (6.51) to Equation (6.43) leads to

$$
\begin{aligned}
A &= z_{11}/z_{21} & (6.52)\\
B &= \Delta_z/z_{21} & (6.53)\\
C &= 1/z_{21} & (6.54)\\
D &= z_{22}/z_{21}, & (6.55)
\end{aligned}
$$

where

$$\Delta_z = z_{11}z_{22} - z_{12}z_{21}. \qquad (6.56)$$

Rearranging,

$$
\begin{aligned}
z_{11} &= A/C \\
z_{12} &= (AD - BC)/C \\
z_{21} &= 1/C \\
z_{22} &= D/C. & (6.57)
\end{aligned}
$$

Now the reciprocity condition in terms of $ABCD$ parameters can be determined. For z parameters, $z_{12} = z_{21}$ for reciprocity; that is, from Equation (6.57), for a reciprocal network,

$$
\begin{aligned}
\frac{AD - BC}{C} &= \frac{1}{C} & (6.58)\\
AD - BC &= 1. & (6.59)
\end{aligned}
$$
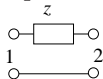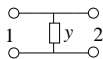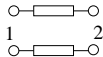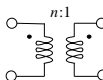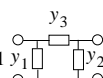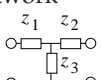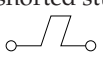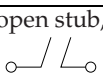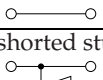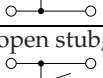
That is, for a two-port network to be reciprocal, $AD - BC = 1$. The utility of these parameters is that the $ABCD$ matrix of the cascade connection of N two-ports in Figure 6-9 is equal to the product of the $ABCD$ matrices of the individual two-ports:

$$
\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix}\begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix}\cdots\begin{bmatrix} A_N & B_N \\ C_N & D_N \end{bmatrix}. \qquad (6.60)
$$

The $ABCD$ parameters of several two-port networks are given in Table 6-1.

Table 6-1 $ABCD$ parameters of several two-ports. The transmission line is lossless with a propagation constant of β.

Series impedance, z	$A = 1 \quad B = z$ $C = 0 \quad D = 1$
Shunt admittance, y	$A = 1 \quad B = 0$ $C = y \quad D = 1$
Transmission line, length ℓ, $Z_0 = 1/Y_0$	$A = \cos(\beta\ell) \qquad B = \jmath Z_o \sin(\beta\ell)$ $C = \jmath Y_o \sin(\beta\ell) \quad D = \cos(\beta\ell)$
Transformer, ratio n:1	$A = n \quad B = 0$ $C = 0 \quad D = 1/n$
Pi network	$A = 1 + y_2/y_3 \qquad\qquad B = 1/y_3$ $C = y_1 + y_2 + y_1 y_2/y_3 \quad D = 1 + y_1/y_3$
Tee network	$A = 1 + z_1/z_3 \quad B = z_1 + z_2 + z_1 z_2/z_3$ $C = 1/z_3 \qquad\qquad D = 1 + z_2/z_3$
Series shorted stub, electrical length θ	$A = 1 \quad B = \jmath Z_0 \tan\theta$ $C = 0 \quad D = 1$
Series open stub, electrical length θ	$A = 1 \quad B = -\jmath Z_0/(\tan\theta)$ $C = 0 \quad D = 1$
Shunt shorted stub, electrical length θ	$A = 1 \qquad\qquad B = 0$ $C = -\jmath/(Z_0 \tan\theta) \quad D = 1$
Shunt open stub, electrical length θ	$A = 1 \qquad\qquad B = 0$ $C = \jmath \tan\theta/Z_0 \quad D = 1$

EXAMPLE 6. 1   Lumped Element Model of a Quarter-Wavelength Long Line

Develop the lumped-element model of a quarter-wavelength long transmission line with impedance Z_0.

Solution:

The model is developed by equating $ABCD$ parameters. From Table 6-1, the $ABCD$ parameters of a quarter-wavelength long line (with $\beta\ell = \pi/2$) are

$$A = \cos(\beta\ell) = 0 \tag{6.61}$$
$$B = \jmath Z_0 \sin(\beta\ell) = \jmath Z_0 \tag{6.62}$$
$$C = \jmath Y_0 \sin(\beta\ell) = \jmath/Z_0 \tag{6.63}$$
$$D = \cos(\beta\ell) = 0. \tag{6.64}$$

The $ABCD$ parameters of a Pi network are

$$A = 1 + y_2/y_3 \tag{6.65}$$
$$B = 1/y_3 \tag{6.66}$$
$$C = y_1 + y_2 + y_1 y_2/y_3 \tag{6.67}$$
$$D = 1 + y_1/y_3. \tag{6.68}$$

Equating Equations (6.62) and (6.66),

$$y_3 = 1/(\jmath Z_0), \tag{6.69}$$

and with Equations (6.61), (6.64), (6.65), and (6.68),

$$y_1 = y_2 = -y_3 = -1/(\jmath Z_0). \tag{6.70}$$

The lumped equivalent circuit of the quarter-wave transformer is shown in Figure 6-11(a), and for any impedance inverter is shown in Figure 6-11(b). An alternative lumped-element model of the inverter is shown in Figure 6-11(c). The lumped-element model of a 50 Ω impedance inverter at 400 MHz is shown in Figure 6-11(b) with $L = 19.89$ nH and $C = 7.968$ pF.

6.3   Scattering Parameters

Direct measurement of the z, y, h, and $ABCD$ parameters requires that the ports be terminated in either short or open circuits. For active circuits, such terminations could result in undesired behavior, including oscillation or destruction. Also, at RF it is difficult to realize a good open or short. Since RF circuits are designed with close attention to maximum power transfer conditions, resistive terminations are preferred, as these are closer to the actual operating conditions, and so the effect of measurement errors will have less impact than when parameter extraction relies on imperfect opens
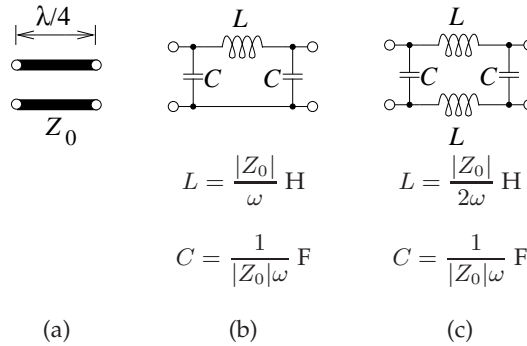
Figure 6-11 Lumped equivalent circuit of a quarter-wavelength long line of characteristic impedance Z_0 which is the same as that of an impedance inverter (of Z_0 Ω): (a) quarter-wavelength long line segment; (b) lumped element equivalent circuit; and (c) alternative lumped element model.

and shorts. The essence of scattering parameters (or S parameters[2]) is that they relate forward- and backward-traveling waves on a transmission line, thus S parameters are related to power flow. All these are good reasons for why RF and microwave engineers prefer the use of S parameters. The discussion of S parameters begins by considering the reflection coefficient, which is the S parameter of a one-port network.

6.3.1 Reflection Coefficient

The reflection coefficient, Γ, of a load, such as that shown in Figure 6-12, can be determined by separately measuring the forward- and backward-traveling voltages on the transmission line:

$$\Gamma(x) \quad = \quad \frac{V^-(x)}{V^+(x)}. \tag{6.71}$$

Γ at the load is related to the impedance Z_L by

$$\Gamma(0) \quad = \quad \frac{Z_L - Z_0}{Z_L + Z_0}, \tag{6.72}$$

where Z_0 is the characteristic impedance of the connecting transmission line. This can also be written as

$$\Gamma(0) = \frac{Y_0 - Y_L}{Y_0 + Y_L}, \tag{6.73}$$

where $Y_0 = 1/Z_0$ and $Y_L = 1/Z_L$. More completely, Γ as defined above is called the voltage reflection coefficient.

---

[2] For historical reasons a capital "S" is used when referring to S parameters. For most other network parameters, lowercase is used (e.g., z parameters for impedance parameters).
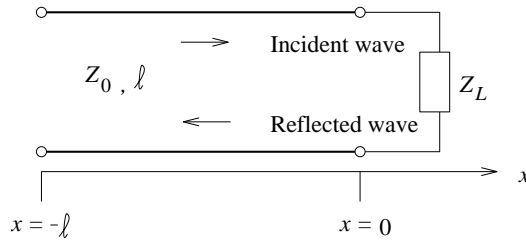
Figure 6-12 Transmission line of characteristic impedance Z_0 and length ℓ terminated in a load of impedance Z_L.
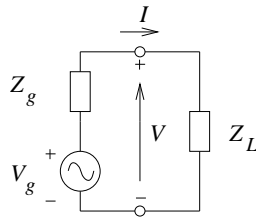
Figure 6-13 A Thevenin equivalent source with generator V_g and source impedance Z_g terminated in a load Z_L.

Consider Figure 6-13, which does not have an explicit transmission line to separate the forward- and backward-traveling waves. Now the total voltage across the load is

$$V = V_g \frac{Z_L}{Z_g + Z_L} \tag{6.74}$$

and the total current is

$$I = \frac{V_g}{Z_g + Z_L}. \tag{6.75}$$

To develop the reflection coefficient, first define equivalent forward- and backward-traveling waves. This can be done by imagining that between the generator and the load there is a transmission line of characteristic impedance Z_g and having infinitesimal length. The incident voltage and current waves (V^+, I^+) are the voltage and current obtained when the generator is conjugately matched to the load (i.e., $Z_L = Z_g^*$). So the equations of forward-traveling voltage and current become

$$V^+ = V_g \frac{Z_g^*}{Z_g + Z_g^*} = V_g \frac{Z_g^*}{2\mathcal{R}\{Z_g\}} \tag{6.76}$$

$$I^+ = V_g \frac{1}{2\mathcal{R}\{Z_g\}}. \tag{6.77}$$

Now return to considering the actual load, Z_L. The reflected voltage and current (V^-, I^-) are obtained by calculating the actual voltage and current at the source using the relationships

$$V = V^+ + V^- \tag{6.78}$$

and

$$I = I^+ + I^- \tag{6.79}$$

to determine the backward-traveling components. In Equations (6.78) and (6.79) the forward-traveling components are those in Equations (6.76) and (6.77). From Equations (6.76) and (6.74),

$$V_{\mathrm{g}} = V^+ \frac{Z_{\mathrm{g}} + Z_{\mathrm{g}}^*}{Z_{\mathrm{g}}^*}, \tag{6.80}$$

and from Equations (6.78), (6.74), and (6.80),

$$V^- = V - V^+ = V_g \frac{Z_L}{Z_g + Z_L} - V^+ \tag{6.81}$$

$$= \left[\frac{Z_g + Z_g^*}{Z_g^*} \frac{Z_L}{Z_g + Z_L} - 1 \right] V^+ \tag{6.82}$$

$$= \left[\frac{Z_g Z_L + Z_g^* Z_L - Z_g^* Z_g - Z_g^* Z_L}{Z_g^* (Z_g + Z_L)} \right] V^+ \tag{6.83}$$

$$= \underbrace{\left(\frac{Z_L - Z_g^*}{Z_L + Z_g} \right) \frac{Z_g}{Z_g^*}}_{\Gamma^V} V^+ = \Gamma^V V^+. \tag{6.84}$$

Similarly

$$I^- = - \underbrace{\left(\frac{Z_L - Z_g^*}{Z_L + Z_g} \right) \frac{Z_g}{Z_g^*}}_{\Gamma^I} I^+ = \Gamma^I I^+. \tag{6.85}$$

Γ^V is called the voltage reflection coefficient, which is usually denoted as Γ, while Γ^I is the current reflection coefficient. It is clear that $\Gamma^V = 0 = \Gamma^I$ when $Z_L = Z_g^*$ and $\Gamma^V = -\Gamma^I = (Z_L - R_g)/(Z_L + R_g)$ when Z_g is purely resistive (i.e., when $Z_g = R_g$). Generally the reflection coefficients are defined using a purely resistive Z_g. This becomes the reference resistance which is more commonly referred to as the reference impedance, Z_0, or more correctly as the system impedance.

6.3.2 Two-Port S Parameters

Two-port S parameters are defined in terms of traveling waves on transmission lines attached to each of the ports of the network as shown

in Figure 6-1(b). Individual elements are determined by measuring the forward- and backward-traveling waves with matched loads at the ports.

Determination of S_{11}: If the output line is matched, $Z_L = Z_0$, the load cannot reflect power and so $V_2^+ = 0$, then

$$S_{11} = \left.\frac{V_1^-}{V_1^+}\right|_{V_2^+=0}. \tag{6.86}$$

By alternately matching the input and output, the remaining three parameters are determined and so S_{22} is found as

$$S_{22} = \left.\frac{V_2^-}{V_2^+}\right|_{V_1^+=0}. \tag{6.87}$$

The transmission parameters are similarly obtained as

$$S_{21} = \left.\frac{V_2^-}{V_1^+}\right|_{V_2^+=0}. \tag{6.88}$$

Sometimes S_{21} is called the transmission coefficient, **T**. In the reverse direction,

$$S_{12} = \left.\frac{V_1^-}{V_2^+}\right|_{V_1^+=0}. \tag{6.89}$$

The relationships between the two-port S parameters and the common network parameters are given in Table 6-2.

A reciprocal network has $S_{12} = S_{21}$. If unit power flows into a two-port, a fraction, $|S_{11}|^2$, is reflected and a further fraction, $|S_{21}|^2$, is transmitted through the network. Fractions less than one can be considered as loss and the reflected loss is known as return loss ($=-20\log_{10}|S_{11}|$). The transmitted loss is known as the insertion loss ($=-20\log_{10}|S_{21}|$).

6.3.3 Evaluation of the Scattering Parameters of an Element

Scattering parameters can be derived analytically for various circuit configurations and in this section the procedure is illustrated for the shunt element of Figure 6-14. The aim is to determine S_{11}, S_{12}, S_{21}, and S_{22}. The procedure to find S_{11} is to match Port 2 so that $V_2^+ = 0$, then S_{11} is the reflection coefficient at Port 1:

$$S_{11} = \frac{Y_0 - Y_{\text{in}}}{Y_0 + Y_{\text{in}}}, \tag{6.90}$$

where $Y_{\text{in}} = Y_0 + Y$, since the matched termination at Port 2 (i.e., Y_o) shunts the admittance Y. Thus

$$S_{11} = \frac{Y_0 - Y_0 - Y}{2Y_0 + Y} = \frac{-Y}{2Y_0 + Y}. \tag{6.91}$$

Table 6-2 Two-port S parameter conversion chart. The z, y, and h parameters are normalized to Z_0. Z', Y', and H' are the actual parameters

	S	In terms of S
z	$Z'_{11} = z_{11}Z_0 \quad Z'_{12} = z_{12}Z_0$	$Z'_{21} = z_{21}Z_0 \quad Z'_{22} = z_{22}Z_0$
	$\delta_z = (1+z_{11})(1+z_{22}) - z_{12}z_{21}$	$\delta_S = (1-S_{11})(1-S_{22}) - S_{12}S_{21}$
	$S_{11} = [(z_{11}-1)(z_{22}+1) - z_{12}z_{21}]/\delta_z$	$z_{11} = [(1+S_{11})(1-S_{22}) + S_{12}S_{21}]/\delta_S$
	$S_{12} = 2z_{12}/\delta_z$	$z_{12} = 2S_{12}/\delta_S$
	$S_{21} = 2z_{21}/\delta_z$	$z_{21} = 2S_{21}/\delta_S$
	$S_{22} = [(z_{11}+1)(z_{22}-1) - z_{12}z_{21}]/\delta_z$	$z_{22} = [(1-S_{11})(1+S_{22}) + S_{12}S_{21}]/\delta_S$
y	$Y'_{11} = y_{11}/Z_0 \quad Y'_{12} = y_{12}/Z_0$	$Y'_{21} = y_{21}/Z_0 \quad Y'_{22} = y_{22}/Z_0$
	$\delta_y = (1+y_{11})(1+y_{22}) - y_{12}y_{21}$	$\delta_S = (1+S_{11})(1+S_{22}) - S_{12}S_{21}$
	$S_{11} = [(1-y_{11})(1+y_{22}) + y_{12}y_{21}]/\delta_y$	$y_{11} = [(1-S_{11})(1+S_{22}) + S_{12}S_{21}]/\delta_S$
	$S_{12} = -2y_{12}/\delta_y$	$y_{12} = -2S_{12}/\delta_S$
	$S_{21} = -2y_{21}/\delta_y$	$y_{21} = -2S_{21}/\delta_S$
	$S_{22} = [(1+y_{11})(1-y_{22}) + y_{12}y_{21}]/\delta_y$	$y_{22} = [(1+S_{11})(1-S_{22}) + S_{12}S_{21}]/\delta_S$
h	$H'_{11} = h_{11}Z_0 \quad H'_{12} = h_{12}$	$H'_{21} = h_{21} \quad H'_{22} = h_{22}/Z_0$
	$\delta_h = (1+h_{11})(1+h_{22}) - h_{12}h_{21}$	$\delta_S = (1-S_{11})(1+S_{22}) + S_{12}S_{21}$
	$S_{11} = [(h_{11}-1)(h_{22}+1) - h_{12}h_{21}]/\delta_h$	$h_{11} = [(1+S_{11})(1+S_{22}) - S_{12}S_{21}]/\delta_S$
	$S_{12} = 2h_{12}/\delta_h$	$h_{12} = 2S_{12}/\delta_S$
	$S_{21} = -2h_{21}/\delta_h$	$h_{21} = -2S_{21}/\delta_S$
	$S_{22} = [(1+h_{11})(1-h_{22}) + h_{12}h_{21}]/\delta_h$	$h_{22} = [(1-S_{11})(1-S_{22}) - S_{12}S_{21}]/\delta_S$
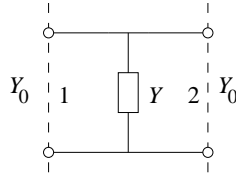
Figure 6-14 Shunt element in the form of a two-port.

From the symmetry of the two-port:

$$S_{22} = S_{11} \tag{6.92}$$

S_{21} is evaluated by determining the transmitted wave, V_2^-, with the output line matched so that again an admittance, Y_0, is placed at Port 2 and so $V_2^+ = 0$. After some algebraic manipulation,

$$S_{21} = 2Y_0/(Y + 2Y_0) \tag{6.93}$$

is obtained. Since this is clearly a reciprocal network, $S_{12} = S_{21}$ and so all four S parameters are obtained. A similar procedure of selectively applying matched loads is used to obtain the S parameters of other networks.
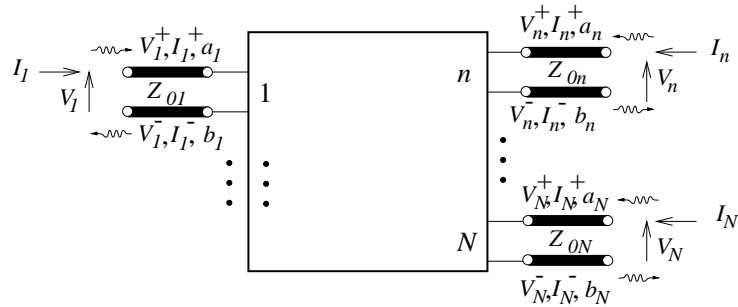
Figure 6-15 N-port network with traveling voltage and current waves.

6.3.4 Scattering Transfer or T Parameters

S parameters relate reflected waves to incident waves, thus mixing the quantities at the input and output ports. However, in dealing with cascaded networks it is preferable to relate the traveling waves at the input ports to the output ports. Such parameters are called the scattering transfer (or ST) parameters, which for a two-port network are defined by

$$\begin{bmatrix} V_1^- \\ V_1^+ \end{bmatrix} = \begin{bmatrix} ^ST_{11} & ^ST_{12} \\ ^ST_{21} & ^ST_{22} \end{bmatrix} \begin{bmatrix} V_2^+ \\ V_2^- \end{bmatrix}, \tag{6.94}$$

where

$$^S\mathbf{T} = \begin{bmatrix} ^ST_{11} & ^ST_{12} \\ ^ST_{21} & ^ST_{22} \end{bmatrix}. \tag{6.95}$$

The two-port ST parameters and S parameters are related by

$$\begin{bmatrix} ^ST_{11} & ^ST_{12} \\ ^ST_{21} & ^ST_{22} \end{bmatrix} = \begin{bmatrix} (-S_{11}S_{22} - S_{12}S_{21}/S_{21}) & (S_{11}/S_{21}) \\ -(S_{22}/S_{21}) & (1/S_{21}) \end{bmatrix}. \tag{6.96}$$

Very often the ST parameters are called T parameters, but there are at least two forms of T parameters, so it is necessary to be specific. If networks A and B have parameters $^S\mathbf{T}_A$ and $^S\mathbf{T}_B$, then the $^S\mathbf{T}$ parameters of the cascaded network are

$$^S\mathbf{T} = ^S\mathbf{T}_A \cdot ^S\mathbf{T}_B. \tag{6.97}$$

6.4 The N-Port Network

The N-port network is a generalization of a two-port, as you may have guessed. A network with many ports is shown in Figure 6-15. Again, each port consists of a pair of terminals, one of which is the reference for voltage. Each port has equal and opposite currents at the two terminals. The incident

and reflected voltages at any port can be related to each other using the
voltage scattering parameter matrix relation:

$$
\begin{bmatrix} V_1^- \\ V_2^- \\ \vdots \\ V_N^- \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} & \ldots & S_{1N} \\ S_{21} & S_{22} & \ldots & S_{2N} \\ \vdots & & \ddots & \\ S_{N1} & S_{N2} & \ldots & S_{NN} \end{bmatrix} \begin{bmatrix} V_1^+ \\ V_2^+ \\ \vdots \\ V_N^+ \end{bmatrix},
\tag{6.98}
$$

or in compact form as

$$
\mathbf{V}^- = \mathbf{S}\mathbf{V}^+.
\tag{6.99}
$$

Notice that

$$
S_{ij} = \left. \frac{V_i^-}{V_j^+} \right|_{V_k^+ = 0 \,\text{for}\, k \neq j}.
\tag{6.100}
$$

In words, S_{ij} is found by driving Port j with an incident wave of voltage
V_j^+ and measuring the reflected wave V_i^- at Port i, with all ports other than
j terminated in a matched load. Reflection and transmission coefficients can
also be defined using the above relationship:

- S_{ii} : reflection coefficient seen looking into Port i

- S_{ij} : transmission coefficient from j to i.

6.4.1 Generalized Scattering Parameter Relations

The S parameters used so far have the same reference impedance at each
port. These can be generalized to have different reference impedances
at each port. These are useful if the actual system being considered has
different loading conditions at various ports. Generalized S parameters
are defined in terms of what are called root power waves, which in
turn are defined using forward- and backward-traveling voltage waves.
Consider the N-port network of Figure 6-15, where the nth port has a
reference transmission line of characteristic impedance Z_{0n}, which can have
infinitesimal length. The transmission line at the nth port serves to separate
the forward- and backward-traveling voltage (V_n^+ and V_n^-) and current (I_n^+
and I_n^-) waves.

The reference characteristic impedance matrix $\mathbf{Z}_0$ is a diagonal matrix, $\mathbf{Z}_0$
$= \text{diag}(Z_{01} \ldots Z_{0n} \ldots Z_{0N})$, and the root power waves at the nth port, a_n and
b_n, are defined by[3]

$$
a_n = V_n^+ / \sqrt{\Re\{Z_{0n}\}} \quad \text{and} \quad b_n = V_n^- / \sqrt{\Re\{Z_{0n}\}},
\tag{6.101}
$$

and shown in Figure 6-16. In matrix form

---

[3]   The symbol $\Re$ indicates that the real part is used.
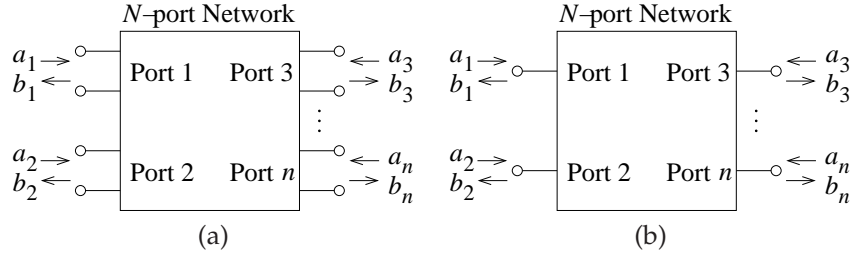
Figure 6-16 N-ports defining the a and b root power waves: (a) two-terminal ports; and ports on their own.

$$\mathbf{a} = \mathbf{Z}_0^{-1/2}\mathbf{V}^+ = \mathbf{Y}_0^{1/2}\mathbf{V}^+, \qquad \mathbf{b} = \mathbf{Z}_0^{-1/2}\mathbf{V}^- = \mathbf{Y}_0^{1/2}\mathbf{V}^-, \quad (6.102)$$

$$\mathbf{V}^+ = \mathbf{Z}_0^{1/2}\mathbf{a} = \mathbf{Y}_0^{-1/2}\mathbf{a}, \qquad \mathbf{V}^- = \mathbf{Z}_0^{1/2}\mathbf{b} = \mathbf{Y}_0^{-1/2}\mathbf{b}, \quad (6.103)$$

where

$$\mathbf{a} = [a_1 \dots a_n \dots a_N]^{\mathrm{T}}, \qquad \mathbf{b} = [b_1 \dots b_n \dots b_N]^{\mathrm{T}}, \qquad (6.104)$$

$$\mathbf{V}^+ = [V_1^+ \dots V_n^+ \dots V_N^+]^{\mathrm{T}}, \qquad \mathbf{V}^- = [V_1^- \dots V_n^- \dots V_N^-]^{\mathrm{T}}, \quad (6.105)$$

and the characteristic admittance matrix is $\mathbf{Y}_0 = \mathbf{Z}_0^{-1}$. It can also be shown that the power wave parameters can be viewed as

$$|a| = \sqrt{\text{incident power}} \quad \text{and} \quad |b| = \sqrt{\text{reflected power}}. \qquad (6.106)$$

It can also be shown that on each reference line,

$$\mathbf{a} = \frac{\mathbf{V} + \mathbf{Z_o}\mathbf{I}}{2\sqrt{\Re\{\mathbf{Z_o}\}}} \qquad (6.107)$$

and

$$\mathbf{b} = \frac{\mathbf{V} - \mathbf{Z_o}^*\mathbf{I}}{2\sqrt{\Re\{\mathbf{Z_o}\}}}. \qquad (6.108)$$

Now, generalized S parameters can be formally defined as

$$\mathbf{b} = \mathbf{S}\mathbf{a}, \qquad (6.109)$$

thus $\mathbf{Y}_0^{1/2}\mathbf{V}^- = \mathbf{S}\mathbf{Y}_0^{1/2}\mathbf{V}^+$, and so $\mathbf{V}^- = \mathbf{Y}_0^{-1/2}\mathbf{S}\mathbf{Y}_0^{1/2}\mathbf{V}^+$. This reduces to $\mathbf{V}^- = \mathbf{S}\mathbf{V}^+$ when all of the reference transmission lines have the same characteristic impedance. However, when the ports have different reference impedances we use $\mathbf{S}^V$ for voltage scattering parameters and $\mathbf{S}^I$ for current scattering parameters, where $\mathbf{V}^- = \mathbf{S}^V\mathbf{V}^+$ and $\mathbf{I}^- = \mathbf{S}^I\mathbf{I}^+$.

The following conversion relationships can also be derived:

$$\mathbf{S^I} = \Re\{\mathbf{Z_o}\}^{-\frac{1}{2}}\mathbf{S}\Re\{\mathbf{Z_o}\}^{\frac{1}{2}} \qquad (6.110)$$

$$\mathbf{S^V} = \mathbf{Z_o}\Re\{\mathbf{Z_o}\}^{-\frac{1}{2}}\mathbf{S}\Re\{\mathbf{Z_o}\}^{\frac{1}{2}}\mathbf{Z_o}^{*\,-1}, \qquad (6.111)$$

where $\Re\{\mathbf{Z_o}\}^{\frac{1}{2}} = \mathrm{diag}\{\sqrt{\Re\{Z_{o1}\}}, \sqrt{\Re\{Z_{o2}\}}, \dots, \sqrt{\Re\{Z_{on}\}}\}$.

Recall that $\mathbf{S}$ is in terms of $\mathbf{a}$ and $\mathbf{b}$; $\mathbf{S^I}$ is in terms of $\mathbf{I}^-$ and $\mathbf{I}^+$; and $\mathbf{S^V}$ is in terms of $\mathbf{V}^-$ and $\mathbf{V}^+$. When port impedances and reference resistances are real, Equations (6.110) and (6.111) assume the simpler forms

$$
\begin{aligned}
\mathbf{S^I} &= -R_o^{-\frac{1}{2}} \, \mathbf{S} \, R_o^{\frac{1}{2}} = -\mathbf{S}, & (6.112) \\
\mathbf{S^V} &= R_o \, R_o^{-\frac{1}{2}} \, \mathbf{S} \, R_o^{\frac{1}{2}} \, R_o^{*-1} & (6.113) \\
&= R_o \, R_o^{*-1} \, \mathbf{S} = \mathbf{S}, & (6.114)
\end{aligned}
$$

where $R_o^{\frac{1}{2}} = \mathrm{diag}\{\sqrt{R_{o1}}, \sqrt{R_{o2}}, \dots, \sqrt{R_{on}}\}$ and R_{on} being the reference resistance at the nth port. In addition, if the reference resistances at each port are the same, all the various scattering parameter definitions become equivalent (i.e., $\mathbf{S^V} = -\mathbf{S^I} = \mathbf{S}$).

6.4.2 Normalized and Generalized S Parameters

S parameters measured with respect to a common reference resistance are often referred to as normalized S parameters. In almost all cases measured S parameters are normalized to $50\ \Omega$, as $50\ \Omega$ cables and components are used in the measurement system. In other words, $Z_{01} = Z_{02} = \dots = Z_{0N} = R_0$ (= $50\ \Omega$). Generalized S parameters can be used to simplify the design process of devices such as amplifiers because they can be used to include complex impedances at the transistor terminals. It is often desirable to be able to convert between measured S parameters (normalized to $50\ \Omega$) and generalized S parameters. Let $\mathbf{S}^0$ be the measured (normalized) S parameter matrix. The procedure is tedious, but it can be shown that the generalized S parameters are

$$
\mathbf{S} = (\mathbf{D}^*)^{-1}(\mathbf{S}^0 - \boldsymbol{\Gamma}^*)(\mathbf{U} - \boldsymbol{\Gamma}\,\mathbf{S}^0)^{-1}\mathbf{D}, \tag{6.115}
$$

where $\mathbf{U}$ is the unit matrix ($\mathbf{U} = \mathrm{diag}(1, 1, \dots, 1)$) and $\mathbf{D}$ is a diagonal matrix with elements

$$
\begin{aligned}
D_{ii} &= |1 - \Gamma_i^*|^{-1}(1 - \Gamma_i)\sqrt{1 - |\Gamma_i|^2} & (6.116) \\
\Gamma_i &= (Z_{0i} - R_0)(Z_{0i} + R_0)^{-1} \qquad i = 1, 2, \dots, N & (6.117)
\end{aligned}
$$

and $\boldsymbol{\Gamma}$ is a diagonal matrix with elements Γ_i. Z_{0i} is the system impedance at Port i to which the generalized S parameters are to be referred. This result is used in Section 11.6.1 on Page 629.

6.4.3 Passivity in Terms of Scattering Parameters

Consider an N-port characterized by its generalized scattering matrix $\mathbf{S}$. The time-average power dissipated in the N-port is

$$P \;=\; \frac{1}{2}\sum_{i=1}^{N}(|\mathbf{a}_i|^2 - |\mathbf{b}_i|^2) \tag{6.118}$$

$$=\; \frac{1}{2}(\mathbf{a}^{*\mathrm{T}}\mathbf{a} - \mathbf{b}^{*\mathrm{T}}\mathbf{b}), \tag{6.119}$$

and so

$$P = \frac{1}{2}\mathbf{a}^{*\mathrm{T}}(\mathbf{U} - \mathbf{S}^{*\mathrm{T}}\mathbf{S})\,\mathbf{a}. \tag{6.120}$$

Above the conjugate, $\mathbf{a}^*$, of the matrix $\mathbf{a}$ is obtained from $\mathbf{a}$ by taking the complex conjugate of each element. For a passive N-port,

$$\mathbf{U} - \mathbf{S}^{*\mathrm{T}}\mathbf{S} \geq 0 \quad \text{for all real } \omega, \tag{6.121}$$

and for a lossless N-port,

$$\mathbf{U} - \mathbf{S}^{*\mathrm{T}}\mathbf{S} = 0 \tag{6.122}$$

or

$$\mathbf{S}^* = \left(\mathbf{S}^{\mathrm{T}}\right)^{-1}. \tag{6.123}$$

This can also be written in the summation form as

$$\sum_{i=1}^{N} S_{ki}S_{kj}^* = \Gamma_{ij} \quad \text{for all } i,j. \tag{6.124}$$

So $\mathbf{S}$ is a unitary matrix if the network is lossless. By examining the S parameters it can be determined quickly whether a network is lossless, lossy, or perhaps has gain.

6.4.4 Impedance Matrix Representation

In this section, N-port S parameters are related to N-port z parameters. The basic relationship of voltage and current at any port using the impedances is given by the matrix equation

$$\begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_N \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1\mathrm{N}} \\ z_{21} & z_{22} & \cdots & z_{2\mathrm{N}} \\ \vdots & & \ddots & \\ z_{\mathrm{N}1} & z_{\mathrm{N}2} & \cdots & z_{\mathrm{NN}} \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_\mathrm{N} \end{bmatrix}, \tag{6.125}$$

or in compact form as

$$\mathbf{V} = \mathbf{Z}\mathbf{I}. \tag{6.126}$$

$\mathbf{Z}$ is said to be symmetric if $Z_{ij} = Z_{ji}$. In other words, if the $\mathbf{Z}$ matrix is reciprocal, then $\mathbf{Z}$ is symmetric. The z parameters defined here are more formally called port-based z parameters, as the voltage and current variables are port quantities.

Relating z and S parameters begins by relating the total voltage and current at the nth terminal plane to the traveling voltage and current waves. From Figure 6-15,

$$V_{\mathrm{n}} = V_{\mathrm{n}}^{+} + V_{\mathrm{n}}^{-} \tag{6.127}$$

$$I_{\mathrm{n}} = I_{\mathrm{n}}^{+} - I_{\mathrm{n}}^{-} \, , \tag{6.128}$$

and in vector form

$$\mathbf{V} = \mathbf{V}^{+} + \mathbf{V}^{-} \tag{6.129}$$

$$\mathbf{I} = \mathbf{I}^{+} + \mathbf{I}^{-} \tag{6.130}$$

$$\mathbf{V}^{+} = \mathbf{Z}_{0}^{*}\mathbf{I}^{+} \tag{6.131}$$

$$\mathbf{V}^{-} = -\mathbf{Z}_{0}\mathbf{I}^{-} \, , \tag{6.132}$$

where $\mathbf{Z}_0 = \mathrm{diag}(Z_{01}, Z_{02}, \ldots, Z_{0N})$. After some algebraic manipulation, the relationships

$$\mathbf{S}^{V} = [\mathbf{U} + \mathbf{Z}\mathbf{Z}_{0}^{-1}]^{-1}[\mathbf{Z}(\mathbf{Z}_{0}^{*})^{-1} - \mathbf{U}] \tag{6.133}$$

$$\mathbf{Z} = [\mathbf{U} + \mathbf{S}^{V}][\mathbf{Z}_{0}^{*-1} - \mathbf{Z}_{0}^{-1}(\mathbf{S}^{V})^{-1}] \tag{6.134}$$

are obtained.

6.4.5 Admittance Matrix Representation

In this section, N-port S parameters are related to N-port y parameters. The relationship of the current and voltage at any port through the admittances is

$$\left[\begin{array}{c} I_1 \\ I_2 \\ \vdots \\ I_N \end{array} \right] = \left[\begin{array}{cccc} y_{11} & y_{12} & \cdots & y_{1\mathrm{N}} \\ y_{21} & y_{22} & \cdots & y_{2\mathrm{N}} \\ \vdots & & \ddots & \\ y_{\mathrm{N}1} & y_{\mathrm{N}2} & \cdots & y_{\mathrm{NN}} \end{array} \right] \left[\begin{array}{c} V_1 \\ V_2 \\ \vdots \\ V_{\mathrm{N}} \end{array} \right] , \tag{6.135}$$

or in compact form the port-based y parameters are

$$\mathbf{I} = \mathbf{Y}\mathbf{V}. \tag{6.136}$$

Again, $\mathbf{Y}$ is symmetric for a reciprocal network.

Using a similar approach to that in the previous subsection, the relationship between S and y parameters can be developed. The development will be done slightly differently and this development is applicable to generalized scattering parameters. First, consider the relationship of the total port

voltage $\mathbf{V} = [V_1 \dots V_n \dots V_N]^{\mathrm{T}}$ and current $\mathbf{I} = [I_1 \dots I_n \dots I_N]^{\mathrm{T}}$ to forward and backward voltage and current waves:

$$\mathbf{V} = \mathbf{V}^+ + \mathbf{V}^- \quad \text{and} \quad \mathbf{I} = \mathbf{I}^+ + \mathbf{I}^-, \tag{6.137}$$

where $\mathbf{I}^+ = \mathbf{Y}_0\mathbf{V}^+ = \mathbf{Y}_0^{1/2}\mathbf{a}$ and $\mathbf{I}^- = -\mathbf{Y}_0\mathbf{V}^- = -\mathbf{Y}_0^{1/2}\mathbf{b}$. The development of the relationship between S parameters and other network parameters is illustrated by considering y parameters defined by

$$\mathbf{I} = \mathbf{YV}. \tag{6.138}$$

Using traveling waves, this becomes

$$\begin{align}
\mathbf{I}^+ + \mathbf{I}^- &= \mathbf{Y}(\mathbf{V}^+ + \mathbf{V}^-) \tag{6.139}\\
\mathbf{Y}_0(\mathbf{V}^+ - \mathbf{V}^-) &= \mathbf{Y}(\mathbf{V}^+ + \mathbf{V}^-) \tag{6.140}\\
\mathbf{Y}_0(\mathbf{1} - \mathbf{Y}_0^{-1/2}\mathbf{SY}_0^{1/2})\mathbf{V}^+ &= \mathbf{Y}(\mathbf{1} + \mathbf{Y}_0^{-1/2}\mathbf{SY}_0^{1/2})\mathbf{V}^+, \tag{6.141}
\end{align}$$

and so

$$\mathbf{Y} = \mathbf{Y}_0(\mathbf{1} - \mathbf{Y}_0^{-1/2}\mathbf{SY}_0^{1/2})(\mathbf{1} + \mathbf{Y}_0^{-1/2}\mathbf{SY}_0^{1/2})^{-1}. \tag{6.142}$$

Alternatively Equation (6.140) can be rearranged as

$$\begin{align}
(\mathbf{Y}_0 + \mathbf{Y})\mathbf{V}^- &= (\mathbf{Y}_0 - \mathbf{Y})\mathbf{V}^+ \tag{6.143}\\
\mathbf{V}^- &= (\mathbf{Y}_0 + \mathbf{Y})^{-1}(\mathbf{Y}_0 - \mathbf{Y})\mathbf{V}^+ \tag{6.144}\\
\mathbf{Y}_0^{-1/2}\mathbf{b} &= (\mathbf{Y}_0 + \mathbf{Y})^{-1}(\mathbf{Y}_0 - \mathbf{Y})\mathbf{Y}_0^{-1/2}\mathbf{a}. \tag{6.145}
\end{align}$$

Comparing this to the definition of S parameters in Equation (6.109) leads to

$$\mathbf{S} = \mathbf{Y}_0^{1/2}(\mathbf{Y}_0 + \mathbf{Y})^{-1}(\mathbf{Y}_0 - \mathbf{Y})\mathbf{Y}_0^{-1/2}. \tag{6.146}$$

For the usual case where all of the reference transmission lines have the same characteristic impedance $Z_0 = 1/Y_0$, and so

$$\mathbf{Y} = Y_0(\mathbf{1} - \mathbf{S})(\mathbf{1} + \mathbf{S})^{-1} \tag{6.147}$$

and

$$\mathbf{S} = (\mathbf{Y}_0 + \mathbf{Y})^{-1}(\mathbf{Y}_0 - \mathbf{Y}). \tag{6.148}$$

6.5 Scattering Parameter Matrices of Common Two-Ports

RF and microwave circuits can generally be represented as interconnected two-ports, as most RF and microwave circuit designs involve cascaded functional blocks such as an amplifier, matching networks, a filter, etc.[4] (see

---

[4] This arrangement tends to maximize bandwidth, minimize losses, and maximize efficiency. Lower-frequency analog design utilizes more complex arrangements; for example, feedback high in the circuit hierarchy improves reliability and robustness of design but comes at the cost of reduced bandwidth and lower power efficiency.
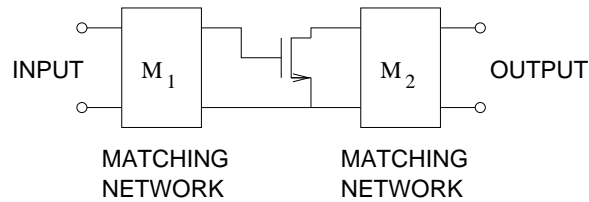
Figure 6-17   An amplifier with an active device and input, M_1, and output, M_2, matching networks. The biasing arrangement is not shown, but usually this is done with appropriate choice of matching network topology.
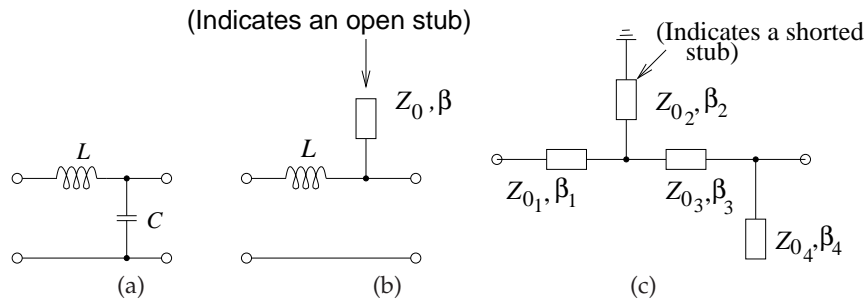
Figure 6-18   Three forms of matching networks with increasing use of transmission line sections.

Figure 6-17). Thus there is great interest in various manipulations that can be performed on two-ports as well as the network parameters of common two-port circuit topologies. As an example, consider the matching networks in Figure 6-17. These are used to achieve maximum power transfer in an amplifier by acting as impedance transformers. Matching networks assume a variety of forms, as shown in Figure 6-18, and all can be viewed as two-port networks and a combination of simpler components. In this section, strategies are presented for developing the S parameters of two-ports.

6.5.1   Transmission Line (Figure 6-19(a))

The traveling waves on a transmission line have a phase that depends on the electrical length, θ, of the line. The transmission line has a characteristic impedance, Z_0, and length, ℓ, which in general is different from the system reference impedance, here Z_{01}. Thus

$$S_{11} = S_{22} = \frac{\Gamma(1 - e^{-2j\theta})}{1 - \Gamma^2 e^{-2j\theta}} \tag{6.149}$$

$$S_{21} = S_{12} = \frac{(1 - \Gamma^2)e^{-j\theta}}{1 - \Gamma^2 e^{-2j\theta}}, \tag{6.150}$$
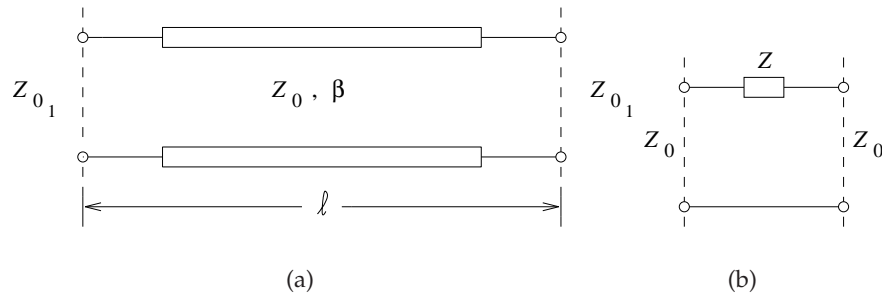
Figure 6-19 Two-ports: (a) section of transmission line; and (b) series element in the form of a two-port.

where $\theta = \beta \ell$ and

$$\Gamma = \frac{Z_o - Z_{o1}}{Z_o + Z_{o1}} \ . \tag{6.151}$$

6.5.2 Shunt Element (Figure 6-14)

The S parameters of the shunt element were developed in Section 6.3.3 on Page 303. In a slightly different form these are

$$S_{11} \quad = \quad S_{22} = -\frac{\overline{y}}{(\overline{y} + 2)} \tag{6.152}$$

$$S_{12} \quad = \quad S_{21} = \frac{2}{(\overline{y} + 2)}, \tag{6.153}$$

where $\overline{y}$ is the admittance normalized to the system reference admittance $(Y_o = 1/Z_o)$:

$$\overline{y} = Y/Y_o. \tag{6.154}$$

6.5.3 Series Element (Figure 6-19(b))

The S parameters of the series element are

$$S_{11} \quad = \quad S_{22} = \frac{\overline{z}}{(\overline{z} + 2)} \tag{6.155}$$

$$S_{12} \quad = \quad S_{21} = \frac{2}{(\overline{z} + 2)}, \tag{6.156}$$

where $\overline{z}$ is the normalized impedance given by

$$\overline{z} = Z/Z_o \ . \tag{6.157}$$
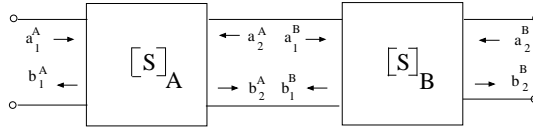
Figure 6-20   Two cascaded two-ports.

6.6 Scattering Parameter Two-Port Relationships

6.6.1 Cascaded Two-Port Networks

Many types of scattering parameters can be used to represent two-port networks. Two two-port networks, A and B, in cascade are shown in Figure 6-20. Here (A) and (B) are used as superscripts to distinguish the parameters of each two-port network, but the subscripts A and B are used for matrix quantities. Since

$$a_2^{(A)} \ = \ b_1^{(B)} \quad \text{and} \quad b_2^{(A)} = a_1^{(B)}, \tag{6.158}$$

it is convenient to put the a and b parameters in cascadable form, leading to the following two-port representation:

$$\begin{bmatrix} a_1 \\ b_1 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} b_2 \\ a_2 \end{bmatrix}, \tag{6.159}$$

where the T-matrix or chain scattering matrix is

$$\mathbf{T} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}. \tag{6.160}$$

$\mathbf{T}$ is very similar to the scattering transfer matrix of Section 6.3.4 on Page 305. The only difference is the ordering of the a and b components. You will come across both forms, so be careful that you understand their placements in the definition. Both forms are used for the same function—cascading two-port networks. The relationships between $\mathbf{T}$ and $\mathbf{S}$ are given by

$$\mathbf{T} = \begin{bmatrix} S_{21}^{-1} & -S_{21}^{-1} S_{22} \\ S_{21}^{-1} S_{11} & S_{12} - S_{11} S_{21}^{-1} S_{22} \end{bmatrix} \tag{6.161}$$

and

$$\mathbf{S} = \begin{bmatrix} T_{21} T_{11}^{-1} & T_{22} - T_{21} T_{11}^{-1} T_{12} \\ T_{11}^{-1} & -T_{11}^{-1} T_{12} \end{bmatrix}. \tag{6.162}$$

For a two-port network, using Equations (6.158) and (6.159),

$$\begin{bmatrix} a_1^{(A)} \\ b_1^{(A)} \end{bmatrix} = \mathbf{T}_A \begin{bmatrix} b_2^{(A)} \\ a_2^{(A)} \end{bmatrix} \tag{6.163}$$

and

$$\left[\begin{array}{c} a_1^{(B)} \\ b_1^{(B)} \end{array} \right] = \mathbf{T}_B \left[\begin{array}{c} b_2^{(B)} \\ a_2^{(B)} \end{array} \right].$$ (6.164)

Thus

$$\left[\begin{array}{c} a_1^{(A)} \\ b_1^{(A)} \end{array} \right] = \mathbf{T}_A \mathbf{T}_B \left[\begin{array}{c} b_2^{(B)} \\ a_2^{(B)} \end{array} \right].$$ (6.165)

For n cascaded two-port networks, Equation (6.165) generalizes to

$$\left[\begin{array}{c} a_1^{(1)} \\ b_1^{(1)} \end{array} \right] = \mathbf{T}_1 \mathbf{T}_2 \ldots \mathbf{T}_n \left[\begin{array}{c} b_2^{(n)} \\ a_2^{(n)} \end{array} \right],$$ (6.166)

and so the **T**-matrix of the cascaded network is the matrix product of the **T**-matrices of the individual two-ports.

6.6.2 Change in Reference Plane

It is often necessary during S parameter measurements of two-port devices to measure components at a position different from that actually desired. An example is shown in Figure 6-21(a). From direct measurement the S parameters are obtained, and thus the **T** matrix at Planes 1 and 2. However, $\mathbf{T}_{\text{DUT}}$ referenced to the Planes $1'$ and $2'$ is required.[5] Now,

$$\mathbf{T} = \mathbf{T}_{\theta_1} \mathbf{T}_{\text{DUT}} \mathbf{T}_{\theta_2},$$ (6.167)

and so

$$\mathbf{T}_{\text{DUT}} = \mathbf{T}_{\theta_1}^{-1} \mathbf{T} \mathbf{T}_{\theta_2}^{-1}.$$ (6.168)

A section of line with electrical length θ and port impedances equal to its characteristic impedance has

$$\mathbf{S} = \left[\begin{array}{cc} 0 & e^{-\jmath\theta} \\ e^{-\jmath\theta} & 0 \end{array} \right]$$ (6.169)

and

$$\mathbf{T}_\theta = \left[\begin{array}{cc} e^{\jmath\theta} & 0 \\ 0 & e^{-\jmath\theta} \end{array} \right].$$ (6.170)

Therefore Equation (6.168) becomes

$$\mathbf{T}_{\text{DUT}} = \left[\begin{array}{cc} T_{11}e^{-\jmath(\theta_1+\theta_2)} & T_{12}e^{-\jmath(\theta_1-\theta_2)} \\ T_{21}e^{\jmath(\theta_1-\theta_2)} & T_{22}e^{\jmath(\theta_1+\theta_2)} \end{array} \right],$$ (6.171)

---

[5] Here, DUT stands for Device Under Test and is the common name given to a device being measured.
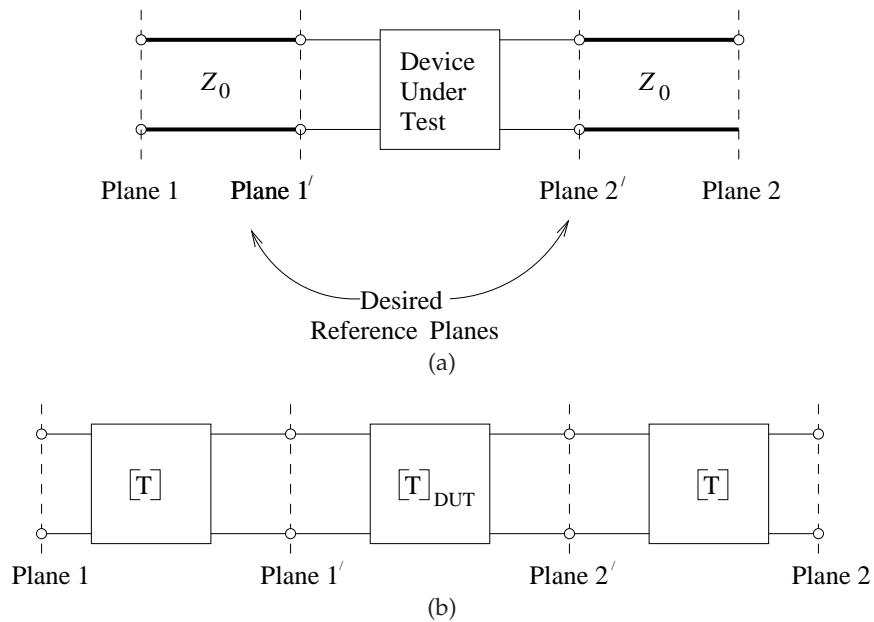
Figure 6-21 Two-port measurement setup: (a) a two-port comprising a Device Under Test (DUT) and transmission line sections that create a reference plane at Planes 1 and 2; and (b) representation as cascaded two-port networks.

and converting back to S parameters, the desired S parameters of the DUT are obtained as

$$\mathbf{S}_{\text{DUT}} = \left[\begin{array}{cc} S_{11}e^{\jmath 2\theta_1} & S_{12}e^{\jmath(\theta_1+\theta_2)} \\ S_{21}e^{\jmath(\theta_1+\theta_2)} & S_{22}e^{\jmath 2\theta_2} \end{array} \right]. \tag{6.172}$$

6.6.3 Conversion Between S Parameters and $ABCD$ Parameters

Figure 6-22 can be used to relate the appropriate parameters for the two views of the network. Assume that the ports are terminated in impedance Z_o, then

$$\left[\begin{array}{c} V_1 \\ I_1 \end{array} \right] = \left[\begin{array}{cc} A & B \\ C & D \end{array} \right] \left[\begin{array}{c} V_2 \\ I_2 \end{array} \right] \tag{6.173}$$

$$\left[\begin{array}{c} b_1 \\ b_2 \end{array} \right] = \left[\begin{array}{cc} S_{11} & S_{12} \\ S_{21} & S_{22} \end{array} \right] \left[\begin{array}{c} a_1 \\ a_2 \end{array} \right]. \tag{6.174}$$
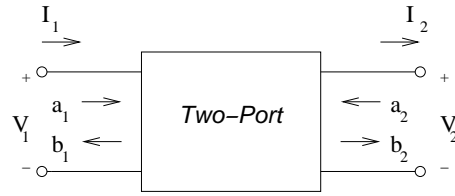
Figure 6-22 Two-port with parameters suitable for defining S and $ABCD$ parameters.

The S parameters are then expressed as

$$S_{11} = \frac{A + B/Z_o - CZ_o - D}{\triangle} \tag{6.175}$$

$$S_{12} = \frac{2(AD - BC)}{\triangle} \tag{6.176}$$

$$S_{21} = \frac{2}{\triangle} \tag{6.177}$$

$$S_{22} = \frac{-A + B/Z_o - CZ_o + D}{\triangle}, \tag{6.178}$$

where

$$\triangle = A + B/Z_o + CZ_o + D. \tag{6.179}$$

The $ABCD$ parameters can be expressed in terms of the S parameters as

$$A = \frac{(1 + S_{11})(1 - S_{22}) + S_{12}S_{21}}{2S_{21}} \tag{6.180}$$

$$B = Z_o \frac{(1 + S_{11})(1 + S_{22}) - S_{12}S_{21}}{2S_{21}} \tag{6.181}$$

$$C = \frac{1}{Z_o} \frac{(1 - S_{11})(1 - S_{22}) - S_{12}S_{21}}{2S_{21}} \tag{6.182}$$

$$D = \frac{(1 - S_{11})(1 + S_{22}) - S_{12}S_{21}}{2S_{21}}. \tag{6.183}$$

6.7 Signal Flow Graph

Signal Flow Graphs (**SFGs**) are convenient ways to represent systems of simultaneous linear equations [80, 81]. They are particularly useful in manipulating scattering parameters. Begin by considering representations of block diagrams before considering their application to S parameter representations. SFGs can be used in many disciplines but they are
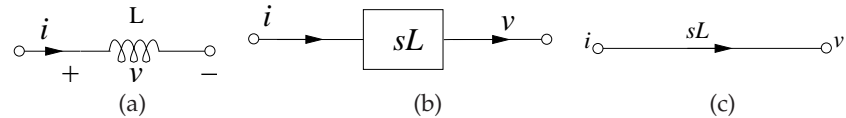
Figure 6-23   An inductor represented as: (a) a two-terminal element; (b) a block diagram with input i and output v; and (c) a signal flow graph.

particularly useful with RF and microwave circuits, as most of these circuits comprise very large numbers of linear components and many useful circuits are entirely linear.

An SFG graphically represents a linear operation on an input. Consider the inductor shown in Figure 6-23(a) when the circuit quantities are v and i, and these are related by the impedance of the inductor sL. The block diagram representation of this relation is given in Figure 6-23(b), with a slight simplification obtained by removing the blocks and writing the operation performed beside a directed edge.[6] The edge is directed from the excitation node to the response node as shown in Figure 6-23(c). The algebraic relationship between v and i is described in the Laplace domain by the constitutive relation

$$v = sL, \tag{6.184}$$

and this is exactly how the SFG of Figure 6-23(c) is interpreted.

A more complicated example is the linearized transistor model shown in Figure 6-24(a) and represented in block diagram form in Figure 6-24(b). Figure 6-24(b) is the single block equivalent of the series connection of blocks in Figure 6-24(c), which in SFG form becomes Figure 6-24(d).

6.7.1   Signal Flow Graph Manipulation

In this section the basic rules for manipulating SFGs are presented and it will be seen why the SFG is an important concept in working with RF and microwave circuits. Keep in mind that SFG representations are used principally in working with S parameters. Think back to when you first started working with circuits. The great abstraction came about when the physical world was represented graphically as connections of circuit elements. Provided that a few simple rules were followed, the graphical representation, coupled with the human aptitude in recognizing patterns, resulted in recognition of circuit topologies and the selection of appropriate solution strategies (e.g., applying the voltage divider rule). When it comes to working with S parameters and interconnections of multiport networks, SFGs serve much the same purpose. As well, SFG analysis enables the development of symbolic expressions. Only a portion of SFG theory is

---

[6]   Edges and nodes are quantities in graph theory of which SFG theory is a subset. A branch connects two nodes. A branch is also called an edge.
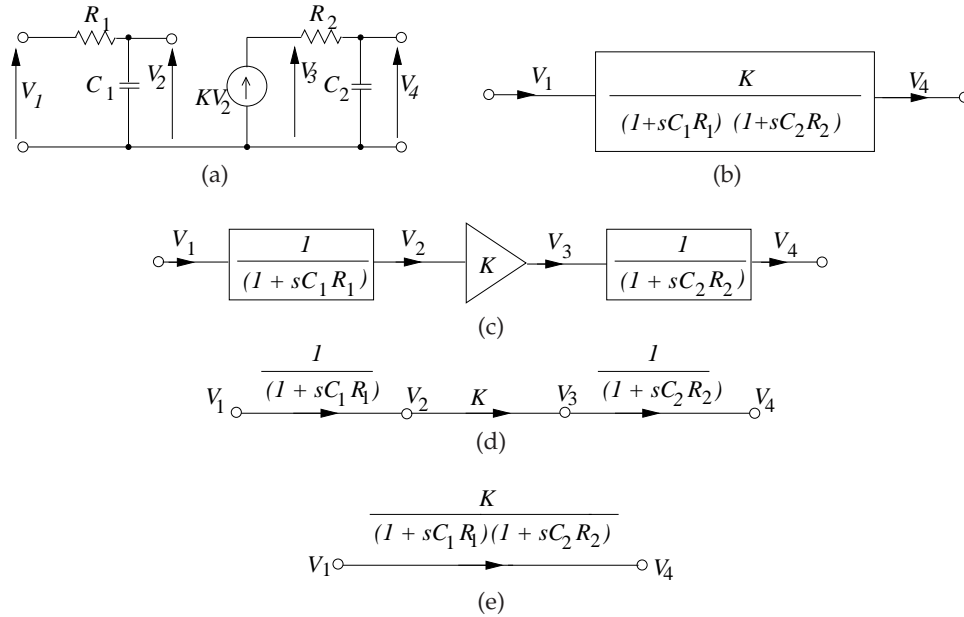
Figure 6-24   A linear transistor model: (a) schematic; (b) block diagram representation; (c) alternative block diagram representation; (d) signal flow graph; and (e) reduced signal flow graph.

considered here—the aspects relevant to manipulating scattering parameter descriptions. Balabanian [82], Abrahams and Coverly [83], and Di Stefano et al. [84] provide a more detailed and general treatment.

Several linear equations are represented in SFG form in Table 6-3. While every attempt is made to draw SFGs so that they can be read from left to right with input quantities on the left and output quantities on the right, this is not always possible or the clearest way to express relationships. Nevertheless, the convention is generally adhered to as an aid to easy recognition of patterns just as circuits are laid out from left to right if possible.

Returning to the SFG of the transistor amplifier model, the SFG of Figure 6-24(d) may be reduced to that shown in Figure 6-24(e), which should be compared to the equivalent block diagram of Figure 6-24(b).
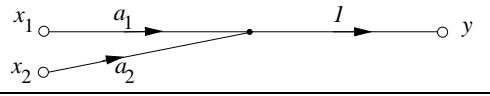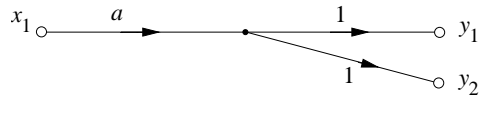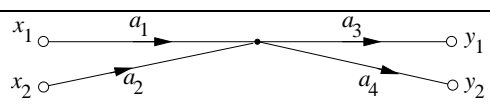
6.7.2   Signal Flow Graph Representation of Scattering Parameters

Scattering parameters relate incident and reflected waves:

$$\mathbf{b} = \mathbf{S}\mathbf{a}, \tag{6.185}$$

where $\mathbf{a}$ and $\mathbf{b}$ are vectors and their ith elements refer to the incident and reflected waves, respectively. These relationships can be represented by

Table 6-3 Mathematical relations expressed as signal flow graphs.

(a)		$y = a_1 x_1 + a_2 x_2$
(b)		$y_1 = a x_1$ $y_2 = a x_1$
(c)		$y_1 = a_3(a_1 x_1 + a_2 x_2)$ $y_2 = a_4(a_1 x_1 + a_2 x_2)$
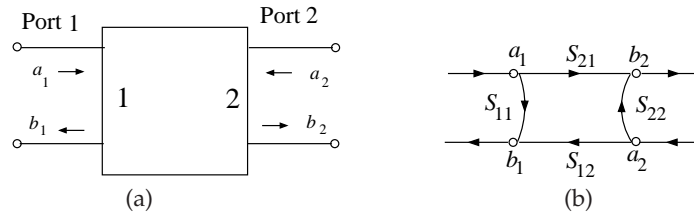
Figure 6-25 Two-port represented as (a) a two-port network with incident and reflected waves; and (b) its SFG representation.

SFGs. Consider the two-port in Figure 6-25(a) described by the equations

$$b_1 = S_{11}a_1 + S_{12}a_2 \tag{6.186}$$
$$b_2 = S_{21}a_1 + S_{22}a_2, \tag{6.187}$$

which are represented in SFG form in Figure 6-25(b).

6.7.3 Simplification and Reduction of Signal Flow Graphs

The power of SFG analysis is that an SFG can be formulated by building up the set of equations describing a network by connecting together the SFGs of sections. Pattern recognition can be used to identify constraints that can be reduced and simplified to arrive at a simple relationship between system input and output.

Addition

Figure 6-26 depicts SFG reduction obtained using addition. Both Figure 6-26(a) and Figure 6-26(b) denote

$$x_2 = G_1 x_1 + G_2 x_2. \tag{6.188}$$
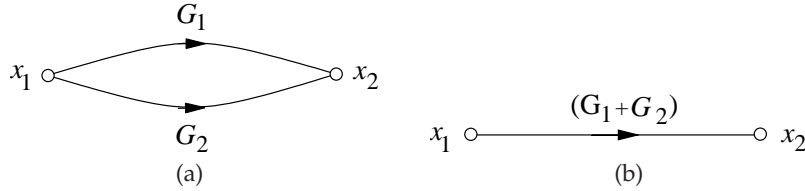
Figure 6-26 Signal flow graph representation of addition.
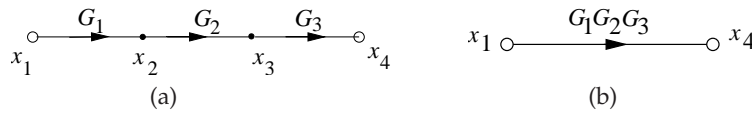
Figure 6-27 Cascade reduction of SFG: (a) three cascaded blocks; and (b) reduced form.
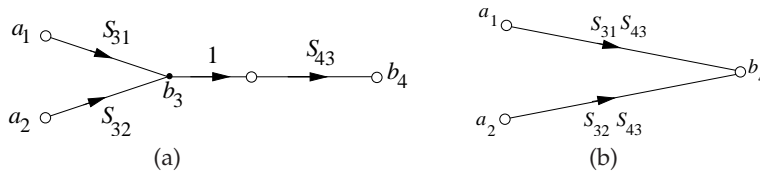
Figure 6-28 Signal flow graph simplification illustrating the elimination of a variable.

Multiplication

Consider the three cascaded blocks represented by the SFG of Figure 6-27(a). Here the output of the first block, x_2, is described by $x_2 = G_1 x_1$. Now x_2 is the input to the second block with output $x_3 = G_2 x_2$, and so on. The total response is obtained by multiplying the individual responses (see Figure 6-27(b)):

$$x_4 = G_1 G_2 G_3 x_1. \tag{6.189}$$

Commutation

Reduction using multiplication and addition are straightforward. The rules that govern the simplification of more complicated graphs use the fact that each graph represents a set of simultaneous equations. Consider the removal of the mixed node, b_3, in Figure 6-28(a). Here

$$b_3 = S_{31} a_1 + S_{21} a_2 \quad \text{and} \quad b_4 = S_{43} b_3. \tag{6.190}$$

Node b_3 is called a mixed node (being both input and output) and can be eliminated, so

$$b_4 = S_{43} S_{31} a_1 + S_{43} S_{21} a_2, \tag{6.191}$$

which has the SFG of Figure 6-28(b). In Figure 6-28(b) the node representing the variable b_3 has been eliminated. Thus elimination of a node from an SFG
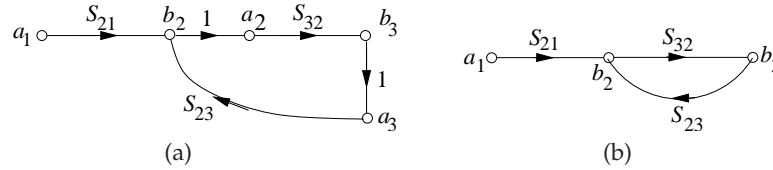
Figure 6-29 Signal flow graphs having self loops.

corresponds to the elimination of a variable in the simultaneous equations which the SFG represents. It is sufficient to recognize the SFG pattern shown in Figure 6-28(a) and replace it with the SFG of Figure 6-28(b) to achieve SFG reduction, and there is no need to write down the equation specifically to eliminate the variable.

Self-loop

Recognizing the self-loop and applying the SFG technique for eliminating it is the best example yet of identifying patterns and direct application of SFG reduction strategies. Consider the SFG of Figure 6-29(a), which has a self-loop. The equations for this graph are

$$b_3 = S_{32}a_2 \tag{6.192}$$
$$b_2 = S_{21}a_1 + S_{23}b_3 \tag{6.193}$$
$$a_2 = b_2 \tag{6.194}$$
$$a_3 = b_3. \tag{6.195}$$

Thus

$$b_3 = S_{32}S_{21}a_1 + S_{32}S_{23}b_3, \tag{6.196}$$

where the variable b_2 has been eliminated. The graph of Equation (6.196) is shown in Figure 6-30(a). The loop attached to node b_3 is called a self-loop. Such loops are not particularly convenient and can be removed by writing Equation (6.196) in the form

$$b_3 = \frac{S_{21}S_{32}}{1 - S_{23}S_{32}}a_1 \ , \tag{6.197}$$

and the SFG for this equation is shown in Figure 6-30(b). Thus the SFGs in Figures 6-29 and 6-30 are equivalent. The rule for removing self-loops follows from the manner in which Figure 6-30(a) was transformed into Figure 6-30(b). As an example, Figure 6-31(a) becomes Figure 6-31(b) and this can be reduced to the SFG of Figure 6-30(b).

Example: Multiple Self-Loops

As a further example, consider the more difficult multiple-loop case shown in Figure 6-32(a). This may be redrawn as Figure 6-32(b). The two self-loops may be directly added and then the graph reduces to Figure 6-32(c).
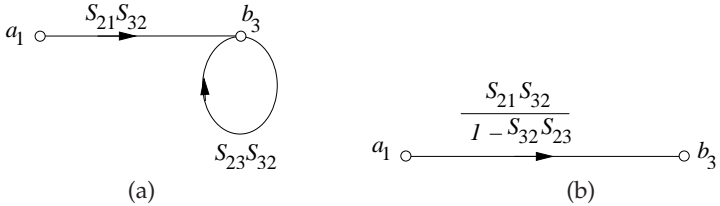
Figure 6-30 Signal flow graph of (a) a self loop; and (b) with the loop eliminated.
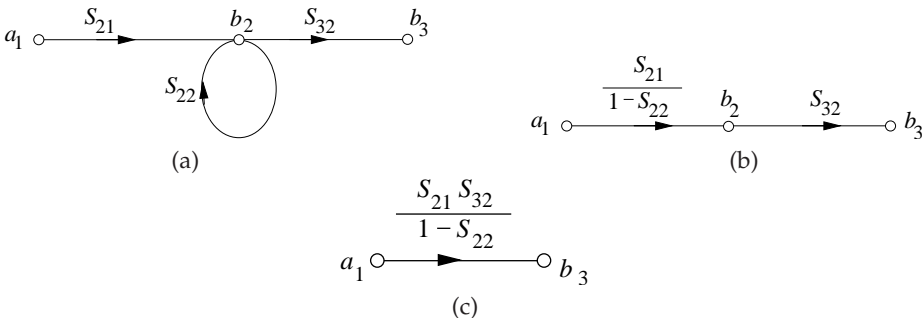
Figure 6-31 Signal flow graph with (a) an embedded loop; (b) the loop eliminated; and (c) variable elimination.
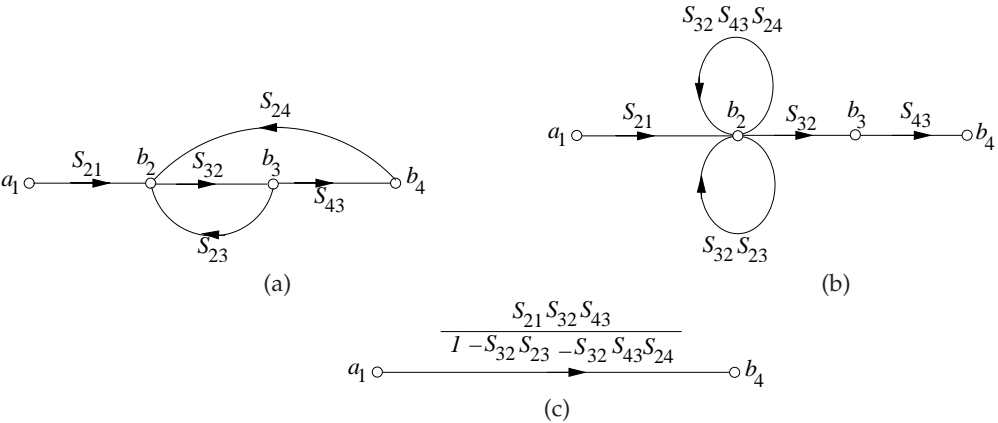
Figure 6-32 Signal flow graph with multiple loops.
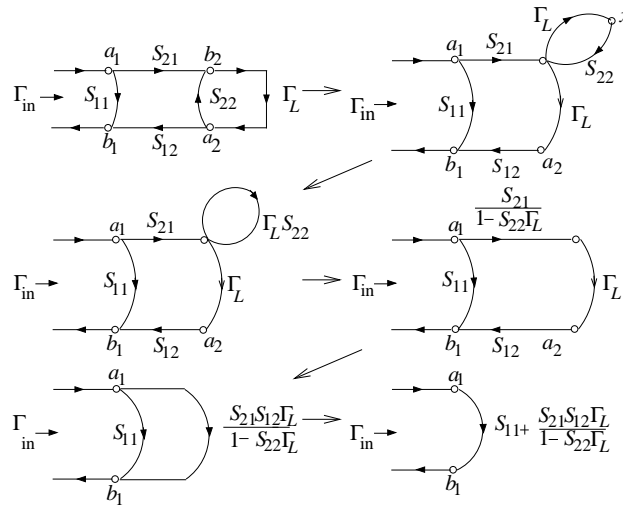
Figure 6-33 Sequence of graphical manipulations in Example 6.2 reducing a terminated two-port to a reflection only.

EXAMPLE 6. 2 　　Signal Flow Graph

Draw the SFG of a two-port with a load (at Port 2) having a voltage reflection coefficient of Γ_L, and at Port 1 a source reflection coefficient of Γ_S. Keep the S parameters of the two-port in symbolic form (e.g. , S_{11}, S_{12}, S_{21}, and S_{22}). Using SFG analysis, derive an expression for the input reflection coefficient looking into the two-port at Port 1.

Solution:

The two-port is shown below with the load attached. The input reflection coefficient is $\Gamma_{\text{in}} = b_1/a_1$ and the properties of the source here has no impact. It is additional information that is not required.

The sequence of SFG manipulations is shown in Figure 6-33 beginning with the SFG in the top left-hand corner. So the input reflection coefficient is

$$\Gamma_{\text{in}} = \frac{b_1}{a_1} = S_{11} + \frac{S_{21}S_{12}\Gamma_L}{1 - S_{22}\Gamma_L} . \tag{6.198}$$

6.7.4 Mason's Rule

Mason's rule is a general procedure for reducing SFGs with multiple loops and has become an essential part of a simple and systematic procedure for the reduction of SFGs to a single branch. Mason [85] derived the formula with the aim of developing a general method for the computer analysis of electrical circuits. First a number of topology definitions must be introduced.

A path begins at one node and traverses a number of successive edges in the direction of the arrows to arrive at the final node or sink. In SFG analysis, a path is a forward path, as the successive edges are traversed in the direction of the arrows. An open path encounters the same node only once, and a closed path (or a loop) ends on the same node at which it is started. The product of the transmittances of all edges in a loop is called the loop transmittance. Two paths, open or closed, are said to be nontouching paths if they have no nodes in common. Similarly, disjoint loops are loops that have no nodes in common.

Mason's rule:

> If T represents the overall graph transfer function and T_k represents the transfer function of the kth forward path from source to sink, then
>
> $$T = \frac{1}{\lambda} \sum_k T_k \lambda_k, \qquad (6.199)$$
>
> where λ is the determinant of the matrix of coefficients of the equations represented by the SFG (usually called the determinant of the graph) and λ_k is the determinant of that part of the graph (subgraph) which does not touch the kth forward path. The determinant is given by
>
> $$\lambda = 1 - \sum_j P_{j1} + \sum_j P_{j2} - \sum_j P_{j3} + \dots, \qquad (6.200)$$
>
> where the first summation is the sum of the loop transfer functions of all the loops in the graph. In the second summation, the products of the transfer functions of all pairs of nontouching loops are added. In the third summation, the products of the transfer functions of disjoint loops taken three at a time are added, and so on.

EXAMPLE 6. 3   Application of Mason's Rule

Use Mason's rule to reduce the SFG shown in Figure 6-34(a). Here input node a_1 is the excitation and the output node b_4 is the response. The transmittance T is equal to b_4/a_1.

Solution:

First consider the determinant of the flow graph. Since all the loops have at least one node in common, the expression for the determinant reduces to

$$\lambda = 1 - \sum_j P_{j1}. \tag{6.201}$$

There are three loops identified in Figure 6-34(b) with transmittances $\mathsf{A} = S_{21}S_{32}S_{13}$, $\mathsf{B} = S_{32}S_{43}S_{24}$, $\mathsf{C} = S_{21}S_{32}S_{43}S_{14}$. Thus the sum of the loop transmittances is given by

$$\sum_j P_{j1} = S_{21}S_{32}S_{13} + S_{32}S_{43}S_{24} + S_{21}S_{32}S_{43}S_{14} \tag{6.202}$$

and the determinant is given by

$$\lambda = 1 - S_{21}S_{32}S_{13} - S_{32}S_{43}S_{24} - S_{21}S_{32}S_{43}S_{14}. \tag{6.203}$$

This rule works well with the human ability to recognize patterns, in this case loops. There is only one forward path, identified in Figure 6-34(b), from source, a_1, to sink, b_4, given by

$$T_1 = S_{21}S_{32}S_{43}. \tag{6.204}$$

Since all the loops of the graph have nodes that touch T_1, then $\lambda_1 = 1$ and

$$\sum_k T_k\lambda_k = T_1\lambda_1 = S_{21}S_{32}S_{43}. \tag{6.205}$$

Thus

$$\frac{b_4}{a_1} = T = \frac{S_{21}S_{32}S_{43}}{1 - S_{21}S_{32}S_{13} - S_{32}S_{43}S_{24} - S_{21}S_{32}S_{43}S_{14}}. \tag{6.206}$$

This can be represented by a single edge from a_1 to b_4 as shown in Figure 6-34(c).

6.8   Polar Representations of Scattering Parameters

A polar plot is a natural way to present S parameters graphically. For example, in Section 6.6.2 on Page 315 it was seen that changing the reference planes of scattering parameters has the effect of rotating the phase of the S parameter while the amplitude stays constant. In this section, the representation of S parameters on an annotated polar plot is discussed. The most complete annotated form is referred to as a Smith chart, introduced in the late 1930s and refined in the early 1940s [86–89], and is the most important graphical tool in RF and microwave engineering.
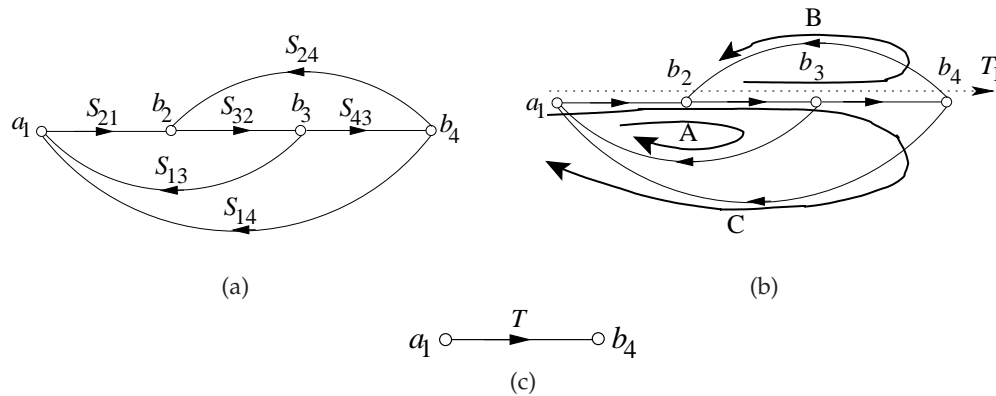
Figure 6-34 Signal flow graph used in Example 6.3 applying Mason's rule: (a) SFG; (b) annotation identifying loops and through-path; and (c) final reduction.

6.8.1 Polar Plot for Reflection Coefficient

Figure 6-35 is used in plotting reflection coefficients and is a polar plot that has a radius of one. So a reflection coefficient with a magnitude of one is on the unit circle. The center of the polar plot is zero so the reflection coefficient of a matched load, which is zero, is plotted at the center of the circle. Plotting a reflection coefficient on the polar plot enables convenient interpretation of the properties of a reflection. The graph has additional notation that enables easy plotting of an S parameter on the graph. Conversely, the magnitude and phase of an S parameter can be easily read from the graph. The horizontal label going from 0 to 1 is used with the magnitude. The three sets of notations arranged on the outer perimeter of the polar plot are used to read off angle information. All three relate to electrical length. One scale is in terms of angle and the other two are in terms of wavelength. You will also notice the additional notation "ANGLE OF REFLECTION COEFFICIENT IN DEGREES" and the scale relates to the polar plot. The $90°$ point is just where one would expect it to be.

The other two scales around the circumference, "WAVELENGTHS TOWARD LOAD" and "WAVELENGTHS TOWARD GENERATOR," are for the electrical length of a transmission line and can be used to rotate the phase of the reflection and transmission coefficients by the electrical length of the line. Care is required to move in the correct direction (toward or away from the load). The notation around the circumference is based on the concept of a reflection coefficient looking into a transmission line terminated in a load. The angular change of the reflection coefficient is twice that of the electrical length of the line. (A generalization for S parameters was given in Section 6.6.2 on Page 315.)
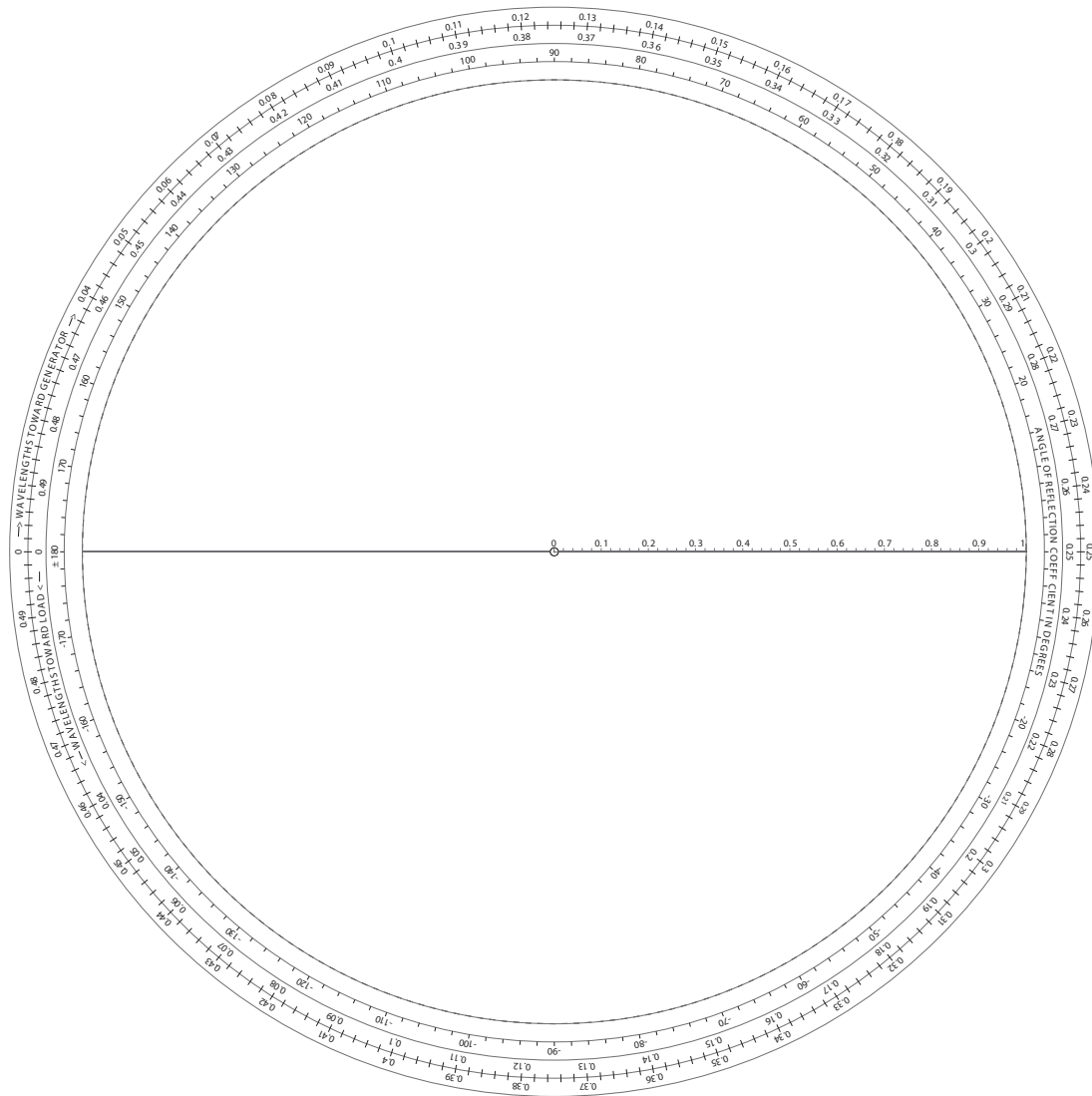
Figure 6-35 Polar chart for plotting reflection coefficient.

The nomograph shown in Figure 6-36 has been developed to aid in interpretation of polar reflection coefficient plots. The nomograph relates the reflection coefficient (RFL. COEFF), ρ; the square of the reflection coefficient, ρ^2; the return loss (RTN. LOSS) (in decibels); the standing wave ratio (SWR); and the standing wave ratio (in decibels) as $20 \log(\text{SWR})$. When printed together with the reflection coefficient polar plot (Figures 6-35 and 6-36
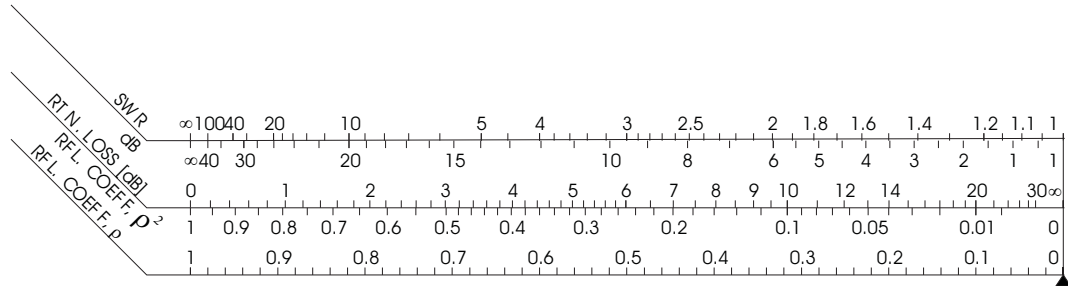
Figure 6-36   Nomograph relating the reflection coefficient (RFL. COEFF), ρ; the square of the reflection coefficient, ρ^2; the return loss (RTN. LOSS) (in decibels); the standing wave ratio (SWR); and the standing wave ratio (in decibels) as $20 \; \log(\text{SWR})$.

combined) the nomograph is scaled properly, but it is expanded here so that it can be read more easily. So with the aid of a compass centered at one point on the zero point of the polar plot and the other at the reflection coefficient, the magnitude of the reflection coefficient is captured. The compass can then be brought down to the nomograph to read ρ, ρ^2, the return loss, and VSWR directly.

6.8.2   Polar Plot for S Parameters

Figure 6-37 is similar to the reflection coefficient plot of Figure 6-35, but now there is an additional angle scale for the transmission coefficient. In particular there are two notations: "ANGLE OF REFLECTION COEFFICIENT IN DEGREES" and "ANGLE OF TRANSMISSION COEFFICIENT IN DE-GREES." There is a factor of two in the ratio of numbers for the two scales. Recall that an open-circuited line with an electrical length of ϕ degrees has an input reflection coefficient of 2ϕ, as the traveling wave must go down the line a distance of ϕ degrees then reflect (with a reflection coefficient of $+1$) and travel back a distance of ϕ degrees. Hence the reflection coefficient has an angle rotation that is twice the angle rotation of the traveling wave. The two angle scales on the polar plot differ by a factor of two in an attempt to capture this effect. Insertion of a line of length ϕ results in a similar rotation of the reflection and transmission coefficients on the polar plot.

The nomograph shown in Figure 6-38 is similar to the reflection nomograph developed for the reflection coefficient. Now the quantities refer to transmission properties. Just as before, a compass reading of a transmission parameter (say, S_{21}) can be directly interpreted a number of ways.
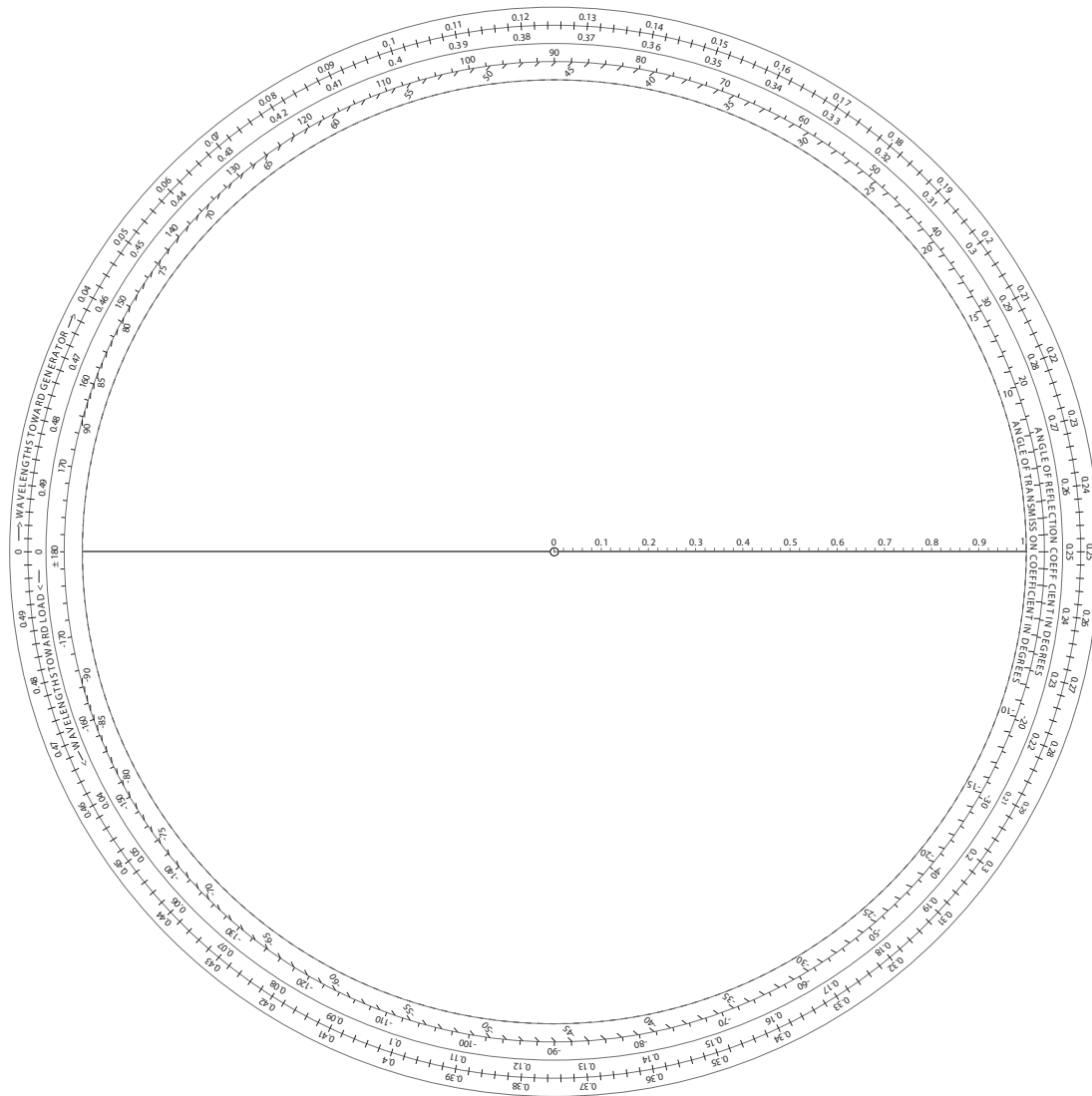
Figure 6-37   Polar plot for plotting reflection and transmission coefficient.

6.9   Smith Chart

The Smith chart[7] is also a polar plot of S parameters, but with the addition of curves relating impedance or admittance to reflection coefficient [86]. The

---

[7]   The Smith chart was invented by Phillip Smith (born April 29, 1905, died August 29, 1987) and presented in close to its current form in 1937. It was described by Smith in several publications [86–89]. Up until the 1970s nomographs and graphical calculators were popular
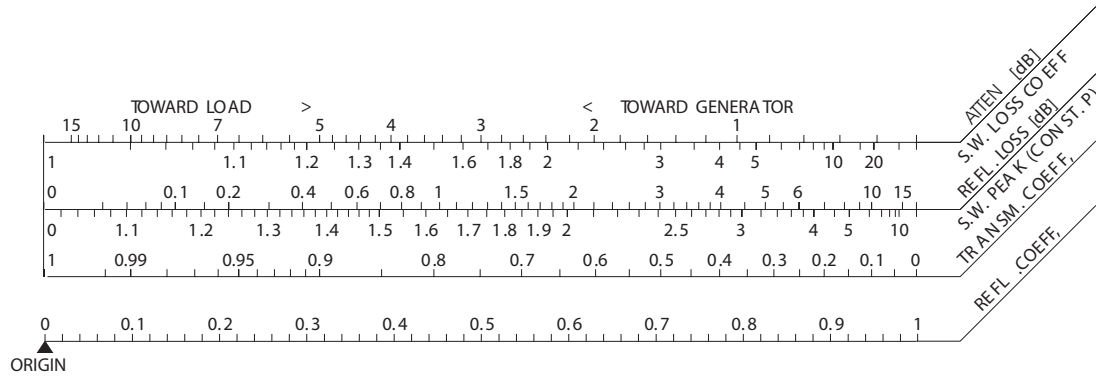
Figure 6-38 Nomograph to be used with transmission parameters relating the magnitude of the transmission line coefficient (TRANSM. COEFF), T); the attenuation (ATEN. [dB]); the standing wave loss coefficient (S.W. LOSS COEFF) to the reflection coefficient (REFL. COEFF, ρ); the reflection loss (REFL LOSS [dB]); and the standing wave peak (S.W. PEAK CONST, ρ).

reflection coefficient, Γ, is related to a load, Z_L, by

$$\Gamma = \frac{Z_L - Z_0}{Z_L + Z_0}, \tag{6.207}$$

where Z_0 is the system reference impedance. With normalized load impedance $z_l = r + \jmath x = Z_L/Z_0$, this becomes

$$\Gamma = \frac{r + \jmath x - 1}{r + \jmath x + 1}. \tag{6.208}$$

Commonly in network design reactive elements are added either in shunt or in series to an existing network. If a reactive element is added in series then the input reactance, x, is changed while the input resistance, r, is constant. So superimposing the loci of Γ (on the S parameter polar plot) with fixed values of r, but varying values of x (x varying from $-\infty$ to ∞), proves useful, as will be seen. Also, plotting the loci of Γ (with fixed values of x and varying values of r (r varying from 0 to ∞) is also useful. The combination of the reflection/transmission polar plots, the nomographs, and the r and x loci is called the impedance Smith chart (see Figure 6-39).

---

engineering tools because there were limited computing tools. Only a few have survived in electrical engineering usage, with Smith charts being overwhelmingly the most important. Smith was active in the IRE which became the IEEE in 1947. In 1952 he was elected IEEE Fellow "for his contributions to the development of antennas and graphical analysis of transmission line characteristics." He was secretary-treasurer of the IEEE Antennas and Propagation Society in 1954. Many resources for the Smith charts are available online with a search on the term Smith chart.
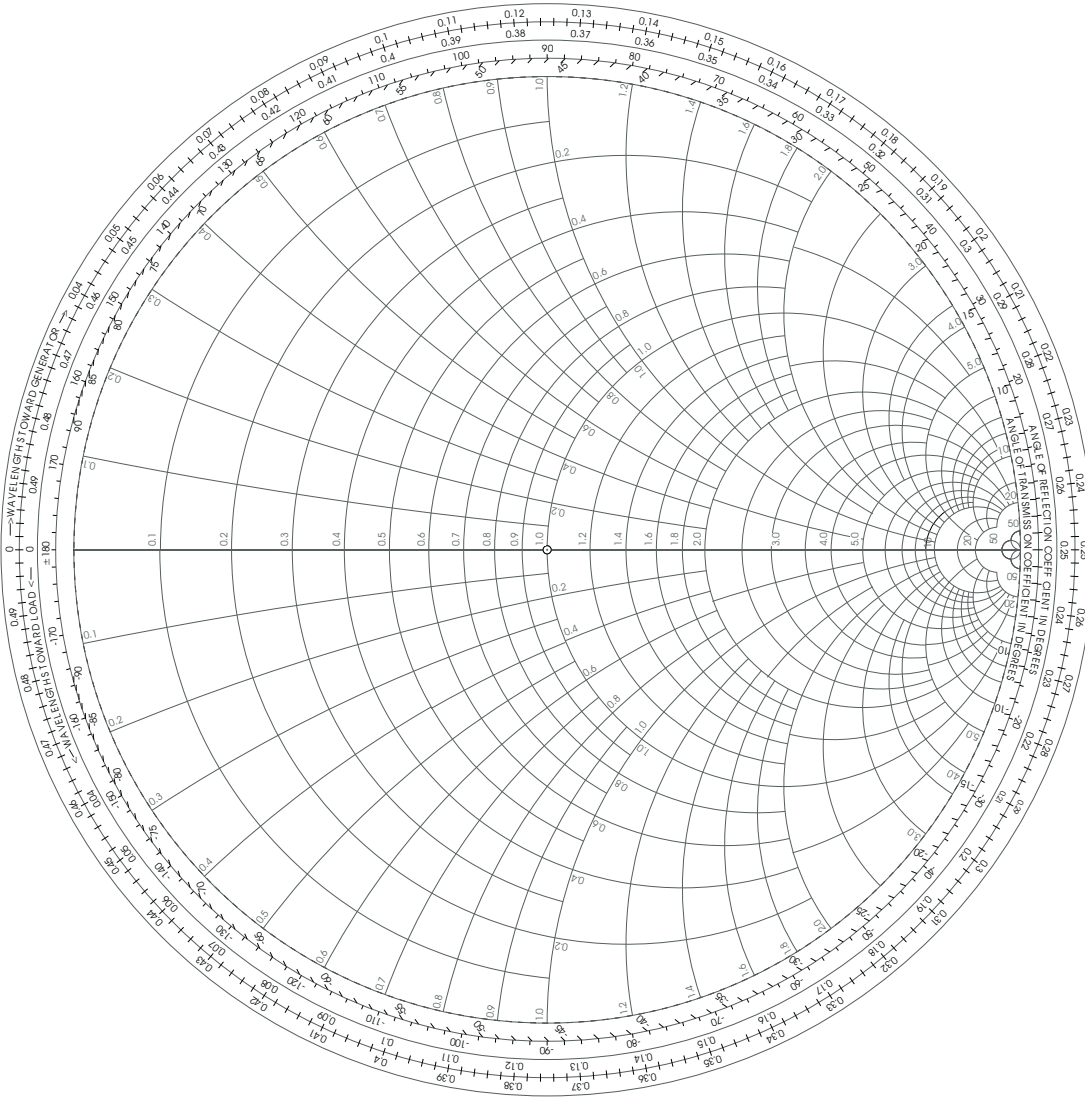
Figure 6-39 Impedance Smith chart.

Repeating this concept, but now for admittance parameters so that the loci superimposed on the Smith chart is for constant conductance and susceptance ranging from $-\infty$ to ∞, and for constant susceptance and conductance ranging from 0 to ∞, yields the admittance Smith chart(see Figure 6-40). The combination of the reflection/transmission polar plots, the nomographs, and the impedance and admittance Smith charts leads to the
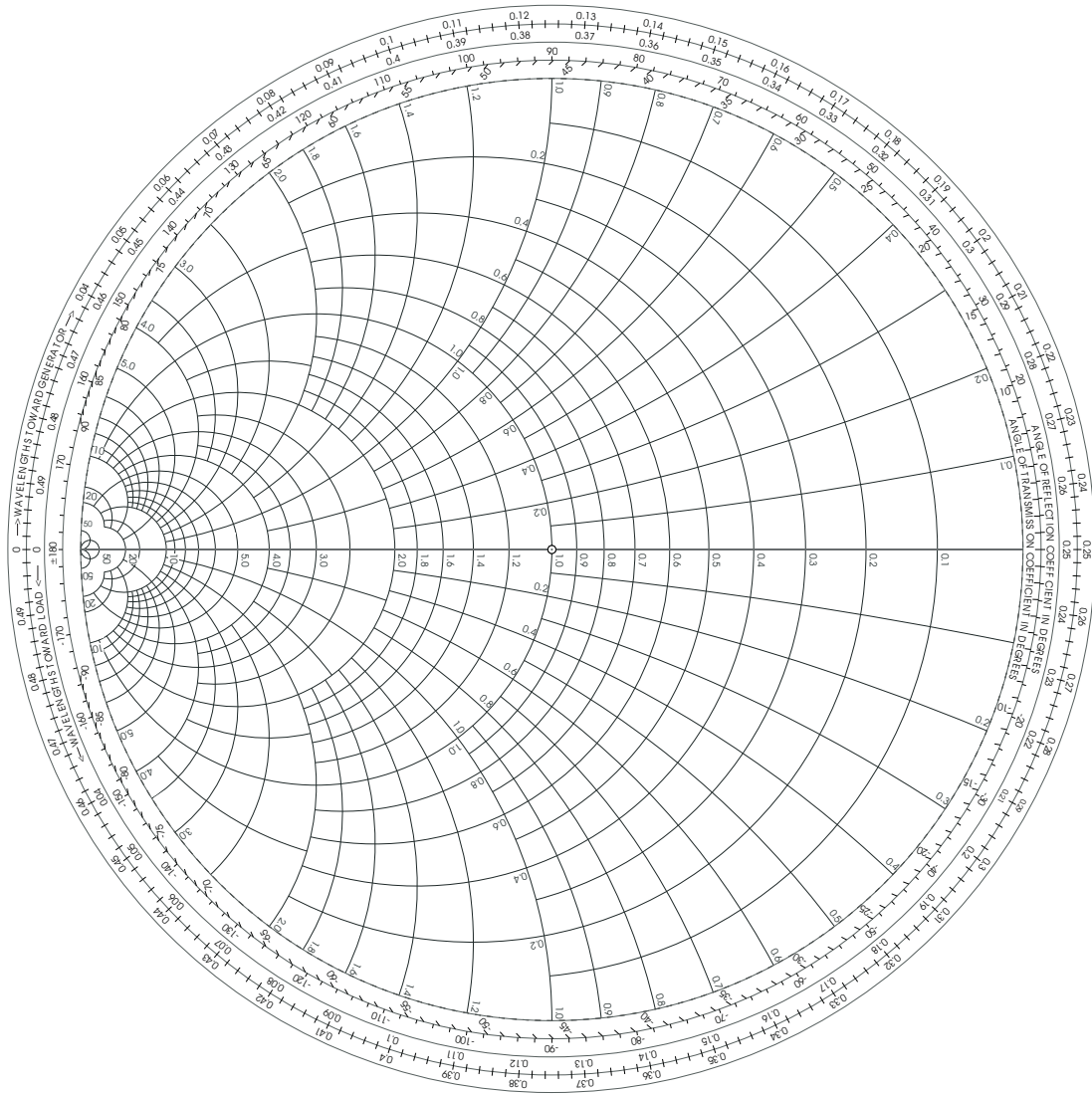
Figure 6-40 Admittance Smith chart.

combined Smith chart (see Figure 6-41).

In summary, forms of Smith charts used in examples presented in this book are shown in Figures 6-42 and 6-43. Note that there are 50 Ω and normalized versions of the Smith chart. The normalized versions, shown here, prove more useful. A combined admittance and impedance Smith chart is shown in Figure 6-44.
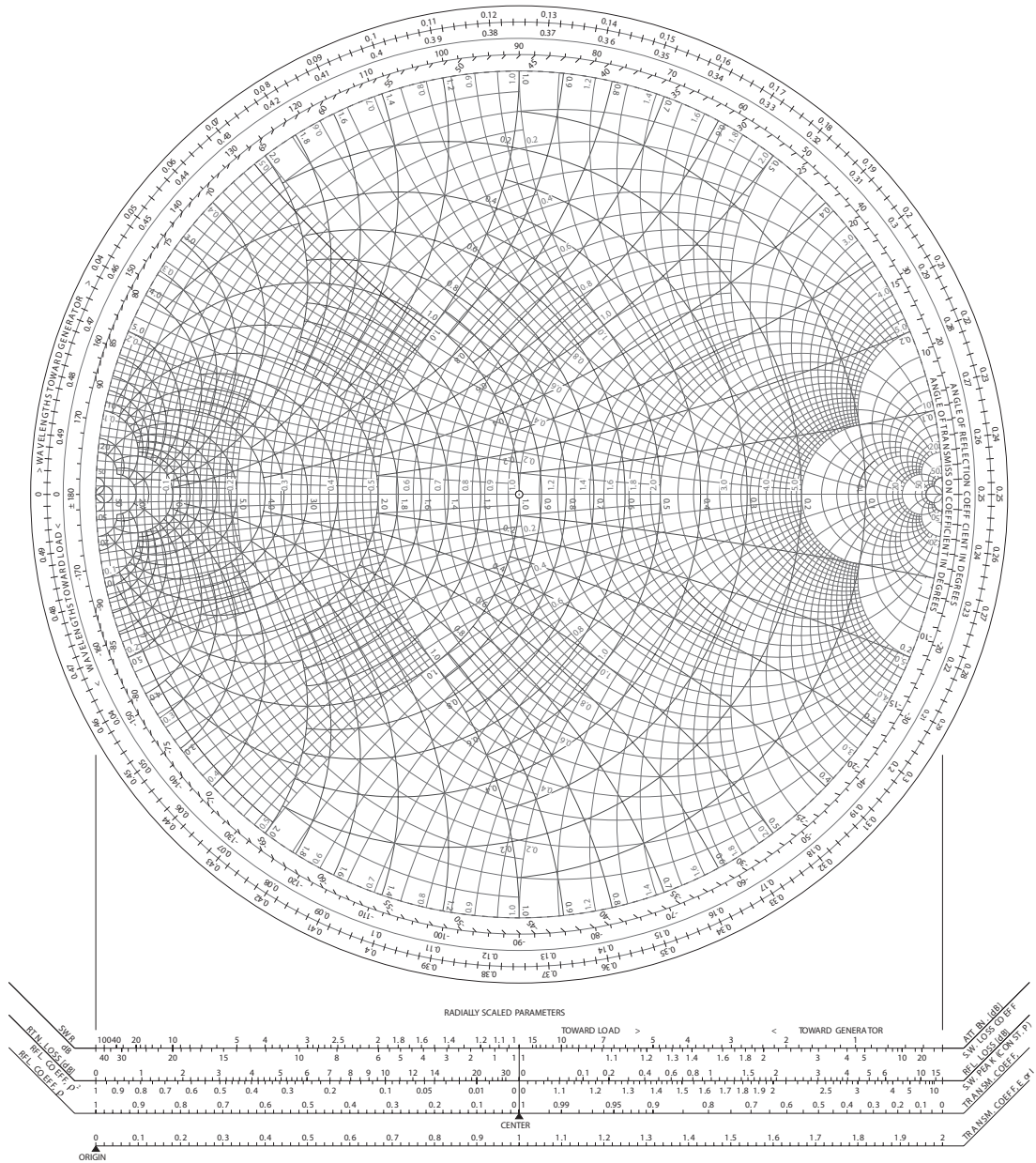
Figure 6-41 Combined impedance and admittance Smith chart.
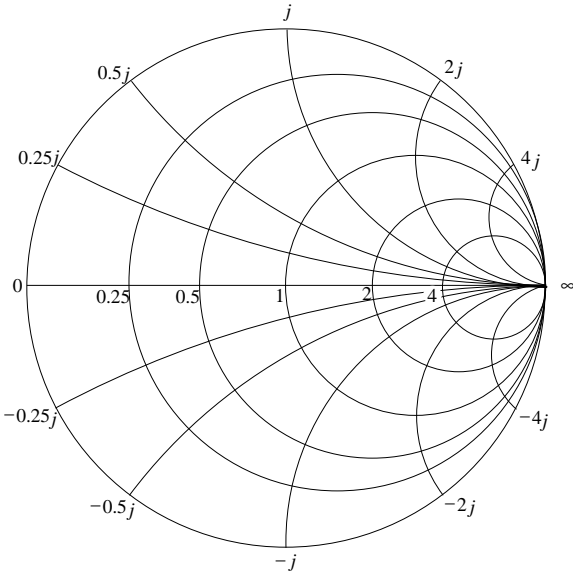
Figure 6-42   Normalized impedance Smith chart.
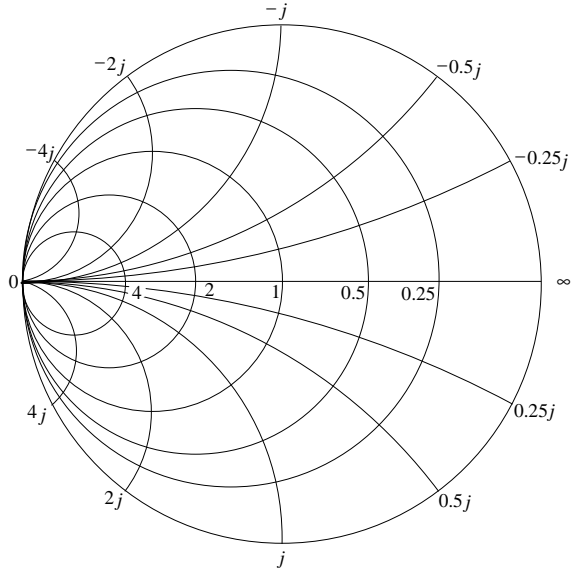
Figure 6-43   Normalized admittance Smith chart.
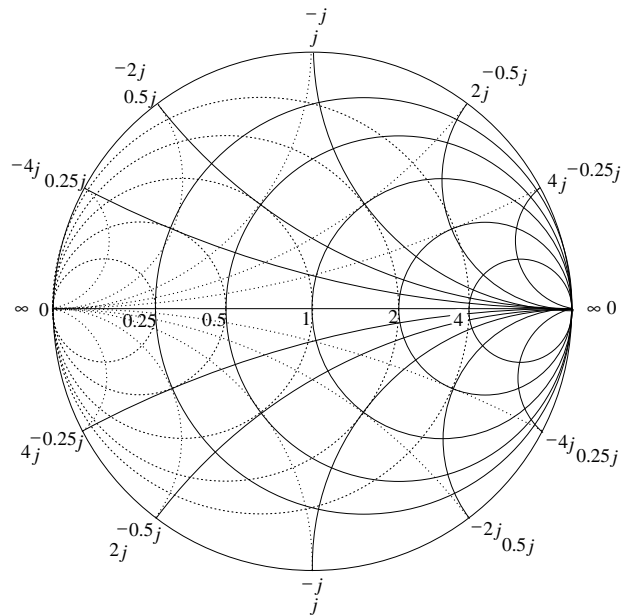
Figure 6-44   Combined impedance and admittance Smith chart.

EXAMPLE 6. 4   Impedance to Admittance Conversion

Use a Smith chart to convert the impedance $z = 1 - 2\jmath$ to an admittance.

Solution:

The impedance $z = 1 - 2\jmath$ is plotted as Point A in Figure 6-45. To read the admittance off the lines of constant conductance and constant susceptance must be interpolated from the arcs and circles provided. The interpolations are shown in the figure, indicating a conductance of 0.2 and a susceptance of $0.4\jmath$. Thus

$$y = 0.2 + 0.4\jmath.$$

(This agrees with the $y = 1/z = 1/(1 - 2\jmath) = 0.2 + 0.4\jmath$ calculation.)

Smith charts are indispensable tools for RF and microwave engineers. Even with the ready availability of CAD programs, Smith charts are generally preferred for portraying measured and calculated data because of the easy interpretation of S parameters. With experience, the properties of circuits can be inferred. Also, with experience it is an invaluable design tool enabling the designer to see the beginning and endpoints of a design and infer the type of circuitry required to go from the start to the end.
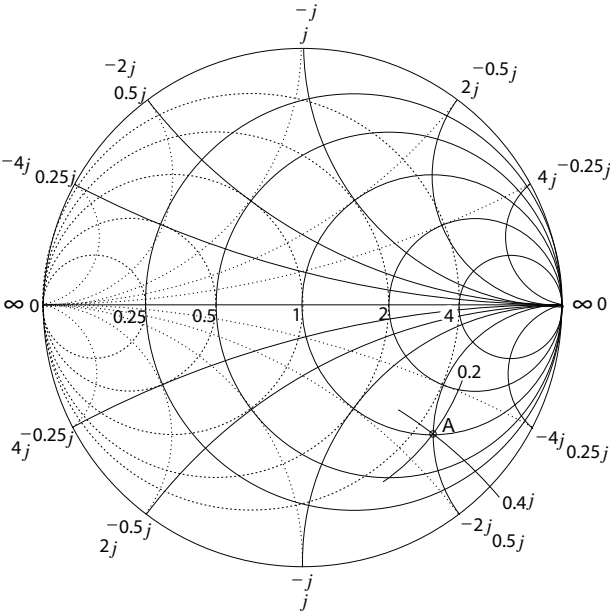
Figure 6-45 Combined impedance and admittance Smith chart.

6.9.1 An Alternative Admittance Chart

In design it is necessary to switch between the admittance and impedance Smith charts. So it is convenient to use a multicolored combined Smith chart, as shown in Figure 6-41. It is also possible to annotate an impedance Smith chart so it can be used to perform admittance-based calculations. Such a chart is shown in Figure 6-46. At first this seems overly detailed. The key is that the same chart can be used for both impedance and admittance calculations. When the impedance form is used, this chart is a polar plot of an S parameter. If the admittance form is used, the reflection coefficient is rotated $180°$.

6.9.2 Transmission Line Stubs and Smith Charts

Transmission lines and Smith charts are closely related, as a length of transmission line results in phase rotation of scattering parameters if the parameters are normalized to the characteristic impedance of the line. Short- and open-circuited transmission lines are called stubs and are commonly used to realize impedances at microwave frequencies. Circuit symbols are used to indicate stubs, as shown in Figure 6-47, with the input impedance of the stub as written. These symbols are used in circuit diagrams.
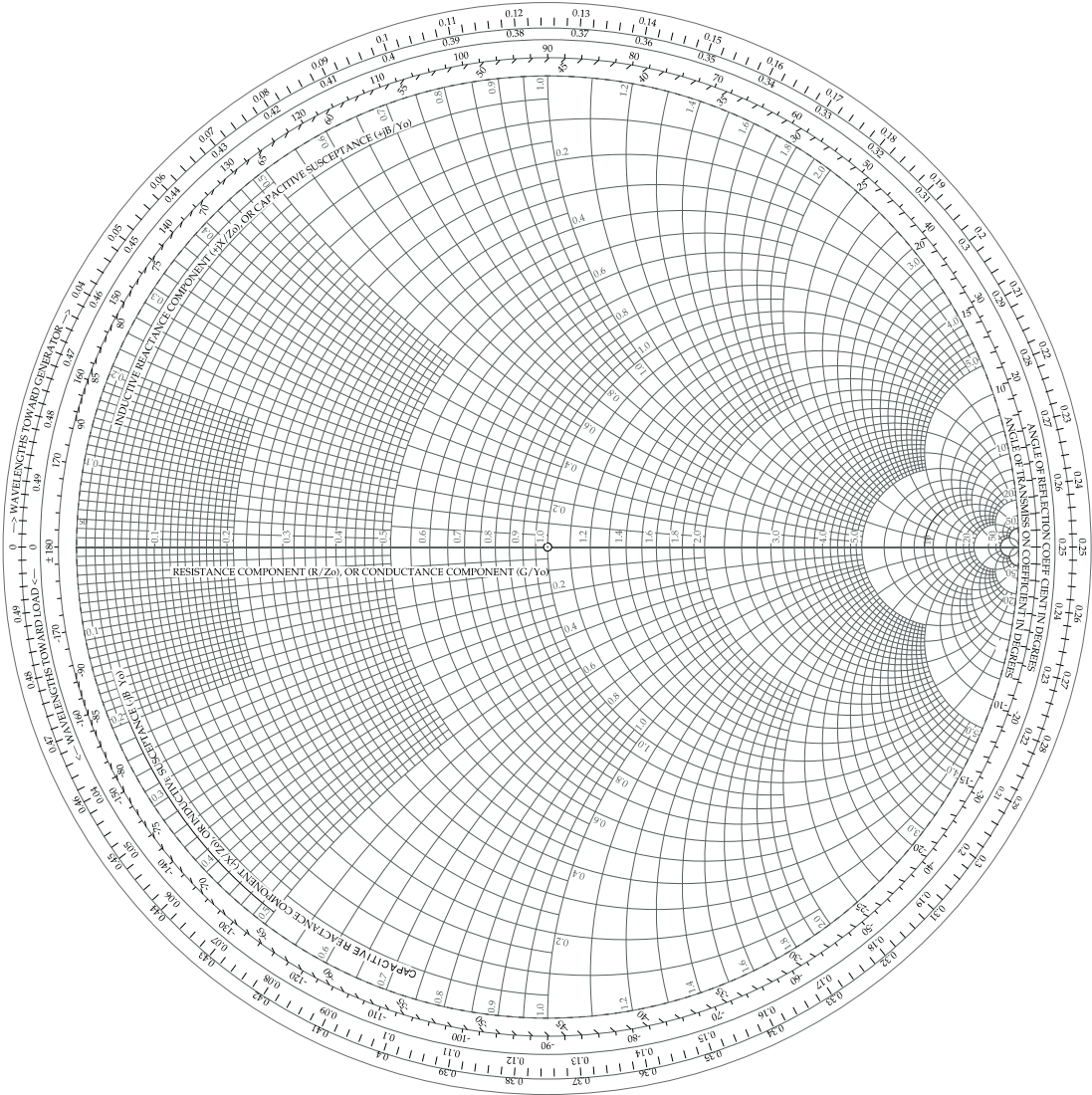
Figure 6-46   Black-magic "poor mans" admittance Smith chart.

Figure 6-47   Circuit symbols for short-circuited and open-circuited stubs.

EXAMPLE 6. 5   Impedance Synthesis

Use a length of a terminated transmission line to realize an impedance of $140 \jmath \ \Omega$.

Solution:

The impedance to be synthesized is reactive so the termination must also be lossless. The simplest termination is with either a short circuit or an open circuit. Both cases will be considered. Choose a transmission line with a characteristic impedance, Z_0, of $100\ \Omega$ so that the desired normalized input impedance is $140\ \Omega / Z_0 = 1.4 \jmath$.

First, the short-circuit case. In Figure 6-48, consider the path AB. The termination is a short circuit and the impedance of this load is Point A with a length of $\ell_A = 0\ \lambda$. The corresponding reflection coefficient angle from the scale is $\theta_A = 180°$. As the line length increases, the input impedance of the terminated line follows the clockwise path to Point B where the normalized input impedance is $\jmath 1.4$. At B, the line length, ℓ_B, is $0.1515\ \lambda$ and the corresponding reflection coefficient angle from the scale, θ_B, is $71.2°$. The reflection coefficient angle and length in terms of wavelengths were read directly off the Smith chart and care needs to be taken that the right sign and correct scale are used. For example, there are two angle scales, one corresponding to the reflection coefficient and the other to the transmission coefficient. A good strategy is to correlate the scales with the easily remembered properties at the open-circuit and short-circuit points. Here the line length is

$$\ell = \ell_B - \ell_A = 0.1515\lambda - 0\lambda = 0.1515\lambda, \tag{6.209}$$

and the electrical length is half of the difference in the reflection coefficient angle,

$$\theta = \frac{1}{2}\left|\theta_B - \theta_A\right| = \frac{1}{2}\left|71.2° - 180°\right| = 54.4°, \tag{6.210}$$

corresponding to a length of $(54.4°/360°)\lambda = 0.1511\lambda$. This is as close as could be expected from using scales. So the length of the stub with a short-circuit termination is 0.1515λ.

For the open-circuited stub, the path begins at impedance infinity point $\Gamma = +1$ and rotates clockwise to Point A (this is a $90°$ or 0.25λ rotation) before continuing on to Point B. So for the open-circuited stub,

$$\ell = 0.1515\lambda + 0.25\lambda = 0.4015\lambda. \tag{6.211}$$
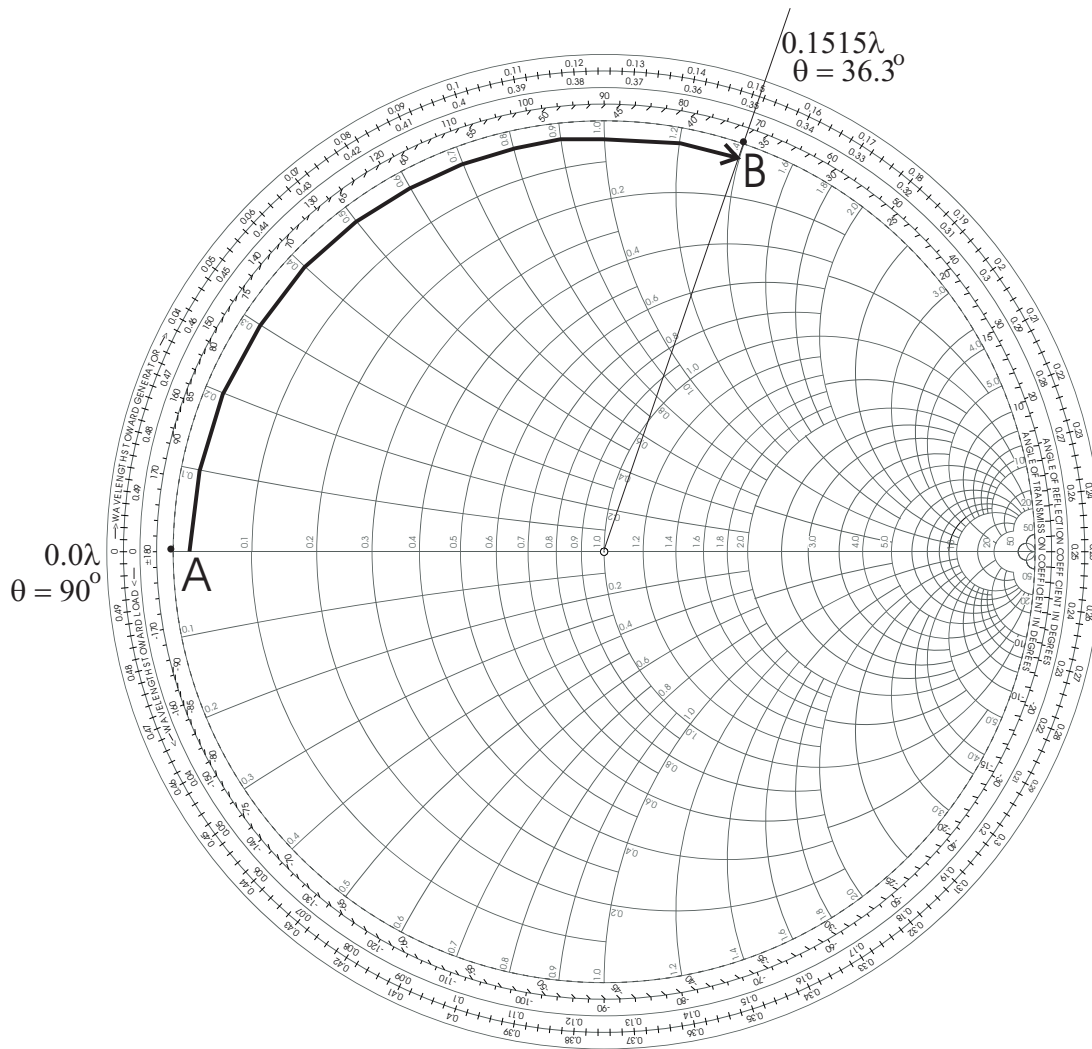
Figure 6-48 Design of a short-circuit stub with a normalized input impedance of $j1.4$.

6.10 Reflection Coefficient and Change of Reference Impedance

6.10.1 Introduction

In this section the input reflection coefficient of a terminated transmission line of characteristic impedance Z_{02} is referenced to a system impedance Z_{01}. The development is based on the properties of the bilinear transform.
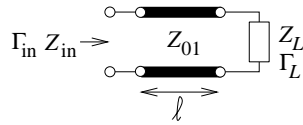
Figure 6-49 Transmission line of characteristic impedance Z_{01} terminated in a load with a reflection coefficient Γ_L.

The generalized bilinear transform of two complex numbers z and w is

$$w = \frac{Az + B}{Cz + D}, \qquad (6.212)$$

where A, B, C, and D are constants[8] which may themselves be complex. This is of interest in dealing with reflection coefficients where w and z are reflection coefficients and A, B, C, and D describe the network between a load with a complex reflection coefficient, z, and w is the reflection coefficient looking into the network. The special property of the bilinear transform is that a circle in the complex plane (here the locus of z) is mapped onto another circle (here the locus of w) in the complex plane. This result is presented in Section A.8 on Page 832.

6.10.2 Reference Impedance Change as a Bilinear Transform

Microwave and RF engineers obtain considerable design and measurement insight by plotting a reflection coefficient on the complex plane, often using a polar plot. In Figure 6-49 a fixed load terminates a transmission line of characteristic impedance Z_{01} that has the input reflection coefficient, Γ_{in}. As will be shown, the locus of Γ_{in}, normalized to any system impedance, is a circle on the complex plane as the electrical length of the line increases. The electrical length of the line increases as the frequency increases (with the physical length of the line held constant), or as the physical length of the line increases (with the frequency held constant).

In Figure 6-49 a transmission line with characteristic impedance Z_{01} is terminated in a load with reflection coefficient Γ_L. From Equation (4.90) the reflection coefficient of a load Z_L referenced to Z_{01} is

$$\Gamma_{L,Z01} = \frac{Z_L - Z_{01}}{Z_L + Z_{01}} \qquad (6.213)$$

and the input reflection coefficient referenced to Z_{01} is

$$\Gamma_{\text{in},Z01} = \Gamma_{L,Z01} e^{-j2\theta}, \qquad (6.214)$$

---

[8]   These are not the cascadable $ABCD$ parameters, but simply the coefficients commonly used with bilinear transforms.

where $\theta = \beta\ell$ is the electrical length of the line. As ℓ increases from zero, $\Gamma_{\mathrm{in},Z01}$ plotted on a polar chart traces out a circle. The important result that will be developed in this section is that when the input reflection coefficient is referenced to another impedance, the new reflection coefficient will also map out a circle. The development of $\Gamma_{\mathrm{in},Z02}$ (the input reflection coefficient referred to Z_{02}) begins by calculating the input impedance of the line (in Figure 6-49):

$$Z_{\mathrm{in}} = Z_{01}\frac{1 + \Gamma_{\mathrm{in},Z01}}{1 - \Gamma_{\mathrm{in},Z01}} \qquad (6.215)$$

so that the reflection coefficient referenced to Z_{02} is

$$\Gamma_{\mathrm{in},Z02} = \frac{Z_{\mathrm{in}} - Z_{02}}{Z_{\mathrm{in}} + Z_{02}} = \frac{(Z_{01} - Z_{02}) + (Z_{01} + Z_{02})\Gamma_{\mathrm{in},Z01}}{(Z_{01} + Z_{02}) + (Z_{01} - Z_{02})\Gamma_{\mathrm{in},Z01}}, \qquad (6.216)$$

and rearranging

$$\begin{aligned}\Gamma_{\mathrm{in},Z02} &= \frac{Z_{\mathrm{in}} - Z_{02}}{Z_{\mathrm{in}} + Z_{02}} = \frac{(Z_{01} + Z_{02})\Gamma_{\mathrm{in},Z01} + (Z_{01} - Z_{02})}{(Z_{01} - Z_{02})\Gamma_{\mathrm{in},Z01} + (Z_{01} + Z_{02})} \\ &= \frac{A\Gamma_{\mathrm{in},Z01} + B}{C\Gamma_{\mathrm{in},Z01} + 1}.\end{aligned} \qquad (6.217)$$

This mapping has the form of the general bilinear transform (Equation (6.212)) where

$$A = 1, \quad B = \frac{Z_{01} - Z_{02}}{Z_{01} + Z_{02}} = C. \qquad (6.218)$$

Since Equation (6.217) is a general bilinear transform, if the locus of $\Gamma_{\mathrm{in},Z01}$ is a circle then the locus of $\Gamma_{\mathrm{in},Z02}$ is also a circle.

EXAMPLE 6. 6　　　Reflection Coefficient, Reference Impedance Change

In the circuit below, a 50 Ω lossless line is terminated in a 25 Ω load. Plot the locus, with respect to the length of the line, of the reflection coefficient, looking to the line referencing it first to a 50 Ω reference impedance and then a 75 Ω reference impedance.
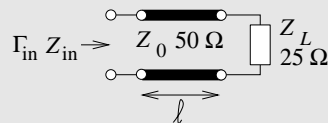
Solution:

Let the input reflection coefficient referenced to a 50 Ω system be $\Gamma_{\text{in},50}$, and when it is referenced to 75 Ω it is $\Gamma_{\text{in},75}$. In the 50 Ω system the reflection coefficient of the load is

$$\Gamma_{L,50} = \frac{Z_L - 50}{Z_L + 50} = \frac{25 - 50}{25 + 50} = -0.3333 \tag{6.219}$$

and

$$\Gamma_{\text{in},50} = \Gamma_{L,50} e^{-\jmath 2\theta}, \tag{6.220}$$

where $\theta = (\beta \ell)$ is the electrical length of the line. The locus, with respect to electrical length, of $\Gamma_{\text{in},50}$ is plotted in Figure 6-50 as the length ℓ ranges from 0 to $\lambda/8$, to $\lambda/2$, and completes the circle at $\lambda/2$. Now

$$Z_{\text{in}} = 50\frac{1 + \Gamma_{\text{in},50}}{1 - \Gamma_{\text{in},50}}, \tag{6.221}$$

so that the 75 Ω reflection coefficient is

$$\Gamma_{\text{in},75} = \frac{Z_{\text{in}} - 75}{Z_{\text{in}} + 75}. \tag{6.222}$$

The locus of this reflection coefficient is plotted on a polar plot in Figure 6-51. This locus is the plot of $\Gamma_{\text{in},75}$ as the electrical length of the line, θ, is varied from 0 to π. The center of the circle in Figure 6-51 is taken from the chart and determined to be -0.177.

A circle is defined by its center and its radius, and Appendix A.8, in particular Equations (A.108) and (A.109), enables the circles for the locus of $\Gamma_{\text{in},Z01}$ and $\Gamma_{\text{in},Z02}$ to be related. The locus of $\Gamma_{\text{in},Z01}$ has a center at the origin and a radius $|\Gamma_L|$, that is,

$$C_{Z01} = 0 \quad \text{and} \quad R_{Z01} = |\Gamma_{L,Z01}|. \tag{6.223}$$

In Equation (A.108) C_{Z01} replaces the center C_z and $|\Gamma_{L,Z01}|$ takes the place of the radius R_z. Similarly the center of the Γ_{Z02} circle, C_{Z02}, replaces C_w and the radius of the Γ_{Z02} circle, R_{Z02}, replaces R_w. So the locus of the reflection coefficient referenced to Z_{02} is described by a circle with center

$$C_{Z02} = \frac{B - A/C}{1 - |C\Gamma_{L,Z01}|^2} + \frac{A}{C} \tag{6.224}$$

with radius

$$R_{Z02} = \left| \frac{(B - A/C)\Gamma_{L,Z01}}{1 - |C\Gamma_{L,Z01}|^2} \right|. \tag{6.225}$$

Since, from Equation (6.218), $A = 1$ and $B = C$, this further simplifies to

$$C_{Z02} = \frac{B - 1/B}{1 - |B\Gamma_{L,Z01}|^2} + \frac{1}{B} \tag{6.226}$$

$$R_{Z02} = \left| \frac{(B - 1/B)\Gamma_{L,Z01}}{1 - |B\Gamma_{L,Z01}|^2} \right|. \tag{6.227}$$
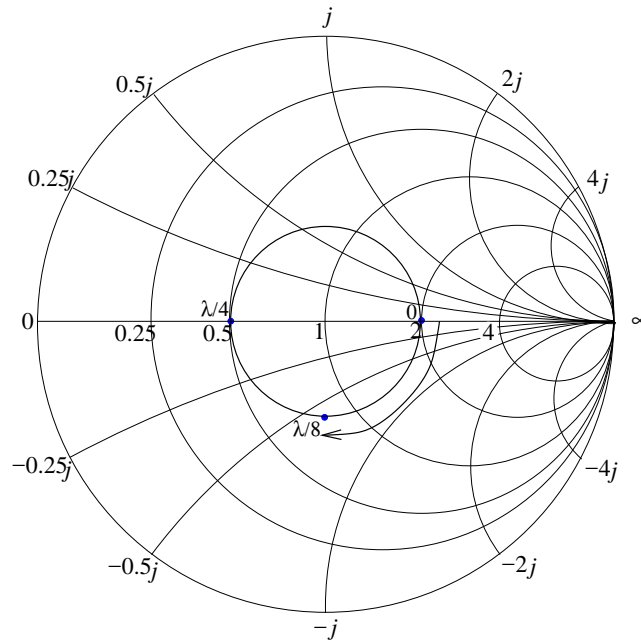
Figure 6-50 A normalized Smith chart with a $50\ \Omega$ normalization impedance with the input reflection coefficient of a $50\ \Omega$ transmission line with a load of $25\ \Omega$.

So the locus (with respect to frequency) of the input reflection coefficient of a terminated transmission line (of characteristic impedance Z_{01}) is a circle no matter what normalization impedance is used. If the normalization impedance is Z_{01}, then the center of the reflection coefficient circle will be zero. It will not be zero if the normalization impedance is other than Z_{01}. The radius of the circle will also change.

6.10.3 Determining the Characteristic Impedance of a Line from the Smith Chart

In measurements of a terminated transmission line of unknown characteristic impedance Z_{01}, the input reflection coefficient circle[9] (normalized to Z_{02}) is known, and from this Z_{01} must be determined. In design, a transmission line circle can be drawn to complete a matching problem and from this the characteristic impedance (Z_{01}) of the line must be found. In both situations C_{Z02}, R_{Z02}, and Z_{02} are known and Z_{01} must be determined. Unfortunately a simple closed form solution cannot be obtained from Equations (6.226) and (6.227). However, by substituting Equation (6.227) in Equation (6.226), and

---

[9] Here the circle is the locus with respect to frequency.
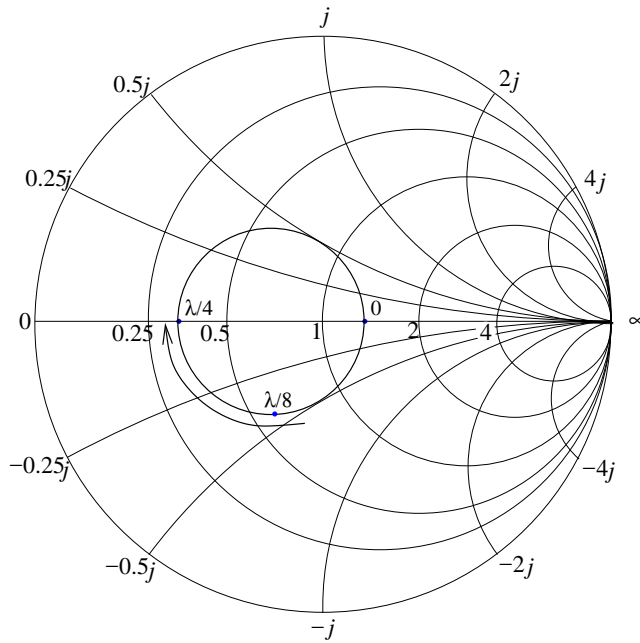
Figure 6-51 A normalized Smith chart with a 75 Ω normalization impedance with the input reflection coefficient of a 50 Ω transmission line with a load of 25 Ω.

if Z_{01} and Z_{02} are close (so that B is small), then

$$C_{Z02} \approx B - \frac{1}{B} + \frac{1}{B} = B. \tag{6.228}$$

The approximation is better for smaller $|\Gamma_{L,Z01}|$. Also

$$R_{Z02} \approx \left| \frac{\Gamma_{L,Z01}}{B} \right|. \tag{6.229}$$

So, provided that the characteristic impedances are close,

$$Z_{01} = Z_{02} \frac{1+B}{1-B} \approx Z_{02} \frac{1+C_{Z02}}{1-C_{Z02}} \approx Z_{02}(1 + 2C_{Z02}) \tag{6.230}$$

and

$$|\Gamma_{L,Z01}| \approx |BR_{Z02}|. \tag{6.231}$$

EXAMPLE 6. 7	Center of Reflection Coefficient Locus

In Example 6.6 the locus of the input reflection coefficient referenced to 75 Ω was plotted for a 50 Ω line terminated in a load with a reflection coefficient (in the 50 Ω system) of $-1/3$. Calculate the center of the input reflection coefficient of the network when it is referenced to 75 Ω.

Solution:

The center of the reflection coefficient normalized to Z_{01} is zero, that is $C_{50} = 0$. From Equation (6.226), the center of the reflection coefficient normalized to $Z_{02} = 75 \ \Omega$ is

$$C_{75} = \frac{B - 1/B}{1 - |B\Gamma_{L,50}|^2} + \frac{1}{B}, \tag{6.232}$$

where

$$B = \frac{Z_{01} - Z_{02}}{Z_{01} + Z_{02}} = \frac{50 - 75}{50 + 75} = -0.2 \ . \tag{6.233}$$

So

$$C_{75} = \frac{-0.2 + 5}{1 - (0.2/3)^2} - 5 = -0.17857 \ , \tag{6.234}$$

which compares favorably to a center of -0.177 determined manually from the polar plot in the previous example.

6.10.4 Reflection Coefficient Locus

The direction of the locus of the input reflection coefficient of a terminated transmission line is always clockwise, even if the Smith chart uses a reference or normalization impedance different from the characteristic impedance of the line. In Figure 6-52, Z_{01} is the characteristic impedance of the line and Z_{02} is the normalization impedance. When $Z_{02} \leq Z_{01}$, the locus rotates in the clockwise direction. The center of the circle of the reflection coefficient normalized to Z_{02} is to the left of the origin if $Z_{02} > Z_{01}$ and to the right of the origin if $Z_{02} < Z_{01}$.

6.11 Measurement of Scattering Parameters

S parameters are measured using a network analyzer called an **Automatic Network Analyzer** (**ANA**), or more commonly a **Vector Network Analyzer** (**VNA**). The VNA was introduced by Hackborn [90] in 1968. A schematic of a modern VNA is shown in Figure 6-53 and comprises:

1. A frequency synthesizer for stable generation of a sinewave.

2. A display plotting the S parameters in various forms, with a polar plot being most commonly used.
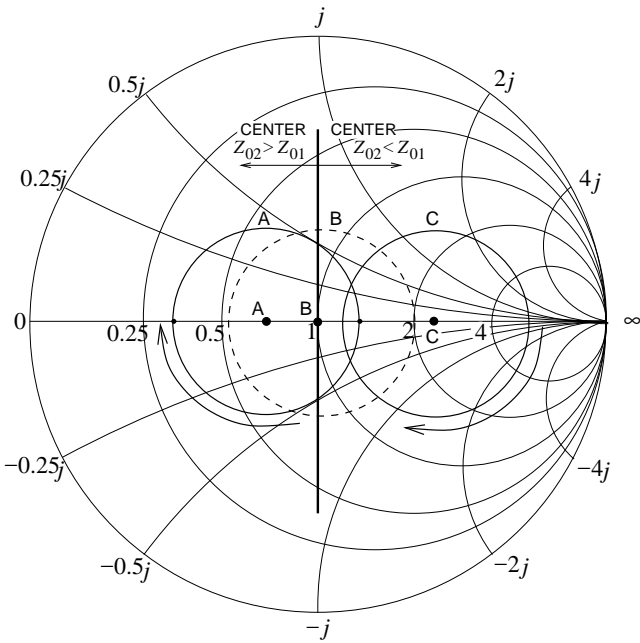
Figure 6-52 The locus of a the input reflection coefficient normalized to Z_{02} of a terminated line of characteristic impedance Z_{01}. The locus is with respect to the length of the line and the arrows show the direction of rotation of the reflection coefficient.

3. An S parameter test set. This device generally has two measuring ports so that when required, S_{11}, S_{12}, S_{21}, and S_{22} can all be determined under program control. The main components are switches and directional couplers. Directional couplers separate the forward- and backward-traveling wave components and will be considered in Section 9.7 on Page 483. Network analyzers with more than two ports—four-port network analyzers are popular—have more switches and directional couplers.

4. A computer controller used in correcting errors and converting the results to the desired form.

For RF measurements up to a few gigahertz, most of these functions are incorporated in a single instrument. At higher frequencies and with older equipment multiple units are used. An outline of such a system is shown in Figure 6-54(a).

An experimental setup for on-wafer measurements is shown in Figure 6-55. Figure 6-55(a) shows a network analyzer in the background with a precision voltage source on top. In the foreground is a probe station that has
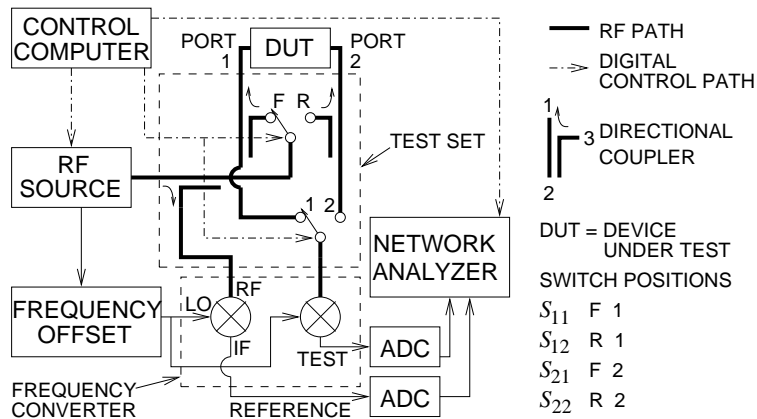
Figure 6-53 Vector network analyzer system. The directional couplers selectively couple forward- or backward-traveling waves. In the directional coupler inset, a traveling wave inserted at Port 2 appears only at Port 1. A traveling wave inserted at Port 1 appears at Port 2 and a coupled version at Port 3. A traveling wave inserted at Port 3 appears only at Port 1. Switch positions determine which S parameter is measured.
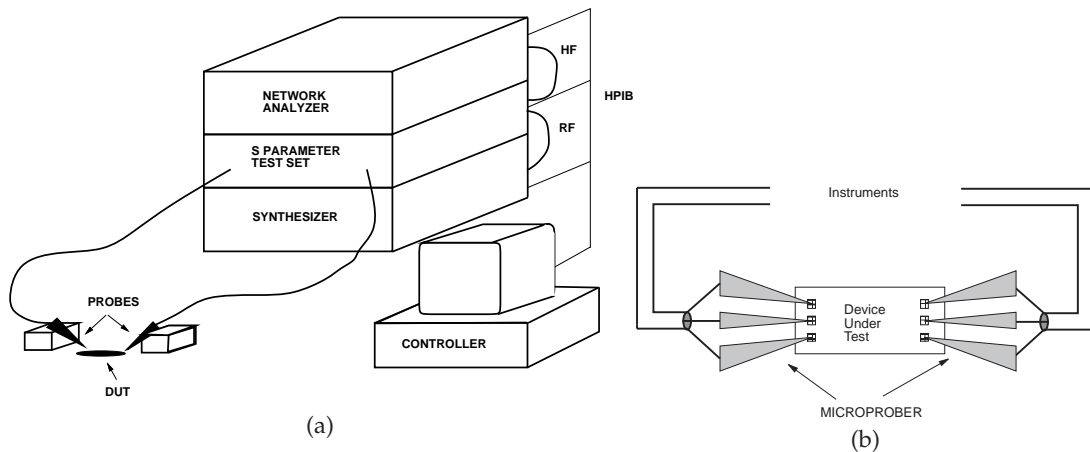
Figure 6-54 S parameter measurement system: (a) shown in a configuration to do on-chip testing; and (b) details of coplanar probes with Ground-Signal-Ground (GSG) configuration and contacting pads.

a stage for an IC and mounts for micropositioners to which **microprobes** are attached. The vertical tube-shaped object is a microscope-mounted camera. Figure 6-55(b) shows a microprobe making a connection to a transmission line on an IC and Figure 6-55(c) shows greater detail of the contact area.

(a)

(b)

(c)
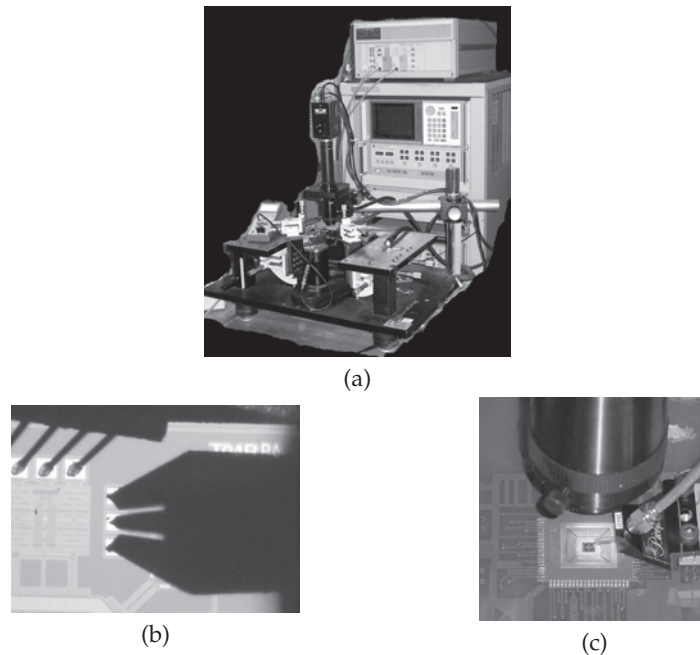
Figure 6-55 Vector network analyzer system for characterizing an IC or module: (a) probe station with video camera; (b) details of coplanar probes with GSG probes used with calibration structures; and (c) side view of an IC under test.

A typical microprobe is based on a micro coaxial cable with the center conductor (carrying the signal) extended a millimeter or so to form a needle-like contact. Two other needle-like contacts are made by attaching short extensions to the outer conductor of the coaxial line on either side of the signal connection. Such probes are called **Ground-Signal-Ground** (**GSG**) probes and transition from a coaxial cable to an on-wafer CPW line, introducing a smooth conversion of the EM fields from those of a coaxial line to those of the CPW line. The on-wafer pads can be as small as 50 μm on a side.

Calibration of the measurement system is critical, as the effect of cabling and connectors can be more significant than that of the device being measured. The process of removing the effect of cabling and connectors is called de-embedding. Ideally, short, open, and match loads would be available, but these can only be approximated, and numerous schemes have been developed for alternative ways of calibrating an RF measurement system. In some cases, for example, in measuring integrated circuits using what is called on-wafer probing, it is very difficult to realize a good match. A solution is to use calibration standards that consist of combinations of transmission line lengths and repeatable reflections. These calibration
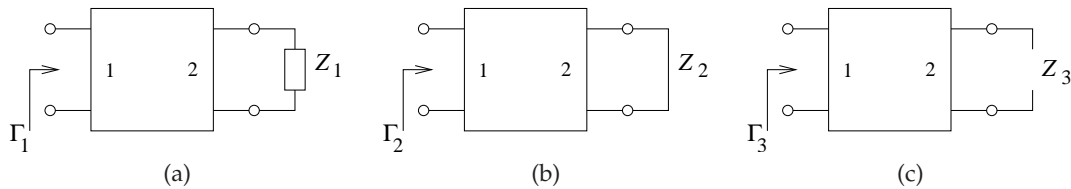
Figure 6-56 Two-ports, each with a load at Port 2.

procedures are known mostly by combinations of the letters T, R, and L such as **Through-Reflect-Line TRL** [98] (which relies on a transmission line of known characteristic impedance to replace the match); or **Through-Line TL**, procedure [91, 92] (which relies on symmetry to replace the third standard).

The essential concept behind calibration is to model errors at each port as a two-port error network between where the measurements are realized inside the instrument and the point at which calibration standards can be placed.

6.11.1 One-Port Calibration

In one-port measurements the desired reflection coefficient cannot be obtained directly. Instead, there is effectively an error network between the measurement plane, where the load is measured, and the ideal internal network analyzer port. The SFG model [93] of the measurement system is shown in Figure 6-56 with three ideal calibration standards: a matched load (Figure 6-56(a)); a short (Figure 6-56(b)); and an open (Figure 6-56(c)). Calibration here is concerned about determining the S parameters of the error two-port.

Various calibration schemes have been developed, some of which are better in particular measurement environments, such as when measuring on-chip structures for example. Calibration schemes share certain commonalities. One of these is that the errors can be represented as a two-port (or several two-ports) that exists between an effective reference plane internal to the measurement equipment and the reference plane at the device to be measured. This device is commonly called the DUT. Known standards are placed at the device reference plane and instrument measurements are made. The concept is illustrated by considering the determination of the S parameters of the two-port shown in Figure 6-56. With Z_1 being a matched load, the input reflection coefficient Γ_1 is S_{11}:

$$S_{11} = \Gamma_1 \ . \tag{6.235}$$

The other two commonly used calibration loads are $Z_2 = 0$ (a short circuit) and $Z_3 = \infty$ (an open circuit).[10] From these, $S_{12} = S_{21}$ and S_{22} can be

---

[10] Previously the undesirability of using opens and shorts at microwave frequencies was

derived as

$$S_{22} = \frac{2S_{11} - \Gamma_2 - \Gamma_3}{\Gamma_3 - \Gamma_2} \qquad (6.236)$$

and

$$S_{21} = S_{12} = [(\Gamma_2 - S_{11})(1 - S_{22})]^{1/2}. \qquad (6.237)$$

The error two-port modifies the actual reflection coefficient to a warped reflection coefficient according to Equation (6.198) on Page 324. In Equation (6.198), the load reflection coefficient, Γ_L, is a complex number that is scaled, rotated, and shifted. Moreover, circles of reflection coefficient are mapped to circles on the warped complex plane; that is, the ideal Smith chart is mapped to a warped Smith chart, as shown in Figure 6-57(a). Here the actual load reflection coefficient, Γ_L, is warped to present a measured reflection coefficient, Γ_M, to the internal network analyzer. Calibration can be viewed as a mapping operation from the complex plane of raw measurements to the complex plane of ideal measurements, as shown in Figure 6-57(a). Figure 6-57(b–e) shows how the mapping is developed using different sets of calibration standards. The mapping operation shown in Figure 6-57(b) uses the short, open, and match loads. These loads are well distributed over the Smith chart and so the mapping can be extracted with good precision. While ideal short, open, and match loads are not available, if the terminations are well characterized then appropriate corrections can be made. The mapping in the one-port case is embodied in the extracted S parameters of the two-port. Figure 6-58(b) shows the longitudinal section of a precision coaxial open. The key feature is that the outer coaxial conductor is extended while the center conductor is terminated. The fringing capacitance at the end of the center conductor can be calculated analytically. Thus the precision open is not a perfect open, but rather an open that can be modeled as a small capacitor whose frequency variation is known. Figure 6-57(c) uses offset shorts (i.e., short-circuited transmission lines of various lengths). This circumvents the problem of difficult-to-produce open and matched loads. Care is required in the choice of offsets to ensure that the mapping can be determined accurately. The calibration scheme shown in Figure 6-57(d) introduces a known load. The mapping cannot be extracted here, but in some situations this is the only option. Figure 6-57(e) introduces a new concept in calibration with the use of a sliding matched load such as that shown in Figure 6-58(a). A perfect matched load cannot be produced and there is always, at best, a small resistance error and perhaps parasitic capacitance. When measured, the reflection will have a small offset from the

---

discussed. However, here the short and open are precision calibration standards and the errors involved are known and incorporated in a more detailed calibration procedure than that presented here. All that is needed to determine the parameters of a two-port are three precisely known loads.
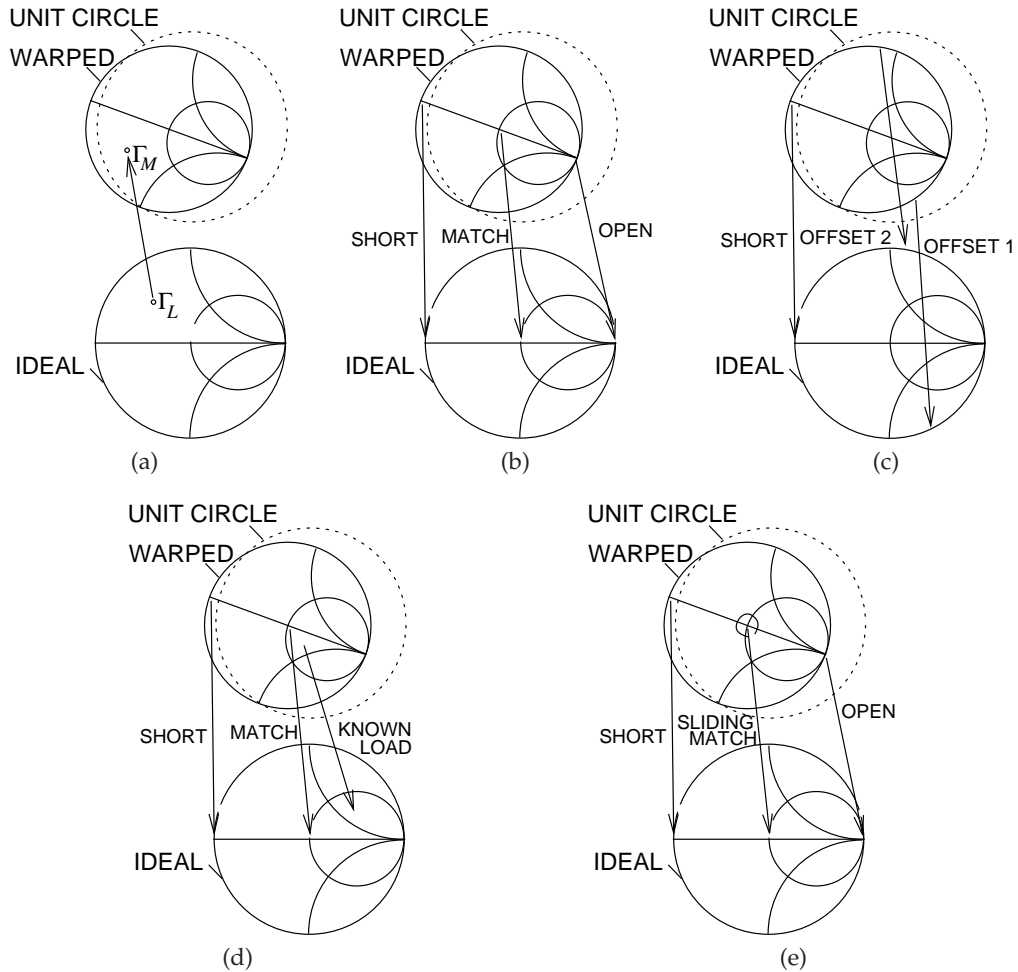
Figure 6-57 Calibration as a mapping operation showing the warped Smith chart, which describes raw measurements, and the ideal Smith chart: (a) warping of a load with reflection coefficient Γ_L to the measured reflection coefficient Γ_M; (b) calibration as a mapping established using short, match, and open calibration standards; (c) mapping using short and two offset short calibration standards; (d) mapping using a short, an offset short, and an offset load calibration; and (e) mapping using a short, an open, and a sliding matched load as calibration standards.
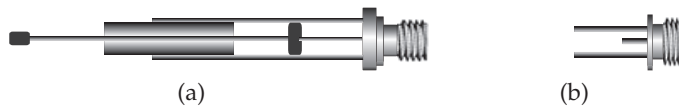
Figure 6-58 Precision calibration standards: (a) sliding load; and (b) precision open.

origin. If the matched load is moved, as with a sliding matched load, a small circle centered on the origin will be traced out and the center of this circle is the ideal matched load.

6.11.2 De-Embedding

The one-port calibration described above can be repeated for two ports, resulting in the two-port error model of Figure 6-59(a). Using the chain scattering matrix (or $\mathbf{T}$ matrix) introduced in Section 6.6.1 on Page 314, the $\mathbf{T}$ parameter matrix measured at the internal reference planes of the network analyzer is

$$\mathbf{T}_{\mathrm{MEAS}} = \mathbf{T}_A \mathbf{T}_{\mathrm{DUT}} \mathbf{T}_B, \qquad (6.238)$$

where $\mathbf{T}_A$ is the $\mathbf{T}$ matrix of the first error two-port, $\mathbf{T}_{\mathrm{DUT}}$ is the $\mathbf{T}$ matrix of the device under test, and $\mathbf{T}_B$ is the $\mathbf{T}$ matrix of the right-hand error two-port. Manipulating Equation (6.238) leads to $\mathbf{T}_{\mathrm{DUT}}$:

$$\mathbf{T}_A^{-1} \mathbf{T}_{\mathrm{MEAS}} \mathbf{T}_B^{-1} = \mathbf{T}_A^{-1} \mathbf{T}_A \mathbf{T}_{\mathrm{DUT}} \mathbf{T}_B \mathbf{T}_B^{-1} \qquad (6.239)$$

$$\mathbf{T}_{\mathrm{DUT}} = \mathbf{T}_A^{-1} \mathbf{T}_{\mathrm{MEAS}} \mathbf{T}_B^{-1}, \qquad (6.240)$$

from which the S parameters of the DUT can be obtained.

6.11.3 Two-Port Calibration

The VNA architecture introduced in 1968 and still commonly used today is shown in Figure 6-53. A key component of the system is the test set that comprises mechanical microwave switches and directional couplers that selectively couple energy in either the forward- or backward-traveling wave. This architecture supports a single signal source and a single test port at which the RF signal to be measured is mixed down to a low frequency, typically 100 kHz, where it is captured by an ADC. Another important component is the frequency converter, which mixes the RF signal with a version of the original RF signal offset in frequency to produce a low-frequency version of the RF signal with the same phase as the RF signal and an amplitude that is linearly proportional to the RF signal. When Hackborn introduced this system in 1968 [90], he also introduced the 8-term error model shown in Figure 6-59(a). This is an extension of the one-port error model discussed in Section 6.11.1, with a two-port capturing the measurement errors for each port.

Closer examination of the VNA system of Figure 6-53 indicates that the signal path varies for each of the four scattering parameters for a DUT. Another imperfection that is not immediately obvious is that there is leakage in the system so that Ports 1 and 2 are coupled by a number of mechanisms which do not involve transmission of the signal through the DUT. Leakage is due to finite isolation in the mixers, and direct coupling between probes used to make contact at Ports 1 and 2, among other possible causes. The first
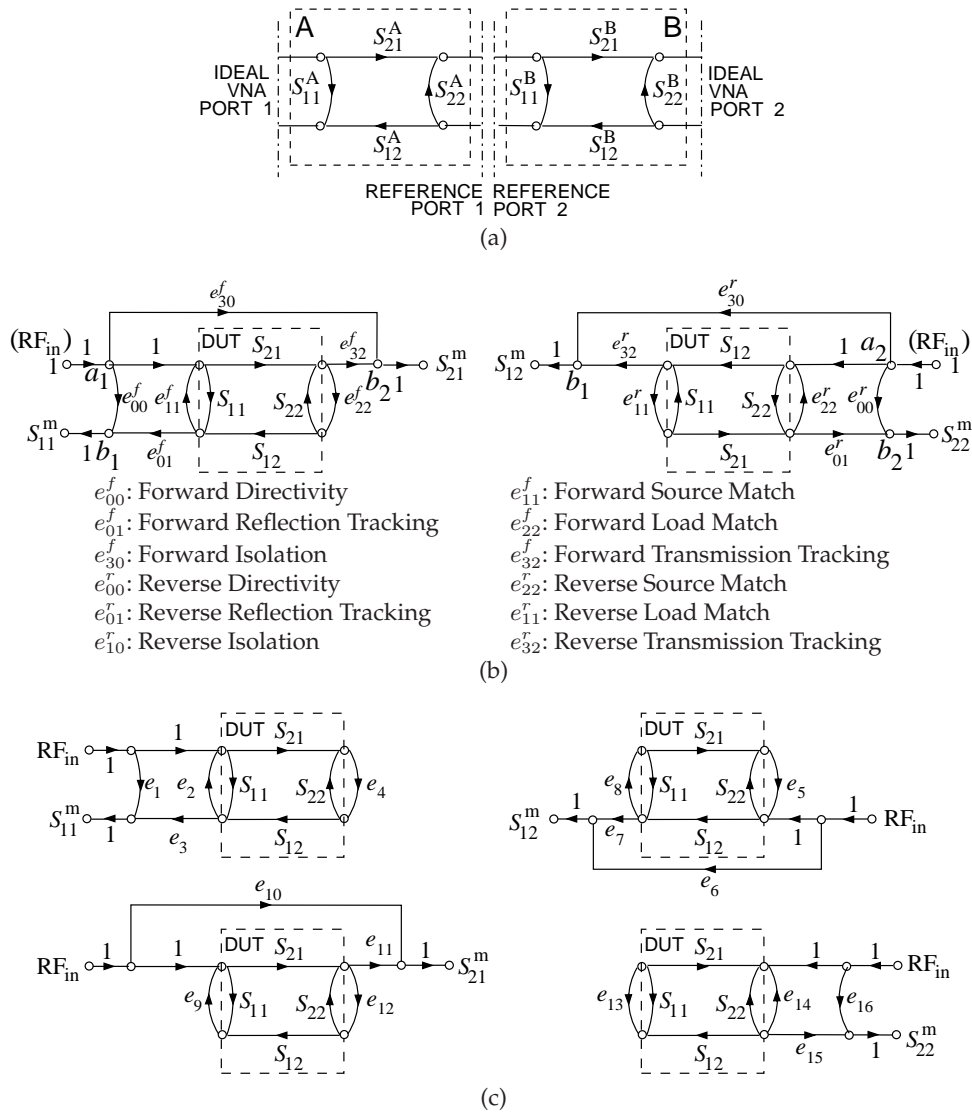
Figure 6-59   Two-port error models: (a) SFG representation of the 8-term two-port error model (1968) [90]); (b) SFG of the 12-term two-port error model (1978) [94]; and (c) SFG of the 16-term two-port error model (circa 1979) [95].

person to identify this was Shurmer [96] in 1973, leading eventually to the 12-term error model [94, 97] shown in Figure 6-59(b). Calibration yields the error parameters, and the S parameters of the DUT are found as follows:

$$S_{11} = \frac{1}{\delta} \left\{ \left[\frac{S_{11}^{\mathrm{m}} - e_{00}^f}{e_{01}^f} \right] \left[1 + \frac{S_{22}^{\mathrm{m}} - e_{00}^r}{e_{01}^r} e_{22}^r \right] - e_{22}^f \left[\frac{S_{21}^{\mathrm{m}} - e_{30}^f}{e_{32}^f} \right] \left[\frac{S_{12}^{\mathrm{m}} - e_{10}^r}{e_{32}^r} \right] \right\}$$
(6.241)

$$S_{21} = \frac{1}{\delta} \left\{ \left[\frac{S_{21}^{\mathrm{m}} - e_{30}^f}{e_{32}^f} \right] \left[1 + \frac{S_{22}^{\mathrm{m}} - e_{00}^r}{e_{01}^r} \left(e_{22}^r - e_{22}^f \right) \right] \right\}$$
(6.242)

$$S_{12} = \frac{1}{\delta} \left\{ \left[\frac{S_{12}^{\mathrm{m}} - e_{10}^r}{e_{32}^r} \right] \left[1 + \frac{S_{11}^{\mathrm{m}} - e_{00}^f}{e_{01}^f} \left(e_{11}^f - e_{11}^r \right) \right] \right\}$$
(6.243)

$$S_{22} = \frac{1}{\delta} \left\{ \left[\frac{S_{22}^{\mathrm{m}} - e_{00}^r}{e_{01}^r} \right] \left[1 + \frac{S_{11}^{\mathrm{m}} - e_{00}^f}{e_{01}^f} e_{11}^f \right] - e_{11}^r \left[\frac{S_{21}^{\mathrm{m}} - e_{30}^f}{e_{32}^f} \right] \left[\frac{S_{12}^{\mathrm{m}} - e_{10}^r}{e_{32}^r} \right] \right\}$$
(6.244)

$$\delta = \left[1 + \frac{S_{11}^{\mathrm{m}} - e_{00}^f}{e_{01}^f} e_{11}^f \right] \left[1 + \frac{S_{22}^{\mathrm{m}} - e_{00}^r}{e_{01}^r} e_{22}^r \right]$$
$$- e_{22}^f e_{11}^r \left[\frac{S_{21}^{\mathrm{m}} - e_{30}^f}{e_{32}^f} \right] \left[\frac{S_{12}^{\mathrm{m}} - e_{10}^r}{e_{32}^r} \right].$$
(6.245)

The most complete model is the 16-term error model introduced in 1979 and shown in Figure 6-59(c) which incorporates a different model for each parameter and fully captures the various signal paths resulting from the different switch positions [95]. Nevertheless, the 12-term error model is the one most commonly used in microwave measurements and is sufficient to reliably extract the desired S parameters of the DUT from the raw VNA measurements. This process of extracting the de-embedded parameters is call de-embedding, or less commonly, unterminating.

The 12-term error model models the error introduced by major leakage or cross-talk internal to the network analyzer. The 16-term error model captures the leakage or cross-talk of the 12-term error model, but in addition includes switch leakage that results in error signals reflecting from the DUT and leaking to the transmission port as well as common-mode leakage. In a coaxial environment, the extra leakage terms are small, provided that the switch has high isolation. However, in wafer probing, where direct coupling between the probes can be appreciable, the transmission-related leakage can be important and the 16-term error model captures errors that the 12-term error model does not.

Coaxial Two-Port Calibration

If the DUT has coaxial connectors then precision open, short, and matched loads are available. While these are not ideal components, the open has fringing capacitance; for example, they are well characterized over frequency. The full set of standards required to develop the 12-term error model includes one-port measurements for each port with open, short, and match connections, a two-port measurement with a through connection, and a two-port measurement with opens at each port. The last measurement determines the internal isolation of the system. Most connectors come in **male** (or **plug**) and **female** (or **jack**) counterparts. As such it is not possible to create a perfect through by joining the connectors and a small delay is introduced. There are, however, sexless connectors, such as the (**Amphenol Precision Connector** (**APC-7**) with a 7 mm outer diameter, and **GR 874** (General Radio model 874) connectors, as they have flush faces that can be connected for a near-perfect through. The APC-7 and GR 874 connectors are no longer commonly used, with most coaxial systems now using SMA connectors or the higher-tolerance APC 3.5 connectors that come in male and female connectors. The majority of the measurements desired these days are of ICs and other planar structures for which there is no clear connector. The connector, or fixture, problem has resulted in a variety of standards being used.

6.11.4 *Transmission Line-Based Calibration Schemes*

A drawback of all de-embedding techniques that use repeated fixture measurements during the de-embedding process is the generation of errors caused by the lack of fixture reproducibility during measurements of each of the required standards. These errors can be amplified during the course of de-embedding.

The **Through-Reflect-Line** (**TRL**) technique [98] is a calibration procedure used for microwave measurements when classical standards such as open, short, and matched terminations cannot be realized. There are now a family of related techniques, here denoted as TxL methods, with the commonality being the two-port measurement of a through connection and of the same structure with an inserted transmission line. The two-port connections required in the TRL procedure are shown in Figure 6-60. The concept behind the representation is that there is a measurement error that is captured by the fixture two-port. The TRL procedure allows the fixtures on either side, Fixtures A and B here, to have different two-port parameters. The SFG representations of the connections are shown in Figure 6-61. Assuming that the fixtures are faithfully reproduced when going from one connection to another, then the SFGs shown in Figures 6-61(a)–(c) need be considered. With each configuration, two-port measurements are made, and as the number of independent S parameter measurements exceed the

(a)   THROUGH CONNECTION

(b)   IDEAL LINE CONNECTION
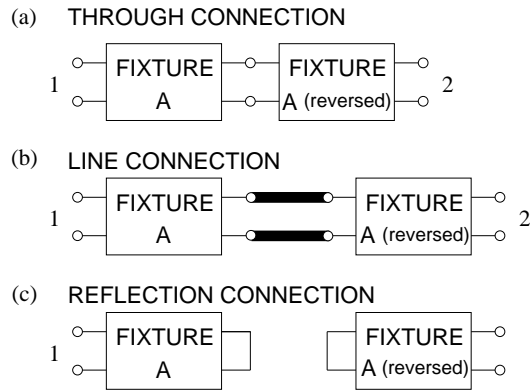

(c)   NONIDEAL LINE CONNECTION

(d)   REFLECTION CONNECTION
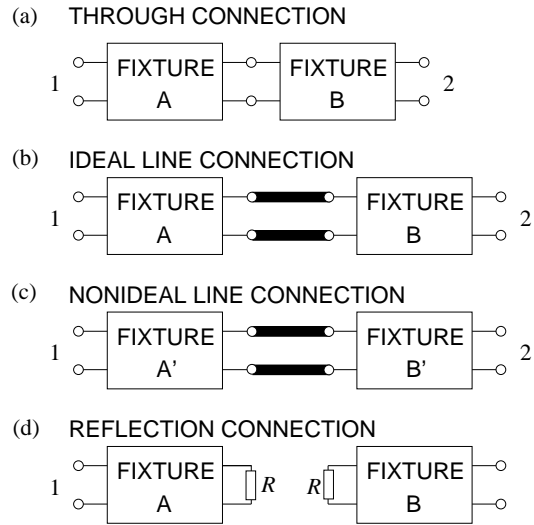
Figure 6-60 Measurement connections in two-port calibration using the TRL calibration procedure: (a) through connection; (b) ideal line connection; (c) realistic configuration showing that the fixtures are not precisely reproduced; and (d) arbitrary reflection configuration.
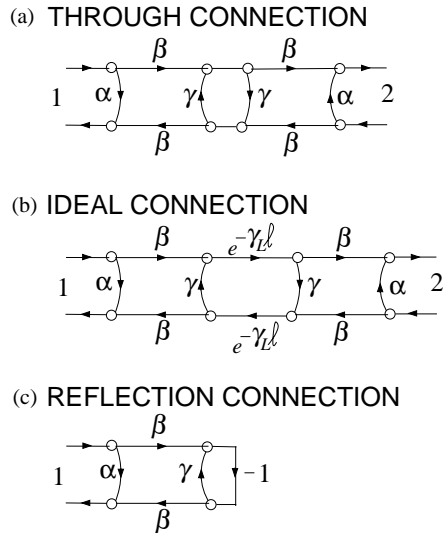
(a) THROUGH CONNECTION

(b) IDEAL CONNECTION

(c) REFLECTION CONNECTION
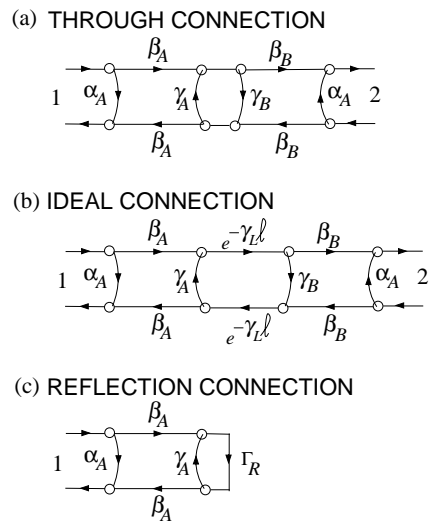
Figure 6-61 SFG representations of the two-port connections shown in Figure 6-60: (a) through connection; (b) ideal line connection; and (c) arbitrary reflection connection.

number of unknowns, it would seem that it may be possible to solve the
SFGs to obtain the unknown parameters of the fixtures. This is all that is
required to de-embed measurements made with a DUT between the fixtures.
Unfortunately this is not possible. However, Engen and Hoer [98] developed
a procedure that enabled the complex propagation constant of the line to be
extracted from the external S parameter measurements of the through and
line connections. This procedure is considered in the next section.

6.11.5 Through-Line Calibration

The through and line measurement structures include fixtures as well as
the direct connection (for the through) and the inserted line (for the line)
(see Figure 6-60). Two-port measurements of these two structures yield the
propagation constant $\gamma_L = \alpha + \jmath\beta$ of the line, as shown by Engen and
Hoer [98] in 1979. Similar but earlier work was presented by Bianco et
al. [101] in 1976. For the through structure, the error networks between
the ideal internal port of a network analyzer and the desired measurement
reference plane are designated as Fixture A at Port 1 and Fixture B at Port 2.
For the line measurement, fixturing is reestablished (following the Through
measurement) with the line inserted. In general, the fixturing cannot be
faithfully reproduced and therefore becomes Fixtures A$'$ and B$'$, respectively
(see Figure 6-60).

The following is based on the derivation of Engen and Hoer [98], but
with a particular extension by Buff et al. [99,100]. The development begins
using cascading matrices, $\mathbf{R}$, each of which is related to two-port scattering
parameters $\mathbf{S}$ by

$$\mathbf{R} = \left[\begin{array}{cc} R_{11} & R_{12} \\ R_{21} & R_{22} \end{array} \right] = \left[\begin{array}{cc} \frac{S_{12}\,S_{21} - S_{11}\,S_{22}}{S_{21}} & \frac{S_{11}}{S_{21}} \\ -\frac{S_{22}}{S_{21}} & \frac{1}{S_{21}} \end{array} \right]. \tag{6.246}$$

Thus the parameters describing the A and B fixtures are $\mathbf{S}_A$, $\mathbf{S}_B$ and $\mathbf{R}_A$, $\mathbf{R}_B$,
respectively. The cascading matrix of the through is simply a unity matrix
and the line of length, ℓ_L, is described by

$$\mathbf{R}_L = \left[\begin{array}{cc} e^{-\gamma_L \ell_L} & 0 \\ 0 & e^{\gamma_L \ell_L} \end{array} \right]. \tag{6.247}$$

The characteristic impedance of the line is taken as the reference impedance,
Z_0, of the measurement system and is reflected by the diagonal zeros
in the above definition for the line standard. Because of this simplifying
assumption, γ_L can be computed using the equations that follow.

The cascading matrix of the through structure is

$$\mathbf{R}_t = \mathbf{R}_A \mathbf{R}_B \quad \text{(through)} \tag{6.248}$$

and the line structure (without fixturing errors such that $\mathbf{R}'_A = \mathbf{R}_A$ and $\mathbf{R}'_B = \mathbf{R}_B$) has the cascading matrix

$$\mathbf{R}_d = \mathbf{R}_A \mathbf{R}_L \mathbf{R}_B \quad \text{(line)}. \tag{6.249}$$

Solving Equation (6.248) for $\mathbf{R}_B$ and substituting into Equation (6.249) yields what will be referred to as the TL equation:

$$\mathbf{T}\mathbf{R}_A = \mathbf{R}_A \mathbf{R}_L, \tag{6.250}$$

where

$$\mathbf{T} = \mathbf{R}_d \mathbf{R}_t^{-1}. \tag{6.251}$$

Expanding the above leads to the system of equations

$$t_{11} R_{A11} + t_{12} R_{A21} = R_{A11} e^{-\gamma_L \ell_L} \tag{6.252}$$
$$t_{21} R_{A11} + t_{22} R_{A21} = R_{A21} e^{-\gamma_L \ell_L} \tag{6.253}$$
$$t_{11} R_{A12} + t_{12} R_{A22} = R_{A12} e^{\gamma_L \ell_L} \tag{6.254}$$
$$t_{21} R_{A12} + t_{22} R_{A22} = R_{A22} e^{\gamma_L \ell_L}, \tag{6.255}$$

which upon solution yields [98]

$$\gamma_L = \frac{1}{2\,\ell_L} \ln \left(\frac{t_{11} + t_{22} \pm \zeta}{t_{11} + t_{22} \mp \zeta} \right), \tag{6.256}$$

where the error term is

$$\zeta = \left(t_{11}{}^2 - 2\,t_{11} t_{22} + t_{22}{}^2 + 4\,t_{21} t_{12} \right)^{1/2}. \tag{6.257}$$

If the fixtures are faithfully reproduced, then networks A and A′ are identical, as are networks B and B′, and so $\zeta = 0$, and the propagation constant of the line γ_L is obtained without error as

$$\gamma_L = \frac{1}{2\,\ell_L} \ln \left(\frac{t_{11} + t_{22}}{t_{11} + t_{22}} \right). \tag{6.258}$$

That is, two-port S parameter measurements of the through connection shown in Figure 6-60(a) and two-port S parameter measurements of the line connection shown in Figure 6-60(b) enable the propagation constant of the line to be determined.

However, the TxL techniques generally result in errors in de-embedding at frequencies where the length of the line standard is a multiple of one-half wavelength. Buff et al. [99,100] showed that small fixture repeatability errors result in errors incorporated in ζ. These errors become evident at frequencies where the length of the line is an odd-integer multiple of a half-wavelength, called critical lengths [102,103]. These errors affect the entire calibration and subsequent de-embedded measurements. These errors are minimized by using measurements where the line standard has an electrical length within 20° of a critical length. Statistical means and the use of multiple lines have been proposed to minimize the error [104,105]. However, these schemes are not necessary to correct the error.
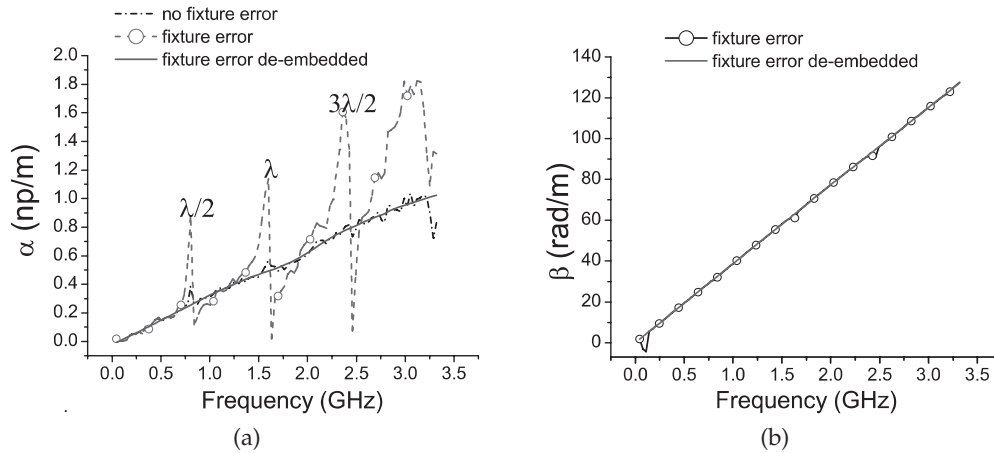
Figure 6-62 Through and line components of TRL with fixture error introduced during the line measurement.
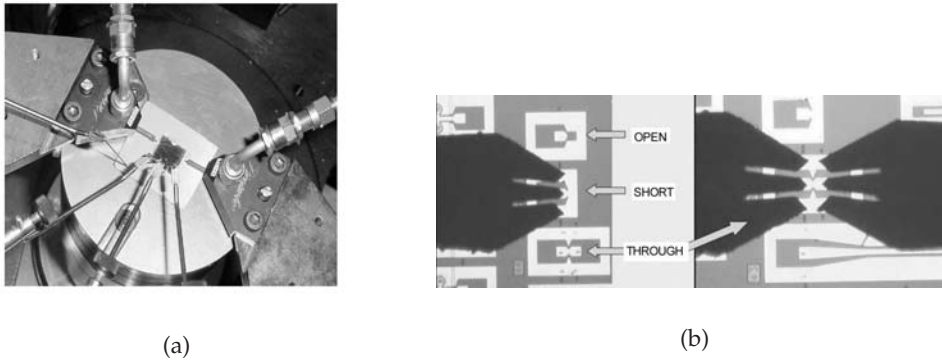
Figure 6-63 Two-port measurements: (a) two-port device under test with DC needle probes at the bottom and GSG probes on the left and right; and (b) open, short, and through calibration standards with GSG probes.

TSD—Through Short Delay

Matched loads can be difficult to realize especially when probing ICs and planar circuits using probes. **Through Short Delay** (**TSD**) uses reflection and transmission measurements [106]. The delay is realized using a line with a 50 Ω characteristic impedance. The through could be another delay, but problems occur when the delay difference (the electrical length difference) of the through and delay is an integer multiple of $180°$. An on-wafer two-port measurement system is shown in Figure 6-63, showing on-chip calibration standards.

TRL—Through Reflect Line

What has become known as the **Through Reflect Line** (**TRL**) calibration technique was developed by Bianco et al. [101] and Engen and Hoer [98], who used a through connection, an arbitrary reflection standard, and a delay or line standard. The line standard can be of arbitrary length with its propagation constant unknown, but it is assumed to be nonreflecting. The result of this assumption is that the characteristic impedance of the line becomes the system reference impedance. Also, the reflection standard, can be any repeatable reflection load, with an open or short preferred. As identified by Engen and Hoer, [98] the transmission line introduced calibration uncertainties when the electrical length was a multiple of one-half wavelength.

TRM—Through Reflect Match

Through Reflect Match (**TRM**) is a variation of TRL (a major variation to be true) that replaces the through by a matched load and is useful only when a matched load can be realized. It retains the arbitrary reflection standard, which is particularly useful as it can be difficult to produce a short or open with planar circuits. A via used in realizing a planar circuit short has finite resistance and inductance and so does not create a good through, but does create a useful repeatable arbitrary reflection.

Other Distributed Standards

TRL and TRM are the most commonly accepted standards used in calibrating measurements of planar circuits. There are several other techniques that also use transmission lines and find usefulness in certain circumstances.

In 1982 Benet [107] developed the open-short-five offset line calibration technique. The key concept is that the multiple offset through lines traced out a circle on the Smith chart and the center of the circle provided an additional error term. Through least squares fitting Benet developed the propagation constant of the through lines and the remaining error terms to develop the 12-term error model.

Pennock et al. [108] introduced the **Double Through Line** (**DTL**) calibration system in 1987. This is useful for characterizing transitions between different media. There is sufficient information to develop the 6-term error model.

6.11.6   Two-Tier Calibration

With measurements in noncoaxial systems such as measurements on ICs, it is difficult to produce precision standards such as shorts, opens, and resistive loads. In these situations, two-tier calibration is sometimes used. In
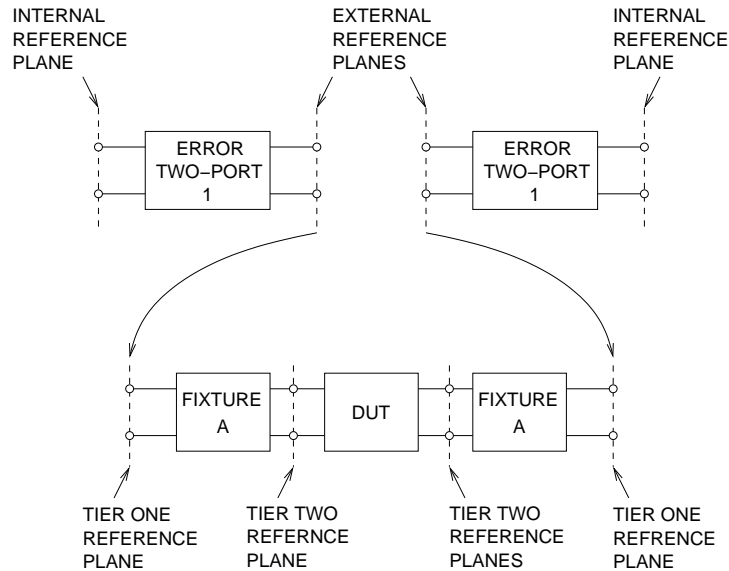
Figure 6-64   Two-tier calibration procedure.

a two-tier calibration procedure, two sets of standards are used to establish first- and second-tier reference planes (see Figure 6-64). In the first-tier calibration, a precision set of standards are normally used. This could be achieved using coaxial standards or perhaps a standard calibration substrate used with probes. Each of the 2 two-ports from the internal reference plane of the network analyzer to an external reference plane is established. Generally with the second tier, insitu calibration standards are used, and the additional fixturing is identical for the two ports. Precision calibration should be done using standards fabricated in the medium of the DUT.

  Two-tier calibration generally results in the fixtures in the second tier being identical, which enables checks to be made for the integrity of the probe connection. Connections to IC pads can be problematic because of oxide build up, especially with aluminum pads. So ensuring symmetry of second-tier calibrations by comparing the two-port measurements of symmetrical structures such as transmission lines can greatly reduce the impact of the fixture-induced errors discussed in Section 6.11.5. That is, small fixturing errors, such as bringing the probes down in slightly different positions, results in measurement errors at the half-wavelength frequencies ($\lambda/2$, λ, $3\lambda/2$, etc.). For example, if the DUT is a device on a silicon wafer, the optimum measurement technique is to realize the calibration standards on the silicon wafer processed in the same way as for the DUT. Typically such measurements would be conducted using probes, and the calibration procedure removes fixturing errors associated with fringing effects at the

probes and with other nonidealities involved in creating the landing pads for the probes. For measurements of microstrip and CPW systems, standard calibration alumina substrates are available. These calibration substrates include TxL standards as well as an open, a short, and a laser-trimmed matched load.

In a two-tier calibration scheme, a first tier is used to establish the 12-term error model. Then a second calibration is used to develop a secondary six-term error model, as in Figure 6-59(a) with $S_{12} = S_{21}$. In such a situation the standards used in developing the 12-term error model would normally include an open, a short, and a matched load. These need not be precise, and so it is not necessary to use a distributed standard such as the TxL family of calibration standards. Thus the 12-term error model would not be precise, but the model does not suffer from the half-wavelength errors of the TxL techniques. For nonplanar measurements using probes, calibration is to the end of the probe tips, and errors with using other substrates will be common to both probe tips in a two-port measurement. In the secondary calibration the measurements of a through and line will be symmetrical with $S_{11} = S_{22}$ and $S_{12} = S_{21}$. This symmetry is exploited in the **Through Line** () technique which exploits symmetry and renders the half-wavelength errors of TxL insignificant [91, 92].

TL—Through Line Symmetry

In a two-tier calibration scheme, a first-tier calibration can result in good calibration to the end of the probe tips. However when these probe tips are used in another configuration additional errors can be introduced because of the fixturing required to connect the DUT. With careful design the fixtures employed in the second tier will be identical and have the same S parameters. The TL calibration procedure exploits this property. The TL calibration connections are shown in Figure 6-65. In the first calibration tier, a 12-term error model is developed using open, short, load, and delay. Consequently there are no half-wavelength glitches that occur when transmission line standards are used. In the second calibration tier, probes are used with two lengths of transmission line fabricated on the same substrate as the DUT. One of the lines becomes a through.

The procedure proceeds by calculating the propagation constant of the line using the technique developed for the TRL procedure (Equation (6.258)). There is fixture symmetry in the TL connections (see Figure 6-65) and this is reflected in the SFG representations of the connections as shown in Figure 6-66. As will be shown, this symmetry enables the arbitrary reflection used in the TRL procedure to be replaced by a precise short or open circuit.

Let the S parameters of Fixture A be $\alpha = S_{11}$, $\beta = S_{12} = S_{21}$, and $\gamma = S_{22}$. Using these parameters with the port reversal required to model the Fixture at Port 2, the SFG shown in Figure 6-66(a) relates the measured S

(a)   THROUGH CONNECTION

(b)   LINE CONNECTION

(c)   REFLECTION CONNECTION

Figure 6-65   Measurement connections in the two-port second-tier calibration using the TL calibration procedure: (a) through connection; (b) ideal line connection; (c) realistic configuration showing that the fixtures are not precisely reproduced; and (d) arbitrary reflection configuration.

(a) THROUGH CONNECTION

(b) IDEAL CONNECTION

(c) REFLECTION CONNECTION

Figure 6-66   SFG representations of the two-port connections shown in Figure 6-65: (a) through connection; (b) ideal line connection; and (c) ideal reflection connection.

parameters to the parameters of the fixtures. Let the measured S parameters of the through connection be designated by a leading superscript T, then the input reflection coefficient of Fixture A with a short circuit placed at Port 2

of A is

$$\rho_{\text{sc}} = \alpha - \frac{\beta^2}{1+\gamma}.$$

Using Mason's rule with the SFG in Figure 6-66(a) relates the measured S parameters to the fixture parameters:

$$^T S_{11} = \alpha + \frac{\beta^2 \gamma}{1 - \gamma^2} \tag{6.259}$$

$$^T S_{21} = \frac{\beta^2}{1 - \gamma^2}. \tag{6.260}$$

Subtracting these expressions yields

$$^T S_{11} - {}^T S_{21} = \alpha + \frac{\beta^2 \gamma}{1 - \gamma^2} - \frac{\beta^2}{1 - \gamma^2} = \alpha + \frac{\beta^2 \gamma - \beta^2}{(1 - \gamma)(1 + \gamma)} \tag{6.261}$$

$$= \alpha - \frac{\beta^2 (1 - \gamma)}{(1 - \gamma)(1 + \gamma)} \tag{6.262}$$

$$= \alpha - \frac{\beta^2}{1 + \gamma} \tag{6.263}$$

$$= \rho_{\text{sc}}. \tag{6.264}$$

That is, the input reflection coefficient of a fixture terminated in a short circuit can be obtained from the through measurement of the back-to-back fixtures in the through configuration:

$$\rho_{\text{sc}} = {}^T S_{11} - {}^T S_{21}.$$

Similar results are obtained for an ideal open circuit placed at Port 2 of Fixture A:

$$\rho_{\text{oc}} = {}^T S_{11} + {}^T S_{21}.$$

Thus it is possible to effectively insert ideal open and short circuits within a noninsertable medium.

6.12 Summary

Scattering parameters describe power flow and traveling waves and are essential to describing distributed circuits as encountered in RF and microwave engineering. RF and microwave design is characterized by conceptual insight and it is essential to use parameters and graphical representations that are close to the physical world. Scattering parameters and the graphical polar representations on a Smith chart are the overwhelming choice of RF and microwave designers. Design is more than evaluating equations and running computer-aided design programs. In RF and microwave engineering there are always considerable approximations

made in design, partly because of necessary simplifications that must be made in modeling, but also because many of the material properties required in a detailed design can only be approximate. An example is permittivity. The permittivity of circuit board material, FR-4, ceramics, and silicon dioxide in silicon ICs depends on the density of the material, which depends on metal density and circuit layout. The variation can be as much as 10%. Most RF and microwave design deals with frequency selective circuits such as line lengths which have an electrical length that is a particular fraction of a wavelength. The relationship to physical length depends on material properties and the ability to model them well. Many designs can require frequency tolerances of as little as 0.1%, and filters can require even tighter tolerances. It is therefore impossible to design exactly. Measurements are required to validate designs and iterate designs to the final frequency centering. Conceptual understanding is essential; the designer must be able to relate measurements, which themselves have errors, with computer simulations. The ability to design circuits with good tolerance to manufacturing variations and perhaps circuits that can be tuned by automatic equipment are skills developed by experienced designers.

6.13 Exercises

1. The scattering parameters of a two-port are $S_{11} = 0.25$, $S_{12} = 0$, $S_{21} = 1.2$, and $S_{22} = 0.5$. The system reference impedance is 50 Ω and the Thevenin equivalent impedance of the source attached to Port 1 is 50 Ω. The power available from the source connected to Port 1 is 1 mW. The load impedance is 25 Ω.

   (a) Is the two-port reciprocal and why?
   (b) What is the voltage of the Thevenin equivalent source?
   (c) What is the power reflected from Port 1?
   (d) Determine the z parameters of the two-port.
   (e) Using z parameters, what is the power dissipated by the load at Port 2?

2. The scattering parameters of a certain two-port are $S_{11} = 0.5 + \jmath 0.5$, $S_{12} = 0.95 + \jmath 0.25$, $S_{21} = 0.15 - \jmath 0.05$, and $S_{22} = 0.5 - \jmath 0.5$. The system reference impedance is 50 Ω.

   (a) Is the two-port reciprocal? Explain your answer.
   (b) Consider that Port 1 is connected to a matched source with an available power of 1 W. What is the power delivered to a 50 Ω load placed at Port 2?

   (c) Using the situation described in (b), what is the power reflected from Port 1?
   (d) What is the value of the load required for maximum power transfer at Port 2? Express your answer as a reflection coefficient.

3. A 50 Ω, 10 dB attenuator is inserted in a 75 Ω system.

   (a) What is the transmission coefficient in the 75 Ω system?
   (b) What is the attenuation in decibels in the 75 Ω system?
   (c) What is the input reflection coefficient at Port 1 including the 75 Ω termination at Port 2?

4. The scattering parameters of a two-port are $S_{11} = 0.25$, $S_{12} = 0$, $S_{21} = 1.2$, and $S_{22} = 0.5$. The system reference impedance is 50 Ω and the Thevenin equivalent impedance of the source attached to Port 1 is 50 Ω. The power available from the source connected to Port 1 is 1 mW. The load impedance is 25 Ω. Using SFGs determine the power dissipated by the load at Port 2.

5. Draw the SFG of a two-port with a load at Port 2 having a voltage reflection coefficient of Γ_L, and at Port 1 a source reflection coefficient of

Γ_S. This is the reflection coefficient looking from Port 1 of the two-port toward the generator. Keep the S parameters of the two-port in symbolic form (e.g., S_{11}, S_{12}, S_{21}, and S_{22}). Using SFG analysis, derive an expression for the reflection coefficient looking into the two-port at Port 2. You must show the stages in collapsing the SFG to the minimal form required. Make sure that you include the impact of the source.

6. A three-port has the scattering parameters

$$\begin{bmatrix} 0 & \gamma & \delta \\ \alpha & 0 & \epsilon \\ \theta & \beta & 0 \end{bmatrix}.$$

Port 2 is terminated in a load with a reflection coefficient Γ_L. Reduce the two-port to a three-port using signal flow graph theory. Write down the four scattering parameters of the final two-port.

7. The S parameters of a three-port are as follows (the S parameters are referred to a 50 Ω system reference impedance):

$$\begin{bmatrix} 0.2\angle180° & 0.8\angle-45° & 0.1\angle45° \\ 0.8\angle-45° & 0.2\angle0° & 0.1\angle90° \\ 0.1\angle45° & 0.1\angle90° & 0.1\angle180° \end{bmatrix}.$$

(a) Is the three-port reciprocal? Explain your answer.

(b) Write down the criteria for a network to be lossless.

(c) Is the three-port lossless? You must show your working.

(d) Draw the SFG of the three-port.

(e) A 50 Ω load is attached to Port 3. Use SFG operations to derive the SFG of the two-port with just Ports 1 and 2. Write down the two-port S parameter matrix of the simplified network.

8. The S parameters of a three-port are as follows (the S parameters are referred to a 50 Ω system reference impedance):

$$\begin{bmatrix} 0.2\angle180° & 0.8\angle-45° & 0.1\angle45° \\ 0.8\angle-45° & 0.2\angle0° & 0.1\angle90° \\ 0.1\angle45° & 0.1\angle90° & 0.1\angle180° \end{bmatrix}.$$

(a) Is the three-port reciprocal? Explain your answer.

(b) Write down the criteria for a network to be lossless.

(c) Is the three-port lossless? Explain your answer.

(d) Draw the SFG of the three-port.

(e) A 150 Ω load is attached to Port 3. Derive the SFG of the two-port with just Ports 1 and 2. Write down the two-port S parameter matrix of the simplified network.

9. A system consists of 2 two-ports in cascade. The S parameter matrix of the first two-port is $\mathbf{S}_A$ and the S parameter matrix of the second two-port is $\mathbf{S}_B$. The second two-port is terminated in a load with a reflection coefficient Γ_L. (Just to be clear, $\mathbf{S}_A$ is followed by $\mathbf{S}_B$ which is then terminated in Γ_L. Port 2 of $\mathbf{S}_A$ is connected to Port 1 of $[S_B]$. The individual S parameters of the two-ports are as follows:

$$\mathbf{S}_A = \begin{bmatrix} a & c \\ c & b \end{bmatrix} \text{ and } \mathbf{S}_B = \begin{bmatrix} d & f \\ f & e \end{bmatrix}.$$

(a) Draw the SFG of the system consisting of the two cascaded two-ports and the load.

(b) Ignoring the load, is the cascaded two-port system reciprocal? (If there is not enough information to decide say so.) Give your reasons.

(c) Ignoring the load, is the cascaded two-port system lossless? (If there is not enough information to decide say so.) Give your reasons.

(d) Consider that the load has no reflection (i.e., $\Gamma_L = 0$). Use SFG analysis to find the input reflection coefficient looking into Port 1 of $[S_A]$. Your answer should be Γ_{in} in terms of $a, b, \ldots, f$.

10. A circulator is a three-port device that is not reciprocal and selectively shunts power from one port to another. The S parameters of an ideal circulator are

$$[S] = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

(a) Draw the three-port and discuss power flow.

(b) Draw the SFG of the three-port.

(c) If Port 2 is terminated in a load with reflection coefficient 0.2, reduce the SFG of the three-port to a two-port with Ports 1 and 3 only.

(d) Write down the 2×2 S parameter matrix of the (reduced) two-port.

11. The S parameters of a certain two-port are $S_{11} = 0.5 + j0.5$, $S_{21} = 0.95 + j0.25$, $S_{12} = 0.15 - j0.05$, $S_{22} = 0.5 - j0.5$. The system reference impedance is 50 Ω.

(a) Is the two-port reciprocal?

(b) Which of the following is a true statement about the two-port.

(A) It is unitary.

(B) Overall power is produced by the two-port.

(C) It is lossless.

(D) If Port 2 is matched, then the real part of the input impedance (at Port 1) is negative.

(E) A and C.

(F) A and B.

(G) A, B, and C

(H) None of the above.

(c) What is the value of the load required for maximum power transfer at Port 2? Express your answer as a reflection coefficient.

(d) Draw the SFG for the two-port with a load at Port 2 having a voltage reflection coefficient of Γ_L and at Port 1 a source reflection coefficient of Γ_S.

12. In the distribution of signals on a cable TV system a 75 Ω coaxial cable is used, as the loss is close to minimum. In a low-loss 75 Ω cable the attenuation constant is 0.1 dB/m at 1 GHz. If a subscriber disconnects a television set from the cable so that the load impedance looks like an open circuit, estimate the input impedance of the cable at 1 GHz and 1 km from the subscriber. An answer within 1% is required. Estimate the error of your answer. Indicate the input impedance on a Smith chart drawing the

locus of the input impedance as the line is increased in length from nothing to 1 km.

13. A load has an impedance $Z = 115 - j20$ Ω.

(a) What is the reflection coefficient, Γ_L, of the load in a 50 Ω reference system?

(b) Plot the reflection coefficient on a polar plot of reflection coefficient.

(c) If a one-eighth wavelength long lossless 50 Ω transmission line is connected to the load, what is the reflection coefficient, Γ_{in}, looking into the transmission line? (Again, use the 50 Ω reference system.) Recall that the reflection coefficient is the ratio of the backward-traveling wave to the forward-traveling wave with phases and amplitudes referred to the same point. Plot Γ_{in} on the polar reflection coefficient plot of part (b). Clearly identify Γ_{in} and Γ_L on the plot.

(d) On the Smith chart, identify the locus of Γ_{in} as the length of the transmission line increases from zero length to one-eighth wavelength long. That is, on the Smith chart, plot Γ_{in} as the length of the transmission line varies.

14. In Figure 6-67 the results of several different experiments are plotted on a Smith chart. Each experiment measured the input reflection coefficient from a low frequency (denoted by a circle) to a high frequency (denoted by a square) of a one-port. Determine the load that was measured. The loads that were measured were one of those shown below.

Load	Description
i	An inductor
ii	A capacitor
iii	A reactive load at the end of a transmission line
iv	A resistive load at the end of a transmission line
v	A parallel connection of an inductor, a resistor, and a capacitor going through resonance and with a transmission line offset

Load	Description
vi	A series connection of a resistor and a capacitor going through resonance and with a transmission line offset
vii	A series resistor and inductor
viii	An unknown load and not one of the above

You should make no assumptions about how low the low frequency was, nor about how high the high frequency was. For each of the measurements below indicate the load using the load identifier above (e.g., i, ii, etc.) There may be more than one correct answer.

(a) What type of load gives rise to the reflection coefficient indicated by curve A?

(b) What type of load gives rise to the reflection coefficient indicated by curve B?

(c) What type of load gives rise to the reflection coefficient indicated by curve C?

(d) What type of load gives rise to the reflection coefficient indicated by curve D?

(e) What type of load gives rise to the reflection coefficient indicated by curve E?

(f) What type of load gives rise to the reflection coefficient indicated by curve F?
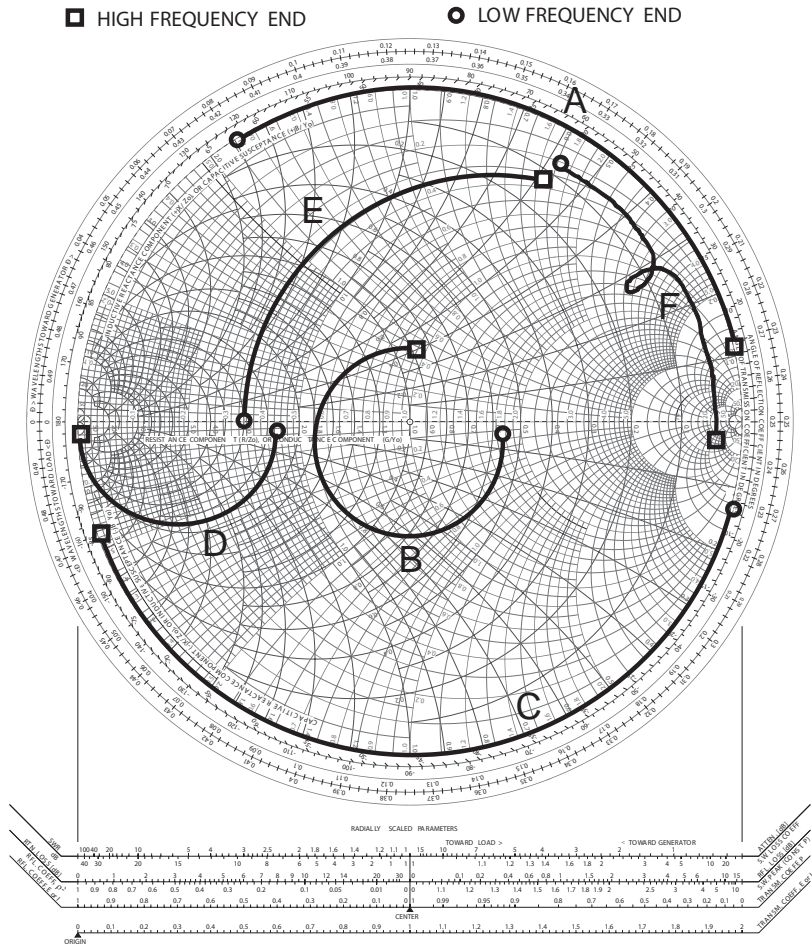
Figure 6-67   The locus of various loads plotted on a Smith chart.

15. A 50 Ω lossy transmission line is shorted at one end. The line loss is 2 dB per wavelength. Note that since the line is lossy the characteristic impedance will be complex, but close to 50 Ω, since it is only slightly lossy. There is no way to calculate the actual characteristic impedance with the information provided. However, this is the common situation that must be dealt with. That is, problems must be solved with small inconsistencies. Full accuracy is not possible and there is often missing information. Microwave engineers do the best they can in design and always rely on measurements to calibrate results.

   (a) What is the reflection coefficient at the load (in this case the short)?
   (b) Consider the input reflection coefficient, Γ_{in}, at a distance ℓ from the load. Determine Γ_{in} for ℓ going from 0.1λ to λ in steps of 0.1λ.
   (c) On a Smith chart plot the locus of the input reflection coefficient from $\ell = 0$ to λ.
   (d) Calculate the input impedance, Z_{in}, when the line is $(3/8)\lambda$ long using the telegrapher's equation.
   (e) Repeat part (d) using a Smith chart.

16. Design a open-circuited stub with an input impedance of $+j75\ \Omega$. Use a transmission line with a characteristic impedance of 75 Ω. [Parallels Example 6.5 on Page 339]

17. Design a short-circuited stub with an input admittance of $-j50\ \Omega$. Use a transmission line with a characteristic impedance of 100 Ω. [Parallels Example 6.5 on Page 339]

18. A load has an impedance $Z_L = 25 - \jmath100\ \Omega$.
   (a) What is the reflection coefficient, Γ_L, of the load in a 50 Ω reference system?
   (b) If a quarter-wavelength long 50 Ω transmission line is connected to the load, what is the reflection coefficient, Γ_{in}, looking into the transmission line?
   (c) Describe the locus of Γ_{in}, as the length of the transmission line is varied from zero length to one-half wavelength long. Use a Smith chart to illustrate your answer

19. A network consists of a source with a Thevenin equivalent impedance of 50 Ω driving first a series reactance of 50 Ω followed by a one-eight wavelength long transmission line with a characteristic impedance of 40 Ω and a reactive element of 25 Ω in shunt with a load having an impedance $Z_L = 25 - \jmath100\ \Omega$. This problem must be solved graphically and no credit will be given if this is not done.

   (a) Draw the network.
   (b) On a Smith chart, plot the locus of the reflection coefficient first for the load, then with the element in shunt, then looking into the transmission line, and finally the series element. Use letters to identify each point on the Smith chart. Write down the reflection coefficient at each point.
   (c) What is the impedance presented by the network to the source?

20. In the circuit below, a 75 Ω lossless line is terminated in a 40 Ω load. On the complex plane plot the locus, with respect to the length of the line, of the reflection coefficient, looking to the line referencing it first to a 50 Ω impedance. [Parallels Example 6.6 on Page 342]
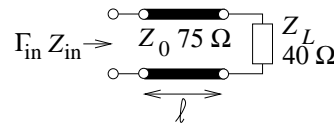
21. Consider the circuit below, a 60 Ω lossless line is terminated in a 40 Ω load. What is the center of the reflection coeffient locus on the complex plane when it is referenced to 55 Ω. [Parallels Example 6.7 on Page 346]
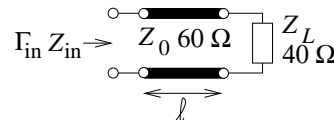
22. In Section 6.11 on Page 346, the S parameters of a reciprocal error network were determined by applying three loads—Z_1, Z_2 and Z_3—and measuring the respective input reflection coefficients. If Z_1 is a matched load, Z_2 is a short circuit and Z_3 is an open, the S parameters of Equations (6.235), (6.236), and (6.237) are found. Use SFG theory to derive these results.

Passive Components

7.1   Introduction

The purpose of this chapter is to introduce a wide variety of passive components. It is not possible to be comprehensive, as there is an enormous catalog of elements and scores of variations, and new concepts are introduced every year.

   At microwave frequencies, wavelengths are in the range of millimeters and centimeters, and distributed components can be constructed that have features with particular properties related to coupling, to traveling waves, and to storage of EM energy. In this chapter we consider many of the passive

elements useful in RF and microwave design. In many cases it is possible to develop lumped-element equivalents of the distributed elements by using the LC ladder model of a transmission line.

7.2 Lumped Elements

7.2.1 Integrated Lumped Passive Components

One of the factors that distinguishes RFICs from other RF and most microwave IC technology is the use of lumped elements rather than distributed elements. RF circuits require many passive components for matching networks, **RF chokes** for bias (i.e., inductors that block RF but provide a lossless DC connection), harmonic tuning, and to ensure stability at frequencies below the frequencies of operation. The primary lumped elements on RFICs are resistors, capacitors, and inductors. One of the other distinguishing features of RFICs is the use of differential signaling for nearly all of the analog signals on-chip and not only for the RF paths. This is necessary to overcome the limitations of silicon MOSFET-based circuitry, including significant substrate coupling.

7.2.2 On-Chip Capacitors

There are three primary forms of on-chip capacitor:

(a) Metal-dielectric-metal capacitor—using the interconnect metallization.

(b) Metal-dielectric-semiconductor capacitor—essentially a MOS transistor.

(c) Semiconductor junction capacitor—either the capacitance of reverse-biased pn junction or Schottky barrier.

In silicon technology it is common to refer to the first two capacitor types as **metal-oxide-metal** (**MOM**) capacitors (or **metal-insulator-metal** (**MIM**) capacitors) and as **metal-oxide-semiconductor** (**MOS**) capacitors. MOM capacitance can be realized as a parallel plate capacitance (see Figure 7-1(a)), but multiple levels of metallization can be used to increase the capacitance density. Relatively low capacitance values of up to $500 \text{ fF}/\mu\text{m}^2$ are available because of the large dielectric thicknesses of 0.5 to 1 μm between metal layers. This thickness is required in normal interconnect circumstances to minimize interconnect-to-interconnect coupling, so it is unlikely to change much. The distance between the bottom metallization layer to the substrate is comparable to the metal layer separation, and so the capacitance between the bottom plate of a capacitor and the substrate leads to significant capacitance—in the range of 10% to 30% of the metal-to-metal capacitance [109]. This series capacitor connection must be considered
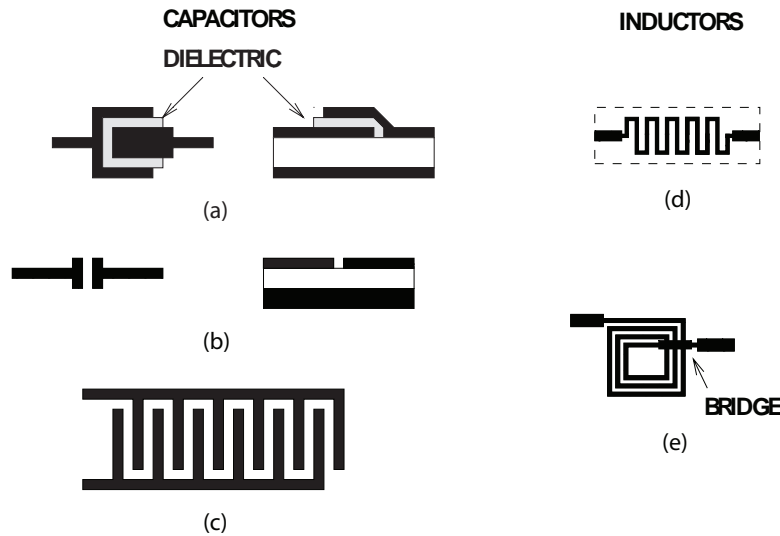
Figure 7-1 Monolithic lumped elements: (a) parallel plate capacitor; (b) gap capacitor; (c) interdigitated capacitor; (d) meander line inductor; and (e) spiral inductor.

in design.

An alternative MOM capacitance is available using lateral arrangements of interconnects on the same layer (see Figure 7-1(b)); that is, adjacent metal structures are separated by a small horizontal gap. Again, there are two distinct metal connections, and a smaller metal separation can be obtained using photolithography than that available using dielectric thickness. However the capacitance density is only increased by a factor of about three. Higher values can be obtained using the **interdigitated capacitor** (**IDC**) of Figure 7-1(c).

Both types of MOM capacitance, parallel plate and lateral, are geometrically defined, are voltage independent, have very low temperature coefficients, and have initial fabrication tolerances of 20% to 30%. At IF, and potentially at RF, many analog designs (e.g., active filters) use transconductance tuning to achieve frequency-response precision. In these cases, the capacitance tolerancing can be compensated for.

MOS capacitors use a MOS transistor with a parallel-plate capacitance between the gate of a MOS transistor and a heavily inverted channel. The drain and source are connected in this configuration and the separation between the conductors is thin, being the gate oxide thickness. This leads to high values of capacitance, although with weak voltage dependence. Junction capacitance is realized as the capacitance of a reverse-biased semiconductor junction. This capacitance can be quite large, but has a strong voltage dependence. This voltage dependence can be utilized to realize
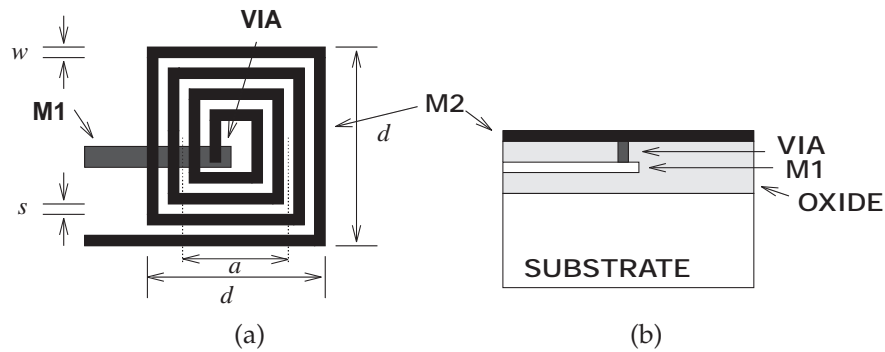
Figure 7-2   An on-chip spiral inductor: (a) plan view; and (b) side view.

tunable circuits (e.g., a voltage-controlled oscillator).

7.2.3   Planar Inductors

Inductors are important components in RF and microwave circuits. In addition to their role in matching networks, they are used to provide bias to active devices while effectively blocking RF signals from the bias circuitry. Inductors of up to 10 nH can be fabricated on-chip. Values above this consume too much die area, and either the entire inductance or the majority of a required inductance are fabricated off-chip, more often on an LTCC substrate or as part of the chip package. Bond wires can also be used to realize small inductances in the 0.5 to 1 nH range. The advantage of having a portion of a large inductance on-chip is reduced sensitivity to die attach (bondwire, etc.) connections.

An on-chip **spiral inductor**, the most common type, is shown in Figure 7-2 in both plan and side views. An approximate expression for the inductance of this structure was developed by Wheeler [110–112]:

$$L \approx \frac{9.4\mu_o n^2 a^2}{11d - 7a}, \tag{7.1}$$

where a is the mean radius of the spiral and n is the number of turns. This formula was derived for circular coils, but its accuracy for square spirals has been determined by Lee [109] to be within 5% of values derived using EM simulation. It is therefore a very useful formula in the early stages of design, but EM analysis is required to obtain the necessary accuracy and frequency dependence of the inductor in design. Alternative inductor designs are shown in Figures 7-1(d) and 7-1(e). The spiral inductor in Figure 7-1(d) uses an overpass or **airbridge** rather than the bottom layer metal of Figure 7-2. Small values of inductance can be realized by the meander line inductor of Figure 7-1(d).

Fields produced by a spiral inductor penetrate the substrate, and as a

ground plane is located at a relatively short distance, the eddy currents on the ground plane reduce the inductance that would otherwise be obtained. The eddy current in the ground conductor rotates in a direction opposite to that of the spiral itself (i.e., counterclockwise). As a result, the inductance of the image inductor in the ground is in the opposite direction to that produced by the spiral itself, with the consequent effect that the effective total inductance is reduced. By creating a broken, or perforated, conductor pattern, the ground inductance is largely eliminated [113], as the eddy currents cannot flow.

One of the major sources of loss for inductors on silicon is substrate loss due to the finite conductivity of the substrate and the resulting current flow. These induced currents follow a path under the conductors of the spiral and, just as with ground plane eddy currents, lower the inductance achieved. However, the resistance of the lines is unchanged. It is difficult to achieve Qs (the ratio of stored energy to energy dissipated per cycle) greater than about 10 above a few gigahertz. Parasitic capacitance both from the line to the ground and between spirals result in resonance of the on-chip inductance structure and hence limit the frequency of operation.

While many formulas and models have been presented for spiral, toroidal, and solenoidal inductors [114–116], the dependencies on a particular process mean that RF inductors must be modeled electromagnetically and modeled for a particular application. Models and design curves are generally provided by the vendor for a particular process.

7.2.4   Transmission Line Discontinuities

Interruptions of the magnetic or electric field create regions where magnetic energy or electric energy is stored. These interruptions are often caused by discontinuities on a transmission line. If the additional energy stored is predominantly magnetic, the discontinuity will introduce an inductance. If the additional energy stored is predominantly electric, the discontinuity will introduce a capacitance. Common microstrip discontinuities were considered in Section 4.12.2 on Page 233. Similar discontinuities occur with all transmission lines. In some cases, transmission line discontinuities introduce undesired parasitics, but they also provide an opportunity to introduce lumped-element effects without the need to attach lumped elements.

7.3   Terminations and Attenuators

7.3.1   Terminations

Terminations are used to completely absorb a forward-traveling wave and the defining characteristic is that the reflection coefficient of a termination is zero. If a transmission line has a resistive characteristic impedance $R_0 = Z_0$,

(a)   (b)
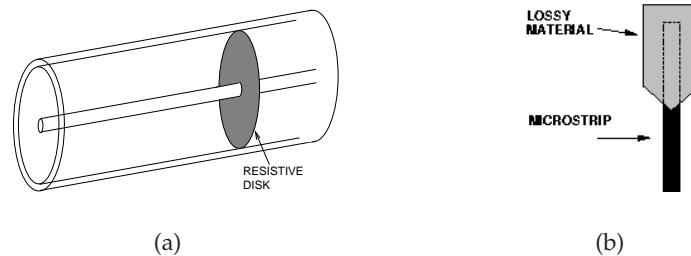
Figure 7-3 Terminations: (a) coaxial line resistive termination; (b) microstrip matched load.

then terminating the line in a resistance R_0 will fully absorb the forward-traveling wave and there will be no reflection. The line is said to be matched. At RF and microwave frequencies some refinements to this simple circuit connection are required. On a transmission line the energy is contained in the EM fields. For the coaxial line shown in Figure 7-3(a), a simple resistive connection between the inner and outer conductors would not terminate the fields and there would be some reflection. Instead, coaxial line terminations generally are comprised of a disk of resistive material. The total resistance of the disk from the inner to the outer conductor will be the characteristic resistance of the line, however, the resistive material is distributed and so creates a good termination of the fields guided by the coaxial line.

Terminations are a problem with microstrip as the characteristic impedance varies with frequency, is in general complex, and the vias that would be required if a lumped resistor was used has appreciable inductance at frequencies above a few gigahertz. A high-quality termination is realized using a section of loss line as shown in Figure 7-3(b). Here lossy material is deposited on top of an open-circuited microstrip line. This increases the loss of the line appreciably without significantly affecting the characteristic impedance of the line. If the length of the lossy line is sufficiently long the forward-traveling wave will be totally absorbed and there will be no reflection. Tapering the line, as shown in Figure 7-3(b), reduces the discontinuity between the microstrip line and the lossy line by ensuring that some of the power in the forward-traveling wave is dissipated before the maximum impact of the loss material occurs. Here a matched termination is achieved without the use of a via, and the termination automatically matches the varying characteristic impedance of the line. The same principle can be used with other transmission structures.

7.3.2 Attenuators

An attenuator is used to reduce the amplitude of a signal and it does this, in most cases, by absorbing power and without distorting the signal.

Table 7-1 IEEE standard symbols for attenuators [117].

Component	Symbol	Alternate
Attenuator, fixed		
Attenuator, balanced		
Attenuator, unbalanced		
Attenuator, variable		
Attenuator, continuously variable		
Attenuator, stepped variable		

Simultaneously the input and output of the attenuator are matched so there is ideally no reflections. They are used to reduce the level of one signal so that two correlated signals of different amplitude can be compared. Attenuators also reduce the effect of mismatch and provide isolation. An attenuator may be fixed, continuously variable, of discretely variable. The IEEE standard symbols for attenuators are shown in Figure 7-1.

When the attenuation is fixed an attenuator is commonly called a **pad**. Resistive pads can be used to minimize the effect of shorts and opens on the integrity of an RF circuit. An example of attenuator use in this situation is in a cable TV system where it is critical that the integrity of the system is not compromised by a consumer disconnecting appliances from a cable outlet. Sometimes the attenuator is combined with a splitter.

Balanced and unbalance resistive pads are shown in Figures 7-4 and 7-5 together with their design equations. The attenuators in Figure 7-4 are T or Tee attenuators, where Z_{01} is the system impedance to the left of the pad and Z_{02} is the system impedance to the right of the pad. The defining characteristic is that the reflection coefficient looking into the pad from the left is zero when referred to Z_{01}. Similarly the reflection coefficient looking into the right of the pad is zero with respect to Z_{02}. Defining K as the attenuation factor,

$$K = \frac{\text{Power in}}{\text{Power out}}. \tag{7.12}$$

T Attenuator

Unbalanced   Balanced

(a)   (b)
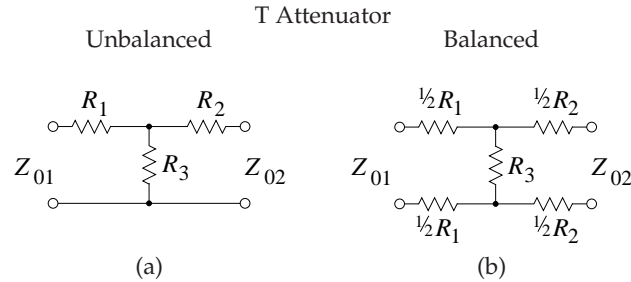
$$R_1 = \frac{Z_{01}(K+1) - 2\sqrt{K Z_{01} Z_{02}}}{K-1} \qquad (7.2)$$

$$R_2 = \frac{Z_{02}(K+1) - 2\sqrt{K Z_{01} Z_{02}}}{K-1} \qquad (7.3)$$

$$R_3 = \frac{2\sqrt{K Z_{01} Z_{02}}}{K-1} \qquad (7.4)$$

If $Z_{01} = Z_{02} = Z_0$, then

$$R_1 = R_2 = Z_0 \left(\frac{\sqrt{K} - 1}{\sqrt{K} + 1} \right) \qquad (7.5)$$

$$R_3 = \frac{2 Z_0 \sqrt{K}}{K-1} \qquad (7.6)$$

Figure 7-4 T (Tee) attenuator: (a) unbalanced T pad; and (b) balanced T pad.

In decibels,

$$K|_{\mathrm{dB}} = 10 \log_{10} K = (\text{Power in})|_{\mathrm{dBm}} - (\text{Power out})|_{\mathrm{dBm}}. \qquad (7.13)$$

If the left and right system impedances are different there is a minimum attenuation factor that can be achieved:

$$K_{\mathrm{MIN}} = \frac{2 Z_{01}}{Z_{02}} - 1 + 2 \sqrt{\frac{Z_{01}}{Z_{02}} \left(\frac{Z_{01}}{Z_{02}} - 1 \right)}. \qquad (7.14)$$

This limitation comes from the simultaneous requirement that the pad be matched. If there is a single system impedance $Z_0 = Z_{01} = Z_{02}$, then $K_{\mathrm{MIN}} = 1$ and so any value of attenuation can be obtained.

   Lumped attenuators are useful up to a few gigahertz above which the size of resistive elements becomes large compared to a wavelength. Also, for planar circuits, vias are required and these are undesirable from a manufacturing standpoint, and electrically they have a small inductance. Fortunately attenuators can be realized using a lossy section of transmission line as shown in Figure 7-6. Here lossy material results in a section of line with a high attenuation constant. Generally the lossy material has little effect on the characteristic impedance of the line so that there is little reflection at

Pi Attenuator

Unbalanced   Balanced

(a)   (b)

$$R_1 = \frac{Z_{01}(K-1)\sqrt{Z_{02}}}{(K+1)\sqrt{Z_{02}} - 2\sqrt{KZ_{01}}} \qquad (7.7)$$

$$R_2 = \frac{Z_{02}(K-1)\sqrt{Z_{01}}}{(K+1)\sqrt{Z_{01}} - 2\sqrt{KZ_{02}}} \qquad (7.8)$$

$$R_3 = \frac{(K-1)}{2}\sqrt{\frac{Z_{01}Z_{02}}{K}} \qquad (7.9)$$
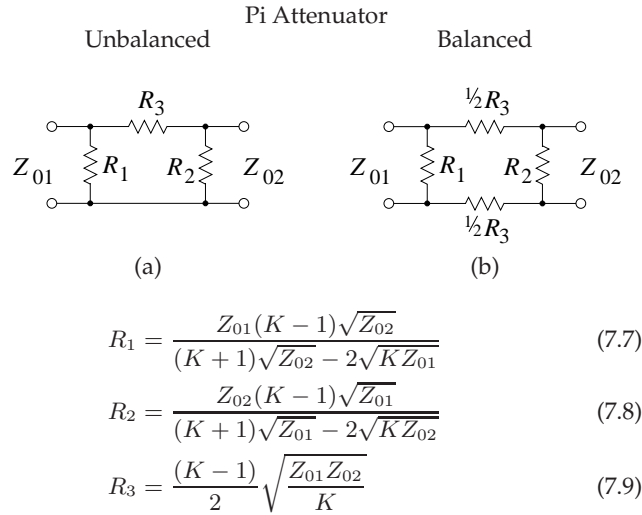
If $Z_{01} = Z_{02} = Z_0$, then

$$R_1 = R_2 = Z_0 \left(\frac{\sqrt{K}+1}{\sqrt{K}-1} \right) \qquad (7.10)$$

$$R_3 = \frac{Z_0(K-1)}{2\sqrt{K}} \qquad (7.11)$$

Figure 7-5   Pi (Π) attenuator: (a) unbalanced pad; and (b) balanced pad.

(a)   (b)
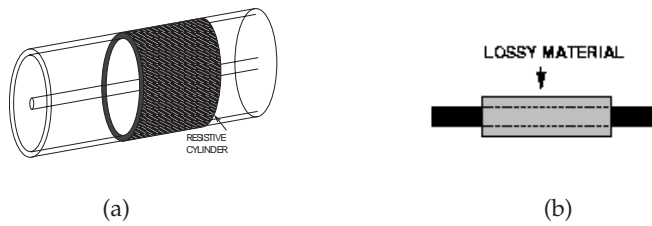
Figure 7-6   Distributed attenuators: (a) coaxial attenuator; and (b) microstrip attenuator.

the input and output of the distributed attenuator. Distributed attenuators can be used at higher frequencies and can be realized with any transmission line structure. The central concept is having a distributed lossy element.

An example of the use of attenuators in the combination of the output of two sources is shown in Figure 7-7. This is a common situation in measurements where the output of two sources is to be combined. The attenuators reduce the level of the signal presented to the output of one source by the other. If the level of the second signal is high, most sources
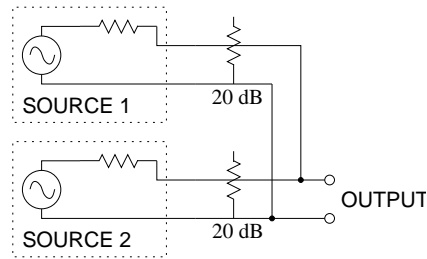
Figure 7-7 The use of attenuators to isolate the outputs of two sources that are combined.

would produce nonlinear distortion, including nonlinear mixing products. Other components such as circulators and hybrids can also be used to reduce unwanted signal levels, but pads are commonly used even then.

EXAMPLE 7. 1   Pad Design

Design a unbalanced 20 dB pad in a 75 Ω system.

Solution:

There are two possible designs using resistive pads. These are the unbalanced Tee and Pi pads shown in Figures 7-4 and 7-5. The Tee design will be chosen. The K factor is

$$K = 10^{(K|_{\mathrm{dB}}/10)} = 10^{(20/10)} = 100. \tag{7.15}$$

Since $Z_{01} = Z_{02} = 75$ Ω, Equations (7.5) and (7.6) become

$$R_1 = R_2 = 75 \left(\frac{\sqrt{100} - 1}{\sqrt{100} + 1} \right) = 75 \left(\frac{9}{11} \right) = 61.4 \ \Omega \tag{7.16}$$

$$R_3 = \frac{2 \cdot 75\sqrt{100}}{100 - 1} = 150\sqrt{\frac{10}{99}} = 15.2\Omega \ . \tag{7.17}$$

The final design is
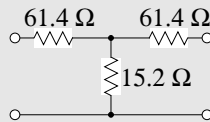
7.4   Magnetic Transformers

In this section, specific realizations of hybrids and baluns will be presented first using magnetic transformers, then transmission lines, and finally using Ls and Cs inspired by lumped-element approximations of transmission

$n_1 : n_2$

(a)

(b)

$\equiv$

(c)

$-n_1 : n_2$

$\equiv$

$n_1 : n_2$

(d)

I_1 $1:n$ I_2

$Z_{in} \rightarrow$ V_1 V_2 R

(e)
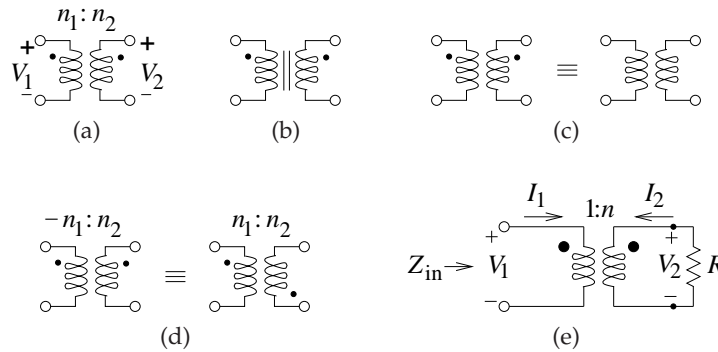
Figure 7-8 Magnetic transformers: (a) a transformer as two magnetically coupled windings with n_1 windings on the primary (on the left) and n_2 windings on the secondary (on the right), the dots indicate magnetic polarity so that the voltages V_1 and V_2 have the same sign; (b) a magnetic transformer with a magnetic core identified by the two long vertical bars; (c) identical representations of a magnetic transformer with the magnetic polarity implied for the transformer on the right; (d) two equivalent representations of a transformer having opposite magnetic polarities (an **inverting transformer**); and (e) and a magnetic transformer circuit.

lines.

A magnetic transformer (see Figure 7-8) couples, via a magnetic field, the current in one wire to current in another. The effect is amplified using coils of wires and using a core of magnetic material (material with high permeability) to create greater magnetic flux density. When coils are used, the symbol shown in Figure 7-8(a) is used with one of the windings called the primary winding and the other called the secondary. If there is a **magnetic core** then the symbol shown in Figure 7-8(b) is used with the vertical lines indicating the core. However, even if there is a core, the simpler transformer symbol in Figure 7-8(a) is more commonly used. Magnetic cores are useful up to several hundreds of megahertz and rely on the alignment of magnetic dipoles in the core material. Above a few hundred megahertz the magnetic dipoles cannot react quickly enough and so the core looks like an open circuit to magnetic flux. Thus the core is not useful for magnetically coupling signals above a few hundred megahertz. As mentioned, the dots above the coils in Figure 7-8(a) indicate the polarity of the magnetic flux with respect to the currents in the coils so that, as shown, V_1 and V_2 will have the same sign. Even if the magnetic polarity is not specifically shown it is implied (see Figure 7-8(c)). There are two ways of showing inversion of the magnetic polarity, as shown in Figure 7-8(d), where a negative number of windings indicates opposite magnetic polarity.

The interest in magnetic transformers is, first, because it is a hybrid in its own right. Also, we can realize configurations of magnetic transformers

using coupled transmission lines and extend their operation to hundreds of gigahertz. The transformer is easy to conceptualize so it is convenient to first develop circuits using the transformer and then translate them to transmission line form. That is, in "back-of-the-envelope" microwave design, transformers can be used to indicate coupling, with the details of the coupling left until later when the electrical design is translated into a physical design. Restrictions must be followed as not all transformer configurations can be translated this way. Also transformers consisting of coils of wire, and perhaps ferrite cores to enhance magnetic coupling, can rarely be used above 1 GHz because of losses. Practicalities will be considered later in this chapter.

The following notation is used with magnetic transformer:

L_1, L_2: self-inductances of the two coils

M: mutual inductance

k: coupling coefficients,

where the coupling factor is

$$k = \frac{M}{\sqrt{L_1 L_2}}. \tag{7.18}$$

Referring to Figure 7-8(e), the voltage transformer ratio is

$$V_2 = n V_1, \tag{7.19}$$

where n is the ratio of the number of secondary to primary windings. An ideal transformer has "perfect coupling" indicated by $k = 1$, and the self inductances are proportional to the number of windings, so

$$\frac{V_2}{V_1} = \sqrt{\frac{L_2}{L_1}}. \tag{7.20}$$

The general equation relating the currents of the circuit of Figure 7-8 is

$$R I_2 + \jmath \omega L_2 I_2 + \jmath \omega M I_1 = 0, \tag{7.21}$$

and so

$$\frac{I_1}{I_2} = -\frac{R + \jmath \omega L_2}{\jmath \omega M}. \tag{7.22}$$

If $R \ll \omega L_2$, then the current transformer ratio is

$$\frac{I_1}{I_2} \approx -\frac{L_2}{M} = -\sqrt{\frac{L_2}{L_1}}. \tag{7.23}$$

Notice that combining Equations (7.20) and (7.23) leads to calculation of the transforming effect on impedance. On the coil 1 side the input impedance is

$$Z_{\text{in}} = \frac{V_1}{I_1} = -\frac{V_2}{I_2}\left(\frac{L_1}{L_2}\right) = R\frac{L_1}{L_2}. \tag{7.24}$$
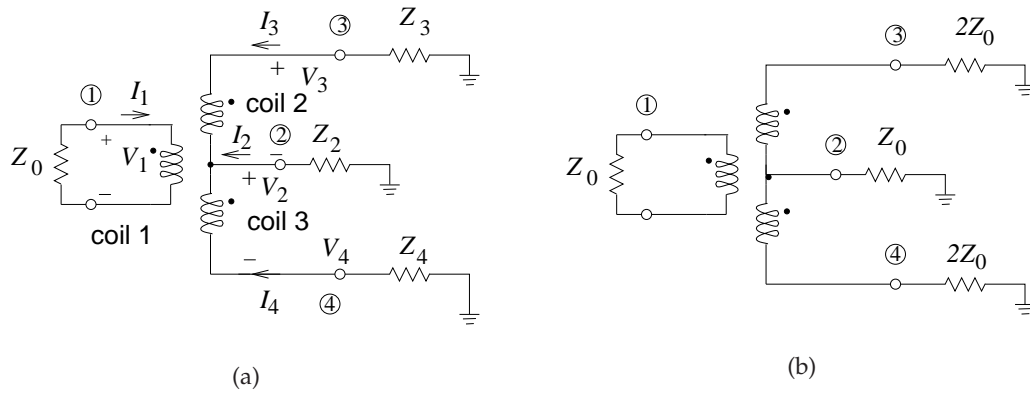
Figure 7-9 Magnetic transformer hybrid with each coil having the same number of windings: (a) showing general loading impedances, and (b) optimum loading to function as a $180°$ hybrid with Port 2 isolated and the input at Port 1.

In practice, however, there is always some magnetic field leakage—not all of the magnetic field created by the current in Coil 1 goes through (or links) Coil 2—and so $k < 1$. Then from Equations (7.20)–(7.24)

$$V_1 = \jmath\omega L_1 I_1 + \jmath\omega M I_2 \tag{7.25}$$
$$0 = R I_2 + \jmath\omega L_2 I_2 + \jmath\omega M I_1. \tag{7.26}$$

Again, assuming that $R \ll \omega L_2$, a modified expression for the input impedance is obtained that accounts for nonideal coupling:

$$Z_{\text{in}} = R\frac{L_1}{L_2} + \jmath\omega L_1(1 - k^2). \tag{7.27}$$

Imperfect coupling, $k < 1$, causes the input impedance to be reactive and this limits the bandwidth of the transformer. Stray capacitance is another factor that impacts the bandwidth of the transformer. The key insight, however, is that a transformer implements resistance transformation. It can also serve as a building block in developing the electrical design of many types of hybrid circuits.

From an RF perspective, one of the main functions of the basic transformer is as an impedance transforming balun. With a tap on the secondary winding, hybrid and alternative balun structures can be obtained, as shown in Figure 7-9 as a hybrid and in Figure 7-10 as a balun.

In the hybrid and balun circuits of Figures 7-9 and 7-10 the number of windings of each coil (there are three in each structure) is the same. The impedance levels given are those required for maximum power transfer and directly indicate the impedance transformations of the structures.
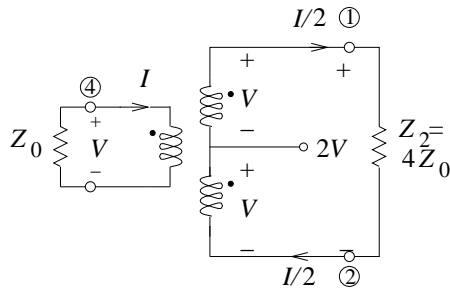
Figure 7-10 Operation of the magnetic transformer of Figure 7-9 as a balun.

Considering Figure 7-10 and equating powers,

$$\frac{1}{2}\frac{(2V)^2}{Z_2} = \frac{1}{2}\frac{V^2}{Z_0} \tag{7.28}$$

$$\frac{2}{Z_2} = \frac{1}{2}Z_0 \tag{7.29}$$

$$Z_2 = 4Z_0. \tag{7.30}$$

EXAMPLE 7. 2　　**Magnetic Transformer**

Consider the magnetic transformer hybrid of Figure 7-9. Determine what type of hybrid this is and calculate the impedance transformations. Assume ideal coupling ($k = 1$).

Solution:

Since there is ideal coupling here and each coil has the same number of windings,

$$(\mathbf{V}_3 - \mathbf{V}_2) = \mathbf{V}_1 \tag{7.31}$$
$$(\mathbf{V}_2 - \mathbf{V}_4) = \mathbf{V}_1. \tag{7.32}$$

The current levels in the transformer depend on the attached circuitry. For this circuit to function as a hybrid with Port 2 isolated, the current I_2 must be zero so that I_3 and I_4 are equal in magnitude but $180°$ out of phase. The loading at Ports 3 and 4 must be the same. Now $V_2 = 0$, since $I_2 = 0$, and so Equations (7.31) and (7.32) become

$$\mathbf{V}_3 = \mathbf{V}_1 \tag{7.33}$$
$$\mathbf{V}_4 = -\mathbf{V}_1, \tag{7.34}$$

so this circuit is a $180°$ hybrid.

To determine loading conditions at Ports 2 through 4 the following transformer equation is used

$$V_3 = \jmath\omega L_2 I_3 + \jmath\omega M I_1 - \jmath\omega L_3 I_4 \tag{7.35}$$

where L_2 is the inductance of Coil 2 and M is the mutual inductance. Now $I_1 = -V_1/Z_0$, and $L_2 = L_3 = M$ since coupling is ideal, and $V_3 = V_1$. So Equation (7.35) becomes

$$V_3 = \jmath\omega M(2I_3 - V_3/Z_0) \tag{7.36}$$

or

$$V_3(1 + \jmath\omega M/Z_0) = \jmath 2\omega M I_3. \tag{7.37}$$

If $\jmath\omega M \gg Z_0$, this reduces to

$$Z_3 = V_3/I_3 = 2Z_0. \tag{7.38}$$

From symmetry, $Z_4 = Z_3$. Also, $I_3 = -I_4 = I_1/2$.

   This above result can also be argued from maximum power transfer considerations. The argument is as follows. An impedance Z_0 is attached to coil 1 and this is an indication that the Thevenin equivalent circuit attached to coil 1 has a Thevenin equivalent impedance of Z_0. Maximum power transfer to the transformer through the coil requires that the input impedance be Z_0. In the ideal hybrid operation the power is split evenly between the power delivered to the loads at Ports 3 and 4, since $V_1 = (V_3 - V_2) = (V_2 - V_4)$ and the power delivered to coil 1 is $V_1^2/(2Z_0)$. The power delivered to Z_3 (and Z_4) is $(V_3 - V_2)^2(2Z_3) = V_1^2/2Z_3 = V_1^2/(4Z_0)$. That is, $Z_3 = 2Z_0 = Z_4$.

The problem is not yet finished, as Z_2 must be determined. For hybrid operation, a signal applied to Port 2 should not have a response at Port 1. So the current at Port 2, I_2, should be split between coils 2 and 3 so that $I_3 = -I_2/2 = I_4$. Thus

$$V_4 = I_2 Z_0/2 = V_3. \tag{7.39}$$

Now

$$V_3 - V_2 = V_1 = 0 = V_2 - V_4, \tag{7.40}$$

and so

$$V_2 = V_4 = -I_2 Z_0 \tag{7.41}$$

or

$$Z_2 = V_2/I_2 = Z_0. \tag{7.42}$$

   The final $180°$ hybrid circuit is shown in Figure 7-9(b) with the loading conditions for matched operation as a hybrid.

   In the example above it was seen that the number of windings in the coils are the same so that the current in Coils 2 and 3 are half that in Coil 1. The general rule is that with an ideal transformer the sum of the amp-turns around the magnetic circuit must be zero. The precise way the sum is calculated depends on the direction of the windings indicated by the "dot" convention. A generalization of the rule for the transformer shown in Figure 7-9 is

$$n_1 I_1 - n_2 I_2 - n_3 I_3 = 0, \tag{7.43}$$

where n_j is the number of windings of coil j with current I_j. The example serves to illustrate the type of thinking behind the development of many RF circuits. The emphasis on maximum power transfer provided an alternative, simpler start to the solution of the problem. Yes, this can be difficult—people write papers on ways of analyzing a new circuit element. This is

why the technical literature is an important source of information for RF
and microwave engineers.

7.5   Hybrids

Previously, coupled transmission lines were examined and it was shown
how the coupling phenomenon can be exploited to realize a directional
coupler. In a sense, directional couplers are transformers that enable a
portion of a signal on one line to be coupled off. Hybrids are transformers
that combine or divide microwave signals among a number of inputs and
outputs. They can be implemented using distributed or lumped elements
and are not restricted to RF and microwave applications. For example, a
magnetic transformer is an example of a hybrid and is used in some audio
equipment to combine the output of two or more transistor stages and
to obtain appropriate impedance levels. Specific implementations will be
considered later in this chapter, but for now the focus is on their idealized
characteristics.

A hybrid is a special type of four-port junction with the property that
if a signal is applied at any port, it emerges from two of the other ports
at half power, while there is no signal at the fourth or isolated port. The
two outputs have specific phase relationships and all ports are matched.
Only two fundamental types of hybrids are used: the $180°$ and $90°$ hybrids.
A hybrid is a type of directional coupler, although the term directional
coupler, or just coupler, is most commonly used to refer to devices where
a small fraction of the power of an input signal is sampled. Also, a balun is a
special type of hybrid. The symbols commonly used for hybrids are shown
in Figure 7-11.

The ideal **$90°$ hybrid**, or **quadrature hybrid**, shown in Figure 7-11(c), has
the scattering parameters

$$S_{90°} = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & -\jmath & 1 & 0 \\ -\jmath & 0 & 0 & 1 \\ 1 & 0 & 0 & -\jmath \\ 0 & 1 & -\jmath & 0 \end{bmatrix}. \tag{7.44}$$

The $90°$ phase difference between the through and coupled ports is
indicated by $-\jmath$. The actual phase shift between the input and output ports
depends on the specific hybrid implementation. Another quadrature hybrid
could have the parameters

$$S_{90°} = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & \jmath & 1 & 0 \\ \jmath & 0 & 0 & 1 \\ 1 & 0 & 0 & \jmath \\ 0 & 1 & \jmath & 0 \end{bmatrix}. \tag{7.45}$$

Recall the requirement of quadrature modulators that the local oscillator be
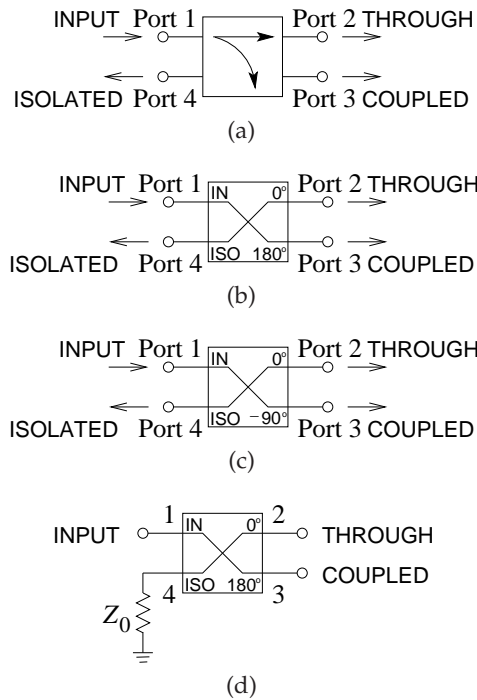
Figure 7-11 Commonly used symbols for hybrids: (a) through and coupled ports; (b) $180°$ hybrid; (c) $90°$ hybrid; and (d) hybrid with the isolated port terminated in a matched load.

supplied as two components equal in magnitude but $90°$ out of phase. A quadrature hybrid is just the circuit that can do this.

To convince yourself that Equation (7.45) describes a network that splits the power, consider the power flow implied by Equation (7.45). The fraction of power transmitted from Port i to Port j is described by $|S_{ji}|^2$. Consider the power that enters Port 1. No power is reflected for an ideal hybrid, as the input at Port 1 is matched, and so $S_{11} = 0$. Port 4 should be isolated so no power will come out of Port 4, and so $S_{41} = 0$. The power should be split between Ports 2 and 3, and these should be equal to half the power entering Port 1. From Equation (7.45),

$$|S_{21}|^2 = \left(\frac{1}{\sqrt{2}}|\jmath|\right)^2 = \frac{1}{2} \quad \text{and} \quad |S_{31}|^2 = \left(\frac{1}{\sqrt{2}}|1|\right)^2 = \frac{1}{2}. \qquad (7.46)$$

The power entering Port 1 is split in half, with half going to Port 2 and half to Port 3.

The scattering parameters of the **$180°$ hybrid**, shown in Figure 7-11(b),

are

$$S_{180°} = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & 1 \\ -1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix},$$ (7.47)

and this defines the operation of the hybrid. In terms of the root power waves a and b, the outputs at the ports are

$$b_1 = (a_2 - a_3)/\sqrt{2}$$ (7.48)
$$b_2 = (a_1 + a_4)/\sqrt{2}$$ (7.49)
$$b_3 = (-a_1 + a_4)/\sqrt{2}$$ (7.50)
$$b_4 = (a_2 + a_3)/\sqrt{2} .$$ (7.51)

Imperfections in fabricating the hybrid will result in nonzero scattering parameters so that it is best to terminate the isolated port of the hybrid, as shown in Figure 7-11(d), to ensure that there is no reflected signal, no a_4, entering the hybrid at the isolated port. Equations (7.48)–(7.51) show that ports are interchangeable. That is, if a signal is applied to Port 3 then it is split between Ports 1 and 4, and Port 2 becomes the isolated port.

A hybrid can be used in applications other than splitting the input signal into a through and a coupled component. An example of another use is in a system that combines or compares two signals, as in Figure 7-12. Here a 180° hybrid with the scattering parameters of Equation (7.47) is shown, with its equivalence as a combiner and a comparator in Figure 7-12. With a signal $x(t)$ applied to Port 2 in Figure 7-11(b), and another signal $y(t)$ applied to Port 3, the output at Port 4 is the sum signal, $x(t) + y(t)$, and the output at Port 1 is the difference signal, $x(t) - y(t)$. More common names for the output ports (when the 180° hybrid is used as a combiner) are to call Port 4 the sigma or sum port often designated using the symbol Σ. Port 1 is called the difference or delta port, $\triangle$. Notice that if a signal is applied to the $\triangle$ port it will generate out-of-phase outputs.

Since a 180° phase difference can be obtained at the output ports, a 180° hybrid can be used in the same manner as a balun. There can be a subtle difference between a balun and a hybrid used as a balun. In many applications, say in driving an antenna, the balanced output of a balun is just one port. However, the nominally balanced output of a hybrid comprises two ports, with each normally referenced to ground. So a hybrid used as a balun has a pseudo-balanced output port.

7.6 Balun

A balun [118, 119] is a structure that joins balanced and unbalanced circuits. The word itself (balun) is a contraction of the phrase balanced-to-unbalanced transformer. It is a key component of RF and microwave
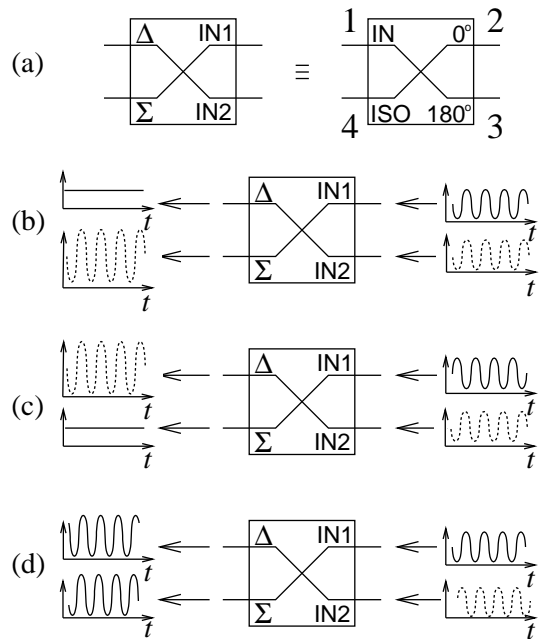
Figure 7-12 A $180°$ hybrid depicted as a combiner (with output at the sum port, Σ) and as a comparator (with output at the difference port, $\triangle$): (a) equivalence between a $180°$ hybrid and a sum-and-difference hybrid; (b) the two inputs in phase; (c) the two inputs $180°$ out of phase; and (d) the two inputs $90°$ out of phase.

communications systems. They are used in balanced circuits, such as double-balanced mixers, push-pull amplifiers, and frequency doublers [120]. Another application of a balun is in a system using RFICs, where a balun transforms the differential outputs of an RFIC to unbalanced microwave circuitry. There are many types of baluns [121, 122] with the Marchand balun [118, 120, 123–125] being the most popular at microwave frequencies, as it can be conveniently realized in planar or coaxial forms [118].

Baluns can be realized using a directional coupler with a coupling factor of 3 dB. One of the terminals of the input pair of terminals is grounded (and so the circuit is unbalanced), whereas the output terminals can be floating or balanced. With these connections the transformer is a balun. Examples of where a balun is required is in feeding an antenna, in separating the transmitted and received signals in a telephone handset, and in interfacing the differential ports of an RFIC to single-ended (i.e., unbalanced) elements. A general representation of a balun is shown in Figure 7-13. If an even-mode excitation is applied to the balanced port (i.e., the signals at terminals B and C have the same voltage to ground) no power is transmitted to the input,
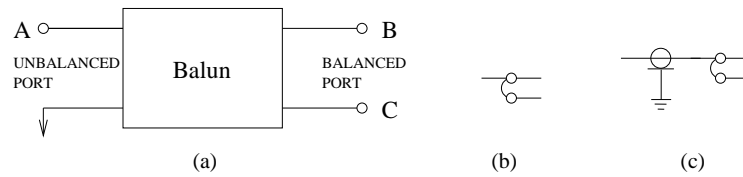
Figure 7-13 A balun: (a) as a two-port with four terminals; (b) IEEE standard schematic symbol for a balun [117]; and (c) an (unbalanced) coaxial cable driving a dipole antenna through a balun.
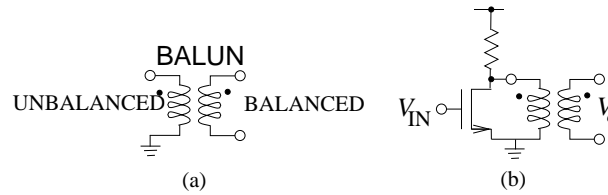
Figure 7-14 Balun: (a) schematic representation as a transformer showing unbalanced and balanced ports; and (b) connected to a single-ended unbalanced amplifier yielding a balanced output.

and the even-mode output port impedance (at terminals B and C) is either an open or short circuit (most often the balun is designed so that it is an open circuit). If an odd mode is applied at the balanced ports, then all of the power will be presented at the unbalanced port. In the reverse operation, a signal applied to the unbalanced port will be split into two equal components (but opposite in phase) at Ports B and C.

The schematic of a balun is shown in Figure 7-14(a) as a transformer connected with one terminal of the unbalanced port grounded. The second port is floating and is not referenced to ground. An example of the use of a balun is shown in Figure 7-14(b), where the amplifier is called a single-ended amplifier and its output is unbalanced, being referred to ground. A balun transitions from the unbalanced input to a balanced output.

7.6.1 Marchand Balun

The most common form of microwave balun is the Marchand balun [124] shown in symbolic form in Figure 7-15(a). An implementation of the Marchand balun using coaxial transmission lines is shown in Figure 7-15(b), [126], and an implementation in planar form using coupled parallel lines is shown in Figure 7-16(a) [123]. Considering the planar form in Figure 7-16, The balun has two single-ended ports with one being identified on the left of Figure 7-16(a) and the other on the right with terminal x. The balun also has a balanced port with characteristic impedance R_B.
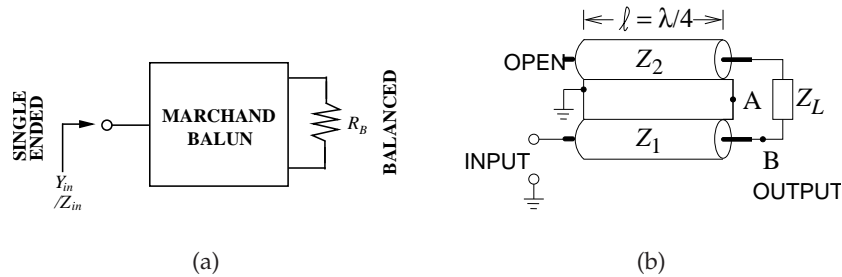
Figure 7-15   Marchand balun: (a) symbolic form; and (b) original coaxial form of the Marchand balun.
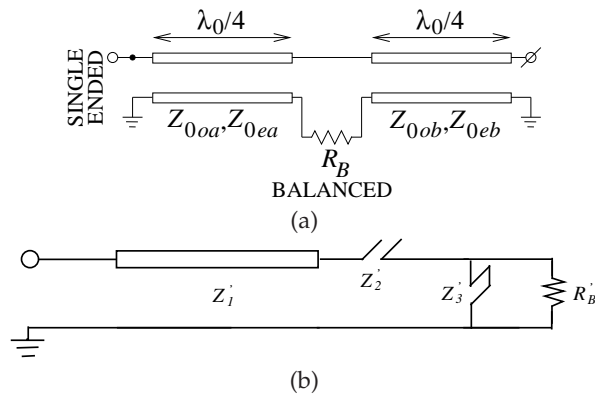
Figure 7-16 Marchand balun: (a) physical layout with balanced load impedance R_B; and (b) S-plane highpass prototype.

The model of the Marchand baluns shown in Figures 7-15(b) and 7-16(a) is the highpass model shown in Figure 7-16(b).[1]

The drawback of the conventional Marchand balun is that the center frequency of the balun is actually the resonant frequency of the transmission line resonators forming the balun. The overall size of a conventional Marchand balun is determined by the transmission line sections, which are one-quarter wavelength long at the balun center frequency. Network synthesis can lead to miniaturized baluns by shifting the quarter-wavelength frequency up while maintaining the balun center frequency. This is presented in Section 13.2.

---

[1]   In network synthesis terminology, Figure 7-16(b) is the single-ended S-plane highpass prototype, where S is Richards' transformation described in Section 10.11. Network synthesis of more advanced miniaturized Marchand baluns is presented in Sections 11.9.3, 13.2, and 11.9.
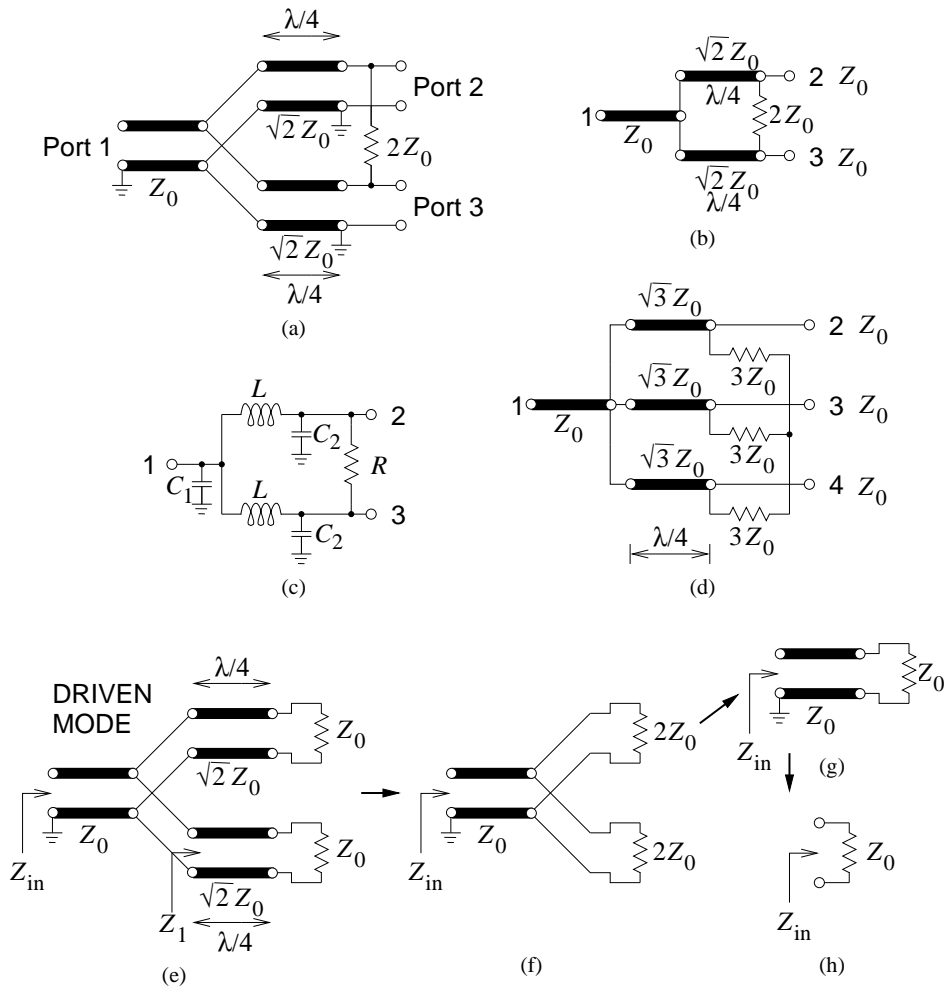
Figure 7-17 Wilkinson combiner and divider: (a) two-way Wilkinson combiner/divider with Port 1 being the combined signal and Ports 2 and 3 being the divided signals; (b) less cluttered representation of the two-way combiner; (c) three-way Wilkinson combiner/divider with Port 1 being the combined signal and Ports 2, 3, and 4 being the divided signals; (d) lumped-element implementation of a two-way combiner; and (e)–(h) steps in the derivation of the input impedance of a Wilkinson divider.

7.7 Wilkinson Combiner and Divider

The Wilkinson divider can be used as a combiner or divider which divides input power among output ports [127]. Figure 7-17(a) is a two-way divider that splits the power equally between the output ports. Power applied at Port 1 is split between Ports 2 and 3. A particular insight that Wilkinson

brought was the resistor between the output ports and this acts to suppress an unbalance between the output signals. If the division is exact, -3 dB of the input power appears at each of Ports 2 and 3, and the phasing matches, the signals are balanced, and no current will flow in the resistor. The circuit works less well as a general-purpose combiner. Ideally power entering Ports 2 and 3 would combine losslessly and appear at Port 1. However, if the signals are not identical, power will be absorbed in the resistor. A typical application is to combine the power at the output of two matched transistors where the amplitude and the phase of the signals can be expected to be closely matched. The bandwidth of the divider is limited by the one-quarter wavelength long lines. However, the bandwidth is relatively large, approaching $\pm 50\%$ [127]. Arbitrary power ratios can also be obtained [128–130].

The operation of the Wilkinson divider can be seen by deriving the input impedance of the two-way Wilkinson divider driven at Port 1 (see Figure 7-17(e)). Since the Wilkinson divider is driven the signals at Ports 2 and 3 will be identical, so it is though the $2Z_0$ resistor in the Wilkinson divider (Figure 7-17(a)) is not there. The input impedance of one of the quarter-wavelength long section is

$$Z_1 = \frac{\left(\sqrt{2}Z_0\right)^2}{Z_0},\tag{7.52}$$

and so the Wilkinson model reduces to that in Figure 7-17(f). The $2Z_0$ resistors are in parallel, resulting in the further model reductions in Figures 7-17(g) and 7-17(h). A similar analysis will show that Ports 2 and 3 are matched (see an impedance of $50\ \Omega$). Treating one-quarter wavelength long transmission line sections as an impedance inverter is key to understanding many distributed passive components. The S parameters of the two-way Wilkinson power divider with an equal split of the output power are

$$\mathbf{S} = \begin{bmatrix} 0 & -\jmath/\sqrt{2} & -\jmath/\sqrt{2} \\ -\jmath/\sqrt{2} & 0 & 0 \\ -\jmath/\sqrt{2} & 0 & 0 \end{bmatrix}.\tag{7.53}$$

Figure 7-17(b) is a more compact representation of the two-way Wilkinson divider which leads to the three-way Wilkinson divider shown in Figure 7-17(d). This pattern can be repeated to produce N-way power combining. The lumped-element version of the Wilkinson divider shown in Figure 7-17(c) is based on the LC model of a quarter-wavelength long transmission line segment. Using only a single element results in a relatively simple circuit. With a $50\ \Omega$ system impedance and center frequency of 400 MHz, the elements of the lumped element are $L = 28.13$ nH, $C_1 = 5.627$ pF, $C_2 = 5.627$ pF, and $R = 100\ \Omega$. Improved bandwidth of the divider can be obtained using the network synthesis techniques described in Chapters 10 and 13.

Figure 7-18(a) is the layout of a direct microstrip realization of a Wilkinson
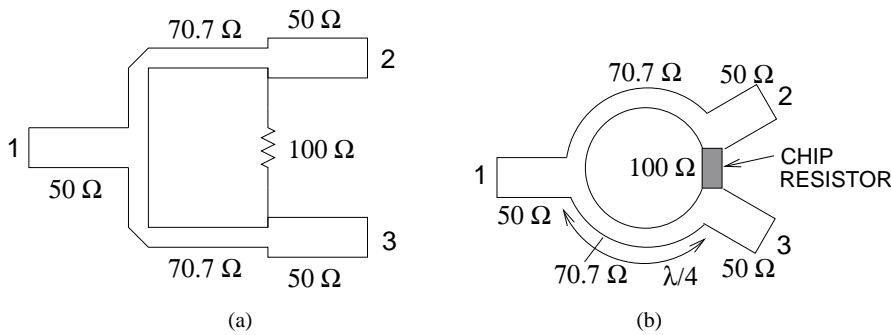
Figure 7-18 Wilkinson combiner and divider: (a) microstrip realization; and (b) higher performance microstrip implementation.
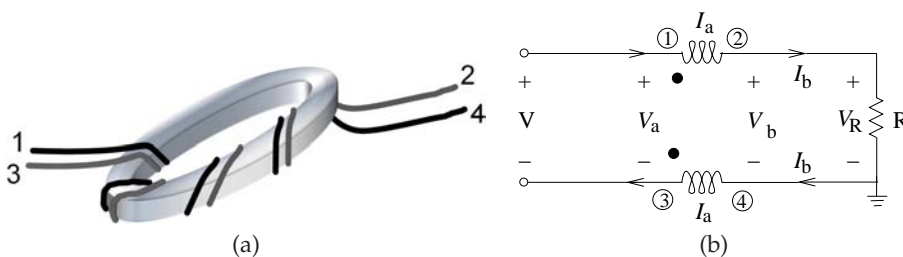
Figure 7-19 A broadband RF balun as coupled lines wound around a ferrite core: (a) physical realization; and (b) equivalent circuit using a wire-wound transformer.

divider. The obvious problem is how to incorporate the resistor. As long as the resistor is placed symmetrically this is not as severe a problem as it would initially seem, as power is not dissipated in the resistor unless there is an imbalance. A higher-performance microstrip layout is shown in Figure 7-18(b), where the transmission lines are curved to bring the output ports near each other so that a chip resistor can be used.

7.8 Transmission Line Transformer

One of the challenges in RF engineering is achieving broadband operation of transformer structures from megahertz up to several gigahertz. In this section, several structures are presented that operate as magnetic transformers at frequencies below several hundred megahertz, but as coupled transmission line structures at high frequencies. A transformer that achieves this is shown in Figure 7-19(a).

At high frequencies the stray capacitance between the windings becomes part of the transmission line capacitance. At several hundred megahertz the ferrite core does not respond to the signal and the magnetic circuit

through the core appears as an open circuit. Then, since the transformer is not operating as a magnetic transformer anymore, magnetic leakage is inconsequential. Also, at high frequencies the wire transmission lines are closely coupled. Transformers with bandwidths from megahertz to several gigahertz can be realized in the same manner. They find use in cable systems and instrumentation.

7.8.1 Transmission Line Transformer as a Balun

The schematic of a broadband 1:1 RF balun is shown in Figure 7-19(b) and a realization of it is shown in Figure 7-19(a). The 1:1 designation indicates that there is no impedance transformation. The circuit equations describing this balun are

$$V_a = V_b \cos(\beta\ell) + \jmath I_b Z_o \sin(\beta\ell) \tag{7.54}$$

$$I_a = I_b \cos(\beta\ell) + \jmath \frac{V_b}{Z_o} \sin(\beta\ell) \tag{7.55}$$

and

$$V_b = I_b R. \tag{7.56}$$

The aim in the following is the development of a design equation that describes the essential properties of the structure. Substituting Equation (7.56) in Equations (7.55) and (7.54) leads to

$$V_a = V_b \left[\cos(\beta\ell) + \jmath \frac{Z_o}{R} \sin(\beta\ell) \right] \tag{7.57}$$

$$I_a = \frac{V_b}{R} \left[\cos(\beta\ell) + \jmath \frac{R}{Z_o} \sin(\beta\ell) \right]. \tag{7.58}$$

Choosing $Z_o = R$ yields

$$V = V_a = V_b e^{\jmath\beta\ell} = V_R e^{\jmath\beta\ell}, \tag{7.59}$$

and so

$$Z_{\text{in}} = R. \tag{7.60}$$

This analysis is idealized, as parasitics are eliminated (mainly parasitic capacitances), but the above equation indicates that the essential function of the structure is as a balun (determined by the structure) with no impedance transformation.

The transformer arrangement shown in Figure 7-19(b) is of particular interest, as it can be realized using coupled transmission lines.
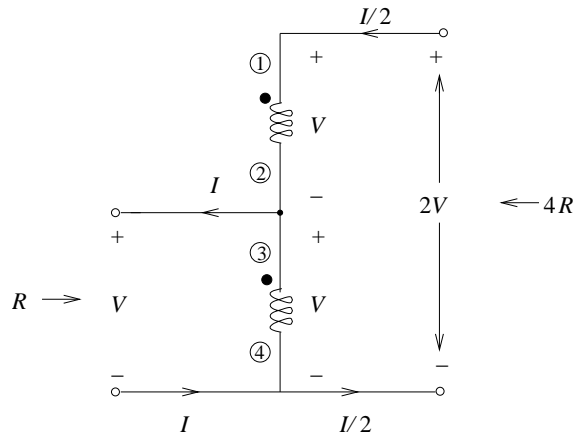
Figure 7-20   Schematic of a 4:1 transformer.
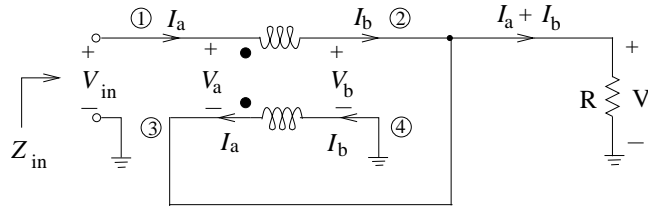
Figure 7-21   A transmission line form of the 4:1 impedance transformer of Figure 7-20.

7.8.2   *Impedance Transformer*

By changing the number of windings or altering the connection of the transformer terminals it is possible to achieve impedance transformation. A 4:1 impedance transformer is shown in Figure 7-20. A specific arrangement of the primary and secondary windings put this into what is called the transmission line form, shown in Figure 7-21. Now

$$V_a \;=\; V_b\cos(\beta\ell) + \jmath I_b Z_o \sin(\beta\ell) \tag{7.61}$$

$$I_a \;=\; I_b\cos(\beta\ell) + j\frac{V_b}{Z_o}\sin(\beta\ell) \tag{7.62}$$

$$V_{\text{in}} \;=\; V_a + R(I_a + I_b) \tag{7.63}$$

and

$$V_b \;=\; (I_a + I_b)R. \tag{7.64}$$

The aim now is to find $Z_{\text{in}} = V_{\text{in}}/I_{\text{a}}$, as this defines the required electrical function. Now

$$V_{\text{in}} = V_{\text{a}} + V_{\text{b}} = V_{\text{b}}(1 + \cos(\beta\ell)) + {\jmath}I_{\text{b}}Z_{\text{o}}\sin(\beta\ell), \qquad (7.65)$$

and substituting Equation (7.64) into this it becomes

$$V_{\text{in}} = I_{\text{b}}\cos(\beta\ell) + {\jmath}(I_{\text{b}} + I_{\text{b}})\frac{R}{Z_0\sin(\beta\ell)}. \qquad (7.66)$$

Thus

$$Z_{\text{in}} = \frac{V_{\text{in}}}{I_{\text{a}}} = Z_{\text{o}}\frac{2R(1 + \cos(\beta\ell)) + {\jmath}Z_{\text{o}}\sin(\beta\ell)}{Z_{\text{o}}\cos(\beta\ell) + {\jmath}Z_{\text{o}}\sin(\beta\ell)}. \qquad (7.67)$$

At very low frequencies the electrical length of the transmission line, $\beta\ell$, is negligibly small and $Z_{\text{in}} = 4R$. To see what happens when the transmission line has a significant effect consider the special case when $Z_o = 2R$; then

$$Z_{\text{in}} = 2R(1 + e^{-{\jmath}\beta\ell}). \qquad (7.68)$$

For a short line, that is $\ell < 0.1\lambda$ or $\beta\ell < 0.2\pi$, Equation (7.68) can be approximated as

$$Z_{\text{in}} \approx 2R[1 + 1 - {\jmath}\beta\ell] = 4R - {\jmath}R(2\beta\ell). \qquad (7.69)$$

The imaginary term, $-{\jmath}R(2\beta\ell)$, is nonzero and so it must be resonated out (e.g., by a series capacitor) to obtain the required resistance transformation.

The general approach described above can be used to design transformers with higher impedance ratios. Two more, a 9:1 transformer and a 16:1 transformer, are shown in Figure 7-22. In all cases, the characteristic impedance of the transmission line is specified by the voltage across the end of the line and the current flowing through the line.

A practical broadband transmission line realization of the 4:1 transformer is shown in Figure 7-23, where the transmission line is a twisted wire.

7.9   Hybrid Transformer Used as a Combiner

The hybrid transformer can do more than combine power. It can also provide impedance transformation if the number of windings are changed. In Figure 7-24, a $180°$ hybrid transformer is used to combine the outputs of two power amplifiers that are driven $180°$ out of phase with respect to each other. This is commonly done when the power available from a solid-state transistor amplifier is not sufficient to meet requirements. Since the sum of the amp-turns of an ideal transformer must be zero,

$$nI_o \;\; = \;\; m(I_1 + I_2) \qquad (7.70)$$
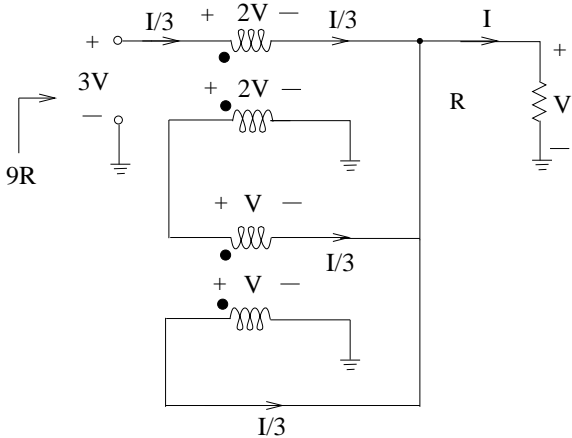
(a)
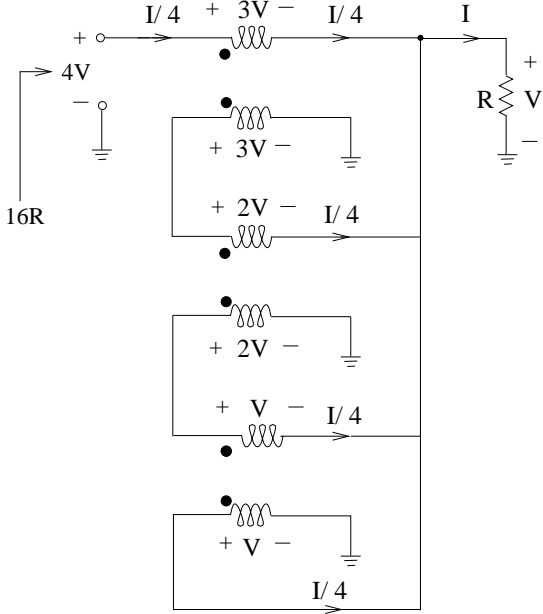
(b)

Figure 7-22   High-order impedance transformers: (a) a 9:1 transformer; and (b) a 16:1 transformer.
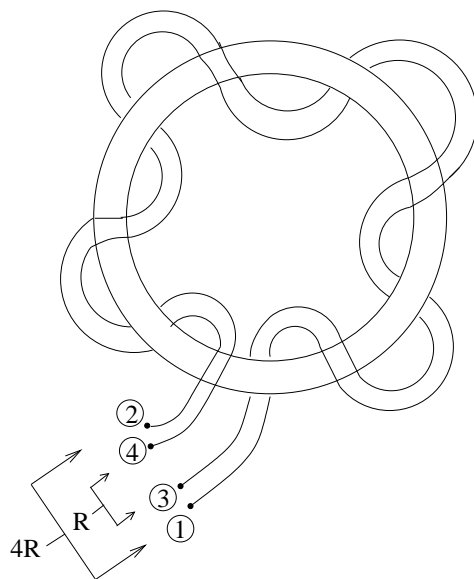
Figure 7-23 Broadband 4:1 impedance transformer with twisted coupled wires twisted around a magnetic core.
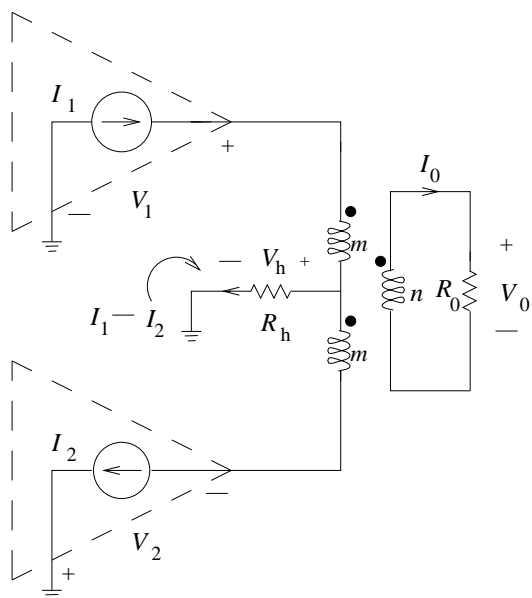
Figure 7-24 A $180°$ hybrid transformer used as a combiner.

and $I_1 = I_2$, since the two amplifiers are driven identically but $180°$ out of phase. Then $V_h = 0$, $V_1 = V_2 = (m/n)V_o$, and both power amplifiers deliver equal power to the load. Also, each amplifier sees a load impedance

$$R_{\text{in}} = \frac{V_1}{I_1} = \left(\frac{m}{n}\right)^2 2R_0. \qquad (7.71)$$

EXAMPLE 7. 3   Transformer Design

The Thevenin equivalent output impedance of the amplifiers of Figure 7-24 is $1\ \Omega$ and the system impedance, R_0, is $50\ \Omega$. Choose the transformer windings for maximum power transfer.

Solution:

For maximum power transfer, $R_{\text{in}} = 1\ \Omega$, and so from Equation 7.71,

$$1 = \left(\frac{m}{n}\right)^2 \cdot 2 \cdot 50 \qquad (7.72)$$

$$\frac{m}{n} = 10. \qquad (7.73)$$

So with a 10:1 winding ratio the required impedance transformation can be achieved.

7.10   Hybrid Transformer Used as a Power Splitter

The hybrid transformer can also be used to split power from the source to drive two loads. The circuit of Figure 7-25 splits power from a current source driver into two loads. With the number of primary and secondary windings equal (i.e., $m = n$), the circuit equations are

$$I = I_1 + I_2 \qquad (7.74)$$
$$V_1 = V + V_h \qquad (7.75)$$
$$V_2 = V_h - V, \qquad (7.76)$$

and so

$$V_1 - V_h = V_h - V_2. \qquad (7.77)$$

Using $V_1 = R_1 I_1$, $V_2 = -I_2 R_2$, and $V_h = (I_2 - I_1)R_h$, the desired electrical characteristics of the splitter are obtained:

$$\frac{I_2}{I_1} = \frac{R_1 + 2R_h}{R_2 + 2R_h}. \qquad (7.78)$$

Several observations can be made about the performance of the power splitter. If $R_1 = R_2$, then $I_1 = I_2$ for any value of R_h. Conversely, if $R_1 \neq R_2$, $I_1 \neq I_2$ for a finite R_h. To obtain equal drive currents in both power amplifiers in spite of variations in R_1 and R_2, the center tap of the transformer needs to be omitted so that $R_h = \infty$.
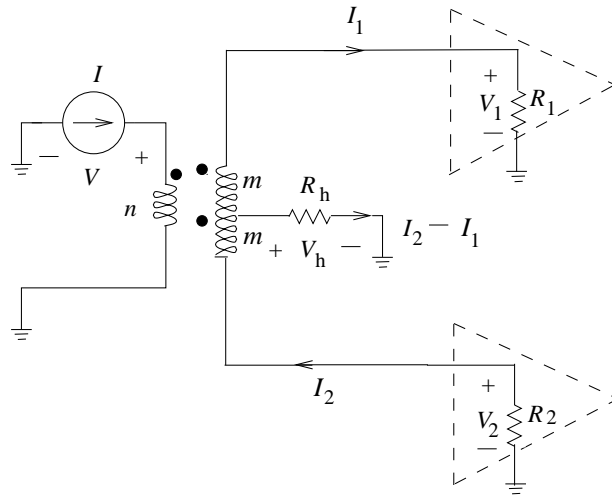
Figure 7-25   A $180°$ hybrid transformer used as a power splitter.

7.11   Broadband Hybrid Combiner

A broadband hybrid combiner is shown in Figure 7-26. In what follows, it is shown that this combiner has the property of accommodating mismatches of the amplifiers. The development begins by assuming that the transformers have an equal number of turns on each winding. These two transformers are used to make a broadband (transmission line transformer) hybrid coupler. The circuit equations are

$$I_1 = I_a + I_b \quad \text{and} \quad I_2 = I_a - I_b \tag{7.79}$$

$$I_a = \frac{1}{2}(I_1 + I_2) \quad \text{and} \quad I_b = \frac{1}{2}(I_1 - I_2) \tag{7.80}$$

$$V_1 = \frac{V_o}{2} + V_h \quad \text{and} \quad V_2 = \frac{V_o}{2} - V_h \tag{7.81}$$

$$\frac{V_1}{I_1} = R_h \left(1 - \frac{I_2}{I_1} \right) + \frac{R_0}{2} \left(1 - \frac{I_2}{I_1} \right) \tag{7.82}$$

$$\frac{V_2}{I_2} = \frac{R_0}{2} \left(1 + \frac{I_1}{I_2} \right) - R_h \left(\frac{I_1}{I_2} - 1 \right). \tag{7.83}$$

A special situation is when $R_h = R_0/2$ and then $V_2/I_2 = R_0$ and $V_1/I_1 = R_0$, thus each of the amplifiers sees a constant load resistance, R_0, even if the amplifiers are mismatched, resulting, for example, when the amplifiers have slightly different gains.
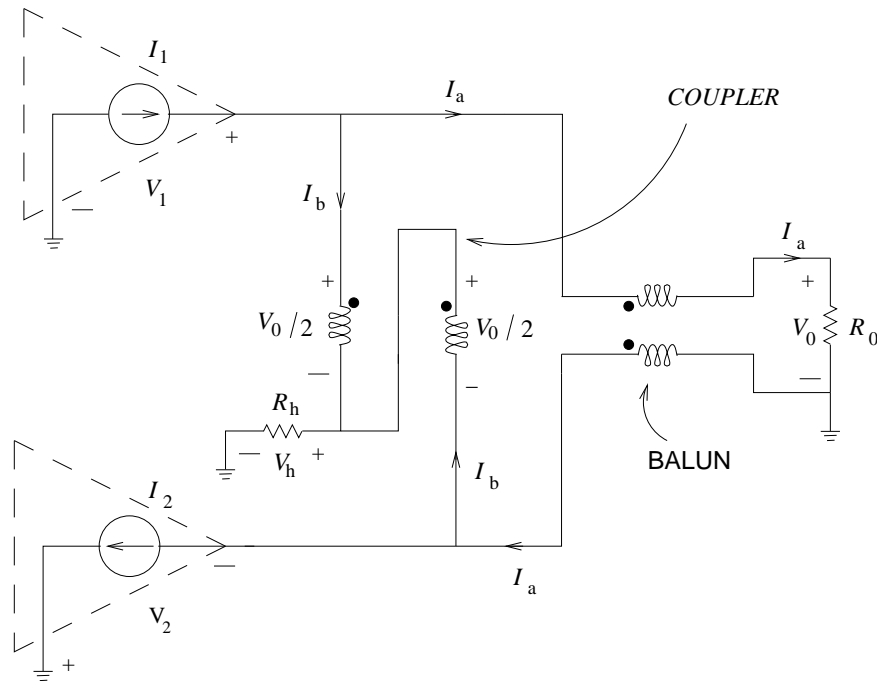
Figure 7-26 Broadband hybrid power combiner.

7.12 Branch-Line Hybrids Based on Transmission Lines

A branch-line hybrid is based on transmission line segments that introduce phase delay. Two such hybrids are shown in Figure 7-27 where the different signal paths result in constructive and destructive interference of signals. This is a very different way of realizing the hybrid function than is obtained using magnetic transformers. These branch-line hybrids may be formed into a ring shape, as shown in Figure 7-28. The operation of the $180°$ hybrid (Figure 7-27(a)) can be readily verified by counting up the total number of one-quarter wavelength ($90°$) phase shifts in each path. It is not so easy to verify the operation of the $90°$ hybrid. The various characteristic impedances of the transmission line segments adjust the levels of the signals. The operation of the $90°$ hybrid is examined in the example in this section.

It is also worth considering the so-called rat-race or hybrid-ring circuit shown in Figure 7-29. Output signals from Ports 2 and 4 differ in phase by $180°$ (in contrast to the branch-line coupler, where the phase difference is $90°$). An interesting and important design feature arises when considering the quarter-wave transformer action of this coupler. Only Ports 2 and 4 exhibit this action, because Port 3 is half-wave separated from the input feeding Port 1. Thus the net effective load on the inner ring lines feeding
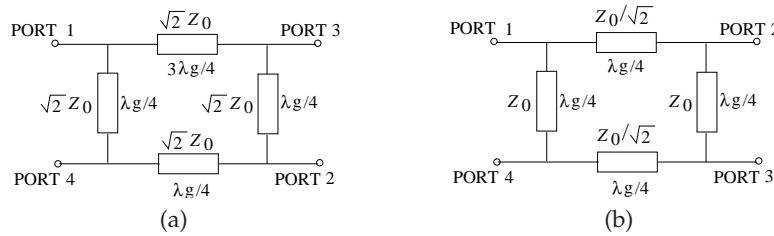
Figure 7-27   Topologies of ring-type hybrids: (a) $180°$ hybrid; and (b) $90°$ hybrid.
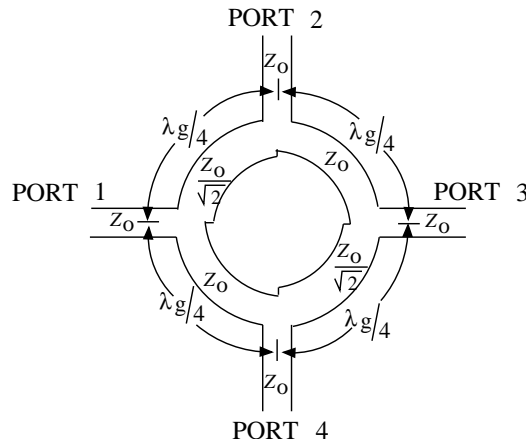
Figure 7-28   Planar implementation of the 3 dB ring-type branch-line hybrid with the topology of Figure 7-27(b).

Ports 2 and 4 amounts to $2Z_0$ (two Z_0 loads appearing, equivalently, in series). Now, the characteristic impedance Z_0 of any quarter-wave transforming line between two impedances, Z_{01} and Z_{02}, is equal to $\sqrt{Z_{01}Z_{02}}$. In this case, the two impedances are Z_0 and $2Z_0$, respectively, so (from Section 4.8.4 on Page 218) the input impedance of the intervening quarter-wave line is

$$Z_0' = \sqrt{Z_0 \cdot 2Z_0} = \sqrt{2}Z_0. \tag{7.84}$$

Thus the characteristic impedance of the line forming the ring itself is $\sqrt{2}$ times that of the feeder line impedances. So when the impedance of all feeders is 50 Ω, the ring characteristic impedance is 70.7 Ω.

7.13   Lumped-Element Hybrids

It is possible to create hybrids using lumped elements just as segments of transmission line can be modeled by a series inductor and a shunt capacitance. The full circle of design can begin with the magnetic transformer
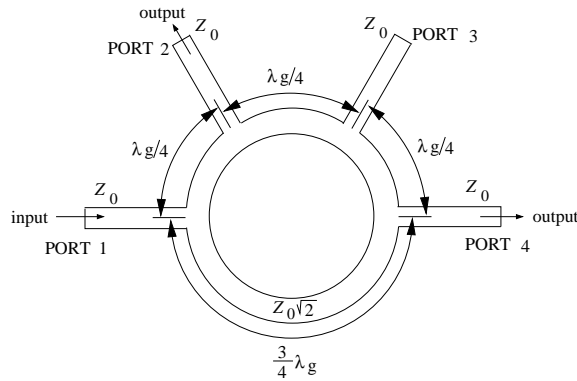
Figure 7-29 Rat-race or hybrid-ring coupler.

conceptualization, followed by a transmission line realization, and then a lumped-element approximation. Even if the magnetic transformer conceptualization is not used, a transmission line structure can be approximated using the lumped-element equivalent. At least this is valid over a small bandwidth centered at a particular frequency, ω_o. The lumped elements are related to the transmission line structure of Figure 7-27(b) by the $180°$ hybrid shown in Figure 7-30(a) with

$$\omega_o L = \frac{1}{\omega_o C} = \sqrt{2} Z_o. \tag{7.85}$$

The lumped-element quadrature hybrid of Figure 7-30(b) has

$$\omega_o Z_o C_a = 1 \tag{7.86}$$

$$C_a + C_b = \frac{1}{\omega^2 L} \tag{7.87}$$

$$\omega_o L = \frac{Z_o}{\sqrt{2}}, \tag{7.88}$$

where Z_o is the port impedance.

7.14 Resonators

Electrical resonators store energy in electric and magnetic forms with the energy alternating between the two forms at the resonant frequency. Most microwave resonators utilize the storage of energy in a volume.

7.14.1 Transmission Line Resonators

Transmission line resonators are either quarter-wavelength or half-wavelength long sections of line depending on whether the resonance
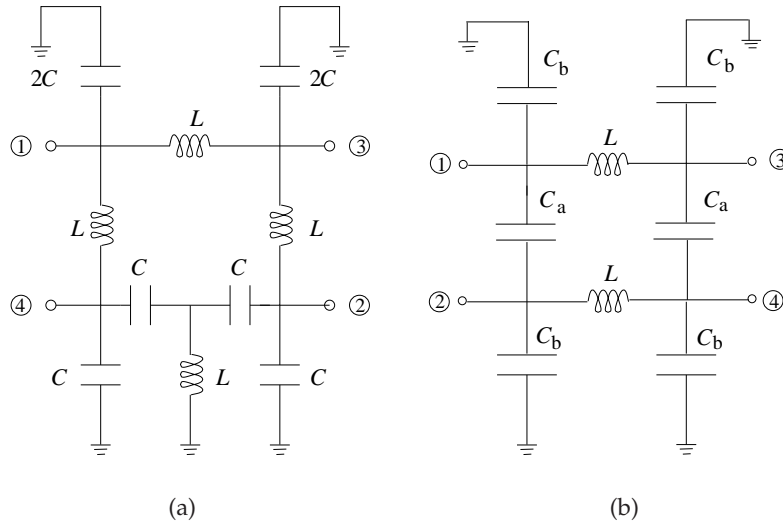
Figure 7-30 Lumped-element hybrids: (a) a $180°$ hybrid; and (b) a quadrature hybrid.

desired is an open or short circuit. The loaded Q of the resonator is an important parameter, so it is important to use a coupling structure such as a cap coupler to connect to the resonator.

7.14.2 Dielectric Resonators

Any cylindrically shaped dielectric structure can store EM energy, with the resonant frequency dependent on both the permittivity and the physical dimensions. The general theoretical expression for such resonant frequencies, applying to a dielectric resonator of radius a and height d, is

$$f_{mnl} = \frac{c}{2\sqrt{\varepsilon_r}}\sqrt{\left(\frac{\chi_{mn}}{\pi a}\right)^2 + \left(\frac{l}{d}\right)^2}, \tag{7.89}$$

where the integer l denotes the number of half wavelengths in the vertical direction and χ_{mn} is the mth extremum of the Bessel function J_n for a TM mode (or alternatively the mth zero for a TE mode).

Low-loss ceramic materials with permittivities in the range of 21 to 150 are generally available and unloaded Q factors are in the 5,000 to 10,000 range. Physical dimensions generally depend upon the resonant frequencies desired.

Figure 7-31(a) shows a common use of a dielectric resonator, often called a **puck**, coupling to the fields of microstrip line. The puck has a resonance defined by the boundary between air and the high permittivity of the puck. There can be several modes of resonance, with some modes introducing a

Figure 7-31 Dielectric resonator: (a) microstrip transmission line with a coupled dielectric resonator; and (b) parallel RLC resonant circuit.

parallel resonant circuit (Figure 7-31(b)) in shunt across the line and others introducing a series resonant circuit (Figure 7-31(c)) in shunt across the line. The strong impedance variation with frequency can be used to define the frequency of the oscillator, for example.

7.15 Circulators and Isolators

Circulators and isolators, shown in Figure 7-32, are nonreciprocal elements and their behavior is conveniently described using S parameters. The essential element of a circulator is a piece of ferrite which when magnetized becomes nonreciprocal, preferring progression of EM fields in one circular direction. An ideal circulator has the scattering matrix

$$\mathbf{S} = \left[\begin{array}{ccc} 0 & 0 & S_{13} \\ S_{21} & 0 & 0 \\ 0 & S_{32} & 0 \end{array} \right]. \tag{7.90}$$

This rather odd behavior is the result of a quantum mechanical effect known as **gyromagnetic** resonance. This resonance occurs when the magnetic dipoles align and progress in unison when excited by an EM field. The reason the effect is only seen in ferrites and related materials is that magnetic dipoles do not align in nonmagnetic materials.

In addition to the insertion and return losses, the performance of a circulator is described by its isolation, which is its insertion loss in the undesired direction. An actual circulator is shown in Figure 7-33.

Isolators are devices that allow power flow in one direction. Ferrite isolators are based on a three-port circulator with one of the ports terminated in a matched load. Figure 7-32(b) shows a microstrip isolator based on a three-port circulator. The puck at the center is magnetic material such as ferrite that when magnetized by a permanent magnet or electromagnet (which is not shown) preferentially supports a clockwise-rotating EM wave. So power entering Port 1 as a traveling wave is transferred to the ferrite and emerges at Port 2. Virtually none of the power emerges at Port 3. A traveling wave signal applied at Port 2 appears at Port 3, where it is absorbed in a termination created by resistive material placed
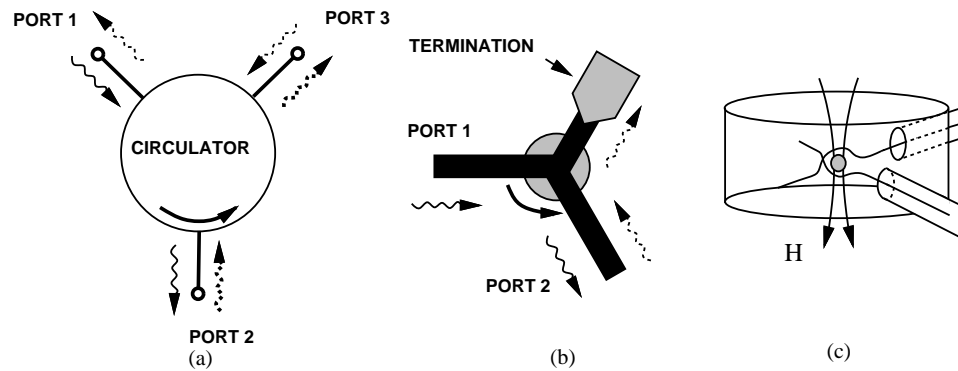
Figure 7-32 Ferrite components: (a) schematic of a circulator; (b) a microstrip isolator, and (c) a YIG-tuned bandpass filter [131].

Figure 7-33 A coaxial circulator showing three ports and the direction of circulation. The ferrite disc is located in the center of the circulator and is topped by an enclosed permanent magnet.

on top of the microstrip. The resistive material forms a lossy transmission line, and provided that the lossy line section is long enough, no power is reflected to be incident at Port 3. Thus power can travel from Port 1 to Port 2, but not in the reverse direction. An isolator is commonly used to protect the output of equipment from high reflected signals. A four-port version implements a duplexer and is used in radar systems and to separate the received and transmitted signals in a transceiver.

The gyromagnetic effect also enables very sharp variable bandpass filters. A **Yttrium-Iron-Garnet** (**YIG**)-tuned filter is shown in Figure 7-32(c) and the result is that Ports 1 and 2 are only coupled when at a precise frequency determined by the state of magnetization of the YIG sphere. With an electromagnet providing a DC magnetic field, the magnetic field can be varied and the resonance frequency, and hence the bandpass frequency, electronically tuned. Most microwave spectrum analyzers use a YIG-tuned filter at the input of the analyzer.

7.16 Summary

Many passive microwave elements exploit particular physical phenomena. Many make use of the characteristics of transmission lines. Each year new variants of microwave elements are developed and documented in patents and publications. Microwave engineers often monitor developments in elements that are being exploited in design.

7.17 Exercises

1. The Thevenin equivalent output impedance of the amplifiers in Figure 7-24 is 5 Ω and the system impedance, R_0, is 75 Ω. Choose the transformer windings for maximum power transfer. [Parallels Example 7.3 on Page 400.]

2. A spiral inductor is modeled as an ideal inductor of 10 nH in series with a 5 Ω resistor. What is the Q of the spiral inductor at 1 GHz?

3. A 50 dB attenuator can be constructed using three resistors. Consider the design of a 50 dB in a 75 system. [Parallels Example 7.1 on Page 380.]

   (a) Draw the topology of the attenuator.
   (b) Write down the design equations for the resistors.
   (c) Complete the design of the attenuator.

4. The circuit below is called a T attenuator and is inserted in a system with real system impedance Z_0; that is, a resistance of value Z_0 is attached to the input and the output.
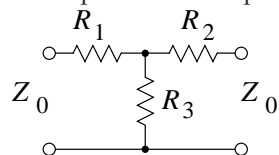
   Design the attenuator so that its attenuation is A dB. Show that

   $$R_1 = R_2 = Z_0\frac{(\sqrt{N}-1}{\sqrt{N}+1)}$$

   $$R_3 = \frac{2Z_0\sqrt{N}}{N-1},$$

   where $N = \log^{-1}(A/10)$.

5. A 20 dB attenuator in a 17 Ω system is ideally matched at both the input and output. Thus

there are no reflections and the power delivered to the load is reduced by 20 dB from the applied power. If a 5 W signal is applied to the attenuator, how much power is dissipated in the attenuator?

6. Write down the 50 Ω scattering parameters of the ideal transformer shown below where the ratio of the number of windings on the secondary side (Port 2) is twice the number of windings on the primary side (Port 1).
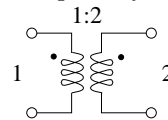
   (a) What is S_{11}? [Hint: Terminate Port 2 in 50 Ω and determine the input reflection coefficient.]
   (b) What is S_{21}?
   (c) What is S_{22}?
   (d) What is S_{12}?

7. Derive the two-port 50 Ω scattering parameters of the magnetic transformer below. The primary (Port 1) has 10 turns, the secondary (Port 2) has 25 turns.
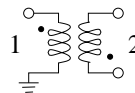
   (a) What is S_{11}?
   (b) What is S_{21}?
   (c) What is S_{22}?
   (d) What is S_{12}?

8. Consider the hybrid shown in Figure ??. If the number of windings of Coils 2 and 3 are twice the number of windings of Coil 1, show that for matched hybrid operation $2Z_2 = Z_3 = 8Z_0$. Also determine the optimum Z_2 for best operation.

9. A signal is applied to Ports 2 and 3 of a $180°$ hybrid, as shown in Figure 7-11(b). If the signal consists of a differential component of 0 dBm and a common mode component of 10 dBm:

    (a) Determine the power delivered at Port 1.
    (b) Determine the power delivered at Port 4. Assume that the hybrid is lossless.

10. The balun of Figure 7-10 transforms an unbalanced system with a system impedance of Z_0 to a balanced system with an impedance of $4Z_0$. The actual impedance transformation is determined by the number of windings of the coils. Design a balun of the type shown in Figure 7-10 that transforms an unbalanced 50 Ω system to a balanced 377 Ω system. [Hint: Find the ratio of the windings of the coils.]

11. A balun can be realized using a wire-wound transformer, and by changing the number of windings on the transformer it is possible to achieve impedance transformation as well as balanced-to-unbalanced functionality. A 500 MHz balun based on a magnetic transformer is required to achieve impedance transformation from an unbalanced impedance of 50 Ω to a balanced impedance of 200 Ω. If there are 20 windings on the balanced port of the balun transformer, how many windings are there on the unbalanced port of the balun?

12. Design a lumped-element 2-way power splitter in a 75 Ω system. Base your design on a Wilkson power-divider. Use a minimum inductance of 2 nH.

13. Design a 3-way power splitter in a 75 Ω system. Base your design on a Wilkson power-

divider using transmission lines and indicate lengths in terms of wavelengths.

14. Design a lumped-element 3-way power splitter in a 75 Ω system. Base your design on a Wilkson power-divider. Use a minimum inductance of 2 nH.

15. Silicon RFICs use differential signal paths to minimize the introduction of substrate noise. As well, differential amplifiers are an optimum topology in current-biased circuits. Off-chip signals are often on microstrip lines and so the source and load, being off-chip, are not differential. The off-chip circuits are then called single-ended. Using diagrams and explanations, outline a system architecture accommodating this mixed differential and single-ended environment.

16. A resistive power splitter is a three-port device that takes power input at Port 1 and delivers power at Ports 2 and 3 that are equal; that is, $S_{21} = S_{31}$. However, the sum of the power at Ports 2 and 3 will not be equal to the input power due to loss. Design a 75 Ω resistive three-port power splitter with matched inputs, $S_{11} = 0 = S_{22} = S_{33}$. That is, draw the resistive circuit and calculate its element values.

17. Develop the electrical design of a rat-race hybrid at 30 GHz in a 50 Ω system.

18. Develop the electrical design of a rat-race hybrid at 30 GHz in a 50 Ω system.

19. Design a lumped-element hybrid at 1900 MHz using 1 nH inductors.

20. Using a schematic, show how an isolator can be obtained from a circulator.

Impedance Matching

8.1   Introduction

The maximum transfer of signal power is one of the prime objectives in RF and microwave circuit design. Power traverses a network from a source to a load generally through a cascade, or at least a sequence, of two-port networks. Maximum power transfer requires that the Thevenin equivalent impedance of a source be matched to the impedance of the load connected to it. More specifically, the load should be the complex conjugate of the impedance (or equivalently the reflection coefficient) that the load sees looking back toward the source. If the impedances are resistive a magnetic transformer can be used to match the impedances. This will not work at radio and microwave frequencies, as the design problem is generally to match complex impedances.

  A transmission line matching problem is illustrated in Figure 8-1, where the matching network is ideally lossless to avoid unnecessary loss of power. Impedance matching or tuning is important for the following reasons:

 (a) Maximum power is delivered when the load is matched to the line
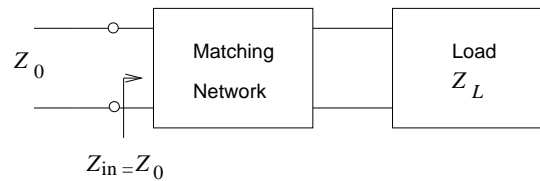
Figure 8-1 A load connected to a transmission line of characteristic impedance Z_0 through a matched load designed to realize maximum power transfer.

(assuming the generator is matched), and power loss in the feed line is minimized. In general we need $Z_{in} = Z_0^*$.

(b) Impedance matching receiver components (antennas, low-noise amplifiers [LNAs], etc.) improves the SNR of the system.

(c) As long as Z_L has some nonzero real part, a matching network can always be found.

Up to a few gigahertz, lumped inductors and capacitors can be used in matching networks, and this is done in cellular phone handsets. Above a few gigahertz, distributed parasitics (loss and additional capacitive or inductive effects) can render lumped-element networks impractical. Then segments of transmission lines are used. They are also used at lower frequencies when loss must be kept to a minimum or power levels are high. This is because the loss of the transmission line component is generally less than the loss of a lumped inductor. Many factors affect the selection of the components of a matching network. The most important of these are size, complexity, bandwidth, and adjustability. The function of a distributed element is directly related to its length compared to a wavelength. A basic functional length of a transmission line element is one-quarter wavelength, so that a characteristic length at 1 GHz is 7.5 cm modified by the square root of the permittivity of the material being used. A circuit board, for example, with a relative permittivity of 4 has one-quarter wavelength of 3.75 cm. This is far too large to fit in consumer wireless products. Thus, at a few gigahertz, lumped-element matching networks are preferred in terminal units.

Over the years the performance of lumped elements has dramatically increased largely in response to the pressure to manage size. Performance generally refers to both the self-resonant frequency of an element and loss. Inductors are a particular case in point. An actual inductor must be modeled using capacitive elements, capturing interwinding capacitance as well as the primary inductive element. At some frequency the inductive and capacitive elements will resonate and establish the self-resonant frequency of the element. The self-resonant frequency is the most common metric used to specify the maximum operating frequency of the element. With regards to complexity, the simplest designs are most often preferred. A simpler

matching network is usually more reliable, less lossy, and cheaper than a complex design. The pursuit of higher bandwidth, however, may necessitate that the simplest circuit not be selected. Any type of matching network can ideally give, ignoring resistive loses, a perfect match at a single frequency. Away from this center frequency the match will not be ideal.

From experience, achieving a reasonable match over a 5% bandwidth based on a single-frequency design is not that hard. However, it is often desirable to match a load over a wider band of frequencies. Thus matching network design has a lot in common with filter design, and many designers prefer to use a filter design approach. Design becomes increasingly more difficult as the required bandwidth increases. Many of the evolving wireless systems require multiple functionality which in turn requires adjustability. In some applications the matching network may require adjustment to match a variable load impedance. Some types of matching networks are more adjustable than others. Such designs require variable components, so matching design can be challenging and a source of competitive advantage.

Depending on the application, an impedance matching network may consist of

(a) Lumped elements only. These are the smallest networks, but have the most stringent limit on frequency of operation. The relatively high resistive losses of inductors limits their performance.

(b) Distributed elements (microstrip-line elements or whatever the appropriate transmission medium might be) only. These have excellent performance, but their size restricts their use to above a few gigahertz.

(c) A combination of lumped and distributed elements, primarily small sections of transmission lines with capacitors. The line lengths are smaller than with the distributed elements only, but higher performance than lumped elements on their own.

(d) Adhoc solutions (suggested by input impedance behavior and features of various components).

This chapter concentrates on matching networks, mostly of the lumped-element type, but all the issues pertinent to matching will be explored. The emphasis here is the development of design equations and synthesis of desired results. An alternative approach, used by some designers, is to choose a circuit topology and then use a circuit optimizer to arrive at circuit values that yield the desired characteristics. This is sometimes satisfactory design technique, but is not a good solution in new designs as it does not provide insight and does not help in choosing new topologies.

Figure 8-2   Resonant circuits: (a) parallel; and (b) series.

8.2 Q **Factor and Resonant Circuits**

Inductors and capacitors, which are the main components in lumped-element matching networks, also have loss and parasitic elements. With inductors there is both series resistance and shunt capacitance mainly from interwinding capacitance, while with capacitors there will be shunt resistance and series inductance generally. A practical inductor or capacitor will be limited to operation below what is called the self-resonant frequency. The impact of loss is quantified by the Q factor (or quality factor). Q is loosely related to bandwidth in general and is based on the response of a series or parallel connection of a resistor (R), an inductor (L), and a capacitor (C). The response of an RLC network is described by a second-order differential equation with the conclusion that the 3 dB fractional bandwidth of the response (i.e., when the power response is at its half-power level below its peak response) is $1/Q$. (The fractional bandwidth is $\Delta f/f_0$ where f_r is the resonant frequency at the center of the band and Δf is the 3 dB bandwidth.) This is not true for any network other than a second-order circuit, but as a guiding principle, networks with higher Qs will have narrower bandwidth.

8.2.1   *Definition*

The Q (quality) factor, of a component or circuit is defined as the ratio of 2π times the maximum energy stored to the energy lost per cycle. For a parallel resonant circuit with elements L, C, and G (see Figure 8-2(a)), the Q is

$$Q = \omega_r C/G = 1/(\omega_r LG), \tag{8.1}$$

where $f_r = \omega_r/(2\pi)$ is the resonant frequency and is the frequency at which the maximum amount of energy is stored. The conductance, G, describes the energy lost in a cycle. A series resonant circuit (Figure 8-2(b)) with L, C, and R elements has a similar expression for Q:

$$Q = \omega_r L/R = 1/(\omega_r CR). \tag{8.2}$$

These resonant circuits have a bandpass transfer characteristic (see Figure 8-3). The Q is approximately the inverse of the fractional bandwidth of the resonator. The fractional bandwidth, Δf, is measured at the half-power points as shown in Figure 8-3. (Δf is also referred to as the two-sided -3 dB
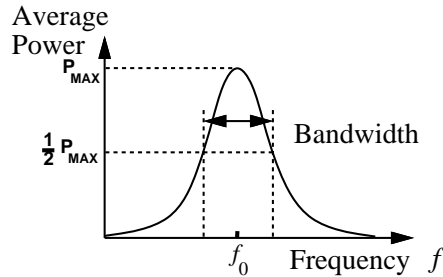
Figure 8-3   Transfer characteristic of a resonant circuit.
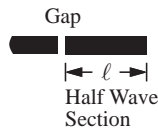
Figure 8-4   Gap coupled half-wave resonator.

bandwidth.) Then

$$Q = f_r/\Delta f. \tag{8.3}$$

Thus the Q is a measure of the sharpness of the bandpass frequency response. The pairs of definitions—Equations (8.1) and (8.3), and Equations (8.2) and (8.3)—are identical for parallel RLC and series RLC circuits, respectively.

   The determination of Q using the measurement of bandwidth together with Equation (8.3) is often not very precise, so another definition that uses the much more sensitive phase change at resonance is preferred. With ϕ being the phase (in radians) of the transfer characteristic, the definition of Q is now

$$Q = \frac{\omega_r}{2} \left| \frac{d\phi}{d\omega} \right| . \tag{8.4}$$

This is another equivalent definition of Q for parallel RLC or series RLC resonant circuits. However, Q is used for more complicated circuits and then its meaning is always a ratio of the energy stored to the energy dissipated. Thus it is meaningful to talk about the Q of circuits other than three-element RLC circuits. The Q of these structures can no longer be determined precisely by bandwidth or by rate of phase change considerations.

8.2.2   *Q of Lumped Elements*

Q is used to characterize the loss of lumped inductors and capacitors. Inductors have a series resistance and the main loss mechanism of a capacitor is a shunt conductance G (see Figure 8-5).

(a)   (b)
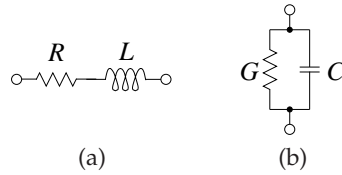
Figure 8-5 Loss elements of practical inductors and capacitors: (a) an inductor has a series resistance R; and (b) for a capacitor, the dominant loss mechanism is a shunt conductance $G = 1/R$.

The Q of an inductor at frequency $f = \omega/(2\pi)$ with a series resistance R and inductance L is

$$Q_{\text{INDUCTOR}} = \frac{\omega L}{R}. \tag{8.5}$$

Since R is approximately constant for an inductor the Q will vary with frequency.

The Q of a capacitor with a shunt conductance G and capacitance C is

$$Q_{\text{CAPACITOR}} = \frac{\omega C}{G}. \tag{8.6}$$

G is due mainly to relaxation loss mechanisms of the dielectric of the capacitor and so varies linearly with frequency. Thus the Q of a capacitor is almost constant with respect to frequency.

8.2.3 Loaded Q Factor

The Q of a component or device as defined in the previous section is called the unloaded Q, Q_U. However, if a component is to be measured or used in any way it is necessary to couple energy in and out of it. This leads to the use of the loaded Q, Q_L. So a parallel LCG circuit with elements L_r, C_r, and G_r (at resonance) loaded by a shunt conductance G_l has

$$Q_U = \omega_r C_r / G_r = 1/(\omega_r L_r G_r) \tag{8.7}$$

and

$$Q_L = \omega_r C_r / (G_r + G_l). \tag{8.8}$$

So

$$\frac{1}{Q_L} = \frac{1}{Q_U} + \frac{1}{Q_X} \tag{8.9}$$

or

$$Q_X = \left(\frac{1}{Q_L} - \frac{1}{Q_U} \right). \tag{8.10}$$

Q_X is called the external Q, and it describes the effect of loading. Q_L is the Q that would actually be measured. Q_U normally needs to be determined, but if the loading is kept very small $Q_L \approx Q_U$.
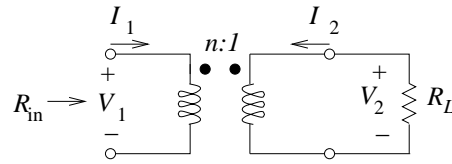
Figure 8-6 A transformer as a matching network.

8.2.4 Microstrip Resonator

Figure 8-4 shows a gap-coupled microstrip resonator comprising a half-wavelength section and coupling to the resonator controlled by the gap. The half-wavelength resonator will have a standing wave established at the resonant frequency. Energy is stored on the line segment and energy is lost to resistive dissipation and radiation. The Q of the half-wavelength section is Q_U. The gap provides a small capacitance that couples energy into the line segment so that the measured or loaded Q is Q_L. The smaller the gap capacitance, the closer Q_L is to Q_U as the loading becomes smaller.

8.3 Impedance Transforming Networks

8.3.1 The Ideal Transformer

The ideal transformer shown in Figure 8-6 can be used to match a load and a source if the source and load impedances are real. This will be shown by starting with the constitutive relations of the transformer:

$$V_1 = nV_2 \ \text{ and } \ I_1 = -\frac{1}{n}I_2. \tag{8.11}$$

Here n is the transformer ratio. For a wire-wound transformer, n is the ratio of the number of windings on the primary side, Port 1, to the number of windings on the secondary side, Port 2. So the input resistance, R_{in}, is related to the load resistance, R_L, by

$$R_{\text{in}} = \frac{V_1}{I_1} = -n^2 \frac{V_2}{I_2} = n^2 R_{\text{L}}. \tag{8.12}$$

The matching problem with purely resistive load and source impedances is solved by choosing the appropriate winding ratio, n. Unfortunately resistive-only problems at RF are rare. However at a few hundred megahertz and below, transformer-based matching sometimes works quite well, as parasitics are relatively small so that resistive load and source impedance are more closely approximated. Also, the parasitics of a wire-wound transformer (such as interwinding capacitance) can then be small.
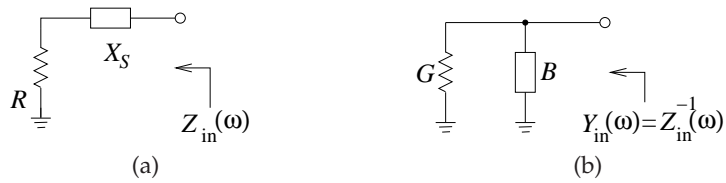
Figure 8-7 Matching using a series reactive element: (a) the series reactive element; and (b) the equivalent transformed shunt circuit.

8.3.2   A Series Reactive Element

Matching using lumped elements is based on the impedance and admittance transforming properties of series and shunt reactive elements. Even a single reactive element can achieve limited impedance matching. Consider the series reactive element shown in Figure 8-7. In Figure 8-7(a), the reactive element, with reactance X_S, is in series with a resistance R. The shunt equivalent of this network is shown in Figure 8-7(b) with a shunt susceptance of B. In this transformation the resistance R has been converted to a resistance $R_P = 1/G$. The mathematics describing this transformation is as follows. The input admittance of the series connection (Figure 8-6(a)) is

$$Y_{\text{in}}(\omega) = \frac{1}{Z_{\text{in}}(\omega)} = \frac{1}{R + jX_S} = \frac{R}{R^2 + X_S^2} - j\frac{X_S}{R^2 + X_S^2}. \tag{8.13}$$

Thus the elements of the equivalent shunt network are

$$G = \frac{R}{R^2 + X_S^2} \quad \text{and} \quad B = -\frac{X_S}{R^2 + X_S^2}. \tag{8.14}$$

The "resistance" of the network, R, has been transformed to a new value,

$$R_P = G^{-1} = \frac{R^2 + X_S^2}{R} > R. \tag{8.15}$$

This is an important start to matching, as X_S can be chosen to convert R (a load, for example) to any desired resistance value (such as the resistance of a source). However, there is still a residual reactance. Before moving on to the solution of this problem consider the following example.
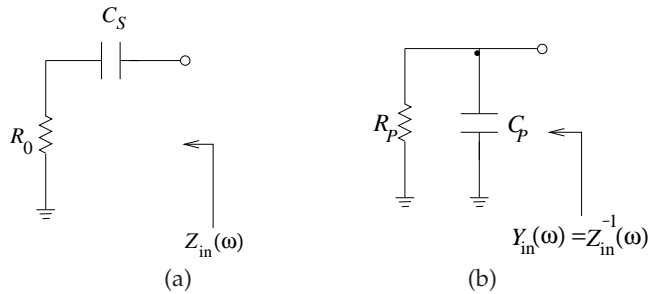
Figure 8-8 Impedance transformation using a series resistive element: (a) a resistor with a series capacitor; and (b) its equivalent shunt circuit.

| EXAMPLE 8. 1 | Capacitive Impedance Transforming |

Consider the impedance transforming properties of a capacitive series element as shown in Figure 8-8. Show that the capacitor can be adjusted to obtain any positive shunt resistance value.

Solution:

The concept here is that the series resistor and capacitor network has an equivalent shunt circuit that includes a capacitor and a resistor, but now the resistor value can be tuned by the series capacitor. The equivalent resistance of the parallel equivalent network is, from Equation (8.15),

$$R_P = G^{-1} = \frac{R_0^2 + (1/\omega^2 C_S^2)}{R_0} = \frac{1 + \omega^2 C_S^2 R_0^2}{\omega^2 C_S^2 R_0} \tag{8.16}$$

and the susceptance is

$$B = \frac{(1/\omega C_S)}{R_0^2 + 1/\omega^2 C_S^2} = \omega \frac{C_S}{1 + \omega^2 C_S^2 R_0^2}. \tag{8.17}$$

Matching, by our definition, is achieved when R_P is equal to the resistance of the source, that is, matching is achieved when $R_P = R_0$. Thus to achieve matching

$$C_P = \frac{B}{\omega} = \frac{C_S}{1 + \omega^2 C_S^2 R_0^2}. \tag{8.18}$$

To match R_0 to a resistive load R_P ($> R_0$) at a given frequency ω_d, then, from Equation (8.16),

$$\omega_d C_S = \frac{1}{\sqrt{R_0 R_P - R_0^2}}, \tag{8.19}$$

and this is the design equation for C_S. To complete the design, place a shunt inductor, as shown in Figure 8-9, where

$$\omega_d C_P = \frac{1}{\omega_d L}, \tag{8.20}$$

the design equation for L. The equivalent impedance in Figure 8-9 is a resistor of value R_P, and the value of this resistor can be adjusted by adjusting C_S.
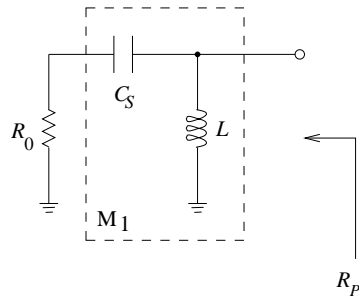
Figure 8-9   An LC matching network.
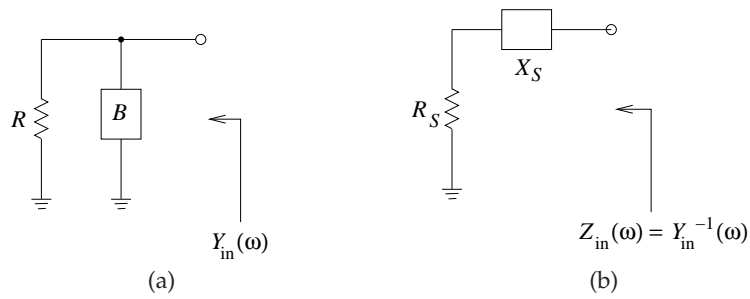
Figure 8-10   A resistor with (a) a shunt parallel reactive element, and (b) its equivalent series circuit.

8.3.3   A Parallel Reactive Element

The dual of the series capacitor matching procedure described in the previous example is obtained with a parallel reactive element, as shown in Figure 8-10(a). In general there are two two-element lumped-element circuits that will enable a particular load resistor to be matched to a source resistor. The input admittance of the shunt circuit is

$$Y_{\text{in}} = \frac{1}{R} + \jmath B. \tag{8.21}$$

This can be converted to a series circuit representation by calculating the impedance, $Z_{\text{in}} = 1/Y_{\text{in}}$:

$$Z_{\text{in}} = \frac{R}{1 + jBR} = \frac{R}{1 + B^2 R^2} - \jmath \frac{BR^2}{1 + B^2 R^2}. \tag{8.22}$$

So

$$R_S = \frac{R}{1 + B^2 R^2} \quad \text{and} \quad X_S = \frac{-BR^2}{1 + B^2 R^2}. \tag{8.23}$$

Notice that $R_S < R$.
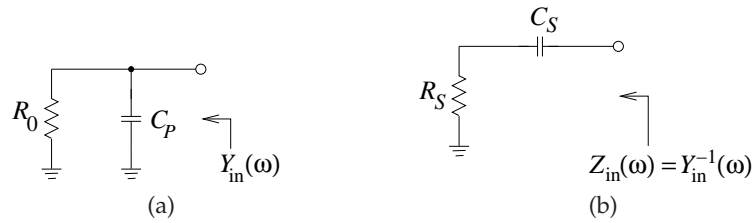
Figure 8-11   A parallel-to-series transformation: (a) a resistor with added shunt resistor; and (b) equivalent series circuit.
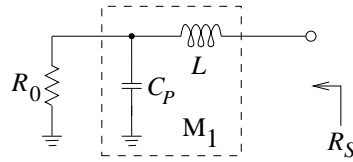
Figure 8-12   Transforming circuit with a series inductor added to the circuit of Figure 8-11(b).

EXAMPLE 8. 2   Parallel Tuning

As an example of the use of a parallel reactive element to tune a resistance value, consider the circuit in Figure 8-11(a) where a capacitor tunes the effective resistance value so that the series equivalent circuit (Figure 8-11(b)) has elements

$$R_S = \frac{R_0}{1 + \omega^2 C_p^2 R_0^2} \tag{8.24}$$

$$X_S = -\frac{\omega C_p R_0^2}{1 + \omega^2 C_p^2 R_0^2} = -\frac{1}{\omega C_S} \tag{8.25}$$

and

$$C_S = \frac{1 + \omega^2 C_p^2 R_0^2}{\omega^2 C_p R_0^2}. \tag{8.26}$$

Now consider matching R_0 to a resistive load R_S, which is less than R_0 at a given frequency ω_d. To complete the design, place a series inductor to remove the reactive effect of the capacitor, as shown in Figure 8-12. The value of the inductor required is

$$\omega_d L = \frac{1}{\omega_d C_S}. \tag{8.27}$$

This leads to the design equation for L:

$$L = \frac{1}{\omega_d^2 C_S}. \tag{8.28}$$

(a)   Lowpass $R_S < R_L$
      $Q = \sqrt{R_L/R_S - 1}$
      $Q = |X_L|/R_S = R_L/|X_C|$

(b)   Lowpass $R_S > R_L$
      $Q = \sqrt{R_S/R_L - 1}$
      $Q = |X_L|/R_L = R_S/|X_C|$

(c)   Highpass $R_S < R_L$
      $Q = \sqrt{R_L/R_S - 1}$
      $Q = |X_C|/R_S = R_L/|X_L|$

(d)   Highpass $R_S > R_L$
      $Q = \sqrt{R_S/R_L - 1}$
      $Q = |X_C|/R_L = R_S/|X_L|$
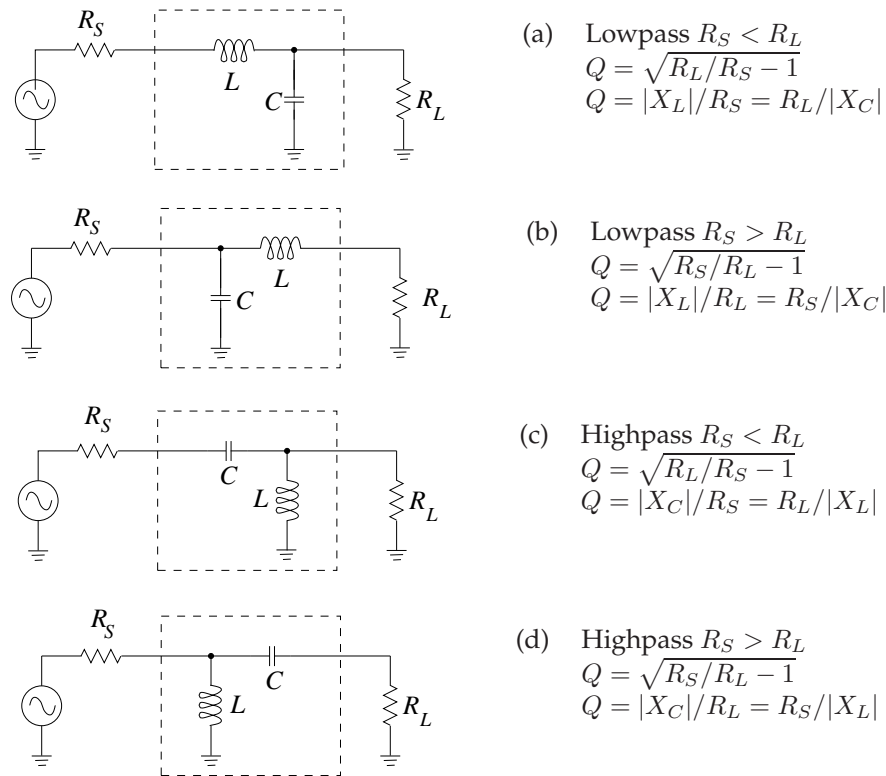
Figure 8-13   L matching networks consisting of one shunt reactive element and one series reactive element.

8.4   The L Matching Network

The examples in the previous two sections suggest the basic concept behind lossless matching of two different resistance levels using an L network:

Step 1: Use a series (shunt) reactive element to transform a smaller (larger) resistance up (down) to a larger (smaller) value with a real part equal to the desired resistance value.

Step 2: Use a shunt (series) reactive element to resonate with (or cancel) the imaginary part of the impedance that results from Step 1.

So a resistance can be transformed to any resistive value by using an LC transforming circuit.

Formalizing the matching approach described above, we note that there are four possible two-element L matching networks (see Figure 8-13). The two possible cases, $R_S < R_L$ and $R_L < R_S$, will be considered in the following subsections.
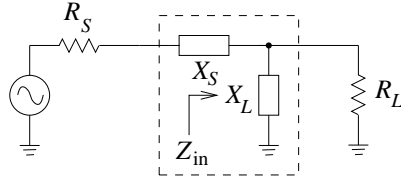
Figure 8-14   Two-element matching network topology for $R_S < R_L$.

8.4.1   Design Equations for $R_S < R_L$

Consider the matching network topology of Figure 8-14. Here

$$Z_{\text{in}} = \frac{R_L(jX_L)}{R_L + jX_L} = \frac{R_L X_L^2}{R_L^2 + X_L^2} + j\frac{X_L R_L^2}{R_L^2 + X_L^2} \tag{8.29}$$

and the matching objective is $Z_{\text{in}} = R_S - jX_S$ (the complex conjugate of the source impedance), so that

$$R_S = \frac{R_L X_L^2}{R_L^2 + X_L^2} \quad \text{and} \quad X_S = \frac{-X_L R_L^2}{X_L^2 + R_L^2}. \tag{8.30}$$

From these,

$$\frac{R_S}{R_L} = \frac{1}{(R_L/X_L)^2 + 1}; \qquad\qquad -\frac{X_S}{R_S} = \frac{R_L}{X_L}. \tag{8.31}$$

Introducing the quantities

$$Q_S = \quad \text{the } Q \text{ of the series leg} = X_S/R_S \tag{8.32}$$
$$Q_p = \quad \text{the } Q \text{ of the shunt leg} = R_L/X_L \tag{8.33}$$

leads to the final design equations for $R_S < R_L$:

$$|Q_S| = |Q_p| = \sqrt{\frac{R_L}{R_S} - 1} \tag{8.34}$$

$$-Q_S = Q_P, \quad Q_S = X_S/R_S, \quad \text{and} \quad Q_P = R_L/X_L. \tag{8.35}$$

The L matching network principle is that X_L and X_S will be either capacitive or inductive and they will have the opposite sign (i.e., the L matching network comprises one inductor and one capacitor). Also, once R_S and R_L are given, the Q of the network is defined; with the L network, the designer does not have a choice of circuit Q.
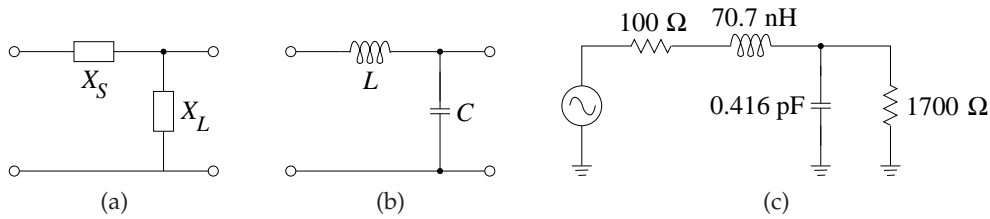
Figure 8-15 Matching network development for Example 8.3.

EXAMPLE 8. 3   Matching Network Design

Design a circuit to match a 100 Ω source to a 1700 Ω load at 900 MHz. Assume that a DC voltage must also be transferred from the source to the load.

Solution:

Here $R_S < R_L$, and so the topology of Figure 8-15(a) can be used, as this enables DC bias to be applied. From Equations (8.34) and (8.35), the design equations are

$$|Q_S| = |Q_p| = \sqrt{\frac{1700}{100} - 1} = \sqrt{16} = 4 \tag{8.36}$$

$$\frac{X_S}{R_S} = 4 X_S = 4 \times 100 = 400. \tag{8.37}$$

This indicates that $\omega L = 400$, and so the series element is

$$L = \frac{400}{2\pi \times 9 \times 10^8} = 70.7 \text{ nH}. \tag{8.38}$$

For the shunt element next to the load, $|R_L/X_C| = 4$, and so

$$|X_C| = \frac{R_L}{4} = \frac{1700}{4} = 425. \tag{8.39}$$

Thus $1/\omega C = 425$ and

$$C = \frac{1}{2\pi \times 9 \times 10^8 \times 425} = 0.416 \text{ pF}. \tag{8.40}$$

The final matching network design is shown in Figure 8-15(c).

8.4.2   L Network Design for $R_S > R_L$

For $R_S > R_L$, the topology shown in Figure 8-16 is used. The design equations for the L network design for $R_S > R_L$ are similarly derived and are

$$|Q_S| = |Q_p| = \sqrt{\frac{R_S}{R_L} - 1} \tag{8.41}$$

$$-Q_S = Q_P, \quad Q_S = \frac{X_S}{R_L}, \quad \text{and} \quad Q_p = \frac{R_S}{X_p}. \tag{8.42}$$
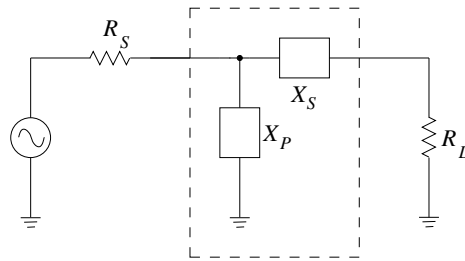
Figure 8-16   Two-element matching network topology for $R_S > R_L$.

| EXAMPLE 8. 4 | Two-Element Matching Network |

Design a passive two-element matching network that will achieve maximum power transfer from a source with an impedance of 50 Ω to a load with an impedance of 75 Ω. Choose a matching network that will not allow DC to pass.

Solution:

$R_L > R_S$, so the appropriate matching network topology is
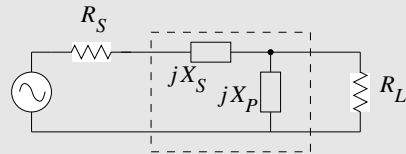
This topology can be either high pass or low pass depending on the choice of X_S and X_P. Design proceeds by finding the magnitudes of X_S and X_P. In two-element matching the circuit Q is fixed. With $R_L = 75$ Ω and $R_S = 50$ Ω, the Q of the matching network is the same for the series and parallel elements:

$$|Q_S| = |X_S|/R_S = \sqrt{\frac{R_L}{R_S} - 1} = 0.7071$$

and

$$|Q_P| = R_L/|X_P| = |Q_S| = 0.7071,$$

and

$$|Q_P| = R_L/|X_P| = |Q_S| = 0.7071,$$

therefore $|X_S| = R_S \cdot |Q_S| = 50 \cdot 0.7071 = 35.35$ Ω. Also

$$|X_P| = R_L/|Q_P| = 75/0.7071 = 106.1 \ \Omega, \quad |Q_P| = 75/0.7071 = 106.1 \ \Omega.$$

Specific element types can now be assigned to X_S and X_P, and note that they must be of opposite type. The lowpass matching network is
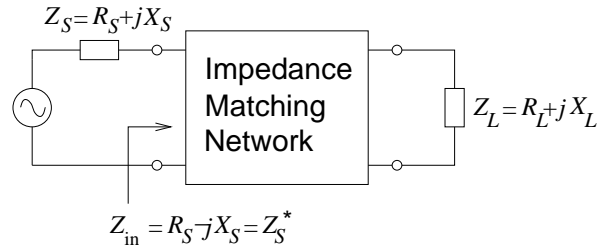
Figure 8-17 A matching network matching a complex load to a source with a complex Thevenin impedance.

$$X_S = +35.35 \ \Omega, \quad X_P = -106.1 \ \Omega.$$

The highpass matching network is

$$X_S = -35.35 \ \Omega, \quad X_P = +106.1 \ \Omega.$$

This final design satisfies the design criterion that DC is not passed, as DC is blocked by the series capacitor.

8.5 Dealing with Complex Loads

8.5.1 Matching

Input and output impedances of transistors, mixers, antennas, etc., contain both resistive and reactive components. Thus a realistic impedance matching problem looks like that shown in Figure 8-17. The matching approaches that were presented in the previous sections can be directly applied if X_S and X_L are treated as stray reactances that need to be canceled or used as part of the matching problem. There are two basic approaches to handling complex impedances:

(a) Absorption: Absorb any stray reactances into the impedance matching network itself. This is done through careful placement of each matching element such that capacitors are placed in parallel with stray capacitances and inductors in series with any stray inductances. The stray values are then subtracted from the L and C values for

the matching network calculated on the basis of the resistive parts of Z_S and Z_L only. The new (smaller) values, L' and C', constitute the elements for the matching network. Sometimes it is necessary to perform a series to parallel, or parallel to series, conversion of the source or load impedances so that the reactive elements are in the correct series or shunt arrangement for absorption.

(b) Resonance: Resonate any stray reactance with an equal and opposite reactance at the frequency of interest.

The very presence of a reactance in the model of a load at a single frequency indicates energy storage and therefore bandwidth limiting of some kind. In the above approaches to handling reactance in the load, the resonance approach could easily result in a narrowband matching solution. The major problem in matching is often to obtain sufficient bandwidth. What is sufficient will vary depending on the situation. To maximize bandwidth, the general goal is to minimize the total energy storage. Roughly the total energy stored will be proportional to the sum of the magnitudes of the reactances in the circuit. Of course, the actual energy storage depends on the voltage or current levels, which will themselves vary in the circuit. A good approach leading to large bandwidths is to incorporate the load reactance into the matching network. Thus the choice of appropriate matching network topology is critical. However, if the stray reactance value is larger than the calculated matching network element value, then absorption cannot take place. In this case resonance must be combined with absorption. Overall, the majority of impedance matching designs are based on some combination of resonance and absorption. A final observation is that for a complex load modeled as a series resistance and reactance, the energy stored is proportional to the reactance value and the energy delivered to the load is proportional to the resistance. If the reactance value is large, so that the energy storage level is relatively high, then it will be difficult to achieve broadband matching.

EXAMPLE 8. 5   Matching Network Design Using Resonance

For the configuration shown in Figure 8-18(a), design an impedance matching network that will block the flow of DC from the source to the load. The frequency of operation is 1 GHz. Design the matching network, neglecting the presence of the 10 pF at the load. Since $R_S = 50\ \Omega < R_L = 500\ \Omega$, consider the topologies of Figures 8-19(a) and 8-19(c). The design criterion of blocking flow of DC from the source to the load narrows the choice to the topology of Figure 8-19(c):

Solution:

Step 1:

$$|Q_S| = |Q_P| = \sqrt{\frac{R_L}{R_S} - 1} = 3. \tag{8.43}$$

Figure 8-18　Matching problem considered in Example 8.5: (a) source; and (b) load.

$$Q_P = \frac{R_L}{X_L}. \tag{8.44}$$

So $X_L = R_L/Q_P = 500/3$. Reducing this gives

$$\omega L = \frac{500}{3}, \tag{8.45}$$

and so

$$L = \frac{500}{3 \times 2\pi \times 10^9} = 26.5 \text{ nH}. \tag{8.46}$$

Similarly

$$-\frac{X_S}{R_S} = 3, \tag{8.47}$$

and so

$$\frac{1/\omega \, C}{R_S} = 3 \quad \text{or} \quad C = \frac{1}{3\omega R_S} \tag{8.48}$$

$$C = \frac{1}{3 \times 2\pi \times 10^9 \times 50} = 1.06 \text{ pF}. \tag{8.49}$$

Step 2:

Resonate the 10 pF capacitor using an inductor in parallel:

$$\left(\omega L'\right)^{-1} = \omega \times 10 \times 10^{-12} \tag{8.50}$$

$$L' = \frac{1}{\omega^2 10^{-11}} \tag{8.51}$$

$$L' = \frac{1}{(2\pi)^2 \, 10^{18} \times 10^{-11}} = 2.533 \text{ nH}. \tag{8.52}$$

Thus we have the circuit of Figure 8-20 as the required matching network. Two inductors are in parallel and the circuit can be simplified to that shown in Figure 8-21, where

$$L_c = \frac{2.533 \times 26.5}{2.533 + 26.5} = 2.312 \text{ nH}. \tag{8.53}$$

Figure 8-19   Matching network topology used in Example 8.5.
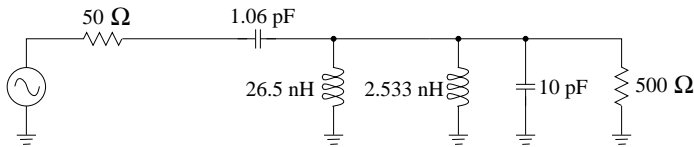
Figure 8-20   Intermediate matching network in Example 8.5.
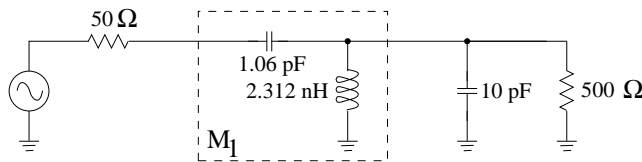
Figure 8-21   Final matching network in Example 8.5.

EXAMPLE 8. 6   Matching Network Design Using Resonance and Absorbtion

For the source and load configurations shown in Figure 8-22, design a lowpass impedance matching network at $f = 1$ GHz.

Solution:

Since $R_S < R_L$, use the topology shown in Figure 8-23(a). For a lowpass response, the topology is that of Figure 8-23(b). Notice that absorption is the natural way of handling the 3 nH at the source and the 5 pF at the load. The design process is as follows:

Step 1:

Design the matching network, neglecting the reactive elements at the source and load:

$$|Q_S| = |Q_P| = \sqrt{\frac{R_L}{R_S} - 1} = \sqrt{10 - 1} = 3 \tag{8.54}$$

$$\frac{X_S}{R_S} = 3, \quad X_S = 3 \times 100, \quad \text{and} \quad \omega L = 300 \tag{8.55}$$

$$L = \frac{300}{2\pi \times 10^9} = 47.75 \text{ nH} \tag{8.56}$$
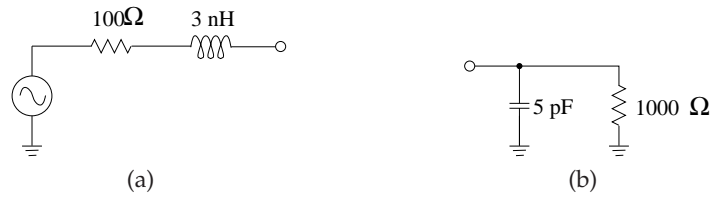
Figure 8-22　Matching problem in Example 8.6: (a) source; and (b) load.

Figure 8-23　Topologies referred to in Example 8.6.

$$\frac{R_P}{X_P} = -3 \quad \text{and} \quad \frac{1000}{-(1/\omega C)} = -3 \tag{8.57}$$

$$C = \frac{3}{1000 \times 2\pi \times 10^9} = 0.447 \text{ pF.} \tag{8.58}$$

This design is shown in Figure 8-24(a). This is the matching network that matches the 100 Ω source resistance to the 1000 Ω load with the source and load reactances ignored.

Step 2:

Figure 8-24(b) is the interim solution to the matching problem. The source inductance is absorbed into the matching network, reducing the series inductance of the matching network by the amount available from the source. The capacitance of the load cannot be fully absorbed. The matching network design for the resistance-only case requires a shunt capacitance of 0.447 pF but 5 pF is available from the load. Thus there is an excess capacitance of 4.533 pF that must be resonated out by the inductance L'':

$$\frac{1}{\omega L''} = \omega 4.533 \times 10^{-12}. \tag{8.59}$$

So

$$L'' = \frac{1}{(2\pi)^2 \times 10^{18} \times 4.533 \times 10^{-12}} = 5.588 \text{ nH.} \tag{8.60}$$

The final matching network design (Figure 8-24(c)) fully absorbs the source inductance into the matching network, but only partly absorbs the load capacitance.
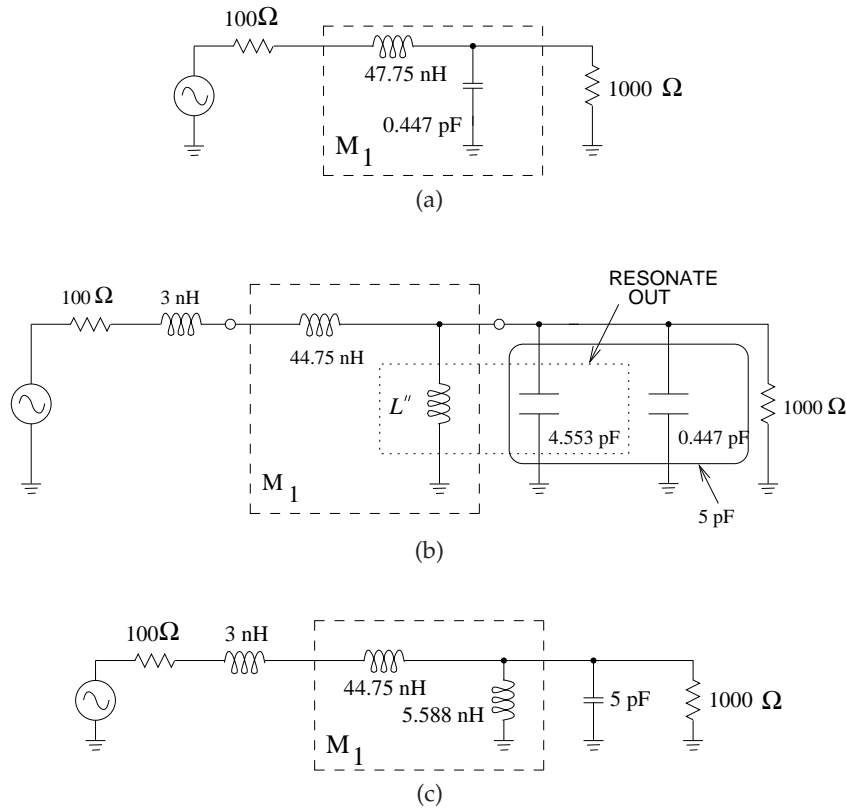
Figure 8-24 Evolution of the final circuit in Example 8.6: (a) matching network design considering only the resistive source and load resistors; (b) matching network with the reactive parts of the source and load impedances included; and (c) final design.

8.5.2 Fano-Bode Limits

A complex load limits the bandwidth of the match that can be achieved by a matching network. Theoretical limits addressing the bandwidth, and the quality of the match that can be achieved, were developed by Fano [132,133], based on earlier work by Bode [134]. These theoretical limits are known as the Fano-Bode criteria or the Fano-Bode limits. The limits for simple loads are shown in Figure 8-25. More general loads are treated by Fano [132]. The Fano-Bode criteria are used to justify the broad assertion that the more reactive energy stored in a load, the narrow the bandwidth of a match.

The Fano-Bode criteria include the term $1/|\Gamma(\omega)|$ which is the inverse of the magnitude of the reflection coefficient looking into the matching network, as shown in Figure 8-25. A matching network provides matching
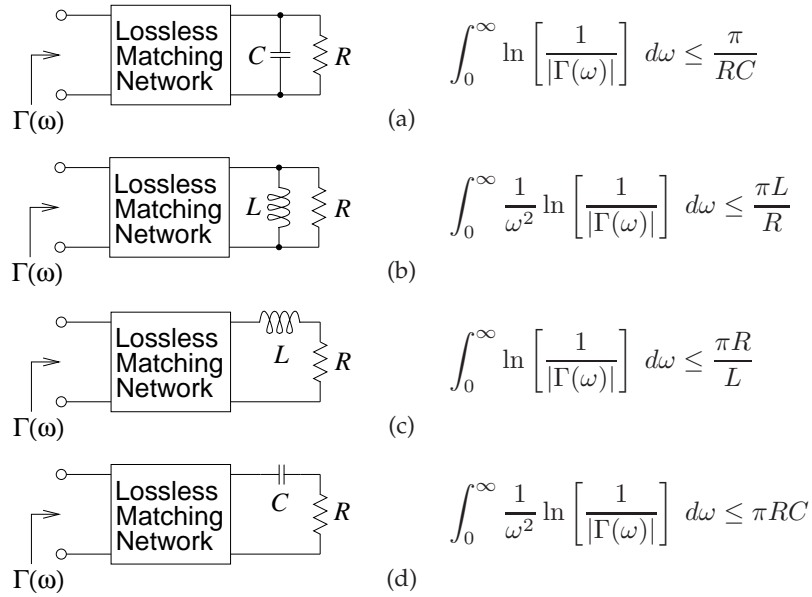
Figure 8-25 Fano-Bode limits for circuits with simple reactive loads: (a) parallel RC load; (b) series RL load; (c) parallel RL load; and (d) series RC load.

over a radian bandwidth BW, and outside the matching frequency band, the magnitude of the reflection coefficient approaches 1. Introducing Γ_{avg} as the average absolute value of $\Gamma(\omega)$ within the passband, and with $f_0 = \omega_0/(2\pi)$ as the center frequency of the match, then, the four Fano-Bode criteria shown in Figure 8-25 can be written as

$$\text{Parallel RC load:} \quad \frac{\mathrm{BW}}{\omega_0} \frac{1}{\Gamma_{\mathrm{avg}}} \leq \frac{\pi}{R(\omega_0 C)} \tag{8.61}$$

$$\text{Parallel RL load:} \quad \frac{\mathrm{BW}}{\omega_0} \frac{1}{\Gamma_{\mathrm{avg}}} \leq \frac{\pi(\omega_0 L)}{R} \tag{8.62}$$

$$\text{Series RL load:} \quad \frac{\mathrm{BW}}{\omega_0} \frac{1}{\Gamma_{\mathrm{avg}}} \leq \frac{\pi R}{(\omega_0 L)} \tag{8.63}$$

$$\text{Series RC load:} \quad \frac{\mathrm{BW}}{\omega_0} \frac{1}{\Gamma_{\mathrm{avg}}} \leq \pi R(\omega_0 C). \tag{8.64}$$

In terms of reactance and susceptance these can be written as

$$\text{Parallel load:} \quad \frac{\mathrm{BW}}{\omega_0} \frac{1}{\Gamma_{\mathrm{avg}}} \leq \frac{\pi G}{B} \tag{8.65}$$

$$\text{Series load:} \quad \frac{\mathrm{BW}}{\omega_0} \frac{1}{\Gamma_{\mathrm{avg}}} \leq \frac{\pi R}{X}, \tag{8.66}$$

where $G = 1/R$ is the conductance of the load, B is the load susceptance, and X is the load reactance. BW/ω_0 is the fractional bandwidth of the matching network. Equations (8.65) and (8.66) indicate that the greater the proportion of energy stored reactively in the load compared to the power dissipated in the load, the smaller the fractional bandwidth, (B/ω_0), for the same average in-band reflection coefficient Γ_{avg}.

Equations (8.65) and (8.66) can be simplified one step further. From Section 8.2.2, the Q of lumped elements can be used to write

$$\frac{\text{BW}}{\omega_0} \frac{1}{\Gamma_{\text{avg}}} \leq \frac{\pi}{Q} , \qquad\qquad (8.67)$$

where Q is that of the load. Several general results can be drawn from Equation (8.67) as follows:

1. If the load stores any reactive energy, so that the Q of the load is non-zero, the in-band reflection coefficient looking into the matching network cannot be zero if the bandwidth of the match is finite.

2. The higher the Q of the load, the narrower the bandwidth of the match for the same average in-band reflection coefficient.

3. The higher the Q of the load, the more difficult it will be to design the matching network to achieve a specified matching bandwidth.

4. A match over all frequencies is only possible if the Q of the load is zero; that is, if the load is resistive. In this case, a resistive load can be matched to a resistive source by using a magnetic transformer. Using a matching network with lumped L and C components will result in a match over a finite bandwidth. However with more L and C elements than just the two considered so far, the bandwidth of the match can be increased.

5. Multi-element matching networks are required to maximize the matching network bandwidth and minimize the in-band reflection coefficient. The matching network design becomes more difficult as the Q of the load increases.

8.6   Multi-Element Matching

With the L network (i.e., two-element matching), the circuit Q is fixed once the source and load resistances, R_S and R_L, are fixed:

$$Q = \sqrt{\frac{R_L}{R_S} - 1}, \qquad\qquad (R_L > R_S). \qquad (8.68)$$

Thus the designer does not have a choice of circuit Q. Introducing a third element in the matching network provides the extra degree of freedom in the design for adjusting Q, and hence bandwidth.
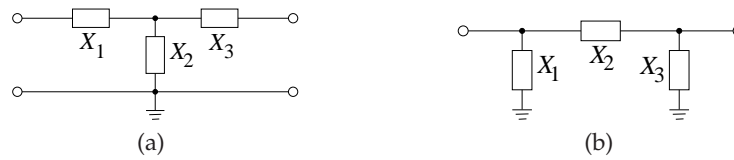
Figure 8-26 Two three-element matching networks: (a) T network; and (b) Pi network.

Two three-element matching networks, the T network and the Pi network, are shown in Figure 8-26. Whether a Pi network or a T network is used depends on

(a) The realization constraints associated with the specific design.

(b) The nature of the reactive parts of the source and load impedances (the type of network that facilitates the handling of these reactive parts using absorption or resonance should be selected).

The three-element matching network comprises 2 two-element (or L) matching networks. Given R_S and R_L, the circuit Q established by an L matching network is the minimum circuit Q available in the three-element matching arrangement. With three-element matching, the Q can only increase, so three-element matching is used for narrowband (high-Q) applications.

8.6.1 The Pi Network

The Pi network may be thought of as two back-to-back L networks that are used to match the load and the source to a virtual resistance, R_V, placed at the junction between the two networks, as shown in Figure 8-27(b). The design of each section of the Pi network is done in exactly the same manner as was done for the L network matching. R_V must be selected smaller than R_S and R_L since it is connected to the series arm of each L section. Furthermore, R_V can be any value smaller than the minimum of R_S, R_L. However, it is customarily used as the design parameter for specifying the desired Q.

As a useful design approximation, we can take the loaded Q of the Pi network as the Q of the L section with the highest Q:

$$Q = \sqrt{\frac{\max(R_S, R_L)}{R_V} - 1}. \qquad (8.69)$$

Given R_S, R_L, and Q, the above equation yields the value of R_V.

(a)

(b)
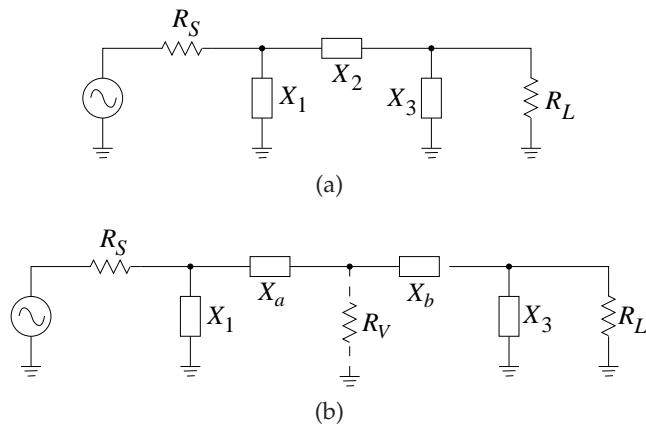
Figure 8-27 Pi matching networks: (a) view of a Pi network; and (b) as two back-to-back L networks with a virtual resistance, R_V, between the networks.

EXAMPLE 8. 7   Three-Element Matching Network Design

Design a Pi network to match a 50 Ω source to a 500 Ω load. The desired Q is 10. A suitable matching network topology is shown in Figure 8-28 together with the virtual resistance, R_V, to be used in design.

Solution:

$$R_s = 50\ \Omega \quad \text{and} \quad R_L = 500\ \Omega, \quad \text{and so} \quad \max(R_S, R_L) = 500\ \Omega. \tag{8.70}$$

Thus the virtual resistor is

$$R_V = \frac{\max(R_S, R_L)}{Q^2 + 1} = \frac{500}{101} = 4.95\ \Omega. \tag{8.71}$$

Design proceeds by separately designing the L networks to the left and right of R_V. For the L network on the left,

$$Q_{\text{left}} = \sqrt{\frac{50}{4.95} - 1} = 3.017, \tag{8.72}$$

and so

$$Q_{\text{left}} = \frac{|X_a|}{R_V} = \frac{R_S}{|X_1|}, \tag{8.73}$$

noting that X_1, X_a must be of opposite types (one is capacitive and the other is inductive). The left L network has elements

$$|X_a| = 14.933\ \Omega \quad \text{and} \quad |X_1| = 16.4\Omega. \tag{8.74}$$
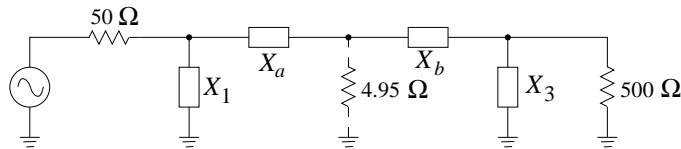
Figure 8-28 Matching network problem of Example 8.7.
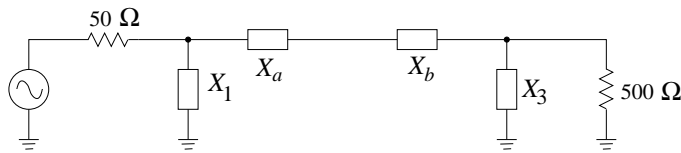
Figure 8-29 Final matching network in Example 8.7.

For the L network on the right of R_V,

$$Q_{\text{right}} = Q = 10, \tag{8.75}$$

thus

$$\frac{|X_b|}{R_V} = \frac{R_L}{X_3} = 10. \tag{8.76}$$

X_b, X_3 are of opposite types, and

$$|X_b| = 49.5\ \Omega \quad \text{and} \quad |X_3| = 50\ \Omega. \tag{8.77}$$

The resulting Pi network is shown in Figure 8-29 with the values

$$|X_1| = 16.4\ \Omega, \quad |X_3| = 50\ \Omega, \quad |X_a| = 14.933\ \Omega, \quad \text{and} \quad |X_b| = 49.5\ \Omega. \tag{8.78}$$

Note that the pair X_a, X_1 are of opposite types and similarly X_b, X_3 are of opposite types. So there are four possible realizations, as shown in Figure 8-30.

In the previous example there were four possible realizations of the three-element matching network, and this is true in general. The specific choice of one of the four possible realizations will depend on specific application-related factors such as

(a) Elimination of stray reactances

(b) The need to pass or block DC current

(c) The need for harmonic filtering.

It is fortunate that it may be possible to achieve multiple functions with the same network.
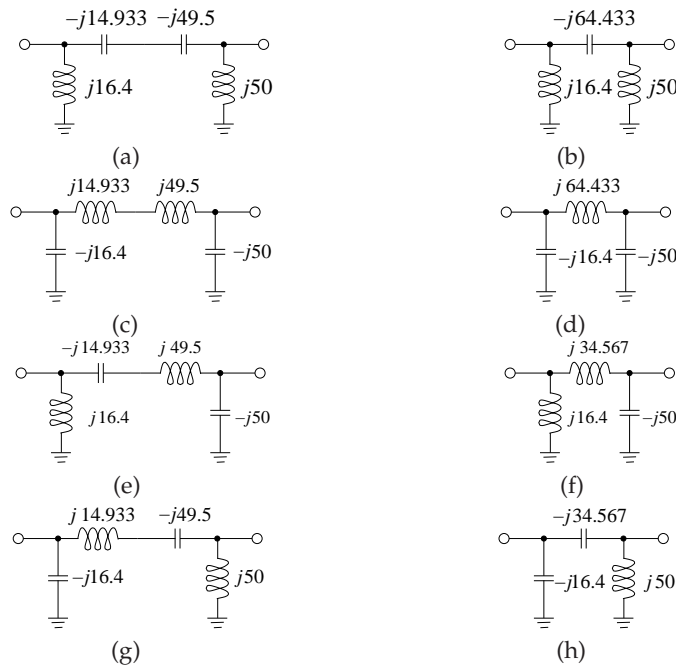
Figure 8-30 Four possible Pi matching networks: (a), (c), (e), (g) conceptual circuits; and (b), (d), (f), (h), respectively, their final reduced Pi networks.
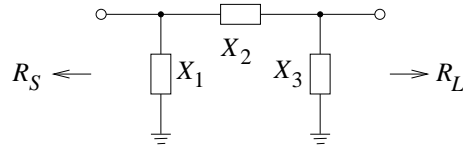
Figure 8-31 A three-element matching network.

8.6.2 Matching Network Q Revisited

To illustrate the fact that the circuit Q established by an L matching network is the minimum circuit Q available in the three-element matching arrangement, consider the design equations for $R_S > R_L$. Referring to Figure 8-31,

$$X_1 = \frac{R_S}{Q} \tag{8.79}$$

$$X_3 = R_L \left(\frac{R_S/R_L}{Q^2 + 1 - R_S/R_L} \right)^{1/2} \tag{8.80}$$

$$X_2 = \frac{QR_S + R_S R_L / X_3}{Q^2 + 1}. \tag{8.81}$$
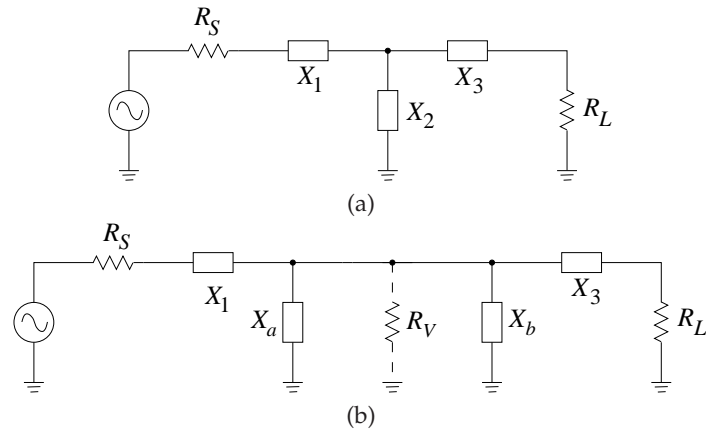
Figure 8-32   T network design approach: (a) T network; and (b) virtual intermediate matching network view.

Notice that the denominator of X_3 can be written as:

$$Q^2 + 1 - \frac{R_S}{R_L} = \left(Q + \sqrt{\frac{R_S}{R_L} - 1} \right) \left(Q - \sqrt{\frac{R_S}{R_L} - 1} \right). \qquad (8.82)$$

Then for a real solution we must have

$$Q \geq \sqrt{\frac{R_S}{R_L} - 1}, \qquad (8.83)$$

and so $Q \geq Q_{\text{L network}}$. For $Q = Q_{\text{L network}}$, $X_3 \to \infty$ and the Pi network reduces to an L network.

8.6.3   The T Network

The T network may be thought of as two back-to-back L networks that are used to match the load and the source to a virtual resistance, R_V, placed at the junction between the two networks (see Figure 8-32. R_V must be selected to be larger than both R_S and R_L since it is connected to the shunt leg of each L section. R_V is chosen according to the equation

$$Q = \sqrt{\frac{R_V}{\min(R_S, R_L)} - 1} \, , \qquad (8.84)$$

where Q is the desired loaded Q of the network. Each L network is calculated in exactly the same manner as was done for the Pi network matching. Once again there will be four possible designs for the T network, given R_S, R_L, and Q.
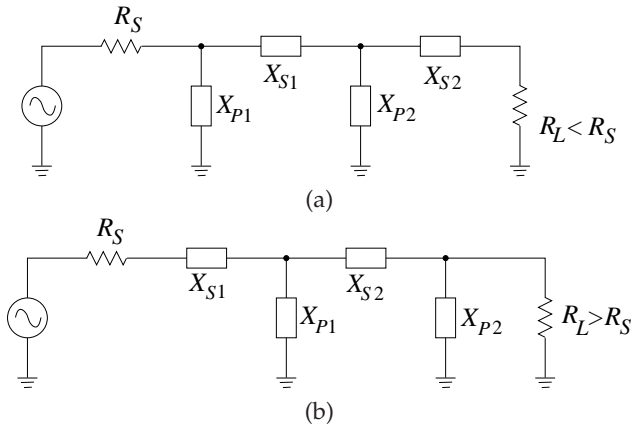
Figure 8-33   Broadband matching networks.
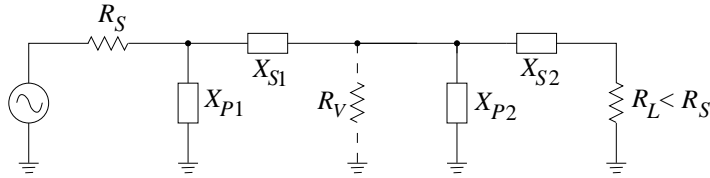
Figure 8-34   Matching network with two L networks.

8.6.4   Broadband (Low Q) Matching

The L network matching does not allow the circuit Q, and hence bandwidth, to be selected. However, Pi network and T network matching allows the circuit Q to be selected independent of the source and load impedances, provided that the chosen Q is larger than that which can be obtained with an L network. Thus the Pi and T networks result in narrower bandwidth designs than of an L network.

One design solution for broadband matching is to use two (or more) series-connected L sections rather than the back-to-back configuration of the Pi or T networks (see Figure 8-33). Design is still based on the concept of a virtual resistor, R_V, placed at the junction of the two L networks (see the circuit of Figure 8-34), but now R_V is chosen to be between R_S and R_L, or mathematically

$$R_{\min} \leq R_V \leq R_{\max}, \tag{8.85}$$

where $R_{\min} = \min(R_L, R_S)$ and $R_{\max} = \max(R_L, R_S)$. Then one of the two networks will have

$$Q_1 = \sqrt{\frac{R_V}{R_{\min}} - 1} \tag{8.86}$$
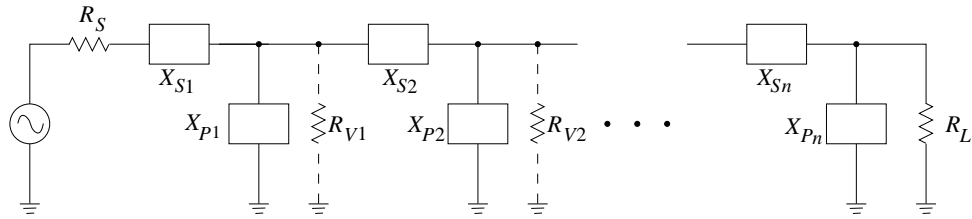
Figure 8-35 Cascaded L networks for broadband matching.

and the other will have

$$Q_2 = \sqrt{\frac{R_{\text{max}}}{R_V} - 1} .$$ (8.87)

The maximum bandwidth (minimum Q) available from this network is obtained when

$$Q_1 = Q_2 = \sqrt{\frac{R_V}{R_{\text{min}}} - 1} = \sqrt{\frac{R_{\text{max}}}{R_V} - 1} .$$ (8.88)

That is, the maximum matching bandwidth is obtained when R_V is the geometric mean of R_S and R_L:

$$R_V = \sqrt{R_L R_S}.$$ (8.89)

Even wider bandwidths can be obtained by cascading more than two L networks, as shown in Figure 8-35. In this circuit

$$R_S < R_{V1} < R_{V2} \ldots < R_{Vn-1} < R_L.$$ (8.90)

For optimum bandwidth the ratios should be equal,

$$\frac{R_V}{R_S} = \frac{R_{V_2}}{R_{V_1}} = \frac{R_{V_3}}{R_{V_2}} = \cdots = \frac{R_L}{R_{V_{n-1}}},$$ (8.91)

and the Q is given by

$$Q = \sqrt{\frac{R_{V_1}}{R_S} - 1} = \sqrt{\frac{R_{V_2}}{R_1} - 1} = \cdots = \sqrt{\frac{R_L}{R_{V_{n-1}}} - 1}.$$ (8.92)

<div style="border:1px solid black">

EXAMPLE 8. 8 Two-Section Matching Network Design

Match a 10 Ω source to a 1000 Ω load using two lumped elements.

Solution:

Here the approximate Qs achieved with a single L matching network and with an optimum two-section design are compared. For a single L network design:

$$Q = \sqrt{\frac{R_L}{R_S} - 1} = 9.95. \tag{8.93}$$

Now consider an optimum two-section design:

$$R_V = \sqrt{R_S R_L}; \ Q_2 = \sqrt{\frac{R_L}{R_V} - 1} = \sqrt{\sqrt{\frac{R_L}{R_S}} - 1} = 3. \tag{8.94}$$

So the Q is 3 compared to the Q of an L section of 9.95. If the fractional bandwidth is inversely proportional to Q then the bandwidth of the two-section design is 9.95/3 = 3.32 times more than that of the L section. Now consider how many sections are required to obtain a Q of 2:

$$(1 + Q^2) = \frac{R_{V_1}}{R_S} = \frac{R_{V_2}}{R_{V_1}} = \ldots = \frac{R_L}{R_{V_{n-1}}} \Rightarrow \tag{8.95}$$

$$\left(1 + Q^2\right)^n = \frac{R_L}{R_S} \Rightarrow n \ln\left(1 + Q^2\right) = \ln \frac{R_L}{R_S} \Rightarrow n = \frac{\ln \frac{R_L}{R_S}}{\ln\left(1 + Q^2\right)}. \tag{8.96}$$

For $Q = 2$ and $R_L/R_S = 100$, $n = 2.86$ is obtained, which rounds to $n = 3$, so three sections are required.

</div>

8.7 Impedance Matching Using Smith Charts

8.7.1 Two-Element Matching

The lumped-element matching networks presented up to now can also be developed using Smith charts which provide a fairly intuitive approach to network design. With experience it will be found that this is the preferred approach to developing designs, as trade-offs can be captured graphically. Smith chart-based design will be presented using examples. This example builds on the preceding lumped-element matching network design and the Smith chart introduction in Section 6.9 on Page 330. Capacitive and inductive regions on the Smith chart are shown in Figure 8-36. In the design examples presented here, circles of constant resistance or constant conductance are followed corresponding to varying reactance.

(a)

(b)
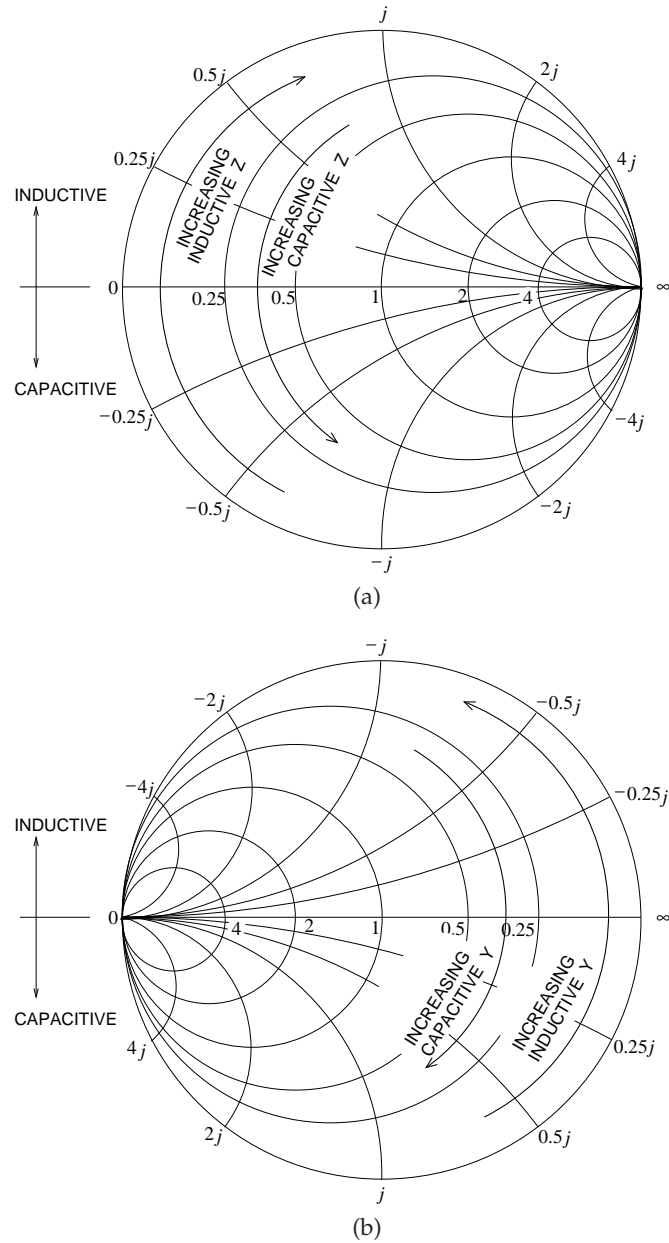
Figure 8-36 Inductive and capacitive regions on Smith charts: (a) impedance Smith chart ; and (b) admittance Smith chart. Increasing capacitive impedance (Z) indicates smaller capacitance; increasing inductive admittance (Y) indicates smaller inductance.

EXAMPLE 8. 9   Two-Element Matching Network Design Using a Smith Chart

Develop a two-element matching network to match a source with an impedance of $R_S = 25$ Ω to a load $R_L = 200$ Ω (see Figure 8-37).

Solution:

The design objective is to present conjugate matched impedances to the source and load; that is, $Z_1 = R_L^*$ and $Z_2 = R_S^*$. Since the source and load impedances are resistive the design can deal with R_S and R_L directly. The load and source resistances are plotted on the Smith chart of Figure 8-38 after choosing a normalization impedance of $Z_0 = 50$ Ω (and so $r_S = R_S/Z_0 = 0.5$ and $r_L = R_L/Z_0 = 4$). The normalized source impedance, r_S, is plotted at Point A, and the normalized load impedance, r_L, is plotted at Point C. The matching network must be lossless, which means that the design must follow lines of constant resistance (on the impedance part of the Smith chart) or constant conductance (on the admittance part of the Smith chart). So Points A and C must be on the above circles and the circles must intersect if a design is possible. The design can be viewed as moving back from the source toward the load or moving back from the load toward the source. (The views result in identical designs.) Here the view taken is moving back from the source toward the load. One possible design is shown in Figure 8-38. From Point A, the line of constant resistance is followed to Point B (there is increasing series reactance along this path). From Point B, the locus follows a line of constant conductance to the final point, Point C. There is also an alternative design that follows the path shown in Figure 8-39. There are only two designs that have a path from A to B following just two arcs. At this point two designs have been outlined. The next step is assigning element values.

The design shown in Figure 8-38 begins with r_S followed by a series reactance, x_S, taking the locus from A to B and then a shunt capacitive susceptance, b_P, taking the locus from B to C and r_L. At Point A the reactance $x_A = 0$, at Point B the reactance $x_B = 1.323$. This value is read off the Smith chart, requiring that an arc as shown be interpolated between the arcs provided. It should be noted that not all versions of Smith charts include negative signs, as they become too complicated. Thus the user needs to be aware and add signs where appropriate. A full Smith chart is much more detailed and any necessary interpolation is much easier. The normalized series reactance is

$$x_S = x_B - x_A = 1.323 - 0 = 1.323; \qquad (8.97)$$

that is,

$$X_S = x_s * Z_0 = 1.323 \times 50 = 66.1 \; \Omega. \qquad (8.98)$$

A shunt capacitive element takes the locus from Point B to Point C and

$$b_P = b_C - b_B = 0 - (-0.661) = 0.661, \qquad (8.99)$$

and so

$$B_P = b_P/Z_0 = 0.661/50 = 132 \text{ mS}, \quad \text{or} \quad X_P = -1/B_P = -75.6 \; \Omega. \qquad (8.100)$$

The final design is shown in Figure 8-40.
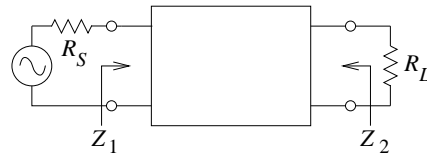
Figure 8-37 Design objectives for Example 8.9.
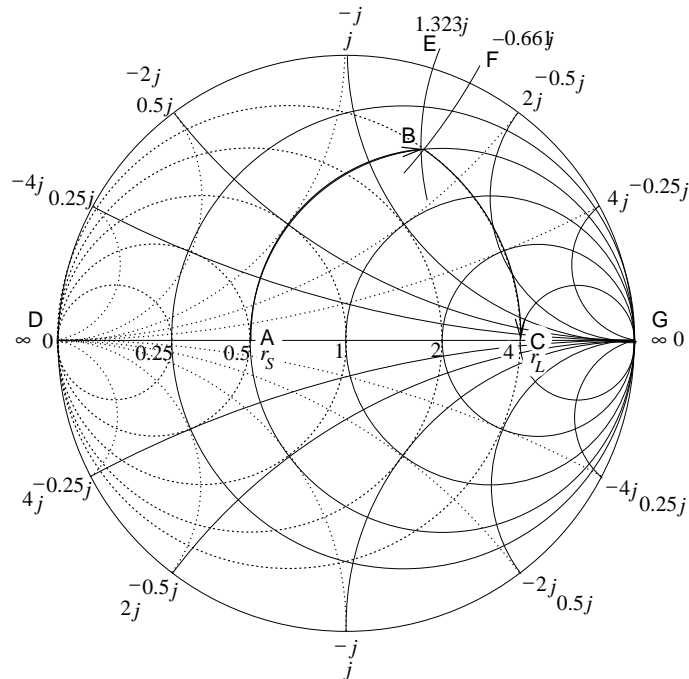
Figure 8-38 Design 1: Smith chart-based design used in Example 8.9. A normalization impedance of $50\ \Omega$ is used.

One of the advantages of using the Smith chart is that design progresses in stages, with the structure of the design developed before actual numerical values are calculated. Of course, it is difficult to extract accurate values from a chart and so designs are regularly roughed out on a Smith chart and refined using CAD tools. Example 8.1 matched resistive source and load to each other. The next example considers matching complex load and source impedances. In the earlier discussion, absorption and resonance were proposed as strategies for dealing with complex terminations. It will be seen that this complication disappears with a Smith chart-based design, as it is conceptually not much different from the resistive problem of Example 8.1.
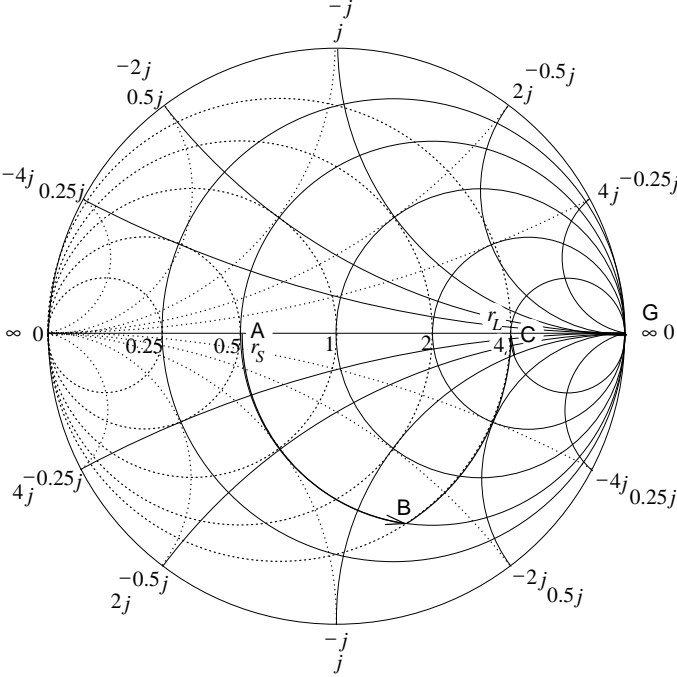
Figure 8-39   Design 2: alternative design for Example 8.9.

$$X_S = 66.1 \ \Omega \ . \quad X_P = -75.6 \ \Omega$$
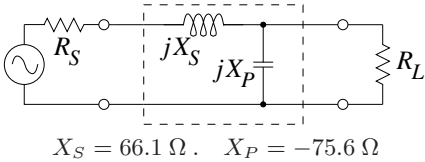
Figure 8-40   Final design for Example 8.9.

EXAMPLE 8. 10   Matching Network Design With Complex Impedances

Develop a two-element matching network to match a source with an impedance of $Z_S = 12.5 + 12.5\jmath \ \Omega$ to a load $Z_L = 50 - 50\jmath \ \Omega$, as shown in Figure 8-41.

Solution:

The design objective is to present conjugate matched impedances to the source and load; that is, $Z_1 = Z_L^*$ and $Z_2 = Z_S^*$. The choice here is to design for Z_1; that is, elements will be inserted in front of Z_L to produce the impedance Z_1. The normalized source and load impedances are plotted in Figure 8-42 using a normalization impedance of $Z_0 = 50 \ \Omega$, so $z_S = Z_S/Z_0 = 0.25 + 0.25\jmath$ (Point S) and $z_L = Z_L/Z_0 = 1 - \jmath$ (Point C).

The impedance to be synthesized is $z_1 = Z_1/Z_0 = z_S^* = 0.25 - 0.25\jmath$ (Point A). (Synthesizing Z_2 results in identical designs.) The matching network must be lossless, which means that the lumped-element design must follow lines of constant resistance (on the impedance part of the Smith chart) or constant conductance (on the admittance part of the Smith chart). Points A and C must be on the above circles and the circles must intersect if a design is possible.

The design can be viewed as moving back from the load impedance toward the conjugate of the source impedance. The direction of the impedance locus is important. One possible design is shown in Figure 8-42. From Point C the line of constant conductance is followed to Point B (there is increasing positive [i.e., capacitive] shunt susceptance along this path). From Point B the locus follows a line of constant resistance to the final point, Point A.

The design shown in Figure 8-42 begins with a shunt susceptance, b_P, taking the locus from Point C to Point B and then a series inductive reactance, x_S, taking the locus to Point C and r_S^*. At Point C the susceptance $b_C = 0.5$, at Point B the susceptance $b_B = 1.323$. This value is read off the Smith chart, requiring that an arc of constant susceptance, as shown, be interpolated between the constant susceptance arcs provided. A full Smith chart is much more detailed and any necessary interpolation is much easier. The normalized shunt susceptance is

$$b_P = b_B - b_C = 1.323 - 0.5 = 0.823, \tag{8.101}$$

that is,

$$B_P = b_P/Z_0 = 0.823/(50 \text{ } \Omega) = 16.5 \text{ mS, or } X_P = -1/B_P = -60.8 \text{ } \Omega. \tag{8.102}$$

A series reactive element takes the locus from Point B to Point A, so

$$x_S = x_A - x_B = -0.25 - (-0.661) = 0.411, \tag{8.103}$$

and so

$$X_S = x_S * Z_0 = 0.411 \times 50 \text{ } \Omega = 20.6 \text{ } \Omega. \tag{8.104}$$

The final design is shown in Figure 8-44.

There are only two designs that have a path from Point C to Point A following just two arcs. In Design 1, shown in Figure 8-43, the Path CBA is much shorter than Path CHA for Design 2 shown in Figure 8-43. The path length is an approximate indication of the total reactance required, and the higher the reactance, the greater the energy storage and hence the narrower the bandwidth of the design. (The actual relative bandwidth depends on the voltage and current levels in the network; the path length criteria, however, is an important rule of thumb.) Thus Design 1 can be expected to have a much higher bandwidth than Design 2. Since designing broader bandwidth is usually an objective, a design requiring a shorter path on a Smith chart is usually preferable.

8.8 Distributed Matching

Matching using lumped elements leads to series and shunt lumped elements. The shunt elements can be implemented using shunt transmission lines, as a short length (less than one-quarter wavelength long) of short-
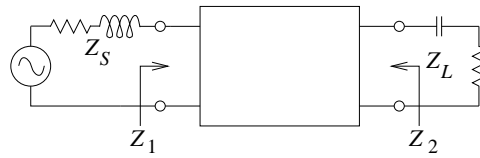
Figure 8-41   Design objectives for Example 8.10.
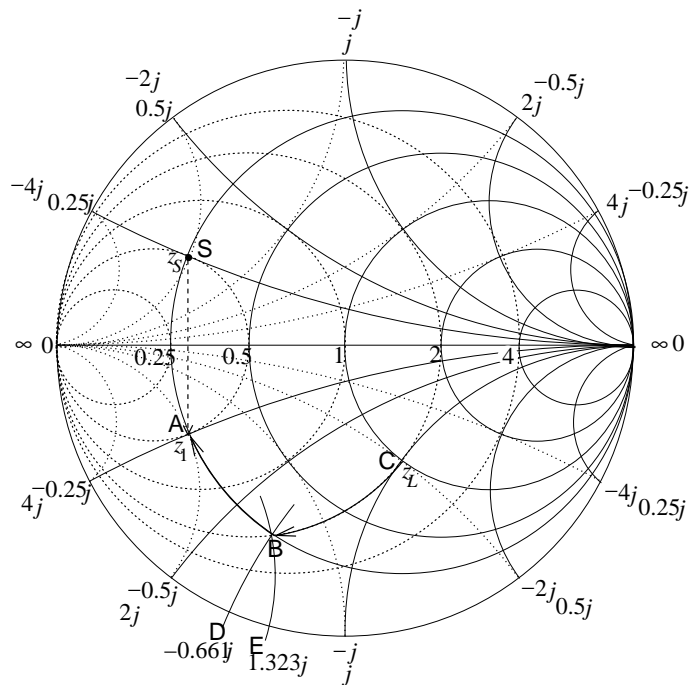
Figure 8-42   Design 1: Smith chart-based design used in Example 8.10. A normalization impedance of 50 Ω is used .

circuited transmission line looks like an inductor and a short section of open-circuited transmission line looks like an capacitor. However, in most transmission line technologies it is not possible to realize the series elements as lengths of transmission lines. While it has been shown that a short length of transmission line is inductive, replacing series inductors by a length of transmission line of high characteristic impedance is not the best approach to realizing networks. The solution is to use the lengths of transmission line together with shunt elements. If space is not at a premium, this is an optimum solution, as transmission lines have much lower loss than a lumped inductor. The series transmission lines rotate the reflection coefficient on the Smith chart. It is also possible to use lumped capacitors as
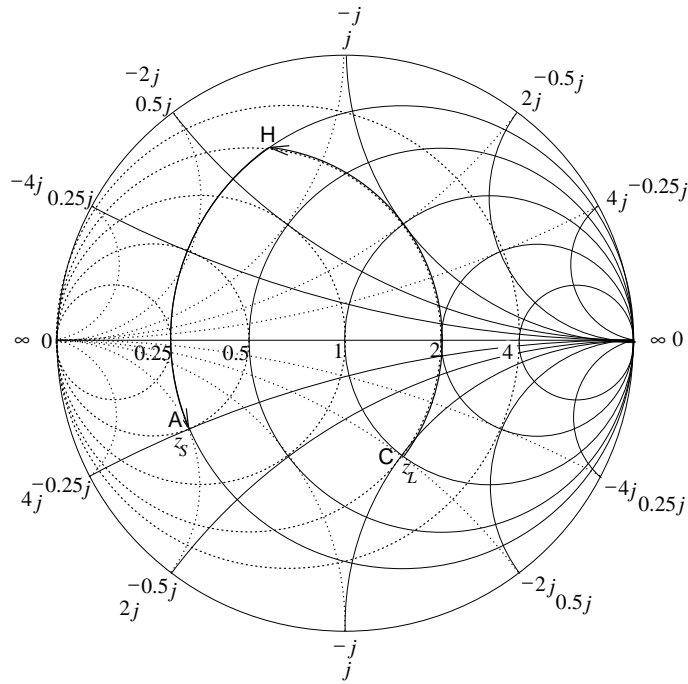
Figure 8-43   Design 2 for Example 8.10.

$$X_S = 20.6 \, \Omega \quad X_P = -60.8 \, \Omega \, .$$
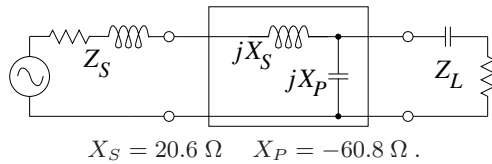
Figure 8-44   Final design for Design 1 of Example 8.10

lumped elements together with lengths of transmission line.

As with all design matching, using transmission lines begins with a topology in mind. Several topologies are shown in Figure 8-45. Figure 8-45(a) is the top view of a microstrip matching network with a series transmission line and stub realized as an open-circuited transmission line. Figure 8-45(b) is a shorthand schematic for this circuit. Matching network design then becomes a problem of choosing the lengths and characteristic impedances of the lines. The stub here is used to realized a capacitive shunt element. This network corresponds to two-element matching with a shunt capacitor. The value of the shunt capacitance can be increased using a dual stub, as shown in Figure 8-45(c), where the capacitive input impedances of each stub are in parallel. The dual circuit to that in Figure 8-
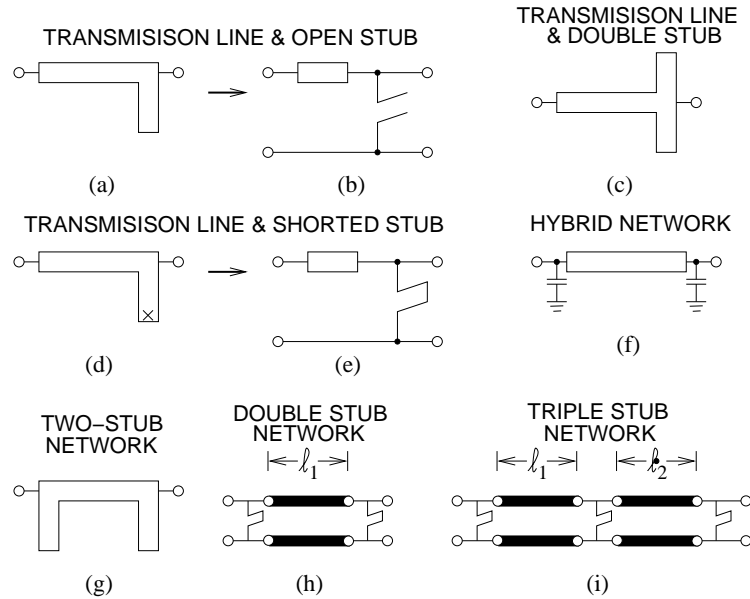
Figure 8-45 Matching networks with transmission line elements:.

45(d) is shown in Figure 8-45(d) together with its schematic representation in Figure 8-45(e). This circuit has a short-circuited stub which realizes a shunt inductance. Mixing lumped capacitors with a transmission line element, as shown in Figure 8-45(f), realizes a much more space-efficient network design. There are many variations to stub-based matching network design, including the two-stub design in Figure 8-45(g). All of these designs require the transmission line lengths and impedances to be designed. A common situation encountered in the laboratory is matching circuits that are in development. Laboratory items available for matching include the stub tuner, shown in Figure 8-46(a), the double-stub tuner shown in Figure 8-46(b), and the double-slug tuner shown in Figure 8-46(c). With this device the length of the series transmission line is fixed, but stubs can have variable length using lengths of transmission lines with sliding short circuits. The double-stub tuner is shown schematically in Figure 8-45(h). Not all impedances can be matched with the circuit, however. The triple-stub tuner shown in Figure 8-45(i) can match all impedances presented to it.

8.8.1 Stub Matching

In this section matching using one series transmission line and one stub will be considered. This corresponds to the microstrip circuit topologies shown in Figures 8-45(a) and 8-45(d). First, consider the terminated transmission line shown in Figure 8-47. When the length, ℓ_1, of the line is zero, the input
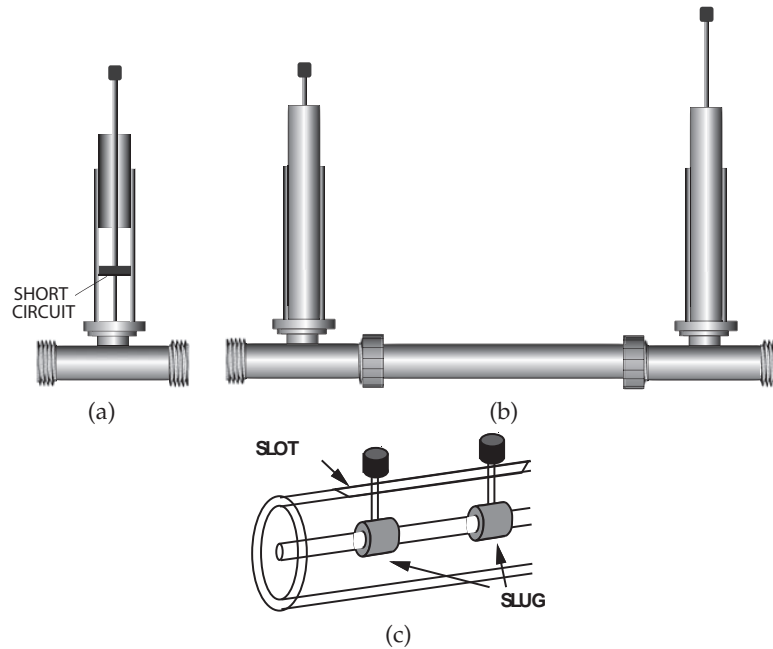
Figure 8-46 Tuning stub: (a) single stub; (b) as parts of a double-stub tuner; and (c) a double-slug tuner where the slugs can be moved up and down the line.
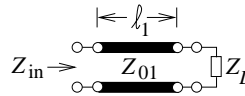
Figure 8-47 Terminated transmission line

impedance of the line, Z_{in}, equals Z_L. How it changes is best described by considering the input reflection coefficient, Γ_{in}, of the line. If the reflection coefficient is normalized to Z_{01} then the magnitude of Γ_{in} and its phase varies as twice the electrical length of the line. This situation is shown in Figure 8-48, where Z_L is chosen arbitrarily. The input reflection coefficient of the line rotates in a clockwise direction as the length of the line increases. One way of remembering this is to consider an open-circuited line. When the line length is zero, Y_{in} is zero and $\Gamma_{\text{in}} = +1$. A short length of line is capacitive so that its reflection coefficient will be in the bottom half of the Smith chart. A length of line can be used to rotate the impedance to an appropriate point to follow a line of constant conductance to the desired input impedance. Single-stub matching will be illustrated by an example.
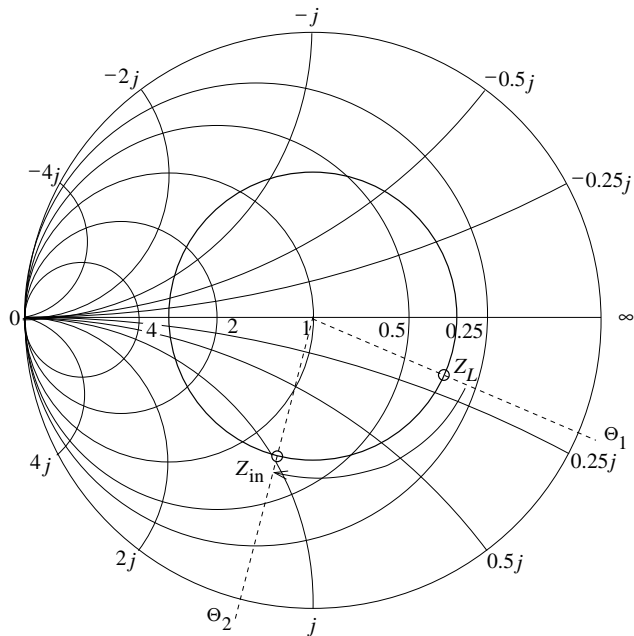
Figure 8-48 Rotation of the input impedance of a transmission line on an admittance Smith chart as the length of line increases.

EXAMPLE 8. 11   Matching Network Design With Transmission Line

Design a two-element matching network to match a source with an impedance of $Z_S = 12.5 + 12.5\jmath\ \Omega$ to a load $Z_L = 50 - 50\jmath\ \Omega$, as shown in Figure 8-41. This example repeats the design in Example 8.10 on Page 445, but now using a transmission line element.

Solution:

As in Example 8.10, choose $Z_0 = 50\ \Omega$ and the design path is from $z_L = Z_L/Z_0 = 1 - \jmath$ to z_s^*, where $z_s = 0.25 + 0.25\jmath$. One possible design solution is indicated in Figure 8-49. The line length, ℓ (taking the locus from Point C to Point B), is

$$\ell = 0.4261\lambda - 0.3125\lambda = 0.1136\lambda,$$

and the normalized shunt susceptance, b_P (taking the locus from Point B to Point C), is

$$b_P = b_A - b_B = 2 - 1 = 1.$$

Thus $X_P = (-1/b_P) \times 50\ \Omega = -50\ \Omega$, so that the final design is as shown in Figure 8-50.
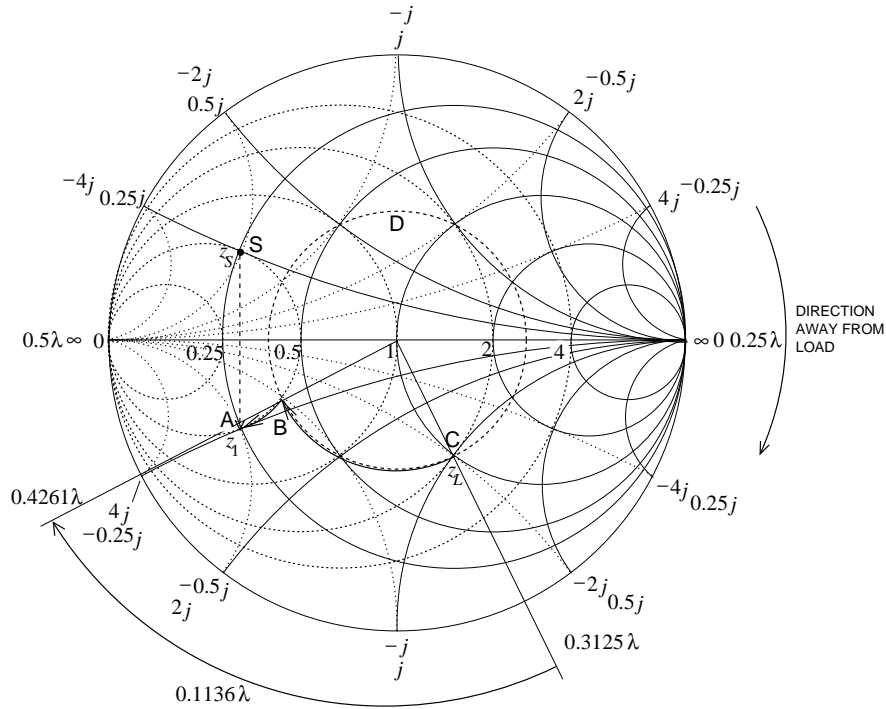
Figure 8-49 Design 1 for Example 8.11.

$$\ell = 0.1136\lambda \quad X_P = -50\,\Omega\,.$$

(a)   (b)
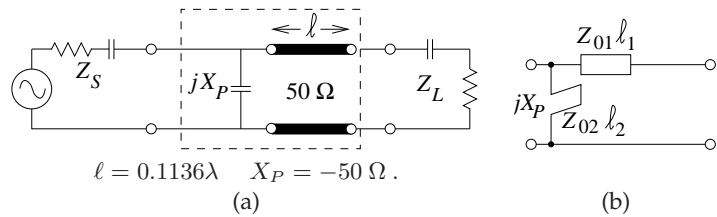
Figure 8-50 Single-stub matching network design of Example 8.11: (a) electrical design; and (b) electrical design with a shunt stub.

8.8.2 Wideband Matching

The strategy for wideband matching is based on the concept of matching to intermediate resistance which is the geometric mean of the source and load impedances. This concept can be represented on a Smith chart using constant Q circles as shown in Figure 8-51. If the load and source impedances, R_L and R_S, are resistive, then the normalizing resistance (R_v) of the Smith chart should be chosen as the geometric mean of the source and load resistance (i.e., $R_v = \sqrt{R_L R_S}$). To maintain a specific circuit Q, and
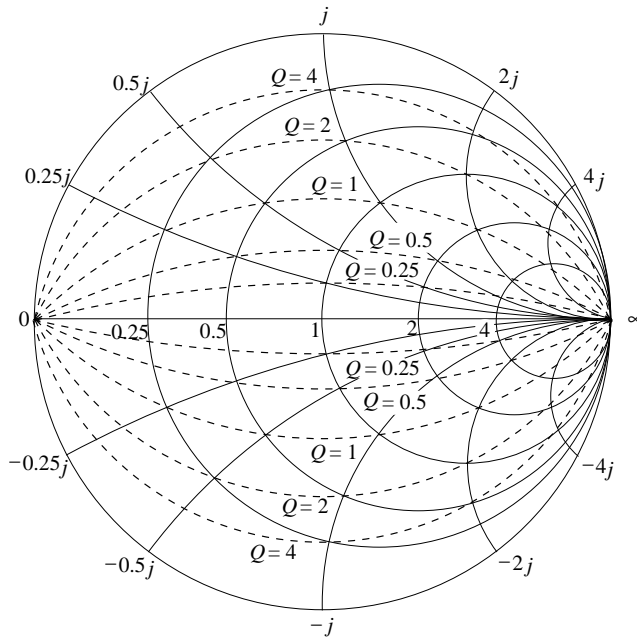
Figure 8-51 Impedance Smith chart with constant Q circles.

hence bandwidth, the locus of impedances in the design must stay with a constant Q circle. It does not matter what elements are used, they can be transmission lines, stubs, or lumped elements.

8.9   Summary

This chapter considered matching using lumped elements. Performance of these matching networks is limited by the self-resonant frequency of lumped elements and by the loss of lumped elements, particularly of inductors. This limits the degree of matching that can be achieved. It is not possible to use an exact synthesis procedure similar to that described here if loss and parasitics are considered. An approach commonly used is to use the synthesis procedure presented in this chapter and then capture the design in a microwave CAD program and use optimization to reoptimize the performance of a matching network. An alternative is to replace the lumped elements by transmission line segments. These generally have lower loss than lumped elements, although their size generally limits operation to a few gigahertz and above.

8.10 Exercises

1. Repeat the analysis in Example 8.1 on Page 419 for the case $B = -1/\omega L$.

2. Develop a two-element matching network for the source/load configuration shown in the Figure below. The matching network must pass DC current. The center frequency of the matching network, $f = 1$ GHz. There are a number of design considerations that should be considered before embarking on network synthesis.

(a) source   (b) load
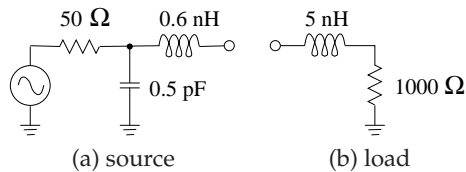
Consider the following:

(i) The source needs to be collapsed to an equivalent circuit with one resistance and one reactance.

(ii) The reactive elements in the source and the load will be accommodated using absorption or resonance. Absorption is preferred, but not always possible.

(iii) The DC current requirements necessitate a lowpass matching network. So there must be a series inductance and a shunt capacitor in the matching network. Since the load resistance is greater than the source resistance, the most likely design has a shunt capacitor on the load side. However, this may change when the transforming properties of the source and load reactances are considered.

(iv) The source reactance should be handled by a series inductor or a shunt capacitor. The input impedance of the source must be considered to determine which.

(v) The load reactance will be resonated out by a shunt capacitor. Looks like absorption will be a possibility at the load.

(a) What is the input impedance of the source? Treat the voltage generator as a short circuit.

(b) What is the reactance of the series element that will resonate the effective input

reactance of the source?

(c) What is the input admittance of the load?

(d) What is the shunt reactance required to resonate the load?

(e) What is the resistive matching problem? That is, since the reactances of the load and source have been resonated out of consideration, what are the effective source and load resistances. Derive the required matching network. Keep the element values of the matching network as reactances.

(f) Draw the complete matching network showing source and load elements required for resonance as well as the matching network for the resistive problem. Keep the element values as reactances.

(g) Draw the final matching network combining all resonant and matching elements. Keep the element values as reactances. This is the electrical design of the matching network.

(h) Calculate the inductance and capacitance values of the matching network.

3. Design a two-element matching network to interface a source with a 25 Ω Thevenin equivalent impedance to a load consisting of a capacitor in parallel with a resistor so that the load admittance is $Y_L = 0.02 + \jmath 0.02$ S. Use the absorption method to handle the reactive load.

4. Design a matching network to interface a source with a 25 Ω Thevenin equivalent impedance to a load consisting of a capacitor in parallel with a resistor so that the load admittance is $Y_L = 0.01 + \jmath 0.01$ S.

(a) If the complexity of the matching network is not limited, what is the minimum Q that could possibly be achieved in the complete network consisting of the matching network and the source and load impedances?

(b) Outline the procedure for designing the matching network for maximum bandwidth if only four elements can be used in the network. You do not need to design the network.

5. A load is modeled as a 50 Ω resistance in series with a reactance of 50 Ω. This load is to be matched to a source with a Thevenin equivalent resistance of 50 Ω. Use the Fano-Bode criteria described in Section 8.5.2 on Page 431 to determine the upper limit on the matching network bandwidth when the average inband reflection coefficient is 1 dB. Note that you must first convert the reflection coefficient to an absolute number.

6. A load is modeled as a 50 Ω resistance in series with a reactance of 50 Ω. This load is to be matched to a source with a Thevenin equivalent resistance of 50 Ω. Use the Fano-Bode criteria described in Section 8.5.2 on Page 431 to determine the upper limit on the matching network bandwidth when the average inband reflection coefficient is 0.1 dB. Note that you must first convert the reflection coefficient to an absolute number.

7. The output of a transistor amplifier operating at 1 GHz is modeled as a 100 Ω resistor in parallel with a 10 pF capacitor. The amplifier must drive the input of a $\lambda/2$ dipole antenna with an input resistance of 73 Ω. To do this efficiently a matching network is required. Consider that the input resistance of the antenna is independent of frequency, and assume that the matching network is lossless. This is the same as assuming that its bandwidth is much greater than the bandwidth required. If the required bandwidth of the matching network is 5%, and using the Fano-Bode criteria (see Section 8.5.2 on Page 431), determine the following:

    (a) The lower limit on the average in-band reflection coefficient of the matching network.

    (b) The upper limit on the average transmission coefficient of the matching network.

8. The output of a transistor amplifier is modeled as a current source in parallel with both a 50 Ω resistor and a 1 pF capacitor. This is to be matched to a load consisting of a 25 Ω resistor in series with a 0.02 nH inductor. The task is to design a matching network that will enable DC bias to be applied from the load to the transistor output, thus the matching network must be a lowpass type. The center frequency of the system is 10 GHz and a bandwidth of 50 MHz is required.

    (a) What is the fractional bandwidth of the system?

    (b) What is the Q of the system?

    (c) Indicate the form of the matching network if no more than four reactive elements are to be used; that is, sketch the matching network.

    (d) Complete the design of the amplifier providing numerical element values.

9. Design a passive matching network that will achieve maximum bandwidth matching from a source with an impedance of 2 Ω (typical of the output impedance of a power amplifier) to a load with an impedance of 50 Ω. The matching network can have a maximum of three reactive elements. You need only calculate reactances and not the capacitor and inductor values.

10. Design a passive matching network that will achieve maximum bandwidth matching from a source with an impedance of 20 Ω to a load with an impedance of 125 Ω. The matching network can have a maximum of four reactive elements. You need only calculate reactances and not the capacitor and inductor values.

    (a) Will you use two, three, or four elements in your matching network?

    (b) With a diagram and perhaps equations, indicate the design procedure.

    (c) Design the matching network. It is sufficient to use reactance values. The design must be complete.

11. Design a passive matching network that will achieve maximum bandwidth matching from a source with an impedance of 60 Ω (typical of the output impedance of a power amplifier) to a load with an impedance of 5 Ω. The matching network can have a maximum of four reactive elements. You need only calculate reactances and not the capacitor and inductor values.

(a) Will you use two, three, or four elements in your matching network?

(b) With a diagram and perhaps equations indicate the design procedure.

(c) Design the matching network. It is sufficient to use reactance values. The design must be complete.

12. Design a three-lumped-element matching network that interfaces a source with an impedance of 5 Ω to a load with an impedance consisting of a resistor with an impedance of 10 Ω. The matching network must have a Q of 6.

13. Design a Pi network to match the source configuration of Figure **??**(a) to the load configuration shown in Figure **??**(b). The design frequency is 900 MHz and the desired Q is 10.

14. Design a T network to match a 50 Ω source to a 1000 Ω load. The desirable loaded Q is 15.

15. A source with a Thevenin equivalent impedance of 75 Ω must drive a load with an impedance of 5 Ω. A matching network with maximum possible bandwidth between the source and the load must be designed to achieve maximum power transfer. Design the matching network for maximum possible bandwidth using no more than four reactive elements.

   (a) Sketch the schematic of the matching network.

   (b) Describe the design procedure.

   (c) Complete the design of the matching network. Determine the values of the elements if the center frequency is 1 GHz.

16. The output of a transistor amplifier is modeled as a current source in parallel with both a 50 Ω resistor and a 1 pF capacitor. This is to be matched to a load consisting of a 25 Ω resistor in series with a 0.02 nH inductor. The task is to design a matching network that will enable DC bias to be applied from the load to the transistor output. Thus the matching network must be a lowpass type. The center frequency of the system is 10 GHz and a bandwidth of 50 MHz is required.

   (a) What is the fractional bandwidth of the system?

(b) What is the Q of the system?

(c) Indicate the form of the matching network if no more than four reactive elements are to be used? That is, sketch the matching network.

(d) Complete the design of the amplifier, providing numerical element values.

17. Repeat Example 8.9 on Page 443 for the case $X_S = \omega L$. That is, show that in a series RL circuit, the inductance can be adjusted to obtain any positive shunt resistance value.

18. A two-port matching network is shown below with a generator and a load. The generator impedance is 40 Ω and the load impedance is $Z_L = 50 - j20$ Ω. Use a Smith chart to design the matching network.

(a) What is the condition for maximum power transfer from the generator? Express your answer using impedances.

(b) What is the condition for maximum power transfer from the generator? Express your answer using reflection coefficients.

(c) What system reference impedance are you going to use to solve the problem?

(d) Plot Z_L on the Smith chart and label the point. (Remember to use impedance normalization if required.)

(e) Plot Z_G on the Smith chart and label the point.

(f) Design a matching network using only transmission lines. Show your work on the Smith chart. You must express the lengths of the lines in terms of electrical length (either degrees or wavelengths long). Characteristic impedances of the lines are required. (You will therefore have a design that consists of one stub and one other length of transmission line.)

19. A two-port matching network is shown below with a generator and a load. The generator impedance is 60 Ω and the load impedance is $Z_L = 30 + j30$ Ω. Use a Smith chart to design a lossless matching network (no credit will be given unless a Smith chart is used to solve this problem). It is important that your solution can be followed, so you must indicate your solution clearly on the chart. Write out your answer on a piece of paper separate from the Smith chart you use.
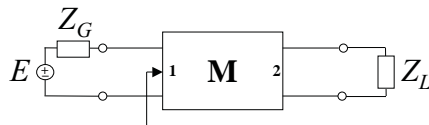
(a) What is the condition for maximum power transfer from the generator? Express your answer using impedances.
(b) What is the condition for maximum power transfer from the generator? Express your answer using reflection coefficients.
(c) What system reference impedance are you going to use to solve the problem?
(d) Plot Z_L on the Smith chart and label the point. (Remember to use impedance normalization if required.)
(e) Plot Z_G on the Smith chart and label the point.
(f) Design a lossless matching network showing your design process on the Smith chart. Label critical points on the Smith chart. Draw the matching network and show the reactance values.

20. A two-port matching network is shown below with a generator and a load. The generator impedance is 30 Ω and the load impedance is $Z_L = 90 - j30$ Ω. Use a Smith chart to design a lossless matching network.
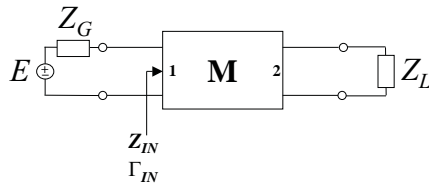
(a) What is the condition for maximum power transfer from the generator? Express your answer using impedances.
(b) What is the condition for maximum power transfer from the generator? Express your answer using reflection coefficients.
(c) What system reference impedance are you going to use to solve the problem?
(d) Plot Z_L on the Smith chart and label the point. (Remember to use impedance normalization if required.)
(e) Plot Z_G on the Smith chart and label the point.
(f) Design a lossless matching network showing your design process on the Smith chart. Label critical points on the Smith chart. Draw the matching network and show the reactance values.

21. Use Smith chart techniques to design a double-stub matching network to match a load with a normalized admittance $y_L = 0.7 - j5$. The stubs are short-circuited and are separated by a length of transmission line of length $\lambda/8$. The load is at the position of the first stub. All transmission lines have the system characteristic impedance. Your design should yield the lengths of the two stubs. (Note that there is a Smith chart for your use on the next page.)

(a) Plot the load on a Smith chart. Clearly indicate the load.
(b) Determine the admittances of each of the stubs. Clearly show and describe your design technique so that it can be understood. Label your efforts on the Smith chart and refer to them here. Note that a description is required and not simply markings on the Smith chart.
(c) Determine the electrical lengths of the stubs (express your answer in terms of wavelengths or degrees).

22. Use a lossless transmission line and a series reactive element to match a source with a Thevenin equivalent impedance of $25 + j50$ Ω to a load of 100 Ω. (That is, use one transmission line and one series reactance only.)

(a) Draw the matching network with the source and load.

(b) What is the value of the series reactance in the matching network (you can leave this in ohms).

(c) What is the length and characteristic impedance of the transmission line?

23. Consider a load $Z_L = 100 - \jmath150\ \Omega$. Use the Smith chart to design a two-stub matching network that will match the load to a generator of 50 Ω. Use 50 Ω transmission lines throughout and assume that the load is immediately next to the first stub. The two stubs are separated by a line with an electrical length of $45°$. Both stubs are short-circuited.

(a) Draw the matching stub system.

(b) What is the normalized load impedance?

(c) Briefly indicate the procedure used to design the two-stub matching network. You will need to use stylized Smith charts.

(d) Plot the load on a Smith chart.

(e) What is the admittance of the first stub (Stub 1)?

(f) What is the electrical length of Stub 1? (Note that the stub is short-circuited.)

(g) What is the admittance of the second stub near the generator (Stub 2)?

(h) What is the electrical length of Stub 2? (Note that the stub is short-circuited.)

24. Consider a load $Z_L = 80 + \jmath40\ \Omega$. Use the Smith chart to design a matching network consisting of only two transmission lines that will match the load to a generator of 40 Ω.

(a) Draw the matching network with transmission lines. If you use a stub, it should be a short-circuited stub.

(b) Indicate your choice of characteristic impedance of your transmission lines. What is the normalized load impedance? What is the normalized source impedance?

(c) Briefly outline the design procedure you will use. You will need to use Smith chart sketches.

(d) Plot the load and source on the Smith chart.

(e) Complete the design of the matching network, providing the lengths of the transmission lines.

25. A two-port matching network is shown below with a generator and a load. The generator impedance is 40 Ω and the load impedance is $Z_L = 20 - \jmath50\ \Omega$. Use a Smith chart to design the matching network.

(a) What is the condition for maximum power transfer from the generator? Express your answer using impedances.

(b) What is the condition for maximum power transfer from the generator? Express your answer using reflection coefficients.

(c) What system reference impedance are you going to use to solve the problem?

(d) Plot Z_L on the Smith chart and label the point. (Remember to use impedance normalization if required.)

(e) Plot Z_G on the Smith chart and label the point.

(f) Design a matching network using only transmission lines, show your work on the Smith chart. You must express the lengths of the lines in terms of electrical length (either degrees or wavelengths long). Characteristic impedances of the lines are required. (You will therefore have a design that consists of one stub and one other length of transmission line.)

CHAPTER 9

Coupled Lines and Applications

9.1 Introduction

This chapter describes what happens when two transmission lines are so close together that the fields produced by one line interfers with the fields produced by the other line. The result is that a portion of the signal energy on one line is transferred to the other. This structure results in significant coupling, as the lines are relatively close. Generally the coupling of the lines is small enough to be ignored if the separation of the strips is, at least, three times the height of the strips.

Two parallel transmission lines are regarded as coupled lines if the EM fields created by the voltage and current on one line induce a voltage or current on another line. With most planar transmission lines the fields extend indefinitely. This is the case with the two parallel microstrip
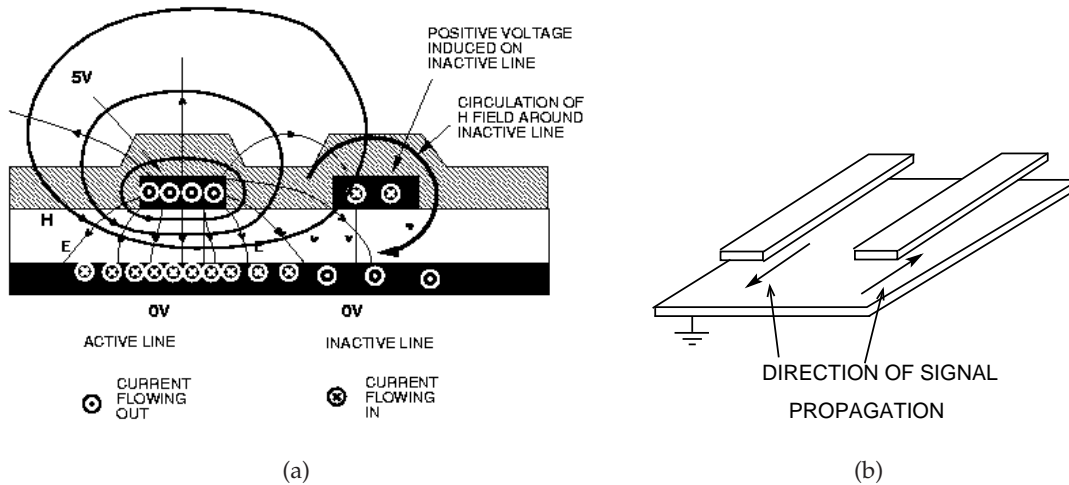
Figure 9-1 Edge-coupled microstrip lines with the left line driven with the forward-traveling field coming out of the page: (a) crosssection of microstrip lines as found on a printed circuit board showing a conformal top passivation or solder resist layer; and (b) in perspective showing the direction of propagation of the driven line (left) and the direction of propagation of the main coupled signal on the victim line (right).

interconnects (see Figure 9-1) ,where only the left line is driven. Part of the magnetic field on the driven or active line (as shown) encompasses the quiet or inactive line on the right, inducing current flow. The electric field also extends from the active interconnect and induces a voltage on the inactive (or victim) line. The result, in the frequency domain, is that a forward-traveling wave on the signal line induces a reverse-traveling wave on the quiet line. A much smaller forward-traveling wave is also induced on the quiet line. Transmission line coupling may be undesirable in many situations, but the phenomenon can be exploited to realize novel RF and microwave elements that selectively tap a portion of the energy of either the forward- or the backward-traveling waves.

9.2 Physics of Coupling

If the fields of one transmission line can intersect the region around another transmission line, then some of the energy propagating on the first line appears on the second. This coupling discriminates in terms of forward- and backward-traveling waves. In particular, consider the coupled microstrips shown in Figure 9-1. The direction of propagation is $\mathbf{E} \times \mathbf{H}$. Using the right-hand rule, the direction of propagation on the left-hand strip is out of the page. This is called the forward-traveling direction. Now consider

the fields around the right-hand strip and note the resulting direction of the induced current and the H fields. By applying $\mathbf{E} \times \mathbf{H}$ to the signal induced on the right-hand strip, it is seen that the induced signal travels into the page or in the backward-traveling direction. The coupling with parallel-coupled microstrip lines is called backward-wave coupling. This is, of course, a frequency-domain view of coupling. In practice, there is also a small component of forward coupling, and in digital systems, this is of most concern. To appreciate this, consider a voltage step traveling in the forward direction on one strip. The step edge couples across to the neighboring line producing a large backward-traveling pulse and a forward-traveling pulse (approximately the derivative of the original step signal). The coupled forward-traveling pulse travels at the same velocity as the step signal so that the coupled forward-traveling signal integrates over time, whereas the backward-traveling signal does not.

Previously forward- and backward-traveling waves were introduced as modes; each is a self-sustaining and propagating field orientation on a transmission line. In understanding coupled line behavior, and to facilitate design using transmission line coupling, it is a good idea to treat a pair of coupled lines as a single transmission structure. This structure supports two modes called the even mode and the odd mode, and both have forward- and backward-traveling components (see Figure 9-2).

In the even mode (Figures 9-2(a) and 9-2(b)), the amplitude and polarity of the voltages on the two signal conductors are the same. In the odd mode (Figures 9-2(c) and 9-2(d)), the voltages on the two signal conductors are equal but have opposite polarity. Any EM field orientation on the coupled lines can be represented as a weighted addition of the even-mode and odd-mode field configurations. The even and odd modes have forward- and backward-traveling wave components.

It is just as valid to consider the field orientations based on each of the lines (taken one at a time) and to consider the effect of the other line as a perturbation of the fields of the first line. However, the odd- and even-mode view of the fields assists in understanding and quantitatively describing coupling. This view facilitates the design of components that exploit coupling. The configuration of the fields supported by the coupled lines depends on how the lines are driven and terminated. The actual field configurations will be a superposition of the even and odd modes. At any instant, the relative amplitudes and polarities of the voltages, taken at any specific crosssectional plane along the structure, can also be categorized as being composed of an alike (or even) component, and an opposite (or odd) component. The actual voltages and currents will be linear combinations of these.

The characteristic impedances of the even and odd modes, on the coupled line system, must differ because of the different field orientations. These impedances are termed the even-mode and odd-mode characteristic
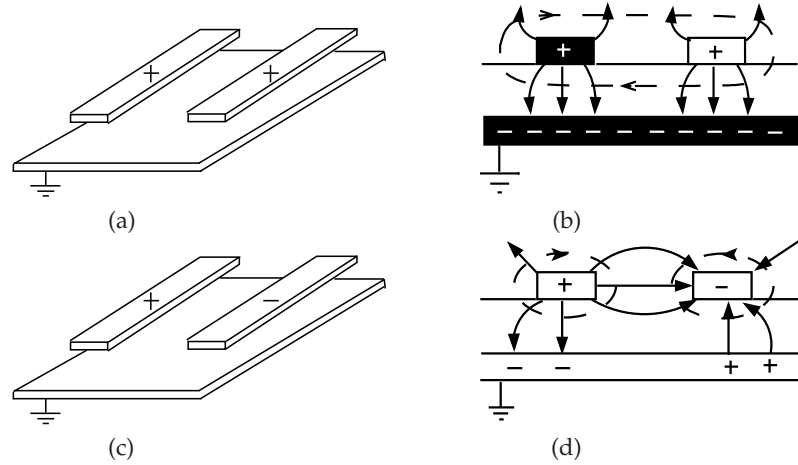
Figure 9-2 Modes on parallel-coupled microstrip lines: (a) and (b) even mode; and (c) and (d) odd mode. In (b) and (d) the electric fields are indicated as arrowed lines and the magnetic field lines are dashed.

impedances, denoted by Z_{0e} and Z_{0o}, respectively. Here the "e" subscript identifies the even mode and the "o" subscript identifies the odd mode. The phase velocities of the two modes may also differ in general, but in the case of a uniform dielectric, the phase velocities will be the same. With coupled microstrip lines they do differ, since the two modes have different amounts of energy in the air and in the dielectric, resulting in different effective permittivities of the two modes.

Coupled lines are modeled by determining the propagation characteristics of the even and odd modes. Also, at the circuit level, coupling can be described by network parameters that relate total voltage and current. Using the definitions shown in Figures 9-3 and 9-4, the total voltages and currents on the original structure are a superposition of the even- and odd-mode solutions:

$$\left.\begin{array}{ll} V_1 = V_{Fe} + V_{Fo} & I_1 = I_{Fe} + I_{Fo} \\ V_2 = V_{Fe} - V_{Fo} & I_2 = I_{Fe} - I_{Fo} \\ V_3 = V_{Be} + V_{Bo} & I_3 = I_{Be} + I_{Bo} \\ V_4 = V_{Be} - V_{Bo} & I_4 = I_{Be} - I_{Bo} \end{array}\right\}, \qquad (9.1)$$

with the subscripts F and B indicating the front and back, respectively, of the even- and odd-mode components. Also

$$\left.\begin{array}{ll} V_{Fe} = (V_1 + V_2)/2 & I_{Fe} = (I_1 + I_2)/2 \\ V_{Fo} = (V_1 - V_2)/2 & I_{Fo} = (I_1 - I_2)/2 \\ V_{Be} = (V_3 + V_4)/2 & I_{Be} = (I_3 + I_4)/2 \\ V_{Be} = (V_3 - V_4)/2 & I_{Be} = (I_3 - I_4)/2 \end{array}\right\}. \qquad (9.2)$$

EVEN ODD
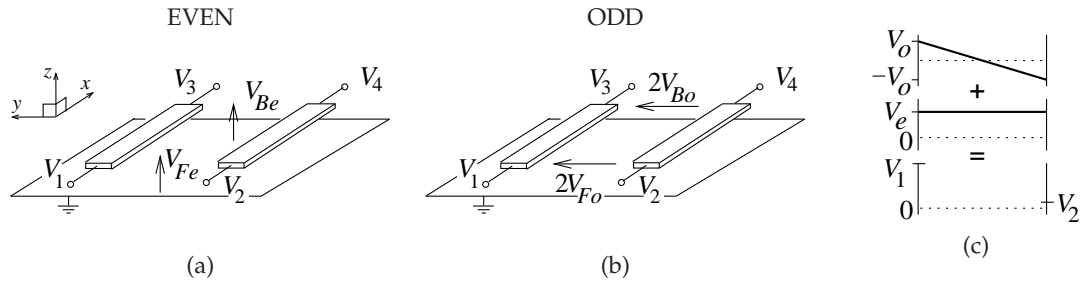
Figure 9-3 Definitions of total voltages and even- and odd-mode voltages on a pair of coupled microstrip lines: (a) even-mode voltage definition; (b) odd-mode voltage definition; (c) depiction of how even- and odd-mode voltages combine to yield the total voltages on individual lines. The subscript "F" indicates front and "B" indicates back.
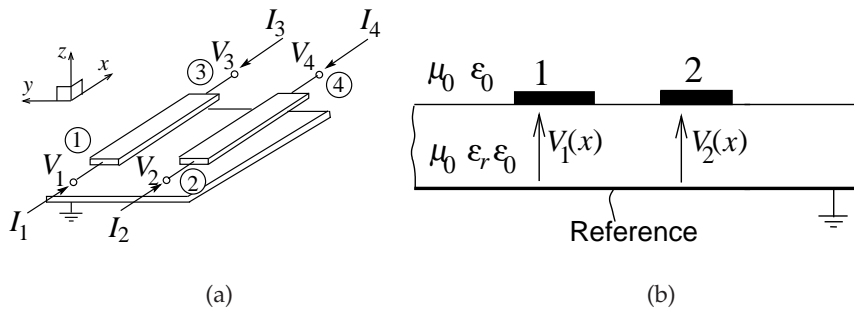
Figure 9-4 Coupled microstrip lines: (a) with total voltages and currents at the four terminals; and (b) crosssection.

Similarly the forward-traveling components can be written as

$$\left.\begin{array}{ll} V_1^+ = V_{Fe}^+ + V_{Fo}^+ & I_1^+ = I_{Fe}^+ + I_{Fo}^- \\ V_2^+ = V_{Fe}^+ - V_{Fo}^+ & I_2^+ = I_{Fe}^+ - I_{Fo}^- \\ V_3^+ = V_{Be}^+ + V_{Bo}^+ & I_3^+ = I_{Be}^+ + I_{Bo}^- \\ V_4^+ = V_{Be}^+ - V_{Bo}^+ & I_4^+ = I_{Be}^+ - I_{Bo}^- \end{array}\right\}. \tag{9.3}$$

The backward-traveling components are written similarly so that the total even- and odd-mode voltages are

$$V_{Fe} = V_{Fe}^+ + V_{Fe}^- \tag{9.4}$$

$$V_{Fo} = V_{Fo}^+ + V_{Fo}^- \tag{9.5}$$

$$V_{Be} = V_{Be}^+ + V_{Be}^- \tag{9.6}$$

$$V_{Bo} = V_{Bo}^+ + V_{Bo}^-. \tag{9.7}$$

With an ideal voltmeter,[1] the V_1, V_2, V_3, and V_4 voltages would be measured from a point on the strip to a point on the ground plane immediately below. V_{Fo} and V_{Bo} are the voltages that would be measured between the strips at the front and back of the lines, respectively. It would not be possible to directly measure V_{Fe} and V_{Be}.

One set of network parameters describing a pair of coupled lines is the port-based admittance matrix equation relating the port voltages, V_1 to V_4, and the port currents, I_1 to I_4:

$$\begin{bmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & y_{13} & y_{14} \\ y_{21} & y_{22} & y_{23} & y_{24} \\ y_{31} & y_{32} & y_{33} & y_{34} \\ y_{41} & y_{42} & y_{43} & y_{44} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix}. \tag{9.8}$$

Since microstrip lines are fabricated (most of the time) on dielectrics that have unity permeability and the permittivity is independent of direction, then the y-parameters are reciprocal so that $y_{ij} = y_{ji}$. Coupling from one line to another is described by the terms: y_{12} ($=y_{21}$), and y_{34} ($=y_{43}$).

9.3 Coupled Transmission Line Theory

In this section coupled transmission line theory is developed in terms of the quantities shown in Figure 9-4. The voltages and currents shown here are phasors that vary along the line and are functions of x.

The quasi-TEM mode of propagation is also assumed, and the transmission line system is completely lossless with perfect conductors and insulators. In phasor form, the generalized telegrapher's equations for a pair of coupled lines are[2]

$$\frac{\mathrm{d}V_1(x)}{\mathrm{d}x} = -\jmath\omega L_{11}I_1(x) - \jmath\omega L_{12}I_2(x) \tag{9.9}$$

$$\frac{\mathrm{d}V_2(x)}{\mathrm{d}x} = -\jmath\omega L_{21}I_1(x) - \jmath\omega L_{22}I_2(x) \tag{9.10}$$

$$\frac{\mathrm{d}I_1(x)}{\mathrm{d}x} = -\jmath\omega C_{11}V_1(x) - \jmath\omega C_{12}V_2(x) \tag{9.11}$$

$$\frac{\mathrm{d}I_2(x)}{\mathrm{d}x} = -\jmath\omega C_{21}V_1(x) - \jmath\omega C_{22}V_2(x). \tag{9.12}$$

These are generalizations of the telegrapher's equations of a single transmission line, and because of reciprocity, $C_{12} = C_{21}$ and $L_{12} = L_{21}$.

---

[1] An ideal voltmeter does not distort the signals and can measure the AC voltage amplitude at any frequency.

[2] In Figure 9-4(a) V_1 and V_2 were the voltages at terminals 1 and 2 of lines 1 and 2. From here on $V_1(x)$ and $V_2(x)$ will be used as the total voltage anywhere on lines 1 and 2, respectively, as shown in Figure 9-4(b). Similar notation will be used for voltages and traveling-wave components.

Compaction of the equations is obtained by introducing the per-unit-length inductance matrix, defined as

$$\mathbf{L} = \begin{bmatrix} L_{11} & L_{12} \\ L_{12} & L_{22} \end{bmatrix}, \qquad (9.13)$$

and the per-unit-length capacitance matrix defined as

$$\mathbf{C} = \begin{bmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{bmatrix}. \qquad (9.14)$$

The next step is to express the voltage on the pair of coupled transmission lines in vector form as

$$\mathbf{V}(x) = \begin{bmatrix} V_1(x) \\ V_2(x) \end{bmatrix} = [V_1(x) \ V_2(x)]^{\mathrm{T}}, \qquad (9.15)$$

where T indicates transpose. Similarly the vector of currents on the coupled transmission lines is

$$\mathbf{I}(x) = [I_1(x) \ I_2(x)]^{\mathrm{T}}. \qquad (9.16)$$

Using the above relations, the telegrapher's equation, from Equations (9.9)–(9.12), is represented in matrix form as

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbf{V}(\mathrm{x}) \;=\; -\jmath\omega\mathbf{L}\mathbf{I}(x) \qquad (9.17)$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbf{I}(\mathrm{x}) \;=\; -\jmath\omega\mathbf{C}\mathbf{V}(x). \qquad (9.18)$$

Rearranging Equations (9.17) and (9.18) and after substitution, the final form is obtained:

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\mathbf{V}(x) + \omega^2\mathbf{L}\mathbf{C}\mathbf{V}(x) \;=\; 0 \qquad (9.19)$$

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\mathbf{I}(x) + \omega^2\mathbf{L}\mathbf{C}\mathbf{I}(x) \;=\; 0. \qquad (9.20)$$

Solving these second-order differential equations yields descriptions of the propagation characteristics. With confidence, a solution in the form of a propagating wave can be assumed:

$$\mathbf{V}(x) = \mathbf{V}_0\, e^{-\jmath\beta x}. \qquad (9.21)$$

Substituting this into Equation (9.19) yields

$$-\beta^2\mathbf{V}_0 + \omega^2\,\mathbf{L}\mathbf{C}\mathbf{V}_0 = 0. \qquad (9.22)$$

For a nontrivial solution of Equation (9.22), the determinant of the matrix equation should be zero:

$$\det\left(\mathbf{L}\mathbf{C} - \frac{\beta^2}{\omega^2}\mathbf{u}\right) = 0, \qquad (9.23)$$

where $\mathbf{u}$ is the unit matrix:

$$\mathbf{u} = \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right]. \tag{9.24}$$

Equation (9.23) is the characteristic equation that can be solved to determine the phase constant, β. There are many possible solutions, and weighted linear combinations of the solutions are also a solution.

If only the quasi-TEM modes are considered, there are two possible sets of solutions for the phase constant, with one set of solutions being

$$\beta_1 = \pm \omega S_1 \tag{9.25}$$

and the other

$$\beta_2 = \pm \omega S_2 . \tag{9.26}$$

The different signs here refer to the forward- and backward-traveling modes.[3] Thus the coupled pair of conductors supports two families of modes (each family comprising forward- and backward-traveling waves) with each set relating to a particular field configuration on the coupled line system.[4] That is, S_1 and S_2 are each single numbers and, just considering the forward-traveling waves, ωS_1 is the propagation constant of one mode, and ωS_2 is the propagation constant of the second mode. There are many possible ways to decompose the fields on the coupled pair of transmission lines into modes. (Note that the total field is the weighted sums of the fields of each mode.) So if $\aleph_1$ describes the fields of one mode and $\aleph_2$ describes the fields of the other mode, the total fields on the coupled line pair is the weighted sum of the modes:

$$\aleph = \delta_1 \aleph_1 + \delta_2 \aleph_2. \tag{9.27}$$

Here δ_1 and δ_2 are weights and their values depend on boundary conditions (the type of source and the line terminations). They could be complex functions of distance along the lines if the modes are not self-sustaining. A mode pair that is naturally self-sustaining is the even- and odd-mode pair defined in Section 9.2 on Page 460. There are other possible descriptions, but the even- and odd-mode pair enables design of coupled lines with defined coupling, as well as providing insight into coupling.

The circuit-level model of a coupled line pair is developed by considering the calculation of $\mathbf{L}$ (inductance) and $\mathbf{C}$ (capacitance) matrices. The elements of the capacitance matrix are obtained in two simulations in which the line charge is calculated. A variety of commercially available software packages

---

[3] Just as with a single transmission line, higher-order modes can be excited if the width of the strips or the height of a strip above a ground plane is equal to one-half wavelength or more.

[4] In general a system with N active conductors (and one reference conductor) will support $2 \times N$ (quasi-) TEM modes.

exist for the extraction of the per-unit-length $\mathbf{L}$ and $\mathbf{C}$ matrices. The matrix $\mathbf{C}$ is calculated, in most packages, from the solutions of a two-dimensional electrostatic problem. The steps involve solving for the charges on the lines with voltages set on the conductors. With total voltages V_1 and V_2 on lines 1 and 2, the charges on lines Q_1 and Q_2 are:

$$\begin{aligned} Q_1 &= C_{11}V_1 + C_{12}V_2 & (9.28) \\ Q_2 &= C_{12}V_1 + C_{22}V_2. & (9.29) \end{aligned}$$

Simulation 1: With $V_1 = 1$ and $V_2 = 0$, the charges are calculated with the result that

$$C_{11} = Q_1 > 0 \qquad (9.30)$$

and

$$C_{12} = Q_2 < 0. \qquad (9.31)$$

(Note that $C_{ij}, i \neq j$, is negative.)
Simulation 2: With $V_1 = 0$ and $V_2 = 1$, the charges are again calculated, and now

$$C_{22} = Q_2 \quad \text{and} \quad C_{12} = Q_1. \qquad (9.32)$$

The characterization of the lines is completed by determining the elements of the inductance matrix. This is done by calculating the capacitances with and without the dielectric. The principle effect of the dielectric is to alter the configuration and magnitude of the electric field. The dielectric has little effect on the magnitude and orientation of the magnetic field. With the same current on the coupled lines, the same magnetic energy is stored, and the inductances of the coupled line are unchanged by the dielectric. The other assumption is that with a TEM mode on the lines, and in the absence of a dielectric, the velocity of propagation is $c = 1/\sqrt{LC}$. Specifically, the assumption is that for a TEM mode and without a dielectric, the phase velocity is just c. This is a very good approximation and is exact if the conductors have infinite conductivity. (If the conductors have finite conductivity there would be field inside the conductors, and the wave would slow down.) Determining the capacitance matrix without the dielectric enables the inductance matrix to be calculated:

$$\mathbf{L} = \frac{1}{c}\mathbf{C}^{-1}. \qquad (9.33)$$

At this stage, the telegrapher's equation has been defined in terms of the phasor transmission line modes $V_1(x)$, $I_1(x)$ and $V_2(x)$, $I_2(x)$. In the next section, the telegrapher's equations are considered in terms of even and odd modes.
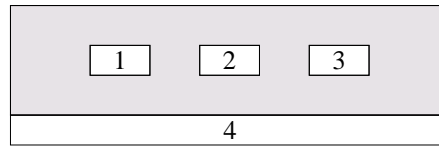
Figure 9-5 Three-line structure with four conductors labelled 1, 2, 3, and 4. The fourth conductor is the ground plane.

9.4 Capacitance Matrix Extraction

This section describes a general procedure for the extraction of the capacitance matrix of a number of coupled conductors [135]. Either measurements or measurement-like simulations can be used. Measurements and many programs that extract capacitance of an interconnect network do so by evaluating one capacitance at a time for different connections of interconnects to each other. This was seen, in the previous section, for two conductors above a ground plane (a three-conductor system).

Without loss of generality, consider the three-line structure shown in Figure 9-5, where there are four conductors including the ground. A 3×3 capacitance matrix is required to describe the capacitive coupling of this structure. Measurements and most EM simulators can only determine one capacitance at a time. The solution is to connect the individual interconnects into two distinct connected structures and determine the capacitance between the pairs of connected conductors. This procedure is repeated until every combination is considered. In this case there are seven possible combinations, as shown on the left-hand side of Figure 9-6. However, because of reciprocity, $C_{ij} = C_{ji}$, there are only six capacitances (C_{11}, C_{12}, C_{13}, C_{22}, C_{23}, and C_{33}) to be determined, and only six of these combinations. The seventh combination serves as a check. The combinations can be realized experimentally by connecting groups of conductors to form the required capacitor. Here capacitance is between the black conductor group and the white group. The measurement of two similar structures, with some known difference, such as lines of different length, can be used to calibrate out the error introduced by the electrical connections, or they can simply be ignored (which is usually reasonable). In an EM simulation, the connections are conveniently realized by holding the black group at one voltage (e.g., 1 V), and the white group at 0 V. The corresponding capacitance connections are shown on the right of this figure: the seven individual capacitance measurements are C_A, C_B, C_C, C_D, C_E, C_F, and C_G. (Note: Because of reciprocity, $C_{ij} = C_{ji}$.) Thus the capacitance matrix can be derived:

$$\mathbf{C} = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{12} & C_{22} & C_{23} \\ C_{13} & C_{23} & C_{33} \end{bmatrix}. \tag{9.34}$$

Conductor Number
1   2   3

$$C_A = C_{11} + C_{12} + C_{13}$$

$$C_B = C_{22} + C_{12} + C_{23}$$

$$C_C = C_{33} + C_{13} + C_{23}$$

$$C_D = C_{11} + C_{13} + C_{22} + C_{23}$$

$$C_E = C_{11} + C_{12} + C_{23} + C_{33}$$

$$C_F = C_{22} + C_{12} + C_{13} + C_{33}$$

$$C_G = C_{11} + C_{22} + C_{33}$$
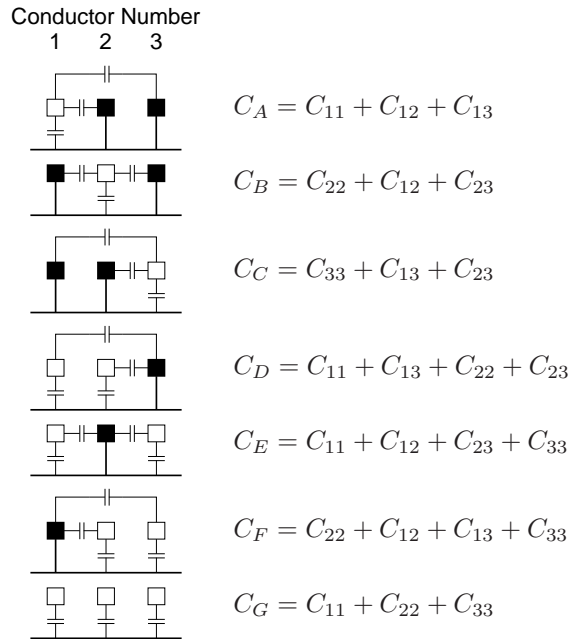
Figure 9-6 Combinations of conductors leading to various capacitance measurements.

There are seven measurements that can be conveniently made, but there are only six unknown quantities in the capacitance matrix. Thus only six measurements are required to determine the capacitance matrix and the redundancy (the extra measurement) can be used as a check on the results. Using the first six measurements,

$$\begin{bmatrix} C_A \\ C_B \\ C_C \\ C_D \\ C_E \\ C_F \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} C_{11} \\ C_{12} \\ C_{13} \\ C_{22} \\ C_{23} \\ C_{33} \end{bmatrix}, \qquad (9.35)$$

thus

$$\begin{bmatrix} C_{11} \\ C_{12} \\ C_{13} \\ C_{22} \\ C_{23} \\ C_{33} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} C_A \\ C_B \\ C_C \\ C_D \\ C_E \\ C_F \end{bmatrix}. \qquad (9.36)$$

It is important to note that the driving point capacitance[5] of an interconnect is affected by the signals or voltages on surrounding structures. For example, and referring to Figure 9-6, if each conductor (other than ground) is driven by a driver, and the output of the driver for the two left conductors is held at 0 V, then this corresponds to the situation for C_C. C_C is the effective capacitance seen by the driver of the right-most interconnect (and is the capacitance that must be charged in changing voltage levels). Thus C_C depends upon the data on the neighboring lines.

9.5   Symmetric Coupled Transmission Lines

In this section, even- and odd-mode descriptions are considered for the special case of a symmetrical pair of coupled lines; that is, each transmission line of the pair is identical. For a pair of microstrip lines, this would mean that the strips have equal width and are in the same plane. Thus the strips have the same self-inductance, L_s, and self capacitance, C_s, where the subscript s stands for "self." L_m and C_m are the mutual inductance and capacitance of the strips, and the subscript m stands for "mutual". Equations (9.9)–(9.12) can thus be written as

$$\frac{\mathrm{d}V_1(x)}{\mathrm{d}x} = -\jmath\omega L_s I_1(x) - \jmath\omega L_m I_2(x) \tag{9.37}$$

$$\frac{\mathrm{d}V_2(x)}{\mathrm{d}x} = -\jmath\omega L_m I_1(x) - \jmath\omega L_s I_2(x) \tag{9.38}$$

$$\frac{\mathrm{d}I_1(x)}{\mathrm{d}x} = -\jmath\omega C_s V_1(x) - \jmath\omega C_m V_2(x) \tag{9.39}$$

$$\frac{\mathrm{d}I_2(x)}{\mathrm{d}x} = -\jmath\omega C_m V_1(x) - \jmath\omega C_s V_2(x). \tag{9.40}$$

The even mode is defined as the mode corresponding to both conductors being at the same potential and carrying the same currents:[6]

$$V_1 = V_2 = V_e \quad \text{and} \quad I_1 = I_2 = I_e. \tag{9.41}$$

The odd-mode is defined as the mode corresponding to the conductors being at opposite potentials relative to the reference conductor and carrying currents of equal amplitude but of opposite sign[7]:

$$V_1 = -V_2 = V_o \quad \text{and} \quad I_1 = -I_2 = I_o. \tag{9.42}$$

---

[5]   The driving point impedance, or here capacitance, is what would be measured without knowledge of the behavior of surrounding structures. It is the Thevenin equivalent input impedance (capacitance) of the structure.

[6]   Here $I_e = (I_1 + I_2)/2$ and $V_e = (V_1 + V_2)/2$. The reason for the supposedly equally valid definition $I_e = I_1 + I_2$ not being used is that the adopted definition results in the desirable form of the even-mode characteristic impedance, as will be seen.

[7]   Here $I_o = (I_1 - I_2)/2$ and $V_o = (V_1 - V_2)/2$.

The characteristics of the two possible modes of the coupled transmission lines are now described. For the even mode, from Equations (9.37) and (9.38),

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(V_1(x) + V_2(x)\right) = -\jmath\omega\left(L_m + L_s\right)\left(I_1(x) + I_2(x)\right), \qquad (9.43)$$

which becomes

$$\frac{\mathrm{d}V_e(x)}{\mathrm{d}x} = -\jmath\omega\left(L_s + L_m\right)I_e(x). \qquad (9.44)$$

Similarly, using Equations (9.39) and (9.40),

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(I_1(x) + I_2(x)\right) = -\jmath\omega\left(C_s + C_m\right)\left(V_1(x) + V_2(x)\right), \qquad (9.45)$$

which in turn becomes

$$\frac{\mathrm{d}I_e(x)}{\mathrm{d}x} = -\jmath\omega\left(C_s + C_m\right)V_e(x). \qquad (9.46)$$

Defining the even-mode inductance and capacitance, L_e and C_e, respectively, as

$$L_e = L_s + L_m, \quad C_e = C_s + C_m \qquad (9.47)$$

leads to the **even-mode telegrapher's equations**:

$$\frac{\mathrm{d}V_e(x)}{\mathrm{d}x} = -j\omega\, L_e I_e(x) \qquad (9.48)$$

and

$$\frac{\mathrm{d}I_e(x)}{\mathrm{d}x} = -j\omega\, C_e V_e(x). \qquad (9.49)$$

From these, the even-mode characteristic impedance can be found,

$$Z_{0e} = \sqrt{\frac{L_e}{C_e}} = \sqrt{\frac{L_s + L_m}{C_s + C_m}} \qquad (9.50)$$

$$(9.51)$$

and also the even-mode phase velocity,

$$v_{Pe} = \frac{1}{\sqrt{L_e C_e}}. \qquad (9.52)$$

The characteristics of the odd-mode operation of the coupled transmission line can be determined in a similar procedure to that used for the even mode. Using Equations (9.37)–(9.40), the **odd-mode telegrapher's equation** becomes

$$\frac{\mathrm{d}V_o(x)}{\mathrm{d}x} = -j\omega\left(L_s - L_m\right)I_o(x) \qquad (9.53)$$

$$\frac{\mathrm{d}I_o(x)}{\mathrm{d}x} = -j\omega\left(C_s - C_m\right)V_o(x). \qquad (9.54)$$

Defining L_o and C_o for the odd mode such that

$$L_o = L_s - L_m, \quad C_o = C_s - C_m, \tag{9.55}$$

then the odd-mode characteristic impedance is

$$Z_{0o} = \sqrt{\frac{L_o}{C_o}} = \sqrt{\frac{L_s - L_m}{C_s - C_m}} \tag{9.56}$$

and the odd-mode phase velocity is

$$v_{Po} = \frac{1}{\sqrt{L_o C_o}}. \tag{9.57}$$

Now for a sanity check. If the individual strips are widely separated, L_m and C_m will become very small and Z_{0e} and Z_{0o} will be almost equal. As the strips become closer, L_m and C_m will become larger and Z_{oe} and Z_{0o} will diverge.

There are no simple formulas for the propagation characteristics of coupled microstrip lines. Use of commercially available software that extracts per-unit-length $\mathbf{L}$ and $\mathbf{C}$ matrices is the easiest way to obtain accurate characteristics. For a more detailed analysis and for design formulas for coupled strip transmission lines, the reader is referred to Edwards and Steer [51].

9.5.1 Odd-Mode and Even-Mode Capacitances

The previous section used the even- and odd-mode capacitances for coupled lines of equal crosssection. In this section these capacitances are related to the elements of the capacitance matrix of the coupled line pair. Repeating Equations (9.28) and (9.29),

$$Q_1 = C_{11}V_1 + C_{12}V_2 \tag{9.58}$$
$$Q_2 = C_{12}V_1 + C_{22}V_2. \tag{9.59}$$

and the capacitance matrix is

$$\mathbf{C} = \left[\begin{array}{cc} C_{11} & C_{12} \\ C_{12} & C_{22} \end{array} \right]. \tag{9.60}$$

In the even mode $V_1 = V_2 = V_e$ and so the charge on strip 1 in the even mode is

$$Q_{1e} = (C_{11} + C_{12}) V_e \tag{9.61}$$

and the charge on strip 2 in the even mode is

$$Q_{2e} = (C_{21} + C_{22}) V_e \tag{9.62}$$

Defining the even-mode charge as

$$Q_e = (Q_{1e} + Q_{2e})/2, \tag{9.63}$$

then the even-mode charge becomes

$$Q_e = V_e(C_{11} + C_{22} + C_{12} + C_{21})/2. \tag{9.64}$$

This leads to the even-mode per unit length capacitance,

$$C_e = Q_e/V_e = (C_{11} + C_{22} + C_{12} + C_{21})/2. \tag{9.65}$$

Similarly, in the odd-mode, $V_o = V_1 = -V_2$, and the odd-mode charge on strip 1 is

$$Q_{1o} = (C_{11} - C_{12})\,V_o, \tag{9.66}$$

and the odd-mode charge on strip 2 is

$$Q_{2o} = (C_{21} - C_{22})\,V_o, \tag{9.67}$$

The odd-mode charge is then

$$Q_o = (Q_{1o} - Q_{2o})/2 = (C_{11} + C_{22} - C_{12} - C_{21})\,V_o/2. \tag{9.68}$$

The odd-mode capacitance is

$$C_o = Q_o/V_o = (C_{11} + C_{22} - C_{12} - C_{21})/2\,. \tag{9.69}$$

This will be clearer in the following example.

EXAMPLE 9. 1   Parallel Line Capacitance

Electromagnetic software can be used to determine the even- and odd-mode parameters of a coupled line. This is usually done by setting phasor voltage on the coupled line and evaluating the phasor charge under each condition. The voltage applied to the left strip is designated as V_1 and the voltage applied to the right strip is V_2. The phasor charge on the strips is Q_1 and Q_2, respectively. One part of the analysis is to redo the calculations, but this time with the substrate removed, so that now we are dealing with the free-space situation. In this case the charges are denoted by Q_{01} and Q_{02}. The matrix of (computer based) measurements is as follows:

Case	V_1 (V)	V_2 (V)	Q_1 (pC/m)	Q_2 (pC/m)	Q_{01} (pC/m)	Q_{02} (pC/m)
A	1	-1	70	-80	22.2	-24.7
B	1	1	30	40	2.82	5.32

(a) What is the two-port capacitance matrix?

(b) What is the even-mode capacitance?

(c) What is the odd-mode capacitance?

(d) What is the free-space (no dielectric) two-port capacitance matrix?

(e) What is the free-space even-mode capacitance?

(f) What is the free-space odd-mode capacitance?

(g) What is the even-mode effective relative permittivity?

(h) What is the odd-mode effective relative permittivity?

(i) Note that both the odd mode and even mode are TEM modes, so the phase velocity in the free-space situation is c. Determine the odd-mode and even-mode inductances per unit length in free space.

(j) Note that the inductances do not change when the dielectric is replaced. What is the even-mode impedance?

(k) What is the odd-mode impedance?

(l) What is the even-mode phase velocity?

(m) What is the odd-mode phase velocity?

Solution:
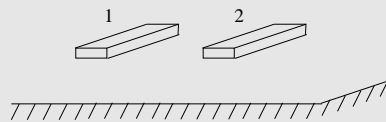
(a) Begin by considering the crosssection of a coupled line shown above and use the basic equations relating the charges on the line to the voltages on them:

$$Q_1 = C_{11}V_1 + C_{12}V_2 \quad \text{and} \quad Q_2 = C_{21}V_1 + C_{22}V_2,$$

and consider two sets of voltage conditions.

Case A: In the first case, referred to as the odd excitation, $V_1 = 1$ V and $V_2 = -1$ V and the charges are

$$Q_{1A} = C_{11} - C_{12} \tag{9.70}$$

and

$$Q_{2A} = C_{21} - C_{22}. \tag{9.71}$$

Case B: In the second condition, referred to as the even excitation, $V_1 = V_2 = 1$ V and the charges are

$$Q_{1B} = C_{11} + C_{12} \tag{9.72}$$

and

$$Q_{2B} = C_{21} + C_{22}. \tag{9.73}$$

Adding Equations (9.70) and (9.72) results in

$$Q_{1A} + Q_{1B} = 2C_{11} \tag{9.74}$$

so

$$C_{11} = (Q_{1A} + Q_{1B})/2 = (70 + 30)/2 \text{ pF/m} = 50 \text{ pF}. \tag{9.75}$$

Subtracting Equation (9.70) from Equation (9.72) yields

$$Q_{1B} - Q_{1A} = 2C_{12}, \tag{9.76}$$

so

$$C_{12} = (Q_{1B} - Q_{1A})/2 = (30 - 70)/2 \text{ pF/m} = -20 \text{ pF/m}. \tag{9.77}$$

Equation (9.71) plus Equation (9.74) yields

$$(Q_{2A} + Q_{2B}) = 2C_{21}, \tag{9.78}$$

so

$$C_{21} = (Q_{2A} + Q_{2B})/2 = (-80 + 40)/2 \text{ pF/m} = -20 \, pF/m \tag{9.79}$$

Equation (9.73) minus Equation (9.71) yields

$$(Q_{2B} - Q_{2A}) = 2C_{22}, \tag{9.80}$$

so

$$C_{22} = (Q_{2B} - Q_{2A})/2 = (40 - (-80))/2 \text{ pF/m} = 60 \text{ pF/m}. \tag{9.81}$$

Thus the per-unit-length capacitance matrix of the coupled line is

$$\mathbf{C} = \begin{bmatrix} 50 & -20 \\ -20 & 60 \end{bmatrix} \text{pF/m}. \tag{9.82}$$

(b) Even-mode capacitance, C_e:

The even mode has $V_1 = V_2$ and the even-mode voltage is $V_e = (V_1 + V_2)/2$.

The even-mode charge is $Q_e = (Q_1 + Q_2)/2 = (30 + 40)/2 \text{ pC/m} = 35 \text{ pC/m}$, so

$$C_e = Q_e/V_e = 35 \text{ pF/m}. \tag{9.83}$$

(c) Odd-mode capacitance, C_o:

Odd-mode voltage, $V_o = (V_1 - V_2)/2$ and

odd-mode charge, $Q_o = (Q_1 - Q_2)/2$. With $V_1 = +1V$ and $V_2 = -1V$, $V_0 = 1$ and

$$
\begin{aligned}
Q_o &= \frac{1}{2}[70 - (-80)] \text{ pC/m} \tag{9.84} \\
&= 75 \text{ pC/m}., \tag{9.85}
\end{aligned}
$$

$$C_o = Q_0/V_0 = 75 \text{ pF/m}. \tag{9.86}$$

(d) Using a similar procedure to that in (a), but now using the free-space charge calculations, Q_{01} and Q_{02} results in the unit capacitance matrix:

$$\mathbf{C}_0 = \left[\begin{array}{cc} 12.5 & -9.69 \\ -9.69 & 15.0 \end{array} \right] \text{ pF/m.} \tag{9.87}$$

(e) $C_{e0} = 4.07 \text{ pF/m}$

(f) $C_{o0} = 23.5 \text{ pF/m}$

(g) $\varepsilon_{re} = C_e/C_{e0} = 35/4.07 = 8.6$

(h) $\varepsilon_{ro} = C_o/C_{o0} = 75/23.5 = 3.2$

(i) Phase velocity, $v_p = 1\big/\sqrt{LC}$.

With no dielectric, the phase velocity is c:

$$v_p = c = 1\big/\sqrt{L_0 C_0} \rightarrow L_0 = 1/(c^2 C_0).$$

Odd-mode inductance:

$$L_{o0} = 1/(c^2 C_{o0}) = 1\big/\left[\left(3 \cdot 10^8\right)^2 23.5 \cdot 10^{-12} \right] \text{ H/m} = 473 \text{ nH/m.} \tag{9.88}$$

The free-space even-mode inductance:

$$L_{e0} = 1/(c^2 C_{e0}) = 1\big/\left[\left(3 \cdot 10^8\right)^2 \cdot 4.07 \cdot 10^{-12} \right] \text{ H/m} = 2.73 \ \mu\text{H/m.} \tag{9.89}$$

(j) $Z_o = \sqrt{L_o/C_o}\,; L_o = L_{o0} \ \ Z_o = \sqrt{473 \cdot 10^{-9}/(75 \cdot 10^{-12})} \ \Omega = 79.4 \ \Omega$

(k) $Z_e = \sqrt{L_e/C_e} \Rightarrow Z_e = \sqrt{2.73 \cdot 10^{-6}/(35 \cdot 10^{-12})} \ \Omega = 279 \ \Omega$

(l) $v_{pe} = 1\big/\sqrt{L_e C_e} = \left(2.73 \cdot 10^{-6} \cdot 35 \cdot 10^{-12}\right)^{-1/2} = 1.023 \times 10^8 \text{m/S}$

(m) $v_{po} = 1\big/\sqrt{L_o C_o} = \left(473 \cdot 10^{-9} \cdot 75 \cdot 10^{-12}\right)^{-1/2} = 1.68 \times 10^8 \text{ m/S}$

9.6 Formulas for Characteristic Impedance of Coupled Microstrip Lines

Formulas for the characteristic impedance and effective permittivity of symmetric coupled microstrip lines, with the cross section shown in Figure 9-7, were developed by Hammerstad and Jensen [66] based on the concept of even and odd modes. The formulas are accurate to better than 1% for $0.1 \leq u \leq 10$ and $g > 0.01$ where u is the normalized width,

$$u = w/h, \tag{9.90}$$

and g is the normalized gap,

$$g = s/h. \tag{9.91}$$
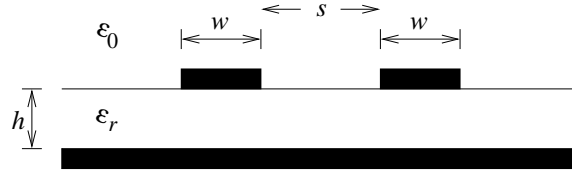
Figure 9-7 Cross section of symmetrically coupled microstrip lines.

In the following, Z_0 and ε_e refer to the characteristic impedance and effective permittivity of an individual microstrip line with a normalized width of u on a substrate with a relative dielectric constant of ε_r.

9.6.1 Even-Mode Coupled-Line Parameters

The even-mode characteristic impedance is

$$Z_{0e}(u,g) = Z_{01e}(u,g)/\sqrt{\varepsilon_{ee}(u,g,\varepsilon)}, \qquad (9.92)$$

where ε_{ee} is the effective relative permittivity of the even mode and Z_{01e} is the even-mode characteristic impedance with the dielectric replaced by free space:

$$Z_{01e}(u,g) = \frac{Z_0(u)}{1 - Z_0(u)\phi_e(u,g)/\eta_0}, \qquad (9.93)$$

where

$$\phi_e(u,g) = \frac{\varphi(u)}{\psi(g)\left\{\alpha(g)u^{m(g)} + [1-\alpha(g)]u^{-m(g)}\right\}} \qquad (9.94)$$

and $\eta_0 = 376.73\ \Omega \approx 377\ \Omega$ is the characteristic impedance of a TEM wave in a vacuum, which is commonly called the characteristic impedance of free space. Now $Z_0(u)$ is the free-space characteristic impedance of an individual microstrip line and is given by Equation (4.182) on Page 224. In Equations (9.92) and (9.93), the effective permittivity of the even mode is

$$\varepsilon_{ee}(u,g,\varepsilon) = \frac{\varepsilon_r+1}{2} + \frac{\varepsilon_r-1}{2}F_e(u,g,\varepsilon_r), \qquad (9.95)$$

where

$$F_e(u,g,\varepsilon_r) = \left[1 + \frac{10}{\mu(u,g)}\right]^{-a(u)b(\varepsilon_r)} \qquad (9.96)$$

$$a(u) = 1 + \frac{1}{49}\ln\left[\frac{u^4 + \{u/52\}^2}{u^4 + 0.432}\right] + \frac{1}{18.7}\ln\left[1 + \left(\frac{u}{18.1}\right)^3\right] \qquad (9.97)$$

$$b(\epsilon_r) = 0.564\left[\frac{\epsilon_r - 0.9}{\epsilon_r + 3}\right]^{0.053} \qquad (9.98)$$

$$\varphi(u) = 0.8645u^{0.1472} \qquad (9.99)$$

$$\psi(g) = 1 + \frac{g}{1.45} + \frac{g^{2.09}}{3.95} \tag{9.100}$$

$$\varphi(u) = 0.8645u^{0.1472} \tag{9.101}$$

$$\psi(g) = 1 + \frac{g}{1.45} + \frac{g^{2.09}}{3.95} \tag{9.102}$$

$$\alpha(g) = 0.5 \exp\left(-g\right) \tag{9.103}$$

$$m(g) = 0.2175 + \left[4.113 + \left(\frac{20.36}{g}\right)^6\right]^{-0.251} + \frac{1}{323} \ln\left[\frac{g^{10}}{1 + (g/13.8)^{10}}\right] \tag{9.104}$$

$$\mu(u,g) = g \exp\left(-g\right) + \frac{u(20 + g^2)}{10 + g^2}. \tag{9.105}$$

9.6.2 Odd-Mode Coupled-Line Parameters

The odd-mode characteristic impedance is

$$Z_{0o}(u,g) = Z_{01o}(u,g)/\sqrt{\varepsilon_{eo}(u,g,\varepsilon)}, \tag{9.106}$$

where ε_{eo} is the effective relative permittivity of the odd mode and Z_{01o} is the odd-mode characteristic impedance with the dielectric replaced by free-space:

$$Z_{01o}(u,g) = \frac{Z_0(u)}{1 - Z_0(u)\phi_o(u,g)/\eta_0}. \tag{9.107}$$

$Z_0(u)$ is the free-space characteristic impedance of an individual microstrip line and is given by Equation (4.182) on Page 224. In Equations (9.106) and (9.107), the effective permittivity of the odd mode is

$$\varepsilon_{eo}(u,g,\varepsilon) = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2} F_o(u,g,\varepsilon_r), \tag{9.108}$$

where

$$F_o(u,g,\varepsilon_r) = f_o(u,g,\varepsilon_r)\left(1 + 10/u\right)^{-a(u)b(\varepsilon_r)} \tag{9.109}$$

$$\phi_o(u,g) = \phi_e(u,g) - \frac{\theta(g)}{\psi(g)} \exp\left[\beta(g)u^{n(g)} \ln\left(u\right)\right] \tag{9.110}$$

$$\theta(g) = 1.729 + 1.175 \ln\left(1 + \frac{0.627}{g + 0.327g^{2.17}}\right) \tag{9.111}$$

$$\beta(g) = 0.2306 + \frac{1}{301.8} \ln\left[\frac{g^{10}}{1 + (g/3.73)^{10}}\right] + \frac{1}{5.3} \ln\left(1 + 0.646g^{1.175}\right) \tag{9.112}$$

$$n(g) = \left\{ \frac{1}{17.7} + \exp\left[-6.424 - 0.76\ln(g) - (g/0.23)^5\right] \right\}$$
$$\times \ln\left(\frac{10 + 68.3g^2}{1 + 32.5g^{3.093}} \right) \tag{9.113}$$

$$f_o(u, g, \varepsilon_r) = f_{o1}(g, \varepsilon_r) \exp\left[p(g)\ln(u) + q(g)\sin\left(\pi\frac{\ln u}{\ln 10} \right) \right] \tag{9.114}$$

$$p(g) = \exp\left(-0.745g^{0.295}\right)/\cosh\left(g^{0.68}\right) \tag{9.115}$$

$$f_{o1}(g, \varepsilon_r) = 1 - \exp\left\{ -0.179g^{0.15} - \frac{0.328g^{r(g,\varepsilon_r)}}{\ln\left[\exp(1) + (g/7)^{2.8}\right]} \right\} \tag{9.116}$$

$$r(g, \varepsilon_r) = 1 + 0.15\left\{ 1 - \frac{\exp\left[1 - (\varepsilon_r - 1)^2/8.2\right]}{1 + g^{-6}} \right\} \tag{9.117}$$

$$q(g) = \exp(-1.366 - g), \tag{9.118}$$

and $a(u)$ and $b(\varepsilon_r)$ are defined in Equations (9.97) and (9.98).

9.6.3 System Impedance of Coupled Lines

The system impedance of a pair of coupled lines is

$$Z_{0S} = \sqrt{Z_{0e}Z_{0o}} \,. \tag{9.119}$$

This is derived in Appendix A of Edwards and Steer [51], where it is shown that there will be no reflections at the ports of a symmetrical coupled line pair if each line of the directional coupler is terminated in an impedance Z_{0S}. That is, Z_{0S} is the impedance required for matching. Z_0 is the characteristic impedance of an individual line of the symmetrical coupled line pair (when there is no coupling) and is only close to Z_{0S} when the separation, s, of the lines is large. They diverge as the lines come closer together.

9.6.4 Discussion

Normalized even- and odd-mode characteristic impedances of a pair of coupled lines are plotted in Figure 9-8 for various normalized widths u (= w/h) as a function of normalized gap width g (= s/h). This plot illustrates the utility of using even- and odd-mode descriptions. In Figure 9-8, the even- and odd-mode impedances are normalized to the characteristic impedance of an individual line, Z_0. When the lines are far apart (i.e., g is large), the even- and odd-mode impedances converge to the characteristic impedance of a single line. As the lines get closer, the gap narrows, and the even- and odd-mode impedances diverge in opposite directions. To obtain the characteristic impedances of a coupled line the characteristic impedance of a single microstrip line must be found. This was given in Section 4.10.2 on Page 223 and the key result is repeated in Figure 9-9. Again, normalization
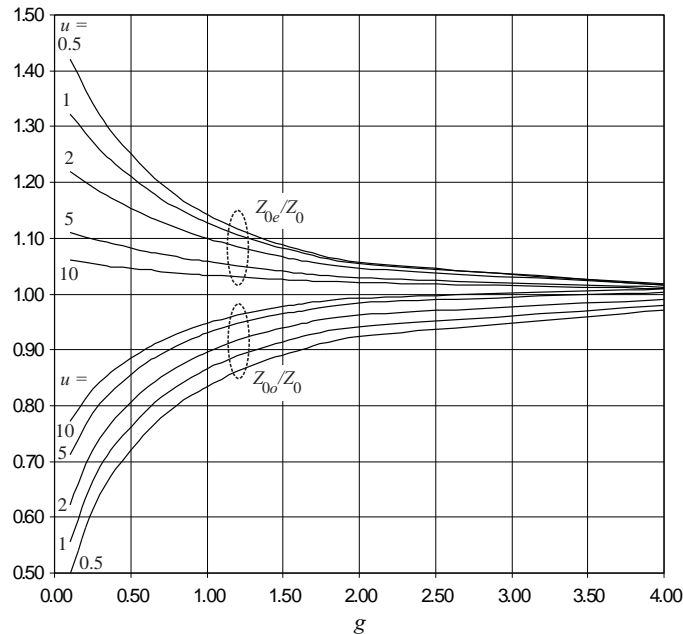
Figure 9-8 Normalized even-mode and odd-mode characteristic impedances of a pair of coupled microstrip lines.

is used to simplify the graphical presentation of data and highlight major dependencies.

In the even mode, more of the field is in the dielectric than with the odd mode. It is therefore not surprising that the effective permittivities of the two modes differ. The normalized coupled line effective permittivities are shown in Figure 9-10. The deviation of the even- and odd-mode permittivities as the gap between the lines narrows is not as large as the change in the characteristic impedance. However, there is a difference in the phase velocities of the two modes along the line, and this has an appreciable effect on the performance of components such as filters that use coupled lines as a functional component. At the first reading of the plot (Figure 9-10) it would seem that there is nonmonotonic behavior at low g. This is an artifact of the normalization used, and the unnormalized permittivities are indeed monotonic with respect to u and g.

Figure 9-11 highlights that the split of the even- and odd-mode characteristics impedances is almost solely dependent on geometry and not the permittivity of the substrate. In the figure, appropriate normalization is used to highlight this fact. There are four families of curves, two for the even-mode characteristic impedances and two for the odd-mode characteristic impedances. Each family comprises the results for three widely different permittivities of the dielectric (specifically $\varepsilon = 4$, 10, and 20).

22t>2ort>2
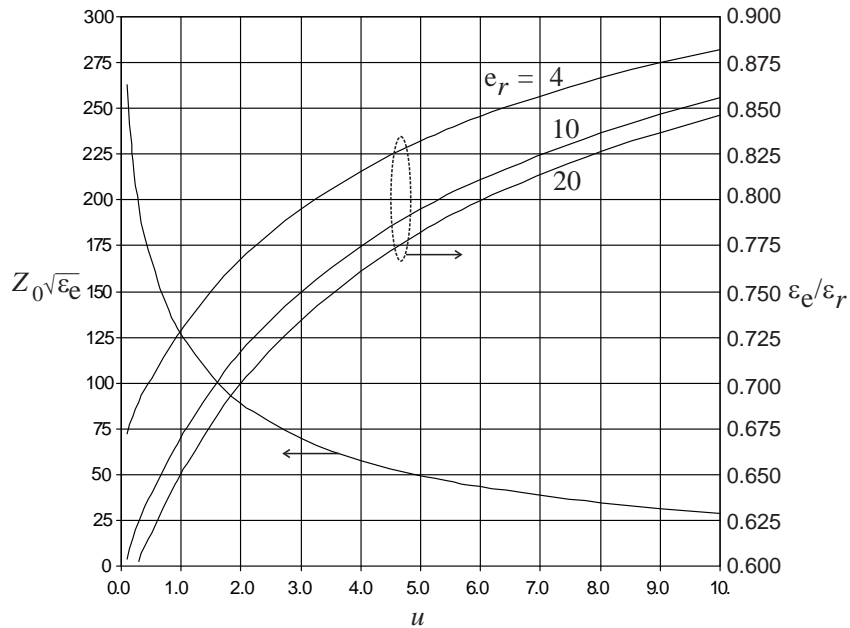
Figure 9-9 Normalized characteristic impedance and normalized effective permittivity of a microstrip line as a function of $u = w/h$. For example, if $u = 1$ and $\varepsilon_r = 10$, then from the figure, $Z_0\sqrt{\varepsilon_e} = 126\ \Omega$ and $\varepsilon_e/\varepsilon_r = 0.671$; thus $Z_0\sqrt{\varepsilon_e} = 48.6\ \Omega$ and $\varepsilon_e = 6.71$.

EXAMPLE 9.2   Even- and Odd-Mode Parameters

A coupled line is constructed on an alumina substrate of thickness 500 μm and relative permittivity $\varepsilon_r = 10$. The lines are 500 μm wide and the gap separation is 250 μm. What are the even- and odd-mode characteristic impedances and effective permittivities of the coupled line?

Solution:

The odd-mode characteristic impedance, Z_{oo}, and even-mode characteristic impedance, Z_{oe}, can be found using Figures 9-8 and 9-9. Now

$$u = w/h = 500\ \mu\text{m}/500\ \mu\text{m} = 1 \quad \text{and} \quad g = s/h = 250\ \mu\text{m}/500\ \mu\text{m} = 0.5 .$$

From Figure 9-8,
$$Z_{0e}/Z_0 = 1.21 \quad \text{and} \quad Z_{oo}/Z_0 = 0.76, \tag{9.120}$$

where Z_0 is the characteristic impedance of an individual line. From Figure 9-9, and using the curve for $\varepsilon = 10$,

$$\varepsilon_e/\varepsilon_r = 0.671 , \quad \text{and so} \quad \varepsilon_e = 10 \times 0.671 = 6.71 . \tag{9.121}$$
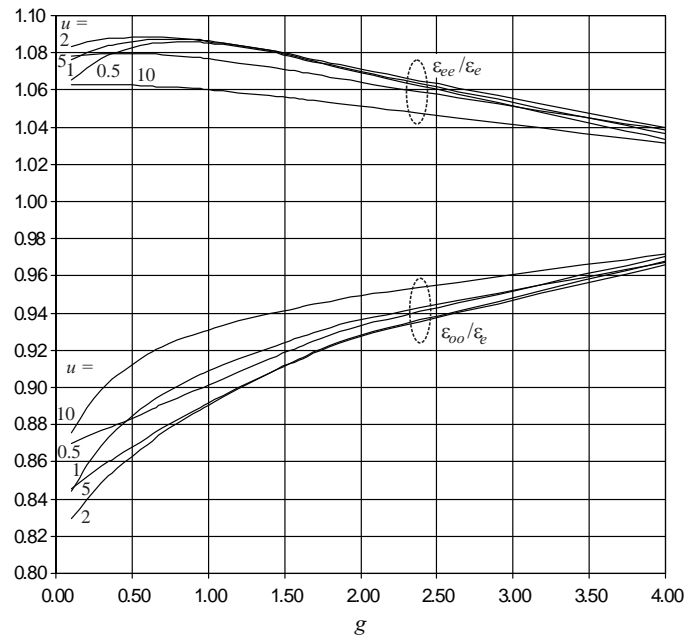
Figure 9-10 Normalized even-mode and odd-mode effective permittivity of a pair of coupled microstrip lines. The effective permittivity of an individual microstrip line with the same normalized width u is ε_e.

Also from Figure 9-9,

$$Z_0\sqrt{\varepsilon_e} = 126, \quad \text{and so} \quad Z_0 = 126/\sqrt{6.71} = 48.6 \ \Omega. \tag{9.122}$$

Consequently, combining Equations (9.120) and (9.122),

$$Z_{0o} = 37.3 \ \Omega \quad \text{and} \quad Z_{oe} = 59.1 \ \Omega. \tag{9.123}$$

The effective odd-mode and even-mode permittivities are obtained from Figure 9-10. The normalized even-mode effective permittivity is

$$\varepsilon_{ee}/\varepsilon_e = 1.086$$

and the normalized odd-mode effective permittivity is

$$\varepsilon_{eo}/\varepsilon_e = 0.868 \ .$$

Since $\varepsilon_e = 6.71$, the final result is

$$\varepsilon_{ee} = 7.28 \quad \text{and} \quad \varepsilon_{eo} = 5.82 \ .$$
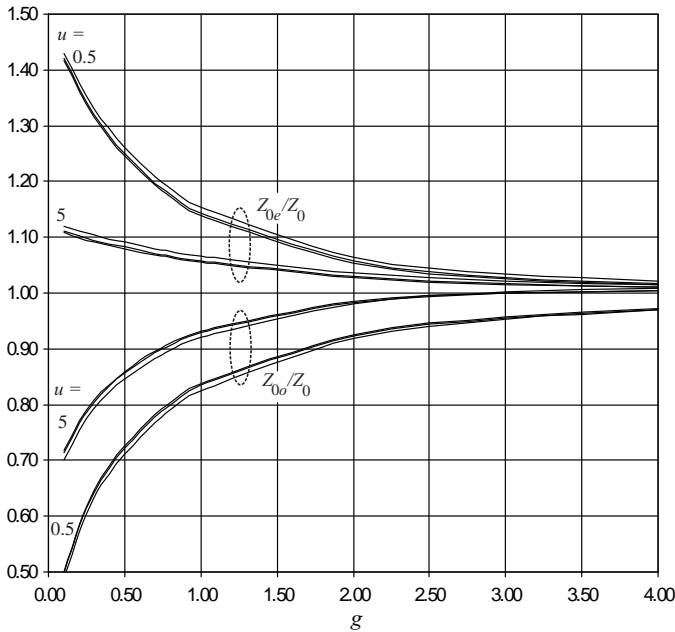
Figure 9-11 Normalized even-mode and odd-mode effective permittivity of a pair of coupled microstrip lines for extremes of u. Each family of three curves is for $\varepsilon_r = 4$, 10, and 20. Z_0 is the characteristic impedance of an individual microstrip line with the same normalized width, $u = w/h$. The strips have equal width.
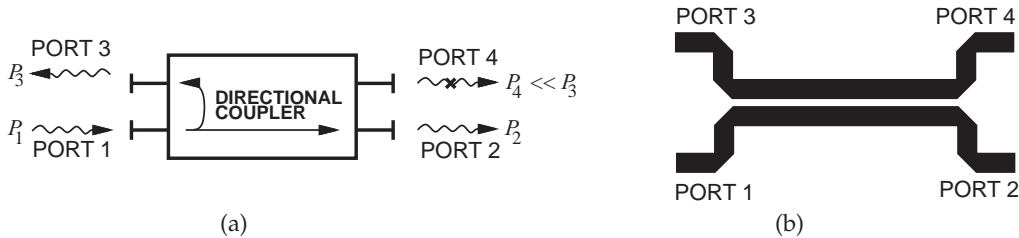
Figure 9-12 Directional couplers: (a) schematic, and (b) backward-coupled microstrip directional coupler.

9.7 Directional Coupler

In analog and RF circuits a designer usually has some freedom to increase the spacing of lines to reduce coupling to acceptable levels. The coupling phenomenon, however, can be exploited to realize a new type of element called a directional coupler. The schematic of a directional coupler is shown in Figure 9-12(a). The top view of a microstrip realization of a directional coupler is shown in Figure 9-12(b). The microstrip realization is typical of most directional couplers in that it comprises two parallel signal lines with

the electric and magnetic fields of a signal on one line inducing currents and voltages on the other. A usable directional coupler has a coupled line length of at least one-quarter wavelength, with longer lengths of line resulting in broader bandwidth operation. Directional couplers are used to sample a traveling wave on one line and to induce a usually much smaller image of the wave on another line. That is, the forward- and backward-traveling modes can be separated. An important application of a directional coupler is as part of a measurement system. Here a prescribed amount of the incident power is required to be coupled out of the system. Thus, for example, a "20 dB microstrip coupler" is a pair of coupled microstrip lines in which $1/100$ of the power input is coupled from one microstrip line onto the another.

Referring to Figure 9-12, a coupler is specified in terms of the following parameters:

- Coupling factor (or coupling parameter):
$C = V_1^+/V_3^-$ =   voltage fraction "transferred" (coupled) across to the opposite arm (greater than 1).

- Transmission factor (also known as insertion gain):
$T = V_2^-/V_1^+$ =   transmission directly through the "primary" arm of the structure (less than 1).

- Directivity factor:
$D = V_3^-/V_4^-$ =   measure of the undesired coupling from Port 1 to Port 4 relative to the signal level at Port 3 (greater than 1).

- Isolation factor:
$I = V_1^+/V_4^-$ =   degree of isolation between Port 4 and Port 1 (greater than 1).

With the exception of the transmission factor, the definitions are designed so that when the factors are expressed in decibels the resulting number is positive.

It is usual to quote all these quantities in decibels. One should be very careful in relating decibels and the numerical values of the factors above, as the negative sign is nearly always dropped from decibel designations. For example, the coupling factor in decibels, $C|_{\mathrm{dB}}$, should be $C|_{\mathrm{dB}} = 20 \log C$, and this would be a negative decibel number, but coupling is almost universally specified using a positive decibel number. So 20 dB coupling indicates that the coupling factor is 10.[8] The appropriate correction of sign is usually required. An ideal quarter-wave coupler has $D = \infty$ (or infinite

---

[8]   Always check the magnitude of the above factors, as some papers and books on couplers use the inverse of C.

directivity) and

$$C = \frac{Z_{0e} + Z_{0o}}{Z_{0e} - Z_{0o}}. \tag{9.124}$$

This result comes from a detailed derivation for the case when the coupled-line section is one-quarter wavelength long—the length when the coupling is maximum. The derivation is given in Appendix A of Edwards and Steer [51]. In decibels the coupling factor is

$$C|_{\mathrm{dB}} = 20\log\left(\frac{Z_{0e} + Z_{0o}}{Z_{0e} - Z_{0o}}\right) \tag{9.125}$$

An ideal coupler does not dissipate power, so

$$|T| = |\sqrt{1 - 1/C^2}|. \tag{9.126}$$

A directional coupler with ports defined as in Figure 9-12 with the ports matched (so that $S_{11} = 0 = S_{22} = S_{33} = S_{44}$) has the following scattering parameter matrix:

$$\mathbf{S} = \begin{bmatrix} 0 & T & 1/C & 1/I \\ T & 0 & 1/I & 1/C \\ 1/C & 1/I & 0 & T \\ 1/I & 1/C & T & 0 \end{bmatrix}. \tag{9.127}$$

There are many types of directional couplers, and the phases of the traveling waves at the ports will not necessarily be in phase as Equation (9.127) implies. The microstrip coupler shown in Figure 9-12(b) has maximum coupling when the lines are one-quarter wavelength long.[9] At the frequency where they are one-quarter wavelength long, the phase difference between traveling waves entering at Port 1 and leaving at Port 2 will be $90°$ so that Equation (9.127) becomes

$$\mathbf{S} = \begin{bmatrix} 0 & -\jmath T & 1/C & 1/I \\ -\jmath T & 0 & 1/I & 1/C \\ 1/C & 1/I & 0 & -\jmath T \\ 1/I & 1/C & -\jmath T & 0 \end{bmatrix}. \tag{9.128}$$

Typical and ideal parameters of a directional coupler are given in Table 9-1.

---

[9] This can be shown through a detailed derivation provided by Edwards and Steer [51] but not repeated here.

Table 9-1   Ideal and typical parameters of a directional coupler.

Parameter	Ideal	Ideal (dB)	Typical
Coupling, C	-	-	3–40 dB
Transmission, T	$\|\sqrt{1 - 1/C^2}\|$	$20 \log \|\sqrt{1 - 1/C^2}\|$	-0.5 dB
Directivity, D	∞	∞	40 dB
Isolation, I	∞	∞	40 dB

EXAMPLE 9. 3   Directional Coupler Isolation

A lossless directional coupler has the following characteristics: coupling factor $C = 20$ dB, transmission factor 0.8, and directivity 20 dB. What is the isolation factor? Express your answer in decibels.

Solution:
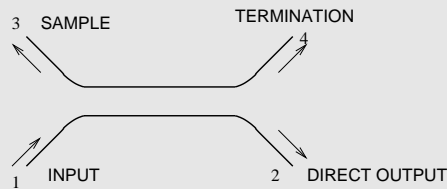
Coupling factor: $C = V_1^+/V_3^-$

Transmission: $T = V_2^-/V_1^+$

Directivity: $D = V_3^-/V_4^-$

Isolation: $I = V_1^+/V_4^-$

Note: Always check the magnitude of the above factors. Now,

$$D = 20\text{ dB} = 10 \quad \text{and} \quad C = 20\text{ dB} = 10,$$

so the isolation is

$$I = \frac{V_1^+}{V_4^-} = \frac{V_3^-}{V_4^-} \times \frac{V_1^+}{V_3^-} = D \cdot C = 10 \times 10 = 100 = 40\text{ dB}.$$

EXAMPLE 9. 4   Directional Coupler Electrical Design

Develop the electrical design of a 15 dB directional coupler using coupled microstrip lines on a substrate with a permittivity of 12, a substrate thickness, h, of 600 μm, and at a center frequency of 10 GHz. Use a system impedance of 50 Ω.

Solution:
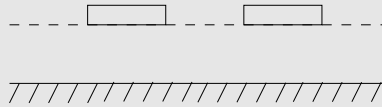
The electrical design of a directional coupler comes down to determining the even- and odd-mode characteristic impedances required. Now the coupling factor, C, is $(Z_{0e} + Z_{0o})/(Z_{0e} - Z_{0o})$ and

$$Z_{0e} = Z_{0S}\sqrt{\frac{1+C}{1-C}}.$$

For a 15 dB directional coupler $\Rightarrow$ $C = 15$ dB $= 5.618$ and so

$$Z_{0e} = 50\sqrt{\frac{5.618 + 1}{5.618 - 1}} = 59.9\ \Omega$$

$$Z_{0o} = Z_{0S}^2/Z_{0e} = 50^2/59.9 = 41.7\ \Omega.$$

The above are the required electrical design parameters.

9.7.1 Characteristic Impedances in Terms of the Coupling Factor

In the previous section the coupling factor was expressed in terms of the even- and odd-mode impedances. However, design starts with the specification of the coupling level and from this the required physical dimensions are derived. A one-quarter wavelength long coupler will be considered, as this is the optimum coupling length. From Equation (9.124), the desired coupling factor is[10]

$$C = \frac{Z_{0e} + Z_{0o}}{Z_{0e} - Z_{0o}}, \tag{9.129}$$

where the coupling factor is an absolute voltage-referenced quantity and usually must be derived from the coupling factor in decibels; let this be $C|_{\text{dB}}$:

$$C = 10^{C|_{\text{dB}}/20}. \tag{9.130}$$

---

[10] These three results, the coupling factor in terms of the even- and odd-mode impedances, the optimum coupling length ($\lambda_g/4$), and the system reference impedance in terms of the even- and odd-mode impedances are derived in Appendix A of Edwards and Steer [51].

The system impedance comes from

$$Z_{0S}^2 \approx Z_{0e} Z_{0o}. \tag{9.131}$$

Z_{0S} is introduced here because Z_0 is used for the characteristic impedance of the individual lines of the coupler; Z_0 is not equal to Z_{0S}. Both should match the characteristic impedance of the transmission lines connected to the coupler. From these expressions, the even- and odd-mode impedances required are

$$Z_{0e} \approx Z_{0S} \sqrt{\frac{C+1}{C-1}} \tag{9.132}$$

$$Z_{0o} \approx Z_{0S} \sqrt{\frac{C-1}{C+1}}, \tag{9.133}$$

and the ratio of impedances is

$$Z_{0e}/Z_{0o} \approx \frac{C+1}{C-1}. \tag{9.134}$$

The last three equations are the basic design expressions, as Z_{0e} and Z_{0o} can be related to physical dimensions. One of the small complications is that the close proximity of the two strips alters the orientation of the field lines and hence the characteristic impedances of the individual lines. That is, the characteristic impedance of each line on its own, Z_0, will differ from the system impedance, Z_{0S}. For a normalized line width of $u = (w/h)$, this effect is shown in Figure 9-13. Here Z_{OS}, the desired system impedance, is the geometric mean of the even- and odd-mode impedances. Z_0 is the characteristic impedance of one of the lines of the coupled line structure. Thus maintaining the desired Z_{0S} requires that Z_0 of the individual lines be greater than Z_{0S}, particularly as the normalized gap size reduces.

A more direct design parameter is to use the ratio of the even- and odd-mode impedances (Equation (9.134)). Microstrip coupler design proceeds as follows. The first step is to examine the specifications and determine the substrate permittivity, ε_r, coupling factor, C, in decibels, and the system characteristic impedance, Z_{0S}. From C find Z_{0e}/Z_0; the data in Tables 9-2 and 9-3 enable some of the physical parameters to be determined, including the normalized gap coupling parameter, g, and the normalized strip width, u. The tables also provide the characteristic impedance, even- and odd-mode characteristic impedances, and the characteristic impedance of the individual strips of the directional coupler. The next step is to design the dimensions of the individual microstrip lines connecting the directional coupler using Table 4-3 on Page 227. At this stage the widths and spacings of the microstrip circuit are normalized. Using the substrate height, these are normalized to obtain the actual physical dimensions. Finally, the length of the coupler is a one-quarter wavelength long, as this was the basis for
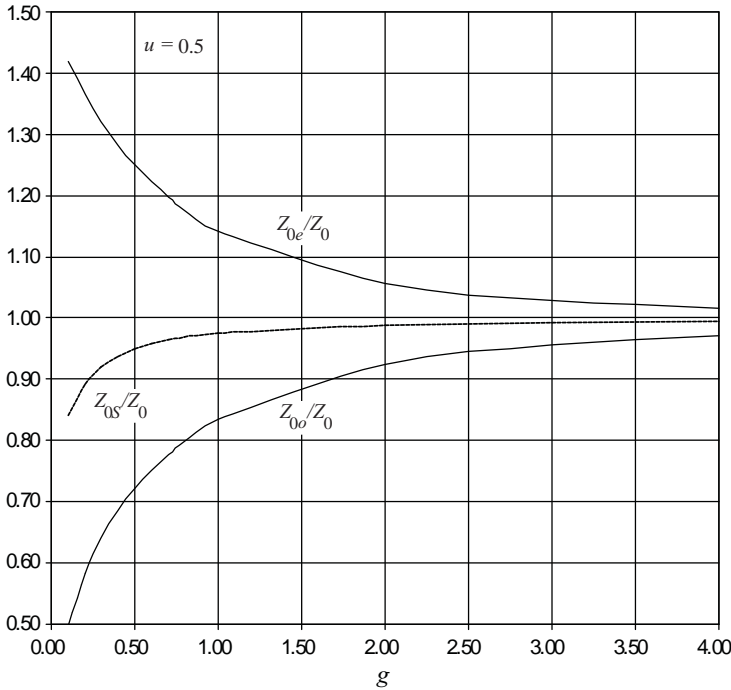
Figure 9-13 Normalized even- and odd-mode effective permittivity of coupled microstrip lines with normalized width, $u = 0.5$, versus normalized gap spacing, g.

the formula relating the even- and odd-mode impedances to the coupling factor in Equation (9.124). The even and odd modes have different effective permittivities and the $\lambda_g/4$ length should apply to both the even and odd modes. Clearly both cannot be satisfied. It is reasonable to use the average of the even- and odd-mode permittivities to establish the coupler length. The length is not a very sensitive parameter anyway. The connection of the individual lines to the coupler is not specifically part of the synthesis described, but if designed they should be designed as a bend.

Often tables do not have the values that are required in design and it is necessary to interpolate the table values using linear interpolation or bilinear interpolation as described in Section A.6 on page 830.

A comment on this design procedure is in order. The design procedure above yields a narrowband directional coupler. A broadband directional coupler, and indeed any component that is desired to have a broad bandwidth, should be designed using filter principles. The filter discussion in Chapter 10 focuses on the design of microstrip filters using coupled lines and so is closely aligned to the design of couplers with desired frequency characteristics. Another comment is that the design flow is one

Table 9-2 Design parameters for coupled lines for $\varepsilon_r = 4$ corresponding to design on an SiO$_2$ or FR-4 (printed circuit board) substrate. The normalized gap, g, is chosen to obtain the desired coupled line mode impedance ratio, z_{0e}/z_{0o}. Data are derived from the analysis in Section 9.6 on Page 476. Z_0 is the characteristic impedance of an individual microstrip line with a normalized width, u, on the same substrate.

$\varepsilon_r = 4$ (SiO$_2$ and FR-4), $Z_{0S} = 50\ \Omega$							
g	z_{0e}/z_{0o}	u	Z_{0e} (Ω)	Z_{0o} (Ω)	ε_{ee}	ε_{eo}	Z_0 (Ω)
0.10	2.15	1.60	73.17	34.10	3.21	2.61	57.96
0.20	1.84	1.74	67.85	36.97	3.24	2.65	55.23
0.30	1.66	1.83	64.29	38.63	3.25	2.68	53.45
0.40	1.55	1.88	62.48	40.18	3.26	2.70	52.77
0.50	1.47	1.92	60.83	41.29	3.27	2.73	52.11
0.60	1.41	1.96	59.31	42.10	3.28	2.75	51.47
0.70	1.36	1.98	58.30	42.92	3.28	2.76	51.16
0.80	1.32	2.00	57.37	43.57	3.28	2.78	50.85
0.90	1.28	2.02	56.51	44.09	3.28	2.79	50.54
1.00	1.25	2.02	56.06	44.75	3.28	2.81	50.54
1.10	1.23	2.04	55.30	45.09	3.28	2.82	50.24
1.20	1.20	2.04	54.93	45.60	3.28	2.83	50.24
1.30	1.19	2.05	54.60	46.05	3.27	2.83	50.24
1.40	1.17	2.06	53.96	46.20	3.27	2.84	49.94
1.50	1.15	2.06	53.68	46.56	3.27	2.85	49.94
1.60	1.14	2.06	53.43	46.87	3.26	2.86	49.94
1.70	1.13	2.06	53.20	47.16	3.26	2.86	49.94
1.80	1.12	2.07	52.99	47.41	3.26	2.87	49.94
1.90	1.11	2.07	52.47	47.37	3.25	2.88	49.64
2.00	1.10	2.08	52.30	47.58	3.25	2.88	49.64
2.50	1.07	2.08	51.61	48.33	3.23	2.91	49.64
3.00	1.05	2.08	51.15	48.80	3.21	2.93	49.64
4.00	1.03	2.08	50.62	49.31	3.18	2.96	49.64

of synthesis. An alternative procedure that is often used is to start with a very approximate design and rely on optimization tools to obtain the desired characteristics. This works in many cases but often does not lead to optimum design. Having said that, the synthesis procedure does not yield a perfect design, as parasitic and dispersive effects are not taken into account. Optimization from the synthesized design requires only a small adjustment. In practice, the uncertainties of physical structures (e.g., variations in the effective permittivity of actual materials) requires experimental iteration.

Table 9-3 Design parameters for a microstrip coupler on a substrate with a relative permittivity, ε_r, of 10. Data are derived from the analysis in Section 9.6 on Page 476. Z_0 is the characteristic impedance of an individual microstrip line with a normalized width, u, on the same substrate.

$\varepsilon_r = 10$ (Alumina), $Z_{0S} = 50\ \Omega$							
g	z_{0e}/z_{0o}	u	Z_{0e} (Ω)	Z_{0o} (Ω)	ε_{ee}	ε_{eo}	Z_0 (Ω)
0.10	2.66	0.66	81.61	30.64	6.95	5.59	59.0
0.20	2.15	0.76	73.38	34.06	7.07	5.65	55.5
0.30	1.90	0.81	69.06	36.39	7.13	5.69	54.0
0.40	1.73	0.85	65.78	38.02	7.17	5.74	52.8
0.50	1.61	0.87	63.66	39.46	7.20	5.77	52.2
0.60	1.52	0.89	61.79	40.55	7.22	5.81	51.7
0.70	1.45	0.91	60.11	41.38	7.23	5.85	51.1
0.80	1.40	0.91	59.25	42.40	7.23	5.88	51.1
0.90	1.35	0.92	58.17	43.07	7.24	5.92	50.9
1.00	1.31	0.92	57.50	43.81	7.23	5.94	50.9
1.10	1.28	0.93	56.60	44.26	7.23	5.98	50.6
1.20	1.25	0.94	55.78	44.62	7.23	6.01	50.3
1.30	1.23	0.94	55.31	45.11	7.22	6.03	50.3
1.40	1.21	0.94	54.90	45.55	7.21	6.06	50.3
1.50	1.19	0.94	54.52	45.93	7.20	6.08	50.3
1.60	1.17	0.94	54.19	46.28	7.19	6.11	50.3
1.70	1.16	0.94	53.89	46.59	7.18	6.13	50.3
1.80	1.14	0.94	53.61	46.86	7.16	6.15	50.3
1.90	1.13	0.95	53.09	46.89	7.16	6.18	50.1
2.00	1.12	0.95	52.87	47.11	7.14	6.20	50.1
2.50	1.09	0.95	52.05	47.93	7.08	6.29	50.1
3.00	1.06	0.95	51.54	48.45	7.02	6.36	50.1
4.00	1.04	0.95	50.97	49.04	6.92	6.46	50.1
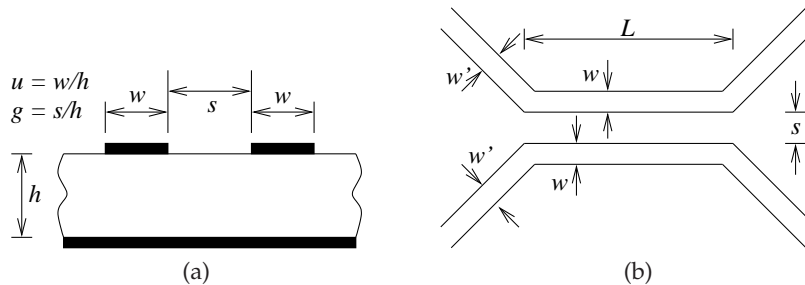
Figure 9-14 Coupler dimensions to be determined in Example 9.11: (a) crosssection; and (b) layout.

EXAMPLE 9. 5   Design of a 10 dB Microstrip Coupler

Design a microstrip directional coupler with the following specifications:

Transmission line technology	Microstrip	
Coupling coefficient	C	= 10 dB
Microstrip characteristic impedance	Z_0	= 50 Ω
Substrate permittivity	ε_r	= 10.0
Substrate thickness	h	= 1 mm
System center frequency (midband for the coupler)	f_0	= 5 GHz

The crosssectional dimensions that must be determined are the strip width, w, and the strip separation, s, as shown in Figure 9-14. The procedure is to first determine the coupling factor:

$$C = 10^{(C|_{\mathrm{dB}}/20)} = 10^{(10/20)} = 3.162 \ . \tag{9.135}$$

The ratio of the even- and odd- mode impedances required to achieve the desired coupling is derived from Equation (9.134):

$$Z_{0e}/Z_{0o} = \frac{C+1}{C-1} = \frac{3.162+1}{3.162-1} = 1.925 \ . \tag{9.136}$$

Solution:

The problem now is to determine the physical geometry (i.e., the line widths and spacing). The data of Table 9-3 apply here (as $\varepsilon_r = 10$), enabling the normalized gap, $g = s/h$, and normalized line width, $u = w/h$, to be determined for a specified impedance ratio, Z_{0e}/Z_{0o}. The table does not contain a line for $Z_{0e}/Z_{0o} = 1.925$ and so the table must be interpolated. The line for $Z_{0e}/Z_{0o} = 2.15$ has $g = 0.2$ and the line for $Z_{0e}/Z_{0o} = 1.90$ has $g = 0.3$. So for $Z_{0e}/Z_{0o} = 1.925$,

$$g = \left(\frac{0.3-0.2}{1.9-2.15}\right) \cdot (1.925 - 2.15) + 0.2 = 0.290 \ , \tag{9.137}$$

thus

$$s = g \cdot h = 0.290 \ \mathrm{mm}. \tag{9.138}$$

The value of u must also be interpolated from Table 9-3 and $u = 0.805$ is obtained; thus

$$w = u \cdot h = 0.805 \ \mathrm{mm}. \tag{9.139}$$

The coupler should be one-quarter wavelength long, so the effective relative permittivity of the even and odd modes is required. From Table 9-3, the interpolated values are

$$\varepsilon_{ee} = 7.124 \quad \mathrm{and} \quad \varepsilon_{eo} = 5.686 \ . \tag{9.140}$$

These effective permittivities are different, so determination of the optimum length of the coupler is not straightforward. The only choice is to use the average of the permittivities:

$$\varepsilon_{e,\mathrm{avg}} = (\varepsilon_{ee} + \varepsilon_{eo})/2 = 6.405. \tag{9.141}$$
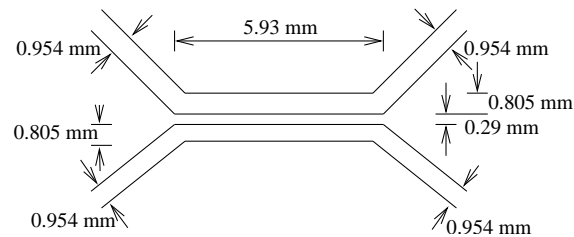
Figure 9-15 Final coupler layout for Example 9.5.
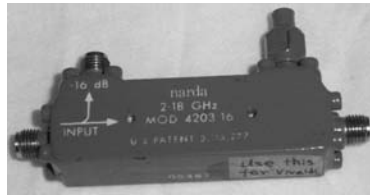
Figure 9-16 A microstrip directional coupler with SMA connectors. The top right-hand connector has a $50\,\Omega$ termination attached.

Thus the average wavelength is

$$\lambda_g = \frac{c}{f\sqrt{\varepsilon_{e,\text{avg}}}} = \frac{3 \cdot 10^8}{5 \cdot 10^9 \sqrt{6.405}} = 2.37 \text{ cm} \qquad (9.142)$$

and the length of the coupler is

$$L = \lambda_g/4 = 5.93 \text{ mm}. \qquad (9.143)$$

Finally, the widths of the feed lines must be determined. The system impedance is $50\,\Omega$ and, from Table 4-3, the width, w', is found to be 0.954 mm. The final layout of the coupler is shown in Figure 9-15.

Electromagnetic analysis is often used to refine this synthesized coupler. There are three main uncertainties in the design. One is the uncertainty in the length of the coupler (due to the different even- and odd-mode effective permittivities). Second is the uncertainty in the connection of the feed lines (of width w') to the coupler. The final uncertainty is that the coupled-line equations come from low-frequency analysis—EM analysis will capture the frequency-dependent effects. However, only minor iteration would normally be required.

The final realization of a microstrip directional coupler as a laboratory component is shown in Figure 9-16, where SMA coaxial connectors are attached to the microstrip lines.

A final comment: the difference in the effective permittivities of the even and odd modes means that the even- (v_{pe}) and odd-mode (v_{po}) phase

velocities are different, with

$$v_{pe} = c/\sqrt{\varepsilon_{ee}} \quad \text{and} \quad v_{po} = c/\sqrt{\varepsilon_{eo}} . \qquad (9.144)$$

This is a problem in coupler design, but can be used to advantage in the design of filters based on coupled lines.

9.8 Common Impedance Coupling

So far coupling has been discussed in terms of the EM fields shared by two transmission lines. This is not the only way coupling of signals occurs. Sharing of a return path results in coupling, often called common impedance coupling, as there is a circuit element common to two or more transmission lines. The simplest situation is a shared impedance rather than a shared transmission line return, so that the return current attributed to one interconnect induces a voltage across the common impedance element. This signal then appears as though it was on the victim line. The common impedance could be the inductance or resistance of the ground conductor in the case of microstrip lines. In general, however, common impedance coupling will occur whenever the current return path is common.

9.9 The Lange Coupler

A directional coupler comprised of two parallel microstrip lines cannot achieve a coupling of 3 dB, which corresponds to splitting the power of a traveling wave into two equal components. This is a very useful function in constructing amplifiers, as this function would enable two amplifiers to be driven with equal amplitude signals derived from the same source and then at the output to be combined back together. Lange [136] introduced a coupler, now known as the Lange coupler, in 1969. The Lange coupler (see Figure 9-17) has a coupling factor of around 3 dB. In this design, true quadrature coupling over an octave is realized as a consequence of the interdigital coupling section, which compensates for the differences of the even- and odd-mode phase velocities over the wide frequency range. Note the use of the center bond wires—this was the key contribution of the Lange coupler. The bonding wires should look, electrically, as close as possible to a short-circuit—or at least as very small lumped inductances. This means that their lengths, l_s, must be kept as short as possible: $l_s \ll \lambda_{gm}/4$, where λ_{gm} is the midband wavelength. In semiconductor technologies, these bond wires are replaced by air bridges, and in structures with two or more metal layers the wirebonds are replaced by vias to another metal layer and a short connection on the second metal layer. In some designs, six coupling fingers are used instead of the four shown in Figure 9-17. Note that the input-to-direct output link meanders through the structure and this DC connection identifies the through connection.

   The physical length of the coupler is approximately one-quarter
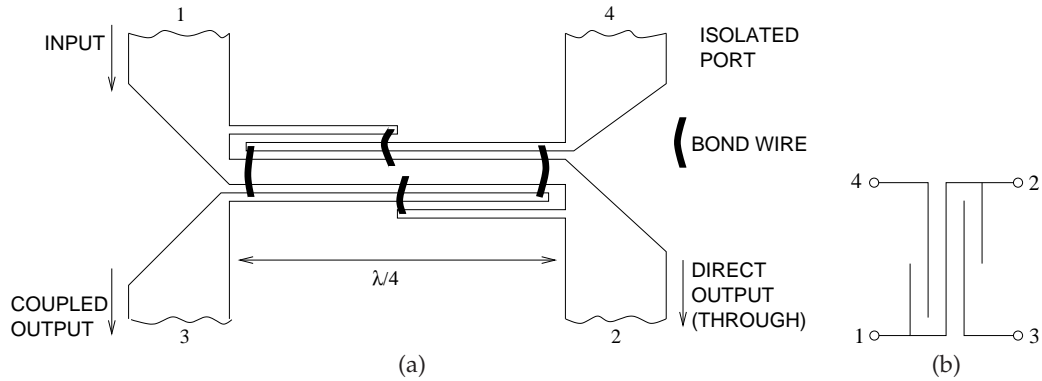
Figure 9-17   A four-finger Lange coupler: (a) microstrip layout; and (b) circuit symbol.
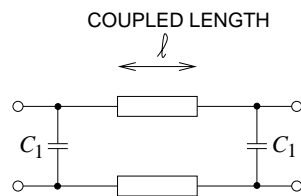
Figure 9-18   Schematic illustration of parallel coupled lines with lumped capacitors bridging the ends to provide compensation.

wavelength long at the center frequency of the coupling band. As with many distributed components, this element was invented using intuition and empirical iterations. Since then, analytic design formulas have been developed to enable synthesis of the electrical parameters of the coupler (see Edwards and Steer [51]). The synthesis is based on even- and odd-mode impedances analogous to those developed in Section 9.7 on Page 483 for a coupler comprised of coupled microstrip lines. Synthesis leads to a design that is close to ideal, and subsequent modeling in an EM simulator can be used to obtain an optimized design accounting for frequency-dependent and parasitic effects.

9.10   Directional Coupler With Lumped Capacitors

Directional couplers using only coupled transmission lines can be large at low frequencies, as the minimum length is approximately one-quarter of a wavelength. This can be a problem at RF and low microwave frequencies, say, below 3 GHz. The length of the line can be reduced by incorporating lumped elements, as shown in Figure 9-18. The development of couplers such as this is best approached using filter design theory, where the lumped elements and transmission line sections can be traded off. The required
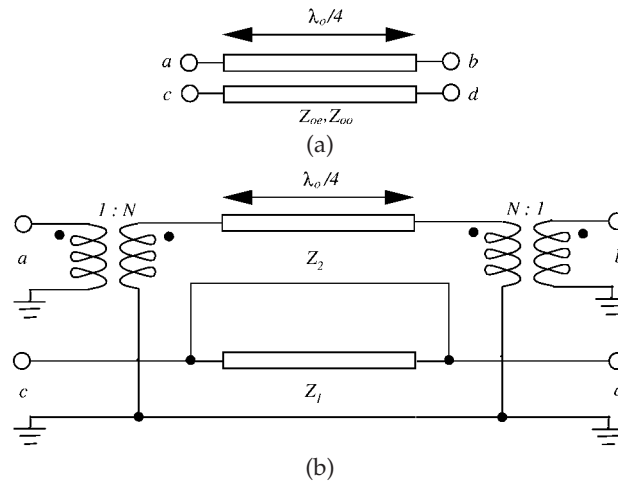
Figure 9-19 A pair of symmetrical coupled lines: (a) physical layout; and (b) its equivalent circuit model with $N = (Z_{0e} + Z_{0o})/(Z_{0e} - Z_{0o})$.

theory is presented in Chapter 10.

9.11　Models of Parallel Coupled Lines

9.11.1　ABCD Parameters of Parallel Coupled Lines in an Inhomogeneous Medium

Much of microwave network synthesis uses the equivalence of transmission line structures with each other and also the equivalence of transmission line structures with lumped-element circuits. Most often the equivalence is established using $ABCD$ parameters. $ABCD$ parameters are transmission parameters, but are in terms of total voltages and currents rather than traveling waves.

A physical pair of coupled lines is shown in Figure 9-19(a) and propagation on these lines is partly described by the even-mode characteristic impedance, Z_{0e}, and the odd-mode impedance, Z_{0o}, with the coupling coefficient defined as[11]

$$K = \frac{Z_{0e} - Z_{0o}}{Z_{0e} + Z_{0o}}. \tag{9.145}$$

The corresponding equivalent network model of the coupled line in Figure

---

[11] Previously a coupling factor, C, was used (see Equation (9.124), however, this introduces confusion since $ABCD$ parameters are being developed. Note that here K = 1/(coupling factor), that is, $K = 1/C$.
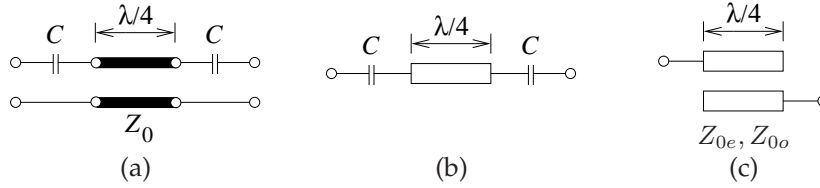
Figure 9-20 Equivalence of a capacitively loaded one-quarter wavelength long line and a parallel coupled line section: (a) a one-quarter wavelength long line with input and output series capacitance; (b) microstrip layout; and (c) coupled lines laid out in microstrip. The network in (a) is equivalent to the network in (c) if $Z_{0e} = (1/C + 2Z_0)$ and $Z_{0o} = Z_0$.

9-19(a) is shown in Figure 9-19(b) with the network parameters:

$$N = \frac{1}{K} \tag{9.146}$$

$$Z_1 = \frac{Z}{\sqrt{1 - K^2}} \tag{9.147}$$

$$Z_2 = Z\frac{\sqrt{1 - K^2}}{K^2}, \tag{9.148}$$

where

$$Z = \sqrt{Z_{0e}Z_{0o}}. \tag{9.149}$$

These lead to the $ABCD$ matrix of a pair of uniform, open-circuited coupled lines in an inhomogeneous media [137]:

$$
\begin{aligned}
A &= \frac{Z_{0e}\cot(\theta_e) + Z_{0o}\cot(\theta_o)}{Z_{0e}\csc(\theta_e) - Z_{0o}\csc(\theta_o)} = D \\
B &= \frac{\jmath}{2}\frac{Z_{0e}^2 + Z_{0o}^2 - 2Z_{0e}Z_{0o}(\cot(\theta_e)\cot(\theta_o) + \csc(\theta_e)\csc(\theta_o))}{Z_{0e}\csc(\theta_e) - Z_{0o}\csc(\theta_o)} \\
C &= \frac{2\jmath}{Z_{0e}\csc(\theta_e) - Z_{0o}\csc(\theta_o)},
\end{aligned}
\tag{9.150}
$$

where Z_{0e} and Z_{0o} are the modal impedances, and θ_e and θ_o are the even- and odd-mode phase lengths.

Developing the $ABCD$ parameters of the capacitively loaded one-quarter wavelength long line[12] in Figure 9-20(a) and equating then with the $ABCD$ parameters in Equation (9.150) leads to an equivalence between the loaded line and the one-quarter wavelength long pair of coupled lines shown in

[12] A one-quarter wavelength long line is an impedance inverter, so the results here apply in general to an impedance inverter with input and output series capacitances.
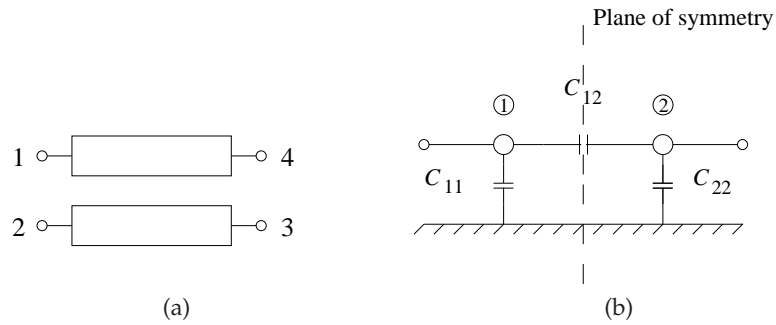
Figure 9-21 Coupled lines: (a) top view of a pair of coupled microstrip lines; and (b) capacitance model of the lines in crosssection.

Figure 9-20(c), where

$$Z_{0e} = (1/C + 2Z_0) \tag{9.151}$$

$$Z_{0o} = Z_0 \tag{9.152}$$

9.11.2 Network Model of Parallel Coupled Lines

The lumped-element filters considered up to now were synthesized using a ladder-type element extraction procedure followed by a variety of transformations for impedance scaling, filter type, and the Richards' transformation to develop distributed forms of a filter. The other important aspect of radio and microwave filter design is identifying structures that inherently have responses that match the synthesized responses. The most important structure that can be utilized to realize frequency selectivity is the coupled line. A pair of coupled lines is shown in Figure 9-21 together with its capacitance model shown in crosssection.

Recall the treatment of coupled lines in Section 9.3 on Page 464, which is repeated and extended here so that we can see how this model relates to a bandpass filter configuration:

$$Q_1 = C_{11}V_1 + C_{12}(V_1 - V_2) = (C_{11} + C_{12})V_1 - C_{12}V_2 \tag{9.153}$$

$$Q_2 = C_{22}V_2 + C_{21}(V_2 - V_1) = -C_{12}V_1 + (C_{22} + C_{12})V_2 \tag{9.154}$$

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \tag{9.155}$$

$$Q_e = C_{11}V_e = C_{22}V_e = C_eV_e \tag{9.156}$$

$$Q_o = (C_{11} + 2C_{12})V_o = C_oV_o \tag{9.157}$$

$$Z_{0e} = \frac{1}{V_{pe}C_e}, \quad Z_{0o} = \frac{1}{V_{po}.C_o} \tag{9.158}$$
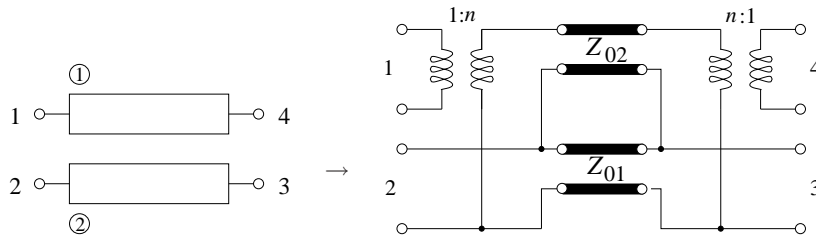
Figure 9-22 Network model of a pair of coupled lines.

Further derivation shows that a pair of coupled lines can be represented by the network model shown in Figure 9-22 [138]. The parameters of the network model are related to the model impedances as follows [138]:

$$n = \frac{1}{K} = \frac{Z_{0e} - Z_{0o}}{Z_{oe} + Z_{0o}} \tag{9.159}$$

$$Z_0 = \sqrt{(Z_{0e}Z_{0o})}, \qquad Z_{01} = \frac{Z}{\sqrt{1 - K^2}}, \qquad Z_{02} = Z\frac{\sqrt{1 - K^2}}{K^2}. \tag{9.160}$$

9.12 Differential and Common Modes

The parameters of coupled lines used in differential mode and common mode can be derived from the odd mode and even mode parameters. The difference is in the definition of the voltage and currents in the modes. These differences are illustrated in Figure 9-23. The even mode is defined with $V_1 = V_2 = V$ and $I_1 = I_2 = I$, while for the common mode $V_1 = V_2 = V$ and $I_1 + I_2 = I$. Thus, in terms of the even-mode characteristic impedance, Z_{0e}, the common-mode characteristic impedance is

$$Z_{0c} = \frac{1}{2}Z_{0e}. \tag{9.161}$$

The odd mode is defined with $V_1 = -V_2 = V$ and $I_1 = -I_2 = I$, while for the differential mode $V_1 - V_2 = V$ and $I_1 = -I_2 = I$. Thus, in terms of the odd-mode characteristic impedance, Z_{0o}, the differential-mode characteristic impedance is

$$Z_{0d} = 2Z_{0o}, \tag{9.162}$$

Other parameters remain unchanged. That is, the propagation constant, phase and group velocities, and wavelengths are the same for common and even modes, as they are for the differential and odd modes.

The driving and termination configurations for differential and common mode signals are shown in Figure 9-24. The reflectionless termination of a differential line is

$$R_L = Z_{0d} = 2Z_{0o}, \tag{9.163}$$

DEFINITION

(a)

EVEN MODE

(b)

COMMON MODE

(c)

ODD MODE

(d)

DIFFERENTIAL MODE

(e)
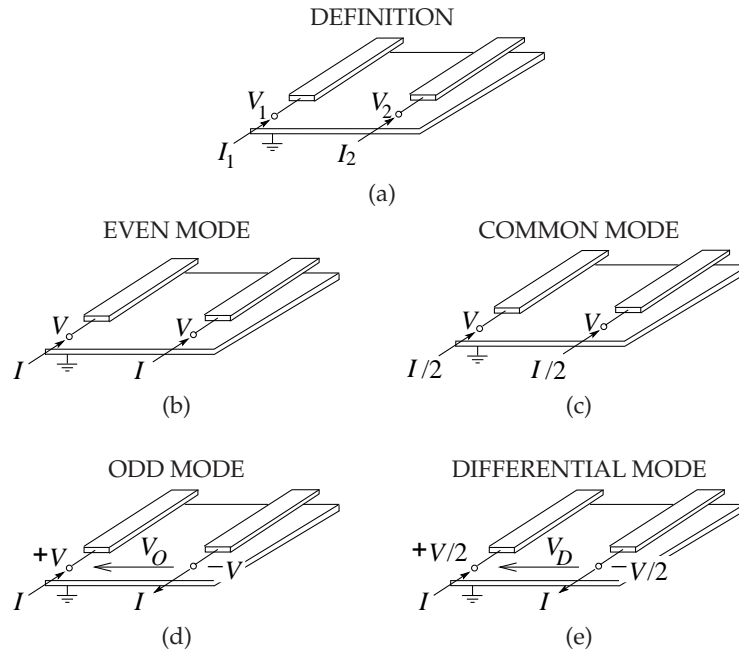
Figure 9-23 Definition of coupled line modes: (a) general definition of voltage and currents; (b) even mode definition; (c) common mode definition; (d) odd mode definition; (e) differential mode definition;

COMMON MODE

(a)

DIFFERENTIAL MODE

(b)
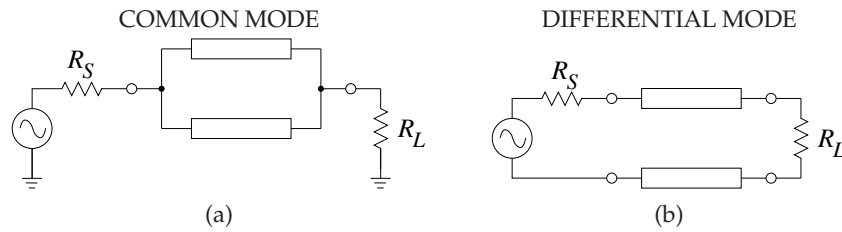
Figure 9-24 Driving configurations: (a) common mode; and (b) differential mode.

and the matched termination of coupled lines used in common mode is

$$R_L = Z_{0c} = \frac{1}{2}Z_{0e}. \tag{9.164}$$

9.13 Summary

Coupling from one transmission line to a nearby neighbor may often be undesirable. However, the effect can be exploited to realize a coupler that does not have a lumped-element equivalent. This is one example of many elements that utilize distributed transmission line effects to obtain novel functionality. Coupled lines are also vital components of many filters. The suite of elements that exploit distributed effects available to a microwave designer is surprisingly large.

9.14 Exercises

1. Consider the crosssection of a coupled transmission line, as shown in Figure 9-1.

   (a) For an even mode on the coupled line we can consider a phasor voltage of 1 V on each of the lines above the ground plane. Sketch the electric fields in the transverse plane (the plane of the crosssection). Treat the ground as being at 0 V.

   (b) For the even mode, sketch the magnetic fields in the transverse plane (the plane of the crosssection).

   (c) For an odd mode on the coupled line we can consider a phasor voltage of $+1$ V on the left line and a phasor voltage of -1 V on the right line. Sketch the electric fields in the transverse plane (the plane of the crosssection). Again, treat the ground as being at 0 V.

   (d) For the odd mode, sketch the magnetic fields in the transverse plane (the plane of the crosssection).

2. Electromagnetic software can be used to determine the even- and odd-mode parameters of a coupled line. This is usually done by setting phasor voltage on the coupled line and evaluating the phasor charge under each condition. The voltage applied to the left strip is V_1 and the voltage applied to the right strip is V_2. The phasor charge on the strips is Q_1 and Q_2, respectively. One part of the analysis is to redo the calculations, but this time with the substrate removed, so that medium is now free-space situation. In this case, the charges are denoted by Q_{01} and Q_{02}. The matrix of (computer-based) measurements is fol-

lows. [This problem parallels Example 9.3 on Page 486.]

Charge	$V_1 = 1$ V; $V_2 = -1$ V	$V_1 = 1$ V; $V_2 = 1$ V
Q_1 (pC/m)	40	20
Q_2 (pC/m)	-50	30
Q_{01} (pC/m)	13.25	6
Q_{01} (pC/m)	-10	2.75

   (a) What is the two-port capacitance matrix?
   (b) What is the even-mode capacitance?
   (c) What is the odd-mode capacitance?
   (d) What is the free-space (no dielectric) two-port capacitance matrix?
   (e) What is the free-space even-mode capacitance?
   (f) What is the free-space odd-mode capacitance?
   (g) What is the even-mode effective relative permittivity?
   (h) What is the odd-mode effective relative permittivity?

3. Two 50 Ω microstrip lines are to be run parallel to each on a printed circuit board with a relative permittivity $\varepsilon_r = 4$. The signal on the lines is 3 GHz. The thickness of the circuit board is 1 mm. This is the distance between the strips and the ground plane. The effective permittivity of the lines is 3.1 and it is determined that the approximate distance over which the lines will be in parallel is 1.42 cm. The coupling of the signals on the lines must be at least 30 dB down.

   (a) What is the free-space wavelength, λ_0, of the signal?

(b) What is the guide wavelength, λ_g, of the signal?

(c) How long is the parallel run of the microstrip lines in terms of λ_g?

(d) What is the maximum parallel-line coupling factor? Explain how you determined this.

(e) What is the minimum separation of the lines? Explain how you determined this distance

4. A pair of cooupled lines has an even-mode effective permittivity, ε_{ee}, of 4.9 and an odd-mode effective permittivity of 5.2.

   (a) What is the even-mode phase velocity?

   (b) What is the odd-mode phase velocity?

5. A directional coupler has the following characteristics: coupling factor C = 20, transmission factor 0.9, and directivity factor 25 dB. Also, the coupler is matched so that $S_{11} = 0 = S_{22} = S_{33} = S_{44}$.

   (a) What is the isolation factor in decibels?

   (b) Determine the power dissipated in the directional coupler if the input power to Port 1 is 1 W.

6. A lossy directional coupler has the following 50 Ω S parameters:

$$S = \begin{bmatrix} 0 & 0.25 & -0.9\jmath & 0.01 \\ 0.25 & 0 & 0.01 & -0.9\jmath \\ -0.9\jmath & 0.01 & 0 & 0.25 \\ 0.01 & -0.9\jmath & 0.25 & 0 \end{bmatrix}.$$

   (a) Which port is the input port (there are two possible answers)?

   (b) What is the coupling in decibels?

   (c) What is the isolation in decibels?

   (d) What is the directivity factor in decibels?

   (e) Draw the signal flow graph of the directional coupler.

7. Develop the design of a 10 dB directional coupler using coupled microstrip lines and a substrate with a permittivity of 10, a substrate thickness, h, of 600 μm, and a center frequency of 1 GHz. Develop the electrical design of the coupler (i.e., find the even- and odd-mode impedances required) and then develop the physical design (with widths and lengths) of the directional coupler. Use a system impedance of 50 Ω.

8. Design a microstrip directional coupler with the following specifications:

   Transmission line technology: Microstrip
   Coupling coefficient, C: 20 dB
   Characteristic impedance, Z_{0S}: 50 Ω
   Substrate permittivity, ε_r: 4.0
   Substrate thickness, h: 635 μm
   Center frequency, f_0: 10 GHz.

9. A directional coupler using coupled line is constructed on an alumina substrate of thickness 300 μm. The lines are 100 μm wide and the gap separation is 100 μm. What are the characteristic impedances, effective permittivities, and even- and odd-mode phase velocities of the even- and odd-mode of the coupled line. Port 1 is the input, Port 2 is the through output, and Port 3 is the coupled output. (Hint: You may need to use interpolation as described in Section A.6 on Page 830.)

   (a) Draw the schematic of the directional coupler and label the ports.

   (b) What is the transmission coefficient of the coupler?

   (c) What is the directivity, D, of the coupler?

   (d) Write down the 4×4 S parameter matrix of the coupler?

10. A directional coupler comprising a coupled pair of microstrip lines is to be designed in a 75 Ω system. The coupling factor is 10.

    (a) What is the odd-mode impedance of the coupler?

    (b) What is the even-mode impedance of the coupler?

11. Consider a pair of parallel microstrip lines separated by a spacing, s, of 100 μm.

    (a) What happens to the coupling factor of the lines as s reduces?

    (b) What happens to the system impedance as s reduces and no other dimensions change?

(a) In terms of wavelengths, what is the optimum length of the coupled lines for maximum coupling?

12. A 1 GHz microstrip directional coupler has a coupling factor of 20 dB. The coupler must have a system impedance of 50 Ω.

(a) Draw the layout of the directional coupler?

(b) What is the even-mode impedance of the coupler?

(c) What is the odd-mode impedance of the coupler?

(d) What is the optimum electrical length of the directional coupler in degrees at the design center frequency?

(e) If, in addition, the isolation of the directional coupler is 40 dB, what is its directivity in decibels.

13. What is the coupling factor of a Lange coupler in decibels.

14. A coupled microstrip line has an odd-mode impedance of 30 Ω and an even-mode impedance of 65 Ω.

(a) What is the differential characteristic impedance of the coupled lines?

(b) What is the common-mode characteristic impedance of the coupled lines?

15. A pair of coupled microstrip lines has an odd-mode characteristic impedance of 60 Ω and an even-mode characteristic impedance of 70 Ω. If a load resistance, R_L is placed at the end of the coupled lines from one line of the pair to the other, what is the value of R_L for no reflection of the even mode.

16. The coupled microstrip lines below have an odd-mode impedance of 30 Ω and an even-mode impedance of 60 Ω.
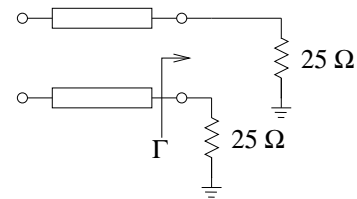
(a) What is the odd-mode reflection coefficient of the ?

(b) What is the common-mode characteristic impedance of the coupled lines?

17. The coupled microstrip lines below have an odd-mode impedance of 60 Ω and an even-mode impedance of 100 Ω. The coupled line have a length ℓ and this is a quarter of the wavelength of the odd-mode.
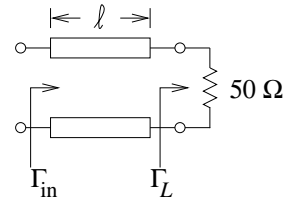
(a) What is the odd-mode reflection coefficient at the load, call this Γ_{Lo}?

(b) What is the odd-mode reflection coefficient at the input to the coupled lines, call this Γ_{ino}?

Filters

with Wael Fathelbab

10.1 Introduction

Filters are the most fundamental of signal processing circuits using energy storage elements such as capacitors, inductors, and transmission lines to obtain frequency-dependent characteristics. Some of the important objectives of a filter are (1) controlling noise by not allowing out-of-band noise to propagate in a circuit; (2) keeping signals outside the transmit band, especially harmonics, from being transmitted; (3) presenting only signals in a specified band to active receive circuitry; and (4) keeping a large transmit signal from the receive circuitry. At RF, a filter can consist solely of lumped elements, solely of distributed elements, or comprise a mix of lumped and distributed elements. The distributed realizations can be transmission line-based implementations of the components of lumped-element filter prototypes or, preferably, make use of particular frequency characteristics found with certain distributed structures. For example, coupled lines have particular frequency selective characteristics that can be exploited. Loss in lumped elements, particularly above a few gigahertz, means that the performance of distributed filters nearly always exceeds that of lumped-element filters. However, since the basic component of a distributed filter is a one-quarter wavelength long transmission line, distributed filters can be prohibitively large at a few gigahertz and lower in frequency.

Only a few basic types of responses are required of most RF filters as follows:

   (a) Lowpass, providing maximum power transfer, or passing, at frequencies below what is called the corner frequency, f_c. Above f_c, transmission is blocked. See Figure 10-1(a).

   (b) Highpass, passing signals at frequencies above f_c. Below f_c, transmission is blocked. See Figure 10-1(b).

   (c) Bandpass, passing signals at frequencies between lower and upper corner frequencies and blocking transmission outside the band. (The frequency span between the lower and upper corner frequencies defines the passband.) This is the most common type of RF filter. See Figure 10-1(c).

   (d) Bandstop (or notch), which blocks signals between lower and upper corner frequencies. See Figure 10-1(d).

   (e) Allpass, which equalizes a signal by adjusting the phase generally to correct for phase distortion in broadband circuits. See Figure 10-1(e).

The above list is not comprehensive, as actual operating conditions may mandate specific frequency profiles. For example, IEEE 802.11a (WiFi) WLAN systems operating at 2.45 GHz are susceptible to potentially large signals being transmitted from nearby cellular phones operating in the
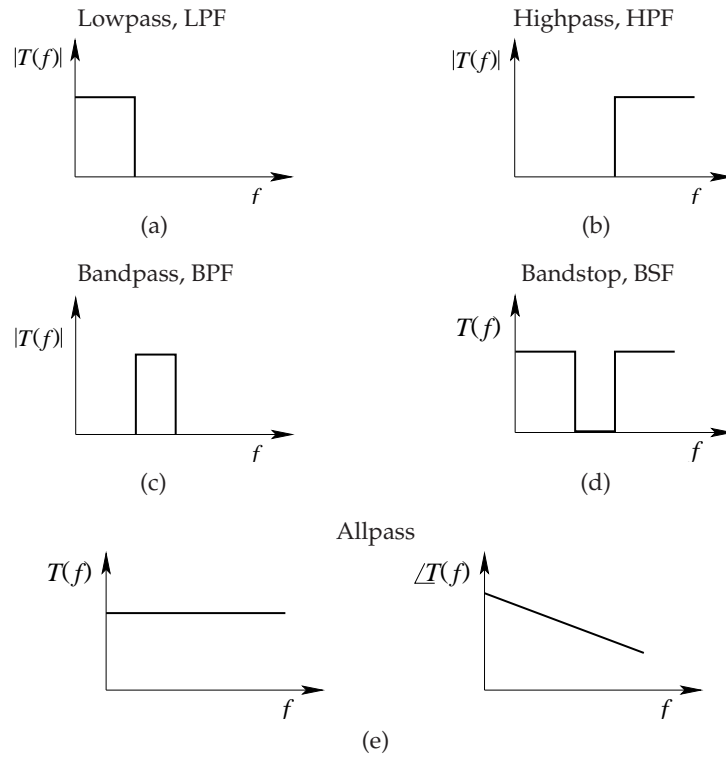
Figure 10-1 Ideal filter responses: (a) lowpass network response; (b) highpass network response; (c) bandpass network response; (d) bandstop network response; and (e) allpass network response.

1700 to 2300 MHz range. Thus a front-end filter for a 2.45 GHz system must ensure very high levels of attenuation over the 1700 to 2300 MHz range while also having low insertion loss in-band at 2.45 GHz. Thus, for a given complexity, the optimum solution probably will have an asymmetrical frequency response with high rejection below the passband obtained by accepting lower rejection on the other side of the passband.

In the 1960s an approach to RF filter design and synthesis was developed and this is still followed. The approach is to translate the response of the RF filter (often a bandpass filter) from that of a lowpass filter. A filter with the desired lowpass response is then synthesized using lumped elements and the resulting filter is called a lowpass prototype. The lowpass filter is then transformed so that the new filter has the desired response, such as highpass or bandpass. In the case of a bandpass filter, each inductor and capacitor of the lowpass prototype becomes a resonator that is coupled to another resonator. In distributed form, this basic resonator is a one-quarter wavelength long transmission line.
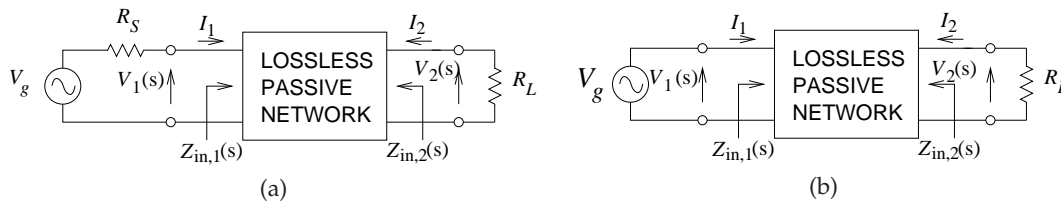
Figure 10-2 Terminated networks: (a) a doubly terminated network; and (b) a singly terminated network.

Filter synthesis is the origin of a systematic approach to realizing circuits with desired frequency characteristics. A filter can also be thought of as a hardware implementation of a differential equation producing a specific impulse, step, or frequency response. It is not surprising that a filter can be described this way, as the lumped reactive elements, L and C, describe first-order differential equations and their interconnection describes higher-order differential equations.

The synthesis of a filter response begins by identifying the desired differential equation response. The differential equations are specified using the Laplace variable, s, so that in the frequency domain (with $s = \jmath\omega$) an nth-order differential equation becomes an nth-order polynomial expression in s. If possible, and this is usually the case, a lowpass version of a filter can be developed. The conversion to the final filter shape, say a bandpass response, proceeds mathematically through a number of stages. Each stage has a circuit form and each stage is called a filter prototype.

The other aspect of filter design is the identification of particular structures that inherently have the desired frequency-shaping attributes. Filter design is both science and art matching synthesis with instinctive knowledge of physical and circuit structures. However, once a suitable filter prototype has been identified it can be optimized using optimization tools incorporated in circuit analysis packages.

Since the beginnings of circuit synthesis, particularly since the 1950s, it has become common to present circuit concepts using examples, very often because the steps in realizing a filter may be too difficult to specify algorithmically and there are many steps which require intuition. This procedure is followed here.

10.2 Singly and Doubly Terminated Networks

Filters are generally two-port networks and they provide maximum power transfer from a load to a source over a specified frequency range while rejecting the transmission of signals over other ranges. Two possible filter networks are shown in Figure 10-2. The network in Figure 10-2(a) is known as a doubly terminated network, as both ports are resistively terminated. The network in Figure 10-2(b) is called a singly terminated network, as only

one port is terminated in a resistor. Most synthesis of singly terminated filter networks requires an analogous procedure to that presented here. It is just that the doubly terminated network is much closer to the type of network required at RF where loads and source impedances are finite. A singly terminated network is applicable in some RFIC applications where very little RF power is involved and often when feedback is used. In such cases the output of an RFIC amplifier can approximate an ideal voltage source, as the Thevenin equivalent source impedance is negligible.

10.2.1 Doubly Terminated Networks

The established synthesis procedure for doubly terminated filter networks focuses on realizing the input reflection coefficient, which is related to the transmission coefficient by the characteristic function. The input reflection coefficient of the doubly terminated network (Figure 10-2(a)) is

$$\Gamma_1(s) = \frac{Z_{\text{in},1}(s) - R_S}{Z_{\text{in},1}(s) + R_S},\tag{10.1}$$

where the reference impedance is the source resistance, R_S, and s is the Laplace variable, so the ideal response in the passband is when the reflection coefficient is approximately zero. There are several other parameters used with filters and these will be introduced now. The **Transducer Power Ratio** (**TPR**) is defined as

$$
\begin{aligned}
\text{TPR} &= \frac{\text{Maximum power available from the source}}{\text{Power absorbed by the load}} \\
&= \frac{\frac{1}{2}(V_g(s)/2)^2/R_S}{\frac{1}{2}V_2^2(s)/R_L} = \left| \frac{1}{2}\sqrt{\frac{R_L}{R_S}}\frac{V_g(s)}{V_2(s)} \right|^2.
\end{aligned}\tag{10.2}
$$

The transmission coefficient, $T(s)$, of the network is

$$T(s) = \frac{1}{\sqrt{\text{TPR}(s)}} = 2\sqrt{\frac{R_S}{R_L}}\frac{V_2(s)}{V_g(s)},\tag{10.3}$$

so that the insertion loss (IL) (or **transducer function**) is:

$$\text{IL}(s) = \text{TPR}(s) = |1/T(s)|^2 .\tag{10.4}$$

Note that this development assumes that the filter is lossless. The next step in development of the synthesis procedure is introduction of the **characteristic function**, defined as the ratio of the reflection and transmission coefficients:

$$K(s) = \frac{\Gamma_1(s)}{T(s)} = \frac{N(s)}{D(s)},\tag{10.5}$$

where N is the numerator function and D is the denominator function. Ideally the filter network is lossless and so, from the principle of energy conservation, the sum of the magnitudes squared of the transmission and reflection coefficients must be unity:

$$|T(s)|^2 + |\Gamma_1(s)|^2 = 1 \tag{10.6}$$

or

$$|T(s)|^2 = 1 - |\Gamma_1(s)|^2. \tag{10.7}$$

Dividing both sides of Equation (10.7) by $|T(s)|^2$ results in

$$1 = \frac{1}{|T(s)|^2} - \frac{|\Gamma_1(s)|^2}{|T(s)|^2}, \tag{10.8}$$

or

$$1 = |\text{IL}(s)|^2 - |K(s)|^2. \tag{10.9}$$

Rearranging Equation (10.9) leads to

$$|\text{IL}(s)|^2 = 1 + |K(s)|^2 \tag{10.10}$$

and so

$$|T(s)|^2 = \frac{1}{1 + |K(s)|^2} \tag{10.11}$$

$$|\Gamma_1(s)|^2 = \frac{|K(s)|^2}{1 + |K(s)|^2}. \tag{10.12}$$

It is seen that both the reflection and transmission coefficients are governed by the characteristic function of the two-port network. A milestone has been reached. In an RF filter, the frequency-dependent insertion loss or transmission coefficient is of most importance, as these are directly related to power flow. Equation (10.11) shows that this can be expressed in terms of another function, $K(s)$, which, from Equation (10.5), can be expressed as the ratio of $N(s)$ and $D(s)$. For lumped-element circuits, $N(s)$ and $D(s)$ are polynomials. Sometimes it is better to focus design effort on realizing $\Gamma_1(s)$, but the quantities are interrelated.

10.2.2   Lowpass Filter Response

As an example of the relationship of the various filter responses consider the typical lowpass filter responses shown in Figure 10-3. This filter has a lowpass response with a corner frequency expressed in radians as $\omega_c = 1\text{ rad}/s$. Ideally $T(s)$ for $s \leq \jmath$ (note $s = \jmath\omega$) would be one, and for $s > \jmath$, $T(s)$ would be zero. It is not possible to realize such an ideal response, and the response shown in Figure 10-3(a) is typical of what can be achieved.
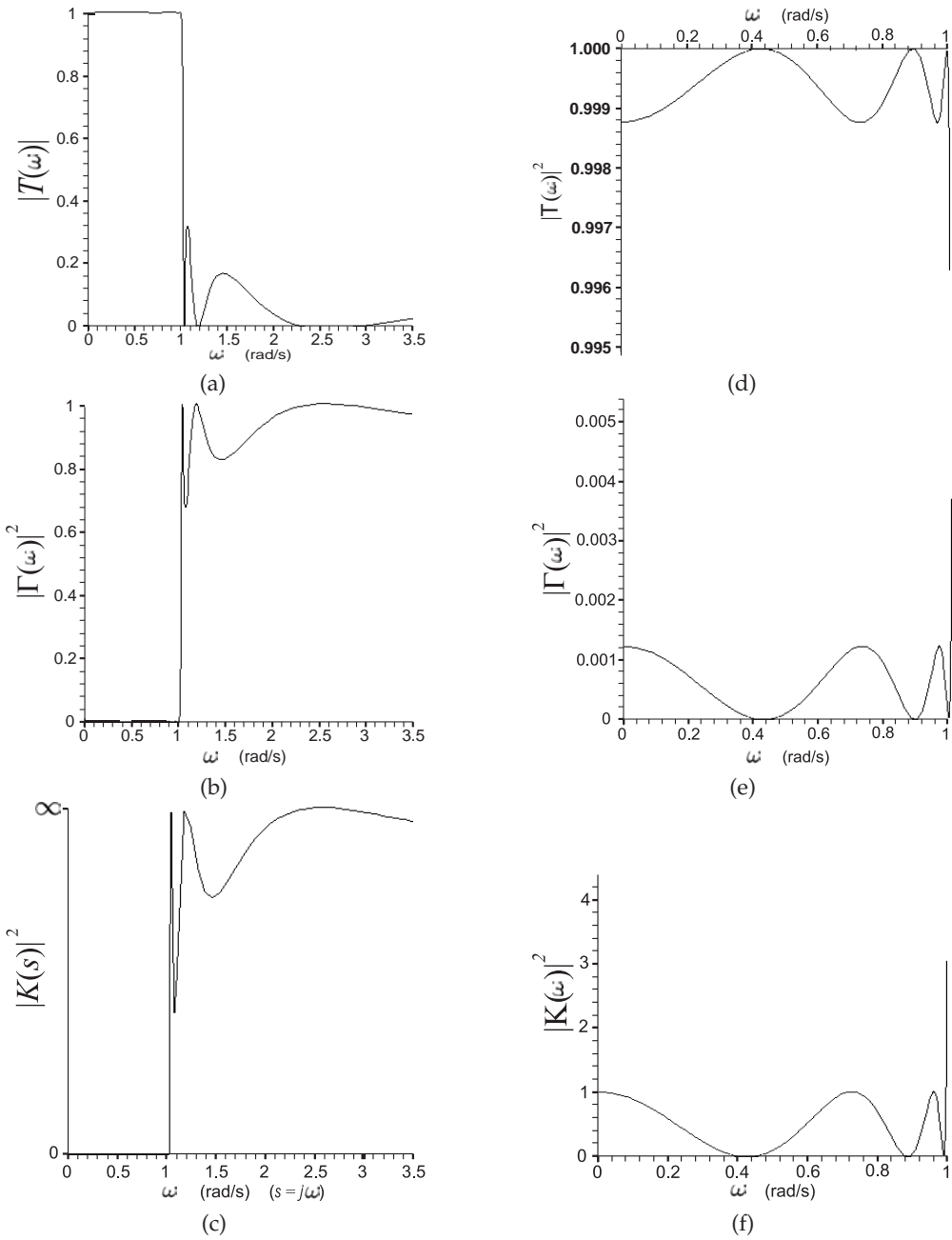
Figure 10-3 An example of a lowpass filter in terms of various responses: (a) transmission coefficient; (b) reflection coefficient response; and (c) characteristic function response. Detailed responses are shown in (d), (e), and (f), respectively.
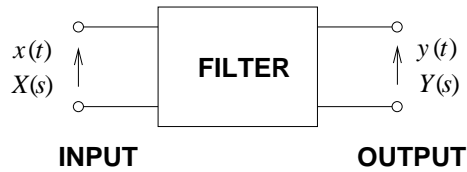
Figure 10-4   Generic filter.

With the filter being lossless, the reflection coefficient is as shown in Figure 10-3(b), with the characteristic function $K(s)$ shown in Figure 10-3(c). If $|K(s)|^2$ is expressed as the ratio of two polynomials (i.e, as $N(s)/D(s)$), then it can be seen from Figures 10-3(e) and 10-3(f) that the zeros of the reflection coefficient, $|\Gamma_1(s)|^2$, are also the zeros of $N(s)$. Also, it is observed that the zeros of the transmission coefficient are also the zeros of $D(s)$, as shown in Figures 10-3(a) and 10-3(c).

   The first objective in lumped-element filter design is development of the transfer function of the network in the frequency domain or, equivalently for the purposes here, the s domain.[1] The input-output transfer function of the generic filter in Figure 10-4 is

$$T(s) = Y(s)/X(s) \tag{10.13}$$

and the design procedure is to match this response to a response defined by the ratio of polynomials:

$$T(s) = \frac{N(s)}{D(s)} = \frac{a_m s^m + a_{m-1} s^{m-1} + \cdots + a_1 s + a_0}{s^n + b_{n-1} s^{n-1} + \cdots + b_1 s + b_0}. \tag{10.14}$$

Here N stands for numerator and D for denominator and are not the same as those in Equation (10.5) (where they also were just labels for numerator and denominator). Both $N(s)$ and $D(s)$ are polynomials in the **Laplace** variable, s. The filter response using a **pole-zero description** can be synthesized so the design process begins by rewriting Equation (10.14) explicitly in terms of zeros, z_m, and poles, p_n:

$$T(s) = \frac{N(s)}{D(s)} = \frac{a_m(s + z_1)(s + z_2) \cdots (s + z_{m-1})(s + z_m)}{(s + p_1)(s + p_2) \cdots (s + p_{n-1})(s + p_n)}. \tag{10.15}$$

Since only the frequency response is of interest, $s = \jmath\omega$, where ω, the angular frequency, is measured in radians per second, thus simplifying the analysis for sinusoidal signals. However, the development here is general and s can be complex, which would correspond to an exotic filter response which is of no interest here.

---

[1]   Care is being taken in the use of terminology as s can be complex with some filter types, but these filters are generally only realized using digital signal processing.
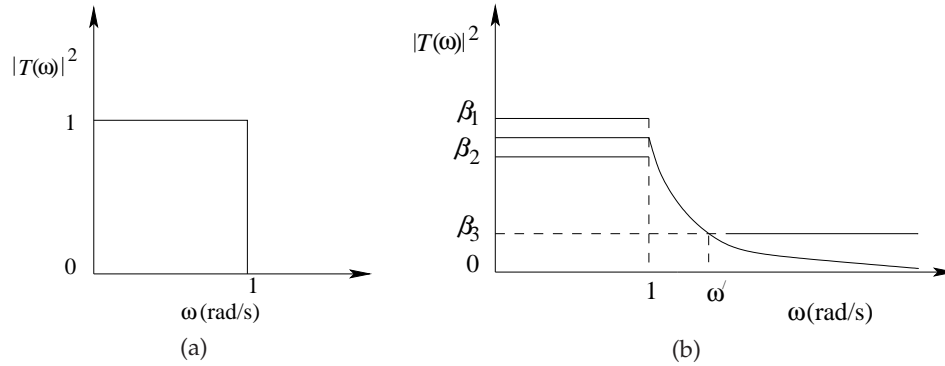
Figure 10-5 Lowpass filter response: (a) ideal "brick-wall" response; and (b) response template for a realizable filter.

The poles and zeroes can be complex numbers and can be plotted on the complex s plane. Conditions imposed by realizable circuits require that $D(s)$ be a **Hurwitz** polynomial which ensures that its poles are located in the left-half plane. $N(s)$ determines the location of the transmission zeros of the filter, and the order of $N(s)$ cannot be more than the order of $D(s)$. That is, $n \geq m$ so that the filter has finite or zero response at infinite frequency.

Two strategies can be employed in deriving the filter response. The first is to derive the polynomials $N(s)$ and $D(s)$ in Equation (10.14). This seems like an open-ended problem, but it turns out that in normal situations there are only a few types of useful responses which are described by a few polynomials, including Butterworth, Bessel, Chebyshev, and Cauer polynomials.

10.3 The Lowpass Filter Prototype

Most filter design is based on the synthesis of a lowpass filter equivalent called a lowpass prototype. Transformations are then used to correct for the actual source and load impedances, frequency range, and the desired filter type, such as highpass or bandpass. An ideal lowpass response is shown in Figure 10-5(a), which shows a transmission coefficient of 1 up to a normalized frequency of 1 rad/sec. This type of response defines what is called a **brick-wall filter**, which unfortunately cannot be physically realized. Instead, the response is approximated and specified in terms of a response template (see Figure 10-5(b)).

The template specifies a passband response that is between β_1 and β_2 at frequencies below the corner radian frequency ($\omega = 1$), and below β_3 above a radian frequency $\omega' = 1$. The lowpass filter is harder to realize the closer the specified response is to the ideal shown in Figure 10-5(a), that is, when

$$(\beta_2 - \beta_1) \rightarrow 0, \qquad \beta_3 \rightarrow 0, \qquad (\omega' - 1) \rightarrow 0. \qquad (10.16)$$

Several polynomials have particularly interesting characteristics that match the requirements of the response template. At first it may seem surprising that polynomial functions could yield close to ideal filter responses. However, as will be seen, there is something special about these polynomials: they are natural solutions to extreme conditions. For example, the nth-order Butterworth polynomial has the special property that the first n derivatives at $s = 0$ are zero. The other polynomials have similar extreme properties, and filters that are synthesized from expressions incorporating them also have extreme properties. As well as the maximally flat property resulting from Butterworth polynomials, filter responses obtained using Chebyshev polynomials have the steepest skirts (or fastest transitions from the frequency region where signals are passed to the region where they are blocked). Usually it is one of these extreme cases that is most desirable.

A streamlining of the filter synthesis procedure is obtained by first focusing on the development of normalized lowpass filters having a 1 rad/sec corner frequency and $1\ \Omega$ reference impedance. Reasonably simple transformations transform a lowpass filter prototype into another filter having the response desired. These transformations, as will be seen, give rise to symmetrical responses.

10.4   The Maximally Flat (Butterworth) Lowpass Approximation

Butterworth filters have a maximally flat response (Figure 10-6) which in the time domain corresponds to a critically damped system. Butterworth filters have the transfer function

$$T(s) = \frac{N(s)}{D(s)} = \frac{k}{s^n + b_{n-1}s^{n-1} + \cdots + b_1 s + b_0}. \tag{10.17}$$

This is an all-pole response. The characteristic polynomial of the Butterworth filter is

$$|K(s)|^2 = |s^{2n}| = \omega^{2n}, \tag{10.18}$$

since $s = \jmath\omega$ and where n is the order of the function. Thus the transmission coefficient is

$$|T(s)|^2 = \frac{1}{1 + |K(s)|^2} = \frac{1}{1 + |s^{2n}|} = \frac{1}{1 + \omega^{2n}}, \tag{10.19}$$

which is often written as

$$|T(s)|^2 = \frac{1}{1 + |K(s)|^2} = \frac{1}{B_n(s)B_n(-s)}. \tag{10.20}$$
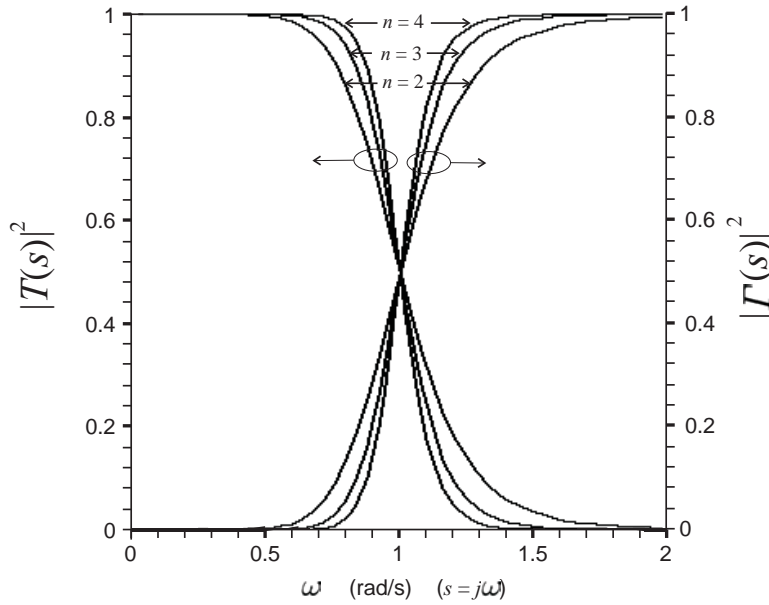
Figure 10-6 Maximally flat, or Butterworth, lowpass filter approximation for various orders, n, of the filter.

In the filter community, $B_n(s)$ is called the Butterworth polynomial and has the general form

$$B_n(s) = \prod_{k=1}^{n/2} \left[s^2 - 2s \cos\left(\frac{2k+n-1}{2n}\pi \right) + 1 \right] \quad \text{for } n \text{ even} \tag{10.21}$$

$$B_n(s) = (s+1) \prod_{k=1}^{(n-1)/2} \left[s^2 - 2s \cos\left(\frac{2k+n-1}{2n}\pi \right) + 1 \right] \quad \text{for } n \text{ odd.} \tag{10.22}$$

In factorized form the $B_n(s)$ polynomial is given in Table 10-1. Further factorization of the second-order factors results in complex conjugate roots, so they are generally left in the form shown.

Insight is gained by examining the characteristic polynomial for a real radian frequency, ω (where $s = \jmath\omega$), at the following frequency points:

$$\text{at } \omega = 0 \quad \rightarrow \quad |T(0)|^2 = \frac{1}{1+0} = 1 \tag{10.23}$$

$$\text{at } \omega = 1 \quad \rightarrow \quad |T(1)|^2 = \frac{1}{1+1} = \frac{1}{2}. \tag{10.24}$$

Thus the transmission response of the filter is at half power at the corner frequency $\omega = 1$. Another observation is that the transmission

Table 10-1   Factors of the $B_n(s)$ polynomial.

n	Factors of $B_n(s)$
1	$(s+1)$
2	$(s^2 + 1.4142s + 1)$
3	$(s+1)(s^2 + s + 1)$
4	$(s^2 + 0.7654s + 1)(s^2 + 1.8478s + 1)$
5	$(s+1)(s^2 + 0.6180s + 1)(s^2 + 1.6180s + 1)$
6	$(s^2 + 0.5176s + 1)(s^2 + 1.4142s + 1)(s^2 + 1.9319s + 1)$
7	$(s+1)(s^2 + 0.4450s + 1)(s^2 + 1.2470s + 1)(s^2 + 1.8019s + 1)$
8	$(s^2 + 0.3902 + 1)(s^2 + 1.1111s + 1)(s^2 + 1.6629s + 1)(s^2 + 1.9616s + 1)$

coefficient at $\omega = 0$ and $\omega = 1$ are independent of the degree of the characteristic polynomial. This is shown in Figure 10-6, where the responses of Butterworth filters are plotted for three different orders.

10.4.1   Construction of the Transfer Function

Often the characteristic function, K, is given in the variable ω, and thus for synthesis purposes the response must be transformed back to the s domain. As an example, consider the maximally flat function of third order with a characteristic polynomial equal to ω^3 (from Equation (10.19)):

$$K(\omega) = \omega^3 \quad \text{and} \quad |T(\omega)|^2 = \frac{1}{1 + \omega^6} . \tag{10.25}$$

Note that the first six derivatives of $|K(s)|^2$ with respect to s are zero at $s = 0$. So, from Equation (10.7),

$$|\Gamma_1(\omega)|^2 = 1 - \frac{1}{1 + \omega^6} = \frac{\omega^6}{1 + \omega^6}. \tag{10.26}$$

At real frequencies (or frequencies lying on the imaginary axis in the s plane)

$$s = \jmath\omega, \tag{10.27}$$

and so to transform from ω to s,

$$\omega = s/\jmath = -\jmath s \tag{10.28}$$

is used. Thus for the example under consideration, the reflection coefficient becomes

$$|\Gamma_1(\jmath\omega)|^2 = |\Gamma_1(\jmath(-\jmath s))|^2 = \frac{(-\jmath s)^6}{1 + (-\jmath s)^6}. \tag{10.29}$$

That is,

$$|\Gamma_1(s)|^2 = \frac{-s^6}{1 - s^6} = \Gamma_1(s)\Gamma_1(-s). \tag{10.30}$$

Factoring the denominator polynomial yields

$$1 - s^6 = (1 - s)(1 + s)(s^2 + s + 1)(s^2 - s + 1). \qquad (10.31)$$

By choosing those factors with roots only in the left-half plane (required for a realizable network), the zeros of the denominator are obtained. Now the numerator of $|\Gamma_1(s)|^2$ is easily factored, and since this is third order it only has three reflection zeros at DC. The resulting function is

$$\Gamma_1(s) = \frac{s^3}{(s+1)(s^2 + s + 1)}, \qquad (10.32)$$

and so

$$T(s) = \frac{1}{(s+1)(s^2 + s + 1)}. \qquad (10.33)$$

This example illustrates how $\Gamma_1(s)$ and $T(s)$ can be obtained from $K(s)$ specified as a Butterworth polynomial. This procedure is generalized in the next section.

10.4.2 nth-Order Reflection Approximation

Following the procedure outlined in the previous section, generalization of the Butterworth response leads to the nth-order response

$$
\begin{aligned}
|\Gamma_1(s)|^2 &= \Gamma_1(s).\Gamma_1(-s) \qquad (10.34)\\
&= \frac{\left(-s^2\right)^n}{1 + \left(-s^2\right)^n} = \frac{\left(-s^2\right)^n}{\displaystyle\prod_{i=1}^{n}(s - s_i).\prod_{j=n+1}^{2n}(s - s_j)}.
\end{aligned}
$$

So $|\Gamma_1(s)|^2$ has n roots (these are the s_is) lying in the left-half s plane, and n roots (the s_js) lying in the right-half s plane. Note that j is used as an index and $\jmath$ ($\jmath$, without the dot) represents $\sqrt{-1}$. It is reasonable to group all of the left-half plane roots together (a circuit could not be synthesized if there were a right-half plane root) so that

$$\Gamma_1(s) = \frac{\pm (s)^n}{\displaystyle\prod_{i=1}^{n}(s - s_i)}. \qquad (10.35)$$

Solving for the roots of the denominator of Equation (10.35) (i.e., finding the roots of $1 + \left(-s^2\right)^n = 0$) yields the following roots in the left-half s plane:

$$s_i = \exp\left\{\jmath(2i - 1 + n)\frac{\pi}{2n}\right\} \quad i = 1, 2, ..., n, \qquad (10.36)$$

where $\exp(\jmath\theta) = \cos(\theta) + \jmath\sin(\theta)$. There are also roots in the right-half plane, but these are just the negative of the left-half plane.

EXAMPLE 10. 1   Reflection Coefficient Derivation for Butterworth Filter

Develop the reflection coefficient from the roots of the third-order ($n = 3$) Butterworth lowpass filter prototype.

Solution: From Equation (10.36) the three roots are

$$s_1 = \exp\left\{\jmath\,(2 \times 1 - 1 + 3)\frac{\pi}{2 \times 3}\right\} = \exp\left\{\jmath\,\frac{2}{3}\pi\right\} \tag{10.37}$$

$$s_2 = \exp\left\{\jmath\,(2 \times 2 - 1 + 3)\frac{\pi}{2 \times 3}\right\} = \exp\left\{\jmath\,\pi\right\} \tag{10.38}$$

$$s_3 = \exp\left\{\jmath\,(2 \times 3 - 1 + 3)\frac{\pi}{2 \times 3}\right\} = \exp\left\{\jmath\,\frac{4}{3}\pi\right\} \tag{10.39}$$

and the input reflection coefficient of the filter is

$$\Gamma_1\,(s) = \frac{s^3}{\left(s + \frac{1}{2} - \jmath\frac{\sqrt{3}}{2}\right)(s + 1)\left(s + \frac{1}{2} + \jmath\frac{\sqrt{3}}{2}\right)} = \frac{s^3}{(s + 1)\,(s^2 + s + 1)}. \tag{10.40}$$

This is the same as Equation (10.32).

10.5   The Chebyshev Lowpass Approximation

The maximally flat approximation to the ideal lowpass filter response is best near the origin but not so good near the band edge. Chebyshev filters have better responses near the band edge, with lower insertion loss near the edges, but at the cost of ripples in the passband. Chebyshev filters are also known as **equiripple** all-pole lowpass filters. Example reflection and transmission responses are shown in Figure 10-7 for a seventh-order and a sixth-order Chebyshev lowpass filter. The general form of the Chebyshev transmission coefficient is

$$|T(s)|^2 = \frac{1}{1 + \varepsilon^2|K(s)|^2}\,, \tag{10.41}$$

where ε is an extra design parameter and defines the **passband ripple (PBR)**:

$$\text{PBR} = \frac{1}{1 + \varepsilon^2} \qquad \text{or} \qquad \text{PBR}\,|_{\text{dB}} = 10\log_{10}\left(\frac{1}{1 + \varepsilon^2}\right). \tag{10.42}$$

For the nth-order Chebyshev (lowpass filter) approximation the square of the characteristic function is

$$|K_n\,(\omega)|^2 = \cos^2\left[n\cos^{-1}\,(\omega)\right], \tag{10.43}$$
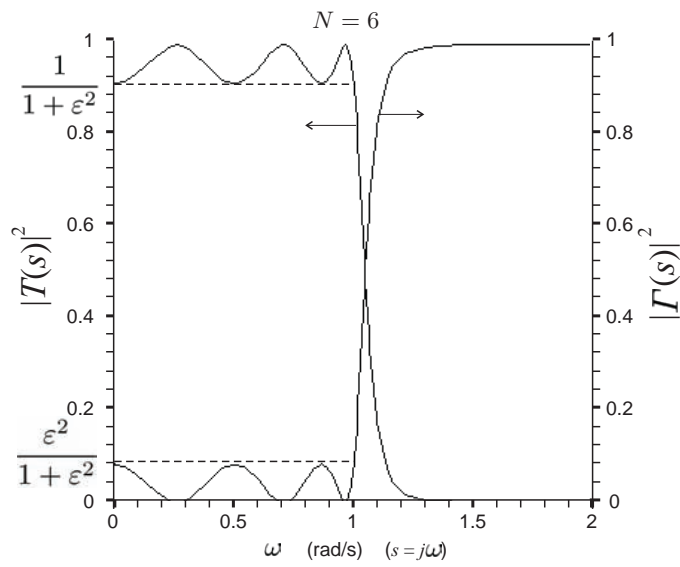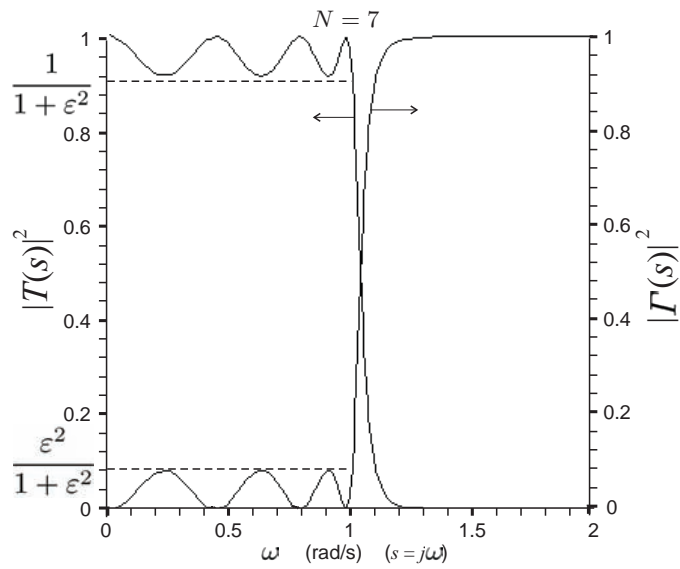
Figure 10-7   Chebyshev lowpass filter approximation: (a) seventh-order filter; and (b) sixth-order filter.

which can be expressed as a polynomial. For example, with $n = 3$,

$$K_3(\omega) = \cos\left[3\cos^{-1}(\omega)\right] = 4\omega^3 - 3\omega, \qquad (10.44)$$

and it is surprising that the trigonometric expression has such a simple polynomial equivalence. From Equation (10.11) the transmission coefficient is

$$|T(\omega)|^2 = \frac{1}{1 + \varepsilon^2 \cos^2\left[n\cos^{-1}(\omega)\right]} \qquad (10.45)$$

and the reflection coefficient is

$$|\Gamma_1(\omega)|^2 = \frac{\varepsilon^2 \cos^2\left[n\cos^{-1}(\omega)\right]}{1 + \varepsilon^2 \cos^2\left[n\cos^{-1}(\omega)\right]}. \qquad (10.46)$$

Factorizing the denominator of either Equation (10.45) or Equation (10.46) yields the following roots (of the denominator of $\Gamma_1(s)$ and $T(s)$):

$$
\begin{aligned}
s_i \;=\; & \sin\left[\frac{(2i-1)\pi}{2n}\right]\sinh\left[\frac{1}{n}\sinh^{-1}\left(\frac{1}{\varepsilon}\right)\right] \\
& + \jmath\cos\left[\frac{(2i-1)\pi}{2n}\right]\cosh\left[\frac{1}{n}\sinh^{-1}\left(\frac{1}{\varepsilon}\right)\right] \qquad i = 1, 2, 3 \,. \;(10.47)
\end{aligned}
$$

The roots of the numerator of $\Gamma_1(s)$ in the s plane are

$$s_j = \jmath\cos\frac{(2j-1)\pi}{2n} \qquad j = 1, 2, .., n \,. \qquad (10.48)$$

Equations (10.47) and (10.48) may be utilized to obtain the reflection and transmission coefficients directly in the s domain.

10.5.1 Chebyshev Approximation and Recursion

It is possible to obtain the characteristic function of the Chebyshev approximation from the recursion formula,

$$K_n(\omega) = 2\omega K_{n-1}(\omega) - K_{n-2}(\omega), \qquad (10.49)$$

with

$$K_1(\omega) = \omega; \qquad K_2(\omega) = 2\omega^2 - 1. \qquad (10.50)$$

For example, with $n = 3$,

$$
\begin{aligned}
K_3(\omega) \;&=\; 2\omega K_{3-1}(\omega) - K_{3-2}(\omega) \qquad\qquad (10.51)\\
&=\; 2\omega\left(2\omega^2 - 1\right) - \omega = 4\omega^3 - 2\omega - \omega = 4\omega^3 - 3\omega.
\end{aligned}
$$

10.6 Element Extraction

In the previous two sections the mathematical responses of Butterworth and Chebyshev filters were derived for various orders. In this section it will be shown how these filters can be implemented with inductors and capacitors using what is called **ladder synthesis** [139]. To obtain the element values yielding the desired transfer function, an impedance or admittance function must first be obtained. The impedance or admittance function can be readily obtained from the input reflection coefficient of a network, but for now the focus is on synthesizing a given impedance function. A general impedance function can be expressed as

$$Z(s) = \frac{a_n \left(s^2 + \omega_1^2\right) \left(s^2 + \omega_3^2\right) \left(s^2 + \omega_5^2\right) \dots}{b_m s \left(s^2 + \omega_2^2\right) \left(s^2 + \omega_4^2\right) \left(s^2 + \omega_6^2\right) \dots}, \tag{10.52}$$

where a_n and b_m are constants. This can be realized by extracting L and C elements, provided that the degree of the numerator and denominator differ by unity $|m-n| = 1$. The lumped-domain elements are inductors, capacitors, and transformers. The poles and zeroes of $Z(s)$ correspond to zeroes of the transmission coefficient. A pole of $Z(s)$ at DC requires the extraction of a series capacitor of value

$$C_0 = \left. \frac{1}{sZ(s)} \right|_{s=0}, \tag{10.53}$$

while a pole at infinity requires the extraction of a series inductor of value

$$L_\infty = \left. \frac{Z(s)}{s} \right|_{s=\infty}. \tag{10.54}$$

Another possibility is a pole at a finite frequency (call this ω_0), which requires the extraction of a series parallel LC block, as shown in Figure 10-9(e), with a capacitor of value

$$C_i = \left. \frac{s}{\left(s^2 + \omega_0^2\right) Z(s)} \right|_{s=j\omega_o} \tag{10.55}$$

and an inductor of value

$$L_i = \frac{1}{\omega_0^2 . C_i}. \tag{10.56}$$

The extraction process can also be carried out on an admittance basis. First,

$$Y(s) = \frac{b_m s \left(s^2 + \omega_2^2\right) \left(s^2 + \omega_4^2\right) \left(s^2 + \omega_s^2\right) \dots}{a_n \left(s^2 + \omega_1^2\right) \left(s^2 + \omega_3^2\right) \left(s^2 + \omega_5^2\right) \dots}. \tag{10.57}$$

Now a pole at zero requires the extraction of a shunt inductor of value

$$L_0 = \left. \frac{1}{sY(s)} \right|_{s=0} \tag{10.58}$$
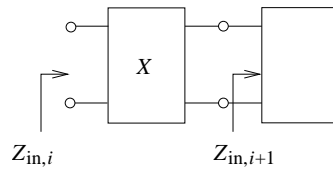
Figure 10-8 Extraction of a network X to reduce an impedance $Z_{\text{in},i}$ to a lower-order impedance $Z_{\text{in},i+1}$.

and a pole at infinity requires the extraction of a shunt capacitor of value

$$C_\infty = \left. \frac{Y(s)}{s} \right|_{s=\infty}. \tag{10.59}$$

A pole at a finite frequency requires the extraction of a shunt-series LC block (as shown in Figure 10-9(f)) with values

$$L_i = \left. \frac{s}{(s^2 + \omega_o^2)\, Y(s)} \right|_{s=j\omega_o} \tag{10.60}$$

and

$$C_i = \frac{1}{\omega_o^2 . L_i}. \tag{10.61}$$

The extractions are summarized in Figure 10-9. After the extraction of an element or a pair of elements an impedance, Z_{rem}, or admittance, Y_{rem}, remains which can be similarly simplified. For example, and referring to Figure 10-9(a), $Z(s) = 1/(sC) + Z_{\text{rem}}$. In general, the synthesis procedure involves extracting a network X from $Z_{\text{in},i}$, leaving a reduced-order impedance to $Z_{\text{in},i+1}$ (see Figure 10-8).

Many aspects of filter synthesis can seem abstract when presented in full generality. Consequently it is common to illustrate concepts using examples. Following this time-honored tradition, an example is now presented. Very often a general approach is developed from a specific example. However, this time a general approach was possible, and this was presented above.

| EXAMPLE 10. 2 | Element Extraction for a Third-Order Lowpass Filter |

A third-order maximally flat filter has the reflection coefficient

$$\Gamma_1(s) = \frac{s^3}{(s+1)(s^2 + s + 1)}. \tag{10.62}$$

Synthesize this filter as a doubly terminated network.

IMPEDANCE BASIS   ADMITTANCE BASIS

C_0

$Z_{\text{in}} = Z_C + Z_{\text{rem}}$ $Y_{\text{in}} = Y_L + Y_{\text{rem}}$

Z_{rem} L_0 Y_{rem}

$Z_C = \dfrac{1}{sC_0}$ $Y_L = \dfrac{1}{sL_0}$

(a)   (b)

L_∞

$Z_{\text{in}} = Z_L + Z_{\text{rem}}$ $Y_{\text{in}} = Y_C + Y_{\text{rem}}$

Z_{rem} C_∞ Y_{rem}

$Z_L = sL_\infty$ $Y_C = sC_\infty$

(c)   (d)

C_i

$Z_{\text{in}} = Z_{LC} + Z_{\text{rem}}$ $Y_{\text{in}} = Z_{LC} + Y_{\text{rem}}$

L_i Z_{rem} L_i Y_{rem}
          C_i

$Z_{LC} = \dfrac{sL_i}{s^2 L_i C_i + 1}$ $Y_{LC} = \dfrac{sC_i}{s^2 L_i C_i + 1}$

(e)   (f)
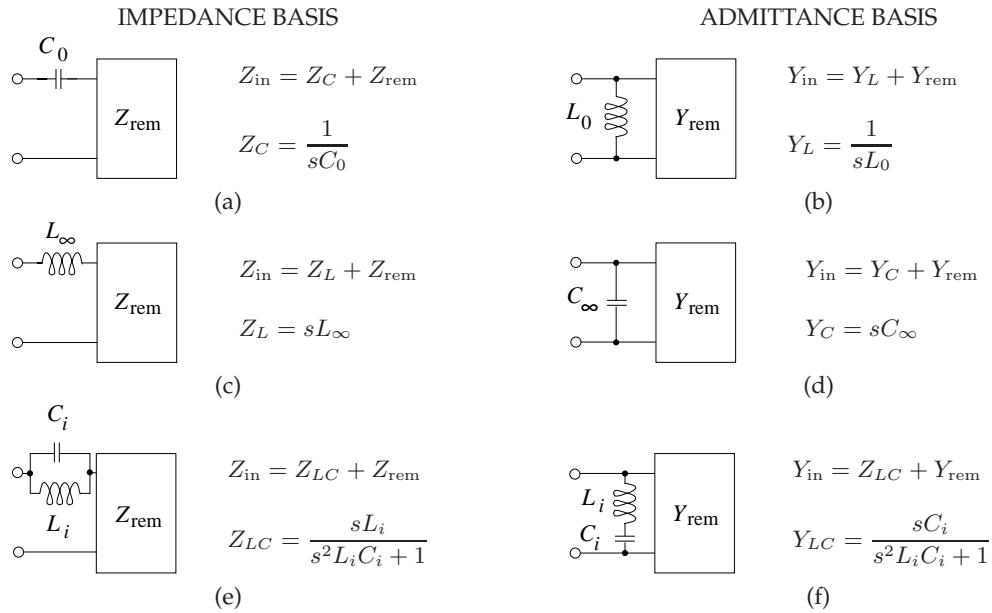
Figure 10-9 Synthesis of impedance and admittance functions. Starting with an impedance function $Z(s)$: (a) extraction of a series capacitor; (c) extraction of a series inductor; and (e) extraction of a series parallel LC block i. Starting with an admittance function $Y(s)$: (b) extraction of a shunt inductor; (d) extraction of a shunt capacitor; and (f) extraction of a shunt series LC block.

Solution:

The reflection coefficient function (Equation (10.62)) has all its poles located at infinity, so the corresponding network realization must be made of simple L or C elements. Hence, referring to Figure 10-2,

$$Z_{\text{in},1}(s) = \frac{1 + \Gamma_1(s)}{1 - \Gamma_1(s)} \quad \text{(in a 1 } \Omega \text{ system)} = \frac{2s^3 + 2s^2 + 2s + 1}{2s^2 + 2s + 1}. \qquad (10.63)$$

Note that the input impedance approaches infinity as the frequency goes to infinity, hence a series inductor must be extracted. The value of this inductor is

$$L_{\infty 1} = \left. \frac{Z_{\text{in},1}(s)}{s} \right|_{s=\infty} = 1 \text{ H}. \qquad (10.64)$$

So the filter is developed by extracting one element at a time. Following the extraction of each element, the stage impedance is left. The stage impedance function here is

$$Z_{\text{in},2}(s) = Z_{\text{in},1}(s) - sL_{\infty 1} = \frac{2s^3 + 2s^2 + 2s + 1}{2s^2 + 2s + 1} - sL_{\infty 1}$$

$$= \frac{2s^3 + 2s^2 + 2s + 1 - s\left(2s^2 + 2s + 1\right)}{2s^2 + 2s + 1} = \frac{s+1}{2s^2 + 2s + 1} \ .$$

Note that the stage impedance above, $Z_{\text{in},2}$, approaches zero as the frequency goes to infinity. However, the stage admittance function,

$$Y_{\text{in},2}(s) = \frac{1}{Z_{\text{in},2}(s)} = \frac{2s^2 + 2s + 1}{s + 1} \tag{10.65}$$

goes to infinity as the frequency approaches infinity and thus a shunt capacitor can be extracted:

$$Y_{\text{in},3}(s) = Y_{\text{in},2}(s) - sC_{\infty 2} = \frac{1}{s+1}, \tag{10.66}$$

where

$$C_{\infty 2} = 2 \text{ F}. \tag{10.67}$$

So sometimes it is more convenient to consider extraction of an admittance and sometimes it is better to consider extraction of an impedance.

By examining the remaining stage impedance, it is seen that a pole exists at infinity, implying that a series inductor, $L_{\infty 3}$, must be extracted. The value of this inductor comes from

$$Z_{\text{in},3} = \frac{1}{Y_{\text{in},3}} = s + 1 \ \Omega \tag{10.68}$$

and so the inductor value is

$$L_{\infty 3} = \left.\frac{s+1}{s}\right|_{S=\infty} = 1 \text{ H}. \tag{10.69}$$

The final step is to extract a load of value 1 as follows:

$$Z_{\text{in},4} = Z_{\text{in},3} - sL_{\infty 3} = 1 \ \Omega. \tag{10.70}$$

This example synthesized a doubly terminated network. The resulting network, called a ladder circuit, is shown in Figure 10-10. The left-most 1 Ω resistor is part of the source. This is inherently behind the conversion, in Equation (10.63), of the input reflection coefficient, $\Gamma_1(s)$, to an input impedance that used a 1 Ω reference impedance.

Note that this circuit has a dual form which consists of two shunt capacitors separated by a series inductor. The dual would be obtained if the admittance function was obtained from the reflection coefficient and then synthesized. Alternative network extraction techniques are presented in Scanlan and Levy [139, 140] and Matthaei et al. [141].

Lowpass Butterworth or Chebyshev filters can be realized by ladder networks of simple series and shunt combinations of individual inductors or capacitors. Such networks can be used, following a number of transformations, to realize highpass filters, and symmetrical bandpass and bandstop filters. Sometimes **asymmetric filters** with asymmetric responses are required, for example, to provide additional suppression of a large
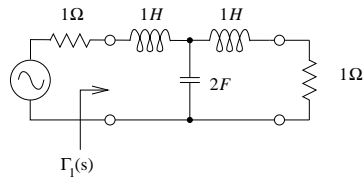
Figure 10-10 Synthesized maximally flat network with a third-order lowpass reflection response.

interferer (e.g., a radar), and then it is necessary to synthesize more complex impedance functions.

The input impedance functions of lumped-element circuits can be expressed as the ratio of two polynomials in s. An important property of the impedance function is that the order of the numerator and that of the denominator polynomial can differ by at most one [140]. If the orders differ by one, then a single inductor or capacitor can be extracted, however, the remaining impedance function may not be realizable. This indicates that a more complex LC (and possibly R) combination is required. To be able to extract arbitrarily complex circuits a long list of possible functions, such as those shown in Figure 10-9, are required. For most of the circuits of interest here the series and shunt LC combinations shown in Figure 10-9 are sufficient. The next example describes impedance function extraction that requires an LC combination.

EXAMPLE 10. 3   Element Extraction of an Impedance Function

Realize the impedance function

$$Z_w = \frac{4s^3 + 4s^2 + 2s + 2}{4s^2 + 2s + 1}.$$

Solution:

The order of the numerator is 1 greater than the order of the denominator and this indicates that a series inductor is perhaps present. The series inductance is

$$L_1 = \left.\frac{Z_w(s)}{(s)}\right|_{z=\infty} = 1 \text{ H.} \tag{10.71}$$

The remaining impedance is

$$Z_{\text{in},2} = Z_w - sL_1 = \frac{4s^3 + 4s^2 + 2s + 2}{4s^2 + 2s + 1} - s = \frac{2s^2 + s + 2}{4s^2 + 2s + 1}. \tag{10.72}$$

The numerator and denominator of $Z_{\text{in},2}$ have the same order and the impedance is not resistive. Therefore a simple L or C element cannot be used to reduce the complexity of the impedance function. Thus the extraction must backtrack and an initial series inductor is not the right choice.

Figure 10-9 shows several element combinations that can be used to reduce the complexity of an impedance function. Insight into which alternative to choose comes from factoring z_w, and real roots are required. So consider

$$Z_w = \frac{4s^3 + 4s^2 + 2s + 2}{4s^2 + 2s + 1} \tag{10.73}$$

$$= \frac{(2s^2 + 1)(2s + 2)}{4s^2 + 2s + 1} \tag{10.74}$$

and the denominator of z_w does not have real roots. Examining Figure 10-9, a ready fit to Z_w is not found. Instead consider the admittance function

$$Y_w = \frac{1}{Z_w} = \frac{4s^2 + 2s + 1}{(2s^2 + 1)(2s + 2)}. \tag{10.75}$$

So the reduction shown in Figure 10-9(f) looks like the right candidate. The general choice for the element is

$$y_x = \frac{as}{bs^2 + 1}. \tag{10.76}$$

Choosing $b = 2$ is the best choice for reducing complexity (since part of the factored denominator of Y_w occurs), so

$$Y_w = \frac{as}{2s^2 + 1} + \left(\frac{4s^2 + 2s + 1}{(2s^2 + 1)(2s + 2)} - \frac{as}{2s^2 + 1} \right) \tag{10.77}$$

$$= \frac{as}{2s^2 + 1} + \left(\frac{(4 - 2a)s^2 + (2 - 2a)s + 1}{(2s^2 + 1)(2s + 2)} \right). \tag{10.78}$$

Choose $a = 1$,

$$Y_w = \frac{s}{(2s^2 + 1)} + \frac{2s^2 + 1}{(2s^2 + 1)(2s + 2)} \tag{10.79}$$

$$= \frac{s}{2s^2 + 1} + \frac{1}{2s + 2} \tag{10.80}$$

$$= \frac{s}{2s^2 + 1} + Y_{\text{in},2}. \tag{10.81}$$

So $C_1 L_1 = b = 1$, $C_1 = a = 1$ F, $L_1 = 2$H, and

$$Y_{\text{in},2} = 1/(2s + 2) \tag{10.82}$$

or

$$Z_{\text{in},2} = \frac{1}{Y_{\text{in},2}} = 2s + 2. \tag{10.83}$$

The final network is as follows:

(a)

(b)
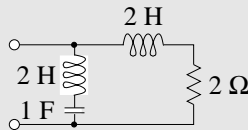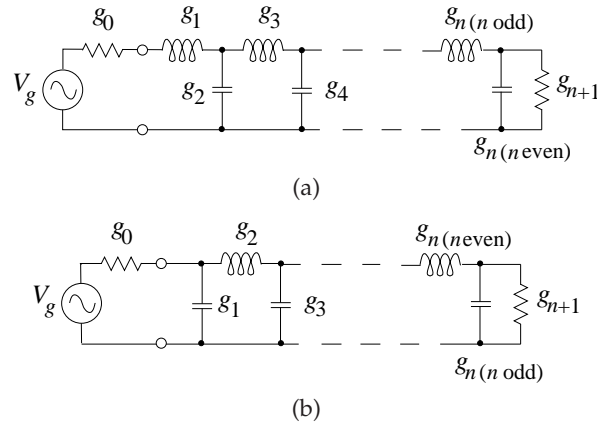
Figure 10-11 Chebyshev filter prototypes in the Cauer topology: (a) of type 1; and (b) of type 2. Here n is the order of the filter.

10.7 Butterworth and Chebyshev Filters

The nth-order lowpass filters constructed from the Butterworth and Chebyshev polynomials have the ladder circuit forms of either Figure 10-11(a) or Figure 10-11(b). Figure 10-11 uses several shorthand notations commonly used with filters. First, note that there are two prototype forms designated Type 1 and Type 2, and these are referred to as **duals** of each other. The two prototype forms have identical responses with the same numerical element values $g_1, \cdots, g_n$. Consider the Type 1 prototype of Figure 10-11(a). The right-most element is the resistive load, which is also known as the $(n+1)$th element. The next element to the left of this is either a shunt capacitor (of value g_n) if n is even, or a series inductor (of value g_n) if n is odd. So for the Type 1 prototype, the shunt capacitor next to the load does not exist if n is odd. The same interpretation applies to the circuit in Figure 10-11(b).

10.7.1 Butterworth Filter

A generalization of the example of the previous section leads to a formula for the element values of a ladder circuit implementing a Butterworth lowpass filter. For a maximally flat or Butterworth response the element values of the circuit in Figures 10-11(a) and 10-11(b) are

$$g_r = 2\sin\left\{(2r-1)\frac{\pi}{2n}\right\} \qquad r = 1, 2, 3, \ldots, n. \qquad (10.84)$$

Equation (10.84) provides the values of g_1 to g_n in Figure 10-11, with g_0 and g_{n+1} both being 1. Table 10-2 lists the coefficients of Butterworth lowpass prototype filters up to ninth order.

Table 10-2 Coefficients of the Butterworth lowpass prototype filter normalized to a radian corner frequency of 1 radian/s and a 1 Ω system impedance.

Order, n	2	3	4	5	6	7	8	9
g_1	1.4142	1	0.7654	0.6180	0.5176	0.4450	0.3902	0.3473
g_2	1.4142	2	1.8478	1.6180	1.4142	1.2470	1.1111	1
g_3	1	1	1.8478	2	1.9318	1.8019	1.6629	1.5321
g_4		1	0.7654	1.6180	1.9318	2	1.9615	1.8794
g_5			1	0.6180	1.4142	1.8019	1.9615	2
g_6				1	0.5176	1.2470	1.6629	1.8794
g_7					1	0.4450	1.1111	1.5321
g_8						1	0.3902	1
g_9							1	0.3473
g_{10}								1
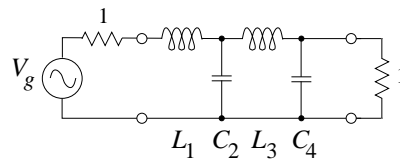
$$L_1 = 0.765369 \text{ H} \quad C_2 = 1.847759 \text{ F}$$
$$L_3 = 1.847759 \text{ H} \quad C_4 = 0.765369 \text{ F}$$

Figure 10-12 Fourth-order Butterworth lowpass filter prototype.

EXAMPLE 10. 4   Fourth-Order Butterworth Lowpass Filter

Derive the fourth-order Butterworth lowpass prototype of type 1.

Solution:

From Equation (10.84),

$$g_1 = 2\sin\left[\pi/(2\cdot 4)\right] = 0.765369 \text{ H} \qquad (10.85)$$
$$g_2 = 2\sin\left[3\pi/(2\cdot 4)\right] = 1.847759 \text{ F} \qquad (10.86)$$
$$g_3 = 2\sin\left[5\pi/(2\cdot 4)\right] = 1.847759 \text{ H} \qquad (10.87)$$
$$g_4 = 2\sin\left[7\pi/(2\cdot 4)\right] = 0.765369 \text{ F}. \qquad (10.88)$$

Thus the fourth-order Butterworth lowpass prototype circuit with a corner frequency of 1 radian/s is as shown in Figure 10-12.

10.7.2   Chebyshev Filter

For a Chebyshev response, the element values of the lowpass prototype shown in Figure 10-11 are found from the following recursive formula [141]:

$$g_0 = 1 \tag{10.89}$$

$$g_{n+1} = \left\{ \begin{array}{ll} 1, & n \text{ odd} \\ \tanh^2(\beta/4), & n \text{ even} \end{array} \right\} \tag{10.90}$$

$$g_1 = \frac{2a_1}{\gamma} \tag{10.91}$$

$$g_k = \frac{4a_{k-1}a_k}{b_{k-1}g_{k-1}}, \quad k = 2, \, 3, \dots, n \tag{10.92}$$

$$a_k = \sin\left[\frac{(2k-1)\pi}{2n}\right], \quad k = 1, \, 2, \dots, n \tag{10.93}$$

and

$$\gamma = \sinh\left(\frac{\beta}{2n}\right), \tag{10.94}$$

where

$$b_k = \gamma^2 + \sin^2\left(\frac{k\pi}{n}\right) \quad k = 1, \, 2, \dots, n \tag{10.95}$$

$$\beta = \ln\left[\coth\left(\frac{R_{\mathrm{dB}}}{17.37}\right)\right] \tag{10.96}$$

$$R_{\mathrm{dB}} = 10\log\left(\varepsilon^2 + 1\right), \tag{10.97}$$

where n is the order of the filter and ε is the ripple expressed in absolute terms. R_{dB} is the ripple expressed in decibels (the ripple is generally specified in decibels).

An interesting point to note here is that the source resistor, the value of which is given by g_0, and terminating resistor, the value of which is given by g_{n+1}, are only equal for odd-order filters. For an even-order Chebyshev filter the terminating resistor, g_{n+1}, will be different and a function of the filter ripple. Because it is generally desirable to have identical source and load impedances, Chebyshev filters are nearly always restricted to odd order. Thus the odd-order Chebyshev prototypes are as shown in Figure 10-13.

Also, for an odd-degree function (n is odd) there is a perfect match at DC,

$$|T(0)|^2 = 1, \tag{10.98}$$

while for an even-degree function a mismatch exists of value

$$|T(0)|^2 = \frac{4R_L}{(R_L + 1)^2} = \frac{1}{1 + \varepsilon^2} \tag{10.99}$$
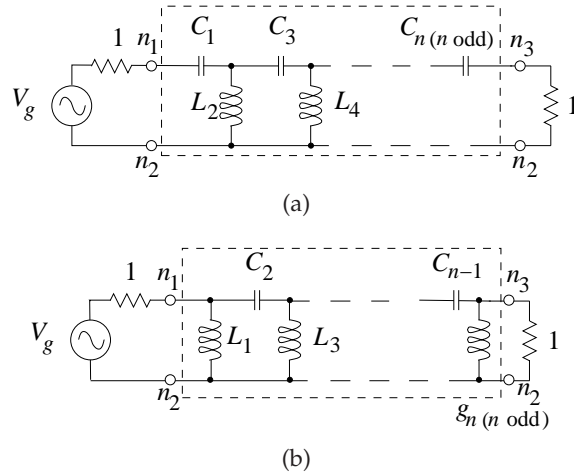
(a)

(b)
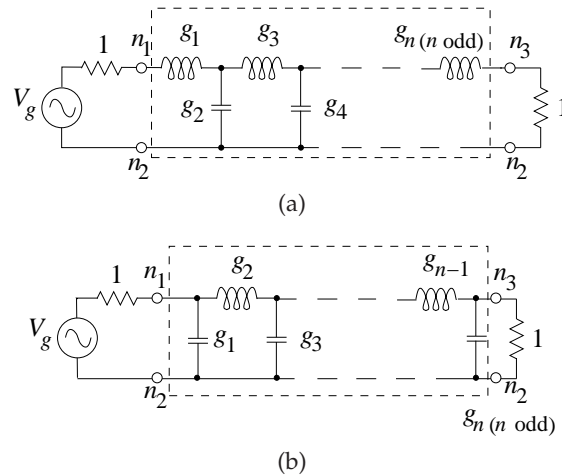
Figure 10-13 Odd-order Chebyshev lowpass filter prototypes in the Cauer topology: (a) of type 1; and (b) of type 2. Here n is the order of the filter.

so that

$$R_L = g_{n+1} = \left[\varepsilon + \sqrt{(1 + \varepsilon^2)} \right]^2 . \qquad (10.100)$$

Coefficients of several Chebyshev lowpass prototype filters with different levels of ripple and of odd orders up to ninth order are given in Table 10-3.

10.7.3   Summary

A Butterworth filter has a monotonic response without ripple but a relatively slow transition from the passband to the stopband. A Chebyshev filter has rapid transition but has ripple in either the stopband or passband. Butterworth and Chebyshev filters are special cases of **elliptical filters**, which are also called **Cauer filters**. An elliptical filter has equiripple (or equalized ripple) in both the stopband and the passband. The level of the ripple can be selected independently in each band. With zero ripple in the stopband, but ripple in the passband, an elliptical filter becomes a **Type I Chebyshev filter**. With zero ripple in the passband, but ripple in the stopband, an elliptical filter is known as a **Type II Chebyshev filter**. With no ripple in either band the elliptical filter becomes a Butterworth filter. With ripple in both the passband and stopband, the transition between the passband and stopband can be made more abrupt or alternatively the tolerance to component variations increased.

Another type of filter is the **Bessel filter** which has maximally flat group delay in the passband, which means that the phase response has maximum linearity across the passband. The **Legendre filter** (also known as the **optimum "L" filter** has a high transition rate from passband to stopband

Table 10-3 Coefficients of Chebyshev lowpass prototype filter normalized to a radian corner frequency of 1 radian/s and a 1 Ω system impedance.

Order	$n = 3$				
Ripple	0.01 dB	0.1 dB	0.2 dB	1.0 dB	3.0 dB
g_1	0.6291	1.0315	1.2275	2.0236	3.3487
g_2	0.9702	1.1474	1.1525	0.9941	0.7117
g_3	0.6291	1.0315	1.2275	2.0236	3.3487

Order	$n = 5$				
Ripple	0.01 dB	0.1 dB	0.2 dB	1.0 dB	3.0 dB
g_1	0.7563	1.1468	1.3394	2.1349	3.4817
g_2	1.3049	1.3712	1.3370	1.0911	0.7618
g_3	1.5573	1.9750	2.1660	3.0009	4.5381
g_4	1.3049	1.3712	1.3370	1.0911	0.7618
g_5	0.7563	1.1468	1.3394	2.1349	3.4817

Order	$n = 7$				
Ripple	0.01 dB	0.1 dB	0.2 dB	1.0 dB	3.0 dB
g_1	0.7969	1.1811	1.3722	2.1664	3.5182
g_2	1.3924	1.4228	1.3781	1.1116	0.7723
g_3	1.7481	2.0966	2.2756	3.0934	4.6386
g_4	1.6331	1.5733	1.5001	1.1736	0.8039
g_5	1.7481	2.0966	2.2756	3.0934	4.6386
g_6	1.3924	1.4228	1.3781	1.1116	0.7723
g_7	0.7969	1.1811	1.3722	2.1664	3.5182

Order	$n = 9$				
Ripple	0.01 dB	0.1 dB	0.2 dB	1.0 dB	3.0 dB
g_1	0.8144	1.1956	1.3860	2.1797	3.5340
g_2	1.4270	1.4425	1.3938	1.1192	0.7660
g_3	1.8043	2.1345	2.3093	3.1215	4.6692
g_4	1.7125	1.6167	1.5340	1.1897	0.8118
g_5	1.9057	2.2053	2.3728	3.1747	4.7272
g_6	1.7125	1.6167	1.5340	1.1897	0.8118
g_7	1.8043	2.1345	2.3093	3.1215	4.6692
g_8	1.4270	1.4425	1.3938	1.1192	0.7660
g_9	0.8144	1.1956	1.3860	2.1797	3.5340

for a given filter order, and also has a monotonic frequency response (i.e., without ripple). It is a compromise between the Butterworth filter with monotonic frequency response but slower transition and the Chebyshev filter, which has a faster transition but ripples in the frequency response.

More in-depth discussions of a large class of filters along with coefficient tables and coefficient formulas are available in Matthaei et al. [141], Hunter [142], Daniels [143], Lutovac et al. [144], and in most other books dedicated
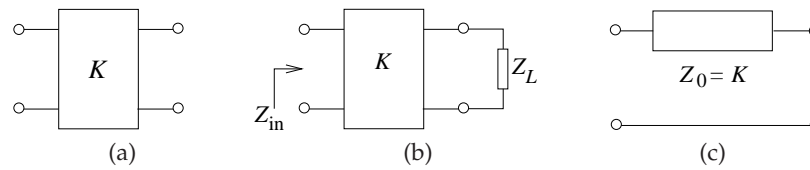
Figure 10-14 Impedance inverter (of impedance K): (a) represented as a two-port; (b) the two-port terminated in a load; and (c) the two-port implemented as a one-quarter wavelength long line of characteristic impedance Z_0.

solely to filters.

10.8 Impedance and Admittance Inverters

Inverters are two-port networks in the prototype networks of many RF and microwave filters. The input impedance of an inverter terminated in an impedance Z_L is $1/Z_L$. Impedance and admittance inverters are the same network, with the distinction being whether siemens or ohms are used to define them. An inverter is sometimes called a **unit element** (**UE**). At frequencies of a few hundred megahertz and below an inverter can be realized using operational and transconductance amplifiers. However, these are not readily available at RF and microwave frequencies. At RF, inverters are of particular interest, as they can be approximately realized by lumped elements, transmission line stubs, and coupled transmission lines. The transmission line realizations are better approximations; they have broader bandwidth than the lumped-element implementations. As a result, inverters are available in RF and microwave filter design and their main application is converting a series element into shunt elements and inverters. It is much easier to realize shunt elements in distributed circuits than series elements. Similar circuit transformations enable inductors to be replaced by capacitors and inverters.

RF filter design could be presented using only impedance inverters, however, admittance inverters are often used as design manipulations more often involve putting elements in shunt (which is preferred). The use of admittance simply unclutters the derivations. One or the other, and sometimes both, are used in various books on filter design and in journals and conference publications. Both will be used here.

The schematic representation of an impedance inverter is shown in Figure 10-14(a) where the distinguishing parameter is K which is in ohms. The constitutive property of the inverter is that the input impedance of the terminated impedance inverter in Figure 10-14(b) is

$$Z_{\text{in}} = \frac{K^2}{Z_L}, \tag{10.101}$$

and so the inverter both inverts the load impedance and scales it. Similarly,
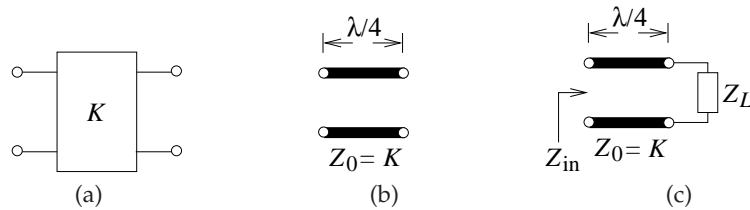
Figure 10-15   Inverter equivalence: (a) two-port impedance inverter (of impedance K): (b) a quarter-wave transmission line of characteristic impedance $Z_0 = 1/K$; a terminated one-quarter wavelength long transmission line of characteristic impedance Z_0.

if Port 1 was terminated in Z_L the input impedance at Port 2 is Z_{in} as defined above. Equation 10.101 does not fully define an inverter which is a two-port. An impedance inverter is fully defined by its $ABCD$ parameters:

$$\mathbf{T} = \begin{bmatrix} 0 & \jmath K \\ \jmath/K & 0 \end{bmatrix}, \tag{10.102}$$

and this is done in the next subsection where equivalence of ABCD parameters is used.

An impedance inverter has the value K, which is in ohms, and sometimes K is called the characteristic impedance of the inverter. Sometimes K is called just the impedance of the inverter.

An admittance inverter is exactly the same element, but now J is used and is called the characteristic admittance of the inverter, and sometimes just the admittance. They are related as $J = 1/K$. At low frequencies they can be implemented as transformers or using operational amplifiers. At RF and microwave frequencies inverters are of such great interest as they can be realized using combinations of transmission lines of reasonable length. Recall that in Section 4.8.4 on Page 218 it was shown that a $\lambda/4$ long line with a load has an input impedance that is the inverse of the load, normalized by the square of the characteristic impedance of the line. So an inverter can be realized at microwave frequencies using a one-quarter wavelength long transmission line (see Figure 10-15(b)). For the configuration shown in Figure 10-15(c),

$$Z_{\text{in}} = \frac{K^2}{Z_L}. \tag{10.103}$$

One of the principal advantages of inverters in developing networks is that they can be used to replace series inductors and capacitors by shunt elements, which are far more easily implemented. This will be seen in the next few sections.

10.8.1 Properties of an Impedance Inverter

At low frequencies it is necessary to use either transformers or active circuits based on operational amplifiers. In this section the impedance inverter can be treated as an abstraction. A two-port lossless, reciprocal, and frequency-independent impedance inverter has the $ABCD$ matrix

$$\mathbf{T} = \left[\begin{array}{cc} 0 & \jmath K \\ \jmath/K & 0 \end{array} \right],$$

(10.104)

where K is called the characteristic impedance of the inverter. With a load impedance, Z_L (at Port 2), the input impedance (at Port 1) is

$$Z_{\text{in}}(s) = \frac{A.Z_L + B}{C.Z_L + D} = \frac{\jmath K}{(\jmath/K)\,Z_L} = \frac{K^2}{Z_L}.$$

(10.105)

Thus the impedance at the input port is proportional to the inverse of the load impedance. (Note that K is in ohms.)

Now the $ABCD$ matrix of the transmission line of Figure 10-14(c) is

$$\left[\begin{array}{cc} \cos\theta & \jmath Z_0 \sin\theta \\ \frac{\jmath}{Z_0}\sin\theta & \cos\theta \end{array} \right],$$

(10.106)

which is identical to Equation (10.104) when the electrical length is $\theta = \pi/2$ (i.e., when the line is $\lambda/4$ long). The inverter is shown in Figure 10-14(a) as a two-port and implemented as a $\lambda/4$ long line in Figure 10-14(c). The bandwidth over which the line realizes an impedance inverter is limited, however, as it is an ideal inverter only at the frequency at which it is $\lambda/4$ long and when it is $(n\lambda/2 + \lambda/4)$ long.

10.8.2 Replacement of a Series Inductor by a Shunt Capacitor

A series inductor can be replaced by a shunt capacitor surrounded by a pair of inverters followed by a negative unity transformer (i.e., an inverter with $K = 1$). This equivalence is shown in Figure 10-16 and will now be shown mathematically.

From Table 6-1, the $ABCD$ matrix of the series inductor shown in Figure 10-16(a) (which has an impedance of sL) is

$$\mathbf{T}_L = \left[\begin{array}{cc} 1 & sL \\ 0 & 1 \end{array} \right]$$

(10.107)

and the $ABCD$ matrix of a shunt capacitor (which has an admittance of sC) is, from Table 6-1,

$$\mathbf{T}_1 = \left[\begin{array}{cc} 1 & 0 \\ sC & 1 \end{array} \right].$$

(10.108)

(a)   (b)   (c)
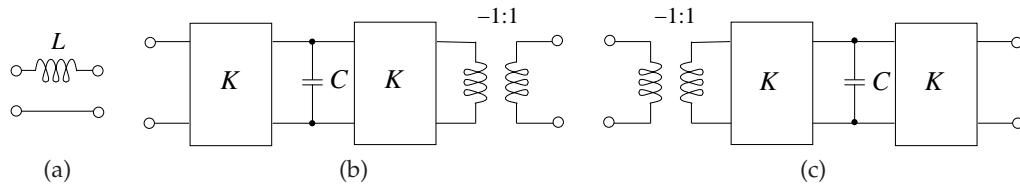
Figure 10-16 Equivalent realizations of a series inductor: (a) as a two-port; (b) its realization using a capacitor, inverters of characteristic impedance K, and a negative unity transformer; and (c) an alternative realization. $C = L/K^2$.

The $ABCD$ matrix of an inverter with $K = 1$ (generally units are dropped and SI is assumed) is

$$\mathbf{T}_2 = \begin{bmatrix} 0 & \jmath K \\ \jmath/K & 0 \end{bmatrix},$$ (10.109)

and finally, the $ABCD$ matrix of a negative unity transformer, $N = -1$, is, from Table 6-1,

$$\mathbf{T}_3 = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.$$ (10.110)

The $ABCD$ matrix of the cascade shown in Figure 10-16(b) is

$$
\begin{aligned}
\mathbf{T}_C &= \mathbf{T}_2 \mathbf{T}_1 \mathbf{T}_2 \mathbf{T}_3 \\
&= \begin{bmatrix} 0 & \jmath K \\ \jmath/K & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ sC & 1 \end{bmatrix} \begin{bmatrix} 0 & \jmath K \\ \jmath/K & 0 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \\
&= \begin{bmatrix} 0 & \jmath K \\ \jmath/K & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ sC & 1 \end{bmatrix} \begin{bmatrix} 0 & -\jmath K \\ -\jmath/K & 0 \end{bmatrix} \\
&= \begin{bmatrix} \jmath sCK & \jmath K \\ \jmath/K & 0 \end{bmatrix} \begin{bmatrix} 0 & -\jmath K \\ -\jmath/K & 0 \end{bmatrix} \\
&= \begin{bmatrix} 1 & sCK^2 \\ 0 & 1 \end{bmatrix}.
\end{aligned}
$$ (10.111)

Thus $T_C = T_L$ if $L = CK^2$ (compare Equations (10.107) and (10.111)). A series inductor can be replaced by a shunt capacitor with an inverter before and after it and with a negative unity transformer. The unity transformer may also be placed at the first port, as in Figure 10-16(c). Thus the two-ports shown in Figure 10-16 are all electrically identical, with the limitation being the frequency range over which the inverter can be realized. An interesting and important observation is that as a result of the characteristic impedance of the inverter (e.g., 50 Ω), a small shunt capacitor can be used to realize a large series inductance value.
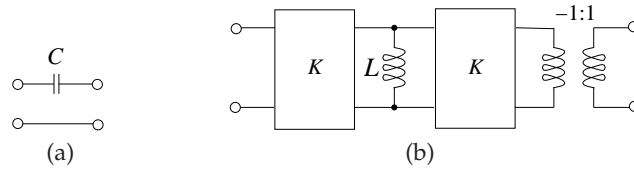
Figure 10-17   A series capacitor: (a) as a two-port; (b) its realization using a shunt inductor, inverters and negative unity transformer. $L = CK^2$

| EXAMPLE 10. 5 | Inductor Synthesis Using an Inverter |

Consider the network of Figure 10-16(c) with inverters having a characteristic impedance of 50 Ω. What value of inductance is realized using a 10 pF capacitor?

Solution:

$K = 50$, so $L = CK^2 = 10^{-11} \times 2500 = 25$ nH.

10.8.3   Replacement of a Series Capacitor by a Shunt Inductor

Analogous to the replacement demonstrated in the previous section, a series capacitor can be replaced by a shunt inductor plus inverters and a negative transformer (see Figure 10-17). The $ABCD$ parameters of the series capacitor in Figure 10-17(a) are

$$\mathbf{T} = \begin{bmatrix} 1 & 1/sC \\ 0 & 1 \end{bmatrix}, \tag{10.112}$$

and here it is shown that the cascade in Figure 10-17(b) has the same $ABCD$ parameters. Choosing the value of the shunt inductor in Figure 10-17(b) as C and using unity impedance inverters, the cascade has the $ABCD$ parameters

$$
\begin{aligned}
\mathbf{T} &= \begin{bmatrix} 0 & \jmath K \\ \jmath/K & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1/sL & 1 \end{bmatrix} \begin{bmatrix} 0 & \jmath K \\ \jmath/K & 0 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \\[2mm]
&= \begin{bmatrix} 0 & \jmath K \\ \jmath/K & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1/sL & 1 \end{bmatrix} \begin{bmatrix} 0 & -\jmath K \\ -\jmath/K & 0 \end{bmatrix} \\[2mm]
&= \begin{bmatrix} \jmath K/sL & \jmath K \\ \jmath/K & 0 \end{bmatrix} \begin{bmatrix} 0 & -\jmath K \\ -\jmath/K & 0 \end{bmatrix} \\[2mm]
&= \begin{bmatrix} 1 & K^2/sL \\ 0 & 1 \end{bmatrix}. \tag{10.113}
\end{aligned}
$$

So the series capacitor, C, can be realized using a shunt inductor, L, inverters, and a negative unity transformer, and $C = L/K^2$. It is unlikely
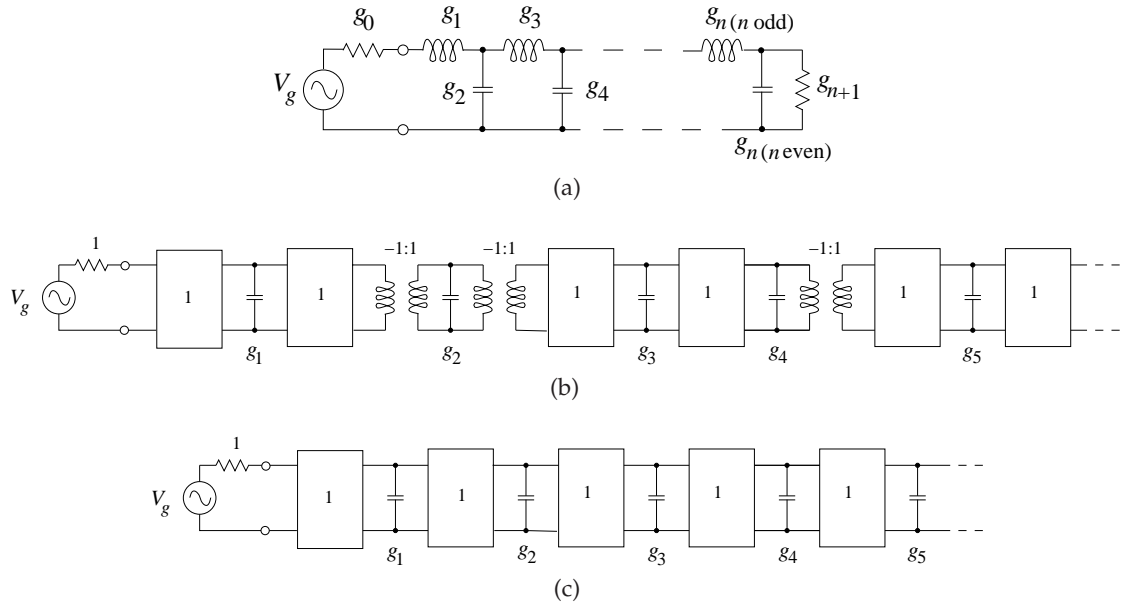
Figure 10-18   Ladder prototype filters using impedance inverters: (a) lumped-element prototype; (b) first stage in transformation using inverters; and (c) final stage.

that this would be desirable, as much better lower loss capacitors can be realized at RF than inductors.

10.8.4   Ladder Prototype with Impedance Inverters

The transformations discussed in the previous two sections can be used to advantage in simplifying filters. In this section the transformations are applied to the prototype of the lowpass ladder filter shown in Figure 10-18(a). The inductors in the ladder circuit are a particular problem as they have considerable resistance at radio and microwave frequencies and limit the frequency selectivity that can be obtained from a filter. Using the concept of inverters, the series inductors can be replaced by a circuit with capacitors, inverters, and transformers, as shown in Figure 10-18(b). This simplifies further to the realization shown in Figure 10-18(c), as the negative unity transformers only affect the phase of the transmission coefficient. So a lowpass ladder filter can be realized using just capacitors and inverters.

10.8.5   Lumped-Element Realization of an Inverter

The admittance inverter is functionally the same as the impedance inverter (see Figure 10-14(a)) and the schematic is the same (see Figure 10-19(a)). As will be shown, this can be realized using frequency-invariant lossless
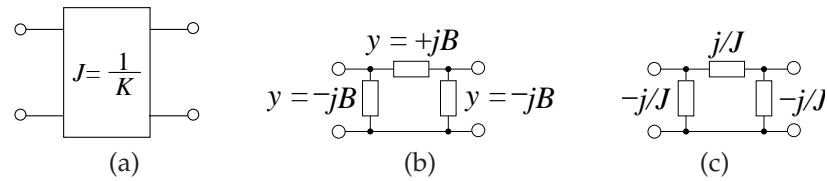
Figure 10-19 Admittance inverter: (a) as a two-port; (b) generalized inverter realized using lumped elements (identical if $B = -J$); and (c) lumped equivalent circuit of the admittance inverter of (a) (the element values in (c) are impedances).

elements (i.e., elements whose reactance or susceptance do not vary with frequency) using the network of Figure 10-19(b). J is generally used with admittance inverters and K with impedance inverters. If not specified by the context, the inverter (with value specified by a number) defaults to being an impedance inverter. Alternatively units can be used to indicate which type of inverter is being used. The function of the inverter is the same in any case; both can be realized by one-quarter wavelength long lines, for example. For the remainder of this chapter it will be more convenient to use the admittance inverter, as many calculations will be in terms of admittances since it is much easier to realize shunt elements in distributed form (e.g., using stubs) than to realize series elements.

Now it will be shown that the lumped-element network of Figure 10-19(b) realizes an inverter. To do this we must show that the inverter and the lumped-element network have the same two-port parameters. First, the $ABCD$ matrix of an inverter of characteristic admittance J is

$$\mathbf{T}_J = \begin{bmatrix} 0 & \jmath/J \\ \jmath J & 0 \end{bmatrix}. \tag{10.114}$$

Referring to Table 6-1, the circuit of Figure 10-19(b) has the ABCD matrix

$$
\begin{aligned}
T &= \begin{bmatrix} 1 & 0 \\ -\jmath B & 1 \end{bmatrix} \begin{bmatrix} 1 & 1/(\jmath B) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\jmath B & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 1/(\jmath B) \\ -\jmath B & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\jmath B & 1 \end{bmatrix} = \begin{bmatrix} 0 & -j/B \\ -\jmath B & 0 \end{bmatrix},
\end{aligned} \tag{10.115}
$$

where B is the susceptance of the frequency-invariant elements. Equation (10.115) is identical to Equation (10.114) if $B = -J$. More practical equivalents of the circuit of Figure 10-19(b) can be derived, as shown later.

For completeness, the lumped-element equivalent of the impedance inverter is shown in Figure 10-20 (derived from Figure 10-19 with $J = 1/K$).
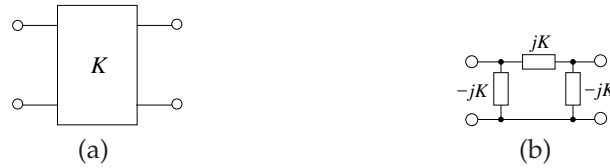
Figure 10-20 Impedance inverter: (a) as a two-port; and (b) its lumped equivalent circuit (the element values in (b) are impedances).
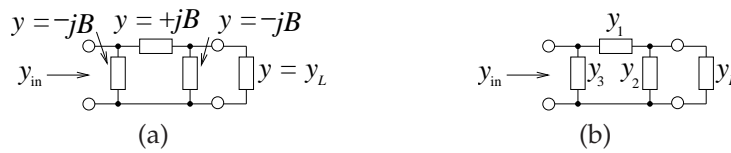
Figure 10-21 Terminated lumped-element admittance inverter.

EXAMPLE 10. 6   Lumped Inverter Analysis

Demonstrate that the lumped-element inverter model in Figure 10-19(b) is an admittance inverter.

Solution:

Terminating the network in Figure 10-19(b) results in the network shown in Figure 10-21(a). This is relabeled for convenience in Figure 10-21(b) where the elements are admittances. Then

$$y_{\text{in}} = y_3 // (y_1 \$ (y_2 + y_L)) , \tag{10.116}$$

where $//$ indicates "in parallel with" and $\$$ indicates "in series with." These are common shorthand notations in circuit calculations. Continuing on from Equation (10.116),

$$y_{\text{in}} = y_3 + \left[\frac{1}{y_1} + \left(\frac{1}{y_2 + y_L} \right) \right]^{-1} = y_3 + \left[\frac{y_2 + y_L + y_1}{y_1 (y_2 + y_L)} \right]^{-1} . \tag{10.117}$$

Substituting $y_1 = \jmath B$, $y_2 = y_3 = -\jmath B$, this becomes

$$y_{\text{in}} = -\jmath B + \frac{\jmath B (y_L - \jmath B)}{y_L} = \frac{-\jmath B y_L + \jmath B y_L + B^2}{y_L} = \frac{B^2}{y_L} . \tag{10.118}$$

Thus the lumped-element of Figure 10-19(b) is an admittance inverter of value B (in Siemens).

10.8.6 Narrowband Realization of an Inverter Using Transmission Line Stubs

In this section it will be shown that an impedance inverter can be implemented using short- and open-circuits stubs. The match is good over

Figure 10-22 Narrowband inverter equivalents at frequency f_0: (a) impedance inverter with characteristic impedance K; (b) lumped-element equivalent network; (c) inverter realized by short- and open-circuited stubs.

a narrow band centered at frequency f_0. An impedance inverter is shown in Figure 10-22(a) and its equivalent lumped-element network in Figure 10-22(b). A stub-based implementation is shown in Figure 10-22(c), where there are short- and open-circuited stubs of characteristic impedance Z_0. The input impedance of the stubs is shown at the inputs of the stubs. The stubs have an electrical length θ at f_0, and the stubs are one-quarter wavelength long (i.e., is resonant) at what is called the commensurate frequency, f_r.

Now it will be shown that the network of Figure 10-22(c) is a good representation of the inverter at f_0. This is done by matching $ABCD$ parameters. The $ABCD$ parameters of an inverter are

$$T = \begin{bmatrix} 0 & jK \\ j/K & 0 \end{bmatrix} \tag{10.119}$$

and, at frequency f_0, the ABCD parameters of the stub circuit of Figure 10-22(c) are

$$T = \begin{bmatrix} 1 & 0 \\ -1/[jZ_0 \tan(\theta)] & 1 \end{bmatrix} \begin{bmatrix} 1 & jZ_0 \tan(\theta) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1/[jZ_0 \tan(\theta)] & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & jZ_0 \tan(\theta) \\ -1/[jZ_0 \tan(\theta)] & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1/[jZ_0 \tan(\theta)] & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & jZ_0 \tan(\theta) \\ j/[Z_0 \tan(\theta)] & 0 \end{bmatrix}. \tag{10.120}$$

Thus, equating Equations (10.119) and (10.120), the stub network is a good representation of the inverter if

$$K = Z_0 \tan(\theta), \tag{10.121}$$

and so the required characteristic impedance of each stub at frequency f_0 is

$$Z_0 = \frac{K}{\tan(\theta)} = \frac{K}{\tan\left(\frac{\pi}{2} \frac{f_0}{f_r}\right)}. \tag{10.122}$$

Special Case $f_r = 2f_0$

The most commonly used designs choose the stub resonant frequency, the commensurate frequency, to be twice that of the center frequency of the design, f_0. So with $f_r = 2f_0$:

$$
\begin{aligned}
Z_0 &= \frac{K}{\tan\left(\frac{\pi}{2}\frac{f_0}{2f_0}\right)} \\
&= \frac{K}{\tan \pi/4} \\
&= K.
\end{aligned}
\tag{10.123}
$$

10.9 Filter Transformations

So far the discussion has centered around lowpass filters. Filter design technology has developed so that the design of lowpass filters is the essential first step and these are used as prototypes to derive filters with other characteristics. The three most important transformations are listed here:

1. Impedance scaling: A lowpass filter prototype is referenced to a standard impedance. Usually $1~\Omega$ is used, so the reference source and load resistances are also $1~\Omega$. To reference to a higher or lower impedance, scaling of the impedance of all elements of the filter is required.

2. Cutoff frequency scaling: The cutoff frequency of a lowpass filter prototype is normalized to 1 radian/s. To reference to another frequency, the values of the elements must be altered so that they have the same impedance at the scaled frequency.

3. Filter type transformation: These transformations enable the circuit of a lowpass filter to be converted to a circuit with a bandpass, bandstop, or highpass response. The concept is that the response at DC is to be replicated at infinite frequency for the highpass (so capacitors become inductors, etc.); to be replicated at the center of the passband for a bandpass filter (so capacitors become a shunt LC circuit); and to be inverted at the center of the bandstop filter (so capacitors become a series LC circuit).

The transformations can be performed in any order. The key point is that the complexity of the lowpass circuit design is minimal since the number of elements does not grow until the filter type transformation is performed, so this step is often done last.

$$L \qquad\qquad\qquad\qquad L \times Z_0$$

(a)

$$C \qquad\qquad\qquad\qquad C / Z_0$$

(b)

$$K \qquad\qquad\qquad\qquad K\, Z_0$$

(c)

Figure 10-23 Impedance transformations: (a) for an inductor; (b) for a capacitor; and (c) for an inverter.

10.9.1 Impedance Transformation

The lowpass prototypes discussed up to now have been referred to a 1 Ω system. The system impedance can be changed to any level by simply scaling the impedances of all the circuit elements in a filter by the same amount, as shown in Figure 10-23. The impedance transformation follows the same procedure for all filter types (the procedure is the same for lowpass, highpass, bandpass, and bandstop filters, for example). It should also be noted that impedance scaling of a transmission line is simply multiplication of the line's characteristic impedance by the scale factor.

EXAMPLE 10. 7   Lowpass Filter Design

Consider the lowpass filter, with an inverter, shown in Figure 10-24(a). This filter is referenced to 1 Ω, as the source and load impedances are both 1 Ω. Redesign the filter so that the same frequency response is obtained with 50 Ω source and load impedances.

Solution:

It is necessary to impedance transform from 1 Ω to 50 Ω. The resulting filter is shown in Figure 10-24(b). Each element has an impedance (resistance or reactance) that is 50 times larger than it had in the prototype.

10.9.2 Frequency Transformation: Lowpass

Frequency transformations differ depending on the type of the filter. So it is normal to frequency transform the lowpass prototype before converting the prototype to another form (such as bandpass). The lowpass prototypes normally have a band-edge or cutoff frequency at an angular frequency of

Figure 10-24 Impedance transformation of an example filter: (a) filter normalized to 1 Ω; and (b) impedance transformed to a reference of 50 Ω.

Figure 10-25 Frequency transformation of a lowpass filter from, (a) one normalized to a corner frequency of 1 rad/s, to (b) one with a radian corner frequency of ω_c.

unity,[2] that is, 1 rad/s. The band edge frequency can be transformed from unity to an arbitrary frequency ω_c, as shown in Figure 10-25, by scaling the reactive elements, as shown in Figure 10-26, so that they have the same impedance at the transformed frequency as they did at the original frequency. The inverter is unchanged, as are the source and load resistances, since these are frequency independent.

---

[2] In this book (and in most treatments) the band-edge frequency is normalized to 1 rad/s. Some books and papers use a 1 Hz normalization, so be wary.

Figure 10-26 Frequency transformations for (a) an inductor; and (b) a capacitor. The impedance of the new (scaled) element at the new (scaled) frequency is the same as it was at the original frequency.

Figure 10-27 Lowpass to highpass transformation.

10.9.3 Lowpass to Highpass Transformation

The transformation of a lowpass filter prototype to a highpass filter is shown diagrammatically in Figure 10-27. Mathematically the transformation is

$$|T(\omega)|^2 = \frac{1}{1 + \varepsilon^2 |K(\omega)|^2} \tag{10.124}$$

$$\omega \rightarrow -\frac{1}{\omega} \tag{10.125}$$

$$\left|T\left(-\frac{1}{\omega}\right)\right|^2 = \frac{1}{1 + \varepsilon^2 \left|K\left(-\frac{1}{\omega}\right)\right|^2} \tag{10.126}$$

so that reactive elements are transformed as shown in Figure 10-30. Here inductors are transformed into capacitors and capacitors into inductors.

For example, if the odd-order lowpass filters are of the types shown in Figure 10-13 then the transformation to highpass is shown in Figure 10-28.

10.9.4 Lowpass to Bandpass Transformation

Understanding the transformation of the lowpass filter into its corresponding bandpass forms requires that the lowpass filter be considered in both its

Figure 10-28 Odd-order Chebyshev highpass filter prototypes in the Cauer topology: (a) of type 1; and (b) of type 2. Here n is the order of the filter.

positive and negative frequency response, as shown in Figure 10-29(a).

This response is shifted in frequency to obtain the bandpass response shown in Figure 10-29(b). Mathematically the radian frequency, ω, in the response function is replaced by its bandpass form:

$$\omega \rightarrow \left[\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega} \right]. \tag{10.127}$$

This separately maps the -1 and $+1$ band-edge radian frequencies of the lowpass response to the bandpass frequencies ω_1 and ω_2:

$$-1 \quad \rightarrow \quad \left[\frac{\omega_1}{\omega_o} - \frac{\omega_o}{\omega_1} \right] \tag{10.128}$$

$$+1 \quad \rightarrow \quad \left[\frac{\omega_2}{\omega_o} - \frac{\omega_o}{\omega_2} \right]. \tag{10.129}$$

Solving the above equations simultaneously yields the center frequency ω_0 and the band-edge frequencies ω_1 and ω_2 with

$$\omega_o = \sqrt{\omega_1 \omega_2} \tag{10.130}$$

and the so-called transformation constant

$$\alpha = \frac{\omega_o}{\omega_2 - \omega_1}. \tag{10.131}$$

The resulting element conversions are given in Figure 10-30.

As an example, a lumped-element Type 2 Cauer bandpass filter is shown in Figure 10-31. The shunt LC combination and the series LC combinations

Figure 10-29 Frequency responses in lowpass to bandpass transformation: (a) two-sided lowpass filter response; and (b) bandpass filter approximation.

are resonators resonant at the center frequency of the filter. Here the filter is normalized to a Z_0 source and load impedances. As these are the same, this filter topology only applies for an odd-order filter. Combining transformations, the element values of a lumped bandpass filter with center radian frequency $\omega_0 = 2\pi f_0$ and radian bandwidth $\omega_{\mathrm{BW}} = 2\pi(f_2 - f_1)$ are as follows:

$$
C_r = \begin{cases} \dfrac{g_r}{\omega_{\mathrm{BW}} Z_0} & r = \text{odd} \\[4mm] \dfrac{\omega_{\mathrm{BW}}}{\omega_0^2 g_r Z_0} & r = \text{even} \end{cases}
\tag{10.132}
$$

$$
L_r = \begin{cases} \dfrac{\omega_{\mathrm{BW}} Z_0}{\omega_0^2 g_r} & r = \text{odd} \\[4mm] \dfrac{g_r Z_0}{\omega_{\mathrm{BW}}} & r = \text{even} \end{cases}.
\tag{10.133}
$$

10.9.5 Lowpass to Bandstop Transformation

Again, consider both the positive and negative frequency responses of the lowpass filter prototype, as shown in Figure 10-32(a). This response is

LOWPASS PROTOTYPE TO LOWPASS TRANSFORMATION		
LOWPASS PROTOTYPE	LOWPASS ELEMENT	REACTANCE TRANSFORMATION ω_0 = corner frequency
C_0	C_1	$C_1 = C_0/\omega_0$
L_0	L_1	$L_1 = L_0/\omega_0$
LOWPASS PROTOTYPE TO HIGHPASS TRANSFORMATION		
LOWPASS PROTOTYPE	HIGHPASS ELEMENT	REACTANCE TRANSFORMATION ω_0 = corner frequency
C_0	L_1	$L_1 = 1/(\omega_0 C_0)$
L_0	C_1	$C_1 = 1/(\omega_0 L_0)$
LOWPASS PROTOTYPE TO BANDPASS TRANSFORMATION		
LOWPASS PROTOTYPE	BANDPASS ELEMENT	REACTANCE TRANSFORMATION ω_0 = center of passband
C_0	L_1 C_1	$C_1 = \alpha C_0/\omega_0 \quad L_1 = 1/(\alpha C_0 \omega_0)$
L_0	$L_1 \quad C_1$	$L_1 = \alpha L_0/\omega_0 \;\; C_1 = 1/(\alpha L_0 \omega_0)$
LOWPASS PROTOTYPE TO BANDSTOP TRANSFORMATION		
LOWPASS PROTOTYPE	BANDSTOP ELEMENT	REACTANCE TRANSFORMATION ω_0 = center of stopband
C_0	$L_1 \quad C_1$	$L_1 = \alpha/(C_0 \omega_0) \quad C_1 = C_0/(\omega_0 \alpha)$
L_0	L_1 C_1	$L_1 = L_0/(\alpha \omega_0) \quad C_1 = \alpha/(L_0 \omega_0)$

Figure 10-30   Transformations of the elements of a prototype lowpass filter to obtain specific filter types. The corner frequency of the lowpass prototype is 1 radian/s. In the transformations to bandpass and bandstop filters, $\omega_0 = 1/\sqrt{L_1 C_1} = \sqrt{\omega_1 \omega_2}$; ω_1 and ω_2 are the band-edge frequencies; and α is the transformation constant, $\alpha = \omega_0/(\omega_2 - \omega_1)$. The corner frequency of the lowpass prototype is 1 radian/s.

Figure 10-31 Lumped-element odd-order (nth-order) Chebyshev bandpass filter prototypes in the type II Cauer topology.

(a)

(b)

Figure 10-32 Frequency responses in lowpass to bandstop transformation: (a) two-sided lowpass filter response; and (b) bandstop filter approximation.

shifted in frequency to obtain the bandstop response shown in Figure 10-32(b). Mathematically the frequency, ω, in the response function is replaced by its bandstop form:

$$ \omega \rightarrow \left[\alpha \left(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega} \right) \right]^{-1}. \tag{10.134} $$

The center frequency (corresponding to DC of the response of the lowpass prototype) is

$$ \omega_o = \sqrt{\omega_1 \omega_2} \tag{10.135} $$

Figure 10-33 Lumped-element odd-order (nth-order) Chebyshev bandpass filter prototypes in the type II Cauer topology.

and the transformation constant is

$$\alpha = \frac{\omega_o}{\omega_2 - \omega_1}, \tag{10.136}$$

where ω_1 and ω_2 are the band-edge frequencies. The resulting element conversions are given in Figure 10-30.

Combining transformations, the element values of a lumped bandstop filter with center radian frequency $\omega_0 = 2\pi f_0$ and radian bandwidth $\omega_{\mathrm{BW}} = 2\pi(f_2 - f_1)$ are as follows:

$$C_r = \begin{cases} \dfrac{g_r \omega_{\mathrm{BW}}}{\omega_0^2 Z_0} & r = \text{odd} \\[4mm] \dfrac{1}{\omega_{\mathrm{BW}} g_r Z_0} & r = \text{even} \end{cases} \tag{10.137}$$

$$L_r = \begin{cases} \dfrac{Z_0}{\omega_{\mathrm{BW}} g_r} & r = \text{odd} \\[4mm] \dfrac{g_r \omega_{\mathrm{BW}} Z_0}{\omega_0^2} & r = \text{even} \end{cases}. \tag{10.138}$$

Similarly a lumped-element type II Cauer bandstop filter is shown in Figure 10-33. The parallel LC combination and the series LC combinations are resonators resonant at the center frequency of the filter. The parallel LC resonator is an open circuit at the center frequency of the stopband and the series LC resonators are short circuits.

10.9.6 *Transformed Ladder Prototypes*

Combining the filter type transformations, and with appropriate use of inverters, the original lowpass prototype ladder filter and its various filter type transforms are shown in Figure 10-34.

Figure 10-34 Ladder prototype filters: (a) lowpass prototype ladder filter; (b) highpass prototype ladder filter; (c) bandstop prototype ladder filter; and (d) bandpass prototype ladder filter.

10.10 Cascaded Line Realization of Filters

In this section, filters are presented that use cascaded sections of transmission line—realizing series inductors and shunt capacitors. For example, a short (less than one-quarter wavelength long) line of relatively high-impedance line behaves predominantly as series inductances. Also, a very short (much less than one-quarter wavelength long) length of relatively low-impedance line acts predominantly as a shunt capacitance. So a Pi network of lumped elements can be realized with alternate sections of low- and high-impedance microstrip lines. Such an inductive line is shown in Figure 10-35(a), which has the two equivalent circuits shown in Figures 10-35(b) and 10-35(c). Basic transmission line theory gives the input reactance

Figure 10-35 Inductive length of line with adjacent capacitive lines: (a) microstrip form; (b) lumped equivalent circuit; and (c) lumped-distributed equivalent circuit.

Figure 10-36 Capacitive length of line with adjacent inductive lines: (a) microstrip; and (b) lumped equivalent (the left-most and right-most series inductors come from the high-impedance lines).

of the line of length, ℓ (with a low-impedance load),

$$X_L = Z_0 \sin\left(2\pi\ell/\lambda_g\right),\qquad (10.139)$$

so that the length of this predominantly inductive line is

$$\ell = \frac{\lambda_g}{2\pi} \sin^{-1}\left(\frac{\omega L}{Z_0}\right).\qquad (10.140)$$

Previously it was shown that a short length of line having a relatively low characteristic impedance yields a capacitive element and this is shown, together with its equivalent circuit, in Figure 10-36. The predominating shunt capacitance is determined by first considering the susceptance

$$B = \frac{1}{Z_0} \sin\left(\frac{2\pi\ell}{\lambda_g}\right)\qquad (10.141)$$

so that

$$\ell = \frac{\lambda_g}{2\pi} \sin^{-1}\left(\omega C Z_0\right).\qquad (10.142)$$

Thus a cascade of low-impedance and high-impedance lines can therefore realize (approximately) an LC ladder network.

It is tempting to take filters synthesized in lumped-element form and realize them in the above manner with transmission lines. The series

element realization concept presented in this section could be extended using shorted and open stubs to better realize shunt elements. The problem with this approach is that the resulting filters are narrowband and the response outside the desired operating range is unpredictable. It is far better to employ the Richards' transformation, considered in the next section, which is a much broader bandwidth technique for realizing filters in distributed form.

10.11   Richards' Transformation

10.11.1   Richards' Transformation and Transmission Lines

The Richards' transformation is a remarkable scheme that takes into account the actual properties of transmission lines, yielding broadband transmission line-based implementations of lumped-element filter prototypes. Consider a section of transmission line of electrical length θ with $ABCD$ parameters

$$T = \left[\begin{array}{cc} \cos(\theta) & \jmath Z_0 \sin(\theta) \\ \jmath/Z_0 \sin(\theta) & \cos(\theta) \end{array} \right].$$

(10.143)

If this line is terminated in a load, Z_L, then its input impedance is

$$Z_{\mathrm{in}}(\theta) = \frac{\cos(\theta).Z_L + \jmath Z_0.\sin(\theta)}{\jmath/Z_0 \sin(\theta).Z_L + \cos(\theta)}.$$

(10.144)

Now examine two extreme conditions. As the load impedance increases, eventually becoming an open circuit, the input impedance of a line with electrical length θ is defined in terms of a cotangent of the electrical length:

$$Z_L \to \infty \quad \Rightarrow Z_{\mathrm{in}}(\theta) \;=\; \frac{Z_0}{\jmath}.\cot(\theta)$$

(10.145)

$$Y_{\mathrm{in}}(\theta) \;=\; \jmath Y_0.\tan(\theta).$$

(10.146)

As the load impedance reduces to become a short circuit, the input impedance of a line with electrical length θ is defined in terms of a tangent of the electrical length:

$$Z_L \to 0 \quad \Rightarrow \quad Z_{\mathrm{in}}(\theta) = \jmath Z_0.\tan(\theta).$$

(10.147)

These results lead to the Richards' transformation, which replaces the Laplace variable, s, by Richards' variable, S, where $S = \jmath\alpha\tan(\theta)$. This transformation is written

$$s \to S = \jmath\alpha\tan(\theta).$$

(10.148)

For now α and θ are constants that can be chosen as design variables. θ, of course, is the electrical length of the line and is not just $90°$ for a $\lambda/4$

line. Also, α must have the units of impedance and it is the characteristic impedance of the transmission line.

Applying the Richards' transformation to a capacitor, the admittance of the element is transformed as follows:

$$y = sC \rightarrow Y = SC = \jmath\alpha C \tan(\theta), \tag{10.149}$$

so that the capacitor is transformed into an open-circuited stub with characteristic admittance

$$Y_0 = \alpha C . \tag{10.150}$$

If a lumped-element capacitor with admittance $y = sC$ is to be realized using a transmission line, the admittance $Y = SC = \jmath\alpha C \tan(\theta)$ is instead realized. There are two parameters to select to realize this admittance. The first, α, is the characteristic admittance of the transmission line (and for any given transmission line topology there is a minimum and maximum characteristic admittance or impedance that can be realized), and θ is the electrical length of the line.

Applying the transformation to an inductor, the impedance of the element is transformed as follows:

$$Z = sL \rightarrow Z = SL = \jmath\alpha L \tan(\theta), \tag{10.151}$$

so that the inductor is transformed into a short-circuited stub with characteristic impedance

$$Z_0 = \alpha L. \tag{10.152}$$

Thus the Richards' transform converts an inductor into a short-circuited stub and a capacitor into an open-circuited stub.

10.11.2 Richards' Transformation and Stubs

There is a duality between stubs and inductors and capacitors; they are coupled by Richard's transformation. One of the important quantities used in the transformation is the commensurate frequency, f_r, which most often is just twice the operating frequency, f_0. Considering stubs that are one-quarter wavelength long at f_r, the duality is as shown in Figure 10-37.

10.11.3 Richards' Transformation Applied to a Lowpass Filter

In this section, Richards' transformation is used to realize a lumped-element filter in distributed form. The design example begins by considering a Chebyshev lowpass filter with the transmission characteristic

$$|T(s)|^2 = \frac{1}{1 + \varepsilon^2 |K(s)|^2}. \tag{10.153}$$

s, Laplace variable	S, Richards' variable		
	Schematic		Equivalence
jX (shunt inductor)	Z_0	$\lambda_r/4$ — Z_0	$X = Z_0$ $f_r = 2f_0$
jX (series inductor)	Z_0	jZ_0 $\lambda_r/4$ — Z_0	$X = Z_0$ $f_r = 2f_0$
jX (shunt capacitor)	Z_0	$\lambda_r/4$ — $-jZ_0\rightarrow$ Z_0	$X = -Z_0$ $f_r = 2f_0$
jX (series capacitor)	Z_0	$-jZ_0$ $\lambda_r/4$ — Z_0	$X = -Z_0$ $f_r = 2f_0$

Figure 10-37 Equivalences resulting from Richards' transformation. With $f_r = 2f_0$ the transmission line stubs are one-eighth wavelength long at f_0.

With the Richards' transformation

$$s \to \jmath\omega \to \jmath\alpha \tan\left(\theta\right)\ , \tag{10.154}$$

the response becomes

$$\left|T\left(\jmath\alpha\tan\theta\right)\right|^2 = \frac{1}{1 + \varepsilon^2\left|K\left(\jmath\alpha\tan\theta\right)\right|^2}. \tag{10.155}$$

Thus the passband edge at $\omega = 1$ is mapped to $\omega = \theta_1$ as

$$1 \quad\to\quad \alpha\tan\left(\theta_1\right) \tag{10.156}$$

$$\alpha \quad=\quad \frac{1}{\tan\left(\theta_1\right)}. \tag{10.157}$$

Recalling that a capacitor is transformed into an open-circuited stub (see Equation (10.150)), Richards' transformation applied to the lowpass filter prototype results in a filter with transmission line elements only, as shown Figure 10-38, provided that the inverters are realized using transmission lines.

Figure 10-38 Lowpass prototypes: (a) lowpass prototype as a ladder filter with inverters; and (b) lowpass distributed prototype with open-circuited stubs (the impedance looking into the stub is indicated).

The implementation of a lowpass filter in distributed form results in passbands and stopbands that repeat in frequency, as shown in Figure 10-39. This occurs because transmission lines are used to realize lumped elements and the impedance of a transmission repeats every wavelength (or half wavelength in some cases). For example, the input impedance of a stub is the same whether it is one-half wavelength long or one wavelength long.

10.11.4 Richards' Transformation Applied to a Highpass Filter

With reference to Figure 10-40(a), the passband edge at $\omega = 1$ is mapped to θ_1, as shown in Figure 10-40(b). This means that the passband at $\omega = 1$ is mapped to θ_1 as

$$1 \quad \rightarrow \quad \alpha \tan(\theta_1) \tag{10.158}$$

$$\alpha \quad = \quad \frac{1}{\tan(\theta_1)}. \tag{10.159}$$

The sequence of steps transforming a lumped lowpass prototype to its distributed highpass prototype are shown in Figure 10-41. As previously discussed, all the inverters can be approximated by transmission lines of length $\pi/2$ at the corner frequency of the filter.

10.11.5 Kuroda's Identities

Kuroda's identities embody a number of specific manipulations using impedance or admittance inverters. They are particularly useful in implementing Richard's transformation as they physically separate transmission

Figure 10-39 Lowpass to distributed lowpass transformation: (a) lowpass filter response; and (b) distributed lowpass filter response.

Figure 10-40 Highpass to distributed highpass transformation: (a) highpass filter response; and (b) distributed highpass filter response.

(a)

(b)

(c)

Figure 10-41 Ladder prototype transformation: (a) lowpass prototype; (b) highpass prototype response; and (c) distributed highpass prototype (the input impedances of the stubs are indicated).

line stubs, transform series stubs into shunt stubs, and can change characteristic impedances that are either too small or too high to practically realize. Kuroda's identities are a number of equivalent two-port networks, as shown in Figure 10-42. The proof is obtained by obtaining the $ABCD$ parameters of the two ports similar to the technique used throughout this chapter. The identities shown in Figure 10-42 are narrow band. With stubs, the Kuroda identities have broader bandwidths. The Kuroda identities with stubs are shown in Figure 10-43. The major use of these identities is to transform designs with series stubs (in addition to possible shunt stubs) into designs with shunt stubs only.

To see how these identities are used consider the identity shown in Figure 10-42(a). The network on the left has a series inductor which, using transmission lines, is realized by a series stub. A series stub is difficult to realize in most transmission line technologies, including microstrip. Using the identity shown on the right in Figure 10-42(a), the series stub is replaced by the shunt stub used to realize the shunt capacitor. At the same time impedance scaling can be used. If the impedance of the inverter in the network on the left is too low, then it can be scaled by a factor n^2, where $n^2 = 1 + Z_2/Z_1$.

Figure 10-42 Kuroda's identities. Here the inverters are impedance inverters and the designation refers to the impedance of the inverter. Recall that an inverter of impedance Z_1 can be realized by a quarter-wavelength long transmission line of characteristic impedance Z_1. (As usual, element impedances are indicated.)

10.12 Coupled Line Configurations

The coupled line model of a pair of coupled lines was presented in Figure 9-22 on Page 499. This model is repeated in Figure 10-44. The parameters of the network model are related to the model impedances as follows, repeating Equations (9.159) and (9.160):

$$n = \frac{1}{K} = \frac{Z_{0e} - Z_{0o}}{Z_{oe} + Z_{0o}} \tag{10.160}$$

$$Z_0 = \sqrt{(Z_{0e}Z_{0o})}, \tag{10.161}$$

$$Z_{01} = \frac{Z}{\sqrt{1 - K^2}}, \tag{10.162}$$

$$Z_{02} = Z\frac{\sqrt{1 - K^2}}{K^2}. \tag{10.163}$$

Various terminating arrangements of the coupled lines result in several useful filter elements. One arrangement is shown in Figure 10-45. Also shown in this figure is the development of the network model based on

$$n^2 = 1 + Z_2/Z_1$$

(a)

(b)

(c)

(d)

Figure 10-43 The stub form of the Kuroda identities with impedance inverters. The stub impedances shown are the input impedances of the stubs.

Figure 10-44 Network model of a pair of coupled lines.

the model in Figure 9-22. The final network model is a transmission line of characteristic impedance Z_1 in cascade with an open-circuited stub. Consider what happens at the resonant frequency, f_0 (the frequency at which the lines are one-quarter wavelength long). At lower frequencies, $f \ll f_0$, the Z_2 line is an open circuit and signals travel along Z_1. At resonance the Z_2 stub becomes a short circuit and signals do not pass. This is a crude verification that this is a lowpass structure. The process is visual

Figure 10-45 Lowpass distributed network section derived from a pair of coupled lines with Port 1 open-circuited. The open circuit is indicated by a node (open circle) with a line through it. The final network model is a transmission line of characteristic impedance A_1 and an open-circuited stub of impedance characteristic impedance Z_2. The lines are one-quarter wavelength long at the corner frequency.

Figure 10-46 Parallel coupled-line section with Ports 1 and 3 open-circuited, and network models.

and is expected to be self-explanatory. Other examples are shown in Figures 10-46 to 10-48.

A detailed combline filter design will be considered in the latter parts of this chapter, so the final model of a **Parallel Coupled Line** (**PCL**) section

Figure 10-47 Interdigital section and network models.

Figure 10-48 Combline section and network models.

Figure 10-49   Equivalent models of a section of combline.

in combline configuration has been taken further in Figure 10-48 than for the other sections, but equivalent reductions can be obtained. The final network reduction is repeated in Figure 10-49, where it will be shown that the model in Figure 10-49(b) is equivalent to the model in Figure 10-49(a). In the combline filter synthesis the network of Figure 10-49(b) is obtained and this can be related back to the dimensions of the coupled line. As is done throughout this chapter, the equivalence is done using $ABCD$ parameters and, as will be seen, the equivalence will not be at just one frequency but will be broadband. The $ABCD$ parameters of the network in Figure 10-49(a) are obtained by cascading the $ABCD$ parameters of three two-ports (the $ABCD$ parameters of which are given in Table 6-1 on Page 298):

$$T_A = T_{\text{TRANSFORMER}} T_{\text{SERIES STUB}} T_{\text{SHUNT STUB}} \tag{10.164}$$

$$= \begin{bmatrix} 1/n & 0 \\ 0 & n \end{bmatrix} \begin{bmatrix} 1 & jZ_{02}\tan\theta \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -j/(Z_{01}\tan\theta) & 1 \end{bmatrix} \tag{10.165}$$

$$= \begin{bmatrix} 1/n & jZ_{02}\tan\theta/n \\ 0 & n \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -j/(Z_{01}\tan\theta) & 1 \end{bmatrix} \tag{10.166}$$

$$= \begin{bmatrix} \frac{1}{n}\left(1 + \frac{Z_{02}}{Z_{01}}\right) & jZ_{02}\tan\theta/n \\ -jn/(Z_{01}\tan\theta) & n \end{bmatrix}. \tag{10.167}$$

The $ABCD$ parameters matrix of the network on Figure 10-49(b):

$$T_B = T_{\text{SHUNT STUB}} T_{\text{SERIES STUB}} T_{\text{SHUNT STUB}} \tag{10.168}$$

$$= \begin{bmatrix} 1 & 0 \\ -j/(Z_{011}\tan\theta) & 1 \end{bmatrix} \begin{bmatrix} 1 & jZ_{012}\tan\theta \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -j/(Z_{022}\tan\theta) & 1 \end{bmatrix} \tag{10.169}$$

$$= \begin{bmatrix} 1 & jZ_{012}\tan\theta \\ -j/(Z_{011}\tan\theta) & 1 + Z_{012}/Z_{011} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -j/(Z_{022}\tan\theta) & 1 \end{bmatrix} \tag{10.170}$$

$$= \begin{bmatrix} 1 + \frac{Z_{012}}{Z_{022}} & jZ_{012}\tan\theta \\ \frac{-j}{\tan\theta}\left(\frac{1}{Z_{011}} + \frac{1}{Z_{022}} + \frac{Z_{012}}{Z_{011}Z_{022}}\right) & 1 + \frac{Z_{012}}{Z_{011}} \end{bmatrix}. \tag{10.171}$$

Equating Equations (10.167) and (10.171) yields

$$1 + \frac{Z_{012}}{Z_{022}} = \frac{1}{n}\left(1 + \frac{Z_{02}}{Z_{01}}\right) \tag{10.172}$$

$$Z_{012} = Z_{02}/n \tag{10.173}$$

$$\left(\frac{Z_{011} + Z_{022} + Z_{012}}{Z_{011} Z_{022}} \right) = \frac{n}{Z_{01}} \tag{10.174}$$

$$1 + \frac{Z_{012}}{Z_{011}} = n, \tag{10.175}$$

which has the solution

$$Z_{012} = \frac{Z_{02}}{n} \tag{10.176}$$

$$Z_{011} = \frac{Z_{012}}{n-1} = \frac{Z_{02}}{n(n-1)} \tag{10.177}$$

$$Z_{022} = \frac{Z_{01} Z_{02}}{Z_{02} - (n-1)Z_{01}}. \tag{10.178}$$

Rearranging these, expressions for Z_{01}, Z_{02}, and n can be obtained, and from Equations (9.159) and (9.160) and the coupled line analysis of Section 9.6, the geometric parameters of the combline coupled line section can be obtained.

10.13 Inverter Network Scaling

A filter that uses transmission lines is nearly always synthesized from a filter prototype that contains inverters. The synthesis of a filter begins with a normalized prototype that is transformed to the desired frequency, impedance, and type. It is the purpose of this section, to enable such scaling when inverters are present in the prototype. In particular, it is shown that scaling the admittance of the network of Figure 10-50(a), a common inverter subnetwork, by a factor x results in the network in Figure 10-50(b).

The nodal admittance matrix of the network in Figure 10-50(a) is

$$Y = \begin{bmatrix} 0 & -\jmath J & 0 \\ -\jmath J & y & -\jmath J \\ 0 & -\jmath J & 0 \end{bmatrix}. \tag{10.179}$$

Assigning nodal voltages to Terminals 1, 2, and 3 the nodal admittance matrix equation is

$$\begin{bmatrix} 0 & -\jmath J & 0 \\ -\jmath J & y & -\jmath J \\ 0 & -\jmath J & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} J_1 \\ 0 \\ J_3 \end{bmatrix}, \tag{10.180}$$

and by eliminating Node 2 using network condensation (see Section A.10 on Page 834) this reduces to

$$\begin{bmatrix} J^2/y & J^2/y \\ J^2/y & J^2/y \end{bmatrix} \begin{bmatrix} v_1 \\ v_3 \end{bmatrix} = \begin{bmatrix} J_1 \\ J_3 \end{bmatrix}, \tag{10.181}$$

Figure 10-50 Inverter network: (a) network with two identical admittance inverters with an inserted shunt element of admittance, y_1; (b) equivalent network using equivalence shown in Figure 10-19; (c) scaled original network; and (d) scaled equivalent network. Element values are impedances except for y and y_1, which are admittances.

and this describes the external characteristics of the subnetwork in Figure 10-50(a). The intent is to show that the network of Figure 10-50(c), which is a scaled version of the original network of Figure 10-50(a), has an identical external characteristic; that is, it has the same reduced nodal admittance matrix. The nodal admittance matrix of the scaled network in Figure 10-50(c) is

$$Y' = \begin{bmatrix} 0 & -\jmath J\sqrt{x} & 0 \\ -\jmath J\sqrt{x} & yx & -\jmath J\sqrt{x} \\ 0 & -\jmath J\sqrt{x} & 0 \end{bmatrix}. \tag{10.182}$$

Assigning nodal voltages to Terminals 1, 2, and 3, the nodal admittance matrix equation is

$$\begin{bmatrix} 0 & -\jmath J\sqrt{x} & 0 \\ -\jmath J\sqrt{x} & yx & -\jmath J\sqrt{x} \\ 0 & -\jmath J\sqrt{x} & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} J_1 \\ 0 \\ J_3 \end{bmatrix}, \tag{10.183}$$

and by eliminating Node 2 this reduces to

$$\begin{bmatrix} (J^2 x)/(yx) & (J^2 x)/(yx) \\ (J^2 x)/(yx) & (J^2 x)/(yx) \end{bmatrix} \begin{bmatrix} v_1 \\ v_3 \end{bmatrix} = \begin{bmatrix} J^2/y & J^2/y \\ J^2/y & J^2/y \end{bmatrix} \begin{bmatrix} v_1 \\ v_3 \end{bmatrix}$$

$$= \begin{bmatrix} J_1 \\ J_3 \end{bmatrix}. \tag{10.184}$$

Thus the original network shown in Figure 10-50(a) has identical external electrical characteristics as the scaled network of Figure 10-50(c), with

Figure 10-51 Realization of a series inductor as a shunt capacitor with inverters.

the characteristic admittance of the inverters scaled by $\sqrt{x}$ and the shunt admittance scaled by x.

A generalization of this result (which is useful when there are additional connections between Nodes 1 and 3, and at Node 2) is that multiplying a row and a column of the nodal admittance matrix by a factor results in identical external characteristics. Note that the element sharing a row and column is multiplied twice. Nodal matrix reduction as used here for a 3×3 nodal admittance matrix can be used to prove this general result.

EXAMPLE 10. 8   Inductor Synthesis

Lumped inductors have low Q. Fortunately, in microwave design they can be realized using inverters and a shunt capacitor which have high Q. Realize the series inductor in Figure 10-51(a) with a shunt capacitor and 10 Ω impedance inverters.

Solution:

The series inductor can be realized by the circuit of Figure 10-51(b), where the shunt capacitor C has the same numeric value as the series inductor. The inverters here can be either impedance inverters or admittance inverters, as their value is equal to one. The design requires that these be realized as 10 Ω series inverters, but scaling is performed on admittance inverters, as shown in Figure 10-51(c), where the value of the admittance inverters is 0.1 S = $\sqrt{x}$. Thus $x = 0.01$ and the admittance of the capacitor must be scaled by 0.01 (thus the capacitance is scaled by 0.01), so $C_1 = C/100 = 10$ pF. The final design is shown in Figure 10-51(d).

Figure 10-52 A lowpass prototype with 1 S admittance inverters.

10.14   Combline Filter

In this section, the combline filter is considered. This is one of the most commonly used transmission line-based filters and is based on the combline section shown in Figure 10-48. One of the important parts of the development of a combline filter is a special property of inverter scaling that will be considered in the next subsection. Following that, combline filter design will be presented by example.

10.14.1   Scaled Prototype with Inverters

This subsection presents an example of how the inverter network scaling of Section 10.13 on Page 563 can be used with combline filters. A lowpass prototype with inverters, suitable for combline filters, is shown in Figure 10-52. Through several transformations this lowpass filter prototype will become a bandpass filter in exactly the right form to be realized by coupled parallel lines in a combline configuration. In the process the capacitors become LC resonators which are realized by a combination of capacitors and stub transmission lines. There are three shunt capacitors in the prototype and, as will be made clearer in the combline filter design that follows, it is desirable that the shunt elements that realize these be identical. To achieve this, the central capacitor must be scaled by a factor of $\frac{1}{2}$. This can be achieved by scaling the impedance of the inverters by $1/\sqrt{2}$. So the network of Figure 10-52 can be replaced by a network where the shunt elements have the desired property of being identical.

10.14.2   Third-Order Chebyshev Combline Filter

The design of a third-order Chebyshev combline bandpass filter is presented here as an example of the filter design procedure based on synthesis. The design is presented in steps and brings together many of the concepts presented in this chapter.

Step 1: Choice of Lowpass Filter Prototype

Design begins with the choice of a prototype that is expected to achieve the required specifications. If the choice is not suitable then another topology

must be chosen. Consider a third-order Chebyshev lowpass filter prototype, shown in Figure 10-53, with a passband ripple of $\varepsilon = 0.1$.

$$C_1 = 0.85239 \text{ F} = C_2 \quad L_1 = 1.10422 \text{ H}$$

Figure 10-53   Step 1. A third-order Chebyshev lowpass filter prototype.

Step 2: Replace Series Elements

The next step in the transformation of the lowpass prototype in Figure 10-53 is to replace the inductor by a shunt capacitor in cascade with admittance inverters, as shown in Figure 10-54. This step uses the series inductor transformation shown in Figure 10-16.

(a)

(b)

$$C_1 = 0.85239 \text{ F} = C_3 \quad L_1 = 1.10422 \text{ F}$$

Figure 10-54   Step 2. Prototype filter with inductor replaced by a capacitor and admittance inverter combination: (a) with unity inverting transformer; and (b) with the transformer eliminated as it has no effect on the filter response. Note that L_1 has the units of farads.

Step 3: Bandpass Filter Transformation

Now the prototype can be scaled to the bandpass circuit. The filter has 10% bandwidth with approximate corner frequencies at $f_1 = 950$ MHz and $f_2 = 1050$ MHz. Using the filter type transformations in Figure 10-30, and $\omega_0 = 2\pi 10^9$ and $\alpha = \omega_0/(\omega_2 - \omega_1) = 1/\text{fractional bandwidth}) = 10$, the first

$C_1' = 1356.6272 \text{ pF} = C_3' \quad L_1' = 0.018718 \text{ nH} = L_3'$

$C_2' = 1757.435 \text{ pF} \quad L_2' = 0.014444 \text{ nH}$

Figure 10-55 Step 3. Bandpass filter prototype with inverters derived from the lowpass filter prototype of Figure 10-54.

$C_1^\sharp = 27.1325 \text{ pF} = C_3^\sharp \quad L_1^\sharp = 0.9359 \text{ nH} = L_3^\sharp$

$C_2^\sharp = 35.1487 \text{ pF} \quad L_2^\sharp = 0.7224 \text{ nH}$

Figure 10-56 Step 4. Bandpass filter approximation where the inverters are now impedance inverters.

capacitor in Figure 10-54, C_1, is transformed to

$$C_1' = \frac{\alpha C_1}{\omega_0} = 1356.6272 \text{ pF} \quad \text{and} \quad L_1' = \frac{1}{\alpha C_1 \omega_0} = 0.018718 \text{ nH}. \quad (10.185)$$

The other capacitors are transformed similarly and the bandpass prototype becomes that shown in Figure 10-55.

Step 4: Impedance Scaling

The system impedance is now scaled from 1 to 50 Ω, leading to the prototype in Figure 10-56.

Step 5: Conversion of Lumped-Element Resonators

Each lumped-element resonator in Figure 10-55 comprises a lumped capacitor and a lumped inductor. The inductors are particularly difficult to realize efficiently and have high loss. The idea here is to replace each lumped resonator (e.g., C_1 and L_1 in Figure 10-55) with a network comprising a lumped capacitor and a transmission line stub. The resonator equivalence is shown in Figure 10-57. The equivalence is developed at the center frequency of the filter passband by making sure that the circuits have the same admittance at the filter center frequency. A broadband approximation is

Figure 10-57 Resonator equivalence: (a) lumped resonator resonant at radian frequency Ω; (b) mixed lumped-distributed resonator containing a resonant short-circuited stub of characteristic impedance Z_0 resonant at radian frequency Ω_r; and (c) short-circuited transmission line stub resonant at radian frequency Ω_r. In (b) and (c) Z_{01} is the characteristic impedance of the transmission line and Z_1 is the input impedance of the shorted transmission line.

Figure 10-58 Short-circuited transmission line used as a stub resonator in a bandpass filter. The line is resonant at the radian frequency ω_r and so the electrical length of the line is $\theta = 90°$. Using the design choice here, the filter center passband frequency $\omega = \omega_r/2$, and so $\theta = 45°$ and so the input impedance of the shorted-stub is $Z_1 = \jmath Z_{01}$.

obtained by equating the derivatives of the admittances as well. The process begins by deriving the input admittance of the lumped-element resonator as follows.

The input admittance of the parallel LC lumped-element resonator is

$$Y_{\text{in}} = \frac{\jmath\left(\omega^2 LC - 1\right)}{\omega L}, \tag{10.186}$$

with frequency derivative at resonance of

$$\frac{dY_{\text{in}}}{d\omega} = \frac{\jmath\left(\omega^2 LC + 1\right)}{\omega^2 L}. \tag{10.187}$$

Now consider a short-circuited stub that is resonant at a radian frequency ω_r; that is, the short-circuited stub is one-quarter wavelength long at ω_r. Thus the input impedance of the stub at ω_r is infinite. From Equation (4.139), the input impedance at ω is

$$Z_1 = \jmath Z_{01} \tan\left(\frac{\pi}{2}\frac{\omega}{\omega_r}\right). \tag{10.188}$$

Since resonant frequency of the stub, f_r, is twice the center frequency of the passband, Equation (10.188) becomes

$$Z_1 = \jmath Z_{01} \tan\left(\frac{\pi}{4}\right) = \jmath Z_{01}. \tag{10.189}$$

Thus for the lumped-distributed element in Figure 10-57(b), the input admittance is

$$Y_{\text{in}}' = \frac{\jmath \left(\omega C_o Z_{01} \tan\left(\frac{\pi}{2}\frac{\omega}{\omega_r}\right) - 1\right)}{Z_{01} \tan\left(\frac{\pi}{2}\frac{\omega}{\omega_r}\right)}, \tag{10.190}$$

where ω_r is the resonant frequency. The frequency derivative at resonance is

$$\frac{dY_{\text{in}}'}{d\omega} = \jmath\frac{1}{2}\frac{\left(2C_o Z_{01}\left(\tan\left(\frac{\pi}{2}\frac{\omega}{\omega_r}\right)\right)^2 \omega\omega_r + \pi + \pi\left(\tan\left(\frac{\pi}{2}\frac{\omega}{\omega_r}\right)\right)^2\right)}{Z_{01}\left(\tan\left(\frac{\pi}{2}\frac{\omega}{\omega_r}\right)\right)^2 \omega_r}. \tag{10.191}$$

For the lumped-element and distributed filters to be identical at the center frequency and approximating each other at nearby frequencies, Equations (10.186) and (10.190) are equated:

$$Y_{\text{in}} = Y_{\text{in}}' \tag{10.192}$$

and Equations (10.187) and (10.191) are also equated:

$$\frac{dY_{\text{in}}}{d\omega} = \frac{dY_{\text{in}}'}{d\omega}. \tag{10.193}$$

Specific Design Choice $\omega_r = 2\omega_0$

At this stage a design choice must be made regarding the relationship of ω_0 to ω_r; that is, the resonant frequency of the stubs in relation to the passband center frequency. Here a shorted stub is considered which will present an open circuit at its input when it is $\lambda/4$ long. The specific design choice is that the resonant frequency of the stubs, f_r, is chosen to be twice the center frequency of the filter passband, that is, $\omega_r = 2\omega_0$. The frequency f_r is generally called the **commensurate frequency** to avoid confusion resulting from the resonant frequency of the filter resonators being different from the resonant frequency of stubs. The stub design proceeds as follows. The equivalence of Equations (10.192) and (10.193) are established at the center of the passband, 1 GHz, and so $\omega = 2\pi10^9$ at the center of the passband and $\omega_r = 2\pi(2 \times 10^9)$ defines the stub resonant frequency as 2 GHz. Now $\omega = \frac{1}{2}\omega_r$, so Equations (10.192), (10.186), and (10.190) become, with $\omega_r = 2\omega$,

$$\frac{\jmath\left[\left(\frac{\omega_r}{2}\right)^2 LC - 1\right]}{\frac{\omega_r}{2}L} = \frac{\jmath\left[\frac{\omega_r}{2}C_0 Z_{01} \tan\left(\frac{\pi}{2}\frac{\omega_r/2}{\omega_r}\right) - 1\right]}{Z_{01} \tan\left(\frac{\pi}{2}\frac{\omega_r/2}{\omega_r}\right)}. \tag{10.194}$$

That is,

$$\frac{\left(\omega_r^2 LC - 4\right)}{2\omega_r L} = \frac{\left(\frac{\omega_r}{2} C_0 Z_0 \tan \frac{\pi}{4} - 1\right)}{Z_{01} \tan \frac{\pi}{4}}. \tag{10.195}$$

Since $\tan \frac{\pi}{4} = 1$,

$$\frac{\left(\omega_r^2 LC - 4\right)}{2\omega_r L} = \frac{\left(\frac{\omega_r}{2} C_0 Z_{01} - 1\right)}{Z_{01}}, \tag{10.196}$$

and rearranging,

$$C_0 = C - \frac{4}{\omega_r^2 L} + \frac{2}{\omega_r Z_{01}}. \tag{10.197}$$

Another relationship comes from equating derivatives. From Equations (10.193), (10.187), and (10.191), and with $\omega = \frac{1}{2}\omega_r$ and $\tan(\frac{1}{2}\frac{\omega}{\omega_r}) = \tan \frac{\pi}{4} = 1$,

$$\frac{\left(\omega_r^2 LC + 4\right)}{\omega_r^2 L} = \frac{1}{2} \frac{\left(2 C_0 Z_{01} \frac{\omega_r^2}{2} + \pi + \pi\right)}{Z_{01} \frac{\omega_r}{2}} = \frac{\left(C_0 Z_{01} \omega_r^2 + 2\pi\right)}{Z_{01} \omega_r}. \tag{10.198}$$

Substituting for C_0 from Equation (10.197) and rearranging, the characteristic impedance of the stub is

$$Z_{01} = \frac{\omega_r L}{4} \left(1 + \frac{\pi}{2}\right). \tag{10.199}$$

In the passband, which is at half the resonant frequency of the stub because of design choice, the input impedance of the stub in the passband is

$$Z_1 = \jmath Z_{01} \tan \beta \ell = \jmath Z_{01}. \tag{10.200}$$

Now consider the first resonator in Figure 10-56 with $L = 0.9359$ nH and $C = 27.1325$ pF. With $\omega_r = 2\pi(2 \text{ GHz})$, then from Equation (10.199),

$$Z_{01} = 7.558 \ \Omega, \tag{10.201}$$

and from Equation (10.197),

$$C_0 = 21.123 \text{ pF}. \tag{10.202}$$

This process is repeated for each lumped-element resonator in Figure 10-56, leading to the prototype shown in Figure 10-59.

Steps 6 and 7: Equating Characteristic Impedances of Stubs

The stubs in Figure 10-59 can be physically realized using transmission line segments. It would be ideal if the impedances of the stubs (i.e., the impedances looking into the stubs) are identical. To achieve this, the method of nodal admittance matrix scaling described in Section 10.13 on Page 563 is used with the inverters scaled by $\sqrt{Z_1/Z_2}$ and the admittance, $C_2''//Z_2$,

$$C_1'' = 21.123 \text{ pF} = C_3'' \qquad C_2'' = 27.363 \text{ pF}$$
$$Z_{01} = 7.558 \ \Omega = Z_{03} \qquad Z_{02} = 5.834 \ \Omega \tag{10.203}$$

Figure 10-59 Step 5. Bandpass combline filter with impedance inverters. The transmission line stubs present impedances $Z_1 = \jmath Z_{01}$, $Z_2 = \jmath Z_{02}$, and $Z_3 = \jmath Z_{03}$ since the resonant frequencies of the stubs are twice that of the design center frequency.

$$C_2''' = 21.1204 \text{ pF} \quad \text{and} \quad Z_{02}' = 7.558 \ \Omega = Z_{01} = Z_{03}$$

Figure 10-60 Step 6. Bandpass combline filter with impedance inverters. Again $Z_1 = \jmath Z_{01}$, $Z_2 \jmath Z_{02}$, and $Z_3 \jmath Z_{03}$ are the input impedances of the stubs since the resonant frequencies of the stubs is twice that of the design center frequency.

Figure 10-61 Inverter translation used with the combline filter design: (a) impedance inverter; (b) realization as a lumped element circuit; and (c) realization using short-circuited stubs resonant at twice the passband center frequency.

scaled by Z_1/Z_2. This step scales each inverter impedance to 56.9102 Ω and also sets the impedance of all the shunt stubs to 7.558 Ω. The combline filter is now as shown in Figure 10-60.

On Page 540 it was shown that a good narrowband model of an inverter comprises a pi network of stubs. So the inverter in the current filter design, shown in Figure 10-61(a), is modeled by the lumped element circuit in Figure 10-61(b) and the stub circuit shown in Figure 10-61(c).

$$C_2''' = 21.1204 \text{ pF} \quad \text{and} \quad Z_2' = 7.558 \ \Omega = Z_1 = Z_3$$

Figure 10-62 Step 6b. Bandpass combline filter with impedance inverters. Again Z_1, Z_2, and Z_3 are the input impedances of the stubs

Stub Impedance When $f_r = 2f_0$

The most commonly used design choice is to choose the resonant (or commensurate) frequency of the resonator to be twice the center frequency of the design. The resonant frequency $f_r = 2\pi\omega_r$ of the shorted transmission line stub is the frequency at which the length of the stub is one-quarter wavelength long.[3] So the common design choice is that $f_r = 2f_0$. Then Equation (10.122) becomes

$$Z_0 = \frac{1}{J \tan\left(\frac{\pi}{2}\frac{\omega_0}{2\omega_0}\right)} = \frac{1}{J \tan\left(\frac{\pi}{4}\right)} = \frac{1}{J} = K. \tag{10.204}$$

Equation (10.204) defines the characteristic impedance of the transmission lines stubs realizing an inverter. Then the elements of the subnetwork in Figure 10-61(b) have the parameters

$$J = 1/K = 56.9102 \ \Omega = 17.5715 \text{ mS} \tag{10.205}$$

---

[3] Of course there are resonant frequencies at every multiple of one-quarter wavelength. The first resonance occurs at f_r, as then the input impedance of the short-circuited stub is an open circuit.

$$C_1'' = 21.1215 \text{ pF} = C_3'' \quad C_2''' = 21.1204 \text{ pF} \qquad Z_{02}'' = 10.2915 \ \Omega$$
$$Z_{01}' = 8.7161 \ \Omega = Z_{03}' \qquad Z_{012} = 56.9102 \ \Omega = Z_{023} \tag{10.207}$$

Figure 10-63 Step 7: Bandpass filter approximation where Z_{01}', Z_{02}', Z_{03}', Z_{012}, and Z_{023} are the characteristic impedances of the stubs which are resonant at frequency f_r—twice the center frequency of the bandpass filter.

$$C_{t1} = 2.7171 \text{ pF} = C_{t3} \qquad C_{t2} = 2.7171 \text{ pF}$$
$$Z_{0t1} = 67.7535 \ \Omega = Z_{0t3} \quad Z_{0t2} = 80 \ \Omega \qquad Z_{0t12} = 442.3836 \ \Omega = Z_{0t23}$$

Figure 10-64 Step 8: Bandpass filter approximation.

$$Z_0 = \frac{1}{J \tan(\theta)} = \frac{K}{\tan(\theta)} = \left. \frac{K}{\tan\left(\frac{\pi}{2}\frac{\omega}{\omega_r}\right)} \right|_{\omega=\omega_0} = 56.9102 \ \Omega. \tag{10.206}$$

At this stage there are several pairs of stubs in parallel. Figure 10-62 illustrates how a pair of parallel stubs is realized by a single stub. Finally, the prototype in Figure 10-63 is developed.

Step 8: Scaling Characteristic Impedances of Stubs

In Figure 10-63 the short-circuited stubs have characteristic impedances of $8.7 \ \Omega$ and $10.3 \ \Omega$, which is low, and will result in wide microstrip lines. These will be scaled to a different impedance level to raise the characteristic impedances of the short-circuited stubs. Note that the characteristic impedance of the middle short-circuited stub can be raised to $80 \ \Omega$ if the system impedance is raised from $50 \ \Omega$ to $388.6681 \ \Omega$. This leads to the element values complying with the bandpass circuit of Figure 10-64.

Step 9: Scaling to 50 Ω System Impedance

For the filter to operate in a $50 \ \Omega$ system, input and output impedance inverters are required to scale the source and load back to $50 \ \Omega$. The

Figure 10-65   Step 9: Bandpass filter approximation.

Figure 10-66   Inverter equivalence: (a) terminated impedance inverter; and (b) its equivalent transmission line form where $Z_0 = K$. The equivalence is exact at the frequency where the transmission line is one-quarter wavelength long.

Figure 10-67   Equivalence of an inverter and a capacitor section. Note that one of the capacitors must be negative.

resulting circuit is shown in Figure 10-65 with input and output inverters of impedance $\sqrt{50 \times 388.61} = 139.404\ \Omega$. Note that inverters with an impedance of $-139.404\ \Omega$ would work just as well.

10.14.3   Realization of the Input/Output Inverters

At this stage the design is as shown in Figure 10-65, where the inverters at the input and output remain to be synthesized. Focusing on just the output inverter with a load R_L, as shown in Figure 10-66, it will be shown here how this can be realized using the capacitive network of Figure 10-67. The input admittance of the inverter plus load resistor of Figure 10-66 is

$$Y_{\text{in}} = \frac{R_L}{K^2}, \tag{10.208}$$

and this must equate to the input admittance of the circuit in Figure 10-67 where $K = 139.404\ \Omega$. The input admittance of the circuit in Figure 10-67 is

$$Y_{\text{in}} = sC_a + \frac{1}{\frac{1}{sC_b} + R_L}. \tag{10.209}$$

$$C_a = -1.0657 \text{ pF} \qquad C_b = 1.22306 \text{ pF}$$

Figure 10-68 The external inverters of the prototype shown in Figure 10-65 replaced by capacitive networks.

$C_1 = C_a + C_{t1} = 1.6515 \text{ pF} = C_3$ $C_{t2} = 2.7171 \text{ pF}$ $C_b = 1.22306 \text{ pF}$
$Z_{0t1} = 67.7535 \ \Omega = Z_{0t3}$ $Z''_{0t2} = 80 \ \Omega$ $Z_{0t12} = 442.3836 \ \Omega = Z_{0t23}$

Figure 10-69 Step 10: The bandpass filter approximation combining the capacitive equivalent of the inverters with the first and last resonators in the circuit of Figure 10-65.

Thus, equating the real parts of Equations (10.208) and (10.209),

$$\Re\left(Y_{\text{in}}\right) = \frac{R_L \omega^2 C_b^2}{R_L^2 \omega^2 C_b^2 + 1} = \frac{R_L}{\left(Z_0\right)^2}, \tag{10.210}$$

and equating the imaginary parts yields

$$\Im\left(Y_{\text{in}}\right) = \frac{\omega \left(C_a \omega^2 C_b^2 R_L^2 + C_a + C_b\right)}{\omega^2 C_b^2 R_L^2 + 1} = 0. \tag{10.211}$$

Solving Equations (10.210) and (10.211) enables us to replace the inverters shown in Figure 10-65 with the capacitive networks as shown in Figure 10-68. This capacitive network is not a general replacement of an inverter. It only works when the inverter is terminated in a resistor. Note that C_a is negative, and this is absorbed in the first and last resonators so that the final lumped-element realization is as shown in Figure 10-69.

10.14.4 Interresonator Coupling

A configuration that can often occur in combline filter design is interresonator coupling, occurring as illustrated for a pair of capacitors separated by an admittance inverter

Figure 10-70   Bandpass filter approximation with an admittance inverter.

Figure 10-71   Bandpass resonator.

A lowpass filter section is shown in Figure 10-70, where, for a bandpass filter, each capacitor is replaced by a resonator, as shown in Figure 10-71. The inverter therefore provides interresonator coupling. The input impedance of the circuit in Figure 10-70 is

$$Z_{\text{in}(s)} = \frac{1}{sC_1 + \frac{J_{12}^2}{sC_2}}. \tag{10.212}$$

With $s = +\jmath\omega$, the poles of Z_{in} occur when

$$J_{12}^2 - \omega^2 C_1 C_2 = 0, \tag{10.213}$$

thus the radian frequencies of the poles are

$$\omega = \pm \frac{J_{12}}{\sqrt{C_1 C_2}}. \tag{10.214}$$

Applying the bandpass transformation (so that the capacitors are replaced by the resonators of Figure 10-71), the frequency is transformed as

$$\omega = \alpha \left(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega} \right) \tag{10.215}$$

and the poles of the bandpass filter occur at

$$\alpha \left(\frac{\omega_1}{\omega_o} - \frac{\omega_o}{\omega_1} \right) = +\frac{J_{12}}{\sqrt{C_1 C_2}} \tag{10.216}$$

and

$$\alpha \left(\frac{\omega_2}{\omega_o} - \frac{\omega_o}{\omega_2} \right) = -\frac{J_{12}}{\sqrt{C_1 C_2}}. \tag{10.217}$$

The bandwidth of the bandpass filter is

$$\omega_1 - \omega_2 = \frac{J_{12}\omega_o}{\alpha\sqrt{C_1 C_2}}. \tag{10.218}$$

Solving the above equations yields

$$f_1 - f_2 = \frac{J_{12}\omega_o}{2\pi\alpha\sqrt{C_1 C_2}}, \tag{10.219}$$

which is called the coupling bandwidth. The coupling bandwidth for the first pair of resonators in the filter of Figure 10-69 is shown in Figure 10-72. This is a measure of the coupling between the first and second resonators when the source and the third resonator are eliminated from the circuit. Since the filter is physically symmetrical, the same plot is valid for the coupling between the second and third resonators.

10.14.5 Input Coupling

Figure 10-71 shows a bandpass resonator with input admittance

$$Y_{\text{in}}(s) = SC + \frac{1}{SL} \tag{10.220}$$

and a reflection coefficient in a 1 Ω system of

$$\Gamma_{(s)} = \frac{1 - Y_{\text{in}}(s)}{1 + Y_{\text{in}}(s)} = a + jb, \tag{10.221}$$

where a and b are the real and imaginary parts. With $s = j\omega$, the phase of Γ is

$$\varphi(\omega) = \tan^{-1}\left(\frac{b}{a}\right) \tag{10.222}$$

and the **group delay** function is the differential of the phase as follows:

$$\text{Group delay} = \frac{d\varphi}{d\omega} = \frac{d}{d\omega}\left[\tan^{-1}\left(\frac{b}{a}\right)\right]. \tag{10.223}$$

The peak of the group delay occurs at the resonant frequency of the resonator. Figure 10-72 shows the typical response of a bandpass filter.

10.14.6 Implementation

The physical layout of the filter designed in the previous sections is shown in Figure 10-73 where there are three parallel lines forming two pairs of coupled lines. The filter synthesis accounts for the effect of parasitic capacitances and the inductive effect of vias among other nonidealities.

Figure 10-72 Interresonator response with two resonators: (a) magnitude of transmission and reflection responses; (b) phase of transmission response; and (c) group delay.

Figure 10-73 Physical layout of the combline bandpass filter designed in Sections 10.14–10.14.5: (a) with discrete capacitors; and (b) with integrated capacitors.

The bandwidth of the filter is particularly sensitive to the strength of the coupling of the parallel lines. Coupling can be thought of as a second-order effect, with the single transmission line propagation on the lines being first order. It is therefore not a surprise that small third-order effects can significantly impact filter performance. Following the synthesis outlined in the previous subsections, the filter layout is typically modeled in an electromagnetic simulator and the lengths adjusted and spacing of the transmission line segments adjusted. However, dimensions are usually only adjusted a few percent for filters with center frequencies of a gigahertz, with greater adjustment at higher frequencies. An exception is the separation of the lines, which can change significantly as small perturbations can result in additional (parasitic) coupling so that even EM simulation is not sufficient to fix the design of a filter. Iterative design, fabrication, and testing is usually required to meet specifications, especially if the filter bandwidth is relatively narrow (say less than 5%). The emphasis placed on the inverters having the same impedance translates into each of the coupled lines having the same length, about one-quarter wavelength at the center of the passband. It would not be possible to realize inverters of different impedances using the layout of Figure 10-73.

10.15 Parallel Coupled Line Filters in an Inhomogeneous Medium

An additional complexity arises when this filter is realized on an inhomogeneous transmission media due to the inequality of the even- and odd-mode phase velocities. This effect can be investigated by utilizing the full $ABCDmatrix$ (see Equation (9.150)) of each PCL section in a filter for a known even-:odd-mode velocity ratio. The inhomogeneous medium results in additional spurious passbands centered approximately at even multiples of f_0 [145–159].

Figure 10-74   Bandpass filter approximation.

$C_1 = 0.85239F = C_3$ $C_2 = 1.10422$ F

$L_1 = L_3 = \frac{1}{C_1} = \frac{1}{C_3} = 1.1743$ H $L_2 = \frac{1}{C_2} = 0.9064$ H

Figure 10-75   Bandstop filter prototype.

$L_1 = L_3 = 1.868$ nH $C_1 = C_3 = 13.587$ pF

$L_2 = 1.442$ nH $C_2 = 17.601$ pF

Figure 10-76   Bandstop filter following transformation from the lowpass prototype.

10.16   Design of a Bandstop Filter

Design of the bandstop filter begins with the lowpass filter prototype shown in Figure 10-74. To the lowpass prototype, the highpass transformation is applied to obtain the highpass prototype of Figure 10-75. Picking a center frequency of approximately 1 GHz and corner frequencies of $f_1 = 950$ MHz and $f_2 = 1050$ MHz, corresponding to a bandwidth of approximately 10%, the bandstop transformation is now applied. This results in the prototype of Figure 10-76. Finally, scaling the system impedance to 50 Ω leads to the prototype of Figure 10-77.

It can be seen that the difference between the capacitor and inductor impedance values is very large. The impedance inverters must remain set at 50 Ω in order to obtain broad match. In fact, approximating the inverters by sections of transmission lines will have virtually no effect on the response in the bandstop region. In the passband, energy will pass at all frequencies.

$$L_1' = L_3' = 1.868 \times 50 = 93.448 \text{ nH}$$
$$C_1' = C_3' = 13.587/50 = 0.2717 \text{ pF}$$
$$L_2' = 1.442 \times 50 = 72.1362 \text{ nH}$$
$$C_2' = 17.601/50 = 0.352 \text{ pF}$$

Figure 10-77   Bandstop filter after impedance transformation.

(a)   (b)

Figure 10-78   Transformations of the resonators in the bandstop filter to obtain realizable values. The series LC resonator in (a) is transformed to the form in (b).

In Figure 10-77, the inductor values are relatively large and the capacitor values are relatively small so that it will be difficult to realize the filter in either lumped or distributed form. These values must be scaled to obtain realizable values. One possible transformation is shown in Figure 10-78. To establish that the left-hand and right-hand networks are equivalent, at least near one frequency, the impedances and derivatives must be matched. For the circuit in Figure 10-78(a),

$$Z_1 = \jmath\frac{\omega^2 LC - 1}{\omega C} \tag{10.224}$$

$$\frac{dZ_1}{d\omega} = \jmath\frac{\omega^2 LC + 1}{\omega^2 C}, \tag{10.225}$$

and for the circuit in Figure 10-78(b),

$$Z_2 = \jmath\frac{\omega^2 L_1 C_1 - 1 + \omega^2 L_1 C_0}{\omega C_0 \left(1 - \omega^2 L_1 C_1\right)} \tag{10.226}$$

$$\frac{dZ_2}{d\omega} = \jmath\frac{\omega^4 L_1^2 C_1^2 - 2\omega^2 L_1 C_1 + \omega^4 L_1^2 C_0 C_1 + \omega^2 L_1 C_0 + 1}{\omega^2 C_0 \left(\omega^2 L_1 C_1 - 1\right)^2}. \tag{10.227}$$

Figure 10-79   Intermediate bandstop filter prototype.

Figure 10-80   Equivalence of shunt bandpass resonator to shunt short-circuited stub.

Equating the above equations enables C_0 and C_1 to be found for a chosen value of L_1. This results in the circuit of Figure 10-79.

At this stage the bandpass resonators are then equated to short-circuited stubs (see Figure 10-80) through the following relationships:

$$Y_1 = \jmath \frac{\left(\omega^2 C L - 1\right)}{\omega L} \tag{10.228}$$

$$Y_2 = \frac{1}{\jmath Z_0 \tan\left(\frac{\pi}{2}\frac{\omega}{\omega_r}\right)}. \tag{10.229}$$

(a)

(b)

(c)

(d)

(e)

Figure 10-81 Bandstop filter prototype using stub approximations. The stub in (b) is the transmission line approximation of the parallel resonant circuit in (a). The stub in (d) approximates the circuit in (c).

The above equations can be solved to obtain ω_r for a selected Z_0. The impedance, Z_0, is selected so that the frequency, ω_r, is not too far away from the upper band-edge frequency of the filter, in this case 1.05 GHz. This results in the bandstop filter prototype with stubs shown in Figure 10-81. The final physical layout of the bandstop filter is shown in Figure 10-82 with the response shown in Figure 10-83.

10.17   Alternative Bandpass Filter Topologies

In the previous sections it was seen that the essential structure of a bandpass filter comprises resonators that are coupled to each other. The combline filter considered previously is one of the commonly used filter topologies and here the coupling is the coupling of parallel lines. Many other types of coupling structures can be used. Figure 10-84 shows another distributed bandpass filter topology utilizing end-coupled microstrip resonators. For these end-coupled resonator filters the coupling must be as strong as possible, so the gaps are usually much smaller than the substrate height.

Figure 10-82 Physical layout of a bandstop filter in microstrip.

Figure 10-83 Response of the bandstop filter shown in Figure 10-82.

Figure 10-84 General microstrip layout for an end-coupled bandpass filter (series coupling gaps between cascaded straight resonator elements).

Filter design using conventional resonators requires several optimally coupled resonators and this is also true with **Dielectric-Resonator** (**DR**)-based filters. A general arrangement of a DR-based filter is shown in Figure 10-85 [160]. The pucks shown are cylinders of high-permittivity

Figure 10-85   A microstrip coupled dielectric resonator bandpass filter configuration.

Figure 10-86   Microstrip layout of a seven-section, parallel coupled bandpass filter.

material (typically having a relative permittivity of 500 to 85,000) with an approximate magnetic wall at the cylindrical surface of the puck. Thus the puck resonates when its diameter is approximately $\lambda/2$.[4]

A variation on parallel coupled resonators is shown in Figure 10-86 [161]. This filter uses edge-coupled microstrip resonators that are $\lambda/2$ long (at the center frequency). All PCL filters, including the combline filter, have spurious passbands [119,157,162]. The root cause of the spurious responses derives from the transformation of the parallel LC resonators into their transmission line form.

10.18   Active Filters

With active circuits, losses in lumped-element components can be compensated for by the gain of active devices resulting in filters with high Q. Inverters can also be realized efficiently. The net result is that active filters provide a performance per unit area advantage over passive element realizations. Of course, active filters are limited to low-level signals (such as in the receive circuitry and prior to the power amplifier stages) as the full power of the signal must be handled by the active devices in the active

---

[4]   A better estimate is developed from the zeroes of Bessel functions, as the fields inside the pucks have a Bessel function dependence (this is the form of the solutions of the wave equation in cylindrical coordinates).

filter. With active devices, additional noise will be introduced and so the noise figure of an active filter must be considered at RF where bandwidths are significant.

Active filters based on traditional low-frequency concepts have relatively low Q. Higher Q and narrower-band applications require active inductors or distributed techniques. With these techniques, negative resistances can be used to compensate for losses. The smallest active filters are the ones that use active inductors and the largest use distributed transmission line elements.

10.18.1 Radio Frequency Active Filters

There are several problems with lumped-element filters. High-order filters require many poles and zeros which must be precisely placed, requiring small tolerances. With active filters it is possible, to some degree, to achieve higher-order filters with fewer reactive elements and so the tolerancing problem is considerably reduced. Also, there are a few designs that can be easily tuned electronically. Lumped-element bandpass filters and highpass filters require inductors. Inductors are a major problem, as RF inductors are lossy and only relatively small values can be obtained. On-chip inductors take considerable area and can thereby dominate the cost of RFICs. With active filters, feedback and capacitors can be used to realize inductor-like characteristics, defined as a positive imaginary impedance increasing with frequency at least over a small frequency range. Many times, and especially at lower radio frequencies, lumped inductors are no longer required. The size of capacitors in filters is another issue, and using feedback and gain, the effective value of capacitors can be magnified. All this comes with limits. The principle limit is that the cutoff frequency of the transistors must be much larger than the operating frequency of the filters. This is a cost whenever feedback is used. A rough rule of thumb is that the transistor cutoff frequency should be about 10 times the operating frequency. So for a 5 GHz bandpass filter, a 50 GHz transistor process is required. The other significant cost is that active filters are only suitable for small signals. The efficiency, a measure of the amount of DC power used in producing the RF output signal, is then not a consideration. Active filters cannot be used where signal levels are likely to be large (such as at the output of a power amplifier or at the receiver input due to out-of-band signals).

The essential aspect of low-RF active filters is the use of feedback, with reactive feedback elements, to realize a frequency-domain transfer function with poles and zeros. The aim is not to synthesize effective capacitors and inductors, but to focus on the overall response. Active filters are generally developed as stages, with each stage realizing a second- or higher-order transfer function. Design is simulation driven, as this is the only way to account for the active device parasitics.

In operational amplifier design, feedback is used to achieve gain stability,

Figure 10-87 Active twin-T bandpass filter: center frequency $f_0 = \pi RC/4$ and $Q = (G+1)/4$.

but at the cost of reduced bandwidth. If reactive feedback elements are used, high-order transfer functions can be obtained with just a few reactive elements. In Figure 10-87, a twin-T notch network is connected in the operational amplifier feedback path to obtain a bandpass filter [163]. Away from the notch frequency the feedback path becomes high impedance and the overall amplifier gain falls off. As the signal frequency approaches the notch frequency the feedback path becomes effective and the amplifier gain increases. The notch frequency, f_n, is proportional to the RC product of the feedback components and the Q is proportional to the amplifier gain. The governing equations for this filter are

$$f_n = 1/(2\pi RC) \qquad \text{and} \qquad Q = (1+G)/4. \qquad (10.230)$$

Active filters are more commonly realized using operational **transconductance amplifiers (OTAs)**, the schematic of which is shown in Figure 10-88(a). Here $I_O = G_M V_{\text{IN}}$, which is also the basic characteristic of an individual transistor. These circuits have characteristics typical of operational amplifiers, including gain stability and resistance to component tolerance variations. OTA-based filters have the advantage of easy tunability, with pole an zero frequencies electronically adjustable. An RF bandpass **biquad** filter is shown in Figure 10-88(b).

10.18.2 Biquadratic Filters

Biquadratic active filter stages are particularly common. These are cascaded to realize high-order filters. A particular form of lowpass filter is described by

$$H(s) = \frac{N(s)}{D(s)} = \frac{a_2 s^2 + a_1 s + a_0}{s^2 + b_1 s + b_0} = \frac{a_2(s+z_1)(s+z_2)}{(s+p_1)(s+p_2)}. \qquad (10.231)$$

This is referred to as a biquadratic function and filters that implement this function are called biquadratic filters or simply biquad filters. There are

Figure 10-88 Operational transconductance amplifier: (a) schematic; and (b) an RF biquad bandpass filter.

several special types of biquad filters.

The form of the biquadratic function for a second-order lowpass filter is

$$H_{\text{LP}}(s) = \frac{a_0}{s^2 + b_1 s + b_0} = \frac{K\omega_p^2}{s^2 + (\omega_p/Q_p)s + \omega_p^2}. \tag{10.232}$$

$H_{\text{LP}}(s)$ has a double zero at $s = \infty$ and thus negligible response at very high frequencies.

The highpass form of the biquadratic filter is described by

$$H_{\text{HP}}(s) = \frac{a_2 s^2}{s^2 + b_1 s + b_0} = \frac{K s^2}{s^2 + (\omega_p/Q_p)s + \omega_p^2}. \tag{10.233}$$

The response of $H_{\text{HP}}(s)$ at high frequencies is K and the response at very low frequencies goes to zero.

The bandpass form of the biquadratic filter is described by

$$H_{\text{BP}}(s) = \frac{a_1 s}{s^2 + b_1 s + b_0} = \frac{K(\omega_p/Q_p)s}{s^2 + (\omega_p/Q_p)s + \omega_p^2}. \tag{10.234}$$

The response of $H_{\text{BP}}(s)$ at high and low frequencies goes to zero, and only at and near the center frequency, $\omega = \omega_p$, is there a reasonable response.

The band-reject or notch form of the biquadratic filter is described by

$$H_{\text{BR}}(s) = \frac{a_2 s^2 + a_0}{s^2 + b_1 s + b_0} = \frac{K(s^2 + \omega_z^2)}{s^2 + (\omega_p/Q_p)s + \omega_p^2}. \tag{10.235}$$

The response of $H_{\text{BR}}(s)$ at high and low frequencies is high, but there is a double zero at the notch frequency, $\omega = \omega_z$, where the response is very low.

With all of the biquadratic filters, the sharpness of the response is determined by Q_p. The edge frequency or center frequency is ω_p for the

Figure 10-89 An active coupled distributed filter: (a) schematic; and (b) equivalent circuit. After Chang and Itoh [166].

bandpass, lowpass, and highpass filters, and the notch frequency is ω_z for the bandstop filter. Biquadratic filters are an important class of filters as they are the basic building block used in active filters. Even when they are used as passive filters, they describe second-order filters which are about the best that can reasonably be obtained using lumped elements at RF.

10.18.3 Distributed Active Filters

At high microwave frequencies, good results can be obtained by combining distributed elements with a gain stage. Commonly the active devices are used as coupling devices. A negative resistance can be produced using the active devices that can compensate for the losses associated with the transmission line elements or lumped passive elements in the filter. The concept uses a resonant tank circuit formed by a transmission line or lumped-element resonator with a negative resistance coupled into the tank circuit by the active devices [164–166]. This concept is illustrated in Figure 10-89. Here a one-half wavelength resonator forms a resonant circuit (commonly called a tank circuit in this context). A negative resistance is coupled into the tank circuit to compensate for losses in the resonator and the result is effectively a lossless tank circuit.

An active resonator circuit is shown in Figure 10-90. The design of the lossless resonator is based on the negative resistance obtained when the capacitor is connected at the source of the MESFET. This element in the source provides a feedback path between the output of the circuit, the drain-to-ground voltage, and the input gate-to-source voltage. An increase in the drain-source current leads to a voltage at the source that changes

Figure 10-90   An active resonator circuit. After Karacaoglu and Robertson [167].

the gate-source voltage. This induces a negative resistance that is adjusted through device selection and feedback capacitance, C_{FB}, to compensate for inductor losses. An additional inductor, L_P, is added at the gate. The inductor resonates with the series combination of C_{FB} and C_{GS}. C_{FB} can be implemented using a varactor diode to enable electronic tuning.

10.19 Summary

Radio frequency and microwave filter design combines the synthesis of required performance and the intuitive realization that particular structures have a desired frequency selectivity. The synthesis approach provides design insight and exploitation of all parameters of a transfer function leading to optimum network topologies. Synthesis can be time consuming and specialized, but it is the only way to develop filters with optimum performance. Insight gain during design identifies yield issues. The other aspect of filter design is identification of particular filter topologies that match the function of the filter being realized. The popular combline filter, for example, exploits the properties of coupled microstrip transmission lines. Once a filter with appropriate topology has been designed, the physical realization is optimized in a circuit simulator to account for parasitics and higher-order EM effects. The same procedure applies to active filter designs where particular circuit topologies, for example, the biquad network blocks, are exploited.

10.20 Exercises

1. Consider the design of a fourth-order lowpass Butterworth filter. This problem follows the development in Section 10.4 on Page 514.

    (a) What is the magnitude-squared characteristic polynomial, $|K(s)|^2$, of the Butterworth filter?

    (b) What is the magnitude-squared transmission coefficient (or transfer function)?

    (c) What is the magnitude-squared reflection coefficient function?

    (d) Derive the reflection coefficient function (i.e., $\Gamma(s)$). Write down the reflection coefficient in factorized form using up to second-order factors.

    (e) What are the roots of the numerator polynomial of the reflection coefficient function?

    (f) What are the roots of the denominator polynomial of the reflection coefficient function?

(g) Identify the conjugate pole pairs in the factorized reflection coefficient.

(h) Plot the poles and zeros of the reflection coefficient on the complex s plane.

2. Synthesize the impedance function

$$Z_x = \frac{s^3 + s^2 + 2s + 1}{s^2 + s + 1}.$$

That is, develop the RLC circuit that realizes Z_x. [Parallels Example 10.2 on Page 522.]

3. Synthesize the impedance function

$$Z_w = \frac{4s^2 + 2s + 1}{8s^3 + 8s^2 + 2s + 1}.$$

That is, develop the RLC circuit that realizes Z_w. [Parallels Example 10.3 on Page 525.]

4. Develop the lowpass prototype of a fifth-order Butterworth lowpass filter. There may be more than one solution. That is, draw the circuit of the lowpass filter prototype with element values.

5. In Section 10.8.2 on Page 534 it was seen that a series inductor can be replaced by a shunt capacitor with inverters and a negative unity transformer. If the inverter is realized with a one-quarter wavelength long transmission line of characteristic impedance 50 Ω:

(a) Derive the $ABCD$ parameters of the cascade of Figure 10-16(c) with 50 Ω inverters.

(b) What is the value of the shunt capacitance in the cascade required to realize a 1 nH inductor?

6. In Section 10.8.3 on Page 536 it was seen that a series capacitor can be replaced by a shunt inductor with inverters and a negative unity transformer. Consider that the inverters are realized with a one-quarter wavelength long transmission line of characteristic impedance 100 Ω.

(a) Derive the $ABCD$ parameters of the cascade of Figure 10-17 with the 100 Ω inverters.

(b) What is the value of the shunt inductance in the cascade required to realize a 1 pH capacitor?

7. A series inductor of 10 pH must be realized by an equivalent circuit using shunt capacitors and sections of one-quarter wavelength long 1 Ω transmission line. Design the equivalent circuit. [Hint: The one-quarter wavelength long lines are impedance inverters.]

8. A series inductor of 10 nH must be realized by an equivalent circuit using shunt capacitors and sections of one-quarter wavelength long 50 Ω transmission line. [Hint: The one-quarter wavelength long lines are impedance inverters.] Design the equivalent circuit.

9. A 0.04 S admittance inverter is to be implemented in microstrip using a single length of transmission line. The effective permittivity of the line is 9 and the design center frequency is 10 GHz.

(a) What is the characteristic impedance of the transmission line?

(b) What is the wavelength in millimeters at the design center frequency in free space?

(c) What is the wavelength in millimeters at the design center frequency in microstrip?

(d) What is electrical length of the microstrip transmission line in degrees at the design center frequency?

(e) What is the length of the microstrip transmission line in millimeters?

10. A 50 Ω impedance inverter is to be realized using three resonant stubs. The center frequency of the design is f_0. The resonant frequency of the stubs is $f_r = 2f_0$.

(a) Draw the circuit using shorted stubs. On your diagram indicate the input impedance and characteristic impedance of each of the stubs if $f_r = 2f_0$.

(b) What is the input impedance of a shorted one-eighth wavelength long transmission if the characteristic impedance of the line is Z_1.

(c) What is the input impedance of an open-ended one-eighth wavelength long transmission if the characteristic impedance of the line is Z_2.

(d) What is the length of each of the stubs in the inverter in terms of the wavelengths at the resonant frequency f_0?

11. At 5 GHz, a series 5 nH inductor is to be realized using one or more 75 Ω impedance inverters, a unity transformer, and a capacitor. What is the value of the capacitor?

12. This problem considers the design of a Butterworth or maximally flat bandpass filter at 900 MHz.

(a) Design an LC second-order Butterworth lowpass filter with a corner frequency of 1 radian/s. (The second-order Butterworth polynomial in s is $c_2 s^s + c_1 s + c_0$, where $c_2 = 1.000$, $c_1 = 1.414$, and $c_0 = 1.000$.) This filter is known as a filter prototype because of the normalized 1 Hz corner frequency.

(b) Using the above filter prototype, design a lowpass filter with a corner frequency of 1800 MHz.

(c) Design a second-order Butterworth bandpass filter at 900 MHz using the above lowpass filter prototype. Use a transformation constant $a = 0.1$. What is the 3 dB bandwidth of this filter?

13. Design a third-order maximally flat bandpass filter prototype in a 50 Ω system centered at 1 GHz with a 10% bandwidth. The lowpass prototype of a maximally flat filter was developed in Example 102 on page 522 The lowpass prototype is shown in Figure 10-10. The problem will parallel the development in Section 10.14 on Page 566 and the end result of this development will be a bandpass prototype filter with the form of that in Figure 10-63. Note that there will be differences as the filter is a different type.

(a) Convert the prototype lowpass filter to a lowpass filter with inverters and capacitors only; that is, remove the series inductors.

(b) Scale the filter to take the corner frequency from 1 rad/s to 1 GHz.

(c) Transform the lowpass filter into a bandpass filter. That is, replace each shunt ca-

pacitor by a parallel LC network. This step will establish the bandwidth of the filter. (The circuit will now be in a form similar to that in Figure 10-55.)

(d) Transform the system impedance of the filter from 1 to 50 Ω. (The circuit will now be in a form similar to that in Figure 10-56.)

(e) Replace the parallel LC circuits by short-circuited stubs in parallel with lumped capacitors. (The circuit will now be in a form similar to that in Figure 10-59.)

(f) For each inverter, derive the three lumped-element equivalent circuit as in Figure 10-20. Do not update the filter prototype yet, but instead draw and label the lumped-element equivalent circuits of each inverter.

(g) Replace each inverter in the bandpass filter prototype by the lumped-element equivalent circuit derived above, but now update the stubs with their new impedances. The final prototype will now have the form of that in Figure 10-63.

14. The lowpass prototype of a lowpass Butterworth filter is shown below.

$$L_1 = 0.765369 \text{ H} \qquad C_2 = 1.847759 \text{ F}$$
$$L_3 = 1.847759 \text{ H} \qquad C_4 = 0.765369 \text{ F}$$

Based on this, develop a fourth-order Butterworth bandpass filter prototype centered at 10^9 rad/s with a fractional bandwidth of 5%.

(a) Scale the lowpass prototype to have a corner frequency of 10^9 rad/s. Draw the prototype with element values.

(b) Draw the schematic of the lumped-element fourth-order Butterworth bandpass prototype based on the above lowpass filter prototype.

(c) Derive the element values of the lumped-element bandpass filter prototype.

15. Design a third-order type 2 Chebyshev high-
    pass filter with a corner frequency of 1 GHz,
    a system impedance of 50 Ω, and 0.2 dB rip-
    ple. There are a number of steps in the design,
    and to demonstrate that you understand them
    you are asked to complete the table below. For
    each stage of the filter synthesis you must in-
    dicate whether the element is an inductance or
    a capacitance by writing L or C in the appro-
    priate cell. Other cells require a numeric value
    and you must include units. The X element is
    identified in the prototype below. A Cauer 2
    lowpass filter prototype is shown with ω_c be-
    ing the corner radian frequency, $f_c = \omega_c/(2\pi)$
    being the corner frequency, and Z_0 being the
    system impedance.

(a) Complete the LPF (lowpass filter) column
    of the table with $\omega_c = 1$ rad/s, $Z_0 = 1\,\Omega$.

(b) Complete the HPF (highpass filter) col-
    umn of the table with $\omega_c = 1$ rad/s, $Z_0 = 1\,\Omega$.

(c) Complete the second HPF column of the
    table with $f_c = 1$ GHz, $Z_0 = 1\,\Omega$.

(d) Complete the third HPF column of the ta-
    ble with $f_c = 1$ GHz, $Z_0 = 50\,\Omega$.

ELEMENT	LPF $\omega_c = 1$ rad/s, $Z_0 = 1\,\Omega$		HPF $\omega_c = 1$ rad/s, $Z_0 = 1\,\Omega$	
	L or C	Value (units)	L or C	Value (units)
X_1				
X_2				
X_3				

ELEMENT	HPF $f_c = 1$ GHz, $Z_0 = 1\,\Omega$		HPF $f_c = 1$ GHz, $Z_0 = 50\,\Omega$	
	L or C	Value (units)	L or C	Value (units)
X_1				
X_2				
X_3				

Amplifiers

11.1   Introduction

In RF and microwave circuits amplifiers increase the power of an RF signal by converting DC power to AC power. Amplifiers can be optimized for low noise, moderate to high gain, or for output power. In this chapter the major active devices and amplifiers based on them are examined. A critical common aspect is minimizing noise, maximizing the efficiency of power conversion to RF, and minimizing insertion loss. Matching networks and hybrids are critical aspects of amplifier design. The most significant change

Figure 11-1 Transistor schematics: (a) pnp bipolar transistor with B for base terminal, C for collector terminal, and E for emitter terminal; (b) n-type MOSFET (nMOS) and (c) n-type JFET (nJFET) with G for gate terminal, D for drain terminal, and S for source terminal. The schematic symbol for a BJT is used for HBTs; and the schematic symbol for a JFET is used for MESFETs, HEMTS, and pHEMTs.

in RF and microwave engineering has been the increasing importance of differential circuits. In part this is because they are very conveniently implemented in silicon technology. It is also a result of the use of monolithic technology and the maturity of semiconductor technologies resulting in very repeatable RF active components.

11.2 Transistor Technology

11.2.1 Transistor Types

Transistors are semiconductor devices with three (and sometimes more) terminals. The third terminal enables output current to be controlled by a relatively small and low-power input signal. In amplifiers, transistors are used to achieve current gain, voltage gain, or power gain. Most often power gain is the objective in RF and microwave design. Most transistors are fabricated using silicon or **compound semiconductors** most often based on **gallium-arsenide** (**GaAs**), **indium phosphide** (**InP**), or **gallium-nitride** (**GaN**). However, the overwhelming trend is to use silicon technology because of the much higher integration density that is possible, with compound semiconductor technology used only when it provides a unique advantage such as high power, good noise performance, or efficient conversion of DC power to RF power. There are three fundamental types of transistors [168, 170]: **Bipolar Junction Transistors**, **BJTs**; **Junction Field Effect Transistors**, **JFETs**; and **Insulated Gate FETs**, **IGFETs,** with the **metal-oxide-semiconductor FETs**, **MOSFETs**, being the most common type of IGFET. The schematics and terminal definitions of the three fundamental types of transistors are shown in Figure 11-1. The three fundamental types of transistors are considered in the following subsections.

Figure 11-2   BJT details: (a) crosssection; (b) crosssection with current flow path; (c) circuit model of fundamental model; and (d) Gummel-Poon model schematic.

11.2.2   BJT and HBT Fundamentals

A bipolar transistor has three semiconductor regions called the collector (C), base (B), and emitter (E) as shown in the BJT crosssection of Figure 11-2(a). An npn BJT has n-type semiconductor at the emitter and collector, and p-type semiconductor forms the base. In this transistor, the positive sense of current flow is from the collector through the base to the emitter (see Figure 11-2(b)) and the dominant carriers in the p-type base region are electrons, and so this is called a **minority carrier device**. The collector current is dependent on the number of carriers injected into the base region from the base terminal. In a pnp BJT the collector, base, and emitter are p-type, n-type, and p-type respectively, and the majority carriers in the base are holes. Current flow is then from the emitter through the base to the collector. If the base region is thin the collector current, I_C, is much greater than I_B with $I_C = \beta_F I_B$ where β_F is called the forward current gain and commonly has a value of several hundred. The key to high performance is a thin base region. When realized in silicon, a bipolar transistor is called a bipolar

junction transistor, **BJT**; and in compound semiconductor technology a **heterostructure bipolar transistor**, **HBT**.

The fundamental operation of a BJT transistor was described by Gummel and Poon [171] using equations that are now implemented in circuit simulators and known as the Gummel-Poon model. The circuit schematic of the Gummel-Poon model is shown in Figure 11-2(d). It is the basis for more sophisticated BJT and HBT models. Both hole and electron charge carriers are involved in current conduction, hence the term bipolar. The Gummel-Poon model is described in Section G.3 on Page 909 and the fundamental operation is described by Equations (G.70)–(G.80). The base-emitter current is

$$I_{BE} = I_{BF}/\beta_F + I_{LE} \tag{11.1}$$

and the base-collector current is

$$I_{BC} = I_{BR}/\beta_R + I_{LC}, \tag{11.2}$$

where β_R is the reverse current gain. The collector-emitter current is

$$I_{CE} = I_{BF} - I_{BR}/K_{QB}. \tag{11.3}$$

The forward diffusion current is

$$I_{BF} = I_S \left(e^{V_{BE}/(N_F V_{TH})} - 1 \right), \tag{11.4}$$

the nonideal base-emitter current is

$$I_{LE} = I_{SE} \left(e^{V_{BE}/(N_E V_{TH})} - 1 \right), \tag{11.5}$$

the reverse diffusion current is

$$I_{BR} = I_S \left(e^{V_{BC}/(N_R V_{TH})} - 1 \right), \tag{11.6}$$

the nonideal base-collector current is

$$I_{LC} = I_{SC} \left(e^{V_{BC}/(N_C V_{TH})} - 1 \right), \tag{11.7}$$

and the base charge factor is

$$K_{QB} = \frac{1}{2} \left[1 - \frac{V_{BC}}{V_{AF}} - \frac{V_{BE}}{V_{AB}} \right]^{-1} \left(1 + \sqrt{1 + 4 \left(\frac{I_{BF}}{I_{KF}} + \frac{I_{BR}}{I_{KR}} \right)} \right). \tag{11.8}$$

Thus the conductive current flowing into the base is

$$I_B = I_{BE} + I_{BC}, \tag{11.9}$$

the conductive current flowing into the collector

$$I_C = I_{CE} - I_{BC}, \tag{11.10}$$

and the conductive current flowing into the emitter

$$I_E = I_{BE} + I_{CE}. \tag{11.11}$$

The forward current gain, β_F, is much greater than the reverse current gain, β_R, and the nonideal base-emitter and base-collector currents are small. Equations (11.1)–(11.11) can then be reduced so that the base current is approximately

$$I_B = \frac{I_S}{\beta_F} \left(e^{V_{BE}/(N_F V_{TH})} - 1 \right), \tag{11.12}$$

the conductive current flowing into the collector is

$$I_C = \beta_F I_B, \tag{11.13}$$

and the conductive current flowing into the emitter is

$$I_E = I_B + I_C. \tag{11.14}$$

From Equation (11.13) it is seen that the fundamental operation of a BJT is as a current-controlled current source. This is the way to view the device when biasing a BJT circuit. For analysis and modeling purposes it is more convenient to view a BJT transistor as comprising a voltage-controlled current source. The circuit model of the fundamental operation a BJT is as shown in Figure 11-2(d), where the resistor r_π relates the base voltage and current.

The schematic symbols used for BJTs are shown in Table 11-1 with the arrow pointing to the n-type semiconductor. The BJT symbol is also the symbol for a HBT.

11.2.3 MOSFET Fundamentals

There are several types of FETs and the MOSFET is the most common. With all FETs there is a channel between two terminals, the source and drain, and an applied field produced by a voltage at a third terminal, the gate, controls the crosssection of the channel, and sometimes the number of carriers in the channel, and hence controls current flow between the drain and the source. With some FETs, the channel does not exist until a gate field is applied and pulls carriers into the channel, and this is called an **enhancement-mode** FET. Sometimes carriers are in the channel even without an applied field and a gate voltage either enhances the crosssection of the channel or closes it off. Most often the gate voltage is used to reduce current conduction, and this type of FET is called a **depletion-mode** FET.

Table 11-1 IEEE standard schematic symbols for bipolar junction transistors (BJTs and HBTs) [117] and commonly used symbols in layouts [174]. The letters indicate terminals: B (base), C (collector), E (emitter). These symbols are used for silicon BJTs and compound semiconductor HBTs.

Transistor	IEEE Symbol	Commonly Used Symbol
BJT, pnp		
BJT, npn		

The enhancement MOSFET is a relatively simple device to fabricate and is the smallest of the semiconductor transistors. It is the preferred technology for high-density integration. The three-dimensional view and crosssections of a MOSFET are shown in Figures 11-3(a) and 11-3(b) for an nMOS transistor. A MOSFET has metal or polysilicon (a reasonable conductor [168–170]) connections at the drain (D), source (S) and gate (G). The MOSFET is nearly always silicon, but possibly GaN [172, 173]. The source and drain connections are highly doped (n+ for the nMOSFET of Figure 11-3(b)) semiconductor regions providing a good ohmic contact rather than forming a Schottky barrier.[1] The gate is not in direct contact with the semiconductor but separated by a thin layer of oxide. With no voltage applied at the gate there are no carriers below the gate oxide that can conduct current between the source and drain. A gate voltage is necessary to draw carriers to the channel region, forming a conducting channel. That is, a voltage applied to the gate creates an electric field that induces electrons (the n carriers for an nMOSFET) to form a conducting channel immediately under the oxide.[2] This process is called inversion. The length of the channel is denoted L_{eff} (the effective gate length), which is less than the gate length L as the highly doped source and drain regions must extend under the gate to ensure good contact to the induced channel. The number of carriers in the channel is controlled by the gate voltage. Higher frequency of operation is obtained by reducing L_{eff}.

Three distinct regions of operation, identified in Figures 11-3(d), are

---

[1] A Schottky barrier occurs at the abrupt interface between a metal and a doped semiconductor, and forms a rectifying diode called a Schottky diode. The Schottky diode is considered in Section 12.2.1 on Page 708.

[2] The discussion is similar for a pMOSFET, but with holes (p-type carriers) forming the channel.

Figure 11-3 MOSFET details: (a) three-dimensional view of a MOSFET; (b) crosssection with metal or **polysilicon** contacts indicated by the black blocks; (c) crosssection illustrating the relationship of the gate length, L, to the effective gate length, L_{EFF}, defined by the channel length (the reduction in effective length results from the diffusion of the highly doped (indicated by n+) source and drain regions; (d) current-voltage characteristics of an enhancement-mode MOSFET; and (e) circuit model of fundamental operation. The **linear region** is sometimes (but less often) called the **triode region** because of similarity to the characteristics of the triode vacuum tube device. Similarly the **saturation region** is also called the **pentode region**.

recognized for a MOSFET. In the linear region the drain-source current, I_{DS}, continues to increase as the drain-source voltage, V_{DS}, increases. I_{DS} depends on both the drain-source and gate-source voltage, V_{DS} and V_{GS}, so the linear region is sometimes exploited in mixers. In the saturation region, I_{DS} is almost independent of V_{DS} and almost solely controlled by V_{GS}. Amplifiers operate in the saturation region. The cutoff region is when there is no drain current, and a FET is particularly good at shutting off conduction and so makes a good voltage-controlled switch.

In design, the mode of fundamental operation must be intuitively understood and so simple models and equations are needed. In contrast, a circuit simulator requires a detailed model capturing subtle physical effects as it is used to validate and optimize a design. A model of a MOSFET that can be used in a circuit simulator is presented in Section G.1 on Page 901. The model presented is known as the Level 3 MOSFET model and captures the fundamental operation of MOSFETs as well as capacitive parasitic

effects. Models are developed using physical insight into semiconductor operation with compromises made for the limited model support of circuit simulators. All semiconductor device models, not just MOSFETs, require extensive fitting to measured data and have limited accuracy. Consequently the design, fabrication, and test cycle are critically important to realizing transistor circuits.

In the saturation region (see Figure 11-3(d)) the fundamental operation of a MOSFET is described by Equation (G.36), which is repeated here:

$$I_{DS} = \frac{W_{\text{eff}}}{L_{\text{eff}}}\, \mu_{\text{eff}}\, C_{ox} \left[(V_{GS} - V_{th}) - 1 + \frac{F_B}{2}\, V_{d\text{sat}} \right] V_{d\text{sat}}\,. \qquad (11.15)$$

Here C_{ox} is the capacitance of the gate oxide, W_{eff} is the effective gate width which is the gate width W modified by fringing and related effects, V_{th} is the **threshold voltage**, and μ_{eff} is the effective mobility[3] of the carriers in the channel (electrons for an nMOSFET and holes for a pMOSFET). $V_{d\text{sat}}$ is the drain saturation voltage and is the drain source voltage at which the device enters the saturation region from the linear region. F_B is the charge in the bulk semiconductor (below the channel) on which the gate-induced electric field terminates. L_{eff} is the effective gate length and this is modulated by the drain-source voltage so that [168–170]

$$L_{\text{eff}} = \frac{L}{1 + \lambda V_{DS}}. \qquad (11.16)$$

Accounting for channel length modulation, described by Equation (11.16), and simplifying [168–170], Equation (11.15) becomes

$$I_{DS} = \frac{W}{L} \frac{\mu_{\text{eff}} C_{ox}}{2} (V_{GS} - V_{th})^2 (1 + \lambda V_{DS})\,. \qquad (11.17)$$

This equation embodies the basic fundamentals we need in developing designs. The key is that the MOSFET can be modeled (at least in the saturation region) as a voltage-controlled current source as shown in the model of Figure 11-3(e). The transconductance, g_m (in saturation), is obtained by differentiating Equation (11.17) so that (ignoring channel length modulation)

$$g_m = \frac{W}{L}\, \mu_{\text{eff}} C_{ox} (V_{GS} - V_{th})\,. \qquad (11.18)$$

This can also be written as

$$g_m = \sqrt{\frac{W}{L} 2 \mu_{\text{eff}} C_{ox} I_{DS}}. \qquad (11.19)$$

---

[3] **Mobility**, μ, is the proportionality of the velocity of carriers to the applied electric field, $v_d = \mu E$, where v_d is the average drift velocity of carriers and E is the applied electric field. Mobility has the units $\text{m}^2/(\text{V} \cdot \text{s})$ or $\text{m}^2 \cdot (\text{V} \cdot \text{s})^{-1}$.

Figure 11-4 Current-voltage characteristics of depletion-mode and enhancement-mode MOSFETs: (a) enhancement-mode; and (b) depletion-mode.

Generally the gate length L is fixed at the minimum supported by a particular process, as this provides the highest frequency of operation. However, both L and W can be selected to control the current, I_{DS}. For example, if $V_{GS} = 0$ (the gate and source are connected), then the MOSFET acts as a current source, with the value of the current adjusted by setting L and W provided that there is sufficient V_{DS}.

The current-voltage characteristics shown in Figure 11-3(d) are those of an enhancement-mode MOSFET which requires the simplest processing. Applying a gate-source voltage enhances the channel and increases I_{DS}. With additional processing [168–170,175] a depletion-mode MOSFET can be fabricated so that the channel exists even without an applied gate voltage. The same equations are used to describe operation with the threshold voltage changed. I_{DS} increases as the gate-source voltage increases, and it reduces as the gate-voltage becomes negative. The contrast between enhancement-mode and depletion-mode MOSFETs is illustrated in Figure 11-4.

The voltage of the bulk semiconductor affects the operation of a MOSFET and is a fourth terminal controlling drain-source conduction but has much a smaller effect than the gate does. Most often the bulk is connected electrically to the most negative voltage in a circuit for an nMOSFET and to the most positive voltage for a pMOSFET. The standard schematic symbols of MOSFETs are shown in Table 11-2.

Table 11-2 IEEE standard schematic symbols for MOSFET transistors [117] and symbols more commonly used in schematics [174]. The MOSFET symbols are for enhancement- and depletion-mode transistors. The letters indicate terminals: G (gate), D (drain), S (source), U (bulk). Four terminal and three terminal common symbols are shown. The three-terminal common symbol is most often used when the bulk is connected to the most negative connection in the circuit, and the pMOSFET symbol is used when the bulk is tied to V_{DD} (the most positive connection).

Transistor	IEEE Symbol	Commonly Used Symbol (3 terminal)	Commonly Used Symbol (4 terminal)
FET, nMOS, depletion			
FET, pMOS, depletion			
FET, nMOS, enhancement			
FET, pMOS, enhancement			

Figure 11-5 JFET details: (a) crosssection; and (b) circuit model of fundamental operation.

11.2.4 MESFET, HEMT, and JFET Fundamentals

The MESFET and HEMT are types of JFETs fabricated using compound semiconductors, with JFET most commonly referring to silicon devices only. The crosssection of a JFET is shown in Figure 11-5(a), where the depth (crosssection) of the conducting channel is varied by the thickness of the depletion region of a reverse-biased junction. With the silicon JFET, the voltage applied to the gate terminal changes the amount of reverse bias and hence the depletion region thickness. Increased reverse bias reduces the crosssection of the current-carrying channel. A JFET looks like a variable

Figure 11-6 Current-voltage characteristics of depletion-mode and enhancement-mode JFETs: (a) enhancement mode; and (b) depletion mode.

conductance. The controlling field of the FET is created at the reverse-biased pn junction at the gate terminal. The term JFET most commonly refers to a silicon junction FET. With compound semiconductors such as GaAs, the pn junction of a silicon JFET is replaced by a Schottky barrier junction and the transistor is called a **metal-epitaxy-semiconductor FET** (**MESFET**). A device similar to the MESFET is the **high electron mobility transistor** (**HEMT**), HEMT, where the field is established at the junction of two compound semiconductor materials with different band gaps (i.e. a heterojunction) which forms the channel instead of a doped region. The HEMT is also called the **heterostructure FET** (**HFET**). A MESFET with a graded junction is called a **modulation-doped FET** (**MODFET**). A **pseudomorphic HEMT** (**pHEMT**) has an extremely thin layer establishing the channel so that the crystal structure stretches and a very high bandgap is established. Enhancement-mode and depletion-mode JFETs are contrasted in Figure 11-6.

The Materka-Kacprzak transistor model was developed for GaAs MESFET transistors [176] but is used to model silicon JFET and HEMT transistors as well. The model is described in Section G.2 on Page 907 and the fundamental operation is described by Equation (G.62), which is repeated

Table 11-3 IEEE standard schematic symbols for JFETs (MESFET, HEMT, JFET) [117] and symbols more commonly used in schematics. The letters indicate terminals: G (gate), D (drain), S (source). Four terminal and three terminal common symbols are shown.

Transistor	IEEE Symbol	Commonly Used Symbol
FET, pJFET		
FET, nJFET, MESFET, HEMT		

here without the area multiplier:

$$I_{DS} = I_{DSS} \left[1 + S_S \frac{V_{DS}}{I_{DSS}}\right] \left[1 - \frac{V_{GS}(t-\tau)}{V_{P0} + \gamma V_{DS}}\right]^{(E + K_E V_{GS}(t-\tau))}$$
$$\times \tanh\left[\frac{S_L V_{DS}}{I_{DSS}(1 - K_G V_{GS}(t-\tau))}\right]. \qquad (11.20)$$

Here I_{DSS} is the drain saturation current, and this, along with all quantities in Equation (11.20) other than V_{DS}, V_{GS}, and I_{DS} are constants and specified as inputs by the user (they are defined in Table G-2). Equation (11.20) indicates that the fundamental operation of a JFET is that of a voltage-controlled current source. Thus the circuit model of fundamental operation is as shown in Figure 11-5(b).

The schematic symbols used for the MESFET, HEMT, and JFET are shown in Table 11-3. The only MESFET type used, however, is the n-type as the p-type MESFET has poor performance due to the low mobility of holes.

11.3 Amplifier Design Strategies

11.3.1 Amplifier Topology

A transistor amplifier requires that the transistor(s) be biased and that input and output matching networks be used to provide good power transfer at the input and output of the transistor stages. This circuit arrangement is shown in the general block form of Figure 11-7. The DC biasing circuit is fairly standard; it does not involve any microwave constraints. The lowpass filters (in the bias circuits) can have one of several forms and are often integrated into the input and output matching networks. Synthesis of the input and output matching networks (and occasionally a feedback network required for stability) is the primary design objective of any amplifier.

Figure 11-7 Block diagram of an RF amplifier including biasing networks.

RF transistors used to amplify small signals should have high maximum available gain and low noise characteristics. For transistors used in transmitters, where the efficient generation of power is critical, it is important that the transistor characteristics be close to linear in the central region of the current-voltage characteristics so that distortion is minimized as the RF voltage variations range over a large portion of the current-voltage characteristics. The ultimate limit on output power is determined by the breakdown voltage at high drain-source voltages and also by the maximum current density. Finally, for efficient amplification of large signals, the knee voltage (where the current-voltage curves bend over and start to flatten at low drain-source voltages) should be low.

11.3.2 Amplifier Efficiency

The efficiency of a circuit is the useful output power divided by the input power. Such a measure of efficiency is called **Power-Added Efficiency** (**PAE**). Two different definitions of PAE are used [177]. The more universal definition that can be used with any two-port network is also called the total power added efficiency or transmit chain efficiency and is denoted here as η_{TOTAL} defined as

$$\eta_{\text{TOTAL}} = \frac{P_{\text{RF,out}}}{P_{\text{DC}} + P_{\text{RF,in}}}. \tag{11.21}$$

It is usually expressed as a percentage in which case Equation (11.21) is multiplied by 100. At RF and microwave frequencies, the most common definition of PAE used with power amplifiers focuses on the additional RF power divided by the DC input power. This is designated as η_{PAE} and is defined as

$$\eta_{\text{PAE}} = \frac{P_{\text{RF,out}} - P_{\text{RF,in}}}{P_{\text{DC}}}. \tag{11.22}$$

For high-gain amplifiers, $P_{\text{RF,in}} \ll P_{\text{DC}}$, and both η_{TOTAL} and η_{PAE}

Table 11-4 Comparison of efficiency metrics for an amplifier producing 1 W RF output power and consuming 2 W of DC power with various power gains.

Power gain (dB)	η_{TOTAL}	η_{PAE}	η
3	40%	25%	50%
6	44%	37%	50%
10	48%	45%	50%
15	49%	48%	50%
20	50%	50%	50%
40	50%	50%	50%

reduce to the efficiency η of the amplifier:

$$\eta = \frac{P_{\text{RF,out}}}{P_{\text{DC}}} \approx \eta_{\text{PAE}} \approx \eta_{\text{TOTAL}} \quad \text{(high gain)}. \tag{11.23}$$

These efficiency metrics are compared in Table 11-4 for an amplifier with 1 W RF output power. The first amplifier has a power gain of 3 dB, which is commonly the gain of the final amplifier stage producing the maximum output power available from a particular transistor technology.

11.3.3 Summary

Manufacturers of discrete transistors and amplifier modules provide substantial information including S parameters and, in some cases reference designs. An extract from the datasheet of a pHEMT transistor is shown in Figure 11-8. The intended application for transistors is provided and the device structure has been optimized for the application. This transistor will be used for examples and calculations throughout this chapter.

Design examples presented in this and the next few sections will use the pHEMT transistor documented in Figure 11-8. This discrete transistor described is a low-noise, high-frequency, packaged pHEMT that can be used in fabricating amplifiers operating at up to 18 GHz. It shares a common characteristic of FET devices in that S_{21} is highest at low frequencies and the feedback parameter, S_{12}, is lowest at low frequencies. This means that gain is harder to achieve at higher frequencies and the higher feedback means that stability is often a problem at higher frequencies. However, the loop effect described by $S_{21}S_{12}$ is large at low frequencies so stability is also a problem at low frequencies.

11.4 Class A, AB, B, and C Amplifiers

Transistor amplifiers follow a number of different biasing strategies. The strategies are identified as classes of amplifiers ranging from Class A to Class G, although the letter designation of the higher classes are less commonly accepted. In this section Class A, AB, B, and C amplifiers will be

Data Sheet Extract.

Transistor technology:	Depletion-mode pHEMT
Model:	FPD6836P70 from RFMD, Inc.
Description:	Low-noise, high-frequency packaged pHEMT
	Optimized for low-noise, high-frequency applications.
Synopsis:	22 dBm output power (P1dB)
	15 dB power gain (G1dB) at 5.8 GHz
	0.8 dB noise figure at 5.8 GHz
	32 dBm output IP3 at 5.8 GHz
	45% power-added efficiency at 5.8 GHz
	Useable gain to 18 GHz

Frequency (GHz)	$|S_{11}|$	$\angle S_{11}$ degrees	$|S_{21}|$	$\angle S_{21}$ degrees	$|S_{12}|$	$\angle S_{12}$ degrees	$|S_{22}|$	$\angle S_{22}$ degrees
0.500	0.976	-20.9	11.395	161.5	0.011	78.3	0.635	-11.5
1.000	0.925	-41.3	10.729	145.1	0.021	67.8	0.614	-22.2
2.000	0.796	-78.2	8.842	116.7	0.034	51.4	0.553	-37.9
3.000	0.694	-106.8	7.180	94.5	0.041	40.4	0.506	-48.9
4.000	0.614	-127.3	6.002	76.7	0.044	33.9	0.475	-57.7
5.000	0.555	-147.0	5.249	60.3	0.048	28.4	0.453	-66.4
6.000	0.511	-170.2	4.729	43.7	0.052	23.3	0.438	-76.0
7.000	0.493	163.9	4.261	26.8	0.057	14.0	0.391	-87.6
8.000	0.486	140.4	3.784	11.2	0.057	6.4	0.340	-99.1
9.000	0.473	122.5	3.448	-2.4	0.059	5.2	0.332	-109.6
10.000	0.488	103.4	3.339	-17.3	0.073	0.9	0.355	-124.8
11.000	0.539	79.8	3.166	-35.0	0.086	-10.1	0.349	-145.6
12.000	0.626	60.8	2.877	-51.9	0.095	-21.4	0.307	-169.6
13.000	0.685	47.6	2.604	-68.2	0.100	-32.5	0.295	165.3
14.000	0.724	36.2	2.392	-83.8	0.106	-43.3	0.312	142.7
15.000	0.787	20.9	2.225	-99.7	0.109	-55.1	0.320	125.4
16.000	0.818	5.2	2.067	-116.6	0.112	-68.4	0.340	103.9
17.000	0.831	-9.6	1.855	-134.4	0.108	-83.5	0.373	76.1
18.000	0.852	-19.5	1.603	-148.6	0.103	-94.2	0.406	54.7
19.000	0.815	-20.5	1.440	-159.3	0.102	-103.0	0.449	43.1
20.000	0.780	-26.8	1.382	-171.2	0.106	-113.5	0.460	37.9
21.000	0.779	-46.8	1.333	171.2	0.109	-130.7	0.438	31.4
22.000	0.786	-62.1	1.195	152.0	0.110	-148.4	0.417	6.0
23.000	0.774	-70.1	1.073	137.2	0.108	-162.4	0.428	-16.5
24.000	0.744	-81.7	1.025	123.5	0.112	-175.2	0.433	-29.0
25.000	0.704	-90.9	1.061	107.3	0.132	170.0	0.396	-46.5
26.000	0.677	-111.1	1.065	85.8	0.148	147.8	0.298	-71.0

Figure 11-8 Scattering parameters of an enhancement mode pHEMT transistor biased at $V_{DS} = 5$ V, $I_D = 55$ mA. Extract from data sheet for the model FPD6836P70 discrete transistor from RFMD [178].

considered and these have the basic topology of Figure 11-9 where input and output matching networks have been omitted. The higher-order classes rely on particular terminations at harmonic frequencies and will be considered in

(a)   (b)

Figure 11-9 Class A single-ended amplifiers: (a) BJT transistor with B for base terminal, C for collector terminal, and E for emitter terminal; (b) MOSFET transistor with G for gate terminal, D for drain terminal, and S for source terminal; (c) single-ended BJT Class A amplifier with resistive bias; and (d) single-ended MOSFET Class A amplifier with resistive bias.

(a)   (b)

- QUIESCENT POINT
  - A   Class A AMPLIFIER
  - B   Class A AMPLIFIER
  - AB Class A AMPLIFIER
  - C   Class A AMPLIFIER

Figure 11-10 Current-voltage characteristics of a transistor used in an amplifier showing the quiescent points of various amplifier classes: (a) output characteristic; and (b) input characteristic.

the next section. Class A–C amplifiers have the same impedance presented to the output of the amplifier at the operating frequency and at harmonics. The output and input characteristics of Class A–C amplifiers are shown in Figure 11-10 for an amplifier built around an FET. Figure 11-10(a) is the output characteristic and shows the distinguishing quiescent points for the

various points are shown. The input characteristics are shown Figure 11-10(b) where the input (I_G) and output (I_D) current waveforms are presented for a sinusoidal input waveform (V_{GS}).

The instantaneous drain current, I_D, and drain source voltage, V_{DS}, are the solution to two constraints: one imposed by the nonlinear transistor characteristics, and the other by the load. If the load is R_L then Ohm's law applied to the circuit is

$$I_D = (V_{DD} - V_{DS})/R_L. \tag{11.24}$$

where V_{DD} is the supply voltage (at the top of the load). Equation (11.24) establishes the **loadline** shown in Figure 11-10(a). The conventional loadline passes through the middle of the output characteristics, while the high-efficiency loadline crosses the top $i - v$ curve at the lowest V_{DS} intercept. The reason this results in higher efficiency will be explained in Example 11.1. The quiescent point (the I_D, V_{DS} solution) is the intersection of the loadline and the transistor curves for a particular input voltage, V_{GS}. The Class A amplifier biased in the middle of the output current-voltage characteristics allows for maximum undistorted output voltage swing. However, the quiescent current (the DC current when there is no RF input) is high.

11.4.1 Class A Amplifier

The Class A amplifier has limited efficiency because there is always substantial quiescent current flowing whether or not RF current is flowing. Higher-order classes of amplifiers achieve higher efficiency, but distort the RF signal. The current and voltage locus of Class A, B, AB, and C amplifiers have a similar trajectory on the output current-voltage characteristics of a transistor. The output characteristics of a transistor are shown in Figure 11-10, showing what is called the linear or DC loadline and the bias points for the various amplifier classes. The loadline is the locus of the DC current and voltage as the DC input voltage is varied. With the Class A amplifier, the transistor is biased in the middle of the transistor characteristics where the response has the highest linearity. That is, if the gate voltage varies from an applied signal the output voltage and current variations are nearly linearly proportional to the applied input. The drawback is that there is always considerable DC current flowing, even when the input signal is very small. That is there is DC power consumption whether or not RF power is being generated at the output of the transistor. This is not of concern if small RF signals are to be amplified, as then a small transistor can be chosen so that the DC current levels are small. It is a problem if an amplifier must handle both large and small signals.

The linear amplifier is generally known as a Class A amplifier and is defined by its ability to amplify small to medium and even large signals with minimal distortion. This is achieved by biasing a transistor in the

Figure 11-11 Current-voltage characteristics of transistor amplifiers shown with a Class A amplifier loadline: (a) bipolar amplifier; and (b) FET amplifier.

middle of its I-V (or current-voltage) characteristics. Figure 11-11 shows the I-V characteristics of the bipolar and FET transistors shown in Figure 11-9, together with the DC loadline. The loadline is the locus of the output current and voltage. For the Class A amplifiers in Figure 11-9(a) and 11-9(b) the loadlines are defined by

$$I_C = (V_{CC} - V_{CE})/R_L \quad \text{and} \quad I_D = (V_{DD} - V_{DS})/R_L \qquad (11.25)$$

respectively. These are called single-ended amplifiers, as the input and output voltages are referred to ground. The opposite type of amplifier is the differential amplifier configurations to be considered shortly. An amplifier using a bipolar transistor (either a BJT or an HBT) is shown in Figure 11-9(a). Referring to Figure 11-11(a), the output voltage of the bipolar amplifier is V_{CE} and this swings from a maximum value of $V_{CE,\max}$ to a minimum of $V_{CE,\min}$. For a bipolar transistor $V_{CE,\min}$ is approximately 0.2 V, while $V_{CE,\max}$ for a resistively biased circuit is just the supply voltage, V_{CC}. The quiescent or bias point is shown with collector-emitter voltage V_Q and quiescent current I_Q. For a Class A amplifier, the quiescent point is just the bias point, and this is in the middle of the output voltage swing and the slope of the loadline is established by the load resistor, R_L.

The I-V characteristics of an FET amplifier are shown in Figure 11-11(b). The notable difference between these characteristics and those of the bipolar transistor is that the curves are less abrupt at low output voltage (i.e., V_D). This results in the FET amplifier's minimum output voltage, $V_{DS,\min}$, being larger than that of a BJT-based amplifier, $V_{CE,\min}$. For a typical RF FET amplifier, as shown in Figure 11-9(b), the supply voltage, V_{DD}, is 3 V while $V_{DS,\min}$ is 0.5 V.

The bipolar and FET amplifiers of Figure 11-9 use resistive biasing so that the maximum output voltage swing is limited. As well, the bias resistor is

Figure 11-12 Class A MOSFET amplifiers with output voltage waveforms: (a) single-ended amplifier with resistive biasing; and (b) single-ended amplifier with inductive biasing. Schematic is shown in (i) and drain voltage waveforms in (ii).

also the load resistor. Various alternative topologies have been developed yielding a range of output voltage swings. The common variations are shown in Figure 11-12 for an FET amplifier. Figure 11-12(a) is a resistively biased Class A amplifier with the output voltage swing between $V_{DS,\min}$ and V_{DD}. The quiescent drain-source voltage is halfway between these extremes. The load, R_L, also provides correct biasing. This amplifier is also called a single-ended amplifier to differentiate it from a differential amplifier. A more efficient Class A amplifier uses inductive biasing, as shown in Figure 11-12(b). Bias current is now provided via the drain inductor and the load, R_L, is not part of the bias circuit. With the inductively loaded Class A amplifier, the quiescent voltage is V_{DD} and the output voltage swing is between $V_{DS,\min}$ and $2V_{DD}$, slightly more than twice the voltage swing of the resistively loaded amplifier.

When manufacturers provide S parameters of transistors, they provide the S parameters of the transistors in class A configuration with a specified bias generally defined by the output voltage and current. For FETs the drain-source voltage, V_{DS}, and drain current, I_D, are specified. See, for example, the datasheet extract in Figure 11-8. For BJTs the collector-emitter voltage and collector current are usually specified.

A linear amplifier is a class A amplifier designed for low distortion and operating with signals small enough that the design can be based on the S parameters of a transistor in a specific configuration. Ideally the effect of bias circuitry would be included in the S parameters of the transistor, but

Figure 11-13 Input and output waveforms for various classes of amplifier: (a) Class A, B, C, and AB amplifiers; and (b) switching amplifiers.

bias circuit design attempts to present RF open or short circuits as required to minimize impact. Generally however design begins with the transistor's S parameters and assuming no bias circuit impact.

11.4.2 Amplifier Efficiency

Since the Class A amplifier is always drawing DC current the efficiency of Class A amplifiers is near zero when the input signal is very small. The maximum efficiency of Class A amplifiers is 25% if resistive biasing is used and 50% when inductive biasing is used. Efficiency is improved by reducing the DC power and this is achieved by moving the bias point further down the DC loadline, as in the Class B, AB, and C amplifiers shown in Figure 11-10. Reducing the bias results in signal distortion for large RF signals. This can be seen in the various output waveforms shown in Figure 11-13(a).

Class A amplifiers have the highest linearity and Class B and C amplifiers result in considerable distortion. As a compromise, class AB amplifiers are used in many cellular applications, although Class C amplifiers are used with constant envelope modulation schemes, as in GSM. Nearly all small-signal amplifiers are Class A. This is also true for most broadband amplifiers, as amplifier stability is more certain. The class A amplifier presents impedances that are almost independent of the level of the signal. However, a Class B, AB, or C amplifier presents an impedance whose value varies depending on the level of the RF signal. Thus design requires more care, as the chances of instability are higher and it is more likely that an oscillation condition will be met. Also, Class B, AB, and C amplifiers are generally not used in broadband applications or at high frequencies mainly

Figure 11-14   DC and RF loadlines of Class A, B and C amplifiers.

because of the problem of maintaining stability. Class A amplifiers are the preferred solution for amplifiers at 10 GHz and above and for broadband amplifiers, again mainly because it is easy to ensure stability, and thus design is much simpler and more tolerant to parasitic effects and variations.

The effect of parasitic capacitances and delay effects (such as those due to the time it takes carriers to move across a base for a BJT or under the gate for an FET) result in the current-voltage locus for RF signals differing from the DC situation. This effect is captured by the **dynamic or AC loadline**, which is shown in Figure 11-14.

The Class A amplifier is biased in the middle of the I-V characteristics and the output from this amplifier has the least distortion, as seen in Figure 11-13. This seems very good, but the drawback is that the Class A amplifier draws DC current even when the input signal is negligible. This is a low-efficiency, but highly linear class. When designers refer to a "linear amplifier" they are referring to a Class A amplifier. The other amplifier classes shown in Figure 11-10 have higher efficiencies but varying degrees of distortion. The outputs are shown in Figure 11-13. The output of the Class B amplifier contains an amplified version of only half of the input signal but draws just a small leakage current when no signal is applied. With the Class C amplifier there must be some positive RF input signal before there is an output: there is more distortion but no current flows, not even leakage current, when there is no RF input signal. The **Class AB** amplifier is a compromise between Class A and Class B amplifiers. Less DC current flows than with Class A when there is negligible input signal, and the distortion is less than with Class B. There are higher classes of amplifier, Class D, E etc., and these rely on resonant circuits to change the shape of the loadline to result in better trade-offs between efficiency and distortion than can be achieved with Class AB.

Class C amplifiers are biased so that there is almost no drain-source (or collector-emitter) current when no RF signal is applied, so the output waveform has considerable distortion, as shown in Figure 11-13. This distortion is important only if there is information in the amplitude of the signal. FM, GMSK, and PM schemes result in signals with constant RF envelopes, thus there is no information contained in the amplitude of the signal. Therefore errors introduced into the amplitude of a signal are of lesser significance and efficient saturating mode amplifiers such as a Class C amplifier can be used. (we are however concerned about AM-PM distortion.) In contrast, MSK, $\pi/4$-DQPSK, $3\pi/8$-8PSK, and QAM modulation schemes do not result in signals with constant RF envelopes and so information is contained in the amplitude of the RF signal. For these modulation techniques reasonably linear amplifiers are required.

EXAMPLE 11. 1   Efficiency of a Class A Amplifier

Determine the efficiency of a Class A FET amplifier with resistive biasing using the FET characteristics shown in Figure 11-11(b).

Solution:

For maximum output voltage swing, the quiescent point should be halfway between the maximum and minimum drain source voltages,

$$V_O = V_{DS,\min} + \left(\frac{V_{DS,\max} - V_{DS,\min}}{2}\right) = \left(\frac{V_{DD} + V_{DS,\min}}{2}\right), \qquad (11.26)$$

since $V_{DS,\max} = V_{DD}$. The quiescent (DC) current through R_L is

$$I_Q = \frac{V_{DD} - V_O}{R_L} = \left(V_{DD} - \frac{V_{DD} + V_{DS,\min}}{2}\right)\frac{1}{R_L} = \frac{V_{DD} - V_{DS,\min}}{2R_L}, \qquad (11.27)$$

and the quiescent (DC) power consumed is

$$P_{DC} = V_{DD} I_Q = V_{DD}\left(\frac{V_{DD} - V_{DS,\min}}{2R_L}\right) = \frac{V_{DD}^2}{2R_L}\left(1 - \frac{V_{DS,\min}}{V_{DD}}\right). \qquad (11.28)$$

Thus the peak voltage of the AC output is

$$V_p = (V_{DD} - V_{DS,\min})/2, \qquad (11.29)$$

so that the output power in the load is

$$P_{RF,\text{out}} = \frac{1}{2}\left(\frac{V_{DD} - V_{DS,\min}}{2}\right)^2\frac{1}{R_L} = \frac{V_{DD}^2}{8R_L}\left(1 - \frac{V_{DS,\min}}{V_{DD}}\right)^2. \qquad (11.30)$$

The efficiency of the amplifier is

$$\eta = \frac{P_{RF,\text{out}}}{P_{DC}} = \frac{1}{4}\left(1 - \frac{V_{DS,\min}}{V_{DD}}\right). \qquad (11.31)$$

Figure 11-15   A two-port network with source and load used in defining stability measures: (a) network; (b) input signal flow graph; and (c) output signal flow graph.

If the minimum drain voltage is ignored ($V_{DS,\mathrm{min}} = 0$), then $\eta = 1/4 = 25\%$. This is the maximum efficiency of a class A amplifier.

11.5   Amplifier Stability

The potential exists for an amplifier to be unstable and oscillate. Generally this is not a constant oscillation at a constant amplitude and frequency but a chaotic response. Oscillator design is not as simple as designing an unstable amplifier, if only we could be so lucky. There is not a simple metric that will indicate whether an amplifier will be stable or not. In the worst case a transistor will oscillate no matter what we do to the external circuitry [179]. A manufacturer could not sell such a transistor and it would not be a useful component in a monolithic integrated process. So it is not surprising that amplifiers can be designed using available transistors. Stability considerations affect the maximum (stable) gain that can be achieved and the ease of design. If the maximum stable gain is too low another transistor needs to be selected. If the specified gain is close to the maximum stable gain, then design will be challenging, especially for broader bandwidths. So design effort increases with simpler (and hence cheaper) transistors. Experience is the best guide to making this tradeoff.

Stability analysis should be applied to the innermost two-port containing the active device and any feedback networks. If the innermost two-port is stable, then the amplifier in which it is embedded will be stable. Internal characteristics of devices could also result in instability, for example due to the internal formation of domains [180, 181]. However, instabilities resulting from such problems are rare, as it would be difficult for a vendor to sell a transistor that oscillated on its own. There is a natural selection process so that available transistors tend not to suffer from internal instabilities. So the active device in a linear amplifier can be treated as a two-port (see Figure 11-15).

Oscillation will initiate if signals reflected at the input port, Γ_{IN} increase

in amplitude as the signal reflects first from the source, Γ_S, and then from the input port. That is, if

$$|\Gamma_S \Gamma_{\text{IN}}| > 1, \tag{11.32}$$

the amplifier will be unstable at the input. This situation is shown in the signal flow graph shown in Figure 11-15(b), where oscillation is initiated by noise. Similarly oscillation will occur if multiple reflections between the output and the input build in amplitude,

$$|\Gamma_L \Gamma_{\text{OUT}}| > 1, \tag{11.33}$$

with the oscillation initiated by noise as shown in Figure 11-15(c). Now

$$|\Gamma_{\text{IN}}| = \left| S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L} \right| \tag{11.34}$$

and

$$|\Gamma_{\text{OUT}}| = \left| S_{22} + \frac{S_{12}S_{21}\Gamma_S}{1 - S_{11}\Gamma_S} \right|. \tag{11.35}$$

Combining Equations (11.32)–(11.35), the amplifier will be unstable if

$$|\Gamma_S \Gamma_{\text{IN}}| = \left| \Gamma_S S_{11} + \frac{S_{12}S_{21}\Gamma_S\Gamma_L}{1 - S_{22}\Gamma_L} \right| > 1 \tag{11.36}$$

or

$$|\Gamma_L \Gamma_{\text{OUT}}| = \left| \Gamma_L S_{22} + \frac{S_{12}S_{21}\Gamma_S\Gamma_L}{1 - S_{11}\Gamma_S} \right| > 1. \tag{11.37}$$

The coupling of Γ_S and Γ_L makes it difficult to independently design the input and output matching networks. It is much more convenient to consider the **unconditionally stable** situation whereby the input is stable no matter what the load and output matching network present, and the output is stable no matter what the source and input matching network present. As a first stage in design, a linear amplifier is designed for unconditional stability. The design space is larger if the more rigorous test for stability, embodied in Equations (11.36) and (11.37), are used to determine stability. The advantage (i.e., a larger design space), however, is often small.

  If the source and load are passive, then $\Gamma_S < 1$ and $\Gamma_L < 1$ so that oscillations will build up if

$$|\Gamma_{\text{IN}}| > 1 \tag{11.38}$$

or

$$|\Gamma_{\text{OUT}}| > 1. \tag{11.39}$$

For guaranteed stability for all passive source and load terminations (i.e., unconditional stability), then

$$|\Gamma_{\text{IN}}| < 1 \quad \text{and} \quad |\Gamma_{\text{OUT}}| < 1. \tag{11.40}$$

Amplifiers are often realized as stages whereby one amplifier stage feeds another. This complicates stability analysis, as it is possible for Γ_S and Γ_L to be more than unity. If Γ_S and Γ_L are both less than one for multiple amplifier stages then the amplifier stability being described here can be used.

For the stability criteria to be used in design they must be put in terms of the scattering parameters of the active device. There are two suitable stability criteria commonly used, the k-factor and the μ-factor, which will now be considered.

11.5.1 Unconditional Stability: Two-Port Stability Circles

The k-factor method, also known as Rollet's stability criterion [182] is the most commonly used stability metric. It is based on the input and output reflection coefficients of an active device. The input reflection coefficient of the active device is determined by the S parameters of the device and the load,

$$\Gamma_{\text{IN}} = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L}, \tag{11.41}$$

and so for stability

$$|\Gamma_{\text{IN}}| = \left| S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L} \right| < 1. \tag{11.42}$$

Similarly, considering the output of the active device, for stability,

$$|\Gamma_{\text{OUT}}| = \left| S_{22} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{11}\Gamma_S} \right| < 1. \tag{11.43}$$

Equations (11.42) and (11.43) must hold for all

$$|\Gamma_S| \leq 1 \quad \text{and} \quad |\Gamma_L| \leq 1. \tag{11.44}$$

When the active is unilateral, $S_{12} = 0$, and Equations (11.42) and (11.43) simplify to the requirement that $S_{11} < 1$ and $S_{22} < 1$. Otherwise, given a device, there will be a limit on the values of Γ_S and Γ_L. The stability criteria are in terms of the magnitudes of complex numbers, and this is known to specify circles in the complex plane. The following development will lead to the center and radius defining the stability circles. These stability circles define the boundary between stable and unstable regions.

For $|\Gamma_{\text{IN}}| = 1$, the output stability circle is defined by

$$\left| S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L} \right| = 1. \tag{11.45}$$

The development puts this into the standard form for a circle, that is, in the form of

$$|\Gamma_L - c| = r. \tag{11.46}$$

where c is a complex number and defines the center of the circle, and r is a real number and is the radius of the circle (see Section A.7). This circle defines the boundary between stable and unstable values of Γ_L. Now Equation (11.45) can be rewritten as

$$|S_{11} - (S_{11}S_{22} - S_{12}S_{21})\Gamma_L| = |1 - S_{22}\Gamma_L|. \tag{11.47}$$

The determinate of the scattering parameter matrix is now

$$\Delta = S_{11}S_{22} - S_{12}S_{21}, \tag{11.48}$$

and so Equation (11.47) becomes

$$|S_{11} - \Delta\Gamma_L| = |1 - S_{22}\Gamma_L|. \tag{11.49}$$

That is, removing the absolute signs by multiplying each side by its complex conjugate

$$(S_{11} - \Delta\Gamma_L)(S_{11} - \Delta\Gamma_L)^* = (1 - S_{22}\Gamma_L)(1 - S_{22}\Gamma_L)^* \tag{11.50}$$

$$S_{11}S_{11}^* + \Delta\Delta^*\Gamma_L\Gamma_L^* - (\Delta\Gamma_L S_{11}^* + \Delta^*\Gamma_L^* S_{11}) = 1 + S_{22}S_{22}^*\Gamma_L\Gamma_L^*$$
$$- (S_{22}\Gamma_L + S_{22}^*\Gamma_L^*) \tag{11.51}$$

$$\left(|S_{22}|^2 - |\Delta|^2\right)\Gamma_L\Gamma_L^* - (S_{22} - \Delta S_{11}^*)\Gamma_L$$
$$- (S_{22}^* - \Delta^* S_{11})\Gamma_L^* = |S_{11}|^2 - 1 \tag{11.52}$$

$$\left(|S_{22}|^2 - |\Delta|^2\right)\Gamma_L\Gamma_L^* - (S_{22} - \Delta S_{11}^*)\Gamma_L$$
$$- (S_{22} - \Delta S_{11}^*)^*\Gamma_L^* = |S_{11}|^2 - 1 \tag{11.53}$$

$$\Gamma_L\Gamma_L^* - \frac{(S_{22} - \Delta S_{11}^*)\Gamma_L - (S_{22}^* - \Delta S_{11}^*)\Gamma_L^*}{|S_{22}|^2 - |\Delta|^2} = \frac{|S_{11}|^2 - 1}{|S_{22}|^2 - |\Delta|^2} \tag{11.54}$$

$$\Gamma_L\Gamma_L^* - \frac{(S_{22} - \Delta S_{11}^*)\Gamma_L}{|S_{22}|^2 - |\Delta|^2} - \frac{(S_{22} - \Delta S_{11}^*)^*\Gamma_L^*}{|S_{22}|^2 - |\Delta|^2}$$
$$+ \frac{(S_{22} - \Delta S_{11}^*)(S_{22} - \Delta S_{11}^*)^*}{\left(|S_{22}|^2 - |\Delta|^2\right)^2}$$
$$= \frac{|S_{11}|^2 - 1}{|S_{22}|^2 - |\Delta|^2} + \frac{(S_{22} - \Delta S_{11}^*)(S_{22} - \Delta S_{11}^*)^*}{\left(|S_{22}|^2 - |\Delta|^2\right)^2} \tag{11.55}$$

$$\left|\Gamma_L + \frac{S_{22} - \Delta S_{11}^*}{|S_{22}|^2 - |\Delta|^2}\right|^2 = \left|\frac{S_{12}S_{21}}{|S_{22}|^2 - |\Delta|^2}\right|^2. \tag{11.56}$$

$|S_{11}| < 1$

$|S_{11}| > 1$

Figure 11-16 Output stability circles on the Γ_L plane: (a) $|S_{11}| < 1$; and (b) $|S_{11}| > 1$. The shaded regions denote the values of Γ_L that will result in unconditional stability at the input indicated by $|\Gamma_{\text{in}}| < 1$.

That is

$$\left| \Gamma_L - \frac{S_{22} - \Delta S_{11}^*}{|S_{22}|^2 - |\Delta|^2} \right| = \left| \frac{S_{12} S_{21}}{|S_{22}|^2 - |\Delta|^2} \right| . \qquad (11.57)$$

This defines a circle, called the output stability circle, in the Γ_L plane with center c_L and radius r_L (see Section A.7):

$$\text{center}: \ c_L = \frac{(S_{22} - \Delta S_{11}^*)^*}{|S_{22}|^2 - |\Delta|^2} \qquad (11.58)$$

$$\text{radius}: \ r_L = \left| \frac{S_{12} S_{21}}{|S_{22}|^2 - |\Delta|^2} \right|, \qquad (11.59)$$

and which is plotted on a Smith chart in Figure 11-16 for the two conditions $|S_{11}| < 1$ and $|S_{11}| > 1$. When $|S_{11}| < 1$ the shaded region in Figure 11-16(a) indicates unconditional stability. That is, as long as Γ_L is chosen to lie in the shaded region, the input reflection coefficient, Γ_{in}, will be less than one. It does not matter what the source impedance is, as long as it is passive there will not be oscillation due to multiple reflections between the input of the amplifier (or any two-port) and the source.

Figure 11-17   Input stability circles on the Γ_S plane: (a) $|S_{22}| < 1$; and (b) $|S_{22}| > 1$. The shaded regions denote the values of Γ_S that will result in unconditional stability at the input indicated by $|\Gamma_{\text{out}}| < 1$.

Similarly an input stability circle can be defined for Γ_S, where

$$\text{center}: \ c_S = \frac{(S_{11} - \Delta S_{22}^*)^*}{|S_{11}|^2 - |\Delta|^2} \tag{11.60}$$

$$\text{radius}: \ r_S = \left| \frac{S_{12} S_{21}}{|S_{11}|^2 - |\Delta|^2} \right|. \tag{11.61}$$

The interpretation of the input stability circles, shown in Figure 11-17, is the same as for the output stability circles.

The stability criterion provided by the input and output stability circles is very conservative. For example, the input stability circle (for Γ_S) indicates the value of Γ_S that will ensure stability no matter what passive load is presented. Thus the stability circles here are called unconditional stability circles. However, an amplifier can be stable for loads (or source impedances) other than those that ensure unconditional stability. However, the use of stability circles provides a good first pass in design of the matching networks between the actual source and load and the amplifier. The stability circles will change with frequency, and so ensuring stability requires a broad frequency view. This is considered in the linear amplifier design example in Section 11.7.

If an amplifier is unconditionally stable, amplifier design is considerably simplified. A more complete stability analysis is presented by Gonzalez [183] and Suárez and Quéré [184].

Figure 11-18   General amplifier configuration.

11.5.2   Rollet's Stability Criterion — k-factor

The most general amplifier is depicted in Figure 11-18. Here the active device has scattering parameters

$$[\mathbf{S}] = \left[\begin{array}{cc} S_{11} & S_{12} \\ S_{21} & S_{22} \end{array} \right]$$

and, as it will be used a lot, define the determinant $\Delta = S_{11}S_{22} - S_{12}S_{21}$. The overall amplifier has scattering parameters S'. Γ_S is the generator reflection coefficient and Γ_L is the reflection coefficient of the load. For unconditional stability, $|S'_{11}| < 1$ and $|S'_{22}| < 1$. If $|S'_{11}| > 1$ or $|S'_{22}| > 1$, then there is a negative resistance and oscillation could possibly occur. Unconditional stability is defined as when $|S'_{11}| < 1$ for all loads Γ_L and $|S'_{22}| < 1$ for all source impedances Γ_S. These inequalities are the same as saying the real part of the input and output impedances of the amplifier are positive resistances. These requirements also describe the unit circle on a Smith chart. Beginning with these definitions and ignoring the feedback network, a stability factor, k, can be defined as

$$k = \left(\frac{1\text{-}|S_{11}|^2\text{-}|S_{22}|^2+|\Delta|^2}{2|S_{12}|\,|S_{21}|} \right), \tag{11.62}$$

where $k > 1$ is required for unconditional stability. This is the Rollet condition. What is done here is rolling two unconditional stability requirements (on $|S'_{11}|$ and $|S'_{22}|$) into one. If $k \leq 1$, the amplifier may not be unstable, but extra care is required when a load is presented to the amplifier. If $k > 1$, the design is relatively straightforward, but if k is near 1 or $k \leq 1$, design will be troublesome. The closer design is to the limits of operation of a device, the more likely k will be near or less than 1. For example, the limits could be the maximum stable gain at the intended frequency of operation.

It has been shown that unconditional stability is assured if

$$k = \left(\frac{1\text{-}|S_{11}|^2\text{-}|S_{22}|^2+|\Delta|^2}{2|S_{12}|\,|S_{21}|} \right) > 1, \tag{11.63}$$

combined with any one of the following auxiliary conditions [185–191]:

$$B_1 = 1 + |S_{11}|^2 - |S_{22}|^2 - |\Delta|^2 > 0 \tag{11.64}$$

$$B_2 = 1 - |S_{11}|^2 + |S_{22}|^2 - |\Delta|^2 > 0 \tag{11.65}$$

$$|\Delta| = |S_{11}S_{22} - S_{12}S_{21}| < 1 \tag{11.66}$$

$$C_1 = 1 - |S_{11}|^2 - |S_{12}S_{21} > 0| \tag{11.67}$$

$$C_2 = 1 - |S_{22}|^2 - |S_{12}S_{21} > 0|. \tag{11.68}$$

The conditions denoted in Equations (11.67) and (11.68) are not independent, and it can be shown that one implies the other if the $k > 1$ [185].

If $k > 1$, then an amplifier is unconditionally stable if any one of Equations (11.64)–(11.68) is satisfied. Rollet's stability criteria, k and $|\Delta|$, are tabulated in Table 11-5 for the pHEMT described in Figure 11-8. The device is unconditionally stable at the frequencies 5–11 GHz and 22–26 GHz. The device could be potentially unstable at frequencies below 5 GHz and from 12 to 21 GHz. At these frequencies design needs to use stability circles in designing matching networks.

11.5.3 Edwards-Sinsky Stability Criterion — μ-factor

Rollet's stability criterion, Equations (11.63)–(11.68), assures unconditional stability but it does not provide a relative measure of stability. That is, the k factor cannot be used to determine how close to the edge of stability a particular design is. There is no ability to compare the relative stability of different designs. Edwards and Sinsky [185] developed a test that can be used to compare the relative stability of different designs. This is called the μ-factor stability criterion, with unconditional stability having

$$\mu = \frac{1 - |S_{11}|^2}{|S_{22} - S_{11}^*\Delta| + |S_{21}S_{12}|} > 1. \tag{11.69}$$

An important result is that larger values of μ indicate greater stability. The μ factor is a single quantity that provides a sufficient and necessary condition for unconditional stability. That is, it does not matter what passive source and load is presented, $\Gamma_S < 1$ and $\Gamma_L < 1$, the amplifier will be stable. This contrasts with Rollet's stability criterion in which two conditions must be met.

Edwards-Sinsky stability parameters for the pHEMT transistor documented in Figure 11-8 are shown in Table 11-6. The unconditionally stable frequencies of operation are 5–11 GHz and 22–26 GHz. The same unconditionally stable frequencies determined by using Rollet's stability criterion (in Table 11-5). The additional information available with the Edwards-Sinsky stability criterion is that μ indicates the relative stability. In Table 11-6, the transistor is unconditionally stable in the 5 to 11 GHz range, and in this

Table 11-5 Rollet's stability factor, k-factor, $|\Delta|$, and stability circle parameters of the pHEMT transistor documented in Figure 11-8. For the active device to be unconditionally stable two conditions must be met: $k > 1$ and $|\Delta| < 1$. The frequencies at which the device is unconditionally stable are in bold.

Freq. (GHz)	k > 1	$|\Delta|$ < 1	C_L	R_L	C_S	R_S
0.5	0.15178	0.62757	$1.0388 + j13.5176$	13.370	$0.92820 + j0.50682$	0.22434
1	0.24895	0.58720	$0.78679 + j7.26968$	6.9988	$0.69258 + j0.96669$	0.44107
2	0.46535	0.46821	$0.73554 + j3.96784$	3.4718	$-0.087892 + j1.481129$	0.72546
3	0.67865	0.37045	$0.63608 + j3.17780$	2.4779	$-0.86485 + j1.46382$	0.85475
4	0.91943	0.29706	$0.52381 + j2.82057$	1.9223	$-1.4232 + j1.2217$	0.91458
5	1.0838	0.24365	$0.32018 + j2.76164$	1.7276	$-1.90410 + j0.77291$	1.0132
6	1.1839	0.18765	$0.065264 + j2.679153$	1.5700	$-2.182123 + j0.024217$	1.0885
7	1.2846	0.14113	$-0.39857 + j2.97836$	1.8266	$-2.06262 - j0.85258$	1.0885
8	1.5225	0.092502	$-0.91418 + j3.21866$	2.0150	$-1.5996 - j1.4914$	0.94750
9	1.6448	0.056060	$-1.2526 + j3.0479$	1.8998	$-1.1418 - j1.8922$	0.92223
10	1.3151	0.072750	$-1.7681 + j2.6944$	2.0189	$-0.54178 - j2.13439$	1.0468
11	1.1043	0.11703	$-2.6666 + j2.4074$	2.5186	$0.24025 - j2.02042$	0.98357
12	0.98784	0.16159	$-4.4423 + j2.2978$	4.0111	$0.73109 - j1.58120$	0.74724
13	0.92131	0.18991	$-5.9502 + j1.0596$	5.1100	$0.96118 - j1.24308$	0.60117
14	0.84098	0.21905	$-5.95467 - j0.75389$	5.1368	$1.10689 - j0.97667$	0.53246
15	0.69555	0.24320	$-6.0313 - j1.9645$	5.6068	$1.20251 - j0.58617$	0.43291
16	0.63420	0.27993	$-5.7310 - j3.8333$	6.2171	$1.26271 - j0.23704$	0.39187
17	0.68792	0.32454	$-3.5163 - j5.6496$	5.9268	$1.257206 + j0.087208$	0.34232
18	0.72764	0.36197	$-1.1258 - j5.5386$	4.8824	$1.17955 + j0.29932$	0.27755
19	0.89194	0.35755	$0.72551 - j2.82689$	1.9913	$1.20751 + j0.32462$	0.27383
20	0.97085	0.32318	$1.1030 - j2.0753$	1.3671	$1.19540 + j0.46897$	0.29068
21	0.97475	0.28626	$1.3859 - j1.8451$	1.3221	$0.94075 + j0.85514$	0.27681
22	1.1014	0.28501	$2.1696 - j1.1961$	1.4187	$0.66834 + j1.07377$	0.24499
23	1.3123	0.29397	$2.35683 - j0.15064$	1.1976	$0.51539 + j1.17421$	0.22605
24	1.4714	0.28083	$2.23754 + j0.46967$	1.0569	$0.28045 + j1.30061$	0.24185
25	1.4273	0.22850	$2.2473 + j1.2506$	1.3389	$0.079906 + j1.412442$	0.31586
26	1.5308	0.17235	$2.6309 + j3.0590$	2.6671	$-0.45588 + j1.43292$	0.36774

range the device is most stable between 8 and 10 GHz. At the high end, above 22 GHz, the device is increasingly more stable. This can be expected to continue as the device capacitive parasitics short out the device. (The result of the low impedance of capacitors at high frequencies is that the effect of feedback will be much less, and smaller RF voltages will be generated for the same drive level.) So this transistor will make a very good 8–10 GHz amplifier provided that appropriate matching networks are chosen to ensure stability below 5 GHz and above 11 GHz. Note, however, that the amplifier can be used up to about 20 GHz, but with extra care in design. Above 20 GHz the maximum unilateral transducer gain is too small to be useful (see Table 11-7 where $G_{TU,\max}$ is tabulated for this transistor).

Table 11-6 Edwards-Sinsky stability parameters for the pHEMT transistor documented in Figure 11-8. For stability, $\mu > 1$ and one of the following factors must be satisfied: that is, $B_1 > 0$, $B_2 > 0$, $|\Delta| < 1$, $C_1 > 0$, $C_2 > 0$. The frequencies at which the device is unconditionally stable are in bold. The device is unconditionally stable at the frequencies 5–11 GHz and 22–26 GHz.

Freq. (GHz)	μ > 1	B_1 > 0	B_2 > 0	$|\Delta|$ < 1	C_1 > 0	C_2 > 0
0.5	0.18785	1.1555	0.056799	0.62757	−0.077921	0.47143
1	0.31338	1.1338	0.17657	0.58720	−0.080934	0.39770
2	0.56362	1.1086	0.45297	0.46821	0.065756	0.39356
3	0.76297	1.0884	0.63717	0.37045	0.22398	0.44958
4	0.94651	1.0631	0.76039	0.29706	0.35892	0.51029
5	1.0526	1.0434	0.83782	0.24365	0.44002	0.54284
6	1.1100	1.0341	0.89551	0.18765	0.49297	0.56225
7	1.1783	1.0703	0.88992	0.14113	0.51407	0.60424
8	1.3310	1.1120	0.87085	0.092502	0.54812	0.66871
9	1.3954	1.1104	0.88335	0.056060	0.57284	0.68634
10	1.2039	1.1068	0.88259	0.072750	0.51811	0.63023
11	1.0739	1.1550	0.81758	0.11703	0.43720	0.60592
12	0.99026	1.2715	0.67626	0.16159	0.33481	0.63244
13	0.93383	1.3461	0.58173	0.18991	0.27037	0.65257
14	0.86542	1.3788	0.52518	0.21905	0.22227	0.64910
15	0.73637	1.4578	0.42389	0.24320	0.13811	0.65507
16	0.67769	1.4752	0.36811	0.27993	0.099372	0.65290
17	0.72762	1.4461	0.34324	0.32454	0.10910	0.66053
18	0.76942	1.4300	0.30791	0.36197	0.10899	0.67006
19	0.92718	1.3348	0.40953	0.35755	0.18890	0.65152
20	0.98311	1.2924	0.49876	0.32318	0.24511	0.64191
21	0.98558	1.3331	0.50306	0.28626	0.24786	0.66286
22	1.0587	1.3627	0.47486	0.28501	0.25075	0.69466
23	1.1641	1.3295	0.49769	0.29397	0.28504	0.70093
24	1.2294	1.2872	0.55509	0.28083	0.33166	0.69771
25	1.2330	1.2866	0.60899	0.22850	0.36433	0.70313
26	1.3676	1.3398	0.60077	0.17235	0.38405	0.75358

11.5.4 Summary

This section presented criteria that can be used in determining the stability of transistor amplifiers and of active devices. Stability tests should be applied to the innermost two-port to provide an enhanced confidence in design. Two criteria were presented for unconditional stability: the k-factor and μ-factor. The k-factor, part of Rollet's unconditional stability criterion, indicates an associated test of whether a device is unconditionally stable or not. It does not provide a relative measure of stability. The μ-factor, derived by Edwards and Sinsky, is a factor that indicates the relative stability of

a network. Stable amplifiers can be designed even if the amplifier is not unconditionally stable. Stability circles aid in the design of such amplifiers. Stability is an extensive topic and the reader is directed to in-depth stability analysis by Suárez and Quéré [184] for further information.

In the following sections a complete set of amplifier gain measures, including maximum stable gain, will be defined. Detailed linear amplifier design will then combine the stability concepts presented here.

11.6 Amplifier Gain Definitions

As with all circuit design, a few figures of merit (FOM) are used to guide design. The most important metric in amplifier design is the gain of the overall amplifier. There are a surprisingly large number of different definitions of gain that are useful at different stages in the design process. Each provides information about the performance of an amplifier and using the full set enables design to be approached in a systematic way. The FOMs are used to describe the performance of an amplifier, to develop an understanding of the active device, to compare different active devices, and, coupled with experience, to formulate an idea of how difficult a design will be.

The quantities used in the various gain definitions are defined in Figure 11-19. The power delivered to the amplifier is P_{in}, and this is equal to the available input power from the source if the source is conjugately matched to the input matching network. The power delivered to the active device, $P_{\text{in}D}$, is equal to the amplifier input power, P_{in}, if the input matching network is lossless. The available output power from the active device, P_{Ao}, is the actual device output power, P_o, delivered to the output matching network if the output of the active device is conjugate matched. This power is also delivered to the load as P_L if M_2 is lossless. In summary,

$$P_{\text{in}} = P_{Ai}, \quad \text{if generator is conjugate matched}$$
$$P_{\text{in}} = P_{\text{in}D}, \quad \text{if } M_1 \text{ is lossless}$$
$$P_o = P_{Ao}, \quad \text{if the output of active device is conjugate matched}$$
$$P_L = P_o, \quad \text{if } M_2 \text{ is lossless.}$$

These power definitions refer to different circuit conditions. This enables a number of different gain definitions to be developed that relate to different stages in the development of an amplifier and indicate the ultimate performance achievable from an amplifier. The basic gain definitions are

- *System gain*:

$$G = \frac{P_L}{P_{\text{in}}}. \tag{11.70}$$

  The system gain is the power actually delivered to the load relative to the input power delivered by the source. This gain is sometimes called

Power	Description
P_{in}	Actual input power delivered to the amplifier.
P_{Ai}	Available input power from the source. $P_{\text{in}} \leq P_{Ai}$. If M_1 provides conjugate matching as seen from the source, then $P_{\text{in}} = P_{Ai}$
$P_{\text{in}D}$	Actual device input power delivered to the active device. $P_{\text{in}D} \leq P_{\text{in}}$. If M_1 is lossless, $P_{\text{in}D} = P_{\text{in}}$.
P_{ADo}	Available device output power of the active device.
P_{Ao}	Available amplifier output power. $P_{Ao} \leq P_{ADo}$. If M_2 is lossless, $P_{Ao} = P_{ADo}$.
P_L	Actual output power delivered to load. Amplifier output power. $P_L \leq P_{Ao}$. If M_2 is lossless and provides conjugate matching, $P_L = P_{Ao} = P_{ADo}$.

Figure 11-19 Parameters used in defining gain measures. The input and output matching networks are lossless so that the actual device input signal power, $P_{\text{in}D}$, is the power delivered by the source. Similarly, the actual output signal power delivered to the load, P_L, is the power delivered by the active device (the transistor including biasing network).

the **actual power gain**.

- *Power gain*:

$$G_P = \frac{P_L}{P_{\text{in},D}}. \tag{11.71}$$

This gain is G, but with the effect of M_1 removed.

- *Transducer gain*:

$$G_T = \frac{P_L}{P_{Ai}}. \tag{11.72}$$

This is the ratio of the power delivered to the load divided by the power available from the source. This gain is G with optimum M_1.

- *Available gain*:

$$G_A = \frac{P_{Ao}}{P_{Ai}}. \tag{11.73}$$

The transducer gain is the power available to the load relative to the input power available from the source. This gain is G_T with optimum M_2.

Note that $G \leq G_P \leq G_T \leq G_A$. These gain definitions are useful at different stages of the design process.

These gains are measures that can be used toward the end of a design iteration but do not indicate the best gains achievable under various conditions. For example, gain measures can be defined that enable the initial choice of a transistor based on the S parameters of the transistor typically provided in manufacturers' data sheets. Developing the generalized scattering parameters of the amplifier normalized to the source and load impedances is key to casting the gains in terms of device S parameters.

11.6.1 Gain in Terms of Scattering Parameters

This section develops the generalized scattering parameters of an amplifier and leads to three important metrics of amplifiers: transducer gain, as defined in Equation (11.72), unilateral transducer gain, and maximum unilateral transducer gain. These gains are expressed in terms of S parameters. A linear amplifier can be represented as a two-port with a Thevenin equivalent source at Port 1 and a load at Port 2, as shown in Figure 11-20(a). This section illustrates the usefulness of generalized S parameters in working with power flow in systems with complex and different system impedances at the ports. Let **S** be the normalized scattering matrix of the two-port, with Z_0 being the normalizing real characteristic impedance. Therefore

$$\mathbf{S} = \left[\begin{array}{cc} S_{11} & S_{12} \\ S_{21} & S_{22} \end{array} \right]. \tag{11.74}$$

The development in this section uses the S parameters of the amplifier in Figure 11-20(a). The aim is to develop an expression for the unilateral transducer gain and for the maximum unilateral transducer gain. The unilateral transducer gain is restricted to the amplifier (Figure 11-20(a)), and the maximum unilateral transducer gain can use the S parameters of either the transistor (Figure 11-20(b)), or the S parameters of the amplifier (Figure 11-20(a)).

The generalized scattering matrix of the amplifier will be normalized to the source impedance, Z_S, and load impedance, Z_L, shown in Figure 11-20.

Figure 11-20 Two-port network with source and load used in defining unilateral gain measures: (a) an amplifier; and (b) active device with input and output matching networks.

First, find the reflection coefficients Γ_S and Γ_L at the source and load using the following equations:

$$\Gamma_S = \frac{Z_S - Z_0}{Z_S + Z_0}, \tag{11.75}$$

$$\Gamma_L = \frac{Z_L - Z_0}{Z_L + Z_0}. \tag{11.76}$$

Now construct

$$\mathbf{D} = \begin{bmatrix} D_{11} & 0 \\ 0 & D_{22} \end{bmatrix}, \tag{11.77}$$

where

$$D_{ii} = \frac{(1 - \Gamma_i)\sqrt{1 - |\Gamma_i|^2}}{|1 - \Gamma_i^*|}. \tag{11.78}$$

From Equation (6.115), the generalized scattering parameters (with Port 1 normalized to Z_S and Port 2 normalized to Z_L) are

$$^G\mathbf{S} = \mathbf{D}^{*-1}(\mathbf{S} - \boldsymbol{\Gamma}^*)(\mathbf{U} - \boldsymbol{\Gamma}\mathbf{S})^{-1}\mathbf{D}. \tag{11.79}$$

Following tedious algebraic manipulations, the following expressions are

obtained:

$$^{G}S_{11} = \frac{1}{W}\frac{1-\Gamma_S}{1-\Gamma_S^*}[(S_{11}-\Gamma_S^*)(1-\Gamma_L S_{22}) + S_{12}S_{21}\Gamma_L] \quad (11.80)$$

$$^{G}S_{12} = \frac{1}{W}\frac{(1-\Gamma_S)(1-\Gamma_L)}{|1-\Gamma_S||1-\Gamma_L|}S_{12}[(1-|\Gamma_S|^2)(1-|\Gamma_L|^2)]^{\frac{1}{2}} \quad (11.81)$$

$$^{G}S_{21} = \frac{1}{W}\frac{(1-\Gamma_S)(1-\Gamma_L)}{|1-\Gamma_S||1-\Gamma_L|}S_{21}[(1-|\Gamma_S|^2)(1-|\Gamma_L|^2)]^{\frac{1}{2}} \quad (11.82)$$

$$^{G}S_{22} = \frac{1}{W}\frac{1-\Gamma_L}{1-\Gamma_L^*}[(S_{22}-\Gamma_L^*)(1-\Gamma_S S_{11}) + S_{12}S_{21}\Gamma_S], \quad (11.83)$$

where

$$W = (1 - \Gamma_S S_{11})(1 - \Gamma_L S_{22}) - S_{12}S_{21}\Gamma_S\Gamma_L \quad (11.84)$$

and

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = {}^{G}\mathbf{S}\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}. \quad (11.85)$$

Here a and b are the root power waves defined on Page 306 and in Figure 6-16.

A number of useful observations can be made. First, in a matched condition where $\Gamma_S = \Gamma_L = 0$ (i.e., $Z_S = Z_L = Z_0$), $^{G}\mathbf{S} = \mathbf{S}$. Second, for a reciprocal two-port network with $S_{12} = S_{21}$, the generalized scattering parameters are also reciprocal (i.e., $^{G}S_{12} = {}^{G}S_{21}$). An amplifier is not reciprocal, however, and for a good amplifier, $^{G}S_{12}$ is approximately zero and $^{G}S_{21}$ is greater than one, indicating power gain.

Now the transducer power gain, G_T, can be expressed in terms of device S parameters. G_T, Equation (11.72), is defined as the ratio of the average power, P_L, delivered to the load Z_L, and the maximum input power available from the generator, P_{Ai}, that is,

$$G_T = \frac{P_L}{P_{Ai}}, \quad (11.86)$$

where the available power from the generator is

$$P_{Ai} = \frac{1}{8}\frac{|V_S|^2}{\Re\{Z_S\}} = \frac{1}{2}|a_1|^2 \quad (11.87)$$

and the power delivered to the load is

$$P_L = \frac{1}{2}\Re\{Z_L\}|-I_2|^2 = \frac{1}{2}|b_2|^2. \quad (11.88)$$

These quantities are defined in terms of the root power waves a_1 and b_2 and these are related by the generalized S parameters. Thus the transducer gain is

$$G_T = \left|\frac{b_2}{a_1}\right|^2 = |^{G}S_{21}|^2, \quad (11.89)$$

and so, using Equation (11.82),

$$G_T = |S_{21}|^2 \frac{(1 - |\Gamma_S|^2)(1 - |\Gamma_L|^2)}{|(1 - \Gamma_S S_{11})(1 - \Gamma_L S_{22}) - \Gamma_S \Gamma_L S_{12} S_{21}|^2}. \tag{11.90}$$

This expression for the gain combines the inherent voltage gain of the device (S_{21}) and the effect of load, and source mismatches through Γ_L and Γ_S. This expression simplifies under particular circumstances which will now be considered.

For a unilateral two-port, $S_{12} = 0$ (and this is a reasonable approximation for many amplifiers, as there is little feedback from the output to the input). Under this circumstance the transducer gain is referred to as the **unilateral transducer gain**, G_{TU}:

$$G_{TU} = |S_{21}|^2 \left(\frac{1 - |\Gamma_S|^2}{|1 - \Gamma_S S_{11}|^2} \right) \left(\frac{1 - |\Gamma_L|^2}{|1 - \Gamma_L S_{22}|^2} \right). \tag{11.91}$$

This is often referred to as just the **unilateral gain**.

From this last expression it can be seen that by choosing $\Gamma_S = (S_{11})^*$ and $\Gamma_L = (S_{22})^*$ (i.e., the complex conjugates), G_{TU} achieves its maximum value, the **maximum unilateral transducer gain**:

$$G_{TU\text{max}} = |S_{21}|^2 \left(\frac{1}{1 - |S_{11}|^2} \right) \left(\frac{1}{1 - |S_{22}|^2} \right). \tag{11.92}$$

If the S parameters in Equation (11.92) are those of the amplifier shown in Figure 11-20(a), then $G_{TU\text{max}}$ is the maximum unilateral transducer gain available from the amplifier. In general, additional matching would be required between the source and the amplifier, and between the amplifier and the load to achieve this gain. If the S parameters in Equation (11.92) are those of the transistor shown in Figure 11-20(b), then $G_{TU\text{max}}$ is the maximum unilateral transducer gain available from the active device. This is a good measure of the maximum power gain readily obtained from a device. However, with feedback (consider the general amplifier configuration of Figure 11-18), the effective $S_{12} \neq 0$, and any gain can be achieved, even oscillation. Also, higher gain is obtained at the expense of reduced bandwidth. As a general design guideline, the closer the gain specified for an amplifier is to $G_{TU\text{max}}$, the more challenging the design task.

The maximum unilateral transducer gain, $G_{TU\text{max}}$, of the pHEMT transistor described in Figure 11-8 in shown in Table 11-7. The maximum unilateral transducer gain is largest at low frequencies and monotonically reduces as frequency increases. This means that it is difficult to design broadband amplifiers. It is also a challenge to ensure stable amplification at low frequencies and matching networks must be chosen to suppress low-frequency gain.

Table 11-7 Maximum unilateral transducer gain, $G_{TU\text{max}}$, of the pHEMT transistor documented in Figure 11-8.

Frequency (GHz)	$G_{TU\text{max}}$ (dB)	Frequency (GHz)	$G_{TU\text{max}}$ (dB)
0.5	36.62	14	11.25
1	31.07	15	11.61
2	24.88	16	11.64
3	21.26	17	11.11
4	18.73	18	10.50
5	16.00	19	8.89
6	15.74	20	7.91
7	14.52	21	7.48
8	13.26	22	6.55
9	12.36	23	5.46
10	12.24	24	4.62
11	12.07	25	4.22
12	11.77	26	3.61
13	11.46		

With the stability criteria established, the full set of gain definitions can be defined in terms of S parameters, with some indicating the best that can be achieved with a stable amplifier. Circuit quantities are defined in Figure 11-19.

- Transducer power gain:

$$G_T = \frac{(1 - |\Gamma_S|^2)|S_{21}|^2(1 - |\Gamma_L|^2)}{|(1 - S_{11}\Gamma_S)(1 - S_{22}\Gamma_L) - S_{12}S_{21}\Gamma_G\Gamma_L|^2}. \qquad (11.93)$$

This is the transducer gain with the effect of matching networks described by Γ_S and Γ_L.

- Transducer power gain in system impedance. Source and load impedances equal to the system impedance:

$$G_T|_{Z_0} = |S_{21}|^2. \qquad (11.94)$$

This is the transducer gain without matching networks, so that $\Gamma_S = 0 = \Gamma_L$. This is nearly always much lower than what can be achieved using matching networks. For one, an active device has significant input and output reactances at microwave frequencies that need to be tuned out.

- Unilateral transducer power gain, G_{TU}:

$$G_{TU} = |S_{21}|^2 \left(\frac{(1 - |\Gamma_S|^2)}{|1 - S_{11}\Gamma_S|^2}\right) \left(\frac{(1 - |\Gamma_L|^2)}{|1 - S_{22}\Gamma_L|^2}\right). \qquad (11.95)$$

This is the transducer gain, G_T, with S_{12} set to zero. This expression has been carefully arranged to show that the unilateral transducer power gain is the product of an active device portion, $|S_{21}|^2$, and an input matching factor, and an output matching factor.

- Power gain with input conjugately matched:

$$G_P = \frac{|S_{21}|^2(1 - |\Gamma_L|^2)}{|(1 - S_{22}\Gamma_L)(1 - |S_{11}|^2)}.$$
(11.96)

With $\Gamma_L = 0$,

$$G_P|_{\Gamma_L=0} = \frac{|S_{21}|^2}{(1 - |S_{11}|^2)}.$$
(11.97)

- Available power gain with output conjugately matched:

$$G_A = \frac{|S_{21}|^2(1 - |\Gamma_S|^2)}{|1 - S_{11}\Gamma_S|^2(1 - |S_{22}|^2)}.$$
(11.98)

With $\Gamma_S = 0$,

$$G_A|_{\Gamma_S=0} = \frac{|S_{21}|^2}{1 - |S_{22}|^2}.$$
(11.99)

- Maximum available power gain, $G_{MA} = G_A$ and G_T with optimum M_1 and M_2:

$$G_{MA} = \left|\frac{S_{21}}{S_{12}}\right| \left(k - \sqrt{k^2 - 1}\right).$$
(11.100)

Here k is Rollet's stability factor (Equation (11.62)).

- Maximum unilateral transducer power gain with input and output conjugately matched so that $\Gamma_S = S_{11}^*$ and $\Gamma_L = S_{22}^*$ (note that we are treating $S_{12} = 0$ in the unilateral approximation):

$$G_{TU,\mathrm{max}} = |S_{21}|^2 \left(\frac{1}{1 - |S_{11}|^2}\right) \left(\frac{1}{1 - |S_{22}|^2}\right).$$
(11.101)

This expression has been carefully arranged to show that the unilateral transducer power gain is the product of an active device portion, $|S_{21}|^2$, and an input factor, and an output factor.

- Maximum stable power gain, G_{MS}. $G_{MS} = G$ with resistive loading, setting $k = 1$, and optimum M_1 and M_2. G_{MS} is useful when $k \geq 1$. This is generally the limit to gain as the input and output mismatch is reduced, with conjugate matching at the input and output oscillation likely to occur.

$$G_{MS} = \left|\frac{S_{21}}{S_{12}}\right|.$$
(11.102)

TUNED · WIDEBAND

Figure 11-21 Low-noise amplifiers: (a) tuned amplifier; and (b) wideband amplifier.

- Unilateral power gain, U. This is the maximum available gain, G_{MA}, with feedback across the active device adjusted so that the effective device feedback parameter $S_{12} = 0$. When $U = 1$, the devices go from being active to being passive:

$$U = \left(\frac{1/2|S_{21}/S_{12} - 1|^2}{k|S_{21}/S_{12}| - Re(S_{21}/S_{12})} \right). \qquad (11.103)$$

11.6.2 Summary

G_{TU} (Equation (11.95)), $G_{TU,\max}$ (Equation (11.101)), G_{MS} (Equation (11.102)), G_{MA} (Equation (11.100)), and U (Equation (11.103)) are used by designers as measures of the ultimate performance of a device. They are important figures of merit as they are defined with optimum input and output matching networks. They are useful at the beginning of a design and lead to the initial active device choice. As well, with experience, comparing these parameters to gain specification indicates the amount of design effort required to achieve a successful design.

G_{TU}, $G_{TU,\max}$, G_{MS}, G_{MA}, and U are tabulated in Table 11-8 for the pHEMT transistor documented in Figure 11-8. As shown in Tables 11-5 and 11-6 the amplifier is conditionally stable from 5–11 GHz and above 22 GHz. G_{MS} is usually taken as the highest gain that can be easily achieved. However, higher gains can be achieved with more attention to stability, but usually amplification is available only over a narrow bandwidth.

Once a design is completed the only gain that matters is the transducer gain, G_T, which is the ratio of the transducer gain power delivered to a load to the power available from the source.

Table 11-8 Device gain metrics for the pHEMT transistor documented in Figure 11-8. Stability factors for the transistor are given in Tables 11-5 and 11-6.

Frequency (GHz)	Max. unilateral transducer gain $G_{TU,\max}$ (db)	Max. available power gain G_{MA} (db)	Max. stable power gain G_{MS} (db)	Unilateral power gain U (db)
0.5	36.6	–	30.1	41.9
1	31.1	–	27.1	39.4
2	24.9	–	24.2	34.4
3	21.3	–	22.4	29.7
4	18.7	–	21.3	25.6
5	17.0	18.6	20.4	23.6
6	15.7	17.0	19.6	22.6
7	14.5	15.5	18.7	20.77
8	13.6	14.0	18.2	17.9
9	12.4	13.0	17.7	16.4
10	12.2	13.2	16.6	17.8
11	12.1	13.7	15.7	19.5
12	11.8	–	14.8	20.5
13	11.5	–	14.2	20.5
14	11.2	–	13.5	21.2
15	11.6	–	13.1	27.6
16	11.6	–	12.7	24.2
17	11.1	–	12.3	21.4
18	10.5	–	11.9	17.0
19	8.88	–	11.5	12.9
20	7.91	–	11.2	11.4
21	7.48	–	10.9	11.0
22	6.55	8.42	10.4	9.22
23	5.46	6.62	9.97	7.43
24	4.62	5.55	9.62	6.22
25	4.23	5.17	9.05	5.73
26	3.61	4.27	8.57	4.79

11.7   Linear Amplifier Design

The design procedure for linear amplifiers is well developed and the strategy forms the basis of all amplifier design. An amplifier has three major components: an input matching network, an active device, and an output matching network (see Figure 11-22). There are a number of design choices to be made and these will be illustrated by considering the design of an amplifier for maximum gain using the discrete pHEMT transistor examined previously (see Figure 11-8 on Page 609). The design specifications are

Gain: maximum gain at 8 GHz
Topology: three two-ports (input and output matching networks, and the active device)

Stability: broadband stability
Bandwidth: maximum that can be achieved using two-
                        element matching networks
Source impedance: $Z_S = 50\,\Omega$
Load impedance: $Z_L = 50\,\Omega$

11.7.1   Bias Circuit Topology

The first design step is to choose a biasing configuration, and this is directly related to the output voltage swing supported. The inductively biased configuration in Figure 11-23 will be used. Here L_1 and L_2 are RF chokes and large enough to block RF. C_1 and C_2 are DC blocking capacitors that block DC but allow RF to pass with negligible impedance. The input matching network is attached to the RF IN terminal and the output matching network is attached to the RF OUT terminal. V_{DD} is the supply voltage and V_G is the DC gate voltage typically provided by an analog integrated circuit available in conjunction with most RF designs. An additional design objective is to absorb the biasing circuit into the matching networks.

11.7.2   Stability Considerations

It is not sufficient to consider a single frequency in design, as stability must be assured at low and high frequencies. The stability factor of the active device is shown in Table 11-6 on Page 626. This indicates that the device is unconditionally stable at 5 to 11 GHz and at 22 to 26 GHz. At the high-frequency end, the gain of the device reduces rapidly with increasing frequency as the capacitive parasitics begin dominating transmission through the device. Therefore it is reasonable to assume that the device is unconditionally stable above 22 GHz.

The first design step is to choose a matching network that will provide the appropriate impedances to ensure stability below 5 GHz and above 11 GHz. To do this the stability circles must be considered, as the device is only conditionally stable at these frequencies. The center and radius of the input and output stability circles are listed in Table 11-5 on Page 625. These are plotted in Figure 11-24 for selected frequencies.

11.7.3   Output Matching Network Design

The output stability circle at 1 GHz (Figure 11-24(a)) indicates that for stable amplification, the output matching network, as seen from the transistor, could look like a short circuit, a matched load, or a capacitor at low frequencies. Figure 11-24(c), the output stability circle at 16 GHz, indicates that for stable amplification the output matching network, as seen from the transistor, could look like an open circuit or a matched load at high frequencies. As expected, the output stability circle at 8 GHz indicates

Figure 11-22   Linear amplifier comprising input and output matching networks and an active device in a specific configuration forming a two-port.

Figure 11-23   Bias configuration for pHEMT amplifier. L_1 and L_2 are RF chokes and large enough to block RF. C_1 and C_2 are DC blocking capacitors that block DC but allow RF to pass with negligible impedance.

unconditional stability. Examining the two-element matching networks in Figure 8-13 on Page 422, there are three candidate output matching networks which are shown in Figure 11-25.

From Figure 11-8, the device S parameters at 8 GHz are as follows:

$$S_{11} = 0.486\angle140.4° \qquad S_{21} = 3.784\angle11.2°$$
$$S_{12} = 0.057\angle6.4° \qquad S_{22} = 0.340\angle-99.1°.$$

Design nearly always commences with the output matching network. To start, ignore S_{12} so that $\Gamma_{\mathrm{OUT}} = S_{22} = 0.340\angle-99.1$. There is little choice here as Γ_{OUT} depends on the input matching network which has not yet been designed. It would have been possible to begin with the input matching network and make this approximation for Γ_{IN}. However, experience is that the error introduced by starting with the output matching network is less. Once the output matching network has been designed, Γ_{IN} can be calculated without approximation. A thorough design would complete the first pass of the design and then make one more pass without the approximation that ignores S_{12}. That is, the output impedance of the active device is

$$Z_{\mathrm{OUT}} = Z_0 \frac{1 + \Gamma_{\mathrm{OUT}}}{1 - \Gamma_{\mathrm{OUT}}}$$

OUTPUT STABILITY CIRCLES

Figure 11-24 Input and output stability circles on the complex reflection coefficient planes for $|S_{11}| < 1$ and $|S_{22}| < 1$: (a), (b), and (c) output stability circle on the Γ_L plane at 1, 8, and 16 GHz; and (c), (d), and (e) input stability circle on the Γ_S plane at 1, 8, and 16 GHz.

$$= (50\,\Omega)\frac{1 + (0.340\angle - 99.1°)}{1 - (0.340\angle - 99.1°)}$$

$$= (50\,\Omega)\frac{1 + (-0.053774 - \jmath 0.335721)}{1 - (-0.053774 - \jmath 0.335721)}$$

$$= 36.153 - \jmath 27.447\,\Omega, \tag{11.104}$$

or $Y_{\text{OUT}} = 1/Z_{\text{OUT}} = 0.017547 + \jmath 0.013322\,S$. The output of the active device looks like a 56.99 Ω resistor in parallel with a capacitor with a reactance of $-75.064\,\Omega$. So taking into account the bias objectives and the available output matching networks in Figure 11-25, the matching network topology of Figure 11-25(c) will be used where the source is the active device here. So the output matching network problem is as shown in Figure 11-26.

The complete output matching problem is shown in Figure 11-26(a). In part, the parallel configuration of the active device output impedance was

Figure 11-25 Output matching network candidates required for out-of-band stability: (a) $R_S > R_L$; (b) $R_S < R_L$; and (c) $R_S > R_L$. (The active device is on the left.)

Figure 11-26 Steps in the design of the output matching network: (a) active device presents itself as a resistance in parallel with a capacitive reactance at the output matching network; (b) with inductor to resonate out active device reactance; (c) simplified matching network problem; and (d) final output matching network design.

chosen, as this is closer to reality since there is a capacitance at the output of the transistor. Resonance, as shown in Figure 11-26(b), will be used to cancel the effect of the active device capacitance, so that the matching problem

reduces to that shown in Figure 11-26(c). Using the procedure outlined in Section 8.4.2 on Page 424,

$$|Q_S| = |Q_P| = \sqrt{\frac{R_S}{R_L} - 1} = \sqrt{\frac{56.99}{50} - 1} = 0.3739 \qquad (11.105)$$

$$Q_S = \left|\frac{X_S}{R_L}\right| = \left|\frac{X_S}{50\,\Omega}\right| = 0.3739 \qquad (11.106)$$

$$Q_P = \left|\frac{R_S}{X_P}\right| = \left|\frac{56.99\,\Omega}{X_P}\right| = 0.3739, \qquad (11.107)$$

so

$$X_S = -18.70\,\Omega \quad \text{and} \quad X_P = 152.4\,\Omega. \qquad (11.108)$$

Now $X_x = -75.064\,\Omega$, so $75.064\,\Omega$ must be added to X_P in parallel, and the reactance of L_0 is $50.29\,\Omega$ thus (at 8 GHz)

$$L_o = 1.00\,\text{nH} \quad \text{and} \quad C_o = 1.064\,\text{pF}. \qquad (11.109)$$

11.7.4 Input Matching Network Design

The input stability circle at 1 GHz (Figure 11-24(d)) indicates that the input matching network, as seen from the transistor, could look like a short circuit, a matched load, or a capacitor at low frequencies. Figure 11-24(f), the input stability circle at 16 GHz, indicates that the output matching network, as seen from the transistor, could look like a short circuit or a matched load at high frequencies. Examining the two-element matching networks in Figure 8-13 on Page 422, three candidate output matching networks are shown in Figure 11-27.

The reflection coefficient looking into the output matching network from the active device is $\Gamma_L = S_{22}^* = 0.340\angle 99.1°$ because of the design decision to ignore S_{12} for the output matching network. Now that the output matching network has been designed, the feedback parameter need no longer be ignored. So

$$\Gamma_{\text{IN}} = S_{11} + \frac{S_{12}S21}{1 - S_{22}\Gamma_L} \qquad (11.110)$$

$$= (0.486\angle 140.4°) + \frac{(0.057\angle 6.4°)(3.784\angle 11.2°)}{1 - (0.340\angle -99.1°)(0.340\angle 99.1°)} \qquad (11.111)$$

$$= -0.1420 + j0.3835\,\Omega. \qquad (11.112)$$

That is, $Z_{\text{IN}} = 28.690 + j26.427\,\Omega$. So, taking into account the bias objectives and the output matching networks shown in Figure 11-27, the matching network topology of Figure 11-27(c) will be used where the source is the

Figure 11-27 Input matching network candidates required for out-of-band stability: (a) $R_S > R_L$; (b) $R_S < R_L$; and (c) $R_S > R_L$. (The active device is on the right.)

active device. The output matching network problem is as shown in Figure 11-28.

$$|Q_S| = |Q_P| = \sqrt{\frac{R_L}{R_S} - 1} = \sqrt{\frac{50}{28.69} - 1} = 0.8618 \qquad (11.113)$$

$$Q_S = \left|\frac{X_S}{R_L}\right| = \left|\frac{X_S}{26.427\ \Omega}\right| = 0.8618 \qquad (11.114)$$

$$Q_P = \left|\frac{R_S}{X_P}\right| = \left|\frac{50\ \Omega}{X_P}\right| = 0.8618 \qquad (11.115)$$

So

$$X_S = -43.09\ \Omega \quad \text{and} \quad X_P = 58.01\ \Omega \qquad (11.116)$$

Now $X_x = 26.427\ \Omega$, so $-26.427\ \Omega$ must be added to X_S, and the reactance of C_i is 69.52 Ω, thus (at 8 GHz),

$$L_i = 1.15\ \text{nH} \quad \text{and} \quad C_i = 0.286\ \text{pF}. \qquad (11.117)$$

11.7.5 Summary

The final schematic of the linear amplifier design is shown in Figure 11-29. The output matching network, L_2 and C_2, enabled the biasing inductor to be replaced by L_2. We were not so lucky with the input matching network, L_1 and C_1. The gate bias network is still required. L_3 should be a large enough value for it to act as an RF choke. A value of 10 nH provides a

Figure 11-28 Steps in the design of the input matching network: (a) active device presents itself as a resistance in parallel with a capacitive reactance at the output matching network; (b) with inductor to resonate out active device reactance; (c) simplified matching network problem; and (d) final output matching network design.

Figure 11-29 Final amplifier schematic.

reactance of approximately 500 Ω at 8 GHz. The value of $C_3 = 100$ pF is chosen large enough to provide an RF short and stabilize the DC bias V_G. This is a surprisingly simple circuit that provides maximum gain, assures out-of-band stability, and provides DC bias. Another design iteration with a more sophisticated input matching network may enable the bias inductor L_3 to be eliminated. As it is, the gate bias circuit further ensures stability at low frequencies, as then the gate tends to be shorted out. The amplifier has a calculated transducer gain of 13.2 dB which compares to the gains reported in Table 11-8 on Page 636, where gains were determined with S_{12} ignored.

Linear amplifier design for a specific gain is also possible. Now the errors involved in ignoring S_{12} during the design process are significant and a full bilateral treatment is required. This is described in references [160, 183, 192].

A linear class A X-band amplifier implemented in GaAs MESFET **Microwave Monolithic Integrated Circuit** (**MMIC**) technology is shown in Figure 11-30. This amplifier operates from 8 to 10 GHz producing 100 mW of total power. This is a two-stage amplifier as this was required to obtain the required gain, power and bandwidth. The input and output pads in a ground-signal-ground (GSG) configuration can be used for microwave probe testing and wire-bonding. The first transistor has two source connections (at the top and bottom), a gate connection on the left, and a drain connection on the right. The layout of the second transistor is the same. The larger second transistor has higher drain current. The source connections of the two transistors are grounded (indicated by the G connection). The matching network is implemented using stubs and capacitors. The second transistor has a feedback network with a spiral inductor and series resistor. Drain bias to the second transistor is through a spiral inductor.

A second, higher-power X-band MMIC using pHEMT transistors is shown in Figure 11-31. There are two stages, but the most striking difference with the previous MMIC is the use of multiple transistors in each of the two stages. The first stage has four transistors in parallel. Input matching network is integrated with a four-way power divider that drives the input of each transistor. An interstage matching network with four two-way power dividers drive the gates of eight pHEMTS. An eight-way power combiner takes the power from the drain of each second-stage transistor bringing them to a single output. Putting transistors in parallel requires close matching of the transistors. Practically the limit is eight transistors and tightly controlled IC technology is required.

11.8 Differential Amplifier

Differential amplifiers are the preferred amplifier topology in silicon monolithically integrated circuits including RFICs. Since substrate noise is common to the nodes of a differential amplifier, there is little differential substrate noise signal. Also, differential circuits lend themselves to current-mode biasing. The defining characteristic of a differential amplifier is that there are two signal paths which are differential. These amplifiers are also (but less commonly) called **balanced amplifiers**. When the RF signal on one of the differential paths is positive, the RF signal on the other is negative.

Figure 11-32(a) shows a **Fully Differential Amplifier** (**FDA**) with resistive biasing in the drain legs. As well as providing biasing current, the resistors are also the loads of the amplifier. The supply voltage of an RFIC can be quite low (a few volts or less), so choosing circuit topologies that provide for large voltage swings is important, particularly for an output amplifier. Differential topologies lead to an almost doubling of the output voltage swing compared to the output voltage swing of a single-ended amplifier.

(a)

(b)

(c)

Figure 11-30 A two-stage, two-transistor X-band (8–12 GHz) MMIC amplifier producing 100 mW of power: (a) photomicrograph with key networks identified, **G** indicates ground; (b) layout of the top spiral inductor; and (c) scanning electron microscope image of the crosssection of the **airbridge**.

Figure 11-31 An 8–12 GHz MMIC amplifier producing approximately 1 W of output power with key networks identified. (Courtesy Filtronic, PLC, used with permission.)

An FDA includes a common current source (see Figures 11-32(a) and 11-32(b)) that can be implemented quite simply using a single FET, as shown in the inset in Figure 11-32(a). Here the gate-source voltage is established by a bias voltage, V_B, which, from Figure 11-11(b), establishes a nearly constant drain current as long as there is sufficient drain-source voltage.

Common-Mode Rejection

One of the attributes that makes FDAs attractive is that they are relatively immune to substrate noise (noise in the substrate produced by other circuits) and they lend themselves to resistive biasing. The signal applied to the inputs of a differential amplifier have differential- and common-mode components. Referring to the differential amplifier in Figure 11-32(c), the

FULLY DIFFERENTIAL AMPLIFIER

(a) CURRENT
SOURCE

(b)

(c)

PSEUDO DIFFERENTIAL AMPLIFIER

(d)

(e)

(f)

Figure 11-32 Differential amplifiers: (a) fully differential amplifier (FDA); (b) FDA with inductive biasing; (c) schematic representation; (d) pseudo-differential amplifier (PDA); (e) PDA with inductive biasing; (f) schematic representation. The inset in (a) shows the implementation of current source as a single enhancement-mode MOSFET with a bias voltage, V_B, at the gate.

differential-mode input signal is

$$V_{id} = V_+ - V_- \tag{11.118}$$

and the common-mode input signal is

$$V_{ic} = \frac{V_+ + V_-}{2}. \tag{11.119}$$

Similarly the differential- and common-mode output signals are

$$V_{od} = V_{o+} - V_{o-} \tag{11.120}$$

and

$$V_{oc} = \frac{V_{o+} + V_{o-}}{2}, \tag{11.121}$$

respectively. The differential-mode voltage gain is

$$A_d = \frac{V_{od}}{V_{id}} \tag{11.122}$$

and the common-mode gain is

$$A_c = \frac{V_{oc}}{V_{ic}}. \tag{11.123}$$

For good noise immunity the common-mode gain should be low and the differential-mode gain should be high. The figure of merit that catches this is the **Common-Mode Rejection Ratio** (**CMRR**):

$$\text{CMRR} = \frac{A_d}{A_c};\tag{11.124}$$

the larger this is, the better. CMRR is usually expressed in decibels, and since CMRR is a voltage gain ratio, CMRR in decibels is $20\log(A_d/A_c)$. The current source at the source of the differential pair of the FDA has a large effect on the CMRR. The current source results in a large CMRR. This is shown in Example 11.2 on Page 650. Without this current source the CMRR of the FDA of Figure 11-32(a) would be one.

Output Voltage Swing

Single-ended amplifiers were discussed in Section 11.4.1 on Page 611, where it was shown that inductive biasing enables higher output voltage swings than possible with resistive biasing. A similar enhancement can be obtained with a differential amplifier. The inductively biased FDA of Figure 11-32(b) has a higher voltage swing than the resistively biased FDA of Figure 11-32(a). More can be achieved, however. The current sources at the common source point of the FDAs in Figures 11-32(a) and 11-32(b) limit the voltage swing, as there is a minimum drain-source voltage drop across the current source. When larger output voltage swings are required, the current source is eliminated and the resulting amplifier is called a **Pseudo-Differential Amplifier** (PDA), as shown in Figure 11-32(d). Again ,inductive biasing (see Figure 11-32(e)) almost doubles the possible voltage swing. The cost, however, is that the CMRR is one.

The output voltage waveforms for single-ended and differential amplifiers with and without inductive biasing are shown in Figure 11-33. Consider the differential resistively biased amplifier shown in Figure 11-33(c). This amplifier topology is also called an FDA. As can be seen in Figure 11-33(c)(ii), the differential voltage swing will be approximately $4V_{DD}$ less than the voltage drop across the current source. The supply voltage of RFICs is limited, so at the final output stage the current source is sometimes sacrificed so that larger voltage swings can be obtained. The resulting amplifier is the PDA, shown in Figure 11-33(d). The maximum output voltage swing is now $4V_{DD} - 2V_{DS,min}$—almost four times the voltage swing (or 16 times the power into the same load) of a single-ended resistively biased Class A amplifier.

Figure 11-33 Class A MOSFET amplifiers with output voltage waveforms: (a) single-ended amplifier with resistive biasing; (b) single-ended amplifier with inductive biasing; (c) fully differential amplifier with inductive biasing; and (d) pseudo-differential amplifier. Schematic is shown in (i), drain voltage waveforms in (ii), and differential output in (iii).

Figure 11-34 FET differential amplifier: (a) schematic; (b) small-signal model; (c) small-signal model for calculating differential gain; and (d) small-signal model for calculating common-mode gain.

EXAMPLE 11. 2   Calculation of Common-Mode Rejection Ratio

Determine the CMRR of the FET differential amplifier shown in Figure 11-34(a).

Solution:

The strategy for solving this problem is to develop the common-mode and differential-mode equivalent circuits and solve for the gain of each. The first step is to develop the small signal model shown in Figure 11-34(b). The differential input signal is V_{id} and the common input signal is V_{ic} so that the input voltage signals are

$$V_{i+} = \frac{1}{2}V_{id} + V_{ic} \quad \text{and} \quad V_{i-} = -\frac{1}{2}V_{id} + V_{ic}. \tag{11.125}$$

The expressions are similar for the output differential- and common-mode signals V_{od} and V_{oc}. This leads to the small-signal differential-mode model of Figure 11-34(c) and the small-signal common-mode model of Figure 11-34(d). From Figure 11-34(c), the output differential signal is

$$V_{od} = \frac{V_{id}}{2}\left[-g_m(r_d//R_L)\right] - \frac{-V_{id}}{2}\left[-g_m(r_d//R_L)\right] = \frac{-V_{id}g_m r_d R_L}{r_d + R_L}, \tag{11.126}$$

so the differential gain is

$$A_d = \frac{V_{od}}{V_{id}} = \frac{-g_m r_d R_L}{r_d + R_L}. \tag{11.127}$$

If, as usual, $r_d \gg R_L$, this becomes

$$A_d = -g_m R_L. \tag{11.128}$$

Focusing on the small-signal common-mode model of Figure 11-34(d) yields the output common-mode signal. The two halves of the circuit are now identical. The sum of the currents at S is zero, so

$$\frac{V_S}{2R_S} + \frac{V_S - V_{oc}}{r_d} - g_m V_{ic} = 0, \tag{11.129}$$

and the sum of the currents at the output terminal is zero, so

$$\frac{V_{oc}}{R_L} + \frac{V_{oc} - V_S}{r_d} + g_m V_{ic} = 0. \tag{11.130}$$

Eliminating V_S from these equations leads to

$$V_{oc} = \frac{-g_m r_d R_L V_{ic}}{(1 + g_m r_d) 2R_S + r_d + R_L}, \tag{11.131}$$

so the common-mode gain is

$$A_c = \frac{V_{oc}}{V_{ic}} = \frac{-g_m r_d R_L}{(1 + g_m r_d) 2R_S + r_d + R_L}. \tag{11.132}$$

If, as usual, $r_d \gg R_L$, this becomes

$$A_c = \frac{-g_m R_L}{1 + 2g_m R_S}. \tag{11.133}$$

The CMRR (when $r_d \gg R_L$) is

$$\text{CMRR} = \frac{A_d}{A_c} = \frac{-g_m R_L (1 + 2g_m R_S)}{-g_m R_L} = (1 + 2g_m R_S). \tag{11.134}$$

So the CMRR depends on the value of R_S. If there is no resistor at the common-mode source point, as it is in a pseudo-differential amplifier, $R_S = 0$ and so

$$\text{CMRR}|_{R_S = 0} = 1. \tag{11.135}$$

If there is an ideal current source at the source node then R_S is effectively infinite, and so

$$\text{CMRR}|_{\text{current source}} = \infty. \tag{11.136}$$

11.8.1 Hybrids and Differential Amplifiers

RFIC's use both fully differential and pseudo-differential signal paths. If signal swing is not a concern, a fully differential signal path would be preferred, particularly because of its immunity to noise and its bias stability. The additional transistors involved in realizing a differential circuit, however, reduce the available voltage swing, and hence power-handling capability. Pseudo-differential signaling uses, in effect, two parallel paths, each referred to ground but of opposite polarity. The signal on one of the parallel paths is the mirror image of the signal on the other. (That is, the signal is not truly differential, which would imply that it was floating or independent of ground.) Each of the parallel paths is unbalanced, but together their RF signal appears to be balanced, or pseudo-balanced. Another consideration is that in working with RFICs it is necessary to interface (unbalanced) microstrip circuits with the input and output of RFICs. The necessary functionality here is the need to split and combine signals, and to convert between balanced and unbalanced signals.

An FDA is shown in Figure 11-35(a). Both the input, $V_i = V_+ - V_-$, and the output, V_o, are differential. Figures 11-35(b) and 11-35(c) show a transformer being used to convert the differential output of the amplifier to an unbalanced signal that, for example, can be connected to a microstrip circuit. The output of many RFIC is pseudo differential as this signaling provides large voltage swings. A pseudo-differential amplifier is shown in Figure 11-36(d), but before dealing with manipulation of the signal path first consider the hybrid on its own.

Figure 11-36(a) shows how two pseudo-balanced signals can be combined to yield a single balanced signal. This $180°$ hybrid function is realized by a **center-tapped transformer**. The signal at Terminal 2 is referenced to ground, and these two terminals are Port 2. The image component of pseudo-differential signal is applied to Port 3 comprising Terminal 3 and ground. The balanced signal at Port 1 can be directly connected to a microstrip line which is, of course, unbalanced. Most implementations of hybrids at RF and microwave frequencies have ports that are referenced to ground. This is emphasized in Figure 11-36(b), making it easier to see how a $180°$ hybrid can be used to combine a pseudo-differential signal, as shown in Figure 11-36(c). This pseudo-differential-to-unbalanced interface is shown in Figures 11-36(e), 11-36(f), and 11-36(g).

Hybrids can be used at the input and output terminals of an RFIC pseudo-differential amplifier so that an unbalanced source can efficiently drive the RFIC and then the output of the RFIC can be converted to an unbalanced port to interface with unbalanced circuitry, such as filters and transmission lines. In the RFIC-based system in Figure 11-37, a $180°$ hybrid is first used as a **splitter** and then at the output as a **combiner**.

FULLY DIFFERENTIAL AMPLIFIER

Figure 11-35 Configurations providing an unbalanced output from the output of a differential amplifier: (a) fully differential amplifier (FDA); (b) FDA configured with a balun; (c) schematic representation; (d) pseudo-differential amplifier (PDA); (e) PDA configured with a $180°$ hybrid to provide an unbalanced output; (f) schematic representation; and (g) differential amplifier (DA) with transformer connection yielding an unbalanced output.

11.9 Distributed Biasing of Differential Amplifiers

In this section a broadband distributed balun-like section is presented as an alternative to inductor-based biasing of active differential circuits. The

(a)   (b)   (c)

Figure 11-36 Equivalent representations of a $180°$ hybrid connected to provide an interface between pseudo-differential balanced port and an unbalanced port: (a) schematic of a $180°$ hybrid with the isolation port terminated in a matched load; (b) hybrid showing two terminal representation of ports; and (c) a transformer configured as a $180°$ hybrid with pseudo-unbalanced-to-balanced configuration.

Figure 11-37 An RFIC with differential inputs and outputs driven by a $180°$ hybrid used as a splitter and followed by another $180°$ hybrid used as a combiner.

distributed biasing circuit discriminates between differential- and common-mode signals, resulting in rejection of common-mode signals.

One of the aims of this section is to show that network synthesis and invention of circuit topologies is an important part of RF and microwave engineering. This section presents one design example of network synthesis. The main sources of novel circuits for particular applications and sources of design concepts include microwave journals, patents, and conferences. This section presents an example of what can be achieved.

On-chip RF power amplifiers are often pseudo differential where the common current source typically used in fully differential circuits is sacrificed to enable a larger voltage swing, see Figure 11-38. Here the inductors present high RF impedance to the transistors (represented as transconductances) while providing a low impedance path for bias currents. With sufficiently high Q inductors, most of the RF energy is delivered to the load rather than being dissipated in the bias circuitry. With the relatively low Q of on-chip planar inductors, inductance peaking near self-resonance is utilized and this leads to narrowband operation. Thus in critical situations off-chip inductors are sometimes used in conjunction with on-chip active devices. In addition, inductive biasing of pseudo-differential circuits presents the same environment to common- and differential-mode signals so that the CMRR is 1.

Figure 11-38 A pseudo-differential circuit without a constant-current source, bias inductors, L, at the drains, parasitic supply resistance, R_{DD}, and single-ended load impedance, R_L.

Figure 11-39 Marchand balun-like biasing circuit with single-ended load resistance R_L (DC bias may be applied at Ports b using decoupling capacitors).

The biasing scheme considered here is suited to broadband biasing pseudo-differential circuits and results in high CMRR [119,193]. The biasing circuitry consists of a coupled line structure analogous to that of a Marchand balun [118] (see Figure 11-39 and Section 7.6.1 on Page 390). The balun structure is comprised of transmission lines connected to a pair of coupled lines that normally transforms a single-ended impedance at Port 1 to a balanced output at Ports 3 and 4. The capacitors in Figure 11-39 result

Figure 11-40   Modal sub-circuits of the inductor-based biasing circuit of Figure 11-38 including single-ended load resistance R_L.

in miniaturization since the distributed transmission lines are one-quarter wavelength long at a frequency well above the passband of the balun. In Section 11.9.1 the equations describing CMRR are first developed, as maximizing CMRR in pseudo-differential circuits is a key design objective. The design equations in the subsequent sections enable the design of the biasing circuit network with specified characteristics. In Sections 11.9.2–11.9.5 the use of the Marchand balun structure as a biasing circuit is investigated and analysis of the even- and odd-mode impedances of the balanced ports is presented.

11.9.1   Conventional Biasing

Multifunctional systems require broadband circuits and, for differential circuits, this requires large differential gain, A_d. At the same time it is desirable to minimize the common-mode gain, A_c, as the resulting high CMRR provides immunity to substrate-induced noise. With transistor transconductance, g_m, and total even- and odd-mode impedances, Z_{EVEN} and Z_{ODD}, presented to the drains of the transistors, the gains are approximately

$$A_d = g_m Z_{\text{ODD}} \text{ and } A_c = g_m Z_{\text{EVEN}}, \tag{11.137}$$

and so

$$\text{CMRR} = A_d/A_c = Z_{\text{ODD}}/Z_{\text{EVEN}}. \tag{11.138}$$

The desired amplifier characteristics are thus obtained by synthesizing the even- and odd-mode impedances.

Modal analysis of the inductor biasing circuit results in the circuit model shown in Figure 11-40 from which the total even-mode impedance is

$$Z_{\text{EVEN}}(s) = (sL + R_{DD}/2) \, //R_L, \tag{11.139}$$

where s is the Laplace operator, $//$ indicates a parallel connection, and the total odd-mode impedance is

$$Z_{\text{ODD}}(s) = sL//R_L. \tag{11.140}$$

Figure 11-41 Coupled lumped-element bandpass resonators: (a) single resonator coupled into by an inverter; and (b) a pair of resonators coupled by inverters.

Thus, unless the inductance is very large, there will be strong frequency variation of Z_{ODD} and Z_{EVEN}. Also, since R_{DD} is often negligible, the CMRR is 1. In differential-mode the gain is maximized over a broad frequency band for a specific single-ended load impedance, R_L. However, if the common-mode gain has prescribed frequency characteristics then the CMRR will be inversely proportional to the even-mode impedance. Thus low even-mode impedance of the biasing network is desirable as this suppresses common-mode noise. This defines a vital design objective of the new biasing circuit.

11.9.2 Design Based on Analogy to Coupled Resonators

Coupled resonator structures are appropriate topology choices that present different impedances for the common- and differential-mode signals. In essence, bias is also a common-mode signal and must be accommodated in the chosen topology. A class of appropriate topologies is based on lumped bandpass resonator(s) coupled by an impedance inverter, K, shown in Figure 11-41(a). (Note that an inverter and unitary element [UE] are synonymous.) The inverter element is an idealized component that is independent of frequency and can be approximated by a one-quarter wavelength long transmission line (see Section 10.8), as well as by various other circuit configurations including coupled lines (see Section 10.12 on Page 558). At resonance the bandpass resonator presents an open circuit and, following impedance inversion, the input impedance of the network is a short circuit. This, then, is the kind of network required to be presented to the common-mode signals. The second part of the network synthesis problem is creation of the high impedance condition for differential-mode signals. These issues will be considered separately. Figure 11-41(b) illustrates a pair of bandpass resonators coupled by inverters. At resonance the input impedance of the network is infinite. This is because the high impedance of Resonator 2 transforms to a short circuit after the second inverter, K_2, and this subsequently appears as an open circuit at the input of the network. So by adding an extra resonator another topology suitable for differential-

Figure 11-42 Modal sub-networks of the balun-like biasing circuit of Figure 11-39 with single-ended load resistance R_L (assuming negligible power supply source impedance, R_{DD}).

mode operation is derived. If a load is now connected in parallel with this network then at resonance this will be the total impedance presented to the output transistors of the pseudo-differential amplifier. In summary, presenting different impedances for common- and differential-mode signals is desired. It will be shown in the next section that the Marchand balun structure has the requisite properties.

11.9.3 Distributed Biasing

Consider now the topology of the Marchand balun of Figure 11-39 with balanced Ports 3 and 4 connected to the output of the pseudo-differential circuit. The aim of this subsection is to show that this topology has the requisite even- and odd-mode impedances. Performing modal analysis at the balanced ports, the modal subnetworks of Figure 11-40 are obtained. In analyzing this structure use is made of the network model of a pair of coupled lines (Figure 9-19). Direct application of the network model and application of the relevant port conditions dictated by the modal subnetworks of the biasing circuit at Point X simplifies the equivalent subnetworks of Figure 11-42 to those of Figure 11-43. Application of the relevant Kuroda transformation (see Section 10.11.5 on Page 555) to distribute the inductance in Figure 11-43(c) results in the transformed subcircuits of Figure 11-43(d). Scrutiny of Figures 11-43(b) and 11-43(d) reveals that they are in fact analogous to the lumped-element coupled-resonator examples discussed earlier. The major differences are that transmission lines replace the inverters and the additional resistor R_c represents the loss associated with each capacitor. Also in the distributed domain each inductor or capacitor represents a short- or open-circuited stub, respectively, which is one-quarter wavelength long at the frequency f_q. The open-circuited stubs may, however, be approximated by lumped

Figure 11-43 Equivalent subcircuits of Figure 11-42: (a) utilizing the network model; (b) even-mode subcircuit; (c) odd-mode subcircuit; and (d) odd-mode subcircuit after Kuroda transformation with $\bar{Z}_1 = 1/Z_{0e}$, $\bar{Z}_2 = (Z_{0e} - Z_{0o})/(2Z_{0e}Z_{0o})$. $R_L = \infty$, $R_L = 100 \; \Omega$, $R_L = 50 \; \Omega$.

capacitors for miniaturization purposes. Thus over the operating band the performance of the even- and odd-mode subcircuits is very close to that of their lumped-element counterparts. In summary, centered at a frequency f_0, the structure of the Marchand balun maximizes differential power delivery to the load while minimizing the gain of common-mode signals.

DC bias may be easily applied through the short-circuited ports (labeled b in Figure 11-39) with the use of decoupling capacitors. The mathematical formulations for the total even- and odd-mode impedances with lumped capacitors are derived from Figure 11-43(b) and 11-43(c) as follows:

$$Z_{\text{EVEN}}(S) = \frac{N_1 S^2 + N_2 S}{D_1 S^2 + D_2 S + D_3} \tag{11.141}$$

with coefficients

$$N_1 = j\left(2Z_{0t}^2 Z_1 \pi f C R_L\right) \tag{11.142}$$

$$N_2 = Z_1 R_L Z_{0t} + Z_{0t}^2 R_L \tag{11.143}$$

$$D_1 = (Z_1 R_L) + j\left(2Z_{0t}^2 \pi f C Z_1\right) \tag{11.144}$$

$$D_2 = \left(Z_{0t}^2 + Z_{0t} Z_1\right) + j\left(2Z_{0t} R_L \pi f C Z_1\right) \tag{11.145}$$

$$D_3 = Z_{0t} R_L, \tag{11.146}$$

and

$$Z_{\text{ODD}}(S) = \frac{N_1 S^3 + N_2 S^2 + N_3 S}{D_1 S^3 + D_2 S^2 + D_3 S + D_4} \tag{11.147}$$

with coefficients

$$N_1 = Z_{0t}^2 R_L Z_1 \tag{11.148}$$

$$N_2 = j\left(2Z_2 Z_{0t}^2 R_L \pi f C Z_1\right) \tag{11.149}$$

$$N_3 = Z_1 Z_2 Z_{0t} R_L + Z_2 Z_{0t}^2 R_L \tag{11.150}$$

$$D_1 = Z_{0t}^2 Z_1 \tag{11.151}$$

$$D_2 = (Z_1 Z_2 R_L + Z_1 Z_{0t} R_L) + j\left(2Z_2 Z_{0t}^2 \pi f C Z_1\right) \tag{11.152}$$

$$D_3 = \left(Z_2 Z_{0t} Z_1 + Z_2 Z_{0t}^2\right) + j\left(2Z_2 Z_{0t} \pi f C Z_1 R_L\right) \tag{11.153}$$

$$D_4 = Z_2 Z_{0t} R_L. \tag{11.154}$$

In Equations (11.141) and (11.147) S is the Richards' variable defined in Section 10.11 on Page 552 as $j\tan(\theta) = j\tan\left((\pi/2)\left(f/f_q\right)\right)$ and f_q is the frequency at which the distributed lines of the balun are one-quarter wavelength long. In the above equations Z_{0t} is the characteristic impedance of the uncoupled transmission lines while

$$Z_1 = \frac{\sqrt{Z_{0e} Z_{0o}}}{\sqrt{1 - \left(\frac{Z_{0e} - Z_{0o}}{Z_{0e} + Z_{0o}}\right)^2}} \tag{11.155}$$

and

$$Z_2 = \sqrt{Z_{0e} Z_{0o}} \sqrt{1 - \left(\frac{Z_{0e} - Z_{0o}}{Z_{0e} + Z_{0o}}\right)^2} \left(\frac{Z_{0e} - Z_{0o}}{Z_{0e} + Z_{0o}}\right)^{-2}, \tag{11.156}$$

with Z_{0e} and Z_{0o} being the even- and odd-mode impedances of the coupled lines forming the balun. The total even- and odd-mode impedances presented to the pseudo-differential active circuit are in general complex, and thus may be written in the following notation:

$$|Z_{\text{EVEN}}(S)| = \sqrt{\Re\left\{Z_{\text{EVEN}}(S)\right\}^2 + \Im\left\{Z_{\text{EVEN}}(S)\right\}^2} \qquad (11.157)$$

$$|Z_{\text{ODD}}(S)| = \sqrt{\Re\left\{Z_{\text{ODD}}(S)\right\}^2 + \Im\left\{Z_{\text{ODD}}(S)\right\}^2}. \qquad (11.158)$$

Hence the procedure for determining the circuit parameters of the new biasing circuit is fairly simple. Suitable choices of the characteristic impedances Z_1 and Z_2 solve Equations (11.155) and (11.156) for the even- and odd-mode characteristic impedances Z_{0e} and Z_{0o} respectively. Selection of the characteristic line impedance Z_{0t} together with the choice of capacitance value, C, fully determines Equations (11.141) and (11.147). The magnitudes of the total even- and odd-mode impedance functions may then be evaluated using Equations (11.157) and (11.158). The commensurate length of the transmission lines will be determined by the value of C. Zero capacitance results in the lines being one-quarter wavelength long at the center of the operating band, otherwise miniaturization of the lines is feasible. An example that shows the overall characteristics of the magnitudes of the even- and odd-mode impedances without lumped capacitors is depicted in Figure 11-44 for a range of single-ended load resistances, R_L.

11.9.4 Operating Bandwidth

The operating bandwidth of the balun is defined here by the crossover frequencies where

$$|Z_{\text{EVEN}}(S)| = |Z_{\text{ODD}}(S)|. \qquad (11.159)$$

At the crossover frequency points (see Figure 11-44), the CMRR of the pseudo-differential circuit is 1 and the balun has similar performance to that of the nondiscriminative inductor-based biasing circuit. For a certain choice of circuit parameters the frequency points at which the even- and odd-mode impedances cross over may be altered, facilitating adjustment of the discriminative bandwidth. This is demonstrated in Figure 11-45 through a set of examples using the design procedure described in the above section. It is worth emphasizing that the total even- or odd-mode impedances of Equations (11.141) and (11.147) are the impedances of the input of the balun in parallel with the single-ended load resistance, R_L. This implies that in order for the pseudo-differential circuit to deliver its power to the load, it is necessary that the output signal at each arm sees the correct impedance. Of course there will only be a perfect match at the center of the band (since the input balun impedance is then infinite) but not at the crossover frequency

Figure 11-44 Calculated magnitudes of the total even- and odd-mode impedances for a range of single-ended load resistances with circuit parameters $Z_{0e} = 208\ \Omega$, $Z_{0o} = 51\ \Omega$, $Z_{0t} = 137\ \Omega$, and $C = 0$ pF.

points. However, a slight mismatch will only degrade the transfer of power slightly and is therefore acceptable. Figure 11-44 illustrates this argument for a 50 Ω system for a particular choice of balun circuit parameters. In this example the magnitudes of the total even- and odd-mode impedances at the crossover frequencies are approximately 43 Ω. However, for a 100 Ω system it is obvious (see Figure 11-44) that the impedance levels are much lower than 100 Ω at the cross over frequencies, causing a big mismatch that needs to be compensated for by adjusting the element values of the balun. Generally the higher the system impedance, the higher the coupling required between the coupled lines of the balun.

11.9.5 Effect of Loss on Distributed Biasing

The remaining design aspect to be discussed is the effect of transmission line loss and also loss of the lumped capacitors if circuit miniaturization is desired. Modification of the total even- and odd-mode impedances to take into account the effect of substrate, conductor, and capacitor shunt

Figure 11-45 Calculated magnitudes of the total even- and odd-mode impedances for an infinite single-ended load resistance, R_L with the following circuit parameters: for Δf_1, $Z_{0e} = 157.32\ \Omega$, $Z_{0o} = 58.45\ \Omega$, $Z_{0,t} = 90\ \Omega$, $C = 0$ pF; for Δf_2, $Z_{0e} = 205.36\ \Omega$, $Z_{0o} = 47.82\ \Omega$, $Z_{0,t} = 100\ \Omega$, $C = 0$ pF; and for Δf_3, $Z_{0e} = 307.26\ \Omega$, $Z_{0o} = 22.24\ \Omega$, $Z_{0t} = 80\ \Omega$, $C = 0$ pF.

resistance, R_c, leads to

$$Z_{\text{EVEN}}(S) = \frac{N_1 S^2 + N_2 S}{D_1 S^2 + D_2 S + D_3} \qquad (11.160)$$

with coefficients

$$N_1 = \left(Z_{0t}^2 Z_1 R_L\right) + \jmath \left(2 Z_{0t}^2 Z_1 \pi f C R_L R_C\right) \qquad (11.161)$$

$$N_2 = Z_1 R_L Z_{0t} R_C + Z_{0t}^2 R_L R_C \qquad (11.162)$$

$$D_1 = \left(Z_1 R_L R_C + Z_{0t}^2 Z_1\right) + \jmath \left(2 Z_{0t}^2 \pi f C Z_1 R_C\right) \qquad (11.163)$$

$$D_2 = \left(Z_{0t}^2 R_C + Z_{0t} Z_1 R_C + Z_{0t} Z_1 R_L\right) + \jmath \left(2 Z_{0t} R_L \pi f C Z_1 R_C\right) \qquad (11.164)$$

$$D_3 = Z_{0t} R_L R_C, \qquad (11.165)$$

and

$$Z_{\text{ODD}}(S) = \frac{N_1 S^3 + N_2 S^2 + N_3 S}{D_1 S^3 + D_2 S^2 + D_3 S + D_4} \qquad (11.166)$$

with coefficients:

$$N_1 = Z_{0t}^2 R_L Z_1 R_C \tag{11.167}$$

$$N_2 = \left(R_L Z_2 Z_{0t}^2 Z_1\right) + \jmath \left(2 Z_2 Z_{0t}^2 R_L \pi f C Z_1 R_C\right) \tag{11.168}$$

$$N_3 = Z_1 Z_2 Z_{0t} R_L R_C + Z_2 Z_{0t}^2 R_L R_C \tag{11.169}$$

$$D_1 = Z_{0t}^2 Z_1 R_C \tag{11.170}$$

$$D_2 = \left(Z_1 Z_2 R_L R_C + Z_1 Z_{0t} R_L R_C + Z_{0t}^2 Z_1 Z_2\right) + \jmath \left(2 Z_2 Z_{0t}^2 \pi f C Z_1 R_C\right) \tag{11.171}$$

$$D_3 = \left(Z_2 Z_{0t} Z_1 R_C + Z_2 Z_{0t}^2 R_C + Z_1 Z_2 Z_{0t} R_L\right) + \jmath \left(2 Z_2 Z_{0t} \pi f C Z_1 R_L R_C\right) \tag{11.172}$$

$$D_4 = Z_2 Z_{0t} R_L R_C \ . \tag{11.173}$$

Now S is the general Richards' transformation, accounting for loss defined as

$$S = \tanh\left(\alpha + \jmath\theta\right) = \tanh\left(\alpha + \jmath\left(\pi/2\right)\left(f/f_r\right)\right), \tag{11.174}$$

and α is the attenuation constant in Nepers/m. With no miniaturization capacitors (i.e., $C = 0$, $R_c = \infty$) and a finite value of α, the same circuit parameters that generated the plots of Figure 11-44 are now fed back into the modified equations (Equations (11.160) and (11.166)) from which their magnitudes are evaluated using Equations (11.157) and (11.158). Figure 11-46 highlights the effect of parasitic loss. This plot assumes infinite single-ended load impedance (i.e., it represents the input impedances of the balun itself). The effect of loss manifests itself as a degradation of the magnitudes of the even- and odd-mode impedance levels. At resonance, the odd-mode impedance is very high, but finite, and likewise the opposite is true for the even-mode impedance. These impedances, in parallel with the single-ended load impedance, directly impact the achievable CMRR of the system and increase insertion loss. However, the plots of Figure 11-46 assume no capacitors for miniaturization purposes. Normally, lumped capacitors will possess finite Q that will further deteriorate the CMRR and reduce power delivery to the next stage. This point is demonstrated by measurements in the next section.

11.9.6 Experimental Results

A differential power amplifier (HELA-10B) from Mini-Circuits [194] with a gain of 10 dB was selected for experimental proof of concept. The chip is powered by a single +12 V DC power supply via a pair of external lumped inductors and consists of a pair of amplifiers in a pseudo-differential amplifier configuration. As the amplifiers are on the same chip, their gains and phase are well matched. The amplifier has an excellent second-order

Figure 11-46 Calculated magnitudes of lossy even- and odd-mode impedances for an infinite single-ended load resistance, R_L, with circuit parameters: $Z_{0e} = 208\ \Omega$, $Z_{0o} = 51\ \Omega$, $Z_{0t} = 137\ \Omega$, $C = 0\ pF$, $\alpha = 0.2$ Nepers/m.

intercept of 88 dBm operating from 50 MHz to 1 GHz in a 50 Ω system. A number of distributed biasing circuits were implemented and integrated with the pseudo-differential amplifier. The design objective was to achieve broadband bandwidth of 1.5:1, and high odd-mode and low even-mode RF bias circuit impedances.

Inductor-Based Biasing

The amplifier was tested first with a conventional inductor-based biasing circuit. A pair of high-Q lumped inductors of value 0.75 μH, were used and the circuit tested in both differential and common modes. The measured common- and differential-mode gains are 10 dB and 9.5 dB, respectively, as shown in Figure 11-47(a). This illustrates the nondiscriminatory bandwidth produced using the conventional inductor-based biasing scheme.

Distributed Biasing

To verify the properties of the differential biasing circuit, the Marchand balun structure with a center frequency of 580 MHz was designed. The optimized balun parameters are $Z_{0e} = 208\ \Omega$, $Z_{0o} = 51\ \Omega$, $Z_{0t} = 137\ \Omega$ and

Figure 11-47 Measured common-mode gain, A_c, and differential-mode gain, A_d: (a) with inductor-based biasing circuit with lumped inductors of 0.75 μ H; and (b) with balun-like biasing circuit without lumped capacitors with circuit parameters $Z_{0e} = 208\,\Omega$, $Z_{0o} = 51\,\Omega$, $Z_{0t} = 137\,\Omega$, and $\tan \delta = 0.016$.

$C = 0\,\mathrm{pF}$. The even- and odd-mode impedances were calculated based on a choice of $Z_1 = 129.5\ \Omega$, $Z_2 = 222.92\ \Omega$ from Equations (11.141)–(11.158). By referring to Figure 11-44 it can be seen that when the single-ended load impedance is 50 Ω the magnitude of the total odd-mode impedance is desirably flat over the specific frequency band. This represents optimum power transfer to the load in a 50 Ω system. The HELA-10B chip was integrated with the balun on an FR-4 PCB with a substrate thickness of 6.2 mil (0.157 mm), relative dielectric constant of 4.7, and loss tangent of 0.016. The above impedances were then converted into physical dimensions

using a commercial CAD tool and the final circuit optimized. The resulting coupled lines were found to be 7 mil (0.177 mm) wide, 7 mil apart, and 3031 mil long. The transmission lines connecting the output ports of the RF chip to the coupled lines were 7 mils wide, and 3031 mil long. The differential- and common-mode gains are shown in Figure 11-47(b). While the differential-mode gain is more than 10 dB in the center of the band, the common-mode gain dips to -10 dB at the resonant frequency. Thus the device now discriminates between the common- and differential-mode signals. The measured bandwidth, defined as $(\Delta f/f_0)$ is 34% of the resonant frequency of 580 MHz, that is 1.4:1 bandwidth.

Distributed Biasing With Miniaturization Capacitors

In another design operating at 580 MHz, the transmission lines were chosen to be resonant at $3\times580 = 1740$ MHz. This yields a reduction in overall size by a factor of three, with $C = 5.6$ pF. In Figure 11-48(a) the measured response (with high-Q capacitors, $Q \approx 400$) is depicted. It is worth investigating the effect of loading the capacitors with resistors, as this drops their effective Q. This was done with the objective of replacing the lumped capacitors with tunable components having moderate Qs to improve overall system functionality. For example, the Q of a BST capacitor ranges between 30 and 60. Figure 11-48(b) illustrates the gains after loading the lumped capacitors with resistors to drop their Qs to about 30. It is obvious from Figure 11-48 that the loss degrades the performance of the circuit. In the case of high-Q capacitors, the measured differential-mode gain is 10 dB. However, the common-mode gain is hardly affected and notches at the center frequency to -10 dB. In the case of low-Q capacitors, the common-mode gain only notches to 0 dB and the differential gain reduces from 10 dB to 9 dB. The measured bandwidth in the two cases is 31% of the resonant frequency of 580 MHz, that is 1.36:1 bandwidth.

It is also observed from Figure 11-48 that the common-mode gain does not notch exactly at the center of the operating band, as was the case without lumped capacitors (see Figure 11-47(b)). This implies that the frequency at which the odd-mode impedance of the balun peaks is different from the frequency at which its even-mode impedance is a minimum. This conflict arises due to the addition of lumped capacitors that cause asymmetry in the characteristics of the even- and odd-mode impedances. This may be resolved by fine-tuning of the lumped capacitors and optimization of the rest of the circuit elements of the equivalent even- and odd-mode subcircuits of the balun structure.

11.9.7 Summary

One of the aims of this section was to show that network synthesis and the invention of appropriate topologies is an essential part of RF and

(a)

(b)

Figure 11-48   Measured common-mode gain, A_c, and differential-mode gain, A_d: (a) with miniaturized balun-like biasing circuit using high Q lumped capacitors with circuit parameters $Z_{0e} = 208\ \Omega$, $Z_{0o} = 51\ \Omega$, $Z_{0t} = 137\ \Omega$, $C = 5.6\ pF$, $R_c = \infty$, and $\tan\delta = 0.016$; and (b) with miniaturized balun-like biasing circuit using high-Q lumped capacitors loaded with resistors with circuit parameters: $Z_{0e} = 208\ \Omega$, $Z_{0o} = 51\ \Omega$, $Z_{0t} = 137\ \Omega$, $C = 5.6\ pF$, $R_c = 1.2\ k\Omega$, and $\tan\delta = 0.016$.

microwave circuit design. The distributed biasing circuit presented here addresses two major problems that result from inductive biasing: the limited Q of inductors and the unity CMRR of pseudo-differential amplifiers. The physical structure comprises transmission lines and a pair of identical

Figure 11-49 DC and RF loadlines: (a) loadlines of Class A, B, and C amplifiers; and (b) loadlines of switching amplifiers.

coupled lines in the form of a Marchand balun. The circuit is inherently discriminatory, yielding finite CMRR. With the distributed biasing circuit, low-loss transmission line sections almost eliminate bias circuit loss and at the same time provides high CMRR. A more sophisticated synthesis resulting in distributed biasing of differential amplifiers with broadband common-mode suppression is presented in Section 13.3 on 766.

11.10   Switching Amplifiers

Switching amplifiers are a conceptual departure from Class A, AB, B, and C amplifiers, as can be seen in the typical AC loadline of a switching amplifier, shown in Figure 11-49(b), compared to the loadlines of Class A, AB, B, and C amplifiers, shown in Figure 11-49(a). This dynamic loadline is obtained by presenting the appropriate harmonic impedances to the transistor amplifier. The particular scheme of harmonic termination (e.g., short or open circuits at the even and odd harmonics) leads to the designation of a switching amplifier as Class D, E, F, etc. The key characteristic of a switching amplifier is that when there is current through the transistor, there is negligible voltage across the output. Also, when there is voltage across the transistor, there is little current through it (see Figure 11-49(b)). The power dissipated by the transistor is approximately the product of the current through it and the voltage across the output. Thus the switching amplifier consumes very little DC power, transferring nearly all of the DC power to the RF signal. Bandpass filtering of the output of the amplifier results in a final RF output with little distortion. Switching amplifiers are emerging as the preferred

Table 11-9   Theoretical maximum efficiencies of amplifier classes.

Amplifier Class	Maximum Efficiency
Class A (resistive bias)	25%
Class A (inductive bias)	50%
Class B	78.53%
Class C	100%
Class E	96%
Class F	88.36%

Table 11-10   Efficiency reductions due to signal type. The class A amplifier uses inductive drain biasing.

Signal	PAR (dB)	Efficiency Reduction Factor	Class A (L bias) PAE	Class E PAE
FSK (MSK, GMSK)	0	1.0	50%	96%
QPSK	3.6	0.437	21.9%	42%
π/4DQPSK	3.0	0.501	25.1%	48.1%
OQPSK	3.3	0.467	23.4%	44.8%
8PSK	3.3	0.467	23.4%	44.8%
64QAM	7.8	0.166	8.3%	15.9%

linear amplifier in both handsets and base stations of cellular systems.

The theoretical maximum power-added efficiencies achieved by the various amplifier classes are given in Table 11-9. With modulated signals, the maximum efficiencies cannot be achieved, so that typically the average input power of the amplifier must be backed off by the PAR of the signal so that the peak carrier portion of the signal has limited distortion. Generally the acceptable distortion of the peak signal occurs at the 1 dB compression point of the amplifier. This is only an approximate guide, but useful. The PARs of several digitally modulated signals are given in Table 11-10, together with their impact on efficiency. If there are two carriers, then the PAR of the combined signal will be higher, requiring greater amplifier backoff [195]. In practice, the efficiencies will differ from these theoretical values because of loss in the amplifier and the trade-off between efficiency and distortion. This is because the PAR does not fully capture the statistical nature of signals, and because of coding and other technologies that can be used to reduce the PAR.

11.11 Noise

11.11.1 Introduction

Noise together with nonlinear distortion set the bounds on the signals that can be processed in an RF circuit. Noise establishes the minimum detectable signal, while nonlinear distortion, by introducing distortion of the constellation diagram, sets the level of the largest signal from which information can be reliably extracted. The range is referred to as dynamic range and is one of the performance limits characterizing analog circuits. In this section, noise is considered first and then in the next expressions for dynamic range are developed.

11.11.2 Source of Noise

Noise is due to random fluctuations. but does not always have a flat frequency spectrum as would be expected from a Gaussian uniformly random process. A comprehensive treatment of noise requires an understanding of random processes [196]. The review here of sources of noise is derived from Kriplani et al. [196], Kriplani [197], and Hartnagel [198].

Noise in RF and microwave systems includes noise from the environment as well as noise generated within the circuitry itself. Noise from the environment can have galactic origins, when it is known as cosmic background noise, from black-body radiation, or can be artificially generated noise. In cellular communication systems the major source of interference is from other phones and base stations in the cellular system. Provided this is uniformly random over the communication band, it can be treated as random noise. Uniformly random noise is also called white noise and can be modeled by a resistor held at a what is called the noise temperature.

A noisy resistor generates **white noise** which has a uniform power spectral density (PSD), as shown in Figure 11-50. A noisy resistor can be modeled by a noise-free resistor and a random voltage or current source denoted by v_n and i_n, respectively (see Figure 11-51). The sources v_n and i_n are random and their PSDs are related by

$$S_{vn}(f) = R^2 S_{in}(f). \qquad (11.175)$$

Thermal noise sources are random up to frequencies of several hundreds of gigahertz (above which quantum effects become important) and the noise PSD of a resistor R at temperature T is [198]

$$S_{vn}(f) = \frac{v_n^2}{B} = 4kTR, \qquad (11.176)$$

so that

$$v_n^2 = 4kTBR\,. \qquad (11.177)$$

Figure 11-50 Normalized power spectral density as a function of frequency. After [197]

(a)   (b)

Figure 11-51 Thermal noise equivalent circuits: (a) noisy resistor modeled as a noise free resistor in series with a random noise voltage source v_n; and (b) noisy resistor modeled as a noise free resistor in parallel with a random noise current source i_n.

Here k is the Boltzmann constant, T is in kelvin, and B is the bandwidth in hertz. Two noise voltage sources, v_{n1} and v_{n2}, are in series and can be partly correlated. The two noise sources can be replaced by a single source V_n where

$$v_n^2 = v_{n1}^2 + v_{n1}^2 + 2C_{n1,n2}v_{n1}v_{n2}. \qquad (11.178)$$

$C_{n1,n2}$ is the **correlation coefficient**, and $-1 \leq C_{n1,n2} \leq 1$. If the sources are uncorrelated, as they would be for two resistors, $C_{n1,n2} = 0$ and

$$v_n^2 = v_{n1}^2 + v_{n1}^2 \qquad (11.179)$$

Correlation of noise sources is important to modeling noise in transistors.

Shot noise is particularly important with semiconductor devices but was first observed by Schottky in 1926 in vacuum tubes. It is due to the discrete nature of current flow and is important when there is a region that is

scarce of free carriers. In transistors and diodes, carriers cross a depletion region and generate a pulse of current. The noise that is created by the separable pulses has a flat frequency spectrum and can be modeled by a white noise source. Shot noise, however, is proportional to current flow and hence biasing of RF circuits should minimize DC current relative to the RF current. The PSD of shot noise current is

$$S_i(f) = 2qI(t), \tag{11.180}$$

where I is the instantaneous current. So shot noise varies during an RF cycle as the current flow varies.

The third type of noise that is of concern with RF and microwave circuits is flicker noise, sometimes called $1/f$ (one over f) noise because of the power spectral density shape. Flicker noise is due to traps in a semiconductor and surface traps. A free carrier is immobilized or trapped when it falls into a trap, that is, a recombination center. When several such carriers are trapped, it means that they are not available for conduction and as a result, the resistance of the semiconductor is modulated. The fluctuations are characterized by multiple relaxation times. The result of a uniform distribution of relaxation times is PSD that has a $1/f$ shape.

Random noise and interference in RF an microwave systems are often treated the same, and they can be if both are uniformly random processes.

11.11.3 Noise Measures

Amplifiers, filters, and mixers in an RF front end process (e.g., amplify, filter, and mix) input noise the same way as an input signal. In addition, these components can contribute excess noise of their own. Without loss of generality, the following discussion will consider noise with respect to the amplifier shown in Figure 11-53(a), where v_s is the input signal. The noise signal, with source designated by v_n, is uncorrelated and random and must be described as an RMS voltage or by its noise power. The most important noise-related metric is the SNR. With the noise power input to the amplifier, N_i, and the signal power input to the amplifier, S_i, the input SNR is $\mathrm{SNR}_i = S_i/N_i$. If the amplifier is noise free then the input noise and signal powers are amplified by the power gain of the amplifier, G. Thus the output noise power is $N_o = GN_i$, the output signal power is $S_o = GS_i$, and the output SNR is $\mathrm{SNR}_o = S_o/N_o = \mathrm{SNR}_i$.

In practice, an amplifier is noisy, with the addition of excess noise, N_e, indicated in Figure 11-53(b). The excess noise originates in different components in the amplifier and is either referenced to the input or the output of the amplifier. Most commonly it is referenced to the output so that the total output noise power is $N_o = GN_i + N_e$. In the absence of a qualifier, the **excess noise** is referred to the output. N_e is not measured directly. Instead, the ratio of the SNR at the input to that at the output is

(a)

Frequency

(b)

Figure 11-52 Shot noise: (a) normalized transient shot noise voltage in time; and (b) normalized power spectral density as a function of frequency. After Kriplani [197].

(a)   (b)   (c)

Figure 11-53 Noise and two-ports: (a) amplifier; (b) amplifier with excess noise; and (c) noisy two-port network.

called the **noise factor**, F:

$$F = \frac{\mathrm{SNR}_i}{\mathrm{SNR}_o},$$

(11.181)

and this is the way it is normally measured. If the circuit is noise free then $\mathrm{SNR}_o = \mathrm{SNR}_i$ and $F = 1$. If the circuit is not noise free, then $\mathrm{SNR}_o < \mathrm{SNR}_i$ and $F > 1$. F can be related to the excess noise produced in the circuit and from this relationship the noise of a cascaded system can be calculated. With the excess noise, N_e, referred to the circuit output,

$$
\begin{aligned}
F &= \frac{\mathrm{SNR}_i}{\mathrm{SNR}_o} = \frac{\mathrm{SNR}_i}{1} \frac{1}{\mathrm{SNR}_o} = \frac{S_i}{N_i} \frac{N_o}{S_o} \\
&= \frac{S_i}{N_i} \frac{GN_i + N_e}{GS_i} \\
&= 1 + \frac{N_e}{GN_i}.
\end{aligned}
\tag{11.182}
$$

One of the conclusions that can be drawn from this is that the noise factor, F, depends on the available noise power at the input of the circuit. As a standard reference, the available noise power, N_R, from a resistor at **standard temperature** T_0 (290 K), and over a bandwidth B is used:

$$
N_i = N_R = kT_0 B,
\tag{11.183}
$$

where k is **the Boltzmann constant**. If the input of an amplifier is connected to this resistor and all of the noise power is delivered to the amplifier then

$$
\begin{aligned}
F &= 1 + \frac{N_e}{GN_i} \\
&= 1 + \frac{N_e}{GkT_0 B}.
\end{aligned}
\tag{11.184}
$$

Several random physical processes inside a circuit contribute to excess noise, and not all of these process vary linearly with temperature. Consequently F is a function of temperature, although usually a weak one. It is also a function of bandwidth, and there is a problem in using F with cascaded systems in which bandwidths vary for different subsystems. Even with all these problems, F is the most important measure used to characterize noise. It can be used to determine the noise performance of a cascade, when the noise factors and gains of the subsystem constituents are known. F is the ratio of powers, and when expressed in decibels, the **noise figure** (**NF**) is used:

$$
\begin{aligned}
\mathrm{NF} &= 10 \log_{10} F \\
&= \mathrm{SNR}_i(\mathrm{dB}) - \mathrm{SNR}_o(\mathrm{dB}).
\end{aligned}
\tag{11.185}
$$

   Development of the noise factor of a cascade begins by considering the noise contributions of the first system, and then the next cascaded system, and so on. The majority of RF and microwave systems are organized as

Figure 11-54   Cascaded noisy two-ports.

cascades of two-port networks with an input port and an output port, as shown in Figure 11-54.

If the excess noise contribution of an amplifier is ignored, the output noise power will be

$$N_o = GkT_0B \ . \tag{11.186}$$

With excess noise, N_e, from the amplifier included, the output noise power is

$$\begin{aligned} N_o &= GkT_0B + N_e \\ &= GkT_0B(1 + N_e/(GkT_0B)) \\ &= FGkT_0B. \end{aligned} \tag{11.187}$$

Rearranging this equation the excess noise power can be written as

$$N_e = (F-1)GkT_0B \ . \tag{11.188}$$

EXAMPLE 11. 3   Noise Figure of an Attenuator

What is the noise figure of a 20 dB attenuator in a 50 Ω system?

Solution:

Denoting the attenuator as being in a 50 Ω system indicates that an appropriate circuit model to use in the analysis consists of the attenuator driven by a generator, with a 50 Ω source impedance, and the attenuator drives a 50 Ω load. Also, the input impedances of the terminated attenuator is 50 Ω, as is the impedance looking into the output of the attenuator when it is connected to the source. The key point is that the noise coming from the source is the noise thermally generated in the 50 Ω source impedance, and this noise is equal to the noise that is delivered to the load, as the impedance presented to the load is also 50 Ω. So the input noise, N_i, is equal to the output noise:

$$N_o = N_i. \tag{11.189}$$

The input signal is attenuated by 20 dB (= 100). So

$$S_o = S_i/100, \tag{11.190}$$

and thus the noise factor is

$$F = \frac{SNR_i}{SNR_o} = \frac{S_i}{N_i}\frac{N_o}{S_o} = \frac{S_i}{N_i}\frac{N_i}{S_i/100} = 100 \qquad (11.191)$$

and the noise figure is

$$NF = 20 \text{ dB}. \qquad (11.192)$$

So the noise figure of an attenuator (or filter) is just the loss of the component. This is not true for amplifiers of course, as there are other sources of noise, and the output impedance of a transistor is not a thermal resistance.

11.11.4 Noise in a Cascaded System

Section 11.11.3 developed the noise factor and noise figure measures for a two-port. This result can be generalized for a system. Considering the second stage of the cascade, the excess noise at the output of the second stage, due solely to the noise generated internally in the second stage, is

$$N_{e,2} = (F_2 - 1)kT_0BG_2. \qquad (11.193)$$

Then the total noise power at the output of a two-stage cascade is

$$N_{o,2} = (F_2 - 1)kT_0BG_2 + N_{o,1}G_2 \qquad (11.194)$$
$$= (F_2 - 1)kT_0BG_2 + F_1kT_0BG_1G_2. \qquad (11.195)$$

The second term above is the noise output from the first stage amplified by the second gain.

Generalizing the above result yields the total noise power at the output of the mth stage:

$$N_{o,m} = \sum_{n=2}^{m}\left[(F_n - 1)\,kT_0B\prod_{i=2}^{n}G_i\right] + F_1kT_0B\prod_{n=1}^{m}G_n. \qquad (11.196)$$

Thus an m-stage cascade has a total cascaded system noise factor of $F^T = N_{o,m}/(G^TN_{i,1})$, with G^T being the total cascaded available gain and $N_{i,1}$ being the noise power input to the first stage. In terms of the parameters of individual stages, the total system noise factor is

$$F^T = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1G_2} + \frac{F_4 - 1}{G_1G_2G_3} + \cdots ; \qquad (11.197)$$

that is,

$$F^T = F_1 + \sum_{n=2}^{m}\frac{F_n - 1}{\prod_{i=n}^{m}G_{i-1}} . \qquad (11.198)$$

This equation is known as **Friis' formula**.

Figure 11-55 Differential amplifier followed by a differential filter.

EXAMPLE 11. 4   Noise Figure of Cascaded Stages

Consider the cascade of a differential amplifier and a filter shown in Figure 11-55.

(a) What is the midband gain of the filter in decibels? Note that IL is insertion loss.
(b) What is the midband noise figure of the filter?
(c) The amplifier has a gain $G_1 = 20$ dB and a noise figure of 2 dB. What is the overall gain of the cascade system in the middle of the band? Express your answer in decibels.
(d) What is the noise factor of the cascade system?
(e) What is the noise figure of the cascade system?

Solution:

(a) $G_2 = 1/\text{IL}$ thus $G_2 = -3$ dB.
(b) For a passive element, $\text{NF}_2 = \text{IL} = 3$ dB.
(c) $G_1 = 20$ dB and $G_2 = -3$ dB and so $G_{\text{TOTAL}} = G_1|_{\text{dB}} + G_2|_{\text{dB}} = 17$ dB.
(d)

$$F_1 = 10^{\text{NF}_1/10} = 10^{2/10} = 1.585$$
$$F_2 = 10^{\text{NF}_2/10} = 10^{3/10} = 1.995$$
$$G_1 = 10^{20/10} = 100$$
$$G_2 = 10^{-3/10} = 0.5 \ .$$

Using Friis' formula

$$F_{\text{TOTAL}} = F_1 + \frac{F_2 - 1}{G_1} = 1.585 + \frac{1.995 - 1}{100} = 1.594 \ . \qquad (11.199)$$

(e)

$$\text{NF}_{\text{TOTAL}} = 10\log_{10}(F_{\text{TOTAL}}) = 10\log_{10}(1.594) = 2.03 \text{ dB}. \qquad (11.200)$$

Figure 11-56 Amplifier model for noise factor calculation. I_S is the source generator and Y_S is the Thevenin equivalent admittance of the source.

11.11.5 Noise Figure of Two-Port Amplifiers

The parameters that define the noise figure were set forth by Haus et al. [199] in 1960 and are used by transistor manufacturers in their datasheets. The amplifier model used in noise factor computation is shown in Figure 11-56. The noise factor of a two-port amplifier is

$$F = F_{\min} + \frac{r_n}{g_s} \left| y_s - y_{\text{opt}} \right|^2, \tag{11.201}$$

where $r_n = (R_n/Z_0)$ is the equivalent noise resistance of the two port. $F_{\min}$ is the minimum noise factor obtained by adjusting tuners at the input of the amplifier. The normalized admittance presented by the tuners at $F_{\min}$ is y_{opt}. With $y_s(= Y_S/Z_0)$ being the actual normalized admittance, Equation (11.201) enables the noise factor to be calculated for an actual design. The parameters describe the effect of internal amplifier noise source and how they are correlated. The parameters can be defined for an amplifier or a transistor in a two-port configuration. More commonly today, noise parameters are reported in terms of reflection coefficients. The source reflection coefficient, Γ_s, comes from

$$y_s = \frac{1 - \Gamma_s}{1 + \Gamma_s} \tag{11.202}$$

and the optimum source reflection coefficient, Γ_{opt}, comes from

$$y_{\text{opt}} = \frac{1 - \Gamma_{\text{opt}}}{1 + \Gamma_{\text{opt}}}. \tag{11.203}$$

Substituting these in Equation 11.201 results in

$$F = F_{\min} + \frac{4r_n \left| \Gamma_s - \Gamma_{\text{opt}} \right|^2}{\left(1 - |\Gamma_s|^2 \right) \left| 1 + \Gamma_{\text{opt}} \right|^2} \tag{11.204}$$

Together $F_{\min}$, r_n, and Γ_{opt} are called the noise parameters of a device, and as indicated, must be measured. The noise parameters of the pHEMT previously considered are given in Table 11-11.

Table 11-11 Noise parameters of the enhancement-mode pHEMT transistor presented in Figure 11-8. $F_{\min}$ is the minimum noise factor, $NF_{\min}$ is the minimum noise figure, and Γ_n is source reflection coefficient yielding $F_{\min}$.

Frequency (GHz)	$F_{\min}$	$NF_{\min}$ (dB)	Γ_n degrees	$\angle\Gamma_n$ (Ω)	$R_n/50$
0.90	1.07	0.29	0.747	15.70	0.165
1.80	1.09	0.38	0.623	24.95	0.176
2.40	1.11	0.44	0.795	37.45	0.158
2.60	1.11	0.47	0.640	47.15	0.159
2.80	1.12	0.49	0.670	47.90	0.160
3.20	1.13	0.53	0.617	51.20	0.156
4.00	1.15	0.61	0.542	68.70	0.141
5.00	1.18	0.72	0.465	85.00	0.120
5.50	1.19	0.77	0.431	91.10	0.114
6.00	1.21	0.83	0.366	101.15	0.107
7.00	1.24	0.93	0.262	122.10	0.096
8.00	1.27	1.04	0.188	153.60	0.100
9.00	1.30	1.14	0.135	-165.60	0.121
10.00	1.33	1.25	0.162	-126.80	0.138
11.00	1.37	1.35	0.183	-85.95	0.187
12.00	1.40	1.46	0.270	-68.40	0.239
13.00	1.43	1.57	0.343	-50.25	0.355
14.00	1.47	1.67	0.431	-43.95	0.461
15.00	1.51	1.78	0.573	-25.80	0.604

The design of a linear amplifier for best noise performance requires that $\Gamma_s = \Gamma_{\mathrm{opt}}$. This changes the input matching network design presented in Section 11.7.4. In general the design for best noise performance does not yield the best gain. The reduction in gain is usually small however. In Section 11.7.4 the input matching network was conjugately matched, yielding maximum amplifier gain (the output matching network had already been designed). For best noise performance the input matching network is not conjugately matched and instead the input reflection coefficient looking into the matching network from the active device is Γ_{opt}. As a result, maximum gain is not obtained. The interpretation of why this is necessary is that a particular mismatch minimizes the noise contributions of partially correlated internal active device noise sources. Techniques have been developed involving noise and gain circles that enable the trade-off in design to be made [183]. In modern RF and microwave design, however, if noise performance is of concern, two or more amplifier stages are used with the first stage designed for optimum noise performance and subsequent stages designed for maximum gain.

Figure 11-57 Nonlinear effects introduced by RF hardware: (a) amplitude (AM-AM) distortion; and (b) phase (AM-PM) distortion.

11.12 Amplifier Nonlinear Distortion

Distortion imposes a fundamental limit to the practical efficiency that can be realized by an amplifier. Distortion originates when the output signal from an amplifier approaches the extremes of the loadline so that the output is not an exact amplified replication of the input signal. For a one-tone signal, the amplitude gain of the signal rolls off as the input power increases, as shown in Figure 11-57(a). This figure plots the amplitude of the output sinewave against the amplitude of the input sinewave, putting both amplitudes in terms of power. The plot is called the AM-to-AM (**AM-AM**) characteristic of the amplifier. The ideal amplifier would follow the linear relationship between the output and input powers. The AM-AM characteristic is linear at low input powers, but eventually the gain compresses and the output power drops in proportion to the input power. At large powers, the parasitic capacitances of the transistors in the amplifier vary the signal phase, and hence phase distortion results. Figure 11-57(b) shows what is called the AM-to-PM (**AM-PM**) characteristic. The AM-AM distortion is generally more significant, and considerable departure from the linear response occurs before the output phase varies appreciably. In Figure 11-57(a), the 1 dB gain compression point is at the point where the difference between the extrapolated linear response exceeds the actual gain by 1 dB. P_{1dB} is the output power at the 1 dB gain compression point and is the single most important metric of distortion, and amplifier designs use P_{1dB} as a point of reference. Phase distortion generally occurs at higher powers (see Figure 11-57(b)). While Figure 11-57 presents the distortion characteristics for a single sinewave, it has proved to be a good indicator of performance with modulated signals. A two-tone signal consisting of two sinusoidal signals is a better representation of modulated signals. A signal linearly combining (adding) two sinusoidal signals of equal amplitude is shown in Figure 11-

(a)   (b)

Figure 11-58 A two-tone signal: (a) a two-tone input waveform; and (b) distorted output showing compression (dashed waveform is undistorted).

Figure 11-59 Spectrum at the output of a nonlinear amplifier with a two-tone input signal.

58(a). When the input signal to a Class A amplifier is large the extremes of the signal on the loadline (see Figure 11-11) are compressed when the signal reaches its extremities. This results in the saturated output waveform shown in Figure 11-58(b). In the frequency domain, this distortion shows up as additional tones so that this distortion is said to produce **intermodulation products** (**IMPs**) as shown in Figure 11-59. Here f_1 and f_2 components have the frequencies of the two tones comprising the two-tone signal. The extra ones in the output, f_3 and f_4, are the intermodulation tones. In Figure 11-59, the tone at $f_3 = 2f_1 - f_2$ is known as the lower **IM3** (or lower intermod) and $f_4 = 2f_2 - f_1$ is known as the upper **intermod**. As well as the amplitude distortion resulting in additional tones, there is phase distortion as captured by the AM-PM characteristic. Amplifiers, however, introduce less phase distortion than amplitude distortion, which is fortuitous since most communication systems encode information in the phase or frequency rather than in the amplitude.

Amplifier design consists of both design for low-power linear operation, requiring maximum power transfer at the input and output of the amplifier, and a trade-off of acceptable distortion and efficiency. In practice, a certain level of distortion must be tolerated, and what is acceptable is embedded in the specifications for the various wireless systems.

For low distortion, the peaks of the RF signal must be amplified linearly, however, the DC power consumed depends on the amplifier class. With Class A amplifiers, the DC power must be sufficient to provide undistorted amplification of the largest RF signal so that the DC power is proportional to the peak AC power. The situation is similar for Class AB amplifiers, with the difference being that the intent is to live with some distortion of the peak signal so that the relationship between peak power and DC power still exists, but the direct proportionality no longer holds. For Class C and higher class amplifiers, the DC power is mostly proportional to the average RF power. So for Class A and Class AB amplifiers, the average operating point must be "backed off" to allow for manageable distortion of the peaks of a signal, with the level of back-off required being proportional to the PAR. For Class C and higher classes, the back-off required comes from experience and experimentation. The characteristics of the signal also determine how much distortion can be tolerated.

The PAR of the signal is an indication of the type and amount of distortion that can be tolerated. The PAR of the two-tone signal is 6 dB, and digitally-modulated signals can have PARs ranging from 0 dB to 10 dB or more. A signal with a higher PAR results in lower efficiency, as more back-off is required. Putting this another way, the DC bias must be set so that there is minimum distortion when the signal is at its peak, but the average RF power produced can be much less than the peak RF power. (It is approximately an amount PAR below.) Thus for a high-PAR signal, generally a higher DC power is required to produce the same RF power. This is especially true for Class A amplifiers. Efficiency can be increased by using switching amplifiers. The PAR of a modulated signal is an indication of how much information is being transferred in the amplitude of the signal. For example, a GMSK signal has a PAR of 0 dB, and there is no information in the amplitude of a signal so that a highly efficient Class C amplifier can be used, as any amplitude introduced does not matter. Signals that have higher PARs contain increasing amounts of information in the amplitude of the signals.

AM-AM and AM-PM distortion and two-tone distortion also provide indications as to the distortion that occurs with both analog and digitally-modulated signals. Distortion with digitally modulated signals consists of in-band and out-of-band distortion. Out-of-band distortion is represented in Figure 11-60, where the spectra at the input and output of a nonlinear system are shown. The process that results in increased levels in the adjacent sidebands is called spectral regrowth. This distortion is approximately captured by the intermodulation distortion with a two-tone signal. The

Figure 11-60 Input and output spectra of a digitally modulated signal: (a) a digitally modulated signal at the input of a nonlinear amplifier; and (b) at the output of the amplifier.

generation of signal in the adjacent channel affects the function of other radios and the level of signal is contained in system specifications. Distortion generated in-band affects the ability to interpret the constellation

Figure 11-61 Output power versus input power of a stage or system plotted as output power in decibels versus input power in decibels. The IP3 response is a result of two-tone intermodulation, and the input power is the combined power of the two signals that have equal amplitude. Extrapolations of the 1:1 linear response and the 3:1 third-order intermodulation response intersect at the IP3 point.

of the signal and hence causes difficulties in demodulating a signal.

11.13 Dynamic Range

While modern communication and radar systems use digitally-modulated signals, two-tone signals are used to approximately characterize nonlinearity and in manual calculations. At low powers before compression becomes a factor, the fundamental response has a 1:1 slope with respect to the input, as shown in Figure 11-61. The IP3 response varies as the cube of the level of input tones when both tones vary by the same amount, as is common in a two-tone test. Thus IP3 has a 3:1 logarithmic slope with respect to the input. Since the relations are linear in a log-log sense, it is possible to describe the nonlinear performance of an amplifier by a quantity called the **Dynamic Range** (**DR**) or by the similar **Spurious Free Dynamic Range** (**SFDR**). SFDR describes the difference between the level at which a signal is distorted and the level of noise (i.e., the noise floor). DR is similar to SFDR except that the level of the **Minimum Discernible Signal** (**MDS**) is used instead of the noise floor. MDS is higher than the level of the noise floor by the minimum acceptable SNR (SNR_{MIN}). SNR_{MIN} is dependent on the type of modula-

tion, on hardware inadequacies (captured by the implementation margin),, on processing gain, and on the error correction coding used in a particular communications protocol.

In the following, an expression for SFDR is developed in terms of input-referenced quantities and this form of the SFDR is called the input referred SFDR, (SFDR_i). A similarly referenced dynamic range (DR_i) is also developed.

Figure 11-61 illustrates the input-output relationship of a one-tone signal to the IP3 response of a subsystem and also graphically defines the dynamic ranges. The point of intersection of the extrapolated linear output (of power P_o) and third-order (IP3) output (of power P_{IP_3}) is called the third-order intercept point (**IP3 intercept**). The point is identified by the output-referred intercept power (OIP3) or by the input referred IP3 intercept power (IIP3), and these are key parameters in describing the linearity of nonlinear subsystems.

In the linear gain region, P_o versus P_i has a slope of 1:1, so that

$$P_{\text{dBm},i} = P_{\text{dBm},o} - G_{\text{dB}}, \tag{11.205}$$

where G_{dB} is the power gain in decibels. P_o is used here as the output power, with $P_{\text{dBm},o}$ indicating the output power in dBm. P_i and $P_{\text{dBm},i}$ are similarly defined. In terms of input quantities

$$\text{IIP3}_{\text{dBm}} = \text{OIP3}_{\text{dBm}} - G_{\text{dB}}, \tag{11.206}$$

where again the dBm subscript indicates that the quantity is expressed in decibels referred to 1 mW. The nonlinearity of RF active components results in harmonics and intermodulation components. With the narrowband amplifiers of communication and radar systems, output filters conveniently filter out harmonics. However, intermodulation distortion cannot be filtered out, as these components are within the main passband. The intermodulation components are therefore spurious tones. Generally just one of these defines the maximum spurious tone and nearly always it is the third-order intermodulation tone with a two-tone input. Consideration of the maximum spurious tone and the noise floor defines the SFDR.

Examining Figure 11-61 leads to the following inequality describing the linear gain of the amplifier:

$$\frac{\text{OIP3}_{\text{dBm}} - P_{\text{dBm},o}}{\text{IIP3}_{\text{dBm}} - P_{\text{dBm},i}} = \frac{\text{OIP3}_{\text{dBm}} - P_{\text{dBm},o}}{(\text{OIP3}_{\text{dBm}} - G_{\text{dB}}) - P_{\text{dBm},i}}$$
$$= 1. \tag{11.207}$$

Here, P_o, P_i, OIP3, and IIP3 are the output power, the input power, the input-referred IP3 intercept, and the output-referred IP3 intercept. The third-order response is characterized by first introducing an equivalent

input power, $P_{\mathrm{dBm,i3}}$ (P_{i3} expressed in dBm), defined as the average power of the two-tone signal that generates an IP3 of power $P_{\mathrm{dBm,o3}}$. Noting that $P_{\mathrm{dBm,o3}}$ varies with a 3:1 logarithmic slope with respect to $P_{\mathrm{dBm,i3}}$, then

$$\frac{\mathrm{OIP3}_{\mathrm{dBm}} - P_{\mathrm{dBm,o3}}}{(\mathrm{OIP3}_{\mathrm{dBm}} - G_{\mathrm{dB}}) - P_{\mathrm{dBm,i3}}} = 3 \tag{11.208}$$

or

$$P_{\mathrm{dBm,i3}} = \frac{1}{3}\left(2 \times \mathrm{OIP3}_{\mathrm{dBm}} + P_{\mathrm{dBm,o3}} - 3G_{\mathrm{dB}}\right). \tag{11.209}$$

The SFDR can now be defined when the third-order intermodulation product of two-tone excitation is the dominant spurious tone. The SFDR is defined as the difference between P_{i3} and P_i when they produce IP3 and linear output respectively that are both equal to the output noise power, N_o (see Figure 11-61); that is, when $P_o = P_{o3} = N_o$. Replacing $P_{\mathrm{dBm,o}}$ in Equation 11.209 with N_o gives

$$P_{\mathrm{dBm,i3}} = \frac{1}{3}\left(2 \times \mathrm{OIP3}_{\mathrm{dBm}} + N_{\mathrm{dBm,o}} - 3G_{\mathrm{dB}}\right) \tag{11.210}$$

and

$$P_{\mathrm{dBm},i} = N_{\mathrm{dBm,o}} - G_{\mathrm{dB}}. \tag{11.211}$$

Note that the difference between the linear output and the third-order intermodulation reduces as the input power increases above P_{i3}. Thus the output-referred SFDR is

$$
\begin{aligned}
\mathrm{SFDR}_{\mathrm{dB},o} &= P_{\mathrm{dBm,i3}} - P_{\mathrm{dBm},i} &&\qquad (11.212)\\
&= \frac{2}{3}\mathrm{OIP3}_{\mathrm{dBm}} + \tfrac{1}{3}N_{\mathrm{dBm,o}} - G_{\mathrm{dB}} - N_{\mathrm{dBm,o}} + G_{\mathrm{dB}}\\
&= \frac{2}{3}\left(\mathrm{OIP3}_{\mathrm{dBm}} - N_{\mathrm{dBm,o}}\right).
\end{aligned}
$$

A similar development can be used to define the input-referred SFDR:

$$\mathrm{SFDR}_{\mathrm{dB},i} = \frac{2}{3}\left(\mathrm{IIP3}_{\mathrm{dBm}} - N_{\mathrm{dBm},i}\right). \tag{11.213}$$

Note that N_i is the input-referred noise and includes noise applied to the subsystem as well as the noise produced internally in the subsystem and referred to the input. The SFDR provides a combined measure of distortion and noise. However, for usable dynamic range the minimum acceptable SNR must be considered. The minimum SNR ($\mathrm{SNR}_{\mathrm{MIN}}$) required is determined by the communication or radar modulation format, error coding, and acceptable BER. So in defining DR the input power of the desired signal must increase sufficiently to produce an SNR of at least $\mathrm{SNR}_{\mathrm{MIN}}$. Since the desired spurious level is still at the noise floor, this

Figure 11-62   Class A inductively-biased MOSFET amplifier.

implies a direct subtraction in decibels of the desired SNR. Therefore the input-referred third-order dynamic range, preferred for receivers, is

$$
\mathrm{DR}_i = \frac{2}{3} \left(\mathrm{IIP3} - N_{\mathrm{dBm},i} \right) - \mathrm{SNR}_{\mathrm{MIN}}
\tag{11.214}
$$

and the output-referred dynamic range, preferred for transmitters, is

$$
\mathrm{DR}_o = \frac{2}{3} \left(\mathrm{OIP3} - N_{\mathrm{dBm},o} \right) - \mathrm{SNR}_{\mathrm{MIN}}.
\tag{11.215}
$$

EXAMPLE 11. 5   Intercept Point

The amplifier shown in Figure 11-62 is modeled, after expansion around the operating point, by a nonlinear transconductance with $i_{DS} = a_1 v_{GS} + a_3 v_{GS}^3$ with $a_1 = 0.01$ A/V and $a_3 = -0.1$ A/V^3. L_1 is an RF choke and can be assumed to present an open circuit at the operating frequency. C_1 is a biasing capacitor and can be treated as a short circuit at the operating frequency. R_2 is the load resistance.

  (a) What is the small signal voltage gain of the amplifier?
  (b) If the input signal is a two-tone signal consisting of two equal amplitude sinewaves at 900 MHz and 901 MHz, what are the frequencies in the spectrum of v_o, the output voltage across R_2?
  (c) If a single sinusoid of amplitude 100 mV is applied to the gate of the MOSFET, what is the level of the fundamental tone at the output of the amplifier? What is the level of the third harmonic?
  (d) What is the input-referred third-order intercept point, IIP3, of the amplifier?
  (e) What is the output-referred third-order intercept point, OIP3, of the amplifier?

Solution:

(a) When the input signal v_i is small, $i_{DS} = a_1 v_i$ so the small-signal output voltage is

$$v_o = -i_{DS} R_2 = -a_1 R_2 v_i,$$

and the small signal voltage gain is

$$A = \frac{v_o}{v_i} = -a_1 R_2 = -0.01 \times 500 = -5.$$

(b) Eventually the symbolic amplitudes of the mixing terms will be required, so a trigonometric expansion is undertaken now. To minimize complexity consider two cosinsoids, $A\cos(x)$ and $B\cos(y)$. The input to the amplifier is

$$v_{GS} = A\cos(x) + B\cos(y).$$

The output of the amplifier is

$$v_o = -i_{DS} R_2 = -R_2 \left(a_1 v_{GS} + a_3 v_{GS}^3 \right)$$
$$= -R_2 \left\{ a_1 \left[A\cos(x) + B\cos(y) \right] + a_3 \left[A\cos(x) + B\cos(y) \right]^3 \right\}$$
$$= -R_2 \left\{ a_1 \left[A\cos(x) + a_1 B\cos(y) \right] + a_3 \left[A\cos(x) + B\cos(y) \right] \left[A\cos(x) + B\cos(y) \right]^2 \right\}$$
$$= -R_2 \left\{ a_1 \left[A\cos(x) + a_1 B\cos(y) \right] \right.$$
$$\left. + a_3 \left[A\cos(x) + B\cos(y) \right] \left[A^2 \cos^2(x) + 2AB\cos(x)\cos(y) + B^2 \cos^2(y) \right] \right\}$$
$$= -R_2 \left\{ a_1 A\cos(x) + a_1 B\cos(y) \right.$$
$$+ a_3 \left[A\cos(x) + B\cos(y) \right]$$
$$\left. \left[\frac{A^2}{2} + \frac{B^2}{2} + \frac{1}{2} A^2 \cos(2x) + \frac{1}{2} B^2 \cos(2y) + AB\cos(x-y) + AB\cos(x+y) \right] \right\}$$

$$\text{(11.216)}$$

$$= -R_2 \left\{ a_1 \left[A\cos(x) + B\cos(y) \right] \right.$$
$$+ a_3 \left[\left(\frac{A^3}{2} + \frac{AB^2}{2} \right) \cos(x) + \frac{1}{2} A^3 \cos(x)\cos(2x) + \frac{1}{2} AB^2 \cos(x)\cos(2y) \right.$$
$$+ A^2 B\cos(x)\cos(x-y) + AB^2 \cos(x)\cos(x+y) + \left(\frac{A^2 B}{2} + \frac{B^3}{2} \right) \cos(y)$$
$$+ \frac{1}{2} B^3 \cos(y)\cos(2y) + \frac{1}{2} A^2 B\cos(y)\cos(2x) + AB^2 \cos(y)\cos(x-y)$$
$$\left. \left. + AB^2 \cos(y)\cos(x+y) \right] \right\}$$

$$= -R_2 \left(a_1 \left[A\cos(x) + a_1 B\cos(y) \right] \right.$$

$$+ a_3 \left\{ \left(\frac{A^3}{2} + \frac{AB^2}{2} \right) \cos(x) + \frac{1}{4} A^3 \left[\cos(x) + \cos(3x) \right] \right.$$

$$+ \frac{1}{4} AB^2 \left[\cos(2y - x) + \cos(x + 2y) \right] + \frac{1}{2} A^2 B \left[\cos(y) + \cos(2x - y) \right]$$

$$+ \frac{1}{2} A^2 B \left[\cos(y) + \cos(2x + y) \right] + \left(\frac{A^2 B}{2} + \frac{B^3}{2} \right) \cos(y)$$

$$+ \frac{1}{4} B^3 \left[\cos(y) + \cos(3y) \right] + \frac{1}{4} A^2 B \left[\cos(2x - y) + \cos(2x + y) \right]$$

$$\left. \left. + \frac{1}{2} AB^2 \left[\cos(x) + \cos(2y - x) \right] + \frac{1}{2} AB^2 \left[\cos(x) + \cos(2y + x) \right] \right\} \right) \quad (11.217)$$

$$= -R_2 \left\{ a_1 A\cos(x) + a_1 B\cos(y) \quad + a_3 \left[\left(\frac{3A^3}{4} + \frac{3AB^2}{2} \right) \cos(x) \right. \right.$$

$$+ \left(\frac{3B^3}{4} + \frac{3AB^2}{2} \right) \cos(y) + \frac{B^3}{4} \cos(3y) + \frac{3}{4} A^2 B\cos(2x - y)$$

$$\left. \left. + \frac{3}{4} A^2 B\cos(2x + y) + A^3 \cos(3x) + \frac{3}{4} AB^2 \cos(2y - x) + \frac{3}{4} AB^2 \cos(2y + x) \right] \right\}$$

$$(11.218)$$

So with x representing 900 MHz and y representing 901 MHz, the frequencies in the putput spectrum are 899 MHz, 900 MHz, 901 MHz, 902 MHz, 2700 MHz, 2701 MHz, 2702 MHz, and 270 3 MHz.

(c) From Equation (11.218) and considering only one tone (i.e., B = 0), the output signal is

$$v_o = -R_2 \left[a_1 A\cos(x) + \frac{3a_3 A^3}{4} \cos(x) + \frac{a_3 A^3}{4} \cos(3x) \right] \quad (11.219)$$

So the coefficient of the fundamental at the output is

$$v_o = -R_2 \left[a_1 A\cos(x) + \frac{3a_3 A^3}{2} \cos(x)) \right] \quad (11.220)$$

Here $A = 100$ mV so

$$\text{Fundamental output level} = -(500\ \Omega) \times \left[(0.01\ \text{A/V}) \times (0.5\ \text{V}) \right.$$
$$+ (-0.1\ \text{A/V}^3) \times (0.1\ \text{V})^3 \right]$$
$$= 0.05 - 0.0125\ \text{V} = 0.0375\ \text{V} = 37.5\ \text{mV} \quad (11.221)$$

$$\text{Third harmonic output level} = \frac{a_3 R_2}{4} v_i^3 = (-500) \times (-0.1) \times (0.1)^3/4 = 0.0125 \text{ V}$$

$$(11.222)$$

(d) To answer this determine the level of the fundamental and the third-order intermodulation product for small v_{GS} and with the two input tones having the same amplitude $A = B = v_{GS}$. For a resistance nonlinearity, such as the transconductance here, the level of the lower and upper third-order intermods are the same. So, after examining Equation (11.218), consider $\cos(x)$ and $\cos(2x - y)$. The fundamental (at $\cos(x)$) is

$$v_o(900 \text{ MHz}) = -R_2 a_1 v_i,$$

and the level of the lower third-order intermod is

$$v_o(899 \text{ MHz}) = -R_2 a_3 \frac{3}{4} (v_i)^3$$

IIP3 is the value of v_i when $v_o(900 \text{ MHz}) = v_o(899 \text{ MHz})$, that is

$$-R_2 a_1 v_i = -R_2 a_3 \frac{3}{4} (v_i)^3 .$$

That is when

$$v_i = \sqrt{\left| \frac{4a_1}{3a_3} \right|} = \sqrt{\frac{4 \cdot 0.01}{3 \cdot 0.1}} = 0.3651 \text{ V} = 365.1 \text{ mV}$$

Thus

$$A_{\text{IIP3}} = 365.1 \text{ mV}.$$

Normally IIP3 is expressed in terms of power. Considering Figure 11-62, $E = 2v_i$, and so the available input power is

$$P_{\text{av}} = \frac{1}{2} \frac{v_i^2}{R_1} = \frac{v_i^2}{2R_1}.$$

Thus

$$\text{IIP3} = \frac{1}{2R_1} A_{\text{IIP3}}^2 = \frac{0.3651^2}{2 \cdot 500} = 0.0001333 \text{ W} = 0.1333 \text{ mW} = -0.875 \text{ dBm}$$

(e) The voltage output-referred intercept point is

$$A_{\text{OIP3}} = |A| A_{\text{IIP3}} = 5 \cdot 0.3651 \text{ V} = 1.8255 \text{ V}$$

and

$$\text{OIP3} = (\text{Power gain}) \cdot \text{IIP3} = \frac{R_1}{R_2} A^2 \cdot 0.1333 \text{ mW} = 3.333 \text{ mW} = 5.23 \text{ dBm}$$

11.14 Distortion and Digitally Modulated Signals

11.14.1 PAR and Probability Density Function

Amplitude variations occur with digital modulation schemes. For example, a QPSK signal consists of two digital data streams, equal in amplitude, modulated in quadrature onto a carrier signal. The resulting signal would have a constant envelope; however, the occupied bandwidth is quite large, as the spectrum of a pulse train is $\sin(x)/x$, the sinc function. The first **sidelobe** of the sinc spectrum is only 13 dB down from the carrier level and is in the middle of the adjacent channel. To reduce the spectrum, a low-pass filter is typically applied to each digital data stream to minimize the out-of-band spectrum of the modulated signal. This comes with a drawback: the filters cause a finite memory effect resulting in amplitude variations as the ringing energy from a previous data pulse adds to the current filtered data pulse.

Amplitude variations of the modulated signal are characterized by measured waveform statistics such as the PAR. A signal with a high PAR requires that the RF system have high linearity to handle the average power requirements and the peak amplitude excursions without generating excessive out-of-band distortion. However, it is possible for a signal with a higher PAR to exhibit less nonlinear distortion than a signal with lower PAR [200]. The reason for this apparent inconsistency is because the signal peak is a singular point measurement with typically a low probability of occurrence. Thus PAR by itself is an incomplete statistic for determining the linearity requirements for a transmitter to carry a signal.

The **Amplitude Probability Density Function** (APDF) is a more complete statistical description of the amplitude variations of a modulated signal. The APDF defines the maximum and minimum variation along with the relative probability of occurrence of amplitudes within the variation. The APDF is typically estimated from a histogram of amplitudes, with a uniform bin size, by

$$f(A) = \frac{N}{\Delta A \times N_c}, \tag{11.223}$$

where N is the number of counts per bin, ΔA is the bin amplitude width, and N_c is the total number of samples. The shape of the amplitude density between the mean and peak amplitude influences the sensitivity of a particular signal to spectral regrowth due to nonlinear gain compression or expansion. For example, Figure 11-63 shows the APDF for a **CDMA** mobile transmitter using OQPSK modulation, the same signal using QPSK modulation, a real **Gaussian signal**, and a complex Gaussian QPSK signal (with I and Q each having Gaussian distribution), where the average power of each signal is set to 0 dBm. Gaussian signals are of particular interest because their simple statistics lend them, and their interaction with nonlinearities, to quasi-analytic treatment. The PAR for each signal is shown

Figure 11-63   Amplitude PDF for CDMA and Gaussian modulation signals. After [201].

Table 11-12   Peak-to-average ratios in decibels for OQPSK and QPSK used in CDMA compared to the PAR's of Gaussian signals.

Signal Modulation	PAR (dB)
OQPSK CDMA	5.4
QPSK CDMA	6.6
Real Gaussian	13.5
Complex Gaussian	11.8

in Table 11-12. The shape of the amplitude density after the mean differs for both signals where it can be seen that it is difficult to determine, a priori, which signal will be more sensitive to nonlinear gain compression. For example, even though QPSK has a higher PAR than OQPSK, the probability that the OQPSK signal is near the peak is higher than for the QPSK signal. It is not surprising then that the measured spectral regrowth of a OQPSK signal is higher than for a QPSK signal. So PAR is only a rough guide to the distortion that is produced.

11.14.2 Design Guidelines

Generally an amplifier is not operated in saturation or near the **Third Order Intercept** (**TOI**) point. An exception is with constant or near-constant envelope modulation schemes such as FM, GMSK, SOQPSK, FOQPSK, and SBPSK. With little variation in the amplitude of the RF signal, saturating amplifiers can be used. With nonconstant envelope modulation schemes (i.e., the signal PAR is more than 0 dB), the variation in the amplitude of the modulated RF carrier will result in signal distortion and spectral regrowth (i.e., power will be dumped into the neighboring communication channel). One general rule of design is to ensure that the peak of the RF pseudo-carrier is at or below the 1 dB gain compression point. The amplifier is said to be backed off by an amount PAR below the 1 dB gain compression point. Generally, however, third-order intermodulation is a much greater concern, as this relates to the amount of power dumped into neighboring channels. Modern communication schemes can require that the adjacent channel spectral regrowth be as much as 80 dB below the power in the main channel. As a result, the signal must be backed off considerably from the TOI point. The TOI is obtained from a two-tone characterization and thus is a weakly accurate indicator of distortion with a digitally modulated signal. However it is a simple measurement to make and understand, so it is widely used. Experience provides a rule-of-thumb for the back-off needed to obtain the necessary spectral regrowth required to meet the specifications for a particular modulation format and transistor technology.

While TOI and the 1 dB compression point only refer to amplitude distortion there will also be phase distortion. Phase distortion will affect the performance of constant envelope modulation schemes so the 1 dB compression point and TOI also provide some indication of possible distortion with constant envelope schemes. EVM and BER are much better metrics for all digital modulation schemes but these are much more difficult to use in guiding initial design of RF hardware.

11.15 Amplifiers and RFICs

Silicon RFICs exploit the high-density integration possible with silicon MOSFET transistors. These transistors can be fabricated with high levels of repeatability so that the transistors in differential designs can be closely matched. As well, the inherent compatibility with digital circuits enables digital control of RF circuits. As far as amplifiers are concerned, there are a few commonly used basic circuits that use complementary MOSFETs (nMOS and pMOS) or CMOS transistors. MOSFET differential amplifiers have been presented throughout this chapter. Other common CMOS circuits are shown in Figure 11-64. The transistors in all the circuits described here operate in the saturation region.

A **cascode amplifier** is shown in Figure 11-64(a). There are two FETs, with

Figure 11-64 MOS analog circuits: (a) cascode amplifier; (b) variable gain cascode amplifier; (c) current mirror; (d) differential pair; (e) multi tanh triplet implementation of a differential pair providing enhanced linearity.

the top FET acting as the drain load of the bottom FET. The gate of the top FET is held at ground so the voltage at the source of the top FET (and the drain of the bottom FET) is held at a nearly constant voltage. Thus the top FET presents a low-resistance load to the bottom FET. The voltage gain of the FET is low, and this reduces the **Miller effect capacitance**, which is the effective input capacitance (being the gate-drain capacitance multiplied by the transistor voltage gain). The cascode topology increases the bandwidth of the circuit. Current gain, and hence power gain, is still realized by the bottom transistor. The voltage gain of the circuit depends on the resistance presented to the drain of the top transistor.

A variable gain **cascode amplifier** is shown in Figure 11-64(b). This is similar to the cascode amplifier of Figure 11-64(a), but now the voltage at the gate of the top transistor is selected so that a variable resistance is presented to the bottom transistor, thus the voltage gain of the circuit can be varied.

FET circuits are nearly always current biased, so circuits that realize current sources and current matching are particularly important. In Section

Figure 11-65 Differential pair with a current mirror: (a) MOSFET circuit; and (b) BJT circuit.

11.8 it was shown how a single MOSFET can be used to realize a current source. If the gate-to-source voltage of a MOSFET is fixed, a near-constant current source is realized (see the drain-source current equation, Equation (11.17)). A differential amplifier circuit with a variable current source is shown in Figure 11-64(d). What is particularly interesting is that the transistor controlling the bias to the current source can be part of a digital circuit enabling digital control of the analog circuit bias. This is a fundamental component of a **software-defined radio** and of **digital control** of analog circuits, including RFICs. For example, modern RFICs incorporate digitally controlled trimming of RFICs to achieve, for example, enhanced IQ balance of quadrature modulators.

   Another circuit that controls current in an RFIC is the **current mirror** shown in Figure 11-64(c). In this circuit, $I_1 = I_2$, as the gate-source voltages of transistors M_1 and M_2 are the same. The drain-gate connection of M_1 ensures that the gate-source voltage will be whatever is needed to support the current (I_1) that is derived from the rest of the circuit.

   A CMOS differential amplifier with current mirror load is shown in Figure 11-65(a). In this configuration the current mirror presents a high differential impedance. If the load impedance presented to the terminal, labeled V_{OUT}, is less than this load, then the current mirror loaded differential amplifier realizes a single-ended output while having the essential functionality of a differential circuit to reject common-mode signals. There is a price to pay for this functionality. The circuit of Figure 11-65(a) has three drain source voltage drops between the rails. This reduces the available voltage swing. This is one of the major problems encountered with RFICs, as the supply voltage is dictated by the relatively low supply voltages that can be supported in a process that is optimized for low-voltage digital circuits. A current mirror can also be realized using BJTs with the BJT-based current mirror loaded differential pair shown in Figure 11-65(b) as an example.

Mathematically the (simplified) input-output characteristics of MOSFETs are rather simple, being essentially a quadratic (see Equation (11.17) on Page 602). It is a challenge to take such fundamental algebraic models and derive the equations that describe the operation of a complete circuit; a challenge that must be addressed in the synthesis of a circuit with required distortion and noise attributes. It can be shown that the relationship between the drain current and the drain source voltage has the form of a tanh function [207]. In terms of the transconductance, g_m, it appears as a quadratic-like function with a peak value at a drain voltage that is controlled by the W/L ratio. By putting several differential pairs in parallel, with each pair having staggered W/L ratios, a compound differential amplifier with enhanced linearity can be realized [209, 210]. This circuit is known as a multi-tanh differential pair. A triplet **multi-tanh differential pair** is shown in Figure 11-64(e). Detailed RFIC design involves the algebraic derivation of the required conditions. This network synthesis applied to RFICs is explored in numerous references [109, 174, 202–206, 208] as well as a large number of papers on RFIC design. Synthesis to control distortion and noise is the heart of RFIC design. Collecting novel circuit topologies, and the techniques to synthesize them, from conference and journal papers and patents is an essential part of RFIC design; which is not that different from all other forms of circuit design.

Figure 11-66 is a photomicrograph of a 90 nm WCDMA transmitter. The design consists of three blocks: an upconverting mixer (MIXER), a variable gain cascode amplifier (VGA), and a two-stage driver amplifier. The mixer circuit is considered in Section 12.3.4 on Page 720. The schematic of the output amplifier stages is shown in Figure 11-67. The variable gain amplifier, the VGA block, is a cascode amplifier with variable biasing of the top FET in the cascode to realize a variable gain. Each of the amplifiers in the two-stage driver amplifier is driven at a different V_{DD} so that the nonlinearities of the two stages cancel and thus the overall performance of the driver amplifier is linearized [209, 211].

11.16   Management of Amplifier Distortion

The circuits considered in this section are examples of how algebriac calculations can be used to manage distortion in a transistor circuit.

11.16.1   *Distortion in a MOSFET Enhancement-Depletion Amplifier Stage*

In this section power series analysis is applied to the MOS enhancement-depletion amplifier stage shown in Figure 11-68(a). The circuit in Figure 11-68(a) is a Class A amplifier with a common-source enhancement gain stage and a depletion transistor as an active load. Recall that with a MOSFET the voltage of the substrate has an important effect on operation of the

VGA   TWO–STAGE DRIVER AMPLIFIER

MIXER

Figure 11-66 Photomicrograph of a 90 nm CMOS WCDMA transmitter showing: (MIXER) the upconversion mixer and quadrature LO divider; (VGA) the cascode variable gain amplifier; and the two-stage driver amplifier with 9.6 dBm output power. Die size is 1.1×1.4 mm. After Yang and Gard [209] and Yang [211]. Copyright K. Gard and X. Yang, used with permission.

Figure 11-67 Three stage amplifier of the WCDMA receiver shown in Figure 11-66: (a) is a variable gain cascode amplifier; and (b) is the two stage driver amplifier. After Yang and Gard [209] and Yang [211]. Copyright K. Gard and X. Yang, used with permission.

Figure 11-68 MOS amplifier stages: (a) enhancement-depletion amplifier stage; and (b) ultra-linear MOS amplifier stage with an enhancement-mode gain stage and an enhancement-mode load.

transistor. With enhancement-mode transistors the substrate or body is typically connected to the most negative voltage in the circuit, in this case ground. So, since M_1 has the body connected to the source, there is no body effect for M_1. However for M_2 the body effect must be accounted for. Using a simple MOSFET quadratic input-output relationship,

$$I_{D1} = \frac{K_1}{2} \left(V_A + V_X - V_{T0E} \right)^2 = \frac{K_2}{2} \left(-V_{TD}^2 \right) = I_{D2} \qquad (11.224)$$

where the threshold voltage of M_2, with the body effect, is

$$
\begin{aligned}
V_{TD} &= V_{T0D} + \gamma \left(\sqrt{2\phi_F + V_B + V_Y} - \sqrt{2\phi_Y} \right) \\
&= V_1 + \gamma \left(\sqrt{2\phi_F + V_B + V_Y} \right).
\end{aligned}
\qquad (11.225)
$$

The aim here is to develop a relationship between the input signal V_X and the output signal V_Y. The first step in developing a simple relationship is to relate the operating point voltage levels. With $V_X = V_Y = 0$

$$\sqrt{\frac{K_1}{K_2}} \left(V_A - V_{TOE} \right) = - \left(V_1 + \gamma \sqrt{2\phi_F + V_B} \right). \qquad (11.226)$$

Combining (11.224) and (11.225) yields

$$\sqrt{\frac{K_1}{K_2}} \left(V_A + V_X - V_{TOE} \right) = - \left(V_1 + \gamma \sqrt{2\phi_F + V_B + V_Y} \right) \qquad (11.227)$$

$$\sqrt{\frac{K_1}{K_2}} \left(V_A + V_X \right) - \sqrt{\frac{K_1}{K_2}} \left(V_{TOE} \right) = - \left(V_1 + \gamma \sqrt{2\phi_F + V_B + V_Y} \right) \quad (11.228)$$

$$\sqrt{\frac{K_1}{K_2}}V_X = \gamma\sqrt{2\phi_F + V_B} - \gamma\sqrt{2\phi_F + V_B + V_Y} \qquad (11.229)$$

$$\left(\sqrt{\frac{K_1}{K_2}}V_X - \gamma\sqrt{2\phi_F + V_B}\right)^2 = \gamma^2\left(2\phi_F + V_B + V_Y\right) \qquad (11.230)$$

and finally

$$\frac{K_1}{K_2}V_X^2 - 2\gamma\sqrt{\frac{K_1}{K_2}}\sqrt{2\phi_F + V_B} = \gamma^2 V_Y. \qquad (11.231)$$

Therefore, the component of the output that differs from the quiescent pont is

$$V_Y = \frac{-2}{\gamma}\sqrt{\frac{K_1}{K_2}}\sqrt{2\phi_F + V_B}V_X + \frac{K_1}{K_2}V_X^2 \qquad (11.232)$$

which simplifies to

$$V_Y = \frac{-1}{b}\sqrt{\frac{K_1}{K_2}}V_X + \frac{K_1}{K_2}V_X^2 \qquad (11.233)$$

with

$$b = \frac{\gamma}{\sqrt{2\phi_F + V_B}}$$

So the second harmonic distortion level is

$$\mathrm{HD}_2 = \left|\frac{\frac{1}{2}\frac{K_1}{K_2}V_X^2}{\frac{-1}{b}\sqrt{\frac{K_1}{K_2}}V_X}\right| = \frac{b}{2}\sqrt{\frac{K_1}{K_2}}|V_X| \qquad (11.234)$$

So, if there is a lot of voltage gain ($\propto \sqrt{K_1/K_2}$) the input signal level must be small to keep the harmonic distortion down. Other design considerations that have the same effect are to use a wider device which increases the transconductance of the transistor so that the drain current will be maintained for the lower input voltage levels. Of course, making the device wider increases capacitive parasitics which will reduce the operating frequency. Making the channel of the device shorter also increases the transconductance while not affecting the frequency performance. The important point is that there are tradeoffs in modifying the performance of the circuit and these are only made apparent using the type of analysis presented here. The type of analysis presented here is the cornerstone of analog IC and RFIC design.

11.16.2 EXAMPLE: DISTORTION IN THE ULTRA-LINEAR MOS CONNECTION

Transistor M_2 in the circuit of Figure 11-68(b) is an n-channel enhancement load with its well tied to its source. That is, in the fabrication of M_2 a well

is formed and the transistor is constructed in the well. The well now serves as the substrate for M_2. Since the well and the source of M_2 are connected together, M_2 is not subject to the body effect. One important property of this circuit is that the characteristics of M_1 and M_2 are matched. Using a similar approach to that used in the previous section, the input/output relationship of the circuit can be developed. Equating drain currents

$$
\begin{aligned}
I_{D1} &= I_{D2} \\
&= \frac{K_1}{2}\left(V_A + V_X - V_{T0E}\right)^2 \\
&= \frac{K_2}{2}\left(V_{DD} - V_B - V_Y - V_{T0E}\right)^2
\end{aligned}
\tag{11.235}
$$

Given $V_X = V_Y = 0$,

$$
\sqrt{\frac{K_1}{K_2}}\left(V_A - V_{T0E}\right) = V_{DD} - V_B - V_{T0E}.
\tag{11.236}
$$

Therefore

$$
\sqrt{\frac{K_1}{K_2}}V_X + \left(V_{DD} - V_B - V_{T0E}\right) = \left(V_{DD} - V_B - V_Y - V_{T0E}\right),
\tag{11.237}
$$

and

$$
V_Y = -\sqrt{\frac{K_1}{K_2}}V_X.
\tag{11.238}
$$

The result is that, provided the transistors are matched, the amplifier is inherently linear. So, within the approximations of the drain current expressions, there is no distortion, including no third-order intermodulation distortion or spectral regrowth.

11.17 Summary and Further Reading

Amplifier design is a major endeavor and many books have been written about particular aspects of RF and microwave amplifier design. This chapter covered the main topics and also presented treatments that are broadly applied. Power amplifier design is a broad topic and one of intense interest. A power amplifier can be a class AB or higher with the switching amplifiers of particular interest for commercial applications below 6 GHz. Power amplifiers are the subject of a very large number of journal papers and several books. The reader is directed to references [183, 186, 217–219, 221–224, 224, 227]. Understanding, analyzing, and characterizing distortion, and there are numerous references available [200, 212–216, 220, 225, 226, 228–230].

11.18 Exercises

1. What is the gain of the following receiver system?

IL = 3 dB   IL = 1 dB
$G_1 = 25$ dB

2. The Class A BJT amplifier in the figure below has an RF choke providing collector current and acting as an open circuit at RF. The load, R_L, is driven through a capacitor C which is effectively a short circuit at RF. The maximum undistorted efficiency of this circuit is 50%. Derive this efficiency. Ignore the base-emitter voltage drop, $V_{CE,\min}$, and note that the maximum of V_O is V_{CC}, allowing a voltage swing of $\pm V_{CC}$ around the collector quiescent operating voltage. [Parallels Example 11.1 on Page 616]

3. The Class A BJT amplifier in the figure below has a load, R_L, and a maximum undistorted efficiency of 25%. Derive the efficiency of this amplifier in terms of R_E and R_L. Asume that V_{CC} is much greater than V_{BE}. [Parallels Example 11.1 on Page 616]

4. A MOSFET amplifier has the small signal S parameters $S_{11} = 0.9\angle 85°$, $S_{12} = 0.05\angle 15°$, $S_{21} = 2.5\angle 100°$, and $S_{22} = 0.85\angle - 50°$. at 5.6 GHz.

(a) What is the k factor of Rollet's stability criterion?

(b) What does Rollet's stability criterion indicate about the stability of the transistor?

(c) What is the μ factor of Edwards-Sinsky stability criterion?

(d) What does Edwards-Sinsky stability criterion indicate about the stability of the transistor?

5. A MOSFET amplifier has the small signal S parameters $S_{11} = 0.9\angle 85°$, $S_{12} = 0.05\angle 15°$, $S_{21} = 2.5\angle 100°$, and $S_{22} = 0.85\angle - 50°$. at 5.6 GHz.

(a) Calculate the radius and center of the output stability circle. Assume that $|\Gamma_{\text{in}}| < 1$.

(b) Draw the output stability circle on a Smith chart.

(c) Draw conclusions from the plot of the output stability circle. That is, what restrictions are placed on the output matching network?

6. A MOSFET amplifier has the small signal S parameters $S_{11} = 0.9\angle 85°$, $S_{12} = 0.05\angle 15°$, $S_{21} = 2.5\angle 100°$, and $S_{22} = 0.85\angle - 50°$. at 5.6 GHz.

(a) Calculate the radius and center of the input stability circle. Assume that $|\Gamma_{\text{out}}| < 1$.

(b) Draw the input stability circle on a Smith chart.

(c) Draw conclusions from the plot of the input stability circle. That is, what restrictions are placed on the input matching network?

7. A MOSFET amplifier has the small signal S parameters $S_{11} = 0.9\angle 85°$, $S_{12} = 0.05\angle 15°$, $S_{21} = 2.5\angle 100°$, and $S_{22} = 0.85\angle - 50°$. at 5.6 GHz.

(a) What is the maximum unilateral transducer gain?

(b) What is the maximum available power gain?

(c) What is the maximum stable power gain?

(d) What is the unilateral power gain?

8. Consider the design of a 15 GHz inductively-biased Class A amplifier using the pHEMT documented in Figure 11-8. Use the topology shown in Figure 11-22.

   (a) If the input of the transistor is terminated in 50 Ω, what is the impedance looking into the output of the transistor?

   (b) Design a two-element output matching network for maximum power transfer.

9. Design an amplifier for maximum stable gain using the discrete pHEMT transistor described in Figure 11-8. The design specifications are

Gain:	maximum gain at 24 GHz
Topology:	three two-ports (input and output matching networks, and the active device)
Stability:	broadband stability
Bandwidth:	maximum that can be achieved using two-element matching networks
Source Z:	$Z_S = 10\,\Omega$
Load Z:	$Z_L = 50\,\Omega$

   Follow the procedure described in Section 11.7 on Page 636.

10. The system shown below is a receiver with band pass filters, amplifiers and a mixer. [Parallels Example 11.4 on Page 678]

$G_1 = -2$ dB $\quad G_2 = 16$ dB $\quad G_3 = -1$ dB $\quad G_4 = 10$ dB $\quad G_5 = 40$ dB
$NF_1 = 2$ dB $\quad NF_2 = 16$ dB $\quad NF_3 = 1$ dB $\quad NF_4 = 10$ dB $\quad NF_5 = 40$ dB

   (a) What is the total gain of the system?

   (b) What is the noise factor of the first band-pass filter?

   (c) What is the system noise factor?

   (d) What is the system noise figure?

11. An amplifier consists of three cascaded stages with the following characteristics:

Characteristic	Stage 1	Stage 2	Stage 3
Gain	10 dB	15 dB	30 dB
NF	0.8 dB	2 dB	2 dB

What is the noise figure (NF) and gain of the cascade amplifier?

12. The front end of a receiver for a cellular phone has a bandpass filter with a 25 MHz passband and a loss of 2 dB and is followed by two amplifier stages. The first stage has a gain of 20 dB and a noise figure of 0.5 dB and the second stage has a gain of 60 dB and a noise figure of 2 dB.

   (a) Sketch the block diagram of the system as described.

   (b) What is the gain of the system?

   (c) What is the noise figure of the bandpass filter?

   (d) What is the noise figure of the system?

   (e) The system is now connected to an antenna with an effective noise temperature of 30 K and which delivers a signal of 10 pW to the bandpass filter. Determine the noise temperature at the output of the system and hence the noise power in the 25 MHz bandwidth. Determine the signal-to-noise ratio at the output of the front-end system.

13. A 75 Ω attenuator has a loss of 16 dB and is between a source with a Thevenin impedance of 75 Ω and a load of 75 Ω.

   (a) What is the noise power, N_i, available from the 75 Ω source resistor at standard temperature (270 K) in a 1 MHz bandwidth?

   (b) Now consider that the source is connected to the attenuator which is also connected to the load. If the source generates a modulated signal that is 1 MHz wide and has an available power, S_i, of 1 fW, what is SNR_i at the input to the attenuator at standard temperature?

   (c) With the attenuator connected to the source, what is the Thevenin equivalent impedance looking into the output of the attenuator?

   (d) Calculate the noise power, N_o, available from the attenuator with the source attached at standard temperature (270 K) in a 1 MHz bandwidth?

(e) What is the signal power, S_o, delivered to the load?

(f) What is the SNR at the load, SNR_o?

(g) What is the noise factor, F, of the attenuator?

(h) What is the noise figure, NF, of the attenuator?

14. A single-stage amplifier has a gain of 16 dB, an output 1 dB gain compression point of 10 dBm, and an output-referred third-order intercept point, OIP3 = 30 dBm. The noise floor at the output of the amplifier is -60 dBm. The communication protocol has a minimum SNR, SNR_{MIN} of 6 dB.

(a) What is the dynamic range of the amplifier?

(b) What is the SFDR of the amplifier?

15. A single-stage amplifier has a gain of 16 dB, an output 1 dB gain compression point of 10 dBm, and an output-referred third-order intercept point, OIP3 = 30 dBm.

(a) What is the maximum sinusoidal input signal when the output of the amplifier is compressed by 1 dB.

(b) What is the input-referred third-order intercept point, IIP3?

16. Short answer questions. Each part requires a short paragraph of about five lines and a figure where appropriate to illustrate your understanding.

(a) Consider a two-tone signal and describe intermodulation distortion.

(b) Describe the effect of a lossy filter on the SNR. Consider signals at the input and output of the filter?

(c) What is meant by 1 dB gain compression?

(d) Consider a digitally-modulated signal and describe the impact of a nonlinear amplifier on the signal. You must include several (at least) negative effects.

17. An inductively biased Class A HBT amplifier is biased with a collector-emitter quiescent voltage of 5 V and a quiescent collector-emitter current of 100 mA. When operated at the 1 dB compression point, the input RF power is 10 mW and the output power is 100 mW. Consider that the RF signal is a sinewave, and note that quiescent collector-emitter voltage will be the supply rail voltage.

(a) What is the quiescent DC power consumed? Express your answer in milliwatts.

(b) What is the output power in dBm?

(c) What is the efficiency of the amplifier? Note that the efficiency of a class A amplifier can be more than 25% if distortion can be tolerated.

(d) What is the power-added efficiency of the amplifier?

(e) If the input power is reduced by 10 dB so that the amplifier is no longer in compression, will the DC quiescent point change? Explain your answer.

(f) If the input power is reduced 10 dB so that the amplifier is no longer in compression, what is the output power in dBm? Ignore any change in the quiescent point.

(g) With 1 mW input power, what is the power-added efficiency of the amplifier if the quiescent point does not change?

18. The distortion properties of the MOSFET circuit below are captured by the nonlinear transconductance equation $i_{DS1} = a_1 V_{GS1} + a_3 V_{GS1}^3$ where $a_1 = 0.01\text{A}/\text{V}$ and $a_3 = -0.1\text{A}/\text{V}^3$. [Parallels Example 11.5 on Page 688]

(a) What is IIP3 in terms of voltage?

(b) What is OIP3 in terms of voltage?

19. The distortion properties of the MOSFET circuit below are captured by the nonlinear transconductance equation $i_{DS2} = b_1 V_{GS1} + b_3 V_{GS2}^3$ where $b_1 = 0.05\text{A/V}$ and $b_3 = -0.2\text{A/V}^3$.

(a) What is the IIP3 in terms of voltage?

(b) What is the OIP3 in terms of voltage?

20. The distortion properties of the MOSFET circuit below are captured by the nonlinear transconductance equations $i_{DS1} = a_1 V_{GS1} + a_3 V_{GS1}^3$ and $i_{DS2} = b_1 V_{GS1} + b_3 V_{GS2}^3$, where $a_1 = 0.01\text{A/V}$, $a_3 = -0.1\text{A/V}^3$, $b_1 = 0.05\text{A/V}$ and $b_3 = -0.2\text{A/V}^3$.

(a) What is the IIP3 in terms of voltage?

(b) What is the OIP3 in terms of voltage?

21. An amplifier consists of two cascaded stages. The first stage has a gain $G_1 = 20$ dB, an output 1 dB gain compression point $P_{1o}(1\text{ dB}) = 0.1$ dBm, and an output-referred third order intercept, $P_{\text{OIP3}} = 0$ dBm. The second stage has a gain $G_2 = 30$ dB, an output 1 dB gain compression point $P_{2o}(1\text{ dB}) = 1$ dBm, and an output-referred third order intercept, $P_{\text{OIP3}} = 20$ dBm.

(a) Develop a symbolic expression for the input referred 1 dB gain compression point of the cascade amplifier.

(b) What is the input-referred 1 dB gain compression point of the cascade amplifier?

(c) Develop a symbolic expression for the IIP3 of the cascade amplifier.

(d) What is IIP3 of the cascade amplifier?

22. A single-stage amplifier has a gain of 16 dB, an output 1 dB gain compression point of 10 dBm, and an output-referred third-order intercept point, OIP3 = 30 dBm. A communication signal with a PAR of 6 dB is used. What is the maximum average power of the input signal before the output suffers significant compression. This is defined at the point at which the peak signal is compressed by 1 dB.

CHAPTER 12

Mixers, Oscillators, and Switches

12.1   Introduction

Mixers, oscillators, and switches utilize strong nonlinear effects to achieve their objectives. The main function of a mixer is to translate the information contained in a finite bandwidth signal from one center frequency to another where it is either more easily processed (if it is lower) or more easily radiated (if it is higher). RF oscillators generate sinusoidal signals that are either used to drive mixers or, if modulated, to produce modulated signals directly. Switches can be realized by a number of elements but the most common switches in RF engineering are based on FETs or diodes, or are miniature (micromachined) mechanical switches controlled by an applied DC voltage. The first section of this chapter examines diodes used in microwave engineering. The next section presents mixers and their applications. Mechanical, transistor-based, and diode switches are then considered. The final section of this chapter presents an in depth design of a VCO.

Table 12-1 IEEE standard symbols for diodes and a rectifier [117]. [†]Symbol is not an IEEE standard symbol but is commonly used.

Component	Symbol
Diode, general (pn junction)	
Gunn diode	
Pin diode	
Light emitting diode (LED)	
Rectifier	
Schottky diode[†]	
Varactor diode	
Zener diode	

12.2 Diodes

Diodes are two-terminal devices that have nonlinear current-voltage characteristics. Diodes are either fabricated using silicon or compound semiconductor, with GaAs being the most common. The most common diodes used in RF and microwave engineering are listed in Table 12-1 along with their standard symbols. Ideally a **rectifier** allows current to flow in one direction and not in the other. A general diode, referred to as a junction diode or a **Schottky diode**, is one type of rectifier, but rectifiers are more general, and, for example, they can be realized using vacuum devices.

12.2.1 Junction and Schottky Diodes

Junction diodes are two-terminal devices that derive their characteristics from the barrier effect that occurs at the junction of two different types of semiconductor (one with excess free holes and one with excess free electrons) or at the interface of a metal and a semiconductor. The result is an asymmetric current-voltage characteristic, as shown in Figure 12-1(a). This is not an ideal rectifier characteristic, as it requires a threshold voltage to be reached before there is appreciable current flow. The diode characteristics are described by the following equation [168–170]

$$I = I_0 \left[\exp\left(\frac{qV}{nkT} \right) - 1 \right] \qquad (12.1)$$

where V is the voltage across the junction, q is the charge on an electron, k is the Boltzmann constant ($1.37 \cdot 10^{-23}$ J/K), and T is the absolute temperature (in kelvin). I_0 is the reverse saturation current and is small,

Figure 12-1 Characteristics of a pn junction diode or a Schottky diode: (a) current-voltage characteristics; capacitance-voltage characteristics; and (c) diode model.

with values ranging from 1 pA to 1 nA. The quantity n is the diode ideality factor, with $n = 2$ for graded junction pn junction diodes, and $n = 1.0$ for step junction diodes where the interface between p-type and n-type semiconductor materials is abrupt. The abrupt junction is most closely realized by a Schottky diode, where a metal forms one side of the interface (typically) replacing the p-type semiconductor. Carriers recombine quickly in the metal, much faster than they would in a p-type semiconductor. Hence a Schottky diode operates at higher frequencies than does a pn junction diode. A unity ideality factor, n, of 1 is the best that can be achieved. When the voltage is sufficiently positive to cause a large current to flow, the diode is said to be forward biased. When the voltage is negative, the current flow is negligible, and the diode is said to be reverse biased. At sufficiently large reverse bias, electrons are ripped from the valence bands of the semiconductor atoms and the current rapidly increases in a process called impact ionization or avalanche. The voltage at which this occurs is called the reverse breakdown voltage.

In a semiconductor diode, charge is separated over distance and so a diode has appreciable capacitance, called the junction capacitance, mathematically modeled as

$$C_j(V) = \frac{C_{j0}}{(1 - (V/\phi))^{\gamma}}, \tag{12.2}$$

where ϕ is the built-in potential difference across the diode. This capacitance profile is shown in Figure 12-1(b). The built-in potential is typically 0.6 V for silicon diodes and 0.75 V for GaAs diodes. The doping profile can be adjusted so that γ can be less than the ideal $\frac{1}{2}$ of an abrupt junction diode.

Current must flow through bulk semiconductor before reaching the active region of the semiconductor diode, and so there will be a resistive voltage drop. Combining effects leads to the equivalent circuit of a pn junction diode or a Schottky diode, shown in Figure 12-1(c).

12.2.2 Varactor Diodes

A varactor diode is a pn junction diode operated in reverse bias and optimized for good performance as a tunable capacitor. Ideally it has low reverse saturation current, high reverse breakdown voltage, and a specific capacitance profile designed for a particular application. A common application is a VCO where the varactor, with voltage-dependent capacitance, C, is part of a resonant circuit (often called a tank circuit) with a lumped inductor, L. The resonant frequency of the tank circuit is proportional to $1/\sqrt{LC}$, so by applying a voltage to a varactor diode, C changes, the resonant frequency of the tank circuit is tuned, and the oscillation frequency is changed. The capacitance versus voltage is described by Equation (12.2) and voltage dependence is shown in Figure 12-1(b).

12.2.3 PIN Diodes

A PIN diode is a variation on a pn junction diode with a region of intrinsic semiconductor (the I in PIN) between the p-type and n-type semiconductor regions. The properties of the PIN diode depend on whether there are carriers in the intrinsic region. The PIN diode has the current-voltage characteristics of a pn junction diode at low frequencies; however, at high frequencies it looks like a linear resistor, as carriers in the intrinsic region move slowly. When a forward DC voltage is applied to the PIN diode, the intrinsic region floods with carriers, and at microwave frequencies the PIN diode is modeled as a low-value resistor. At high frequencies there is not enough time to remove the carriers in the intrinsic region, so even if the total voltage (DC plus RF) across the PIN diode is negative, there are carriers in the intrinsic region throughout the RF cycle. If the DC voltage is negative, carriers are removed from the intrinsic region and the diode looks like a large-value resistor. The PIN diode can be used as a microwave switch controlled by a DC voltage.

12.2.4 Other Diodes

Any nonlinear two-terminal device can be called a diode, so it is not surprising that there are a large number of different types. The more common diodes are shown in Table 12-1. **Gunn diodes** are more correctly called **Gunn devices** and are sometimes referred to as **transferred electron devices** [168, 170]. They are n-type semiconductors and their properties derive from the complex band structure of compound semiconductors such as GaAs and InP. Under the right conditions, a **charge domain** (a local separation of net positive and negative charge) can form that establishes a negative differential resistance region. They are commonly used in oscillators. While they can operate at microwave frequencies, they

are not as frequently used today as they were, but they can operate at 100s of gigahertz.

Transit time devices [234] exploit the finite time for carriers to cross semiconductor material. An **IMPATT** (impact ionization avalanche transit-time) diode is a pn junction diode with a depletion region forming the transit time regions adjacent to the junction [232, 233]. When the diode is biased beyond the avalanche breakdown voltage, carriers are injected into the transit time region. A similar device is the **TRAPATT** diode (trapped plasma avalanche transit time diode), a pn junction diode, where the carrier injection results from a trapped space-charge plasma formed within the junction region [168, 238].

As a result of the transit time delay, the RF current through a DC-biased transit time device and the RF voltage across it can become out-of-phase to the extent that, at microwave frequencies, a transit time device can look like a negative resistance and so can be used as the active component of an oscillator [235]. They can be used as amplifiers as a reflection device with a reflection coefficient greater than one [231]. They are not commonly used today.

Zener diodes are pn junction or Schottky diodes that have been specially design to have sharp reverse breakdown characteristics [168]. They can be used to establish a voltage reference or, used as a limiter diode, to provide protection of more sensitive circuitry. As a limiter they are found in communication devices in a back-to-back configuration to limit the voltages that can be applied to sensitive RF circuitry.

In semiconductor diodes, the recombination of holes and electrons can result in photons being generated. This effect is enhanced to form **light emitting diodes** (**LEDs**) [168, 170, 234, 236]. **Tunnel diodes** create a dynamic negative resistance through the physical process of electrons tunneling from one energy band to another in a highly doped semiconductor [168, 170, 237]. This occurs if the valence bands and conduction bands are sufficiently close in energy-momentum space. In the late 1950s, 1960s and 1970s, tunnel diodes found use as active components of microwave and millimeter-wave oscillators. They have been replaced by more conventional semiconductor devices.

12.3 Mixers

Frequency conversion, mixing or heterodyning, is the process of converting information at one frequency (present in the form of a modulated carrier) to another frequency. The second frequency is either higher, in the case of frequency **up-conversion**, where it is more easily transmitted, or lower, when mixing is called frequency **down-conversion**, where it is more easily captured. Capture of the down-converted signal is nearly always by an ADC. Frequency conversion occurs with any nonlinear element. In Figure

Figure 12-2 Diode mixer: (a) circuit; (b) diode current-voltage characteristic; (c) spectrum across the nonlinear device; and (d) schematic symbol for a mixer.

12-2(a) a nonlinear device is driven by two signals at ω_m and ω_c. The larger signal, the LO, is called the **pump** and the other signal is called the RF. The spectrum of the signals present in the circuit is shown in Figure 12-2(c). The aim here is to produce a signal at the difference frequency (or intermediate frequency (IF)) with the same modulation, and hence the same information, as the original RF signal. Two mixers based on transistors are shown in Figure 12-3. The transistor mixer shown in Figure 12-3(a) uses filtering to separate the RF, LO, and IF components. Filters can be large, so one of the particular advantages of the **Gilbert mixer** shown in Figure 12-3(b) is that it is a balanced mixer and filtering is not required to separate the various signals. The symmetrical (or balanced) nature of this circuit means that only differential mode signals at the input of the common source differential pairs can appear at the output. Thus the largest signal present, the LO, is suppressed. A balanced mixer can also be realized using diodes arranged in a ring to form the **diode ring double-balanced mixer** shown in Figure 12-4. This mixer has the advantage of being bidirectional, whereas the transistor mixer circuits of Figure 12-3 are unidirectional.

(a)

(b)

Figure 12-3 Transistor-based mixer circuits: (a) single-ended **FET mixer** with LO, RF, and IF bandpass filters; (b) Gilbert mixer with baluns producing differential LO and RF signals.

Figure 12-4 Diode ring double-balanced mixer.

12.3.1 Mixer Analysis

A mixer can be designed around any nonlinear device. Using an operational amplifier, an ideal multiplier can be designed which will multiply two signals, each described as cosinusoids. If the LO cosinusoid is $\cos(\omega_1 t)$ and the input RF cosinusoid $\cos(\omega_2 t)$, the multiplication of the two signals will produce an output,

$$y(t) = [\cos(\omega_1 t)] \cdot [\cos(\omega_2 t)]$$
$$= \frac{1}{2} \left\{ \cos[(\omega_1 - \omega_2)t] \cos[(\omega_1 + \omega_2)t] \right\}. \qquad (12.3)$$

This has two components: one at the radian frequency $(\omega_1 - \omega_2)$; and the other at $(\omega_1 + \omega_2)$. If the LO and RF signal are close, then the component at

$(\omega_1 - \omega_2)$ will be at a frequency much lower than either the LO or the RF, and the component at $(\omega_1 + \omega_2)$ will be at almost twice the input frequencies. Appropriate filtering will choose one of these components depending on whether the application is up-conversion or down-conversion. Operational amplifiers are not available at microwave frequencies and at RF they do not have sufficient dynamic range. So circuits that do not realize ideal multiplication must be used. In this section, what happens when two components are applied to an arbitrary nonlinearity will be considered. The result is that a large number of tones will be generated in the mixing process. Balanced circuit designs can eliminate many of these tones, greatly reducing the filtering required. Filtering is of concern because efficient filters are large and communication devices must be miniaturized.

A two-tone input

$$x(t) = |X_1| \cos(\omega_1 t + \phi_1) + |X_2| \cos(\omega_2 t + \phi_2)$$

can be written using complex notation as

$$x(t) = \frac{1}{2} \left[\tilde{X}_1 e^{\jmath\omega_1 t} + \tilde{X}_1^* e^{-\jmath\omega_1 t} + \tilde{X}_2 e^{\jmath\omega_2 t} + \tilde{X}_2^* e^{-\jmath\omega_2 t} \right].$$

Note that the coefficient of the positive exponential frequency component is one-half that of the phasor. Thus the phasor of the ω_1 component is $\hat{X}_1 = |X_1| e^{(\jmath\phi)} = 2\tilde{X}_1$ and the phasor of the ω_2 component is $\hat{X}_2 = |X_2| e^{(\jmath\phi)} = 2\tilde{X}_2$. So the first three powers of x can be easily expanded manually; for example, expanding x^2 gives

$$
\begin{aligned}
x^2(t) \quad = \quad & \left(\frac{1}{2}\right)^2 \left[X_1^2 e^{\jmath 2\omega_1 t} + 2X_1 X_1^* + 2X_1 X_2 e^{\jmath(\omega_1 + \omega_2)t} + 2X_1 X_2^* e^{\jmath(\omega_1 - \omega_2)t} \right. \\
& + (X_1^*)^2 e^{-\jmath 2\omega_1 t} + 2X_1^* X_2 e^{\jmath(\omega_2 - \omega_1)t} + 2X_1^* X_2^* e^{-\jmath(\omega_1 + \omega_2)t} + X_2^2 e^{\jmath 2\omega_2 t} \\
& \left. + 2X_2 X_2^* + (X_2^*)^2 e^{-\jmath 2\omega_2 t} \right],
\end{aligned}
$$
(12.4)

and similarly, expanding x^3 yields

$$
\begin{aligned}
x^3(t) \quad = \quad & \left(\frac{1}{2}\right)^3 \left[X_1^3 e^{\jmath 3\omega_1 t} + 3X_1^2 X_1^* e^{\jmath\omega_1 t} + 3X_1^2 X_2 e^{\jmath(2\omega_1 + \omega_2)t} \right. \\
& + 3X_1^2 X_2^* e^{\jmath(2\omega_1 - \omega_2)t} + 3X_1(X_1^*)^2 e^{-\jmath\omega_1 t} + 6X_1 X_1^* X_2 e^{\jmath\omega_2 t}
\end{aligned}
$$

$$
\begin{aligned}
& + 6X_1 X_1^* X_2^* e^{-\jmath\omega_2 t} + 3X_1 X_2^2 e^{\jmath(\omega_1 + 2\omega_2)t} \\
& + 6X_1 X_2 X_2^* e^{\jmath\omega_1 t} + 3X_1(X_2^*)^2 e^{\jmath(\omega_1 - 2\omega_2)t} + (X_1^*)^3 e^{-\jmath 3\omega_1 t} \\
& + 3(X_1^*)^2 X_2 e^{\jmath(\omega_2 - 2\omega_1)t} + 3(X_1^*)^2 X_2^* e^{-\jmath(2\omega_1 + \omega_2)t} + 3X_1^* X_2^2 e^{\jmath(2\omega_2 - \omega_1)t} \\
& + 3X_1^*(X_2^*)^2 e^{-\jmath(\omega_1 + 2\omega_2)t} + X_2^3 e^{\jmath 3\omega_2 t} + 6X_1^* X_2^* X_2 e^{-\jmath\omega_1 t} + 3X_2^2 X_2^* e^{\jmath\omega_2 t} \\
& \left. + 3X_2(X_2^*)^2 e^{-\jmath\omega_2 t} + (X_2^*)^3 e^{-\jmath 3\omega_2 t} \right],
\end{aligned}
$$
(12.5)

so that the output of the cubic equation,

$$y(t) = a_0 + a_1 x(t) + a_2 x^2(t) + a_3 x^3(t),$$

can be calculated for a two-tone input. Table 12-2 lists these phasors and groups them according to frequency. The phasors of the various intermodulation products resulting from x, x^2, and x^3 can be taken as the coefficients of the positive exponential frequency components after the factor of two correction for terms other than DC. Terms of the same frequency are summed to obtain the output at a particular frequency. For example, the phasor output at ω_1 is given by the sum of three intermodulation products:

$$Y_{\omega_1} = a_1 \left(\frac{1}{2}\right) X_1 + 3a_3 \left(\frac{1}{2}\right)^3 X_1^2 X_1^* + 6a_3 \left(\frac{1}{2}\right)^3 X_1 X_2 X_2^*. \qquad (12.6)$$

So the level of the output of a mixer is related directly to the strength of the LO signal (with amplitude $|X_1|$), the strength of the nonlinearity (captured by the a_n coefficients), and the level of the original signal $|X_3|$. Unfortunately, low-order power series analysis as used here is not sufficient to model practical mixers and computer-aided modeling tools are necessary.

12.3.2   Mixer Performance Parameters

The main characteristics that define the performance of a mixer are the conversion gain or loss and the noise figure. There are others that define other attributes. Mixing results from a nonlinear process, so most of the additional performance parameters derive from the generation of nonlinear responses in addition to the desired response. The mixer performance parameters are as follows.

Conversion loss: This is the ratio of the available power of the input signal to that of the output signal after mixing. It is usually expressed in decibels. In the diode mixer shown in Figure 12-5, the conversion loss is

$$L_C = \frac{P_{\text{in}}(\text{RF})}{P_{\text{out}}(\text{IF})}. \qquad (12.7)$$

In decibels, the conversion loss is

$$L_C|_{\text{dB}} = 10 \log_{10} \left[\frac{P_{\text{in}}(\text{RF})}{P_{\text{out}}(\text{IF})}\right]. \qquad (12.8)$$

Noise figure: The NF is 10 times the log of the noise factor F. The NF is the ratio of the SNR at the RF input to the SNR at the IF output (using the input noise generated by a resistor at standard temperature, 290K). However,

Table 12-2   The intermodulation products resulting from x, x^2, and x^3, where x is a two-tone signal, showing only the positive frequencies. The first column gives the complex amplitudes of the frequency components.

Intermodulation Product	Frequency	Order
$\frac{1}{2}X_1X_1^*$	0	2
$\frac{1}{2}X_2X_2^*$	0	2
$2\frac{1}{2}X_1$	ω_1	1
$2(\frac{1}{2})^33X_1^2X_1^*$	ω_1	3
$2(\frac{1}{2})^36X_1X_2X_2^*$	ω_1	3
$2\frac{1}{2}X_2$	ω_2	1
$2(\frac{1}{2})^33X_2^2X_2^*$	ω_2	3
$2(\frac{1}{2})^36X_1X_1^*X_2$	ω_2	3
$2(\frac{1}{2})^2X_1^2$	$2\omega_1$	2
$2(\frac{1}{2})^2X_2^2$	$2\omega_2$	2
$2(\frac{1}{2})^3X_1^3$	$3\omega_1$	3
$2(\frac{1}{2})^3X_2^3$	$3\omega_2$	3
$2\frac{1}{2}X_1X_2$	$\omega_1+\omega_2$	2
$2\frac{1}{2}X_1X_2^*$	$\omega_1-\omega_2$	2
$2(\frac{1}{2})^33X_1^2X_2$	$2\omega_1+\omega_2$	3
$2(\frac{1}{2})^33X_1^2X_2^*$	$2\omega_1-\omega_2$	3
$2(\frac{1}{2})^33X_1X_2^2$	$\omega_1+2\omega_2$	3
$2(\frac{1}{2})^33X_1^*X_2^2$	$2\omega_2-\omega_1$	3

there are qualifications for mixers. The first is that two input signals (at the RF and image frequencies) can produce noise and signal power at the IF. The double-sideband (DSB) NF includes signal and noise contributions at both the RF and the image frequencies. Single-sideband (SSB) NF includes the input signal at the RF only, but includes noise at both the RF and at the image frequencies.

Image rejection: If a mixer has an LO of 1.1 GHz and an RF of 1500 MHz, then the IF (the difference frequency) will be at 400 MHz. This IF can also be generated by a signal, called the image, at 700 MHz. Numerically this is the reflection of the RF in the LO. Image rejection refers to the ability of a mixer

to reject the image signal. This can be achieved using an input RF bandpass filter.

If the level of the output signal (at the IF) produced by the intended RF signal is P_{out}, and that produced by the image signal is $P_{\text{out,image}}$, then the **image rejection ratio** (**IRR**) is

$$\text{IRR} = \frac{P_{\text{out}}}{P_{\text{out,image}}}. \tag{12.9}$$

This is typically expressed in decibels and

$$\text{IRR} = 10 \log_{10} \left(\frac{P_{\text{out}}}{P_{\text{out,image}}} \right). \tag{12.10}$$

12.3.3 Mixer Waveforms

This section presents the waveforms of two diode mixers. Diode mixers, such as the balanced mixer in Figure 12-4, are bilateral. In a communication transceiver such a mixer can be used for both receive and transmit, greatly simplifying filtering requirements. Most mixers in cell phones, however, are based on transistors.

The first mixer to be considered is the single-ended mixer shown in Figure 12-5(a). Filters are not included in the circuit to simplify the circuit and to see the full complexity of the spectra in the circuit. The transient waveform across the diode (measured at the test point) is shown in Figure 12-5(b). The turn-on transient is due to both the turn-on ramp of the LO and RF sources and to the internal capacitance of the diode. The spectra of the signal at the test point is shown in Figures 12-5(c) and 12-5(d). The desired signal here is the IF which is at the difference frequency of the LO and RF (i.e., at 400 MHz). Extracting just the IF in the response shown in Figure 12-5(d) would require significant filtering. The level of the IF tone here is 11.5 mV$_{\text{peak}}$. If a lossless 400 MHz filter is attached to the test point and the IF load (on the other side of the filter) is 50 Ω, the IF power delivered to the load is $P_{\text{out}} = \frac{1}{2}(11.5 \text{ mV})^2/(50 \text{ }\Omega) = 1.322 \text{ }\mu\text{W}$. The available RF power is $P_{\text{in}} = \frac{1}{2}(100 \text{ mV})^2/(50 \text{ }\Omega) = 100 \text{ }\mu\text{W}$. So the conversion loss is

$$L_C = \frac{100 \text{ }\mu\text{W}}{1.332 \text{ }\mu\text{W}} = 75.08 = 18.8 \text{ dB}. \tag{12.11}$$

Similar results are presented for the diode ring mixer shown in Figure 12-6. Waveforms and spectra are shown for LO, IF, and RF test points in Figure 12-7. Note how relatively clean the spectra at each of the test points. The reduction in clutter is the result of the balanced circuit. This is known as a double-balanced mixer and the same concept can be used with transistor mixers. Focusing on just the spectra at the IF test point (Figure 12-6(f)), it can be seen that the LO signal and its harmonics are non

(a)

(b)

(c)

(d)

Figure 12-5 Diode mixer: (a) schematic; (b) waveform at test point; (c) spectrum; and (d) expanded spectrum at test point.

Figure 12-6 Diode double-balanced mixer schematic.

existent. The components other than the 400 MHz IF are other sum and difference products of the RF and LO. However, compared to the spectra of the single-ended diode mixer in Figures 12-5(c) and 12-5(d), many of these are eliminated as well. This is a very attractive circuit, as it significantly reduces the specifications required of an output filter. This is one of the

Figure 12-7 Waveforms and spectra of the diode double-balanced diode ring mixer of Figure 12-6: (a) spectrum and (b) waveform at LO test point; (c), (d) at RF test point; and (e), (f) at IF test point.

Figure 12-8   Gilbert cell double-balanced mixer: (a) basic circuit; and (b) folded cascode implementation with larger voltage swing.

special characteristics of RF and microwave design. Much can be gained by being creative—a designer gets better with experience. The level of the IF signal at the RF test point is 26.7 mV$_{\text{peak}}$. The IF power delivered to the 50 Ω load is $P_{\text{out}} = \frac{1}{2}(26.7\,\text{mV})^2/(50\,\Omega) = 7.126\,\mu\text{W}$. The available RF power is $P_{\text{in}} = \frac{1}{2}(100\,\text{mV})^2/(50\,\Omega) = 100\,\mu\text{W}$. So the conversion loss is

$$L_C = \frac{100\,\mu\text{W}}{7.126\,\mu\text{W}} = 14.03 = 11.47\,\text{dB}. \qquad (12.12)$$

The conversion loss is much lower than was obtained with the single-ended mixer. In part this is because power was not dissipated in a large number of spurious tones.

12.3.4   Mixers and RFICs

The mixer circuit used in RFICs is most commonly a variation of the Gilbert mixer. This is a four-quadrant mixer that has an almost ideal multiplier response limited by how closely transistors can be matched [239]. In RFICs the transistors can be matched very well. A double-balanced Gilbert cell mixer is shown in Figure 12-8(a). This circuit has good performance and rejects RF and LO feed-through to the output, and this is dependent on the

Figure 12-9 Quadrature up-converting mixer of the WCDMA receiver shown in Figure 11-66. BB is the low frequency baseband signal with (I) and (Q) components. $\overline{BB}$ is the inverse (negative) of BB. LO and $\overline{LO}$ are the differential local oscillator signal, and RF is the single-ended output signal. The LO is at 2 GHz, the baseband input frequency is 5±1.72 MHz, and the output frequency is 1.95 MHz. After Yang and Gard [209] and Yang [211]. Copyright K. Gard and X. Yang, used with permission.

quality of the transistor match. One problem with the circuit is the reduced voltage swing resulting from three drain-source voltage drops between the supply rails. The classic technique for solving this problem is to use a folded cascode design. The Gilbert cell double-balanced mixer then becomes the circuit of Figure 12-8(b) [209, 211]. This merges the cascode amplifier designs in Figures 11-64(a) and 11-64(b) with the double-balanced Gilbert cell mixer in Figure 12-8(a). This design is used in the up-converting mixer of the WCDMA transmitter RFIC presented in Section 11.15 and Figure 12-9 on Page 721. The circuit schematic of the upconverting mixer is shown in Figure 12-8(b). The double-balanced Gilbert cell mixer in Figure 12-8(a) is duplicated to realize a quadrature mixer. The differential amplifier of the double balanced Gilbert cell mixer in Figure 12-8(a) is replaced by a multi-tanh triplet differential amplifier as described in Section 11.15 (see Figure 11-64(e)).

12.4 Switches

Microwave switches are commonly used to alternately connect an antenna to a transmitter or a receiver. In some communication systems, such as GSM, a phone does not transmit and receive simultaneously. Consequently a switch can be used to separate the transmitted and received signals. In multiband phones, a switch is used to connect the right transmitter and receiver, which are band specific, to the antenna. In radar systems, switches are used to steer an antenna beam by changing the phase of the microwave signal delivered to each antenna in an array of antennas. An ideal microwave switch is shown in Figure 12-10(a), where an input port, $\mathrm{RF_{IN}}$, and an output port, $\mathrm{RF_{OUT}}$, are shown. For maximum power transfer between the ports, the switch should have little loss and thus low on resistance. At microwave frequencies, switches must be modeled with parasitics and have finite on and off resistance. A realistic model applicable to many switch types is shown in Figure 12-10(b). The capacitive parasitics, the C_Ps, limit the frequency of operation of the switches, and the on resistance, R_{ON}, impacts the switch loss. Ideally the off resistance, R_{OFF}, is very large, however, the parasitic shunt capacitance, C_{OFF}, is nearly always more significant. The result is that at high frequencies there is an alternative capacitive connection between the input and output through C_{OFF}. The on resistance of the switch introduces voltage division which can be seen by comparing the ideal connection shown in Figure 12-10(c) and the more realistic connection shown in Figure 12-10(d). From the voltage division ratio, the loss of the switch can be calculated.

Switches are configured to provide connection from one or more inputs to one or more outputs. The configuration of a switch is indicated by poles and throws, and several configurations are shown in Figure 12-11. Most commonly in microwave applications single-pole switches are used and this input is connected to an antenna. The throws would be connected to different bands of a multiband phone for example.

There a four types of microwave switches: mechanical, PIN diode, FET, and **microelectromechanical system** (**MEMS**) switches. Mechanical switches are nearly ideal but tend to be large, relatively expensive, and are mostly used in laboratory settings. The other switches are of most interest for use in systems. The PIN diode, FET, and MEMS switches are shown in Figures 12-10(e), 12-10(f), and 12-10(g), respectively. With these technologies, most higher-order switches are based on interconnections of **SPST switches**. The attributes of these switches is summarized in Table 12-3 for switches that are suitable for cell phone applications. PIN diode switches are the most robust, handling the most RF power, and operating at higher frequencies than either FET- or MEMS-based switches. However, this comes at a price. The PIN diode used is similar to a pn junction diode with the addition of an intrinsic layer between the p- and n-type materials. With applied forward bias the diode has low series resistance. In reverse bias

Figure 12-10 Microwave switches: (a) ideal switch connecting RF_{IN} and RF_{OUT} ports; (b) model of a microwave switch; (c) ideal circuit model with switch on and with source and load; (d) realistic low-frequency circuit model with switch on and with source and load; (e) switch realized using a PIN diode; (f) switch realized using an FET; and (g) switch realized using a MEMS switch.

the diode resistance is large. Forward bias requires DC current and voltage, so control power is consumed when a PIN diode switch is on. The circuit configuration for an SPST PIN diode switch is shown in Figure 12-10(e). Series bias decoupling capacitors are required at the RF ports.

A FET makes a good electronic switch; with the correct bias applied to the gate, the drain-source connection looks like a small resistance. Changing the bias to the other extreme removes free carriers from the channel between the drain and source, and a large resistance is the result. Both Si and GaAs

Figure 12-11   Switch configurations.

Table 12-3   Typical properties of small microwave switches. (Sources: [1]Radant MEMS, [2]RF Micro Devices, and [3]Tyco Electronics.)

Switch Type	Configuration	Power Handling	Loss at 2 GHz	Operating Frequency	Actuation Voltage	Response Time
MEMS[1]	SPST	1 W	0.15 dB	to 12 GHz	40–120 V	5 μs
MEMS[1]	SPST	0.5 W	0.27 dB	to 40 GHz	40–120 V	5 μs
pHEMT[2]	SPDT	2 W	0.5 dB	to 2.5 GHz	6 V	0.5 μs
pHEMT[2]	SPDT	10 W	0.5 dB	to 6 GHz	6 V	0.5 μs
PIN[3]	SPDT	13 W	0.35 dB	to 2 GHz	12 V	0.5 μs
PIN[3]	SPDT	10 W	0.4 dB	to 6 GHz	12 V	0.5 μs

switches are used at cellular frequencies, with GaAs switches operating at extended frequencies approaching 6 GHz. The operation of a FET can be described as a variable drain-source resistance with the gate-source voltage controlling the crosssection of the channel. The circuit for an FET-based SPST switch is shown in Figure 12-10(f). Series bias blocking capacitors are required at the RF ports. Control power is only required to change the state of the switch; negligible power is required to maintain the switch state.

A MEMS switch is fabricated using photolithographic techniques similar to those used in semiconductor manufacturing [240–242]. They

are essentially miniature mechanical switches with a voltage used to control the position of a shorting arm, which is usually a cantilever or a membrane. As there is no direct connection between the RF signal path and the control circuitry, MEMS switches have inherently high operating frequencies. Power is required to change the switch, but once switching has been accomplished negligible DC power is required to maintain the connection. The basic structure of a MEMS switch that makes a metal-to-metal contact is shown in Figure 12-12. A large number of variations on this structure have been developed. One of the major problems has been creating the contact with enough force to reduce the series resistance. Figure 12-13 shows the concept of a cantilever beam, and a switch fabricated using this concept is shown in Figure 12-14. Generally it is not necessary to make a DC contact, and creating a structure that has a large difference between the on capacitance and off capacitance is sufficient for applications such as tunable matching networks [240].

12.5 Oscillators

RF and microwave oscillators can be designed using either a two-port or a one-port approach. The classic treatment of oscillators is based on a two-port gain device in a feedback loop, but the oscillator can nearly always be viewed, and thus more conveniently designed, as a one-port in which a resonant circuit, called the tank circuit, is connected to an active circuit which presents a negative resistance. Microwave oscillator design is almost always based on one-port circuits. Oscillators in RFICs, however, are just as often designed by considering a two-port with feedback.

12.5.1 Two-Port Oscillators

The two-port view of oscillators is based on a feedback loop, as shown in Figure 12-15. Here $L(A, \omega)$ describes the characteristics of an amplifier whose transfer function is both amplitude, A, and frequency, ω, dependent and has a magnitude greater than 1. $H(\omega)$ describes the characteristics of a linear feedback network. The output of the feedback system is described by

$$V_{\text{out}} = L(A, \omega)V_{\text{in}} + L(A, \omega)H(\omega)V_{\text{out}} \qquad (12.13)$$

or

$$V_{\text{out}} = \frac{L(A, \omega)V_{\text{in}}}{1 - L(A, \omega)H(\omega)}. \qquad (12.14)$$

The aim of most oscillator design is to use an active device whose characteristics are independent of frequency but whose transfer response depends on amplitude. Then Equation (12.14) becomes

$$V_{\text{out}} = \frac{L(A, \omega)V_{\text{in}}}{1 - L(A, \omega)H(\omega)}. \qquad (12.15)$$

Figure 12-12 Traditional architecture of metal-to-metal contact switches: (a) a schematic of the switch structure before the bias voltage is applied; (b) an electric field is generated between the movable beam and the actuation electrode immediately after the bias voltage is applied to the structure. This field results in the electrostatic actuation force; and (c) after sufficient time (typically on the order of a few microseconds) the movable beam contacts the designated metal and a short circuit is created. The electrostatically generated force is used as the contact force between the two metals. After Katehi and Peroulis [240].

Oscillation begins with input noise when the oscillator is powered on. If the denominator of Equation (12.15) is close to zero, oscillations build up at a frequency determined by the feedback network. As the amplitude of the oscillations build up, $L(A)$ compresses until the denominator is finite

Movable cantilever beam   Static cantilever beam

Actuation pad covered by dielectric layer

RF line   RF line

Si substrate

Vbias = 0V   Vbias > 14 V

(a)   (b)

Figure 12-13 Basic switch concept and operation: (a) when no bias is applied, the movable cantilever touches the stiff static beam leading to metal-to-metal contact; and (b) if the DC bias is higher than the actuation voltage of the movable beam, this beam deflects downward leading to an open circuit. After Katehi and Peroulis [240].

CPW ground planes   Movable cantilever beam

High resistivity bias lines   DC bias pads

Acc.V Spot Magn Det WD   100 µm
5.00 kV 3.0 250x   SE   36.2

Figure 12-14 Scanning electromicrograph (SEM) pictures of fabricated metal-to-metal switches. After Katehi and Peroulis [240].

$H(\omega)$

V_{in} ⊕ → $L(A,\omega)$ → V_{out}

Figure 12-15 Representation of a feedback oscillator based on a two-port active device.

and there are stable oscillations.[1] Stable oscillation is no accident and is the

---

[1] This description of oscillation, based on Equation (12.15), is behind the erroneous Barkhausen stability criterion, which is also known as the Barkhausen oscillation criterion. Barkhausen himself used the criterion, known as the Barkhausen criterion, to establish the frequency of oscillation as $L(A,\omega)H(\omega) = 1$. This was misinterpreted as a stability or

Figure 12-16   Feedback oscillator with Pi- and T-type feedback networks: (a) Pi feedback network; and (b) T feedback network

Figure 12-17   Basic oscillator feedback networks: (a) Hartley; (b) Colpitts; and (c) Clapp circuits.

result of careful design, and much depends on the nature of the feedback network. There are two dominant types of feedback networks, the Pi- and T-type networks, shown in Figure 12-16. Three Pi-type networks have proven to be particularly suited to the amplitude saturation characteristics of FET and BJT active devices. The particular forms of the networks that have the right characteristics to ensure stable oscillation are the Hartley, Colpitts, and Clapp circuits shown in Figure 12-17. Single transistor oscillators using these feedback networks can be arranged in common gate, common source or common drain configurations (see Figure 12-18). Crystal references are commonly used for fixed-frequency oscillators, as they have high Q values. A circuit with a crystal-based feedback loop is shown in Figure 12-19. The Hartley, Colpitts, and Clapp oscillators are widely used as the basis of oscillators in CMOS RFICs where the feedback network can be closely approximated using differential circuits and designs with high transistor

oscillation criterion. It is a necessary criterion for two-port feedback oscillation, but not sufficient. It does not indicate whether a system is unstable. Instead the Nyquist criterion is the necessary and sufficient criterion for oscillation in feedback oscillators [243–246]. The Barkhausen criterion should not be used in determining whether oscillation will occur.

Common Gate   Common Drain   Common Source

Figure 12-18 Circuit schematics for FET feedback oscillators using Pi-type feedback network: (a) common gate; (b) common drain; and (c) common source. The current sources provide bias.

Figure 12-19 Crystal oscillator: (a) schematic symbol for a crystal; (b) equivalent circuit ; and (c) FET crystal oscillator.

counts [109, 202–206].

12.5.2 Phase Noise in Oscillators

The performance of most RF and microwave systems is limited by oscillator noise. The two main sources of noise in electronic devices are white noise, which has a frequency-independent power spectral density (PSD), and flicker noise, which has a PSD that varies as the inverse of frequency.[2] In

---

[2] In reality, flicker noise may not be exactly $1/f$, as there are a variety of effects that can produce flicker noise, including carrier recombination [196, 197]. However, traditionally flicker noise is treated as having the inverse frequency dependence.

Figure 12-20   Log-log plot of oscillator noise spectra: (a) open loop noise showing flicker noise $(1/f_m)$ and white noise regions; and (b) closed loop noise with low Q loop; and (c) closed loop noise with high Q loop.

RF amplifiers flicker noise is often of less concern than it is with oscillators, however, in CMOS amplifiers it can be appreciable and related to random variations in the crosssection of the inversion layer and random changes in mobility caused by charge trapping. In an oscillator the noise close to the carrier, or oscillation frequency, affects system performance and flicker noise is significant. Noise close to the oscillation center frequency, called the carrier (tens of hertz to hundreds of kilohertz away for an RF and microwave oscillators), manifests itself as a random variation of the phase of the carrier. Thus close-in noise is called phase noise. Phase noise on the oscillator limits the ability to demodulate signals and is one of the main reasons for a separation of the frequency of a local oscillator of a mixer from the frequency of the RF signal.

Without feedback, the noise close to the carrier is dominated by flicker noise, as shown in Figure 12-20(a) for an oscillator with the feedback loop open. This describes the intrinsic noise property of the active device (and surrounding circuitry) and the white noise and flicker noise responses are

clearly seen. When the loop is closed, the loop transfer characteristic shapes the noise response, producing noise that has regions close to the carrier that has a $1/f_m^3$ shape, and further from the carrier it varies as $1/f_m^2$ if the Q of the loop is low [247]. The switch from $1/f_m^3$ to $1/f_m^2$ dependence is at what is called the transition frequency, f_c (see Figure 12-20(b)). If the Q of the loop is high, the transition from the $1/f_m^3$ regime will be directly to a $1/f_m$ regime, and again the transition frequency is f_c, see Figure 12-20(c). The transition frequency is usually around a few kilohertz for microwave transistors. The feedback analysis that describes how $1/f$ and white noise are converted to the $1/f^3$ and $1/f^2$ or $1/f$ characteristics was developed by Leeson in 1966 [247]. Microwave and RF oscillators have low loop Q and the region of interest is the $1/f^2$ segment, as this affects the performance of current RF, radio, and radar systems.

Phase noise is defined as the ratio of the phase noise power in a 1 Hz bandwidth of a single sideband (SSB) to the total single power. This is measured at a frequency f_m offset from the carrier and denoted $\mathcal{L}(f_m)$. with the units of dBc/Hz (i.e., decibels relative to the carrier per hertz).

The phase noise performance of RF and microwave oscillators (having relatively low Q) are dominated by the $1/f^2$ shape of the phase noise. This enables comparison of different oscillators. The phase noise at 1 MHz is related to the phase noise measured at f_m by

$$\mathcal{L}(1\,\text{MHz}) = \mathcal{L}(f_m) - 10\log\left(\frac{1\,\text{MHz}}{f_m}\right)^2. \qquad (12.16)$$

Another commonly used quantitative assessment of oscillator performance is provided by the oscillator figure of merit, FOM_1, which accounts for DC power consumed [248]:

$$\text{FOM}_1 = \mathcal{L}(f_m) - 10\log\left(\frac{1\,\text{MHz}}{f_m}\right)^2 + 10\log\left(\frac{P_{\text{DC}}}{P_{\text{ref}}}\right), \qquad (12.17)$$

where P_{ref} is conventionally taken as 1 MAW and FOM_1 is referenced to the phase noise at 1 MHz. For silicon monolithic oscillators it is conventional to use just the power drawn by the oscillator core while for other technologies, including hybrid oscillators, it is not possible to separate out the oscillator core.

Most RF and microwave oscillators are VCOs). It is one of the most competitive aspects of RF design, as every decibel reduction in phase noise increases overall system performance. A high-performance VCO also relaxes demands on other system components. While FOM_1 serves as a useful metric to compare VCOs, another FOM provides bandwidth weighting is

$$\text{FOM}_2 = \mathcal{L}(f_m) - 10\log\left(\frac{1\,\text{MHz}}{f_m}\right)^2 - 10\log\left(\frac{f_{\text{BW}}}{f_{\text{ref}}}\right) \qquad (12.18)$$

Table 12-4 Comparison of RF VCOs. Phase noise is worst case over tuning range; RF output power is the minimum. All oscillators are hybrids unless indicated by IC. If f_m is not 1 MHz then a $1/f^2$ dependence is assumed for the phase noise to calculate the phase noise at 1 MHz. VCO is a quadrature VCO producing two outputs $90°$ apart. After Victor and Steer [249] with correction for FOM$_1$.

f_0 GHz	f_{BW} MHz	P_{RF} dBm	P_{DC} mW	f_m MHz	$\mathcal{L}(f_m)$ dBm /Hz	$\mathcal{L}$(1 MHz) dBm /Hz	FOM$_1$ dBm /Hz	FOM$_2$ dBm /Hz	Reference
4.92	770	0	150	1	−128		−106	−157	SiGe HBT hybrid [249]
5.05	500	0	150	1	−130		−106	−155	SiGe HBT hybrid [249]
5.16	229	−0.43	24	1	−111		−98	−136	InGaP/GaAs HBT [250]
11.5	550	9		0.1	−91	−111		−138	GaAs MESFET [251]
9.33	440	3.3	30.5	1	−102		−87	−128	GaN HEMT [252]
6.40	150	5.5	173	0.1	−105	−105	−85	−127	SiGe HBT [253]
5.94	166	−4.0	8.1	1	−110		−94	−134	CMOS IC [254]
4.87	70	−4.0	4.8	1	−131		−124	−149	GaInP/GaAs HBT [255]
5.38	120	−4.0	12.8	1	−127		−108	−148	GaInP/GaAs HBT [256]
5.29	270	−5.5	14	1	−106		−94	−130	SiGe HBT [257]
2.17	385	11.2	1.9	0.6	−120	−125	−122	−150	CMOS IC [258]
1.72	262	−11.5	75	1	−129		−111	−153	InGaP/GaAs HBT [259]
4.80	1200	4.8	36	1	−111		−95	−141	SiGe BiCMOS IC [260]
9.35	2500	18.3	570	1	−110		−82	−144	GaN/SiC pHEMT [261]
1.72	261	−10.3	55	1	−120		−103	−144	InGaP/GaAs HBT [262]
4.17	70	−6.1	102	1	−116		−96	−134	GaInP/GaAs HBT [263]
2.09	360		20.8	3	−140	−130	−117		CMOS VCO [264]
1.53	330		21.2	0.6	−133.5	−138	−125		CMOS VCO [265]
4.89	650		22	1	−124		−111		CMOS VCO [266]
1.85	280		20	3	−143	−133	−120		CMOS VCO [266]

where f_{BW} is the tuning bandwidth, and f_{ref} is the reference bandwidth taken here as 1 MHz. Again the phase noise is referenced to 1 MHz. A number of high-performance RF and microwave oscillators are compared in Table 12-4. The best phase noise that can typically be achieved by VCOs operating in the 1 to 10 GHz range is −130 dBc/Hz at 1 MHz. This compares to white noise at standard temperature which, is −174 dBc/Hz (this can be calculated from the thermal noise, kTB, of a resistor and dividing by the bandwidth, B) held at standard temperature. The performance of oscillators in the 1 to 20 GHz range designed as on-chip oscillators is surveyed in Kinget [248].

12.5.3 One-Port Oscillators

Design of stable negative-resistance oscillators traditionally uses the one-port oscillator stability requirement outlined by Kurokawa [267]. Most two-port oscillators can also be designed by casting them in the form of a one-port oscillator. In applying the criterion, each of the networks—the active device, the resonator load, and the device termination—are characterized

Figure 12-21 One-port oscillator operation: (a) as the amplitude of the oscillation increases the magnitude of the device conductance, $|G_d|$ decreases while the conductance of the tank circuit, G, is constant; and (b) as the frequency of the oscillation increases the susceptance of the tank circuit, B, changes while, B_d (ideally) does not change.

as one-ports. When a device with admittance $G_d - jB_d$ is connected to a loaded resonator of admittance $G + jB$, the voltage amplitude A and radian frequency ω of the resulting equilibrium oscillation are determined when $-G_d(A) = G(\omega)$ and $B_d(A) = B(\omega)$. In this procedure, the assumption is that the device admittance at a single frequency is a strong function of voltage amplitude, while the resonator admittance is a function only of angular frequency. This condition can be represented graphically by first denoting the locus of the negative of the device's complex admittance as $-Y_d(A) = -[G_d(A) - jB_d(A)]$ (also referred to as the inverse device reflection coefficient, or $1/S$ locus) and the locus of the resonator admittance as $Y(\omega) = -G(\omega) + jB(\omega)$. Then, for stable single-frequency oscillation, the intersection of these loci occurs at a single point. In most oscillator design the aim is to make the device admittance independent of frequency, and, of course, the admittance of the linear tank circuit is independent of the amplitude of the oscillating signal. This oscillation condition is depicted in Figure 12-21. In Figure 12-21(a), as the amplitude of the oscillation increases the magnitude of the device conductance, $|G_d|$, decreases while the conductance of the tank circuit, G, is constant. In Figure 12-21(b), as the frequency of the oscillation increases, the susceptance of the tank circuit, B, changes while B_d (ideally) does not change. The intersections here define the amplitude and frequency of the oscillation.

  If the device admittance is dependent on frequency, then it is difficult to avoid multiple intersections of the $-Y_d$ and Y as viewed in the complex plane. The angle of intersection of the Y and Y_d loci is an important indicator of stability leading to multiple oscillations, oscillator startup problems, and excess noise [268]. Thus the appropriate angular intersections of these loci are critical. It is difficult to achieve all of these objectives in design unless Y_d

is frequency independent.

Resonator design requires that Q be maximized while achieving the desired admittance change with tuning voltage. Furthermore, for a VCO, voltage tuning of the resonator must satisfy the specific stability criteria, including single point of intersection and appropriate angle of intersection, over the tuning range. With emphasis on these characteristics and the presence of parasitic elements, a proper stable resonator-device interface is troublesome. An alternative and equally viable approach to stability analysis of a broad class of oscillators, particularly for those using three-terminal devices, is application of the two-port criteria developed for amplifier stability assessment. However, the one-port approach is preferred by designers because the one-port connection is closer to the intended operation. The one-port assessment of oscillator stability is not unlike the Bode criteria applied to two-port feedback systems [269, 270]. However, unlike the two-port open-loop assessment of stability, the one-port characterization technique is conveniently aligned with measurements made by a VNA [271, 272]. As well, the nonlinear limiting effect of the active device is readily measured.

12.6 Design of a Voltage-Controlled Oscillator

12.6.1 Introduction

This section presents a detail design of a high-performance microwave VCO first reported in Victor and Steer [249]. In this approach, the design objective is the generation of a frequency-dependent negative conductance, $G_d(A, \omega)$, with a prescribed reflection coefficient shape, Γ_d, using a three-terminal active device in a common-base (series-inductive feedback) configuration. Reactive loading modifies the effective device conductance so that it becomes frequency dependent. Alternatively, modifications can be incorporated in the resonator load, but then it is seen that the frequency-dependent behavior of the tank circuit is inappropriate, resulting in multiple oscillations and other instabilities. In addition the resonator Q is compromised. Small-signal S parameters are generally good indicators of oscillator operation, particularly for the frequency of oscillation [273], however, they do not provide sufficient information to determine if stable oscillation will occur.

There are several aspects to this design that should be noted:

1. Circuit modifications are introduced that facilitate design for correct operation of the active device-resonator combination. The technique uses measured reflection coefficients and compensates for the effect of parasitics at the interface between the active device and the resonator.

2. A device mapping technique is used to modify the active device characteristic. The mapping is achieved using a combination of

additional capacitive reactive loading at the emitter-base terminals
and at the collector. The net result is an effective active device
characteristic that is largely a function of signal amplitude, while the
frequency-dependent characteristics are properly modified.

So the central idea in the design is to develop a frequency-independent
effective device admittance. As a result, the conventional one-port approach
to oscillator design can be used. Design is both an art and a science.
Sometimes the problem must be simplified for the designer to be able to
conceptualize and synthesize the required circuit.

12.6.2 Series Feedback Oscillators

The oscillator circuit used here is a common-base HBT oscillator similar
to the common-gate oscillator presented in Figure 12-18(a). A negative
resistance oscillator is normally realized using a series capacitor in the
emitter (for Z_1) and a negative conductance oscillator is realized using a
series inductor in the base lead. Both oscillator types use feedback to obtain
a negative real component. In Gonzalez and Sosa [274] the value of series
feedback reactance required is found in terms of device impedances, and
in general this can be extended for all passive terminations and applied
to any terminal of the active device. An interesting observation for both
configurations is that the resulting reflection coefficient is optimum over
a restricted region of the Smith chart. Here optimum is in the sense
that the resulting real part of the resonator series resistance (or shunt
conductance) for a series-tuned (or shunt-tuned) resonator is minimized
(maximized) to meet the criteria for oscillator startup. This criteria simply
stated, is that $|G_d(A, \omega)| > G(\omega)$ or $|R_d(A, \omega)| > R(\omega)$. Compliance with
these requirements requires that the complex reflection coefficient of the
active device, Γ_d, be greater than unity. Furthermore, there is a specific
angular range of active device reflection coefficient that is found to assist
in providing these conditions. However, it is not sufficient to simply have
large values of $|\Gamma|$. The reflection coefficient angle must be constrained
to minimize the losses associated with the resonator, at least to ensure
oscillator startup. Thus a specific angular range of active device reflection
coefficients is found to provide these conditions. Figure 12-22 plots the
equivalent parallel resistance, R_p, of the resonator as a function of the
reflection coefficient angle, $\angle\Gamma$, for several values of $|\Gamma|$. Also shown in
Figure 12-22 is the equivalent oscillator Q expressed as $|B_d/G_d|$ for $|\Gamma| = 2$.
Returning to the R_p curves, it can be seen that the point where families
of Γ values converge for a reasonable range of device $\angle\Gamma$ is approximately
$140°$. Reflection coefficient angles that are less would require resonators with
higher unloaded Q, Q_{UL}, in order to satisfy oscillator starting conditions.
If a large tuning range is required then reflection coefficient angles greater
than $100°$ are desired. Thus it is clear that the design of the tank circuit

Figure 12-22 The resistance, R_p, of a parallel (or shunt-tuned) resonator required to satisfy the condition of oscillation for (a) $|\Gamma_L| = 1.4$, (b) $|\Gamma_L| = 2$, and (c) $|\Gamma_L| = 4$ versus the reflection coefficient angle. Curve (d) is the oscillator equivalent Q for $|\Gamma| = 2$.

(or resonator) and the active device interface is a methodical process to provide appropriate admittance (or impedance) variation over the tuning bandwidth of the VCO. It is not possible therefore to simply embed parasitics in the tank circuit and design an oscillator with the required attributes.

The common base configuration used here is shown in Figure 12-23. The resonator, to the left of (x-x) in Figure 12-23, uses a tapped transmission line to improve the loaded Q and series back-to-back varactors to increase the AC breakdown voltage [275] and the unloaded Q. The series feedback inductance, to the right of the (x-x) interface, includes device mounting pads, printed board traces, and film inductors. The capacitors at (a) and (b) in Figure 12-23 are auxiliary compensating capacitors whose selection and function will now be described.

The values of the capacitors at (a) and (b) are derived using an iterative approach that involves finding the complex load required to obtain the necessary frequency and amplitude dependence of Γ_d, the reflection coefficient required of the device network presented at the (x-x) cut. Referring to Figure 12-23, the development begins by assigning $\hat{S}$ to the S-parameter matrix of the active circuit to the right of the (x-x) line. This S-parameter network comprises the small signal parameters of the transistor modified by the addition of series feedback and normalized to 50 Ω source and load terminations. The network at this point does not include the effects of a complex load Γ_L and does not include the capacitors at (a) and (b). Then

Figure 12-23 Common base oscillator configuration: (a) oscillator showing partition (x-x) between tank circuit (on the left) and the device; (b) schematic of circuit representing active device; and (c) schematic of circuit representing tank circuit. Capacitors C_a and C_b compensate for parasitic inductances (not shown). The choke inductor, L_{CHOKE}, presents an RF open circuit and is part of the bias circuit.

the device input reflection looking to the right of the (x-x) cut is modified to

$$S'_{11} = \hat{S}_{11} + \frac{\hat{S}_{12}\hat{S}_{21}\Gamma_L}{1 - \hat{S}_{22}\Gamma_L}. \qquad (12.19)$$

S'_{11} is similar to the reflection coefficient used in the oscillator design

Figure 12-24   Resonator, left of cutaway line (x-x), is separated from the active network to the right.

Figure 12-25   Negative conductance network loaded with 50 Ω termination. Active device and series feedback are centered on card. A 35 ps delay is required to reference measurement at the circuit card edge. Scale is 5:1.

approach of Gonzalez and Sosa [274] except that the terminations are not restricted to 50 Ω. This enables the loci of S'_{11} to be conveniently plotted, as the values of the capacitors at location (a) and (b) vary. The effects of the capacitors are incorporated in Γ_L and the source termination. The reflection locus curve that has the proper dependence on amplitude and frequency sets the values of the capacitors and then S'_{11} becomes Γ_d, the device reflection coefficient with the required attributes.

   As previously discussed, the oscillator design process uses Figure 12-22 as a guide in selecting the required input reflection coefficient of the active circuit. The next step in design is using a VNA to measure the reflection coefficient of the active device network to the right of (x-x) in Figures 12-23 and 12-24, and shown again in its measurement configuration in Figure 12-25. Measurement of the active device is through the tapered tap line and includes the emitter return resistor and bypass capacitor. Here it is imperative that the correct reference plane be established. Use of SOL (short open load) calibration permits the reference plane to be set correctly for a

Figure 12-26 Small-signal measurement of the active device on a compressed Smith chart: (a) proper calibration and delay; and (b) delay not applied. Marker 1: $133 - j155\ \Omega$ (uncalibrated) at 3.8 GHz; Marker 2: $55.2 - j64.9\ \Omega$ at 4.4 GHz; Marker 3: $14.1 - j47.1\ \Omega$ at 4.8 GHz; Marker 4: $20.95 - j29.8\ \Omega$ at 5.4 GHz.

3.5 mm SMA connector. The connector center pin is located right of the center cut line (x-x), as shown in Figure 12-24, requiring that 35 ps of additional delay be incorporated in calibration. This delay accounts for the offset location of the SMA open and the length of the connector center pin. The small value of corrective delay is significant, as it represents a major shift in the reflection coefficient phase required of the resonator. The resulting inverse reflection coefficient, or $1/S$, locus of the active device is curve (a) in Figure 12-26, thus requiring a capacitive resonator load.

The resonator or tank circuit is shown to the left of the (x-x) line in the oscillator schematic of Figures 12-23 and 12-24. Measurement of the tank circuit, using a similar procedure to that described above for the active device, yielded the resonator locus shown in Figure 12-27. Again, SOL calibration and correct delay adjustment is required. The resonator locus is seen to have significant parasitic series inductance that is attributed to the varactor interconnections as well as the pads. It is this parasitic that prevents straight-forward VCO design.

Design proceeds by matching the characteristics of the active device (Figure 12-26) and the tank circuit (Figure 12-27). First, the small signal device $1/S$ should provide

$$|\Gamma_d(A,\omega)| < |\Gamma(\omega)| \qquad (12.20)$$

Figure 12-27 Resonator locus on a compressed Smith chart showing that the resonator is dominantly inductive over the voltage tuning range. Varactor voltage increases in the direction of the arrow, with increasing frequency marked from (a) at 4.5 GHz to (b) at 5.3 GHz.

of the resonator for all tuned frequencies. Second, the rotation of $B(\omega)$ should be positive and in the opposite direction of the $\Gamma_d(A, \omega)$ locus. As device self-limiting occurs with an increase in drive signal to the active device, the argument of $\Gamma_d(A, \omega)$ and $\Gamma(\omega)$ should sum to zero degrees. This should be unique for each tuning voltage, and thus oscillation frequency. Finally, the trajectory of the limiting $\Gamma_d(A, \omega)$ locus should intersect the $\Gamma(\omega)$ locus at right angles to minimize phase noise [267, 272]. In this work these requirements are referred to as a complement relationship between the active device and the resonator reflection coefficient locus.

Curve (a) in Figure 12-26 and the resonator locus in Figure 12-27 illustrate the problem in achieving the single-frequency stable oscillation condition at all tuning voltages. That is, as limiting occurs, the trajectory of the negative conductance of the device intersects the resonator locus at multiple points, particularly around 5 GHz (marker points 3 and 4 on curve (a) in Figure 12-26). These conditions lead to multioscillation. A technique for addressing this problem is presented in the next section.

Figure 12-28 Modified active device reciprocal reflection $1/S$ curve which rotates counterclockwise as the device limits. The incident power measurement is at +10 dBm. Marker 1: at 3.5 GHz; Marker 2: at 4.5 GHz; Marker 3: at 4.8 GHz; and Marker 4: at 5.3 GHz.

12.6.3 Reflection Coefficient Shaping

This section presents a technique for modifying the active device network that enables straight-forward design of a single-frequency, stable, wideband, VCO. Previously it was pointed out that the input reflection coefficient of the active device network can be represented as a mapping of $1/S$ of the active device as a function of collector termination. Next, additional device modification is used to modify the map. The input termination is next added to the device. The objective here is to find the appropriate terminations at the collector and emitter terminals of the active device for a given series feedback impedance. The corresponding $-G_d(A)$ that results must yield a locus, $-G_d(A)$ versus frequency, that is, $\Gamma_d(A, \omega)$, that provides the proper interface to the resonator. If possible, the network modification should position the reflection coefficient of the modified active device in the region of the chart in Figure 12-22 above $100°$. The trajectory of the negative conductance as device limiting occurs and where $1/S$ just intersects the unit circle, must complement the argument of the resonator. This is the situation shown in Figure 12-28 where the new modified device characteristic was achieved by adding capacitive terminations to the collector and the emitter base terminals. Here, unlike the conventional common base series feedback oscillator situation, the input of the active device network is now capacitive

Figure 12-29 Multioscillation at 5.1 GHz prior to reflection coefficient shaping. Resolution BW: 3 MHz, Video BW: 1 MHz.

(see Figure 12-28). Consequently a portion of the parasitic inductance of the resonator is successfully absorbed. Thus the small signal $1/S$ one-port reflection coefficient is inductive initially.

Normally, with a common base oscillator, limiting at increasing power levels results in the device's $1/S$ locus moving along lines of constant susceptance as the negative conductance of the active device decreases. Instead, with the modified network here, there is a counterclockwise rotation of the active network's inverse reflection coefficient as limiting occurs.

The discussion can now return to the oscillation condition as determined by matching the resonator locus in Figure 12-27 to the modified active device characteristic shown in Figure 12-28. Oscillation occurs when a point on the resonator locus in Figure 12-27 corresponds to the point of the same frequency on the modified-device locus in Figure 12-28. Under small-signal conditions, the loci may not coincide, but the important point is that they do when limiting occurs, as well as providing for the startup of oscillation. The counterclockwise rotation of the modified active device locus, as described above, ensures stable, single-frequency oscillation. In particular, oscillation over the frequency range from 4.5 to 5.3 GHz follows the trajectory from point (a) to point (b) in Figure 12-27. Multioscillation as demonstrated in Figure 12-29 is suppressed in this technique. Note that in effect the resonator is operated as a shunt tunable inductance as opposed to a tunable capacitive reactance. Here is a case where the use of two-port small-signal S parameters to manage the resonator design would not be appropriate, providing little useful design insight.

Figure 12-30 Tuning curve showing oscillation frequency and VCO sensitivity as a function of tuning voltage.

12.6.4 Oscillator Performance

The oscillator design procedure outlined above was followed in implementing a VCO operating from 4.5 to 5.5 GHz using a SiGe HBT. The oscillator schematic is shown in Figure 12-23 and includes the active device modified by additional capacitors, (a) and (b). Device characterization and circuit operation was at 5 V and 30 mA bias current. In characterizing the oscillator, the varactor tuning voltage was verified against the desired frequency range by comparing the resonator locus with the $1/S$ sweep of the active device. Open-loop one-port measurements were done with +10 dBm of incident power. The resonator tuning characteristics are trimmed against those of the active device. This ultimately sets the oscillator tuning gain, K_0. Additional tuning gain adjustment is controlled by the coupling between the varactor stack and the microstrip line. Average tune gain is 120 MHz/V. The tuning performance of the oscillator is shown in Figure 12-30. The tuning characteristic is monotonic, with no jumps or discontinuities in the tuning curve as the oscillator was tuned over the full voltage tuning range. Figure 12-31 presents the fundamental output power and harmonics. The fundamental output varies by less than 2 dB over the full tuning range and the harmonic levels are relatively low.

The measured phase noise is shown in Figure 12-32 at the ends of the tuning range, 4.5 GHz (corresponding to a tuning voltage of 0 V), and 5.3 GHz (a tuning voltage of 9 V), as well as at 5.1 GHz, where the best phase noise was obtained. Phase noise was measured using a Rohde & Schwarz FSUP26 signal source analyzer and a test set loop bandwidth of 5 kHz. The phase noise is approximately the same across the tuning range with a $1/f$ noise corner frequency of 30 kHz. The phase noise at 10 kHz offset, $\mathcal{L}(f_m)$

Figure 12-31 Output power and harmonics, demonstrating low level harmonic content.

Figure 12-32 Phase noise measured at the top and bottom of the tuning range as well as at 5.1 GHz where phase noise is optimum. Minimum phase noise floor -116 dBc/Hz at 1 kHz offset, -160 dBc/Hz at 10 MHz offset.

(10 kHz), is better than -85 dBc/Hz, while at 1 MHz $\mathcal{L}(f_m)$ (1 MHz) is better than -130 dBc/Hz. The best measured phase noise near band center (5.1 GHz) is -135 dBc/Hz.

12.6.5   Bandwidth and RF Output

The standard oscillator design procedure matches the inverse reflection coefficient $1/S$ of the active device to the reflection coefficient of a tank circuit. Design, however, is often complicated by resonator parasitics so that the effective negative admittance of the active device satisfies the condition of oscillation at multiple frequency points. The Kurokawa oscillator condition establishes that for stable oscillation at the operating point of a negative conductance oscillator

$$\frac{\partial G_d}{\partial V_r}\frac{\partial B}{\partial \omega_r} - \frac{\partial B_d}{\partial V_r}\frac{\partial G}{\partial \omega_r} > 0, \tag{12.21}$$

where the subscript r refers to the operating point. In the standard approach to oscillator design the device susceptance is assumed to be independent of signal amplitude (i.e., $(\partial B_d/\partial V_r) = 0$) and the loaded resonator conductance to be independent of frequency (i.e. $(\partial G/\partial \omega_r)$), so that the stability condition becomes the much simpler

$$\left(\frac{\partial G_d}{\partial V_r}\right)\left(\frac{\partial B}{\partial \omega_r}\right) > 0. \tag{12.22}$$

The focus of this section was managing the third term of Equation (12.21), $(\partial B_d/\partial V_r)$, while the fourth term, $(\partial G/\partial \omega_r)$, is addressed by proper design of the resonator.

12.6.6   Summary

The VCO design here used reactive compensation elements at the device-resonator interface that resulted in the reflection coefficient of the augmented active device having the necessary frequency dependence to compensate for the nonideal resonator characteristic. That is, the technique results in both the effective negative resistance (conductance) and susceptance of the device to properly complement the frequency-dependent admittance of the resonator, including parasitics. Equally important is that the standard one-port approach to stable oscillator design can be used. The topology developed is suited to realizing stable, spurious-free wideband VCOs using three-terminal devices in common base configuration. However, the concept of an augmentation network is applicable to oscillators using three-terminal active devices.

12.7   Summary and Further Reading

As with most circuit design, the most current work is presented in journals, patents, and conference papers. Many of the nuggets are not the topics of books. However, there are several that will provide excellent development in the subjects. RF oscillator design in silicon and silicon

germanium technology can be explored using various references [109, 202–206]. These designs are characterized by using differential signal paths, feedback, and many transistors. Numerous references discuss microwave oscillator design, generally using compound semiconductors and single-ended designs [245,246,248,276,277] and microwave mixer design [192,215, 278,279]. Greater understanding of RF and microwave MEMS can be gained from Rebeiz [242], and Katehi and Peroulis [240].

12.8 Exercises

1. Consider the single-ended diode mixer in Figure 12-5(a).

    (a) Develop a symbolic expression for the voltage at the test point. The diode is modeled by $i_D = a_1 v_D + a_2 v_D^2 + a_3 v_D^3$.

    (b) What are the frequencies and amplitudes of the components of the spectrum at the test point.

2. The double-balanced ring diode mixer shown below has the special characteristic that the LO and RF tones are suppressed at the IF output port. Develop a symbolic expression for the voltage at the IF port. The diodes are matched and are modeled by $i_D = a_1 v_D + a_2 v_D^2 + a_3 v_D^3$. The LO voltage, at the LO terminal, is $v_{\mathrm{LO}} = A\cos(\omega_{\mathrm{LO}}t)$ and the RF voltage, at the RF terminal, is $v_{\mathrm{RF}} = B\cos(\omega_{\mathrm{RF}}t)$. Consider a 1:1 winding ratio. That is, the number of windings on the secondary on each side of the center tap is equal to the number of windings on the primary.

3. A two-port feedback oscillator is shown in Figure 12-15(a).

    (a) Draw the schematic of a feedback Colpitts oscillator.

    (b) Considering that the amplifier in the feedback oscillator has a gain that is in-dependent of frequency, what is the oscillation frequency if the components of the Colpitt's feedback network are $C_1 = C_2 = 2$ pF, and $L_3 = 5$ pF.

    (c) On a Smith chart draw the locus of the open loop transfer function of the Colpitt's oscillator from a frequency one-fifth of the oscillation frequency to five times the oscillation frequency. (Scale the open loop transfer funcion by treating $L(A)$ as one.)

4. The phase noise of an oscillator was measured as -120 dBm/Hz at 10 kHz. What is the normalized phase noise at 1 MHz assuming that the phase noise varies as the inverse of frequency.

5. The RF front end of a communications unit consists of a switch, then an amplifier, and then a mixer. The switch has a loss of 0.5 dB, the amplifier has a gain of 20 dB, and the mixer has a conversion gain of 3 dB. What is the overall gain of the cascade.

6. The RF front end of a communications unit consists of an amplifier followed by a mixer. The amplifier has a gain of 20 dB and a noise figure of 4 dB. The mixer has a conversion loss of 6 dB and a double-sideband noise figure of 8 dB.

    (a) Why is the double-sideband noise figure used with a mixer but not with an amplifier.

    (b) If the amplifier has sufficient bandwith to amplifier both the RF and image frequency, what is the noise figure of the cascade. Note that the overall noise figure is a single-sideband noise figure.

Synthesis

with Wael Fathelbab

13.1 Introduction

The optimum design of microwave networks is best approached using network synthesis. While direct realization of lumped-element circuits using transmission line stubs is possible, this yields a design that is accurate at just one frequency. The microwave network synthesis approach provides design insight, leading to novel topologies. These enable reasonably broad bandwidth networks to be realized that need minimum optimization in a microwave circuit simulator to account for parasitics and fine tune responses. The art of microwave design is the choice of a topology that has just the right characteristics so that the synthesis of a microwave network is not a matter of "fighting mother nature." Network synthesis decomposes a specified transfer function into its constituents (i.e., the elements of the synthesized prototype), thus enabling control of all available design parameters. Aided by the use of appropriate circuit transformations, novel topologies can be created that achieve performances close to the theoretical optimum in terms of size and bandwidth.

Realization in planar form most often results in the use of **Parallel Coupled Line** (**PCL**) transmission line sections. PCL filters [283] do not necessarily require ground connections and thus are easily fabricated in

planar form. They have been used to realize passband bandwidth ratios ranging from as small as 1.02:1 to as large as 3:1 [284]. Another advantage of PCL filters is that their resonators can be folded, thus reducing size [285, 286]. However, when a bandpass filter is realized using transmission lines, it suffers from poor upper stopband performance, typically having spurious passbands centered at odd harmonics of the center frequency of the fundamental passband. This is the result of the classic realization of networks in transmission line form using one-quarter wavelength long sections. At three times the operating frequency the transmission line section is three-quarters of a wavelength long and so is electrically identical to that at the operating frequency. It will be shown in this chapter how shorter sections of line can be used to realize networks that have improved out-of-band response, as well as making the distributed structures smaller.

Different orientations and connections of PCL sections can realize networks having particular forms of frequency responses. A key feature is that a PCL section can realize an inverter with lumped-element components in series or parallel. For example, Figure 9-20 on Page 497 presented the equivalence of a PCL section and an inverter with series capacitive elements. Figure 9-22 on Page 499 presented an equivalent circuit of a PCL section that consists of transformer-coupled transmission lines. From this equivalent circuit, relatively simple individual transmission line section models with stub loading can be developed. If these stub-loaded transmission lines are the result of synthesis, then the PCL equivalents can be used. So network synthesis becomes the art of merging a desired implementation configuration (e.g. a PCL section or sections), with mathematical synthesis beginning with initial transfer function descriptions.

Synthesis is best described through design examples as there is no formulaic process that can be followed and many design decisions must be made during synthesis itself. Being an art, network synthesis cannot be scheduled and the process improves with experience. This chapter is derived from designs described in a number of papers [125, 157, 193, 281, 282, 287–292] and book chapters [119, 280]. The broadband network design concepts in this chapter exploit the intrinsic frequency shaping properties of PCLs. A number of broadband networks are presented: miniature Marchand baluns (Section 13.2); differential networks that can be used in biasing differential amplifiers yielding broadband common-mode rejection (Section 13.3); and networks with combined balun and matching functionality (Section 13.4). The final development is presented in Section 13.5 with the design of bandpass filters with enhanced stopband performance.

13.2   Design of Miniature Planar Marchand Baluns

In this section three new classes of miniaturized Marchand balun are defined following the development in Fathelbab and Steer [125]. The

Figure 13-1 Bandpass Marchand balun: (a) PCL sections are one-quarter wavelength long at the center frequency, f_0, of the Balun passband (from Figure 7-16); and (b) with stub loading so that the PCL sections are one-quarter wavelength long at the commensurate frequency, $f_r = 2f_0$. The coupled line sections are defined by their odd-mode (Z_{0oa}, Z_{0ob}) and even-mode (Z_{0ea}, Z_{0eb}) characteristic impedances.

baluns are suited to mixed lumped-distributed planar realizations with small size resulting from transmission line resonators being one-quarter wavelength long at frequencies higher than the passband center frequency. Each class corresponds to an S-plane bandpass prototype derived from the specification of transmission zero locations. A tunable 50:100 Ω balun is realized at 1 GHz.

13.2.1 Derivation of a Bandpass Marchand Balun

This subsection demonstrates that the procedure of element extraction from a prototype results in a general bandpass balun topology. By appropriately loading the highpass balun of Figure 7-16(a), repeated here as Figure 13-1(a), with open-circuited stubs, the bandpass balun of Figure 13-1(b) results [123].

The procedure that will be used in the following is to take the parallel-coupled balun topology shown Figure 13-1 and develop a simple form that is the starting prototype for synthesis of a balun with specified mathematical characteristics. Then this prototype, and thus the final balun, is synthesized. In particular, the approach presented here closely follows that outlined in [123]. The design of a balun using this approach is presented in Section 13.2.3 on Page 756. This process of beginning with a topology and developing a simplified prototype that can be designed using mathematical techniques,

(a)

(b)

Figure 13-2 Derivation of the S-plane bandpass prototype of the balun of Figure 13-1: (a) after substitution of the network model of Figure 9-19(b) for each pair of coupled lines; and (b) after scaling up the series open-circuited stub, Z_{0c3}, by n^2 and $n_a = n_b = n$.

is central to microwave network synthesis. Developing the target topology is the result of experience and invention.

Section 9.11.2 on Page 498, and in particular Figure 9-21 on Page 498, presented a network model of a pair of coupled lines as transformer-coupled transmission lines. This leads to the transmission line model shown in Figure 13-2(a). Recall that stubs, inductors, and capacitors are related simply by Richard's transformation, see Figure 10-37 on Page 554, when the commensurate frequency is $f_r = 2f_0$, where f_0 is the center frequency of the equivalence. So in Figures 13-1(b) and 13-2, the stubs are one-quarter wavelength long at f_r and one-eighth wavelength long at f_0. The commensurate frequency does not always need to be twice f_0, but this should be indicated if it is not the default $2f_0$.

Further simplification of the prototype in Figure 13-2(a) would require that the pair of transformers in parallel with the series stub, Z_{0c3}, have identical turns ratios (i.e. $n = n_a = n_b$). This leads to the loss of one degree of freedom, since the pair of coupled lines comprising the balun must now have the same coupling coefficient. Proceeding with this assumption results in the circuit in Figure 13-2(b) with the impedance of the series stub, Z_{0c3},

Figure 13-3 Derivation of the S-plane bandpass prototype of the balun of Figure 13-1: (a) after shifting the input transformer throughout the network and absorbing it into the load resistance; and (b) final prototype.

scaled up to $Z_{0c3} n^2$. Upon further simplification, the prototype of Figure 13-3(a) is obtained after shifting the input transformer to the far right and absorbing it into the balanced load impedance. This is a redundant S-plane bandpass prototype with two zeroes at $S = 0$ (i.e., at DC), two zeroes at infinity, and a redundant quarter-wave long transmission line as an inverter. So given the impedances of the final prototype shown in Figure 13-3(b), the PCL parameters of the balun can be extracted after solving the following equations:

$$n_a = \frac{1}{K_a}, \quad n_b = \frac{1}{K_b}, \quad n = n_a = n_b \tag{13.1}$$

$$Z_{0c1} = \frac{Z_{02a}}{n^2} = Z_{0a}\sqrt{1 - K_a^2} \tag{13.2}$$

$$Z_{0c2} = \frac{Z_{02b}}{n^2} = Z_{0b}\sqrt{1 - K_b^2} \tag{13.3}$$

$$Z_{0c3} = \frac{Z_{01a} + Z_{01b}}{n^2} = (Z_{0a} + Z_{0b})\frac{K^2}{\sqrt{1 - K^2}}. \tag{13.4}$$

This provides the coupled line parameters in terms of the prototype's circuit elements as follows:

$$K = K_a = K_b = \sqrt{\frac{Z_{0c3}}{Z_{0c1} + Z_{0c2} + Z_{0c3}}} = \frac{1}{n} \tag{13.5}$$

$$Z_{0a} = \frac{Z_{0c1}}{\sqrt{1 - K^2}} \tag{13.6}$$

and

$$Z_{0b} = \frac{Z_{0c2}}{\sqrt{1 - K^2}}. \tag{13.7}$$

Once K, Z_{0a}, and Z_{0b} are determined, the even- and odd-mode impedances of each individual coupled line pair comprising the balun can be found after solving Equations (9.149) and (9.145) to give

$$Z_{0ea} = Z_{0a}\sqrt{\frac{1 + K_a}{1 - K_a}} \text{ and } Z_{0oa} = \frac{Z_a^2}{Z_{0ea}} \tag{13.8}$$

$$Z_{0eb} = Z_{0b}\sqrt{\frac{1 + K_b}{1 - K_b}} \text{ and } Z_{0ob} = \frac{Z_b^2}{Z_{0eb}}. \tag{13.9}$$

With knowledge of Z_{0ea}, Z_{0oa}, Z_{0eb}, Z_{0ob}, Z_{0a}, and Z_{0b}, the physical parameters of the coupled line can be calculated. Also, for further clarity, the coupling coefficients used in Equations (13.8) and (13.9) are, as stated earlier, of identical values and defined using Equation (9.145) as

$$K_a = \frac{Z_{0ea} - Z_{0oa}}{Z_{0ea} + Z_{0oa}} = K_b = \frac{Z_{0eb} - Z_{0ob}}{Z_{0eb} + Z_{0ob}} = K = \frac{1}{n}. \tag{13.10}$$

Now the impedances of the open-circuited stub, Z_{0c4}, and the balanced load, R_B, can be found from

$$Z_{0c4} = Z_{c4}/K^2 \tag{13.11}$$

and

$$R_B = R_{oB}/K^2. \tag{13.12}$$

Finally, without any further scaling, the impedances of the remaining stubs are found to be

$$Z_{0c1} = Z_{01}, \quad Z_{0c2} = Z_{03}, \text{ and } Z_{0c3} = Z_{02}. \tag{13.13}$$

Although the synthesis equations developed above are based on a specific prototype corresponding to a bandpass balun topology, the procedure is general and is adaptable to other classes of baluns.

In summary, the prototype shown in Figure 13-3(b) is synthesized and then Equations (13.5)–(13.13) are used to develop the parameters describing the coupled line balun shown in Figure 13-1(b).

13.2.2 S-Plane Bandpass Prototypes

The development of generations of prototypes culminating in the final implementation form uses the transformations shown in Figure 13-4. A logical step in deriving miniaturized baluns is to apply the half-angle transform [293] to the synthesized S-plane highpass prototype [123,126,294]

Figure 13-4 Stub network transformations: (a) open- and (b) short-circuited stubs after application of the half-angle transform [293]; (c) equivalence between series connection of open- and short-circuited stubs and stepped impedance transmission line [138]; (d) first Kuroda transformation of a series short-circuited stub; and (e) second Kuroda transformation of a series open-circuited stub from Section 10.11.5 on Page 555.

of Figure 7-16. This transform is shown in Figure 13-4(a) for an open-circuited stub. To elaborate on the mathematics involved in this transform,

Figure 13-5 Class A balun: (a) prototype derived by transforming the highpass prototype of Figure 7-16(b) using the half-angle transformation; (b) physical topology loaded with open-circuited stubs; and (c) miniaturized mixed lumped-distributed physical topology. (The distributed prototype has two zeroes at DC, two zeroes at infinity, and two nonredundant inverters).

consider the impedance of an open-circuited stub [293]:

$$Z_{0c} = Z_0/S = -\jmath Z_0 \cot(\theta) = \jmath (Z_o/2) \tan(\theta/2) - \jmath (Z_0/2)/\tan(\theta/2) \ ,$$
(13.14)

where $\theta = (\pi/2)(f/f_0)$. Equation (13.14) describes the series connection of short- and open-circuited stubs having characteristic impedances of $Z_0/2$ with half the original electrical length. This implies that the resulting transmission line resonators are one-quarter wavelength long at $2f_0$ (i.e. they are one-eighth wavelength long at f_0). Through a similar analytical treatment the short-circuited stub has the equivalence shown in Figure 13-4(b). Application of this transform to the highpass prototype of Figure 7-16 leads to the transformed prototype of Figure 13-5(a) with the input inverter now split into two equal sections. While the half-angle transformation does not affect the transmission characteristics of the network, it actually changes the distribution of transmission zeroes of the prototype. Now the prototype possesses two zeroes at DC, two zeroes at infinity, and two nonredundant inverters (i.e. it is now a bandpass prototype with a commensurate frequency $f_r = 2f_0$). Thus the balun equivalent of this prototype, shown in Figure 13-5(b), is reduced in size by almost a factor of 2 compared to a conventional balun design. However, the series open-

Figure 13-6 Class B balun: (a) prototype with two zeroes at DC, two zeroes at infinity, and a redundant inverter matching the single-ended source impedance; and (b) miniaturized mixed lumped-distributed physical topology.

circuited stub is difficult to realize in planar technology such as microstrip, and there is also scope for further size reduction if the open-circuited stubs are approximated by lumped capacitors. This is shown in Figure 13-5(c). This prototype is therefore identified as Class A. In an attempt to derive other prototypes suitable for realization as Marchand balun structures, different sets of transmission zeroes were tried and various prototype networks were synthesized. Application of the classical synthesis technique in the transformed variable [295, 296] for an arbitrarily selected f_0, and f_q with specific sets of transmission zeroes results in the prototype Classes B and C shown in Figures 13-6 and 13-7, respectively. In fact, Class A prototypes can also be directly synthesized using the same technique [295, 296].

In subsequent design steps the synthesized network prototypes undergo several network transformations in the S-plane. This allows scaling of different elements of the prototype in order to obtain realizable even- and odd-mode impedances after using Equations (13.5)–(13.7). These circuit transformations also lead to various balun topologies and allow flexible designs to be achieved. Essential circuit transformations are shown in Figure 13-4. As an example, transformed Class B prototypes are presented in Figure 13-8. Similar transformed Class A prototypes may be obtained in the same way.

So the design strategy is to first classify the S-plane bandpass prototypes from a set of transmission zeroes, apply circuit transformations, and then convert the specific prototypes to their physical balun realizations. The last step is performed by inspection in a backward manner, as was done in Figure 13-2. Subsequently the open-circuited stubs loading the balun are all approximated by lumped capacitors at the center of the passband in

Figure 13-7 Class C baluns: (a) C1 prototype with two zeroes at DC, one zero at infinity, and a nonredundant inverter (one-quarter wavelength long, $\lambda_0/4$, transmission line); (b) its corresponding miniaturized mixed lumped-distributed physical topology; (c) C2 prototype with a zero at DC, two zeroes at infinity, and a nonredundant inverter; and (d) corresponding miniaturized mixed lumped-distributed physical topology.

the f-plane to obtain mixed lumped-distributed baluns. The next section demonstrates this process using a step-by-step numerical example.

13.2.3 Design Example: Class B Balun

Descriptions of network synthesis can be involved and so here a design example is presented to illustrate the procedure. Here a 50:100 Ω Class B balun with a center frequency, f_0, of 1 GHz and a commensurate frequency, f_r, of 2 GHz is synthesized. The objective is to design a balun with a return loss greater than 15 dB at the single-ended port and a bandwidth ratio of 1.54:1, giving lower and upper passband edge frequencies of 0.82 GHz and

Figure 13-8 Transformed Class B prototypes of Figure 13-6(a): (a) after application of the stepped-impedance transmission line transformation to the series branch; (b) after application of the first Kuroda transformation; (c) after splitting the series connection of the open- and short-circuited stubs, application of the first Kuroda transformation to part of the series short-circuited stub, and subsequently transforming the rest of the series branch to a stepped impedance transmission line; (d) after splitting the series connection of open- and short-circuited stubs, application of the second Kuroda transformation to part of the series open-circuited stub, and subsequently transforming the rest of the series branch to a stepped-impedance transmission line (and then absorbing the input transformer into the load impedance); and (e)–(h) corresponding miniaturized mixed lumped-distributed physical topologies.

1.26 GHz respectively. The balun is to be implemented on an FR4 PCB with:

$$
\begin{aligned}
\text{Substrate thickness} &: \quad 62 \text{ mil } (1.57 \text{ mm}) \\
\text{Relative dielectric constant} &: \quad 4.7 \\
\text{Metal thickness} &: \quad 2.2 \text{ mil } (0.05 \text{ mm}) \\
\text{Minimum spacing} &: \quad 7 \text{ mil } (0.17 \text{ mm}) \\
\text{Loss tangent} &: \quad 0.016 \,.
\end{aligned}
\qquad (13.15)
$$

Now the characteristic polynomial, $K(S)$, of the prototype may be constructed for a passband ripple of 0.05 dB using the procedure described in Orchard and Temes [295] (or Wenzel [296]) with two zeroes at DC and two zeroes at infinity leading to[1]

$$K(S) = \left(0.7232261254S^4 + 1.868909798S^2 + 0.9448913438\right)/\left(2S^2\right),$$
(13.16)

from which the square of the magnitude of the reflection transfer function is evaluated using

$$|S_{11}(S)|^2 = \frac{|K(S)|^2}{1 + |K(S)|^2}.$$
(13.17)

$S_{11}(S$ may then be found with the knowledge of

$$|S_{11}(S)|^2 = S_{11}(S) \cdot S_{11}^*(S) = S_{11}(S) \cdot S_{11}(-S)$$
(13.18)

leading to

$$S_{11}(S) = \frac{\left(0.7232261254S^4 + 1.868909798S^2 + 0.9448913438\right)}{\left(\begin{array}{c}0.7232261254S^4 + 1.611778003S^3 + 3.664909798S^2 \\ +1.842294719S + 0.9448913438\end{array}\right)}.$$
(13.19)

The input impedance is then evaluated in a 1 Ω system from

$$Z_{\text{in}}(S) = \frac{1 + S_{11}(S)}{1 - S_{11}(S)},$$
(13.20)

which is synthesized using the standard element extraction as described in Section 10.6. Steps in the lumped-element extraction are shown in Figure 13-9. The addition of the 1 Ω transmission line has no effect on the magnitude of the reflection and transmission responses but enables subsequent Kuroda transformations that will remove the series element.

Using Richards' transformation, the prototype shown in Figure 13-10(a) is obtained where the circuit parameters come from Equations (13.2)–(13.4):

$$Z_{c1} = 50\ \Omega, \quad Z_{c2} = 44.863\ \Omega, \text{ and } Z_{c3} = 60.434\ \Omega.$$
(13.21)

Substituting these into Equations (13.5)–(13.7) gives the coupled-line parameters

$$K_a = K_b = K = 0.624, \quad Z_a = 63.974, \text{ and } Z_b = 57.401\ \Omega.$$
(13.22)

Now the parameters in Equation (13.22) are substituted in Equations (13.8)–(13.9) yielding

$$Z_{0ea} = 132.915\ \Omega, \quad Z_{0oa} = 30.792\ \Omega$$
(13.23)

---

[1]   High precision of the numerical values must be retained throughout the synthesis process.

$L_1 = 0.89726$ H, $C_1 = 1.02546$ F, $L_2 = 1.20868$ H, $C_2 = 1.38155$ F, $\omega_0 = 1$, $Z_0 = 1\ \Omega$

(a)

$L_1 = 0.89726$ H, $C_1 = 1.02546$ F, $L_2 = 1.20868$ H, $C_2 = 1.38155$ F, $\omega_0 = 1$, $Z_0 = 1\ \Omega$

(b)

$L_1 = 0.89726$ H, $C_1 = 1.02546$ F, $L_2 = 1.20868$ H, $C_2 = 1.38155$ F, $\omega_0 = 1$, $Z_0 = 1\ \Omega$

(c)

$X_1 = 44.863\ \Omega$, $X_2 = 51.273\ \Omega$, $X_3 = 60.434\ \Omega$, $X_4 = 69.0777\ \Omega$, $Z_0 = 50\ \Omega$

(c)

Figure 13-9   Steps in the extraction of the lumped-element balun network: (a) element extraction for impedance function in Equation (13.20); (b) addition of a transformer to reference the network to a 1 Ω system impedance; (c) addition of 1 Ω transmission line; (d) network scaled to a system impedance of 50 Ω.

and

$$Z_{0eb} = 119.259\ \Omega, \quad Z_{0ob} = 27.628\ \Omega. \qquad (13.24)$$

Conversion of the above set of even- and odd-mode impedances into physical dimensions yields width, w_i, and spacing, S_i, of the strips:

$$w_1 = 29.88\ \text{mil} (0.75\ \text{mm}), \quad S_1 = 3.17\ \text{mil} (0.08\ \text{mm}) \qquad (13.25)$$

Figure 13-10 Transformations for a 50:100 Ω, 1 GHz balun: (a) Class B prototype with passband ripple of 0.05 dB; (b) after equally splitting the series short-circuited stub and applying the first Kuroda transformation; (c) after transforming the series branch into a stepped impedance transmission line; (d) resynthesized and retransformed Class B prototype with passband ripple of 0.0806 dB; and (e) after manual optimization applied to the prototype of (d) to obtain a return loss of approximately 20 dB.

and

$$w_2 = 38.15 \text{ mil} \, (0.96 \text{ mm}), \quad S_2 = 2.49 \text{ mil} \, (0.06 \text{ mm}) \,. \tag{13.26}$$

Unfortunately the spacing between the coupled lines violates the minimum fabrication spacing (see Equation (13.15)). This is when the circuit transformations become useful. By transforming the prototype of Figure 13-10(a) to that of Figure 13-10(b) and making use of the stepped impedance transmission line equivalence, the transformed prototype of Figure 13-10(c) results. Performing similar manipulation to that above yields new coupled-line parameters:

$$Z_{0ea} = 157.76 \ \Omega, \quad Z_{0oa} = 47 \ \Omega \tag{13.27}$$

and

$$Z_{0eb} = 160.548 \ \Omega, \quad Z_{0ob} = 47.836 \ \Omega. \tag{13.28}$$

This results in new physical dimensions of

$$w_1 = 17.14 \text{ mil} \, (0.43 \text{ mm}), \quad S_1 = 8.79 \text{ mil} \, (0.22 \text{ mm}) \tag{13.29}$$

and

$$w_2 = 16.16 \text{ mil} \, (0.41 \text{ mm}), \quad S_2 = 9.01 \text{ mil} \, (0.22 \text{ mm}). \tag{13.30}$$

The above dimensions are realizable, but it is always desirable to obtain identical coupled-line parameters to construct a physically symmetrical balun structure. In terms of circuit synthesis it is possible to obtain a prototype with element values such that this condition is satisfied. This is done by iterating on the synthesis cycle for different values of passband ripple until the values of the resulting even- and odd-mode impedances of the pair of coupled lines become identical. Figure 13-10(d) shows another synthesized and transformed prototype achieving a passband ripple of 0.0806 dB (i.e., a return loss of approximately 17.37 dB with the same bandwidth ratio of 1.53:1). It can be seen from the figure that the impedances of the inverter and the input series stepped transmission line are virtually equal ($80 \approx 79.993 \ \Omega$).

Now for Figure 13-10(d) (referencing Equations (13.2)–(13.4) and Figure 13-3(b)),

$$Z_{c1} \cong Z_{c2} = 80 \ \Omega, \text{ and } Z_{c3} = 55.978 \ \Omega, \tag{13.31}$$

and substituting these into Equations (13.5)–(13.7) yields the coupled-line parameters:

$$K_a = K_b = K = 0.5091, \quad Z_a = Z_b = 92.946 \ \Omega. \tag{13.32}$$

Evaluating Equations (13.8)–(13.9) using Equation (13.32) gives identical coupled-line parameters of values

$$Z_{0ea} = Z_{0eb} = 162.966 \ \Omega, \quad Z_{0oa} = Z_{0ob} = 53.011 \ \Omega. \tag{13.33}$$

Figure 13-11 The physical balun layout based on the prototype in Figure 13-10 with lumped capacitors (in pF) and additional matching quarter-wavelength lines for impedance matching scaling the balanced load to 100 Ω. (The dimensions are in mils.)

At this stage a manual optimization step improves the return loss level resulting in the circuit of Figure 13-10(e) with a return loss of approximately 20 dB. However, the values of the impedances of the inverter and input series stepped transmission line must remain fixed (80 Ω) to guarantee the condition of physical symmetry. The prototype of Figure 13-10(e) results in the physical balun layout of Figure 13-11 with even- and odd-mode impedances as shown in the same figure. The following physical dimensions were determined:

$$w_1 = w_2 = 10.2 \ \mathrm{mil} \,(0.25 \ \mathrm{mm}), \quad S_1 = S_2 = 9.08 \ \mathrm{mil} \,(0.23 \ \mathrm{mm}). \quad (13.34)$$

The shunt open-circuited stub at the balanced output and the balanced load, R_B, must now be scaled accordingly using Equations (13.9) and (13.12) with K evaluated as 0.561 to give 270.51 Ω and 221.707 Ω respectively. The balanced load impedance is then scaled to 100 Ω by a pair of transmission lines of impedance 74.449 Ω each. The remaining stubs are left unscaled.

Since a bandpass prototype is being used it is possible to approximate any or all of its open-circuited stubs by lumped capacitors with very little deterioration of the circuits passband performance. Now each open-circuited stub is approximated by equating its input admittance to that of a lumped capacitor using the relationship

$$\frac{\jmath}{Z_i} \tan\left(\frac{\pi}{2} \frac{f_0}{f_r}\right) = \jmath 2\pi f_0 C_i, \qquad (13.35)$$

where Z_i is the impedance of the open-circuited stub and C_i is its corresponding capacitance value. This approximation is valid over the vicinity of the passband of the filter but will effect the upper stopband characteristics. However, in many cases such a step will lead to better stopband performance, suppressing upper undesired passbands. Performing the approximation using Equation (13.35) gives values of capacitors as shown in Figure 13-11. A final minor optimization step is required to adjust the return loss level, giving lines that are 15 mil (0.38 mm) wide, 7 mil (0.17 mm) apart, and 950 mil (24.13 mm) long. Also each output matching line is 49 mil (1.24 mm) wide and 1636 mil (41.55 mm) long.

13.2.4   Implementation of the Class B Balun

The layout in Figure 13-11 (without the high-impedance transmission lines) was implemented and its measured and simulated performance, using high quality lumped capacitors with $Q = 600$, are shown in Figures 13-12 and 13-13. In Figures 13-12(a) and 13-12(b), the measured midband insertion loss is about 3.6 dB. The excess 0.6 dB of loss is mainly attributed to substrate and conductor losses. The maximum measured amplitude error, from Figure 13-13(a), is about 0.2 dB over the passband, which is just slightly better than the simulation by about 0.05 dB. Figure 13-13(b) shows a plot of transmission phase error over the passband indicating a maximum measured phase error of approximately $2°$ at 1.2 GHz in contrast to approximately $4°$ according to simulation. It is fair to note at this point that the performance of the balun is excellent. The practical capacitor values used were 0.8 pF, 0.7 pF, and 0.6 pF as opposed to the simulated values of (from Figure 13-11) 0.75 pF, 0.675 pF, and 0.588 pF, respectively. The difference between the derived capacitance values and the actual values used, is because the parasitic via inductance was ignored in simulation, and also because of the differences of the even- and odd-mode phase velocities (and hence electrical lengths of the PCLs in even- and odd-mode).

The new baluns introduced in this section lend themselves to flexible designs since the design approach is based on bandpass prototypes with selection of the commensurate frequency, f_r. Figure 13-13(c) shows the simulated f-plane response of the circuits of Figures 13-10(a) to 13-10(e) (i.e., before the approximation of the open-circuited stubs by lumped capacitors). The first higher stopband is centered at f_r (2 GHz), while the following stopbands occur periodically at nf_r, where $n = 2, 3, 4, ...$ The second passband occurs at $2f_r - f_0$, also recurring periodically at $(2f_r - f_0) \times m$, where $m = 2, 4, 6, ...$ Also shown in Figure 13-13(c) is the measured performance of the balun with all the open-circuited stubs approximated by lumped capacitors. Now it is clear that the approximation step benefits the stopband rejection of the balun since the second passband now occurs at about 4 GHz with wider stopband performance around 3 GHz. Thus

(a)

(b)

Figure 13-12 Measured and simulated balun performance of the network in Figure 13-10(f) without bias lines and with high-Q lumped capacitors: (a) $|S_{21}|^2$ and $|S_{11}|^2$ (M1: 1 GHz, -3.6 dB); and (b) $|S_{31}|^2$ and $|S_{11}|^2$ (M2: 1 GHz, -3.5 dB).

size reduction and control of the stopband performance are simultaneously made possible through the design approach based on bandpass prototypes.

13.2.5 Summary

In this section synthesis-based design of baluns was considered. The balun prototypes were synthesized from a transfer function specification so that the functionalities of a bandpass filter and a balun were integrated into a single network. This reduces the overall system losses compared to separate realizations of a balun and a filter.

Figure 13-13 Measured and simulated balun performance using the configuration of Figure 13-10(f): (a) transmission coefficient amplitude error; (b) transmission coefficient phase error; and (c) broadband transmission coefficients.

13.3   Biasing of Differential Circuits with Broadband Common-Mode Suppression

Section 11.9 on Page 653 presented the design of a distributed biasing structure. This design is elaborated on here by combining the biasing function with **differential filtering**. In this section two classes of four-port microwave networks are presented with discriminative even- and odd-mode transmission characteristics. Design of the networks is based on filter prototypes with predefined sets of transmission zeroes. These network prototypes were specifically developed for active differential circuits. Differential circuit design leads to stable, noise-tolerant monolithically integrated analog circuits and is compatible with current-mode biasing. Symmetrical differential circuits are generally less sensitive to component variations and when appropriately designed can reject common-mode signals, particularly coupled noise. However, pseudo-differential circuits, as shown in the dotted box in Figure 13-14, typically have unity CMRR when biased by a pair of drain inductors. In this section two classes of four-port networks are introduced that are suited to biasing differential circuits and obtaining broadband common-mode signal suppression. DC biasing is achieved by appropriate selection of the set of transmission zeroes of the odd-mode subnetwork of the four-port interstage circuit.

13.3.1   Class A Four-Port Interstage Network

The Class A network is shown in Figure 13-15(a) and comprises a pair of coupled-line sections connected to a pair of matching or filtering networks. The topology of the network may consist of coupled-line sections, transmission lines, and lumped components depending on the choice of transmission zeroes required to satisfy an electrical specification. Application of modal analysis to this structure results in the subnetworks of Figure 13-15(b) with equivalent even- and odd-mode prototypes as shown in Figures 13-15(c) and 13-15(d) respectively. These equivalent prototypes result because the plane of symmetry, $X - X'$, transforms to either an open or short circuit for common- or differential-mode signal excitations, respectively, at Ports 1 and 2. In the odd-mode effective network differential-mode signals are passed to single-ended loads, R_L. In the even-mode equivalent circuit it is seen that common-mode signals do not pass through the network by virtue of the effective series open circuit present in the circuit. Thus the operation of Class A interstage networks in suppressing common-mode signals relies on the isolating property of the microwave section shown in the dashed box in Figure 13-15(c) [137, 298]. Also, Class A networks possess short-circuited points, marked **b** in Figure 13-15(a), where DC bias of the active devices can be applied with the use of decoupling capacitors.

The square of the modulus of the transmission transfer function of the

Figure 13-14 A pseudo-differential circuit commonly used in RFICs connected to a pair of single-ended loads, R_L, separated by a four-port interstage network. Drain voltage, V_{DD}, is applied at relevant points within the interstage network to bias the active devices.

reciprocal odd-mode network of Figure 13-15(d) may now be defined in terms of its characteristic polynomial, $K(S)$, as

$$|S_T(S)|^2_{\text{ODD}} = \frac{1}{1 + |K^2(S)|}, \tag{13.36}$$

where $|S_T(S)|^2_{\text{ODD}}$ replaces $|S_{31}(S)|^2_{\text{ODD}}$ or $|S_{42}(S)|^2_{\text{ODD}}$. With the even-mode network, ideally there is no transmission through the network and thus the design objective is

$$|S_T(S)|^2_{\text{EVEN}} = 0, \tag{13.37}$$

where $|S_T(S)|^2_{\text{EVEN}}$ replaces $|S_{31}(S)|^2_{\text{EVEN}}$ or $|S_{42}(S)|^2_{\text{EVEN}}$. In Equations (13.36) and (13.37) S is the Richards' variable. Now the square of the

Figure 13-15 Generalized topology of Class A interstage network made of pairs of coupled lines and matching/filtering networks: (a) physical topology; (b) even- and odd-mode subnetworks; (c) even-mode prototype; and (d) odd-mode prototype with at least one zero at DC and a single inverter.

modulus of the CMRR of the four-port interstage network may be defined as

$$|\text{CMRR}(S)|^2 = \frac{|S_T(S)|^2_{\text{ODD}}}{|S_T(S)|^2_{\text{EVEN}}}, \tag{13.38}$$

Figure 13-16 An example of Class A interstage network made of a pair of appropriately connected bandpass interdigital filters: (a) physical topology; (b) even-mode prototype; and (c) odd-mode prototype.

and consequently the resulting CMRR of the sub-system configuration consisting of the pseudo-differential circuit in cascade with the four-port interstage network (the dashed box in Figure 13-14(a)) is

$$\text{CMRR}(S) = \frac{A_d(S) \times |S_T(S)|_{\text{ODD}}}{A_c(S) \times |S_T(S)|_{\text{EVEN}}} \qquad (13.39)$$

where $A_c(S)$ and $A_d(S)$ are the single-ended common- and differential-mode gains, respectively, of the pseudo-differential circuit. For a pseudo-differential circuit, $A_c(S)$ and $A_d(S)$ are identical and thus the sub-system CMRR is dominated by the performance of the interstage network.

An example of a Class A interstage network is shown in Figure 13-16. It consists of a $2n$th order network, where n is the number of resonators in either the even- or odd-mode subnetworks. It is seen that the nth resonators of the even-mode prototype (Figure 13-16(b)), are completely decoupled from each other, thus offering optimum isolation of common-mode signals. The odd-mode prototype, shown in Figure 13-16(c), consists of cascades of short-circuited interdigital coupled-line sections [118, 137, 142, 296, 297] loaded by open-circuited stubs. Thus the odd-mode network [296] possesses

a zero at DC, n zeroes at infinity, and $(n-1)$ zeroes at $S = 1$, leading to the existence of $(n-1)$ quarter-wave long transmission lines (i.e., inverters).

The general design procedure for the interstage network starts with the synthesis of an nth-order S-plane odd-mode prototype corresponding to an f-plane response with a specific passband centered at f_0, lower- and upper-band edge frequencies, f_1 and f_2, respectively, and a commensurate frequency, f_r. A set of transmission zeroes is selected to suit a particular electrical specification with a minimum of one zero at DC, and an inverter may be chosen to generate the appropriate characteristic polynomial [294–296]. Then the square of the modulus of the reflection coefficient of the odd-mode prototype is constructed. Assuming a lossless system, the square of the modulus of the odd-mode reflection transfer function is, from Equation (13.36),

$$|S_R(S)|^2_{\text{ODD}} = 1 - |S_T(S)|^2_{\text{ODD}}, \tag{13.40}$$

which can be redefined in terms of $K(S)$ as

$$|S_R(S)|^2_{\text{ODD}} = \frac{|K(S)|^2}{1 + |K(S)|^2}. \tag{13.41}$$

$S_R(S)$ may then be found with the knowledge that

$$|S_R(S)|^2_{\text{ODD}} = S_R(S)|_{\text{ODD}} \cdot S_R(-S)|_{\text{ODD}}. \tag{13.42}$$

From this, the input impedance of the odd-mode prototype can be derived in a $1\,\Omega$ system using

$$Z_{\text{in}}(S)|_{\text{ODD}} = \frac{1 + S_R(S)|_{\text{ODD}}}{1 - S_R(S)|_{\text{ODD}}} \tag{13.43}$$

and synthesized using standard element extraction.

A possible source of CMRR degradation in Class A networks is the leakage of common-mode energy through the network. This is particularly a concern with inhomogeneous media such as microstrip or coplanar waveguide. To investigate this further, consider the inhomogeneous transfer admittance of the isolating section (the dashed box in Figure 13-15(c)) [137]:

$$Y_{12} = -\jmath \frac{\cos(\theta_e) - \cos(\theta_o)}{Z_{0e}\sin(\theta_o)\cos(\theta_e) + Z_{0e}\cos(\theta_o)\sin(\theta_e)} = Y_{21}, \tag{13.44}$$

where Z_{0e} and Z_{0o} are the even- and odd-mode impedances of the coupled lines and θ_e and θ_o are the even- and odd-mode phase lengths, respectively. It can be seen from Equation (13.44) that the numerator is zero only when the even- and odd-mode phase lengths are equal, resulting in the equivalence of Figure 13-15(c). Thus the deterioration in subsystem performance due to inhomogeneous media will be manifested in lower achievable CMRR since the transmission of the even-mode network, $|S_T(S)|^2_{\text{EVEN}}$, will have a finite value.

Figure 13-17   A quadrature hybrid coupler: (a) physical topology; (b) even-mode prototype; and (c) odd-mode prototype.

13.3.2   Class B Interstage Matching Network

Another four-port circuit with valuable characteristics is the quadrature coupler of Figure 13-17(a), with even- and odd-mode prototypes illustrated in Figures 13-17(b) and 13-17(c). The even- and odd-mode prototypes typically consist of sections of shunt one-eighth wavelength long open- and short-circuited stubs, respectively, separated by one-quarter wavelength long sections of transmission lines of uniform impedance, Z_i. Each quarter-wave transmission line section may then be split into two equal sections and the resulting $ABCD$ matrix of a cascade pair of one-eighth wavelength long transmission lines is

$$\left[\begin{array}{cc} A & B \\ C & D \end{array} \right] = \left\{ \frac{1}{\sqrt{1-S^2}} \left[\begin{array}{cc} 1 & \frac{S}{Z_i} \\ Z_i S & 1 \end{array} \right] \right\}^2 = \frac{1}{1-S^2} \left[\begin{array}{cc} 1+S^2 & \frac{2S}{Z_i} \\ 2Z_i S & 1+S^2 \end{array} \right].$$
(13.45)

From Equation (13.45) it can be seen that the resulting matrix is invariant with respect to the transformation

$$S \rightarrow 1/S$$
(13.46)

except for a change of sign [299]. This leads to a very interesting property.

The above implies that the odd-mode prototype of the coupler can be derived from its even mode, and vice versa, by simply replacing the open-circuited stubs by short-circuited stubs [299]. Now, in the S-plane, the even- and odd-mode prototypes have lowpass and highpass responses respectively. In the f-plane the lowpass and highpass responses become periodic functions with bandstop and bandpass responses, respectively, at the center frequency, f_0. Thus, over a frequency band centered at f_0, the even-mode network suppresses common-mode signals while the odd-mode prototype passes differential-mode signals. So the CMRR of this network has the relationship already defined in Equation (13.38). In this case the square of the modulus of the even-mode transmission transfer function, $|S_T(S)|^2_{\mathrm{EVEN}}$, only dips to zero at the center of the band, and thus the CMRR will have a varying characteristic with enhanced performance in the middle of the operating band.

The quadrature coupler utilized as an interstage network has a maximum bandwidth ratio (f_2/f_1:1, where f_1 and f_2 are the passband edge frequencies) of 3:1 over which there is discrimination of common- and differential-mode signals. This condition is satisfied only if the cutoff frequency, f_1, of the even-mode prototype is chosen to be

$$f_1 = f_0/2, \tag{13.47}$$

leading to an upper cutoff frequency of

$$f_2 = 3f_1. \tag{13.48}$$

In principle the quadrature coupler can be used as an interstage network but, in contrast to a Class A network, it is a lowpass structure. This structure can be modified for the purpose of biasing active devices with the inclusion of a pair of wideband bias-Ts [300] at the input arms. However, by proper selection of the relevant set of transmission zeroes a coupler-like network that supports DC bias can be designed. This leads to the Class B interstage network shown in Figure 13-18. The generalized odd-mode S-plane prototype of a Class B circuit is shown in Figure 13-18(b) and consists of a minimum of five zeroes at DC and i inverters, where $i > 5$ is an even number to maintain physical symmetry [294]. This has the benefit of providing DC bias to other parts of the system if required. It is now intuitive to conclude that the even- and odd-mode subnetworks of the coupler-like network will provide the necessary attenuation and matching of common- and differential-mode signals. This is illustrated in Figure 13-19 for a set of synthesized Class B prototypes with a 3:1 passband centered at 500 MHz for various filter degrees. It is seen that the even-mode and odd-mode responses are considerably different from each other over the frequency range of interest. The main advantage of a Class B network is simplicity in design, control of the even-mode attenuation level, and inherent broadband

Figure 13-18
Generalized topology of Class B interstage network: (a) physical topology; and (b) odd-mode prototype with five zeroes at DC and $i > 5$ inverters (the one-quarter wavelength long transmission lines).

(a)

(b)

Figure 13-19 Simulated transmission characteristics of Class B interstage network for different degrees: (a) odd-mode and (b) even-mode responses centered at 500 MHz with a 3:1 bandwidth.

operation. It is worth pointing out that the effective electrical lengths of the transmission lines forming the branches of the coupler-like section of Class B networks are one-half wavelength long, whereas the branches of the true quadrature coupler are one-quarter wavelength long. In general, Class B networks are particularly suited for systems operating at very high frequency, as electrical miniaturization of a Class B structure is not possible as is the case with Class A networks.

13.3.3 Implementation of a Class A Interstage Network

A pseudo-differential amplifier (HELA-10B) [194] was integrated with a Class A network. The amplifier has a gain of 11 dB and broadband operation from 50 MHz to 1 GHz in a 50-Ω system. The chip requires a single +12 V DC power supply. An odd-mode S-plane bandpass filter prototype was designed corresponding to an f-plane response with a center frequency,

f_0, of 500 MHz, band-edge frequencies (f_1 and f_2) of 250 MHz and 750 MHz, a commensurate frequency, f_r, of 2 GHz, and a return loss of 16 dB. The set of transmission zeroes chosen for this filter are three zeroes at DC, three zeroes at infinity, and two nonredundant inverters. Based on this the characteristic polynomial, $K(S)$, of the filter prototype is constructed using a classic procedure [295, 296]. This gives

$$K(S) = \frac{\left(2\,S\,\sqrt{1-S^2}\right)}{D_K(S)}, \qquad (13.49)$$

where

$$D_K(S) = -56.51122588\,S^4 + 0.01613466\,S^3 - 13.39620723\,S^2$$
$$+ 0.0001404614\,S - 0.629533102 \qquad (13.50)$$

from which the input reflection coefficient is evaluated using Equations (13.41) and (13.42). The input impedance, $Z_{\text{in}}(S)|_{\text{ODD}}$, is then evaluated using Equation (13.43). The synthesized prototype is shown in Figure 13-20(a) in a 50 Ω system. Construction of the four-port interstage network is now possible after transformation of parts of the odd-mode prototype to a structure made of coupled lines and lumped capacitors. Figure 13-20(b) shows the final layout.

Implementation on an FR4 PCB with a substrate thickness of 62 mil (1.57 mm), relative dielectric constant of 4.7, and loss tangent of 0.016 translate coupled-line even- and odd-mode impedances of 276.589 Ω and 48.69 Ω to a pair of PCLs that are 10 mil (0.254 mm) wide and 7 mil (0.177 mm) apart. The coupled lines are 910 mil (23.1 mm) long. The even- and odd-mode transmission characteristics of the physical layout of Figure 13-20(b) were measured and are presented in Figure 13-21(a).

In the odd mode, a bandpass response with a 3:1 bandwidth ratio is observed when the even-mode isolation is at least 22 dB. The finite even-mode isolation level is primarily due to the difference between the even- and odd-mode phase velocities of the microstrip media, leading to the transmission of common-mode signals through the isolating sections. Another source of energy leakage contributing to this finite isolation level is the parasitic coupling beyond nearest neighboring resonators leading to unwanted couplings throughout the even-mode networks of the interstage network. This is particularly true since there is weak coupling between closely spaced pairs of resonators of isolating sections giving rise to cross-coupling routes.

The overall measured characteristics of the interstage network coupled to the pseudo-differential amplifier are shown in Figure 13-21(b). Very low common-mode gain, A_c, of less than -12 dB was recorded over the operating bandwidth while a 3:1 bandwidth ratio was achieved with a differential-mode gain, A_d, of about 10 dB. The excess 1 dB of loss is

Figure 13-20 A Class A interstage network in a 50 Ω system: (a) synthesized odd-mode prototype; and (b) physical layout with capacitors in picofarads.

attributed to loss in the interstage network. Also it is observed that the differential-mode gain ripples over the measured passband and rolls off rapidly toward the upper band-edge frequency (i.e., 750 MHz). This is primarily due to the impedance level mismatches since the measured return loss of the interstage circuit is only 6 dB (see Figure 13-20(c)). The deteriorated return loss is due to the restriction imposed by the FR4 board manufacturing process that only allowed a minimum of 7 mil spacing between the coupled lines. To achieve a better return loss, for example, 12 dB, a 5 mil spacing between the lines is required.

13.3.4 Implementation of a Class B Interstage Network

A Class B interstage network was also designed on an FR4 board with the same specification to that used in the above sub-section. In this case the objective is to highlight the characteristics of the even- and odd-mode networks of a Class B circuit without integration with the pseudo-differential amplifier. The odd-mode prototype was designed for an f-plane response identical to that defined in the above sub-section but with a return loss (RL) of 36 dB. Here the commensurate frequency, f_r, of the f-plane prototype coincides with the center frequency, f_0, since the prototype

Figure 13-21 Measured characteristics of the interstage network shown in Figure 13-20: (a) measured even- and odd-mode transmission characteristics; and (b) measured common- and differential-mode gains of the pseudo-differential amplifier in cascade with the interstage network.

corresponds to an S-plane highpass network. The set of transmission zeroes of the odd-mode prototype of the coupler-like network are chosen to be five zeroes at DC and four inverters. The characteristic polynomial, $K(S)$, was directly generated using the technique described by Horton and Wenzel [294] to give

$$K\left(S\right) = \frac{\left(\begin{array}{c} -.3234410760S^{12} - 4.831961462S^{10} - 12.43391743S^{8} \\ +6.369465650S^{6} + 29.01088612S^{4} - 1.294045780S^{2} \\ -16.49698604 \end{array}\right)}{2S^{5}\left(S^{2} - 1\right)}.$$

(13.51)

In a manner similar to that used previously, $S_R(S)|_{\text{ODD}}$ and $Z_{\text{in}}(S)|_{\text{ODD}}$ are constructed using Equations (13.41)–(13.43).

Direct synthesis of the above transfer function leads to the prototype of Figure 13-22(a) in a 50 Ω system. Application of the appropriate Kuroda transformations to the circuit of Figure 13-22(a) results in the circuit of Figure 13-22(b). It is seen from Figure 13-22(b) that the characteristic impedances of the inner pair of shunt short-circuited stubs are 33.27 Ω. This pair of stubs was forced to have a value of 50 Ω by application of circuit optimization, leading to a degradation of the return loss from 36 dB to about 15 dB with a slight increase in bandwidth. The prototype is shown in Figure 13-22(c). At this point the layout of the interstage network is constructed from the odd-mode prototype. However, the input and output sections (series open-circuited stubs separated by a inverter) are difficult to realize using microstrip technology and the equivalence to coupled lines must be made, leading to the final network of Figure 13-22(d). Now the even- and odd-mode impedances of values 133.741 Ω and 41.306 Ω of each coupled-line section are translated into physical dimensions, resulting in a width of 27 mil (0.685 mm), spacing of 7 mil (0.177 mm), and length of 3483 mil (88.5 mm). The width of the other transmission lines of 50 Ω, 70.047 Ω, and 83.594 Ω are 110 mil (2.794 mm), 57 mil (1.447 mm), and 38 mil (0.965 mm), respectively. Also, the lengths of all the transmission lines were 3483 mil (88.5 mm).

The measured even- and odd-mode transmission characteristics of the interstage network are shown in Figure 13-23, showing significant difference as expected.

13.3.5   Summary

This section presented the derivation of and design procedures for two classes of four-port microwave interstage networks that block common-mode signals over a very broad frequency range. The design procedure presented for the new networks is based on the synthesis of an odd-mode prototype with broad bandwidth and out-of-band rejection. Subsequent integration of either class of these networks with pseudo-differential circuits leads to dramatic improvement of the subsystem CMRR over large bandwidth. Both classes of networks possess fundamentally different even- and odd-mode transmission characteristics and reflect undesired common-mode energy. Class B networks could be designed to posses a finite attenuation level to reject common-mode signals. This feature does not exist with Class A networks, and basically the rejection level of common-mode signals will depend on how homogeneous the realization medium is. The choice of which class to use is dependent on the system application.

Figure 13-22 A Class B interstage network: (a) odd-mode prototype: (b) after application of Kuroda transformations; (c) after circuit optimization; and (d) physical layout.

Figure 13-23   Measured even- and odd-mode transmission characteristics of the interstage network shown in Figure 13-22(d).

13.4   Tapped Marchand Baluns for Stable Matching Applications

Subsystems with active devices generally require matching networks which, as well as providing optimum transfer of power within an operating band, often determine the stability of the subsystem by their out-of-band characteristics. Essentially a matching network must extract maximum power from the source for transfer to the load, but must also present the source with a specific immittance profile both in and out of its operating band. For example, the impedance level of a matching network connected to the output of a power amplifier approximating a current source must drop to a low level in the out-of-band region to ensure device stability. These design criteria impose requirements on the matching network, necessitating that its design be based on a singly terminated [301–304] rather than a doubly terminated network [125]. In the case of a singly terminated matching network, the real part of its input immittance function becomes the design parameter and hence is tailored appropriately to suit the characteristics of the source.

   This section presents structures that combine the functionality of baluns and matching networks. The structures are derived from several filter prototypes [125] that can be realized in the form of Marchand baluns. The baluns represent a solution to the problem of matching between terminals of extreme impedance (or admittance) values without using elements of extreme characteristic impedances in the balun structures. This is by virtue of tapping the balanced resonators of the baluns, thus allowing a wide range of load-to-source immittance ratios to be realized.

   The presentation here begins in Sections 13.4.1 and 13.4.2 by highlighting the difference in design objectives between doubly and singly terminated networks. Section 13.4.4 describes a method of scaling the balanced load

of a balun to specific values. Finally, Sections 13.4.6 and 13.4.7 present the synthesis and implementation of two tapped baluns. Size miniaturization and specific load-to-source immittance ratios are achieved.

13.4.1 Doubly Terminated Matching Network

The fundamental difference between doubly and singly terminated matching networks concerns the design parameter synthesized to satisfy an electrical specification. First, the doubly terminated matching network will be considered.

A doubly terminated network attempts to achieve a constant return-loss level at its terminals over a specified passband. Referring to Figure 7-15, the design focus is the input reflection coefficient

$$\Gamma_{\text{in}} = \frac{Z_{\text{in}} - R_S}{Z_{\text{in}} + R_S} \tag{13.52}$$

for an impedance-based network, where R_S is the source impedance. The input reflection coefficient can also be expressed as

$$\Gamma_{\text{in}} = \frac{Y_S - Y_{\text{in}}}{Y_S + Y_{\text{in}}} \tag{13.53}$$

for an admittance-based network where Y_s $(= 1/R_s)$ is the source admittance. In a lossless system the transmission characteristic is

$$|S_{21}|^2 = 1 - |\Gamma_{\text{in}}|^2. \tag{13.54}$$

13.4.2 Singly Terminated Matching Network

An ideal impedance-based singly terminated network is driven by a current source and presents a constant-impedance level to its driving source over its operating band. It can be shown that the transmission coefficient of a singly terminated network is actually equal to the real part of its input impedance [303, 304]:

$$|S_{21}|^2 = \Re\{Z_{\text{in}}\}. \tag{13.55}$$

Also, for an admittance-based network driven by a voltage source, the transmission characteristic is

$$|S_{21}|^2 = \Re\{Y_{\text{in}}\} .. \tag{13.56}$$

From Equations (13.54)–(13.56) it is evident that the transfer of power from the source to the load is different in the two cases.

13.4.3 A Balun Example

To demonstrate the above points, the performance of a third-order Class A Marchand balun [125] designed from the doubly and singly terminated prototypes presented in Sections 13.4.1 and 13.4.2 is considered.

Figure 13-24 Input admittance of Class A Marchand balun normalized to a 1 Ω system: (a) real part of the input admittance of a balun designed from a doubly terminated prototype for various single-ended return-loss levels; and (b) the same balun designed utilizing a singly terminated prototype with a specific passband ripple.

Figure 13-24(a) presents plots of the real part of the balun's input admittance function at the single-ended port. Here it is shown how the characteristic of the real part changes as the return-loss level is varied from 10 dB to 20 dB. As expected, the passband ripple decreases as the return loss improves. However, it is also seen, from Figure 13-24(a), that the real part peaks at frequencies outside the passband (defined by the vertical dotted lines) as the return loss improves, resulting in an uncontrolled increase in bandwidth. Thus, in order to improve the in-band ripple of the input admittance level, the skirt selectivity must be degraded. In almost every application the skirt selectivity of a matching network is equally as important as the quality of the in-band match and has direct influence on the overall system performance.

Now the same type of balun is synthesized on a singly terminated basis

Table 13-1 Type of parameter synthesized for each class of baluns.

Prototype classes [9]	Distribution of zeroes	Type of parameter realized at single-ended port	Type of parameter realized at balanced ports
A	2 zeroes at S=0 2 zeroes at $S = \infty$ 2 zeroes at S=1	$\Re\{Y_{in}\}$	$\Re\{Z_{in}\}$
C1	2 zeroes at S=0 1 zero at $S = \infty$ 1 zero at S=1	$\Re\{Y_{in}\}$	$\Re\{Z_{in}\}$
C2	1 zero at S=0 2 zeroes at $S = \infty$ 1 zero at S=1	$\Re\{Z_{in}\}$	$\Re\{Z_{in}\}$
B	2 zeroes at S=0 2 zeroes at $S = \infty$	$\Re\{Z_{in}\}$	$\Re\{Z_{in}\}$

(see the characteristic in Figure 13-24(b)). Here it is seen that the transition from passband to stopband is much faster than that of any of the functions presented in Figure 13-24(a). In addition, there is a more controlled in-band ripple variation. The function depicted in Figure 13-24(b) is a Chebyshev which is well known for its optimum performance in terms of trading off passband ripple and skirt selectivity [303].

Thus, based on the proceeding comparison, it is possible to conclude that a singly terminated balun integrated as a matching network in a subsystem configuration will extract maximum available power from its source and will possess the desired out-of-band characteristics for stability requirements of active circuits. Table 13-1 summarizes the types of parameters that are synthesized at either port with reference to the sets of transmission zeroes used.

13.4.4 Tapping the Balanced Resonators of the Balun

This section shows that the balun of Figure 13-25(a) [125] can be designed for specific load-to-source immittance ratios. The equivalent prototype of this balun, Figure 13-25(b), can be synthesized for different commensurate frequencies, f_r. The commensurate frequency determines the electrical lengths of the transmission lines forming the balun and thus the larger the value of f_r, the smaller the size of the resulting balun structure. In addition to fixing the overall size of the balun, the choice of f_r also impacts the values of the immittances terminating the balun. In general the inherent values of the load and source immittances of a Marchand balun

Figure 13-25   A bandpass Marchand balun with balanced load impedance R_B: (a) its physical layout based on Figure 13-1; and (b) its corresponding S-plane bandpass prototype (from Figure 13-3(b)).

are functions of the passband ripple, operating bandwidth, and also f_r. An example of a synthesized balun with an octave bandwidth centered at 500 MHz is shown in Figure 13-26. It is seen that the inherent values of the single-ended and balanced impedances drop as f_r increases from 5 GHz to 20 GHz. Thus miniaturizing such a structure has an impact on its terminal impedances. It is also observed that the inherent balanced-to-single-ended impedance ratios and the modal-impedances in the two cases remain practically unaffected with different values of f_r. The focus will now shift to tapping the balanced resonators of the balun to achieve a wide range of balanced-to-single-ended immittance ratios.

   The detailed analysis of the generalized Marchand balun presented in Section 13.2 demonstrated that the balanced load is always much higher in value than the single-ended source impedance. It is common practice to scale down the balanced load impedance by using a pair of one-quarter wavelength long transmission lines (e.g., see Figure 13-11). While this is normally an acceptable technique, the pair of transmission lines drastically increased the overall size of the balun. Note that these matching lines resonated at the passband center frequency, f_0, while the balun structure comprised transmission lines that resonated at the commensurate

(a)

f_r (GHz)	Z_s (Ω)	Z_{0c1} (Ω)	Z_{0c3} (Ω)	Z_{0c4} (Ω)	$Z_{0ea} = Z_{0eb}$ (Ω)	$Z_{0oa} = Z_{0ob}$ (Ω)	R_s (Ω)	R_B (Ω)	R_B/R_s
5	47.55	1.56	3.117	4.744	164.946	58.253	6.166	59.988	9.728
10	48.028	0.385	0.787	1.184	166.842	59.421	3.127	30.567	9.775
20	48.148	0.095	0.197	0.294	167.502	59.827	1.569	15.373	9.797

(b)

f_r (GHz)	Z_s (Ω)	Z_{0c1} (Ω)	Z_{0c3} (Ω)	Z_{0c4} (Ω)	$Z_{0ea} = Z_{0eb}$ (Ω)	$Z_{0oa} = Z_{0ob}$ (Ω)	R_s (Ω)	R_B (Ω)	R_B/R_s
5	35.695	0.985	3.345	5.245	189.411	50.789	2.748	39.173	14.255
10	36.257	0.245	0.845	1.310	191.724	52.069	1.396	19.988	14.318
20	36.399	0.061	0.211	0.326	192.152	52.307	0.701	10.035	14.315

(c)

Figure 13-26 Class A balun: (a) the electrical layout with an octave bandwidth centered at 500 MHz; (b) element values for a prescribed impedance function at the balanced ports with a passband ripple of 0.5 dB; and (c) element values for a prescribed admittance function at the single-ended port with a passband ripple of 0.5 dB.

frequency, f_r, where $f_r > f_0$. This situation is undesirable in circumstances where size reduction is the primary objective. This drawback is overcome while simultaneously achieving a specific load-to-source immittance ratio by tapping the balanced resonators of the balun. It is common practice to alter the system impedance at either or both ports of a microwave filter [305] to yield a realizable circuit. Consequently, tapping the resonators at the scaled node(s) of the filter matches it to the desired system impedance with the advantage of no added hardware. This approach is used here in designing Marchand baluns.

To begin the description of the tapped balun design, consider Figure 13-27(a), which illustrates the basic output subsection of a bandpass Marchand balun. The subsection consists of a short-circuited stub, representing the balanced resonators, and an open-circuited stub connected to the balanced load, R_B. To scale the load impedance from R_B to R_D ($R_D < R_B$) an

Figure 13-27 Balanced resonator tapping: (a) typical output section of a bandpass balun prototype after scaling its balanced load using an ideal transformer; (b) identification of the virtual ground plane; and (c) and (d) balanced resonators after tapping.

ideal 1:X transformer $\left(X = \sqrt{R_D/R_B}\right)$ is inserted at the output, as shown in Figure 13-27(a). If the balun is designed such that the virtual ground plane bisects the subcircuit in the center, then it is possible to exactly represent the subcircuit of Figure 13-27(a) by that of Figure 13-27(b). The subcircuit of Figure 13-27(b) (excluding the open-circuited stub) can then be approximated by that of Figure 13-27(c). Examination of the subcircuit of Figure 13-27(c) reveals that the total length of the balanced resonators is still equal to l_r, but now the short-circuited stubs of length l_1 represent the parts of each resonator seen by each half of the scaled balanced load to the virtual ground. Also the existence of the lumped capacitor C_0 is required for fine-tuning purposes.

For a balun realizing a prescribed immittance function at the single-ended port, the following approximation is valid for passband bandwidths up to an octave. At the center frequency, f_0, the impedance, Z_T, in Figure 13-27(b) is

$$Z_T = \Re_T + \jmath \Im_T, \tag{13.57}$$

where

$$\Re_T = 2R_D^2 \left[R_B \left(\frac{4R_D^2}{Z_{\text{sc}}^2 \left(\tan\left(\beta l_r\right)\right)^2} + \frac{4R_D^2}{R_B^2} \right) \right]^{-1} \tag{13.58}$$

and

$$\Im_T = 2R_D^2 \left[Z_{\text{sc}} \left(\tan\left(\beta l_r\right)\right) \left(\frac{4R_D^2}{Z_{\text{sc}}^2 \left(\tan\left(\beta l_r\right)\right)^2} + \frac{4R_D^2}{R_B^2} \right) \right]^{-1}. \tag{13.59}$$

Now the ABCD matrix of half of the circuit of Figure 13-27(c) (without including the load impedance) can be derived based on the following matrix manipulations:

$$\begin{aligned}
\begin{bmatrix} A & B \\ C & D \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ \jmath(2\pi f)C_0 & 1 \end{bmatrix} \\
&\times \begin{bmatrix} \cos\left(\beta\left(l_r - l_1\right)\right) & \jmath\left(\frac{Z_{sc}}{2}\right)\sin\left(\beta\left(l_r - l_1\right)\right) \\ \jmath\left(\frac{2}{Z_{sc}}\right)\sin\left(\beta\left(l_r - l_1\right)\right) & \cos\left(\beta\left(l_r - l_1\right)\right) \end{bmatrix} \\
&\times \begin{bmatrix} 1 & 0 \\ \jmath\left(\frac{Z_{sc}}{2}\right)\tan\left(\beta l_1\right) & 1 \end{bmatrix}.
\end{aligned} \tag{13.60}$$

Using Equation (13.60), the input impedance, Z_t, is evaluated at f_0 to give

$$Z_t = \frac{A\left(R_D/2\right) + B}{C\left(R_D/2\right) + D} = \Re_t + \jmath \Im_t. \tag{13.61}$$

In Equations (13.58)–(13.60), $\beta\left(= \left(2\pi/\lambda\right) = \left(2\pi f/v_p\right)\right)$ is the propagation constant and v_p is the phase velocity in the transmission medium. The real

and imaginary parts of Equations (13.57) and (13.61) may now be equated and solved simultaneously for the two unknowns, l_1 and C_0. It should be appreciated, however, that there is a limit to how much the balanced load impedance can be lowered from R_B to R_D using the tapping approach. For example, forcing R_D to be a very low value will likely degrade the ripple level of the real part of the input immittance at the single-ended port. In general, the approximation is increasingly valid as R_D approaches $R_B/2$.

The above procedure can also be used to scale the balanced load of a balun realizing a prescribed immittance function at its balanced ports. In this case, it is more appropriate to approximate the subcircuit of Figure 13-27(b) by that of Figure 13-27(d) in a similar fashion as was just described. The process of load scaling will be demonstrated as part of the examples discussed in the next section.

13.4.5 Design Example

The general form of the characteristic polynomial of an S-plane highpass network comprising m open- or short-circuited stubs, and n one-quarter wavelength transmission lines is [138, 294]

$$K(S) = \sqrt{\cos\left(m\cos^{-1}\left(\frac{S_c}{S}\right) + n\cos^{-1}\left(\sqrt{\frac{1-S_c^2}{1-S^2}}\right)\right)}, \qquad (13.62)$$

where S is the Richards' variable. The prototypes to be synthesized in the next two subsections are chosen to have a pair of zeroes at DC ($m = 2$) and a single zero at $S = 1$, so ($n = 1$). For an octave bandwidth, Equation (13.62) becomes

$$K(S) = \frac{-S^2\left(2S_c^2 + S_c^2\sqrt{1-S_c^2}\right) + \left(2\sqrt{1-S_c^2} + 2\right)}{S^3\left(S_c^3\sqrt{1-\frac{1}{S^2}}\right)}, \qquad (13.63)$$

with S_c equal to $\jmath 0.57735$. The FR4 PCB substrate utilized for the fabrication of the baluns had a thickness of 1.57 mm, relative dielectric constant of 4.7, and loss tangent of 0.016. The synthesis concept is implemented in the next two subsections for two types of baluns.

13.4.6 Balun With Prescribed Single-Ended Admittance

A Class A balun is required to have a prescribed immittance function at its single-ended port with a passband ripple, ε, of 0.508 (i.e., $10\log_{10}\left[1/\left(1+\varepsilon^2\right)\right] = 1$ dB) centered at 500 MHz. According to Table 13-1, the Class A balun implements an admittance function at the single-ended port. The balun is also required to transform a single-ended source impedance of 7 Ω to a balanced load of 100 Ω. A typical situation that

requires this type of balun is in the matching stage between the output of a power amplifier having low output impedance and a balanced circuit such as an antenna.

Using Equation (13.63), the real part of the input admittance is derived from

$$\Re\left\{Y_{\text{in}}(S)\right\} = \frac{1}{1 + \varepsilon^2 \left|k(S)\right|^2} = \left|S_{21}(S)\right|^2 \ , . \qquad (13.64)$$

Partial fraction expansion of Equation (13.64) is then applied after which the input admittance is extracted with the knowledge of the relationship [303, 304]

$$\Re\left\{Y_{\text{in}}(S)\right\} = 0.5 \times \left\{Y_{\text{in}}(S) + Y_{\text{in}}(-S)\right\}. \qquad (13.65)$$

This gives the input admittance,

$$
Y_{\text{in}}(S) \ = \ \frac{(0.1662867861 + j0.3615083714)}{(S + 0.4719034609 - j1.681902406)} +
$$
$$
\frac{(0.1662867847 - j0.361508366)}{(S + 0.4719034609 + j1.681902406)} - \frac{2.438018122}{(S + 3.734004591)} + 1.
$$
$$(13.66)$$

Note that all of the poles of Equation (13.66) are located on the left-hand side of the S-plane to ensure that the input impedance is positive real. The above function is then synthesized to obtain an S-plane singly terminated prototype with element values as shown in Figure 13-28(a).

It is worth mentioning that each S-plane capacitor and inductor of Figure 13-28(a) becomes an open- and a short-circuited stub, respectively, in the f-plane with characteristic impedances of $1/(\omega_0 C)$ and $\omega_0 L$, respectively (with $\omega_0 = 1$). The system impedance of the prototype is then raised from 1 Ω to 7 Ω, after which the half-angle and relevant Kuroda transformations (see Figures 13-4(a) and 13-4(b)) are applied, resulting in a transformed bandpass prototype. This leads to the f-plane prototype of Figure 13-28(b). The element values of the circuit of Figure 13-25(b) may now be obtained from the circuit of Figure 13-28(b). They are as follows:

$$
\begin{aligned}
&Z_{01} = 28.023 \ \Omega, \ Z_{c1} = 77.777 \ \Omega, \ Z_{02} = 0, \ Z_{c2} = 78.287 \ \Omega, \\
&Z_{03} = 502.364 \ \Omega, \ Z_{c3} = 70.025 \ \Omega = Z_{c4}, \ R_{oB} = 60.669 \ \Omega.
\end{aligned} \qquad (13.67)
$$

Substituting Equation (13.67) into Equations (13.5)–(13.7), the parameters of the balun of Figure 13-28(c) can be derived using Equations (13.8)–(13.13). Figure 13-28(c) depicts a load impedance of 100 Ω after the inclusion of a 1:0.714 transformer. This implies that the balun has an inherent balanced-to-single-ended impedance ratio of

$$\frac{R_D}{R_S} = \frac{195.879}{7} = 27.98 \qquad (13.68)$$

Figure 13-28 Class A balun with prescribed admittance function at the single-ended port: (a) S-plane singly terminated prototype; (b) admittance-scaled and circuit-transformed f-plane prototype; (c) layout; and (d) layout with tapped resonators.

Figure 13-29   Class A balun fabricated from the layout shown in Figure 13-28(d).

From Figure 13-28(c) it can be seen that the pairs of coupled-lines of the balun possess identical modal impedances. This is a result of the appropriate sequence of circuit transformations applied to the initial prototype of Figure 13-28(a). Due to this, the balanced load and the output shunt stubs are symmetrically bisected by a virtual ground plane (in the center). Thus the resonator tapping approach, discussed in Section 13.4.4, can be implemented. For this balun the elements of the output subsection (corresponding to Figure 13-27(a)) are

$$Z_{\text{oc}} = 225.579\,\Omega, \ Z_{\text{sc}} = 225.579\,\Omega, \ R_B = 195.879\,\Omega,$$
$$R_D = 100\,\Omega, \ l_r = 75\,\text{mm}, \ X = 0.714, \tag{13.69}$$

which, upon solving Equations (13.57) and (13.61), yields the tapping height and the value of the capacitor as follows:

$$l_1 = 55.787\,\text{mm}, \ C_0 = 1.313\,\text{pF}. \tag{13.70}$$

At this stage the open-circuited stubs in the balun structure of Figure 13-28(c) are approximated by lumped capacitors leading to the electrical layout of Figure 13-28(d). Subsequently the electrical layout of the balun is then converted into a physical layout and fabricated as shown in Figure 13-29.

   For practical purposes, additional sections of 50 Ω lines interconnecting the SMA connectors to the actual balun ports are added, as shown in Figure 13-29. The extra transmission lines at the balanced ports have no effect on the measured response at the single-ended port, however, the delay introduced between the SMA connector and the actual single-ended port of the balun must be de-embedded in order to accurately measure the input admittance. A pair of 50 Ω loads is connected at the balanced ports and the single-ended port of the balun connected to a digital sampling oscilloscope. The measured time delay due to the transmission line section at the single-ended port is 82.5 ps.

   The measured performance of the balun is shown in Figure 13-30. Here the center frequency of the passband has been offset 6.6% (shifted down by 33 MHz) and is now centered at 467 MHz. This offset is due to the

Figure 13-30   Measured characteristics of the fabricated balun of Figure 13-28(e).

parasitic inductances of the via holes, which are not accounted for, and can be corrected by slightly shortening the transmission line resonators of the balun. The measured amplitude and phase imbalances are less than 0.2 dB and 1°, respectively.

13.4.7   Balun With Prescribed Balanced Impedance

Here a balun is synthesized with a prescribed immittance function at the balanced ports of a Marchand balun. The Class A balun of the previous subsection transformed a balanced source impedance of 100 Ω to a single-ended load of 50 Ω with a ripple, ε, of 0.349 (0.5 dB). According to Table 13-1, the structure realizes an impedance function. A typical application that requires this type of balun is the matching stage between the differential outputs of an RFIC feeding a single-ended circuit such as a filter.

The real part of the balanced impedance function was developed, as previously, using Equations (13.63)–(13.65), but on an impedance basis. The synthesized S-plane prototype is shown in Figure 13-31(a). The prototype is then scaled up in impedance so that a 50 Ω load results at the single-ended port, leading to the f-plane prototype of Figure 13-31(b). Going through identical steps as described in Section 13.4.4, the tapped balun of Figure 13-31(c) results, with its simulated differential impedance depicted in Figure 13-31(d). The fabricated balun is shown in Figure 13-32.

Balanced operation of the balun was tested by measuring the real and imaginary parts of the impedance function to ground at Port 1 of Figure 13-31(c) with Port 2 terminated to ground in a 50 Ω load. A similar measurement at Port 2 with Port 1 terminated in 50 Ω is shown in Figure 13-33.

Figure 13-31 Class A balun with prescribed impedance function at the balanced ports: (a) S-plane singly terminated prototype; (b) impedance-scaled and circuit-transformed f-plane prototype; (c) electrical layout with tapped resonators depicting the modal impedances of the coupled lines; (d) simulation of the differential impedance between Ports 1 and 2; and (e) the fabricated balun.

Figure 13-32 Photograph of the fabricated balun with the layout shown in Figure 13-31.

Figure 13-33 Measured characteristics of the balun of Figure 13-32: (a) and (b) are the real and imaginary parts of the driving point impedances to ground at each balanced port with the other terminated in a 50 Ω load.

13.4.8 Summary

Singly terminated prototypes were used to derive new immittance-transforming Marchand baluns suitable for matching purposes. This led to miniaturized lumped-distributed baluns realizing specific immittance

profiles at either their single-ended or balanced ports. The stability of active devices such as single-ended or differential amplifiers could be improved by controlling the immittance they see. This is achieved by utilizing the proposed baluns in those situations. For a specific operating bandwidth and passband ripple the commensurate frequency of the presented baluns can be varied, resulting in a wide range of terminal immittance values. Additional tapping of the balanced resonators of the baluns leads to the realization of specific balanced-to-single-ended immittance ratios at no increase in overall size. Two baluns were designed using a synthesis procedure. Measured and simulated specifications were in close agreement.

13.5 Filters With Enhanced Stopband Performances

A new class of parallel-coupled line filters with broad stopband response is introduced in this section. The design is based on the synthesis of bandpass prototypes with predefined upper stopband characteristics. The new filters have uniform- and stepped-impedance resonators, some of which are loaded by open-circuited stubs at their open-circuited ends.

Considerable efforts to improve the stopband performance of microstrip PCL filters have largely been directed at suppressing the harmonic spurious passband located at $2f_0$ (where f_0 is the passband center frequency). This has been done by compensating for the difference between the even- and odd-mode phase velocities of the inhomogeneous media. In one approach, the coupled lines were loaded by lumped capacitors [145] or meandered [146]. Other approaches have introduced ground plane apertures underneath the coupled lines [147–149] or utilized suspended substrates with dielectric overlay [150, 151]. Other attempts to suppress the undesired spurious passbands have introduced geometrical perturbations to the uniform-impedance resonators in the coupling regions [152–156], loaded them by split ring resonators [158], or utilized stepped-impedance resonators [159]. Using uniform- or stepped-impedance resonators is the most successful strategy for enhancing the stopband performance of PCL filters.

13.5.1 Uniform- and Stepped-Impedance PCL Filters

It was pointed out in Section 9.11 that filters using transmission lines have spurious passbands. To some extent, it is possible to alter the positions of the spurious passbands of conventional PCL filters by using stepped-impedance resonators. An excellent account of this technique can be found in Makimoto and Yamashita [159]. Figure 13-34(a) shows a Type 1 stepped-impedance resonator where $Z_s > Z_r$ [159]. The design approach using Type 1 resonators is briefly explained as follows. The uniform resonators of each PCL section in a conventional filter are staggered by a defined amount. The impact of this is subsequently compensated for by setting the

Figure 13-34 Stepped-impedance resonators: (a) A Type 1 stepped-impedance resonator from Makimoto and Yamashita [159]; (b) a pair of coupled stepped-impedance resonators.

characteristic impedances of the uncoupled line sections of the resonators to a fixed high value, Z_s. Consequently the length, l_s, of each uncoupled line in the filter is then optimized until the return loss of the fundamental passband is restored. A pair of stepped-impedance resonators is shown in Figure 13-34(b). Intuitively one could conclude that a change in the stopband performance of a filter depends on the amount of the stagger and the impedance ratio, Z_s/Z_r. This process can be iterated until a suitable stopband performance is obtained. In general this technique offers limited stopband improvement and is best suited to applications where a small shift in the spurious passbands would block harmonics of the input signal from passing through the filter. Thus the spurious passbands of a filter are not eliminated, but in effect the stopband characteristic is altered to suit a particular specification [159]. There is also a Type 2 stepped-impedance resonator with $Z_s < Z_r$ resulting in filters with slightly different stopband behavior to those utilizing Type 1 resonators [159]. In the following it is shown that design of PCL filters based on S-plane bandpass prototypes leads to better stopband performances.

13.5.2   Design of PCL Filters With Enhanced Stopband

In Cheong et al. [162] a PCL filter with uniform-impedance resonators loaded by open-circuited stubs at all its open-circuited ends suppresses the second- and third-ordered harmonic passbands. This approach was generalized in Fathelbab and Steer [157] and this development is described here. The new PCL filters are derived from synthesized bandpass prototypes

with controlled upper stopband responses. This is done by systematic application of appropriate circuit transformations on the initial bandpass prototype resulting in a class of PCL filters comprising uniform- and stepped-impedance resonators, some of which are loaded by open-circuited stubs at their open-circuited ends.

The prototype used here is shown in Figure 13-35(a). It is classified as an S-plane bandpass prototype with the Richards' variable redefined as

$$S = \jmath\Omega_{BP} = \jmath\tan\left(\theta_{BP}\right) = \jmath\tan\left(\frac{\pi}{2}\frac{f}{f_r}\right). \tag{13.71}$$

In Equation (13.71) f_r is the center frequency of the first upper stopband in the f-plane. The general design approach is to synthesize the prototype of Figure 13-35(a) in the S-plane by exact methods [138, 295, 296] and then transform it to the f-plane using Equation (13.71). In the f-plane the inverters, capacitors, and inductors of Figure 13-35(a) become transmission lines, open- and short-circuited stubs, respectively, and are all one-quarter wavelength long at f_r. It has been shown [303] that the f-plane response of an S-plane bandpass prototype is periodic, with a fundamental passband centered at f_0 ($f_0 < f_r$) and the spurious passbands centered at

$$2if_r - f_0; \text{ and } 2if_r + f_0 \quad (i = 1, 2, 3, \dots .) . \tag{13.72}$$

Thus synthesis based on an S-plane bandpass prototype has great flexibility in controlling the location of the first pair of spurious passbands in the f-plane. This is possible by selecting a large commensurate frequency, f_r, to shift the upper spurious passbands further away from the fundamental passband.

A typical f-plane response of the proposed prototype is depicted in Figure 13-35(b), where it can be seen that the series S-plane LC sections (i.e., f-plane short- and open-circuited stubs) contribute to multiple transmission zeroes at if_r, where $i = 1, 2, 3, \dots .$ Doubling the commensurate frequency (i.e., moving from f_r to $2f_r$) leads to the response shown in Figure 13-35(c) where the spurious responses are shifted, leaving behind a broad stopband region.

Now attention is focused on the circuit structure of Figure 13-35(a). Through a number of circuit transformations this is converted into a form that can be implemented using distributed elements. Consider the S-plane subcircuit shown in Figure 13-36(a), which comprises an inverter separating a pair of series capacitors followed by a series inductor. It is feasible to transform the series LC section across the inverter by using relevant Kuroda transformations [138, 303]. This leads to the subcircuit of Figure 13-36(b) whose element values are related to those of Figure 13-36(a) through the

(a)

(b)

(c)

Figure 13-35 Bandpass PCL filter response: (a) S-plane bandpass prototype; (b) real frequency for one commensurate frequency, f_r; and (c) response after doubling f_r.

Figure 13-36 Exact circuit transformations: (a) basic S-plane subcircuit; (b) application of relevant Kuroda transformations; (c) splitting of the 1:N_i transformer; (d) distribution of the series capacitor; and (e) f-plane subsection comprising a parallel coupled line loaded by an open-circuited stub at the left open-circuited end.

following relations:

$$N_i = 1 + \frac{1}{Z_{0i}C_{oi}}$$

$$C_{i1} = \frac{C_{oi}}{N_i\left(N_i - 1\right)C_{oi}Z_{0i} + N_i^2}$$

$$C_{i2} = \frac{L_{o(i+1)}}{N_i Z_{0i} \left(N_i Z_{0i} + L_{o(i+1)} \right)}$$

$$Z_{it} = N_i Z_{0i} + L_{o(i+1)}. \tag{13.73}$$

Now the series capacitor, C_{i1} and the inverter of characteristic impedance Z_{it} form a single PCL section if a 1:N_{i2} transformer is extracted from the existing 1:N_i transformer. The value of the turns ratio of this transformer is

$$N_{i2} = \left(\sqrt{\frac{Z_{it}}{Z_{it} + \frac{1}{C_{i1}}}} \right)^{-1}, \tag{13.74}$$

leaving a 1:N_{i1} transformer of turns ratio

$$N_{i1} = \frac{N_i}{N_{i2}}. \tag{13.75}$$

This step leads to the circuit of Figure 13-36(c). The 1:N_{i2} transformer together with the series capacitor and inverter is the S-plane model of an open-circuited coupled-line section in homogenous media [138]. So for now, ignore the shunt capacitor C_{i2} and investigate the left-hand subcircuit of Figure 13-36(d). If a relevant Kuroda transformation is applied to spread the series capacitor C_{i1}, then the relations

$$C_{i1}' = \frac{1}{N_{i2} \left(N_{i2} - 1 \right) Z_{it}}$$

$$Z_{it}' = N_{i2} Z_{it} \tag{13.76}$$

hold for the circuits of Figure 13-36(d). Then the modal impedances of the PCL section loaded by an open-circuited stub (of characteristic impedance $1/C_{i2}$) at the left open-circuited end are evaluated [303] using Equation (13.76) to give

$$Z_{0ei} = \frac{(1/C_{i1}' + 2Z_{it}')}{N_{i2}^2}, \quad \text{and} \quad Z_{0oi} = \frac{(1/C_{i1}')}{N_{i2}^2}. \tag{13.77}$$

This concludes the transformation of the networks in Figure 13-36(a) into that of Figure 13-36(e), which is almost in the final desired form for implementation using transmission lines only. The ultimate transformation follows.

Figure 13-37(a) shows an S-plane bandpass prototype where the networks enclosed in the dashed boxes can be systematically transformed as just described. The transformation results in the f-plane filter shown in Figure 13-37(b). This is only one among many that could have been obtained. In practical terms it might be desirable not to transform all of the inductors across the inverters to form parallel coupled lines loaded by open-circuited

Figure 13-37 Filter 1: (a) *S*-plane bandpass prototype; and (b) its corresponding *f*-plane filter comprising PCLs loaded by open-circuited stubs at the left open-circuited ends. Filter 2: (c) *S*-plane bandpass prototype; and (d) its corresponding *f*-plane filter comprising PCLs loaded by open-circuited stubs at either the left or the right open-circuited ends. The filters in (b) and (d) are derived after application of the circuit transformation scheme described in Figure 13-36 on the subcircuits enclosed in the dashed boxes of the *S*-plane prototypes.

stubs. This is typically due to extreme element values that may be required that would lead to realization difficulties on a PCB.

To a great extent these constraints can be resolved dependent on how the initial S-plane bandpass prototype is transformed. For this reason, consider application of the transformation steps to the networks (enclosed in the dashed boxes) of the S-plane prototype shown in Figure 13-37(c). This prototype is identical to that of Figure 13-37(a) except that some of the inductors are now divided into pairs. Upon transformation, the f-plane filter illustrated in Figure 13-37(d) results. As seen in Figure 13-37(d), the filter inevitably has some series short-circuited stubs in its main signal path. These can be approximated by sections of high-impedance transmission lines, as described below.

A section of transmission line of short electrical length at f_0 and of high characteristic impedance Z_0 has an input impedance of

$$Z_{\text{in}} \simeq \jmath\,(Z_0)\,(\theta') = \jmath(Z_0)\left(\frac{\pi}{2}\frac{f_0}{f_r'}\right), \qquad (13.78)$$

where f_r' is the one-quarter wavelength frequency of the transmission line. In comparison, a series short-circuited stub of characteristic impedance L is resonant at f_r and has an input impedance at the center of the fundamental passband equal to

$$Z_{\text{in}} = \jmath\,(L)\tan\left(\theta_{BP}\right) = \jmath\,(L)\tan\left(\frac{\pi}{2}\frac{f_0}{f_r}\right). \qquad (13.79)$$

The above equations are approximately equal for passband bandwidths up to an octave (i.e., a 2:1 bandwidth ratio) and hence can be simultaneously solved for the required resonant frequency of the transmission lines, yielding

$$f_r' = \frac{\pi}{2}\frac{Z_0 f_0}{L\tan\left(\frac{\pi}{2}\frac{f_0}{f_r}\right)}. \qquad (13.80)$$

Subsequently the physical length of the transmission line can be evaluated for a given substrate specification. In general the higher the line impedance that can be realized, the better the approximation. However, half of the S-plane inductors in a prototype are transformed into open-circuited stubs and thus the shape of the f-plane transfer function of a prototype in the stopband region is maintained after the above approximation. This completes the filter design approach and now it is time for an example.

13.5.3   Example: PCL Filter With Enhanced Stopband

In this example a seventh-order Chebyshev bandpass filter at 1 GHz with a broad stopband response is designed. The design begins with a seventh-order approximation function in the S-plane [138, 295, 296]. Derivation of

Figure 13-38 Development of a seventh-order parallel coupled line filter: (a) the synthesized S-plane prototype in a 50 Ω system; (b) after splitting some of the inner inductors into pairs; and (c) the topology of the f-plane filter depicting inter resonator couplings. (The X-X' is the plane of symmetry).

the function from the outline procedure presented in the references cited is straight forward but lengthy. From the final function and using standard synthesis, the physically symmetrical S-plane network results and half of it is shown in Figure 13-38(a). In the f-plane, the response of the network has a

fundamental passband center frequency, f_0, of 1 GHz, and a commensurate frequency, f_r, of 2.75 GHz which, according to Equation (13.72), leads to spurious passbands centered at 4.5 GHz and 6.5 GHz. The return loss of the fundamental passband is 16 dB and its bandwidth ratio is 1.22:1 (i.e., 20%), leading to lower and upper band-edge frequencies of 0.9 GHz and 1.1 GHz, respectively.

The next stage of design is transforming all of the inductors in the prototype into open-circuited stubs to yield a PCL filter similar in topology to that of Figure 13-37(b). From the first box (of the prototype of Figure 13-38(a)), the following parameters are obtained:

$$C_{01} = 1/119.025; \ Z_{01} = 34.52; \ L_{02} = 713.332, \tag{13.81}$$

corresponding to the network as shown in Figure 13-36(a). Now using Equation (13.73), a set of element values are evaluated to give

$$\begin{aligned} N_1 &= 4.448, \ C_{11} = 1/2884.297 \\ C_{12} &= 1/186.595, \ Z_{1t} = 866.876 \ . \end{aligned} \tag{13.82}$$

This transforms the subsection to that of Figure 13-36(b). A $1{:}N_{12}$ transformer is then extracted from the $1{:}N_1$ transformer using Equation (13.74). The value of the turns ratio of this transformer is

$$N_{12} = 2.08, \tag{13.83}$$

which, from Equation (13.75), leaves a $1{:}N_{11}$ transformer of turns ratio

$$N_{11} = 2.138 \ . \tag{13.84}$$

This transforms the network to the topology of Figure 13-36(c). The next step is to evaluate the element values of the network in Figure 13-36(d) using Equation (13.76) to give

$$C_{11}' = 1/1947.35, \ Z_{1t}' = 1803.102 \ . \tag{13.85}$$

Now the modal impedances of the first pair of parallel coupled lines, loaded by an open-circuited stub of characteristic impedance $1/C_{12} = 186.595 \ \Omega$ at the left open-circuited end, are calculated using Equation (13.77) as

$$Z_{0e1} = 1283.643 \ \Omega; \ Z_{0o1} = 450.108 \ \Omega \ . \tag{13.86}$$

This concludes the transformation of the first network to a topology similar to that of Figure 13-36(e). In a similar fashion, the networks enclosed in the second and third boxes are transformed, leading to the following element values:

$$\begin{aligned} N_{21} &= 2.14, \ 1/C_{22} = 176.964 \ \Omega, \\ Z_{0e2} &= 1351.136 \ \Omega, \ Z_{0o2} = 403.706 \ \Omega, \end{aligned} \tag{13.87}$$

and

$$N_{31} = 1.737, \ 1/C_{32} = 231.266 \ \Omega,$$
$$Z_{0e3} = 765.535 \ \Omega, \ Z_{0o3} = 336.182 \ \Omega. \tag{13.88}$$

It must be appreciated that each transformed network has a $1{:}N_{i1}$ transformer associated with it. Each of these transformers is eliminated by scaling down the modal impedances of the PCL sections together with the characteristic impedances of the stubs in the filter. This step leads to a symmetrical f-plane filter compatible with that of Figure 13-37(b) having the following element values:

$$
\begin{aligned}
L_{o1} &= 364.926 \ \Omega \\
Z'_{0e1} &= 280.749 \ \Omega, \ Z'_{0o1} = 98.454 \ \Omega, \ 1/C'_{12} = 40.805 \ \Omega \\
Z'_{0e2} &= 64.922 \ \Omega, \ Z'_{0o2} = 19.72 \ \Omega, \ 1/C'_{22} = 8.442 \ \Omega \\
Z'_{0e3} &= 12.093 \ \Omega, \ Z'_{0o3} = 5.31 \ \Omega, \ 1/C'_{32} = 3.653 \ \Omega. \tag{13.89}
\end{aligned}
$$

The design now enters a final phase taking into account fabrication constraints. Here the choice is to construct the filter on an FR4 PCB with a substrate thickness of 62 mil (1.57 mm), relative dielectric constant of 4.7, and loss tangent of 0.016. Practical constraints impose a minimum track width and spacing between tracks of 7 mil (0.177 mm). From Equation (13.89), the coupling coefficients of the parallel coupled lines are calculated and the physical dimensions determined. Unfortunately the resulting dimensions are unrealizable on this type of board. Also, as seen in Equation (13.89), some of the impedances of the stubs are too low. Therefore transformation of the initial prototype to the proposed network topology of Figure 13-37(d) is incomplete. The steps involved are shown in Figures 13-38(b) and 13-38(c). It can now be seen from the circuit of Figure 13-38(c), that the modal impedances are actually high and would still be unrealizable on the PCB. Scaling down the impedance matrix representing this circuit is therefore necessary. This step changes the system impedance of the filter from 50 Ω to 50/2.5 Ω and imposes the placement of an impedance inverter of value 31.62 Ω at each port of the filter. Each inverter is then approximated by a transmission line and the scaled series short-circuited stubs (near the filter ports in Figure 13-38(c)) of characteristic impedance 364.926/2.5 Ω are subsequently transformed into open-circuited stubs using a Kuroda transformation.

For a seventh-order filter there is only a pair of series short-circuited stubs left in the main line of the filter that must be approximated by high-impedance transmission lines. From Figure 13-38(c), the scaled characteristic impedance of one of these short-circuited stubs is 737.934/2.5. The highest characteristic impedance of a transmission that is realizable on the board is 137 Ω, corresponding to a 7 mil (0.177 mm) wide track. Thus, using Equation (13.80), the one-quarter wavelength frequency of the line is evaluated as

Figure 13-39   Electrical layout of the filter in Figure 13-38. (The X-X′ is the plane of symmetry).

Figure 13-40   The PCB layout of the implemented filter having an overall length of 9449 mil (240 mm).

follows:

$$f'_r = \frac{\pi}{2} \frac{137 \left(1 \times 10^9\right)}{(737.934/2.5) \tan\left(\frac{\pi}{2} 12.75\right)} = 1.134 \, \text{GHz}. \tag{13.90}$$
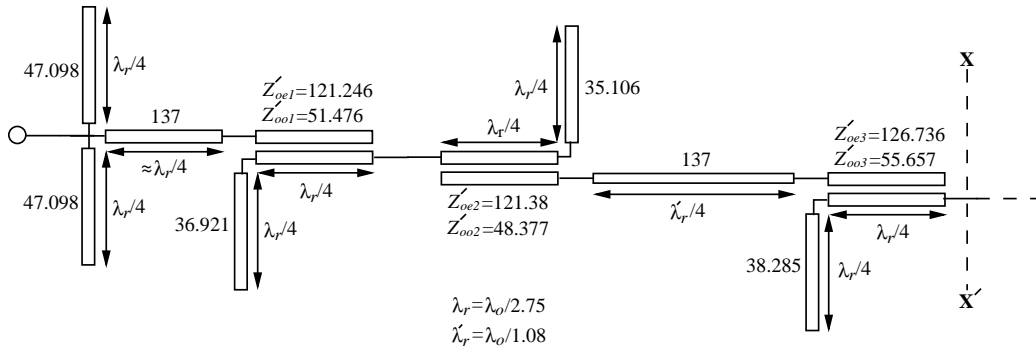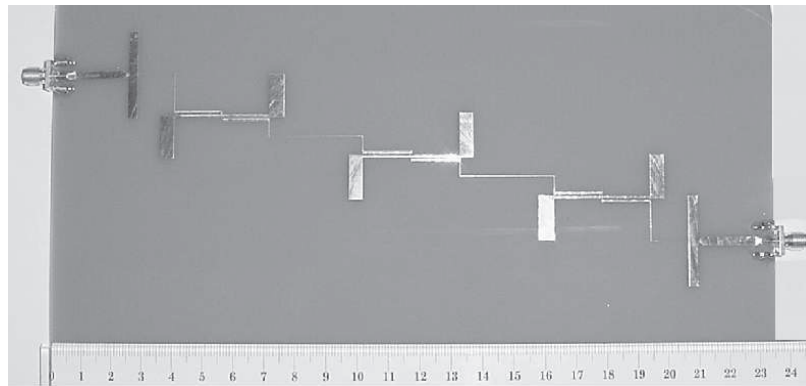
At this stage, tuning of some of the element values of the filter is required to obtain a flat return loss over the passband. This leads to the electrical layout of the filter shown in Figure 13-39. Examination of the inter-resonator couplings and stub impedances show that they can now be conveniently realized on the PCB. The electrical layout of the filter is then converted into physical dimensions (see Figure 13-40).

The measured in-band performance of the filter is shown in Figure 13-41. The center frequency of the passband is offset by 65 MHz (6.5%) from the design objective, which can be compensated for by shortening the physical lengths of the transmission line resonators. The midband insertion loss was 5.1 dB and the return loss is greater than 10 dB (see Figure 13-41(a)). The
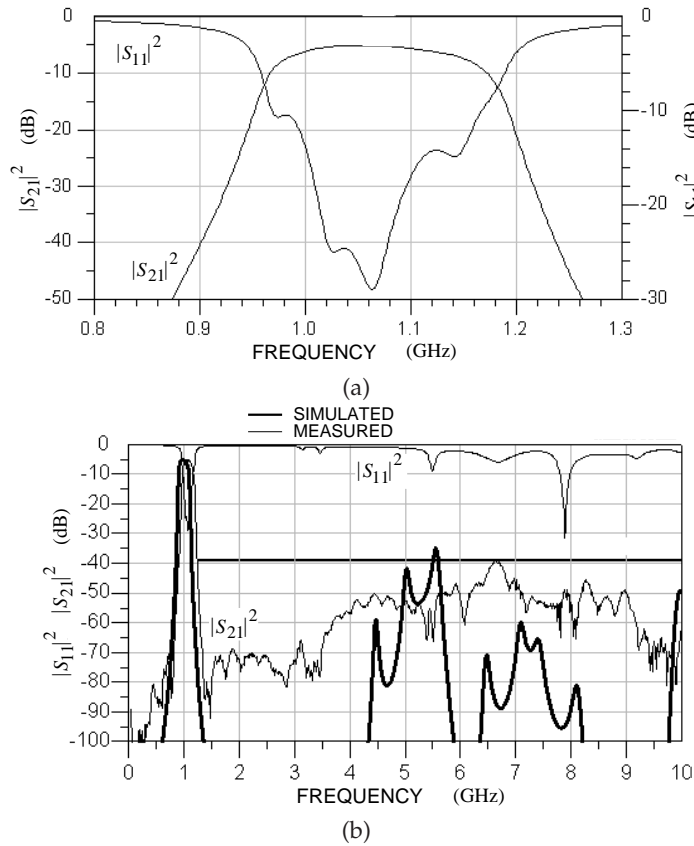
Figure 13-41 Measured and simulated performances of the implemented parallel coupled line filter of Figure 13-40: (a) in-band frequency response; and (b) wideband frequency response.

simulated and measured wideband scattering parameters of the filter from DC up to 10 GHz are shown in Figure 13-41(b).

This filter was designed to have its first pair of spurious passbands centered at 4.5 GHz and 6.5 GHz if it were realized on a homogeneous media. However, the simulated plot of Figure 13-41(b) predicts that the first two spurious passbands are centered at 5.25 GHz and 7.25 GHz instead. This difference is associated with the difference of the even- and odd-mode phase velocities in microstrip. The difference is also partially due to approximating the inner pair of short-circuited stubs in the f-plane filter by high-impedance transmission lines. In summary, the measured transmission performance of the filter followed the simulation and achieves an attenuation level exceeding 40 dB (i.e., an effective rejection level of about

35 dB) up to 10 GHz. As seen from Figure 13-41(b), the stopband of the filter has a tendency to show a spurious passband at the region where the theoretical spurious passbands would have existed. This is around 6 GHz.

13.5.4   Summary

A circuit-oriented approach to the design of a new class of parallel coupled line filters was presented in this section. The parallel coupled line filters are driven from synthesized S-plane bandpass prototypes after appropriate circuit transformations. As a result, the filters have uniform- and stepped-impedance resonators, some of which are loaded by open-circuited stubs at their open-circuited ends. In comparison with the performances of conventional filters utilizing uniform- or stepped-impedance resonators, the new filters have broader stopband characteristics.

13.6   Summary

The purpose of this chapter was to present, by example, advanced microwave circuit design based on network synthesis. To develop competitive designs requires a synthesis approach with an inherent appreciation of what can be achieved by distributed structures. Many lumped-element designs at RF are developed using RF lumped-element equivalents of distributed designs, and in many cases it would not have been conceivable to realize the functionality beginning with lumped elements alone. Sure, single-frequency designs can be done, but these will usually not achieve a competitive product or the best achievable performance.

   RF and microwave network synthesis builds on a solid understanding of filter synthesis. The designs presented in this chapter can be viewed as bandpass or bandstop filters synthesized with particular characteristics. To delve further into this topic means embracing microwave technology. The art of design involves conceptualizing topologies and the science is the knowledge and perseverance to follow through with the detailed mathematical derivations. After this step, computer-aided design tools can be used to achieve the final 5% or 10% of performance by taking into account parasitics and second-order effects that are not captured in the synthesis approach.

References

[1] M.C. Jeruchim, P. Balaban, and K.S. Shanmugan, *Simulation of Communication Systems: Modeling, Methodology, and Techniques,* Kluwer Academic, 2000.

[2] J.R. Carson, "Notes on the theory of modulation," *Proc. IRE*, Vol. 10, Iss. 1, Feb. 1922, pp. 57–64.

[3] L.W. Couch III, *Digital and Analog Communication Systems*, 6th ed., Prentice-Hall, 2001.

[4] E.P.J. Tozer, *Broadcast Engineer's Reference Book*, Focal Press, 2004.

[5] 3rd Generation Partnership Project (3GPP) http://www.3gpp.org

[6] A. Richardson, *WCDMA Design Handbook*, Cambridge University Press, 2005

[7] F.M. Colebrook, "Homodyne," *Wireless World and Radio. Rev.*, Vol. 13, Feb. 1924, pp. 645–648.

[8] D.G. Tucker, "The synchrodyne," *Electronic Engineering*, Vol. 19, March 1947, pp. 75–76.

[9] E.H. Armstrong, "Some recent developments in the audion receiver," *Proc. IRE*, Vol. 3, Iss. 3, Sept. 1915, pp. 215–248.

[10] L. Lessing, *Man of High Fidelity: Edwin Howard Armstrong, a Biography*, J.B. Lippincott, 1956.

[11] S. Ramo, J.R. Whinnery, and T. Van Duzer, *Fields and Waves in Communication Electronics*, John Wiley & Sons, 1965.

[12] C.A. Balanis, *Antenna Theory: Analysis and Design*, 2nd ed., John Wiley & Sons, 2005.

[13] V. Fusco, *Foundations of Antenna Theory and Techniques*, Prentice-Hall, 2004.

[14] M. Born and E. Wolf, "Two-dimensional diffraction of a plane wave by a half-plane," Section 11.5 in *Principles of Optics: Electromagnetic Theory of Propagation, Interference, and Diffraction of Light*, 7th ed., Cambridge University Press, 1999.

[15] J.B. Andersen, T.S. Rappaport, and S. Yoshida, "Propagation measurements and models for wireless communications channels," *IEEE Communications Magazine*, Vol. 33, Iss. 1, Jan. 1995, pp. 42–49.

[16] T.K. Sarkar, Z. Ji, K. Kim, A. Medouri, and M. Salazar-Palma, "A survey of various propagation models for mobile communication," *IEEE Antennas and Propagation Magazine*, Vol. 45, Iss. 3, June 2003, pp. 51–82.

[17] J.S. Seybold, *Introduction to RF Propagation*, 2005, John Wiley & Sons.

[18] "IEEE Virtual Museum" at http://www.ieee-virtual-museum.org Search term: 'Faraday'

[19] J.C. Rautio, "Maxwell's legacy," *IEEE Microwave Magazine*, Vol. 6, Iss. 2 (June 2005), pp. 46–53.

[20] J.C. Maxwell, *A Treatise on Electricity and Magnetism*, Vol. 1, Clarendon Press, 1873. Reprinted, Oxford University Press, 1998.

[21] D.H. Ring, "Mobile Telephony – Wide Area Coverage," *Bell Laboratories Technical Memorandum*, Dec. 11, 1947.

[22] W.D. Lewis, "Coordinated broadband mobile telephone system," *IRE Trans. Vehicle Communications*, Vol. 9, Iss. 1, June 1957, pp. 43–48.

[23] H.J. Schulte and W. A. Cornell, "Multi-area mobile telephony system," *IRE Trans. Vehicle Communications*, May 1960, pp. 49–53.

[24] A.E. Joel, US Patent 3,663,762, filed May 16, 1972.

[25] F. Ikegami, "Mobile radio communications in Japan," *IEEE Trans. Communications*, Vol. 20, Iss. 4, Aug 1972, pp. 738–746.

[26] Cost Final Report, http://www.lx.it.pt/cost231

[27] J. Walfisch and H.L. Bertoni, "A Theoretical model of UHF propagation in urban environments," *IEEE Trans. Antennas Propagation*, Vol.36, 1988, pp.1788–179.

[28] F. Ikegami, T. Takeuchi, and S. Yoshida, "Theoretical prediction of mean field strength for Urban Mobile Radio," *IEEE Trans. Antennas Propagation*, Vol.39, Iss. 3, March 1991, pp. 299–302.

[29] J. Doble, *Introduction to Radio Propagation for Fixed and Mobile Communications*, *Artech House*, Boston-London, 1996.

[30] J.R. Barry, E.A. Edwards, and D.G. Messerschmitt, *Digital Communication*, 3rd ed., Kluwer Academic Publishers, 2004.

[31] D.R. Smith, *Digital Transmission Systems*, 3rd ed., Kluwer Academic Publishers, 2004.

[32] G.E. Lewis, *Communications Systems: Engineers' Choices*, 3rd ed., Focal Press, 1999.

[33] H.K. Markey and G. Antheil, "Secret communication system," US patent 397,412, April. 11, 1942 (filed June 10, 1941).

[34] W.C.Y. Lee, "Spectrum efficiency in cellular [radio]," *IEEE Trans. Vehicular Technology*, Vol. 38, Iss. 2, May 1989, pp. 69–75.

[35] International Telecommunications Union, http://www.itu.int

[36] Third Generation Partnership Project specifications, http://www.3gpp.org/specifications

[37] IEEE Std. 802.11a-1999, Supplement, "Supplement to IEEE standard for information technology telecommunications and information exchange between systems - local and metropolitan area networks - specific requirements. Part II: wireless LAN medium access control (MAC) and physical layer (PHY) specifications: high-speed physical layer in the 5 GHz band," Dec. 30, 1999.

[38] IEEE Std. 802.11a-1999, Appendment, "IEEE standard for information technology—telecommunications and information exchange between systems—local and metropolitan area networks—specific requirements Part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications," 2003.

[39] IEEE Std. 802.11a-1999, "Information technology—telecommunications and information exchange between systems- local and metropolitan area networks- specific requirements part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications amendment 1: high-speed physical layer in the 5 GHz band," 2000.

[40] C. Eklund, R.B. Marks, K.L. Stanwood, S. Wang, "IEEE Standard 802.16: a technical overview of the WirelessMAN Air interface for broadband wireless access," *IEEE Commun. Mag.*, Vol. 40, Iss. 6, June 2002, pp. 98–107.

[41] IEEE Std. 802.16–2001, "IEEE Standard for local and metropolitan area networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems," 2002.

[42] IEEE Std 802.16c-2002, "IEEE Standard for local and metropolitan area networks - Part 16: air interface for fixed broadband wireless access systems-amendment 1: detailed system profiles for 10–66 GHz (Amendment to IEEE Std 802.16–2001)," 2002.

[43] G.J. Foschini, "Layered space-time architecture for wireless communication in a environment when using multi-element antennas," *Bell Labs Technical J.*, Vol. 1, Iss. 2, Autumn 1996, pp. 41–59.

[44] G.G. Raleigh and J.M. Cioffi, "Spatio-temporal coding for wireless communications," *Global Telecommunications Conf. (GLOBECOM) 1996*, Vol. 3, Nov. 1996, pp. 1809–1814.

[45] A. Goldsmith, S.A. Jafar, N. Jindal, S. Vishwanath, "Capacity limits of MIMO channels," *IEEE J. Selected Areas in Communications*, Vol. 21, Iss. 5 (June 2003), pp. 684–702.

[46] D. Gesbert, M. Shafi, D. Shiu, P.J. Smith, and A. Naguib, "From theory to practice: an overview of MIMO space-time coded wireless systems," *IEEE J. Selected Areas in Communications*, Vol. 21, Iss. 3 (April 2003), pp. 281–302.

[47] W. He, C.N. Georghiades, "Computing the capacity of a MIMO fading channel under PSK signaling," *IEEE Trans. Information Theory*, Vol. 51, Iss. 5 (May 2005), pp. 1794–1803.

[48] T. Abe, T. Asai; K. Suda, "A practical throughput comparison of MIMO-CDMA and MIMO-OFDM," *IEEE 60th Vehicular Technology Conf. (VTC2004-Fall. 2004)*, Vol. 2, 26–29 Sept. 2004, pp. 1431–1438.

[49] "Wireless transmitting and receiving mechanism for electric waves," US patent 810,150, Jan. 16, 1906 (filed March 14, 1904). Patented in Dusseldorf as "Telemobiloskop" in 1904.

[50] O. Heaviside, 'Electromagnetic induction and its propagation,' *Electrician*, 3 June 1887.

[51] T.C. Edwards and M.B. Steer, *Foundations of Interconnect and Microstrip Design*, John Wiley & Sons, Chichester, 2000.

[52] R.M. Barrett, "Microwave printed circuits—a historical survey," *IEEE Trans. Microwave Theory Techn.*, Vol. 3 , Iss. 2 (March 1955), pp. 1–9.

[53] C. P. Wen "Coplanar waveguide: a surface strip transmission line suitable for nonreciprocal gyromagnetic device applications," *IEEE Trans. Microwave Theory and Techniques*, Vol. 17, Iss. 12 (Dec. 1969), pp. 1087–1090.

[54] N. Simons, N.I. Dib, and L.P.B. Katehi, "Modeling of coplanar stripline discontinuities," *IEEE Trans. Microwave Theory and Techn.*, vol. 44, Iss. 5 (May 1996) pp. 711–716.

[55] D. Prieto *et al.*, "CPS structure potentialities for MMICs: a CPS/CPW transition and a bias network," *1998 IEEE MTT-S Int. Microwave Symp. Dig.*, June 1998, pp. 111–114.

[56] K. Goverdhanam, R.N. Simons and L.P.B. Katehi, L. P. B., "Micro-coplanar striplines—a new transmission media for microwave applications," *1998 IEEE MTT-S Int. Microwave Symp. Dig.*, June 1998 pp. 1035–1038.

[57] K. Goverdhanam, R.N. Simons and L.P.B. Katehi, L. P. B., 'Coplanar stripline components for high-frequency applications,' *IEEE Trans. Microwave Theory and Techn.*, Vol. 45, Oct. 1997, pp. 1725–1729.

[58] J.B. Knorr and K.-D. Kuchler, "Analysis of coupled slots and coplanar strips on dielectric substrate," *IEEE Trans. Microwave Theory and Techn.*, Vol. 23, July 1975, pp. 541–548.

[59] S.G. Pintzos, "Full-wave spectral-domain analysis of coplanar strips," *IEEE Trans. Microwave Theory and Techn.*, Vol. 39, Feb. 1991, pp. 239–246.

[60] M.A.R. Gunston, *Microwave Transmission-line Impedance Data*, Van Nostrand Reinhold Company, London, 1973.

[61] P.A.R. Holder, "X-band microwave integrated circuits using slotline and coplanar waveguide," *The Radio and Electronic Engineer*, vol. 48, Iss. 1/2 (Jan./Feb. 1978), pp. 38–43.

[62] K.C. Gupta, R. Garg, and I.J. Bahl, *Microstrip Lines and Slotlines*, 2nd ed., Artech House, 1996.

[63] B.E. Spielman, "Computer-aided analysis of dissipation losses in isolated and coupled transmission lines for microwave and millimetre-wave applications," *NRL Formal Report*, Iss. 8009, 1976.

[64] A.K. Ganguly and B.E. Spielman, "Dispersion characteristics for arbitrarily configured transmission media," *IEEE Trans. Microwave Theory and Tech.*, Vol. 25, Iss. 12 (Dec. 1977), pp. 1138–1141.

[65] C.B. Boyer and U.C. Merzbach, "Invention of logarithms" *A History of Mathematics*, 2nd ed., John Wiley & Sons, pp. 312–313, 1991.

[66] E.O. Hammerstad and Ø Jensen, "Equations for microstrip circuit design," *1980 IEEE MTT-S Int. Microwave Symposium Dig.*, 1980.

[67] E.O. Hammerstad and F. Bekkadal, 'A microstrip handbook,' ELAB Report, STF44 A74169, University of Trondheim, Norway, Feb. 1975.

[68] E.O. Hammerstad, "Equations for microstrip circuit design," *Proc. 5th European Microwave Conf.*, Sept. 1975.

[69] H.A. Wheeler, "Transmission-line properties of parallel wide strips by a conformal mapping approximation," *IEEE Trans. Microwave Theory and Tech.*, vol. 12, Iss. 12, May 1964, pp. 280–289.

[70] H.A. Wheeler, "Transmission-line properties of parallel strips separated by a dielectric sheet," *IEEE Trans. Microwave Theory and Tech.*, vol.13, Iss. 3, Mar. 1965, pp. 172–185.

[71] Cahill, L. W., 'Approximate formulas for microstrip transmission lines,' *Proc. Institute of Radio Engineers*, vol. 3, Oct. 1974, pp. 317–321.

[72] R.P. Owens, "Accurate analytical determination of quasi-static microstrip line parameters," *The Radio and Electronic Engineer*, Vol. 46, Iss. 7 (July 1976), pp. 360–364.

[73] P.B. Katehi and N.G. Alexopoulos, "Frequency-dependent characteristics of microstrip discontinuities in millimeter-wave integrated circuits," *IEEE Trans. Microwave Theory Techn.*, Vol. 33, Iss. 10,

Oct. 1985, pp. 1029–1035.

[74] G.D. Vendelin, "Limitations on stripline Q," *Microwave J.,* May 1970, pp. 63–69.

[75] H. Hasegawa et al. *IEEE Trans. Microwave Theory Tech.*, Vol. 19, Iss. 11 (Nov. 1971), pp. 869–881.

[76] D. Jäger, "Slow-wave propagation along variable Schottky-Contact Microstrip Line," *IEEE Trans. Microwave Theory Tech.* Vol. 24, Iss. 9 (Sept. 1976), pp. 556–573.

[77] Y.R. Kwon et al, "Quasi-TEM analysis of 'Slow- Wave mode' propagation on coplanar microstructure MIS transmission lines", *IEEE Trans. Microwave Theory Tech.* Vol.35, Iss. (June 1987), pp. 545–551.

[78] A. Stogryn, "Equations for calculating the dielectric constant of saline water," *IEEE Trans. Microwave Theory and Tech.*, 1971, pp. 733–736.

[79] H. Marshall and V. Yakovlev, *36 th Microwave Power Symposium*, April 2001.

[80] S.J. Mason, "Feedback Theory - some properties of signal flow graphs," *Proc Institute of Radio Engineers*, vol. 41, 1953, pp. 1144–1156.

[81] A. Tustin, *Direct Current Machines for Control Systems*, The Macmillan Company, 1956.

[82] Balabanian,*Fundamentals of circuit theory*

[83] Abrahams and Coverly,*Signal Flow Analysis*

[84] Di Stefano et.al., *Theory and problems of feedback control systems*

[85] S.J. Mason, "Feedback theory-further properties of signal flow graphs," Proc. IRE, Vol. 44, Iss. 7, July 1956, pp. 920–926.

[86] P.H. Smith, *Electronic Applications of the Smith Chart: In Waveguide, Circuit, and Component Analysis*, 2000, SciTech Publishing, ISBN 1–884932–39–8.

[87] P.H. Smith, "Transmission line calculator," *Electronics*, Vol. 12, Jan. 1939, pp. 29–30.

[88] P.H. Smith, "An improved transmission-line calculator," *Electronics*, Vol. 17, Jan. 1944, pp. 130–133, 318, 320, 322, 324, 325.

[89] P.H. Smith, "R.F. transmission line nomographs," *Electronics*, Vol. 22, Feb. 1949, pp. 112–117.

[90] R.A. Hackborn, "An automatic network analyzer system," *Microwave J.*, May 1968, pp. 45–52.

[91] S.B. Goldberg, M.B. Steer, P.D. Franzon and J.S. Kasten, "Experimental electrical characterization of interconnects and discontinuities in high speed digital systems," *IEEE Trans. Components Hybrids and Manufacturing Tech.*, Vol. 14, Dec., 1991, pp. 761–765.

[92] M.B. Steer, S.B. Goldberg, G. Rinne, P.D. Franzon, I. Turlik and J.S. Kasten, "Introducing the through-line deembedding procedure," *1992 IEEE MTT-S Int. Microwave Symp. Dig.*, June 1992, pp. 1455–1458.

[93] J.K. Hunton, "Analysis of microwave measurement techniques by means of signal flow graphs," *IRE Trans. Microwave Theory and Tech.*, March 1960, pp. 206–212.

[94] J. Fitzpatrick, "Error models for systems measurement," *Microwave J.*, May 1978, pp. 63–66.

[95] A.M. Saleh, "Explicit formulas for error correction in microwave measuring sets with switching-dependent port mismatches," *IEEE Trans. Instrumentation and Measurement*, Vol. 28, Iss. 1 (March 1979), pp. 67–71.

[96] H.V. Shurmer, "Calibration procedure for computer corrected s parameter characterization of devices mounted in microstrip," *Electron. Lett.*, Vol. 9, Iss. 14, July 1973, pp. 323–324.

[97] *De-embedding and Embedding S-paramater Networks Using a Vector Network Analyzer*, Agilent Application Note 1363–1, May 2004.

[98] G.F. Engen and C.A. Hoer, "Thru-reflect-line: an improved technique for calibrating the dual six-port automatic network analyzer," *IEEE Trans. Microwave Theory Tech.*, Vol. 27, Iss. 12, Dec. 1979, pp. 987–993.

[99] P.M. Buff, *Characterization of propagation on wires over lossy earth,* Ph.D. dissertation, North Carolina State University, 2006.

[100] P.M. Buff, J. Nath and M.B. Steer, "Origin of the half-wavelength errors in microwave measurements using through-line calibrations," *IEEE Trans. Instrumentation and Measurement*, Vol. 55, Iss. 10 (Oct. 2007), pp 1610–1615.

[101] B. Bianco, M. Parodi, S. Ridella and F. Selvaggi, "Launcher and microstrip characterization," *IEEE Trans. Instrumentation and Measurement*, Vol. 25, Iss. 4, Dec. 1976, pp. 320–323.

[102] C.A. Hoer, "Choosing line lengths for calibrating network analyzers," *IEEE Trans. Microwave Theory Tech.,* Vol. 31, Iss. 1, Jan. 1983, pp. 76–78.

[103] C.A. Hoer, "Some questions and answers concerning air lines as impedance standards," *Proc. 29th Automatic RF Tech. Group Conf.,*, June 1987, pp. 161–173.

[104] R. Marks, "A multiline method of network analyzer calibration," *IEEE Trans. Microwave Theory Tech.,* Vol. 39, Iss. 7, July 1991, pp. 1205–1215.

[105] D. Williams, "De-embedding and unterminating microwave fixtures with nonlinear least squares," *IEEE Trans. Microwave Theory Tech.,* Vol. 38, Iss. 6, June 1990, pp. 787–791.

[106] N. R. Franzen and R. A. Spaciale, "A new procedure for system calibration and error removal in automated s-parameter measurements," *Proc. 5th European Microwave Conf.,* 1975, pp. 69–73.

[107] J.A. Benet, "The design and calibration of a universal MMIC test fixture," *1982 IEEE MTT-S Int. Microwave Symp. Dig.,* 1982, pp. 36–41.

[108] S.R. Pennock, C.M.D. Rycroft, P.R. Shepherd and T. Rozzi, "Transition characterisation for de-embedding purposes," *Proc. 17th European Microwave Conf,* 1987, pp. 355–360.

[109] T.H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*, Second Edition, Cambridge University Press, 2004.

[110] H.A. Wheeler, "Simple inductance formulas for radio coils," *IRE Proc.,* Vol. 16, Iss. 10, Oct. 1928, pp. 1398–1400.

[111] H.A. Wheeler, "Inductance formulas for circular and square coils," *Proc. IEEE*, Vol. 70, Iss. 12, Dec. 1982, pp. 1449–1450.

[112] H.A. Wheeler, "Inductance chart for solenoid coil," *Proc. IRE*, Vol. 38, Iss. 12, Dec. 1950, pp.1398–1400.

[113] C.P. You and S.S. Wong, 'On-chip spiral inductors with patterned ground shields fo Si-based RF IC's,' *Dig. Tech. Papers Sym. VLSI Circuits*, 1997, pp. 85–86.

[114] T.H. Lee, The Design of CMOS Radio-Frequency Integrated Circuits, Second Edition, Cambridge University Press, 2004.

[115] Z. Feng, C. A. Bower, J. Carlson, M. Lueck, D. Temple and M. B. Steer, "High-Q solenoidal inductive elements," *2007 IEEE MTT-S Int. Microwave Symp. Dig.,* June 2007, pp. 1905–1908.

[116] W. Y. Liu, J. Suryanarayanan, J. Nath, S. Mohammadi, L. P. B. Katehi and M. B. Steer, "Torroidal inductors for radio frequency integrated circuits," *IEEE Trans. Microwave Theory and Tech.,* Vol. 52, Iss. 2, Feb. 2004, pp. 646–654.

[117] IEEE Standard 315-1975, Graphic Symbols for Electrical and Electronics Diagrams (Including Reference Designation Letters)," Institute of Electrical and Electronics Engineers, USA, Adopted Sept. 1975, Reaffirmed Dec. 1993. Approved by American National Standards Institute, Jan. 1989. Approved adopted for mandatory use, Department of Defense, United States of America, Oct. 1975. Approved by Canadian Standards Institute, Oct. 1975.

[118] R. Mongia, I. Bahl, and P. Bhartia, *RF and Microwave Coupled-line Circuits*, Artech House, pp. 411–435, 1999.

[119] W.E. Fathelbab and M.B. Steer, "Broadband network design," Chapter 8 in *Multifunctional Adaptive Microwave Circuits and Systems*, Edited by M. Steer and W.D. Palmer, SciTech Publishing, Inc., 2008.

[120] I. D. Robertson and S. Lucyszyn, *RFIC and MMIC Design and Technology*, IEE Circuits, Devices and Systems Series 13, 2001.

[121] D. Kuylenstierna, and P. Linner, "Design of broad-band lumped-element baluns with inherent impedance transformation," *IEEE Trans. Microwave Theory Tech.,* Vol. 52, Iss. 12, Dec. 2004, pp. 2739–2745.

[122] B. J. Minnis, and M. Healy, "New broadband balun structures for monolithic microwave integrated circuits," 1991 *IEEE MTT-S Int. Microwave Symp. Dig.*, June 1991, pp. 425–428.

[123] C. L. Goldsmith, A. Kikel, and N. L. Wilkens, "Synthesis of Marchand baluns using multilayer

microstrip structures," *Int. J. of Microwave & Millimeter-Wave Computer-Aided Eng.*, Vol. 2, Iss. 3, 1992, pp. 179–188.

[124] N. Marchand, "Transmission line conversion transformers," *Electronics*, pp. 142–145, Dec. 1944.

[125] W. M. Fathelbab and M. B. Steer, "New classes of miniaturized planar Marchand baluns," *IEEE Trans. Microwave Theory and Tech.*, Vol. 53, Iss. 4, April 2005, pp. 1211–1220.

[126] J. Cloete "Graphs of circuit elements for the Marchand balun," *Microwave J.*, Vol. 24, Iss. 5, May 1981, pp. 125–128.

[127] E.J. Wilkinson, "An N-way hybrid power divider," *IRE Trans. Microwave Theory and Techniques*, Vol. 8, Iss. 1, Jan. 1960, pp. 116–118.

[128] J.-S. Lim, K.-S. Choi, D. Ahn, S. Oh, J.-J. Koo M.-S. Hwang, C. Park, and Y.-C. Jeong, "An unequal Wilkinson power divider with variable dividing ratio," *2007 IEEE MTT-S Int. Microwave Symposium*, 3–8 June 2007, pp. 411–414.

[129] Y.-L. Wu, H. Zhou; Y.-X. Zhang, and Y.-A. Liu, "An unequal Wilkinson power divider for a frequency and its first harmonic," *IEEE Microwave and Wireless Components Letters*, Vol. 18, Iss. 11, Nov. 2008, pp. 737–739.

[130] J.-S. Lim; S.-W. Lee; C.-S. Kim; J.-S. Park, D. Ahn, S. Nam, "A 4.1 unequal Wilkinson power divider," *IEEE Microwave and Wireless Components Letters*, Vol. 11, Iss. 3, March 2001, pp. 124–126.

[131] J. Helszajn, *YIG Resonators and Filters*, Wiley, 1985.

[132] R.M. Fano, *Theoretical Limitations on the Broadband Matching of Arbitrary Impedances*, Massachusetts Institute of Technology, Research Laboratory of Electronics, Technical Report 42, Jan. 2, 1948.

[133] R.M. Fano "Theoretical limitations on the broad-band matching of arbitrary impedances," *J. Franklin Institute*, Vol. 249, Jan. 1950, pp. 57–83, and Feb. 1950, pp. 139-154.

[134] H.W. Bode, *Network Analysis and Feedback Amplifier Design*, Van Nostrand, 1945.

[135] B. Biswas, *Modeling and simulation of high speed interconnects*, M.S. Thesis, North Carolina State University, 1998.

[136] L. Lange, "Interdigitated strip-line quadrature hybrid,' *IEEE Trans. Microwave Theory Tech.*, Vol. 17, Dec. 1969, pp. 1150–1151.

[137] G.I. Zysman and A.K. Johnson, "Coupled transmission line networks in an inhomogeneous dielectric medium," *IEEE Trans. Microwave Theory Tech.*, Vol. 17, Iss. 10, Oct. 1969, pp. 753–759.

[138] J. A. G. Malherbe, *Microwave Transmission Line Filters*, Artech House, 1979.

[139] J.O. Scanlan and R. Levy, *Circuit Theory*, Edinburgh, Oliver & Boyd, 1973.

[140] J.O. Scanlan and R. Levy, *Circuit Theory*, Vol. 2, 1970.

[141] G.L. Matthaei, L. Young and E.M.T. Jones, *Microwave Filters, Impedance-Matching Networks and Coupling Structures*, McGraw-Hill, New York, 1965. Reprinted in 1980, Artech House.

[142] I.C. Hunter, *Theory and Design of Microwave Filters*, IEE Press, 2001.

[143] R.W. Daniels, *Approximation Methods for Electronic Filter Design*, McGraw-Hill, 1974.

[144] M.D. Lutovac, D. Miroslav, D.V. Tosic, V. Dejan, and B.L. Evans, *Filter Design for Signal Processing using MATLAB and Mathematica*, Prentice Hall, 2001.

[145] I. Bahl, "Capacitively compensated high performance parallel coupled microstrip filters," *IEEE MTT-S Int. Microwave Symp. Dig.*, June 1989, pp. 679–682.

[146] S. M. Wang, C. H. Chi, M. Y. Hsieh, C. Y. Chang, "Miniaturized spurious passband suppression microstrip filter using meandered parallel coupled lines," *IEEE Trans. Microwave Theory Tech.*, Vol. 53, Iss. 2, Feb. 2005, pp. 747–753.

[147] M. C. Velazquez-Ahumada, J. Martel, and F. Medina, "Parallel coupled microstrip filters with ground-plane aperture for spurious band suppression and enhanced coupling," *IEEE Trans. Microwave Theory Tech.*, Vol. 52, Iss. 3, Mar. 2004, pp. 1082–1086.

[148] M. C. Velazquez-Ahumada, J. Martel, and F. Medina, "Parallel coupled microstrip filters with floating ground-plane conductor for spurious-band suppression," *IEEE Trans. Microwave Theory Tech.*, Vol. 53, Iss. 5, May 2005, pp. 1823–1828.

[149] L. Zhu, H. Bu, and K. Wu, "Broadband and compact multi-pole microstrip bandpass filters using

ground plane aperture technique," *IEE Proc.-Microw. Antennas Propag.*, Vol. 149, Iss. 2, Feb. 2002, pp. 71–77.

[150] J. T. Kuo, M. Jiang, and H. J. Chang, "Design of parallel-coupled microstrip filters with suppression of spurious resonances using substrate suspension," *IEEE Trans. Microwave Theory Tech.*, Vol. 52, Iss. 1, Jan. 2004, pp. 83–89.

[151] J. T. Kuo, and M. Jiang, "Enhanced Microstrip Filter Design with a Uniform Dielectric Overlay for Suppressing the Second Harmonic Response," *IEEE Microwave and Wireless Components Letters*, Vol. 14, Iss. 9, Sept. 2004, pp. 419–421.

[152] T. Lopetegi, M. A. G. Laso, J. Hernandez, M. Bacaicoa, D. Benito, M. J. Garde, M. Sorolla, and M. Guglielmi, "New microstrip "wiggly-line" filters with spurious passpand suppression," *IEEE Trans. Microwave Theory Tech.*, Vol. 49, Iss. 9, Sept. 2001, pp. 1593–1598.

[153] T. Lopetegi, M. A. G. Laso, F. Falcone, F. Martin, J. Bonache, J. Garcia, L. Perz-Ceivas, M. Sorolla, and M. Guglielmi, "Microstrip "wiggly-line" filters with multispurious rejection," *IEEE Microwave and Wireless Components Letters*, Vol. 14, Iss. 11, Nov. 2004, pp. 531–533.

[154] B. S. Kim, J. W. Lee, and M. S. Song, "An implementation of harmonic-suppression microstrip filters with periodic grooves," *IEEE Microwave and Wireless Components Letters*, Vol. 14, Iss. 9, Sept. 2004, pp. 413–415.

[155] J. T. Kuo, W. H. Hsu, and W. T. Huang, "Parallel coupled microstrip filters with suppression of harmonic response," *IEEE Microwave and Wireless Components Letters*, Vol. 12, Iss. 10, Oct. 2002, pp. 383–385.

[156] J. T. Kuo, W. H. Hsu, and W. T. Huang, "Tapped wiggly-coupled technique applied to microstrip bandpass filters for multi-octave spurious suppression," *Electronics Letters*, Vol. 40, Iss. 1, Jan. 2004, pp. 46–47.

[157] W. M. Fathelbab and M. B. Steer, "Parallel-coupled line filters with enhanced stopband performance," *IEEE Trans. Microwave Theory and Tech.*, Vol. 53, Iss. 12, Dec. 2005, pp. 3774–3781.

[158] J. Garcia-Garcia, F. Martin, F. Falcone, J. Bonache, I. Gil, T. Lopetegi, M. Laso, M. Sorolla, and R. Marques, "Spurious passband suppression in microstrip coupled line band pass filters by means of split ring resonators," *IEEE Microwave and Wireless Components Letters*, Vol. 14, Iss. 9, Sept. 2004, pp. 416–418.

[159] M. Makimoto and S. Yamashita, "Bandpass filters using parallel coupled striplines stepped impedance resonators," *IEEE Trans. Microwave Theory Tech.*, Vol. 28, Iss. 12, Dec. 1980, pp. 1413–1417.

[160] I. Bahl and P. Bhartia, *Microwave Solid State Circuit Design*, Wiley-Interscience, John Wiley & Sons, 1988.

[161] A.H. Nakamura *et al.*, 'Expert system for microwave filter design,' *1990 IEEE MTT-S Int. Microwave Symp.*, May 1990, pp. 1183–1186.

[162] P. Cheong, S. W. Fok, and K. W. Tam "Miniaturized parallel coupled-line bandpass filter with spurious-response suppression," *IEEE Trans. Microwave Theory Tech.*, Vol. 53, Iss. 5, May 2005, pp. 1810–1816.

[163] F. Rosenbaum, R.O. Gregory, W.D. Richard, W. Ou, F.G. Kuhns, and T.M. Trimble, "An MMIC twin-tee active bandpass filter," *1993 IEEE MTT-S Int. Microwave Symposium Dig.*, vol. 1, 1993, pp. 361–364.

[164] B.Y. Kapilevich, "Active microwave filters," *Telecommunication Radio Engineering*, Vol. 4, Iss. 2, Feb. 1985, pp. 51–58.

[165] C.-Y. Chang and T. Itoh, "A varactor-tuned, active microwave band-pass filter, *Electronics Letters*, Vol. 25, Iss. 16, Aug. 1989, pp. 1228–1229.

[166] C.-Y. Chang and T. Itoh, "A varactor-tuned, active microwave band-pass filter, *1990 IEEE MTT-S Int. Microwave Symposium Dig.*, Vol. 1, May 1990, pp. 499–502.

[167] U. Karacaoglu and I.D. Robertson, "High selectivity varactor-tuned MMIC band-pass filter using lossless active resonators," *1994 IEEE MTT-S Int. Microwave Symposium Dig.*, Vol. 2, May 1994, pp.

613–615.

[168] S.M. Sze and K.K. Ng, *Physics of Semiconductor Devices*, Third Edition, John Wiley & Sons, 2007.

[169] D.K. Schroder, *Semiconductor Material and Device Characterization*, IEEE Press and Wiley, 2006.

[170] B. Streetman and S. Banerjee, *Solid State Electronic Devices*, sixth edition, Pearson Prentice Hall, 2006.

[171] H.K Gummel, H.C Poon, "An integral charge control model of bipolar transistors," Bell Syst. Tech. J, vol. 49, MayJune 1970, pp. 827852

[172] M.A.L. Johnson, D.W. Barlage, and D. Braddock, "Prospect for III-nitride heterojunction MOSFET structures and devices," *Materials Research Society Symposium Proceedings*, 2004.

[173] C. Roff, P. McGovern, J. Benedikt, P.J. Tasker, M.A. Johnson, D.W. Barlage, W. Sutton, and D. Braddock, "Pulsed-IV and RF waveform measurements of unique high-K dielectric GaN MOSFETs," *IEEE Int. Conf. on Microwaves, Communications, Antennas and Electronic Systems, 2008 (COMCAS 2008)*, 13–14 May 2008, pp. 1–4.

[174] R.J. Baker, *CMOS Circuit Design, Layout, and Simulation*, Second Edition, Wiley-Interscience, IEEE Press, 2008.

[175] Y. Bito, N. Iwata, and M. Tomita, "64% efficiency enhancement-mode power heterojunction FET for 3.5 VLi-ion battery operated personal digital cellular phones," *1998 IEEE MTT-S Int. Microwave Symposium Dig.*, Vol. 2, 7–12 Jun 1998, pp. 439–442.

[176] A. Materka and T. Kacprzak, "Computer calculation of large-signal GaAs FET amplifier characteristics," Vol. 33, Iss. 2, Feb. 1985, pp. 129–135.

[177] S. Lucyszyn, "Power-added efficiency errors with RF power amplifiers," *Int. J. of Electronics*, Vol. 82, Iss. 3, Mar. 1997, pp. 303–312.

[178] RFMD, FPD6836P70 Data Sheet, low noise high frequency packaged enhancement modee pHEMT transistor. http://www.rfmd.com

[179] L. Esaki and R. Tsu, "Superlattice and nagative differential conductivity ion semiconductors," IBM J. Research and Development, Jan. 1970, pp. 61–65.

[180] R. J. Trew and M. B. Steer, "Millimeter-wave performance of state-of-the-art MESFET, MODFET and PBT transistors," *Electronics Letters,* Vol. 8, 12 th Feb. 1987, pp. 149–151.

[181] M. B. Steer and R. J. Trew, "High frequency limits of millimeter-wave transistors," *IEEE Electron Device Letters,* Vol. 7, Nov. 1986, pp. 640–642.

[182] J. M. Rollett, "Stability and power gain invariants of linear two-ports," *IRE Trans. Circuit Theory*, Vol. 9, Mar. 1962, pp. 29–32.

[183] G. Gonzalez, *Microwave Transistor Amplifiers: Analysis and Design*, second edition, Prentice Hall, 1997.

[184] A. Suárez and R. Quéré, *Stability Analysis of Nonlinear Microwave Circuits*, Artech House, 2003.

[185] M.L. Edwards and J.H. Sinsky, "A new criterion for linear 2-port stability using a single geometrically derived parameter," *IEEE Trans. Microwave Theory and Techniques*, Vol. 40, Iss. 12, Dec. 1992, pp. 2303–2311.

[186] T.T. Ha, *Solid State Microwave Amplifier Design*, Wiley, 1981.

[187] D. Woods, "Reappraisal of the unconditional stability criteria for active 2-port networks in terms of *S*-parameters," *IEEE Trans. Circuits Systems*, Vol. 23, Iss. 2, Feb. 1976, pp. 73–81.

[188] W.H. Ku, "Unilateral gain and stability criterion of active two-ports in terms of scattering parameters," *Proc. IEEE*, Vol. 54, Nov. 1966, pp. 1617–1618.

[189] D.C. Youla, "A note on the stability of linear nonreciprocal *n*-ports," *Proc. IRE*, Vol. 48, Iss. 1, Jan. 1960, pp. 121–122.

[190] R.P. Meys, "Review and discussion of stability criteria for linear 2-ports," *IEEE Trans. Circuits Systems*, Vol. 37, Iss. 11, Nov. 1990, pp. 1450–1452.

[191] G.E. Bodway, "Two port power flow analysis using generalized scattering parameters," *Microwave J.*, May 1967, pp. 61–69.

[192] G.D. Vendelin, A.M. Pavio and U.L. Rohde,*Microwave Circuit Design Using Linear and Nonlinear Techniques*, Wiley, 1990.

[193] W. M. Fathelbab and M. B. Steer, "Distributed biasing of differential RF circuits " *IEEE Trans. Microwave Theory and Tech.,* Vol. 52, Iss. 5, May 2004, pp. 1565 – 1572.

[194] Mini-Circuits HELA-10B, 50 MHz to 1 GHz amplifier.

[195] J. F. Sevic and M. B. Steer, "On the significance of envelope peak-to-average ratio for estimating the spectral regrowth characteristics of an RF/Microwave power amplifier," *IEEE Trans. Microwave Theory Techniques,* Vol. 48, Iss. 6, June 2000, pp. 1068–1071.

[196] N.M. Kriplani, S. Luniya, and M.B. Steer, "Integrated deterministic and stochastic simulation of electronic circuits: application to large signal noise analysis," Int. J. Numerical Modeling, 2008.

[197] N.M. Kriplani, *Modelling Colored Noise under Large-Signal Conditions*, PhD Dissertation, North Carolina State University, Raleigh, North Carolina, 2005.

[198] H.L. Hartnagel, R. Katilius, and A. Matulionis, *Microwave Noise in Semiconductor Devices*, Wiley, 2001.

[199] H.A. Haus, W.R. Atkinson, G.M. Branch, W.B. Davenport, W.H.Fonger, W.A Harris, S.W.; Harrison, W.W. McLeod, E.K. Stodola, and T.E. Talpey, "Representation of noise in linear two-ports," *Proc. IEEE*, Vol. 48, Iss. 1, Jan. 1960, pp. 69–74.

[200] J.F. Sevic and M.B. Steer, "Volterra series analysis of MESFET spectrum regeneration driven by a $\pi/4$-DQPSK modulated source," *1995 IEEE MTT-S Int. Microwave Symp. Dig.*, May 1995.

[201] K.G. Gard, *Autocorrelation Analysis of Spectral Regrowth Generated by Nonlinear Circuits in Wireless Communication Systems*, Ph.D. dissertation, University of California at San Diego, 2003.

[202] L. Dai and R. Harjani, *Design of High Performance CMOS Voltage-Controlled Oscillators* Kluwer Academic Publishers, 2003.

[203] M. Tiebout, *Low Power VCO Design in CMOS*, 2006.

[204] A. Aktas and M. Ismail, *CMOS PLLs and VCOs for 4G wireless*, Kluwer, 2004.

[205] D. O. Pederson and K. Mayaram, *Analog Integrated Circuits for Communication: Principles, Simulation and Design*, Springer, 2008.

[206] B. Razavi, *Design of Analog CMOS Integrated Circuits*, McGraw-Hill, 2001.

[207] P R Gray, P J Hurst, S H Lewis, and R G Meyer, *Analysis and Design of Analog Integrated Circuits*, Fourth Edition, Wiley, 2001.

[208] B. Leung, *VLSI for Wireless Communications*, Prentice Hall, 2002.

[209] X. Yang, A. Davierwalla, D. Mann, and K.G. Gard, "A 90nm CMOS Direct Conversion Transmitter for WCDMA," *2007 IEEE Radio Frequency Integrated Circuits (RFIC) Symposium*, 3-5 June 2007, pp. 17–20.

[210] B. Gilbert, "The multi-tanh principle: a tutorial overview," *IEEE J. Solid State Circuits*, Vol. 33, Iss. 1, Jan 1998, pp. 2–17.

[211] X. Yang, *90nm CMOS Transmitter Design for WCDMA*, PhD Dissertation, North Carolina State Unievrsity, 2009.

[212] M. B. Steer and K. M. Gharaibeh, "Volterra Modeling for Analog and Microwave Circuits," in *Encyclopedia of RF and Microwave Engineering,* John Wiley, 2005, pp. 5507–5514.

[213] J.C. Pedro and N.B. Carvalho, *Intermodulation Distortion in Microwave and Wireless Circuits*, Artech House, 2003.

[214] T.R. Turlington, *Behavioral Modeling of Nonlinear RF and Microwave Devices*, Artech House, 2000.

[215] S.A. Maas, *Nonlinear Microwave and RF Circuits*, Second Edition, Artech House, 2003.

[216] J. Vuolevi and T. Rahkonen, Distortion in RF Power Amplifiers, Artech House, 2003.

[217] A. Grebennikov, *RF and microwave power amplifier design*, McGraw-Hill, 2005

[218] A. Shirvani, *Design and Control of RF Power Amplifiers*, Kluwer Academic, 2003.

[219] P. B. Kenington, *High-Linearity RF Amplifier Design*, Artech House, 2000.

[220] J. Hu, K.G. Gard, N.B. Carvalho and M.B. Steer, "Time-frequency characterization of long-term memory in nonlinear power amplifiers," *2008 IEEE MTT-S Int. Microwave Symp.* June 2008, pp. 269–272.

[221] A. Grebennikov, *Switchmode RF Power Amplifiers*, Elsevier/Newnes 2007.

[222] S. Cripps, *RF Power Amplifiers for Wireless Communications*, Artech House, 1999.

[223] S. Cripps, *Advanced techniques in RF Power Amplifiers Design*, Artech House, 2002.

[224] P. Reynaert, *RF Power Amplifiers for Mobile Communications*, Springer, 2006.

[225] K. M. Gharaibeh and M. B. Steer, "Characterization of cross modulation in multichannel amplifiers using a statistically based behavioral modeling technique," *IEEE Trans. Microwave Theory and Tech.*, Vol. 51, Iss. 12, Dec. 2003 pp. 2434 – 2444.

[226] M. B. Steer, J. W. Bandler and C. M. Snowden, "Computer-aided design of RF and microwave circuits and systems," *IEEE Trans. Microwave Theory Techniques,* Vol. 50, Iss. 3, Mar. 2002, pp. 996–1005,

[227] M. B. Steer, J. Sevic and B. Geller, Special issue on RF power amplifier, *IEEE Trans. Microwave Theory and Techniques,* Vol. 49, Iss. 6, June 2001.

[228] H. Gutierrez, K. Gard, and M. B. Steer, "Nonlinear gain compression in microwave amplifiers using generalized power series and transformation of input statistics," *IEEE Trans. Microwave Theory Techniques,* Vol. 48, Iss. 11, Nov. 2000, pp. 1774–1777.

[229] G.W. Rhyne and M. B. Steer, "Generalized power series analysis of intermodulation distortion in a MESFET amplifier: simulation and experiment," *IEEE Trans. Microwave Theory Techniques,* Vol. 35, Iss. 12, Dec. 1987, pp. 1248–1255.

[230] K. Gard, H. Gutierrez and M. B. Steer, "Characterization of spectral regrowth in microwave amplifiers based on the nonlinear transformation of a complex gaussian process," *IEEE Trans. Microwave Theory Techniques,* Vol. 47, Iss. 7, July 1999, pp. 1059–1069.

[231] M.-S. Gupta, "Large-signal equivalent circuit for IMPATT-diode characterization and its application to amplifiers," *IEEE Trans. Microwave Theory and Techniques*, Vol. 21, Iss. 11, Nov. 1973, pp. 689–694.

[232] M.-S. Gupta, R.J. Lomax, and G.I. Haddad, "Noise considerations in self-mixing IMPATT-diode oscillators for short-range doppler RADAR applications," *IEEE Trans. Microwave Theory and Techniques*, Vol. 22, Iss. 1, Jan. 1974, pp. 37–43.

[233] W.J. Evans and G.I Haddad, "A large signal analysis of IMPATT diodes," *IEEE Trans. Electron Devices*, Vol. 15, Iss. 10, Oct. 1968, pp. 708–717.

[234] K.K. Ng, *Complete Guide to Semiconductor Devices*, IEEE Press and Wiley Interscience, 2002.

[235] K. Kurokawa and F.M. Magalhaes, "An X-band 10-watt multiple-IMPATT oscillator," *Proceedings of the IEEE* Vol. 59, Iss. 1, Jan. 1971, pp. 102–10.

[236] A.A. Bergh and P.J. Dean, "Light-emitting diodes," *Proceedings of the IEEE*, Vol. 60, Iss. 2, Feb. 1972, pp. 156–223.

[237] L. Esaki, "Discovery of the tunnel diode," *IEEE Trans. Electron Devices*, Vol. 23, Iss. 7, July 1976, pp. 644–647.

[238] B.C. DeLoach Jr. and D.L. Scharfetter, "Device physics of TRAPATT oscillators," *IEEE Trans. Electron Devices*, Vol. 17, Iss. 1, Jan. 1970, pp. 9–21.

[239] B. Gilbert, "A precise four-quadrant multiplier with subnanosecond response," *IEEE J. Solid-State Circuits*, Vol. 3, Iss. 4, Dec. 1968, pp. 365–373.

[240] L.P.B. Katehi and D. Peroulis, "RF MEMS components: switches and varactors," Chapter 2 in *Multifunctional Adaptive Microwave Circuits and Systems*, Edited by M. Steer and W.D. Palmer, SciTech Publishing, Inc., 2008.

[241] L.P.B. Katehi and D. Peroulis, "RF MEMS for reconfigurable circuits and antennas," Chapter 3 in *Multifunctional Adaptive Microwave Circuits and Systems*, Edited by M. Steer and W.D. Palmer, SciTech Publishing, Inc., 2008.

[242] G.M. Rebeiz, *RF MEMS: Theory, Design, and Technology*, Wiley, 2003.

[243] E.A. Faulkner, *Introduction to the Theory of Linear Systems*, Chapman & Hall, 1969.

[244] A.B. Pippard, *Response and Stability*, Cambridge University Press, 1985.

[245] M. Odyniec (editor), *RF and Microwave Oscillator Design*, Artech House, 2002.

[246] U.L. Rohde, A.K. Poddar, and G. Bock, *The Design of Modern Microwave Oscillators for Wireless Applications*, Wiley, 2005.

[247] L.D. Leeson, "A simple model of feedback oscillator noise spectrum," Proc. IEEE, Vol. 54, Iss. 2, 1966, pp. 329–330.

[248] P. Kinget, *Integrated GHz Voltage Controlled Oscillators*, Norwell, MA: Kluwer, 1999, pp. 353–381

[249] A. Victor and M. B. Steer, "Reflection coefficient shaping of a 5 GHz voltage tuned oscillator for improved tuning," *IEEE Trans. Microwave Theory and Tech.*, Vol. 55, Iss. 12, Dec. 2007, pp. 2488–2494.

[250] S.-S. Myoung and J.-G. Yook, "Low-phase-noise high-efficiency MMIC VCO based on InGaP/GaAs HBT with the LC filter," *Microwave and Optical Technology Letters*, Vol. 44, Iss. 2, Jan 20 2005, pp. 123–126.

[251] C.-H. Lee, S. Han, B. Matinpour, and J. Laskar, "Low phase noise X-band MMIC GaAs MESFET VCO," *IEEE Microwave and Guided Wave Letters*, Vol. 10, Iss. 8, Aug. 2000, pp. 325–327.

[252] Z.Q. Cheng, Y. Cai, J. Liu, Y. Zhou, K.M. Lau, and K.J. Chen, "A low phase-noise X-band MMIC VCO using high-linearity and low-noise composite-channel $Al_{0.3}Ga_{0.7}N/Al_{0.05}Ga_{0.95}N/GaN$ HEMTs," *IEEE Trans. Microwave Theory Techn.*, Vol. 55, Iss. 1, Jan. 2007, pp. 23–28.

[253] H. Zirath, R. Kozhuharov, and M. Ferndahl, "Balanced Colpitt oscillator MMICs designed for ultra-low phase noise," *IEEE J. Solid-State Circuits*, Vol. 40, Iss. 10, Oct. 2005, pp. 2077–2086.

[254] Y.-K. Chu and H.-R. Chuang, "A fully integrated 5.8-GHz U-NII band 0.18-μm CMOS VCO," *IEEE Microwave Wireless Compon Lett.*, Vol. 13, Iss. 7, July 2003, pp. 287–289.

[255] C.C. Meng, Y.W. Chang, and S.C. Tseng, "4.9-GHz low-phase-noise transformer-based superharmonic-coupled GaInP/GaAs HBT QVCO," *IEEE Microwave and Wireless Components Letters*, Vol. 16, Iss. 6, June, 2006, pp. 339–341.

[256] C. C. Meng, C. H. Chen, Y. W. Chang, and G. W. Huang, "5.4 GHz–127 dBc/Hz at 1 MHz GaInP/GaAs HBT quadrature VCO using stacked transformer," *Electron. Lett.*, Vol. 41, Iss. 16, Aug. 2005, pp. 906–908.

[257] T. M. Hancock and G. Rebeiz, "A novel superharmonic coupling topology for quadrature oscillator design at 6 GHz," in *Proc. IEEE Radio Freq. Integr. Circuit Symp.*, June 2004, pp. 285–288.

[258] S.-W. Yoon, S. Pinel, J. Laskar, "A 0.35-μm CMOS 2-GHz VCO in wafer-level package," *IEEE Microwave and Wireless Components Letters*, Vol. 15, Iss. 4, April 2005, pp. 229–231.

[259] J.-H. Yoon, S.-H. Lee, A-R. Koh, Kennedy, P. Gary, N.-Y. Kim, "Optimized phase noise of LC VCO using an asymmetric-inductance tank in InGaP/GaAs HBT technology," *Microwave and Optical Technology Letters*, Vol. 48, Iss. 6, June 2006, p 1035–1040.

[260] O.T. Esame, I. Tekin, A. Bozkurt, and Y. Gurbuz, "Design of a 4.2–5.4 GHz differential LC VCO using 0.35 μm SiGe BiCMOS technology for IEEE 802.11a applications," *Int. J. of RF and Microwave Computer-Aided Engineering*, Vol. 17, Iss. 2, March, 2007, pp. 243–251.

[261] A.P.M. Maas, and F.E. van Vliet, "A low-noise X-band microstrip VCO with 2.5 GHz tuning range using GaN-on-SiC p-HEMT," *GAAS 2005 Conf. Proceedings — 13th European Gallium Arsenide and other Compound Semiconductors Application Symp.*, Oct. 2005, pp. 257–260.

[262] J.-H. Yoon, S.-H. Lee, A.-R. Koh, B. Shrestha, S.-H. Cheon, G.P. Kennedy, and N.-Y. Kim, "A novel harmonic noise frequency filtering VCO for optimizing phase noise," *2006 IEEE MTT-S Int. Microwave Symp. Dig.*, June 2006, pp. 1805–1808.

[263] C.C. Meng, S.C. Tseng, Y.W. Chang, J.Y. Su, and G.W. Huang, "4-GHz low-phase-noise transformer-based top-series GaInP/GaAs HBT QVCO," *2006 IEEE MTT-S Int. Microwave Symp. Dig.*, June 2006, pp. 1809–1812.

[264] P. Andreani and X. Wang "On the phase-noise and phase-error performances of multiphase LC CMOS VCOs," *IEEE J. of Solid-State Circuits*, Vol. 39, Iss. 11, Nov. 2004, pp. 1883–1893.

[265] P. Vancorenland and M.S.J. Steyaert, "A 1.57-GHz fully integrated very low-phase-noise quadrature VCO," *IEEE J. of Solid-State Circuits*, Vol. 37, Iss. 5, May 2002, pp. 653–656.

[266] S.L.J. Gierkink, S. Levantino, R.C. Frye and V. Boccuzzi, "A lowphasenoise 5GHz quadrature CMOS VCO using commonmode inductive coupling," *Proceedings of the 28th European Solid-State Circuits Conf.*, 24–26 Sept. 2002, pp. 539–542.

[267] K. Kurokawa, "Some basic characteristics of broadband negative resistance oscillator circuits", *The*

Bell System Technical J., Vol. 48, Iss. 6, July-Aug. 1969, pp.1937–1955.

[268] C. Rauscher, "Large-signal technique for designing single-frequency and voltage-controlled GaAs FET oscillators," *IEEE Trans. Microwave Theory and Techn.*, Vol. 29, Iss.4, April 1981, pp. 293–304.

[269] R.W. Jackson, "Criteria for the onset of oscillation in microwave circuits," *IEEE Trans. On Microwave Theory and Techniques*, Vol. 40, Iss.3, March 1992, pp. 566–569.

[270] D.J.H. Maclean, *Evaluating Feedback in Amplifiers and Oscillators: Theory, Design and Analogue Applications*, Research Studies Press LTD., Baldock, Hertfordshire, England, 2004.

[271] J.W. Boyles, "The oscillator as a reflection amplifier: an intuitive approach to oscillator design," *Microwave J.*, Vol.29, Iss.6, June 1986. pp. 83–98.

[272] D. Esdale, and M. Howes, "A reflection coefficient approach to the design of one-port negative resistance oscillators," *IEEE Trans. Microwave Theory Techn.*, Vol. 29, Iss. 8, Aug. 1981, pp. 770–776.

[273] P.J. Topham, A. Dearn, and G. Parkinson, "GaAs bipolar wideband oscillators," *IEE Colloquium on Characterization of Oscillators Design and Measurement*, Feb. 3, 1992, pp. 2/1–2/4.

[274] G. Gonzalez, and O.J. Sosa, "On the design of a series-feedback network in a transistor negative-resistance oscillator," *IEEE Trans. Microwave Theory and Techn.*, Vol. 47, Iss. 1, Jan. 1999, pp. 42–47.

[275] A.P. Knights and M.J. Kelly, "Laterally stacked varactor formed by ion implantation," *Electronics Lett.*, Vol. 35, Iss. 10, May 13 1999, pp. 846–847.

[276] G. Gonzalez, *Foundations of Oscillator Circuit Design*, Artech House, 2007.

[277] R. Gilmore and L. Besser, *Practical RF Circuit Design for Modern Wireless Systems*, Artech House, 2003.

[278] S.A. Maas, *Microwave Mixers*, Artech House, 1986.

[279] F. Giannini, Nonlinear microwave circuit design, Wiley, 2004

[280] W. M. Fathelbab and M. B. Steer, "Tunable filters," Chapter 9 in *Multifunctional Adaptive Microwave Circuits and Systems*, Edited by M. B. Steer and W. D. Palmer, 2008.

[281] W. M. Fathelbab and M. B. Steer, "Parallel-coupled line filters with enhanced stopband performance," *IEEE Trans. Microwave Theory and Tech.,* Vol. 53, Iss. 12, Dec. 2005, pp. 3774 – 3781.

[282] W. Fathelbab and M. B. Steer, "Filter prototypes comprising singlet and/or inline sections," *Int. J. of Electronics*, Vol. 94, Iss. 9, Oct. 2007, pp. 925–934.

[283] S.B. Cohn, "Parallel-coupled transmission-line resonator filters," *IRE Trans. Microwave Theory Tech.*, Vol. 6, Iss. 4, Apr. 1958, pp. 223–231.

[284] B.J. Minnis, "Printed circuit coupled-line filters for bandwidths up to and greater than an octave," *IEEE Trans. Microwave Theory Tech.*, Vol. 29, Iss. 3, Mar. 1981, pp. 215–222.

[285] E.G. Cristal, and S. Frankel, "Hairpin-line and hybrid hairpin-line/half-wave parallel-coupled-line filters," *IEEE Trans. Microwave Theory Tech.*, Vol. 20, Iss. 11, Nov. 1972, pp. 719–728.

[286] U.H. Gysel, "New theory and design for hairpin-line filters," *IEEE Trans. Microwave Theory Tech.*, Vol. 22, Iss. 5, May 1974, pp. 523–531.

[287] W. M. Fathelbab and M. B. Steer, "Tapped marchand baluns for matching applications," *IEEE Trans. Microwave Theory and Tech,* Vol. 54, Iss. 6, June 2006, pp. 2543–2551.

[288] J. Nath, D. Ghosh, J.-P. Maria, A. I. Kingon, W. Fathelbab, P. D. Franzon and M. B. Steer, "An electronically-tunable microstrip bandpass filter using thin-film Barium Strontium Titanate (BST) varactors," *IEEE Trans. Microwave Theory and Tech.,* Vol. 53, Iss. 9, Sep. 2005, pp. 2707–2712.

[289] W. M. Fathelbab and M. B. Steer, "New classes of miniaturized planar Marchand baluns," *IEEE Trans. Microwave Theory and Tech.,* Vol. 53, Iss. 4, April 2005, pp. 1211–1220.

[290] W. M. Fathelbab and M. B. Steer, "A re-configurable bandpass filter for RF/microwave multifunctional systems," *IEEE Trans. Microwave Theory and Tech.,* Vol. 53, Iss. 3, Part 2, March 2005, pp. 1111–1116.

[291] W. M. Fathelbab and M. B. Steer, "Four-port microwave networks with intrinsic broadband suppression of common-mode signals," *IEEE Trans. Microwave Theory and Tech,* Vol. 53, Iss. 5, May 2005, pp. 1569–1575.

[292] W. Fathelbab and M. B. Steer, "Design of bandstop filters utilizing circuit prototypes," *IET Microwave, Antennas and Propagation*, Vol. 1, Iss. 2, April 2007, pp 523–526.

[293] J. A. G. Malherbe, "Realization of elliptic function bandstop filters by means of resonated prototype," *IEEE Trans. Microwave Theory Tech.*, Vol. 25, Iss. 8, Aug. 1977, pp. 717.

[294] M. Horton and R. Wenzel, "General theory and design of optimum quarter-wave TEM filters," *IEEE Trans. Microwave Theory Tech.*, Vol. 13, Iss. 5, May 1965, pp. 316–327.

[295] H. J. Orchard and G. C. Temes, "Filter design using transformed variable," *IEEE Trans. Circuit Theory*, Vol. 15, Iss. 12, Dec 1968, pp. 385–408.

[296] R. J. Wenzel, "Synthesis of combline and capacitively loaded interdigital bandpass filters of arbitrary bandwidth," *IEEE Trans. Microwave Theory Tech.*, Vol. 19, Iss. 8, Aug. 1971, pp. 678–686.

[297] G. Matthaei, L. Young and E.M.T. Jones, *Microwave Filters Impedance-Matching Networks, and Coupling Structures,* Artech House, 1980.

[298] H. Ozaki and J. Ishii, "Synthesis of a class of strip-line filters," *IRE Trans. Circuit Theory*, Vol. 5, June 1958, pp. 104–109.

[299] R. Levy and L.F. Lind, "Synthesis of symmetrical branch-guide directional couplers," *IEEE Trans. Microwave Theory Tech.*, Vol. 16, Iss. 2, Feb. 1968, pp. 80–89.

[300] B.J. Minnis, "Decade bandwidth bias Ts for MIC applications up to 50 GHz," *IEEE Trans. Microwave Theory Tech.*, Vol. 35, Iss. 6, June 1987, pp. 597–600.

[301] R. J. Wenzel, "Application of exact synthesis methods to multi-channel filter design," *IEEE Trans. Microwave Theory Tech.*, Vol. 13, Iss. 1, Jan. 1965, pp. 5–15.

[302] R. Levy, "Synthesis of general asymmetric singly- and doubly-terminated cross-coupled filters," *IEEE Trans. Microwave Theory Tech.*, Vol. 42, Iss. 12, Dec. 1994, pp. 2468–2471.

[303] B. J. Minnis, *Designing Microwave Circuits by Exact Synthesis*, Artech House, 1996.

[304] H. J. Carlin and P. P. Civalleri, *Wideband Circuit Design. CRC Press LCC*, 1998.

[305] E. Cristal, "Tapped-line coupled transmission lines with applications to interdigital and Combline filters," *IEEE Trans. Microwave Theory Tech.*, Vol. 23, Iss. 12, Dec. 1975, pp. 1007–1012.

[306] S. Ponnusamy and H. Silverman, "Bilinear transformations and mappings," Chapter 8 in *Complex Variables with Applications*, C Birkhuser Boston, pp. 61–90, 2006.

[307] Standard Reference Data database, U.S. National Institute of Standards and Technology and References, http://www.nist.gov/srd .

[308] International Council for Science : Committee on Data for Science and Technology, see http://www.codata.org/resources/databases/index.html .

[309] J.D. Cox, D.D. Wagman, and V.A. Medvedev, *CODATA Key Values for Thermodynamics*, Hemisphere Publishing Corp., New York, 1989.

[310] G. Elert, The Physics Hypertextbook, 2007, http://hypertextbook.com .

[311] Reference Data for Radio Engineers, Second Edition, Federal Telephone and Radio Corporation, 1946, available at http://www.pmillett.com/Books/FTR_ref_data.pdf .

[312] S. Uma, A.D. McConnell, M. Asheghi, K. Kurabayashi and K.E. Goodson, "Temperature-dependent thermal conductivity of undoped polycrystalline silicon layers," *Int. J. Thermal Physics*, Vol. 22, Iss. 2, March 2001, pp. 605–616.

[313] M. Born and E. Wolf, "Basic properties of the electromagnetic field," Chapter 1 in *Principles of Optics: Electromagnetic Theory of Propagation, Interference, and Diffraction of Light*, 7th ed., Cambridge University Press, 1999.

[314] C.C. Enz and E.A. Vittoz, *Charge-Based MOS Transistor Modeling: the EKV Model for Low-Power and RF IC Design*, Wiley, 2006.

[315] W. Grabinski, B. Nauwelaers and D. Schreurs, (editors) *Transistor Level Modeling for Analog/RF IC Design*, Springer, 2006.

[316] P. Yang, B.D. Epler, and P.K. Chatterjee, "An investigation of the charge conservation problem for MOSFET circuit simulation," *IEEE J. Solid-State Circuits*, Vol. 18, Iss. 1, Feb. 1983, pp. 128–138.

[317] D.P. Foty, *MOSFET Modeling With SPICE: Principles and Practice*, Prentice-Hall, 1997.

[318] W. Liu, *Mosfet Models for Spice Simulation, Including BSIM3v3 and BSIM4*, John Wiley & Sons, 2001.

[319] I. Angelov, H. Zirath, and N. Rosman, "A new empirical nonlinear model for HEMT and MESFET

devices," *IEEE Trans Microwave Theory and Techniques*, Vol. 40, Iss. 12, Dec. 1992, pp. 2258–2266.

[320]   R.S. Pengelly, *Microwave Field-Effect Transistors: Theory, Design, and Applications*, 3rd ed., Noble, 1994.

[321]   F. Schwierz and J.J. Liou, *Modern Microwave Transistors: Theory, Design, and Performance*, Wiley, 2003.

[322]   M. Rudolph, *Introduction to Modeling HBTs*, Artech House, 2006.

[323]   J.S. Yuan, *SiGe, GaAs, and InP Heterojunction Biopolar Transistors*, Wiley, 1999.

Mathematical Identities and Relationships

A.1   Complex Numbers and Phasors

A complex number z is an ordered pair of two real numbers x and y, and has special meaning when working in the frequency domain. The rectangular form of the complex number is $z = x + \jmath y$, where $\jmath = \sqrt{-1}$ is an imaginary number.[1]. The polar form of the complex number $z = r \angle \varphi$ relates directly to the amplitude and phase of a cosinusoid. In determining phase, reference is made to a cosinusoid of the form $u(t) = r \cos(\omega t + \varphi)$, as this has a value of r when the radian frequency $\omega = 0$ (i.e., at DC). The cosinusoid is shown in Figure A-1(a), where r is the amplitude and φ is the phase of the cosinusoid. The cosinusoid can be presented as the complex valued function

$$f(t) = z e^{\jmath \omega t}, \tag{A.1}$$

where

$$z = r e^{\jmath \varphi} \tag{A.2}$$

---

[1] Outside electrical engineering i is used for $\sqrt{-1}$, but this is replaced by $\jmath$ in electrical engineering to avoid confusion with the usage of i for current.

(a)   (b)
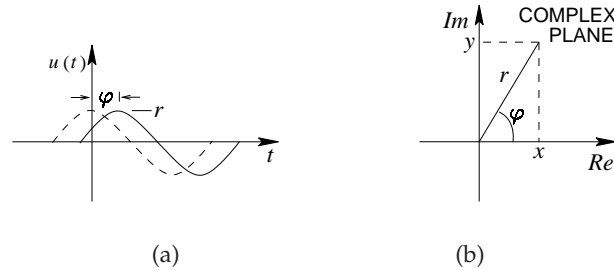
Figure A-1 Sinusoidal signal representation: (a) waveform referenced to the zero-phase cosinusoid shown as the dashed waveform; and (b) representation as a phasor on the complex plane where Re is the real part and Im is the imaginary part.

is called the phasor of $u(t)$. The phasor contains the amplitude and phase of the cosinusoid, and given the frequency of the cosinusoid, the original time-domain waveform can be reconstructed. The time-domain form of the complex function is

$$u(t) = \Re\{f(t)\} = r\cos(\omega t + \varphi). \tag{A.3}$$

$\Re\{\}$ indicates that the real part of a complex number is taken, and similarly $\Im\{\}$ indicates that the imaginary part of a complex number is taken. The complex number can be represented in rectangular or magnitude-phase form so that

$$z = x + \jmath y = r\angle\varphi = re^{\jmath\varphi} \tag{A.4}$$

with the relationship shown in Figure A-1(b). Note that in Equation (A.4), φ must be expressed in radians and whenever it is used in calculations. Using trigonometry,

$$x = r\cos(\varphi) \quad \text{and} \quad y = r\sin(\varphi). \tag{A.5}$$

Basic calculations using the complex numbers

$$w = a + \jmath b = q\angle\theta \quad \text{and} \quad z = c + \jmath d = r\angle\varphi \tag{A.6}$$

(with a, b, c, d, θ, φ, r and q being real numbers) are

- Addition: $w + z = (a + \jmath b) + (c + \jmath d) = (a + c) + \jmath(b + d)$

- Subtraction: $w - z = (a + \jmath b) - (c + \jmath d) = (a - c) + \jmath(b - d)$

- Multiplication: $w \cdot z = (a + \jmath b) \cdot (c + \jmath d) = (ac - bd) + \jmath(bc + ad)$

- Multiplication (alternative): $w \cdot z = qr\angle(\theta + \varphi)$

- Division: $w/z = (a + \jmath b)/(c + \jmath d) = \left(\frac{ac+bd}{c^2+d^2}\right) + \jmath\left(\frac{bc-ad}{c^2+d^2}\right)$

- Division (alternative): $w/z = q/r\angle(\theta - \varphi)$

- Square root: $\sqrt{w} = \sqrt{q}\angle(\theta/2)$

- nth root: $\sqrt[n]{w} = \sqrt[n]{q}\angle(\theta/n)$

- Negative: $-w = -q\angle\theta = q\angle(\theta + \pi)$

- Complex conjugate: $w^* = q\angle(-\theta)$

- Magnitude: $|w| = \sqrt{a^2 + b^2} = q^2$

- Angle: $\theta = \arctan(b/a)$ (correction may be required to get the phase in the right quadrant)

- Squared magnitude: $|z|^2 = z \cdot z^*$

- $\jmath \cdot \jmath = \jmath^2 = -1$

EXAMPLE A. 1   Complex numbers

Consider the complex numbers $w = -0.5 - \jmath 1.6$ and $z = 5 - \jmath 3$, and the waveform $v(t) = 2.3\cos(2\pi 10 - 1.2)$.

- In polar form, $w = 1.676\angle -1.874$, where 1.874 is in radians; alternatively $w = 1.676\angle -107.4°$

- In polar form, $z = 5.831\angle -0.54$

- $2w = -1 - \jmath 3.2 = 3.353\angle -1.874$

- $z \cdot w = w \cdot z = w \cdot z = (1.676 \cdot 5.831)\angle(-1.874 - 0.54) = 9.774\angle -2.414 = -7.3 - \jmath 6.5$

- $\sqrt{z} = 2.327 - \jmath 0.645$

- Alternatively $\sqrt{z} = \sqrt{5.831}\angle(-0.54/2) = 2.415\angle -0.27$

- $w/z = (1.676/5.831)\angle[-1.874 - (-0.54)] = 0.287\angle -1.333 = 0.068 - \jmath 6.5$

- The radian frequency of $v(t)$ is 20π rads/s

- The frequency of $v(t)$ is 10 Hz

- The phasor of $v(t)$ is $v = 2.3\angle -1.2 = 2.3\angle -68.755° = 0.833 - \jmath 2.144$

- If w is a phasor and its frequency is 1 GHz, the waveform equivalent of w is $w = 1.676\cos(2\pi \cdot 10^9 t - 1.874)$ and the phase is -1.874 radians or $-107.35°$ or $252.65°$

A.2   Del (∇) Vector Operator

A.2.1   Operations in Cartesian, Cylindrical, and Spherical Coordinate Systems

This section defines the del, or the nabla ∇, a calculus operator that operates on vector quantities and also on a scalar to create a vector. The various del operators in the Cartesian (rectangular), spherical, and cylindrical

coordinate systems are presented. Here a scalar is denoted by f and a vector by $\mathbf{F}$:

- grad, ∇, is the gradient of a scalar with direction.

- div, $\nabla\cdot$, is a measure of how a vector is spreading out from a point.

- curl, $\nabla\times$, is a measure of the rotation of a vector.

Cartesian (rectangular) coordinates (x, y, z):
$\hat{x}$ is the unit vector in the x direction and F_x is the component of the field in the x-direction. Similarly for the other components.

$$\mathbf{F} = F_x\hat{x} + F_y\hat{y} + F_z\hat{z} \tag{A.7}$$

$$\text{grad } \nabla f = \frac{\partial f}{\partial x}\hat{x} + \frac{\partial f}{\partial y}\hat{y} + \frac{\partial f}{\partial z}\hat{z} \tag{A.8}$$

$$\text{div } \nabla \cdot \mathbf{F} = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z} \tag{A.9}$$

$$\text{curl } \nabla \times \mathbf{F} = \left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z}\right)\hat{x} + \left(\frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x}\right)\hat{y}$$

$$+ \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right)\hat{z} \tag{A.10}$$

$$\nabla^2\mathbf{F} = \frac{\partial^2\mathbf{F}}{\partial x^2} + \frac{\partial^2\mathbf{F}}{\partial y^2} + \frac{\partial^2\mathbf{F}}{\partial z^2} \tag{A.11}$$

$$\text{differential length, } d\ell = dx\,\hat{x} + dy\,\hat{y} + dz\,\hat{z} \tag{A.12}$$

$$\text{differential normal area, } d\mathbf{s} = dy\,dz\,\hat{x} + dx\,dz\,\hat{y} + dx\,dy\,\hat{z} \tag{A.13}$$

$$\text{differential volume, } d\mathbf{v} = dx\,dy\,dz \tag{A.14}$$

Cylindrical coordinates (ρ, ϕ, z):
The cylindrical coordinates are ρ for radius, ϕ for angle, and z for height. These are related to the Cartesian coordinates by

$$x = \rho\cos\phi\,, \quad y = \rho\sin\phi\,, \quad z = z \tag{A.15}$$

$$\rho = \sqrt{x^2 + y^2}\,, \quad \phi = \text{atan2}(y, x)\,, \quad z = z \tag{A.16}$$

$$\text{vector } \mathbf{F} = F_\rho\hat{\rho} + F_\phi\hat{\phi} + F_z\hat{z} \tag{A.17}$$

$$\text{grad } \nabla f = \frac{\partial f}{\partial\rho}\hat{\rho} + \frac{1}{\rho}\frac{\partial f}{\partial\phi}\hat{\phi} + \frac{\partial f}{\partial z}\hat{z} \tag{A.18}$$

$$\text{div } \nabla \cdot \mathbf{F} = \frac{1}{\rho}\frac{\partial\rho F_\rho}{\partial\rho} + \frac{1}{\rho}\frac{\partial F_\phi}{\partial\phi} + \frac{\partial F_z}{\partial z} \tag{A.19}$$

$$\text{curl } \nabla \times \mathbf{F} = \left(\frac{1}{\rho}\frac{\partial F_z}{\partial\phi} - \frac{\partial F_\phi}{\partial z}\right)\hat{\rho} + \left(\frac{\partial F_\rho}{\partial z} - \frac{\partial F_z}{\partial\rho}\right)\hat{\phi}$$

$$+ \frac{1}{\rho}\left(\frac{\partial(\rho F_\phi)}{\partial\rho} - \frac{\partial F_\rho}{\partial\phi}\right)\hat{z} \tag{A.20}$$

$$\text{differential length, } d\ell = d\rho \, \hat{\rho} + \rho d\phi \, \hat{\phi} + dz \, \hat{z} \qquad (A.21)$$

$$\text{differential normal area, } d\mathbf{s} = \rho d\phi \, dz \, \hat{\rho} + d\rho \, dz \, \hat{\phi} + \rho d\rho \, d\phi \, \hat{z} \qquad (A.22)$$

$$\text{differential volume, } d\mathbf{v} = \rho d\rho \, d\phi \, dz \qquad (A.23)$$

Spherical coordinates (ρ, θ, ϕ):

$$x = r \sin\theta \cos\phi \,, \quad y = r \sin\theta \sin\phi \,, \quad y = r \cos\theta \qquad (A.24)$$

$$r = \sqrt{x^2 + y^2 + z^2} \,, \quad \theta = \arccos(z/r) \,, \quad \phi = \mathrm{atan2}(y, x) \qquad (A.25)$$

$$\text{vector } \mathbf{F} = F_r \hat{r} + F_\theta \hat{\theta} + F_\phi \hat{\phi} \qquad (A.26)$$

$$\text{grad } \nabla f = \frac{\partial f}{\partial r} \hat{r} + \frac{1}{r} \frac{\partial f}{\partial \theta} \hat{\theta} + \frac{1}{r \sin\theta} \frac{\partial f}{\partial \phi} \hat{\phi} \qquad (A.27)$$

$$\text{div } \nabla \cdot \mathbf{F} = \frac{1}{r^2} \frac{\partial (r^2 F_r)}{\partial r} + \frac{1}{r \sin\theta} \frac{\partial (F_\theta \sin\theta)}{\partial \theta}$$
$$+ \frac{1}{r \sin\theta} \frac{\partial F_\phi}{\partial \phi} \qquad (A.28)$$

$$\text{curl } \nabla \times \mathbf{F} = \frac{1}{r \sin\theta} \left(\frac{\partial (F_\phi \sin\theta)}{\partial \theta} - \frac{\partial F_\theta}{\partial \phi} \right) \hat{r}$$
$$+ \frac{1}{r} \left(\frac{1}{\sin\theta} \frac{\partial F_r}{\partial \phi} - \frac{\partial (r F_\phi)}{\partial r} \right) \hat{\theta}$$
$$+ \frac{1}{r} \left(\frac{\partial (r F_\theta)}{\partial r} - \frac{\partial F_r}{\partial \theta} \right) \hat{\phi} \qquad (A.29)$$

$$\text{differential length } d\ell = dr \, \hat{r} + r \, d\theta \, \hat{\theta} + r \sin\theta d\phi \, \hat{\phi} \qquad (A.30)$$

$$\text{differential area } d\mathbf{s} = r^2 \sin\theta \, d\theta \, d\phi \, \hat{r} + r \sin\theta \, dr \, d\phi \, \hat{\theta}$$
$$+ r \, dr \, d\theta \, \hat{\phi} \qquad (A.31)$$

$$\text{differential volume } d\mathbf{v} = r^2 \sin\theta \, dr \, d\theta \, d\phi \qquad (A.32)$$

A.2.2 Identities

Identities for the del, ∇, operator:

$$\text{div grad } f = \nabla \cdot (\nabla f) = \nabla^2 f = \triangle f (\text{Laplacian}) \qquad (A.33)$$

$$\triangle(fg) = f \triangle g + 2 \nabla f \cdot \nabla g + g \triangle f \qquad (A.34)$$

$$\text{curl grad } f = \nabla \times (\nabla f) = 0 \qquad (A.35)$$

$$\text{div curl } \mathbf{F} = \nabla \cdot (\nabla \times \mathbf{F}) = 0 \qquad (A.36)$$

$$\text{curl curl } \mathbf{F} = \nabla \times (\nabla \times \mathbf{F}) = \nabla(\nabla \cdot F) - \nabla^2 F \qquad (A.37)$$

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = \mathbf{B}(\mathbf{A} \cdot \mathbf{C}) - \mathbf{C}(\mathbf{A} \cdot \mathbf{B}) \,. \qquad (A.38)$$

A.3 Trigonometric Identities

$$\cos(-x) = \cos(x) \qquad \sin(-x) = -\sin(x) \qquad \tan(-x) = -\tan(x) \qquad (A.39)$$

$$\cos(x) = \tfrac{1}{2}\left(e^{\jmath x} + e^{-\jmath x}\right) \quad \sin(x) = \tfrac{1}{2}\jmath\left(e^{-\jmath x} - e^{\jmath x}\right) \qquad\qquad (A.40)$$

$$\sin x = 1/\csc x \qquad\qquad \cos x = 1/\sec x \qquad\qquad \tan x = 1/\cot x \qquad (A.41)$$
$$\tan x = \sin x / \cos x \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (A.42)$$

$$\sin^2 x + \cos^2 x = 1 \qquad 1 + \tan^2 x = \sec^2 x \qquad 1 + \cot^2 x = \csc^2 x \qquad (A.43)$$

$$\sin\left(\tfrac{\pi}{2} - x\right) = \cos x \qquad \cos\left(\tfrac{\pi}{2} - x\right) = \sin x \qquad \tan\left(\tfrac{\pi}{2} - x\right) = \cot x \qquad (A.44)$$
$$\csc\left(\tfrac{\pi}{2} - x\right) = \sec x \qquad \sec\left(\tfrac{\pi}{2} - x\right) = \csc x \qquad \cot\left(\tfrac{\pi}{2} - x\right) = \tan x \qquad (A.45)$$

$$\sin(-x) = -\sin x \qquad \cos(-x) = \cos x \qquad \tan(-x) = -\tan x \qquad (A.46)$$
$$\csc(-x) = -\csc x \qquad \sec(-x) = \sec x \qquad \cot(-x) = -\cot x \qquad (A.47)$$

$$\sin(x + y) = \sin x \cos y + \cos x \sin y \qquad\qquad\qquad\qquad\qquad\qquad (A.48)$$
$$\sin(x - y) = \sin x \cos y - \cos x \sin y \qquad\qquad\qquad\qquad\qquad\qquad (A.49)$$
$$\cos(x + y) = \cos x \cos y - \sin x \sin y \qquad\qquad\qquad\qquad\qquad\qquad (A.50)$$
$$\cos(x - y) = \cos x \cos y + \sin x \sin y \qquad\qquad\qquad\qquad\qquad\qquad (A.51)$$

$$\sin(2x) = 2\sin x \cos y \qquad \tan(2x) = \tfrac{2\tan x}{1 - \tan^2 x} \qquad \cos(2x) = \cos^2(x) - \sin^2(x) \quad (A.52)$$

$$\sin^2 x = \tfrac{1}{2}\left[1 - \cos(2x)\right] \quad \cos^2 x = \tfrac{1}{2}\left[1 + \cos(2x)\right] \quad \tan^2 x = \tfrac{1 - \cos(2x)}{1 + \cos(2x)} \qquad (A.53)$$

$$\sin x + \sin y = 2\sin\left(\tfrac{x+y}{2}\right)\cos\left(\tfrac{x-y}{2}\right) \qquad\qquad\qquad\qquad\qquad (A.54)$$
$$\sin x - \sin y = 2\cos\left(\tfrac{x+y}{2}\right)\sin\left(\tfrac{x-y}{2}\right) \qquad\qquad\qquad\qquad\qquad (A.55)$$
$$\cos x - \cos B = 2\cos\left(\tfrac{x+y}{2}\right)\cos\left(\tfrac{x-y}{2}\right) \qquad\qquad\qquad\qquad\qquad (A.56)$$

$$\sin x \sin y = \tfrac{1}{2}\left[\cos\left(x - y\right) - \cos\left(x + y\right)\right] \qquad\qquad\qquad\qquad\qquad (A.57)$$
$$\cos x \cos y = \tfrac{1}{2}\left[\cos\left(x - y\right) + \cos\left(x + y\right)\right] \qquad\qquad\qquad\qquad\qquad (A.58)$$
$$\sin x \cos y = \tfrac{1}{2}\left[\sin\left(x + y\right) + \sin\left(x - y\right)\right] \qquad\qquad\qquad\qquad\qquad (A.59)$$
$$\cos x \sin y = \tfrac{1}{2}\left[\sin\left(x + y\right) - \sin\left(x - y\right)\right] \qquad\qquad\qquad\qquad\qquad (A.60)$$

A.4 Hyperbolic Functions and Complex Numbers

$$\cosh(x) = \tfrac{1}{2}(e^x + e^{-x}) = \cos(\jmath x) \qquad \text{(A.61)}$$
$$\cosh(\jmath x) = \tfrac{1}{2}(e^{\jmath x} + e^{-\jmath x}) = \cos(x) \qquad \text{(A.62)}$$
$$\cosh(-x) = \cosh(x) \qquad \text{(A.63)}$$

$$\sinh(x) = \tfrac{1}{2}(e^x - e^{-x}) = -\jmath \sin(\jmath x) \qquad \text{(A.64)}$$
$$\sinh(\jmath x) = \tfrac{1}{2}(e^{\jmath x} - e^{-\jmath x}) = \jmath \sin(x) \qquad \text{(A.65)}$$
$$\sinh(-x) = -\sinh(x) \qquad \text{(A.66)}$$

$$\cosh^2(x) - \sinh^2(x) = 1 \qquad \text{(A.67)}$$

$$\tanh(x) = \sinh(x)/\cosh(x) \qquad \text{(A.68)}$$
$$\tanh(x) = -\jmath \tan(\jmath x) \qquad \text{(A.69)}$$
$$\tanh(\jmath x) = \jmath \tan(x) \qquad \text{(A.70)}$$
$$\tanh(-x) = -\tanh(x) \qquad \text{(A.71)}$$

$$\sin(\jmath x) = \jmath \sinh(x) \qquad \text{(A.72)}$$
$$\cos(\jmath x) = \cosh(x) \qquad \text{(A.73)}$$
$$\tan(\jmath x) = \jmath \tanh(x) \qquad \text{(A.74)}$$

$$e^x = \cosh(x) + \sinh(x) \qquad \text{(A.75)}$$
$$e^{-x} = \cosh(x) - \sinh(x) \qquad \text{(A.76)}$$
$$e^{\jmath x} = \cos(x) + \jmath \sin(x) \qquad \text{(A.77)}$$
$$e^{-\jmath x} = \cos(x) - \jmath \sin(x) \qquad \text{(A.78)}$$

$$e^{\gamma} = e^{(\alpha + \jmath \beta)} = e^{\alpha} e^{\jmath \beta}, \text{ where } \gamma = \alpha + \jmath \beta \qquad \text{(A.79)}$$

A.5 Volumes and Areas

	Circle	Sphere	Cylinder	Cone
	r = radius	r = radius	r = radius h = height	r = radius of base h = height
Area	πr^2	$4\pi r^2$	$2\pi rh$ (cylinder) $2\pi r^2$ (base & top)	
Volume		$4\pi r^3/3$	$\pi r^2 h$	$\pi r^2 h/3$

A.6 Interpolation

A.6.1 Linear Interpolation

Linear interpolation can be used to extract data from tables. In the table below there are two known points, (x_0, y_0) and (x_1, y_1), and the linear interpolant is the straight line between them. The unknown point (x, y) is found by locating it on this straight line.

Variable	
Independent	Dependent
x_0	y_0
x	y
x_1	y_1
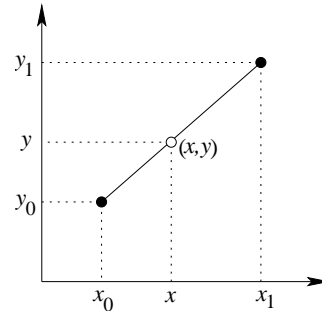
Thus

$$\frac{y - y_0}{y_1 - y_0} = \frac{x - x_0}{x_1 - x_0} \tag{A.80}$$

so that

$$y = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0}. \tag{A.81}$$

A.6.2 Bilinear Interpolation

Bilinear interpolation is a two-dimensional generalization of linear interpolation. In the table below, $f(\mathsf{D}) = f(x_0, y_0)$, $f(\mathsf{A}) = f(x_0, y_1)$, $f(\mathsf{F}) = f(x_1, y_0)$, and $f(\mathsf{C}) = f(x_1, y_1)$ are known and $f(x, y)$ is to be found. The bilinear interpolation technique is illustrated in the figure.

Variable		
Independent	Independent	Dependent
x_0	y_0	f_0
x	y	f
x_1	y_1	f_1
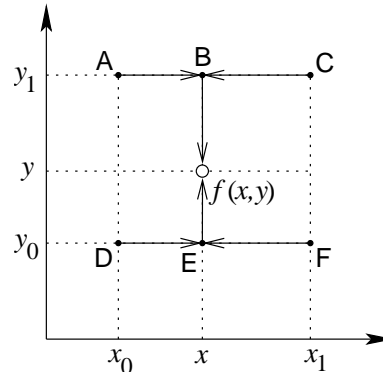
First, find the value of the function at point B from the values at Point A and Point C. Thus, from Equation (A.81),

$$f(B) = f(x, y_1) = f(A) + (x - x_0)\frac{f(C) - f(A)}{x_1 - x_0}. \qquad (A.82)$$

Similarly

$$f(E) = f(x, y_0) = f(D) + (x - x_0)\frac{f(F) - f(D)}{x_1 - x_0}. \qquad (A.83)$$

Linear interpolation between Point B and Point E yields

$$f(x, y) = f(E) + (y - y_0)\frac{f(B) - f(E)}{y_1 - y_0}. \qquad (A.84)$$

Combining these, the function obtained from bilinear interpolation is

$$\begin{aligned}
f(x, y) &= \frac{f(x_0, y_0)}{(x_1 - x_0)(y_1 - y_0)}(x_1 - x)(y_1 - y) \\
&+ \frac{f(x_1, y_0)}{(x_1 - x_0)(y_1 - y_0)}(x - x_0)(y_1 - y) \\
&+ \frac{f(x_0, y_1)}{(x_1 - x_0)(y_1 - y_0)}(x_1 - x)(y - y_0) \\
&+ \frac{f(x_1, y_1)}{(x_1 - x_0)(y_1 - y_0)}(x - x_0)(y - y_0). \qquad (A.85)
\end{aligned}$$

When interpolating from a table, choosing the points closest to the final point will yield greatest accuracy.

A.7 Circles on the Complex Plane

A common situation that occurs when working with reflection coefficients and complex numbers in general involves the magnitude of a complex number being equal to a real number. As will be shown, this defines a circle in the complex plane. Consider the relation

$$|S - c| = r, \qquad (A.86)$$

where $S = x + \jmath y$ and $c = a + \jmath b$ are complex numbers and r is a real number. Thus Equation (A.86) becomes

$$|(x - a) + \jmath(y - b)| = r; \qquad (A.87)$$

that is

$$[(x - a) + \jmath(y - b)]\,[(x - a) - \jmath(y - b)] = r^2 \qquad (A.88)$$

$$(x - a)^2 + (y - b)^2 = r^2. \qquad (A.89)$$

The above equation is a general equation for a circle with center (a, b) and radius r (see Figure A-2).
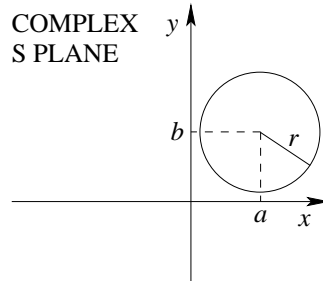
Figure A-2 A circle in the S plane defined by the relation $|S - c| = r$, where $c = a + \jmath b$ and $S = x + \jmath y$.

A.8 Bilinear Transform

The bilinear transformation is used in many domains and it has several special properties [306]. In RF and microwave engineering the special property that proves useful is that a bilinear transformation maps a circle in the complex plane onto another circle in the complex plane.

The bilinear transformation of a complex variable z is

$$w = \frac{Az + B}{Cz + 1}. \tag{A.90}$$

In RF and microwave usage z and w can be reflection coefficients or impedances. Rearranging Equation (A.90) results in

$$w = \frac{A}{C} + \frac{B - \frac{A}{C}}{Cz + 1} \tag{A.91}$$

which can be rearranged again as

$$\frac{w - \frac{A}{C}}{B - \frac{A}{C}} = \frac{1}{Cz + 1}. \tag{A.92}$$

The complexity of the derivation is reduced by introducing the intermediate variables W and Z:

$$W = \frac{w - \frac{A}{C}}{B - \frac{A}{C}} \tag{A.93}$$

and

$$Z = Cz + 1 \tag{A.94}$$

and noting that, from Equation (A.92),

$$W = 1/Z. \tag{A.95}$$

If, in Equation (A.94), the locus of z describes a circle, then the locus of Z will also be a circle since C is a constant. The center of the circle in Z-space

is C_Z and the distance from any point on the circle to the center will be the radius R_Z so that

$$|Z - C_Z| = R_Z. \tag{A.96}$$

Considering Equation (A.94), then in z space the center of the circle is

$$C_z = \frac{C_Z}{C} - 1 \tag{A.97}$$

and its radius is

$$R_z = R_Z/C. \tag{A.98}$$

Removing the magnitude signs in Equation (A.96) leads to

$$(Z - C_Z)(Z^* - C_Z^*) = R_Z^2. \tag{A.99}$$

or

$$ZZ^* - C_Z Z^* - C_Z^* Z + (|C_Z|^2 - R_Z^2) = 0. \tag{A.100}$$

which is the general equation for a circle in the complex plane.

Now, the circle in the Z plane will be related to what happens in the W plane. Substituting Equation (A.95) into Equation (A.100) results in

$$\frac{1}{WW^*} - C_Z \frac{1}{W^*} - C_Z^* \frac{1}{W} + (|C_Z|^2 - R_Z^2) = 0. \tag{A.101}$$

and rearranging,

$$WW^* - \frac{C_Z}{(|C_Z|^2 - R_Z^2)} W - \frac{C_Z^*}{(|C_Z|^2 - R_Z^2)} W^* + \frac{1}{(|C_Z|^2 - R_Z^2)} = 0. \tag{A.102}$$

This is in the same form as Equation (A.100), so Equation (A.102) describes a circle in W-space with center

$$C_W = \frac{C_Z^*}{(|C_Z|^2 - R_Z^2)} \tag{A.103}$$

and radius

$$R_W = \left| \frac{R_Z}{|C_Z|^2 - R_Z^2} \right|. \tag{A.104}$$

Now the locus of w will also be a circle; from Equation (A.93),

$$w = W \left(B - \frac{A}{C} \right) + \frac{A}{C} \tag{A.105}$$

so that W is scaled and a constant added. The center of the circle in w space is

$$C_w = C_W \left(B - \frac{A}{C} \right) + \frac{A}{C} = \frac{C_Z^* (B - A/C)}{(|C_Z|^2 - R_Z^2)} + \frac{A}{C} \tag{A.106}$$

and the radius is

$$R_w = R_W \left| B - \frac{A}{C} \right| = \left| \frac{(B - A/C)}{|C_Z|^2 - R_Z^2} \right|. \tag{A.107}$$

Substituting Equations (A.97) and (A.98) in the above relates the centers and radii in the w and z spaces:

$$C_w = \frac{(CC_z + 1)^*(B - A/C)}{\left(|CC_z + 1|^2 - |CR_z|^2 \right)} + \frac{A}{C} \tag{A.108}$$

$$R_w = \left| \frac{(B - A/C)CR_z}{|CC_z + 1|^2 - |CR_z|^2} \right| = \left| \frac{(BC - A)R_z}{|CC_z + 1|^2 - |CR_z|^2} \right|. \tag{A.109}$$

A.9 Quadratic Equation

A quadratic equation is a second-order polynomial equation. The general form of the quadratic equation in x is

$$ax^2 + bx + c = 0 \tag{A.110}$$

where $a \neq 0$. There are two solutions, called roots, given by the quadratic formula:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{A.111}$$

and $\pm$ indicates two possible roots:

$$x_+ = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{and} \quad x_- = \frac{-b - \sqrt{b^2 - 4ac}}{2a}. \tag{A.112}$$

Even if a, b, and c are real, the roots may be complex and they could be degenerate, that is, the same.

The factors of the quadratic equation come from the roots, that is,

$$ax^2 + bx + c = a\left(x - x_+ \right)\left(x - x_- \right). \tag{A.113}$$

A.10 Network Condensation

Network condensation is used in developing equivalent networks of large linear circuits with algebraic y parameters, that is, for resistive networks or general linear circuits in frequency-domain analysis. Network condensation is a numerical approach particular to circuits. It is of use in reduced-order modeling of linear circuits.

The indefinite nodal admittance formulation of a network with four external terminals (see Figure A-3) is

Figure A-3   A four terminal network.

$$
\begin{bmatrix}
y_{11} & y_{12} & \cdots & y_{1,k} & y_{1,k+1} \\
y_{21} & y_{22} & \cdots & y_{2,k} & y_{2,k+1} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
y_{k1} & y_{k2} & \cdots & y_{k,k} & y_{k,k+1} \\
y_{k+1,1} & y_{k+1,2} & \cdots & y_{k+1,k} & y_{k+1,k+1}
\end{bmatrix}
\begin{bmatrix}
v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ \vdots \\ v_k \\ v_{k+1}
\end{bmatrix}
=
\begin{bmatrix}
J_1 \\ J_2 \\ J_3 \\ J_4 \\ 0 \\ \vdots \\ 0 \\ 0
\end{bmatrix},
\quad \text{(A.114)}
$$

as there are no external current sources at the internal terminals $5, 6, \cdots, k+1$. Since the $(k+1)$th current source is zero, the $(k+1)$th row and column from $\mathbf{Y}$ can be eliminated, yielding

$$
\begin{bmatrix}
y'_{11} & y'_{12} & \cdots & y'_{1,k} \\
y'_{21} & y'_{22} & \cdots & y'_{2,k} \\
\vdots & \vdots & \ddots & \vdots \\
y'_{k1} & y'_{k2} & \cdots & y'_{k,k}
\end{bmatrix}
\begin{bmatrix}
v'_1 \\ v'_2 \\ v'_3 \\ v'_4 \\ v'_5 \\ \vdots \\ v'_k
\end{bmatrix}
=
\begin{bmatrix}
J_1 \\ J_2 \\ J_3 \\ J_4 \\ 0 \\ \vdots \\ 0
\end{bmatrix},
\quad \text{(A.115)}
$$

where

$$
y'_{ij} = y_{ij} - y_{k+1,j}\frac{y_{i,k+1}}{y_{k+1,k+1}} \tag{A.116}
$$

$$
v'_i = v_i - v_{k+1}. \tag{A.117}
$$

It is of no concern that the terminal voltages have been altered, as they have all had the same quantity added to them. This has the same effect as choosing another reference terminal. The reference terminal is arbitrary until it is connected to the full circuit. The process can be continued until the network equations are reduced to that of a four-terminal element, that is

$$
\begin{bmatrix}
y''_{11} & y''_{12} & y''_{13} & y''_{14} \\
y''_{21} & y''_{22} & y''_{23} & y''_{24} \\
y''_{31} & y''_{32} & y''_{33} & y''_{34} \\
y''_{41} & y''_{42} & y''_{43} & y''_{44}
\end{bmatrix}
\begin{bmatrix}
v''_1 \\ v''_2 \\ v''_3 \\ v''_4
\end{bmatrix}
=
\begin{bmatrix}
J_1 \\ J_2 \\ J_3 \\ J_4
\end{bmatrix}.
\quad \text{(A.118)}
$$

The y-parameter matrix in Equation (A.118) is that of the four-terminal element shown in Figure A-3.

Power Descriptions

At radio and microwave frequencies signal levels are usually expressed in terms of the power of a signal. While power can be expressed in absolute terms such as watts (W) or milliwatts (mW), it is much more useful to use a logarithmic scale. The ratio of two power levels in bels (B) is

$$P(B) = \log_{10} \frac{P}{P_{\mathrm{REF}}}, \qquad (B.1)$$

where P_{REF} is a reference power. Human senses have a logarithmic response and the minimum resolution tends to be about 0.1 B, so it is most common to use decibels (dB); 1 B = 10 dB. Common designations are shown in Table B-1. Also, 1 dBm is a very common power level in RF and microwave power circuits.

Table B-1   Common reference power designations.

P_{REF}	Bell Units	Decibel Units
1 W	BW	dBW
1 mW	Bm	dBm

EXAMPLE B. 1   Power Calculations

The output stage of an RF front end consists of an amplifier followed by a filter and then an antenna. The amplifier has a gain of 33 dB, the filter has a loss of 2.2 dB, and of the power input to the antenna, 45% is lost as heat due to resistive losses. If the power input to the amplifier is 1 W, calculate the following:
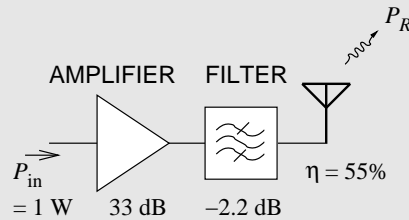
(a) What is the power input to the amplifier expressed in dBm?

$P_{\text{in}} = 1\text{ W} = 1000\text{ mW}, \quad P_{\text{dBm}} = 10\log_{10}(1000/1) = 30\text{ dBm}$

(b) Express the loss of the antenna in dB.
45% of the power input to the antenna is dissipated as heat.

The antenna has an efficiency, η, of 55%. $P_2 = 0.55 P_1$
Loss $= P_1/P_2 = 1/0.55 = 1.818 = 2.60$ dB

(c) What is the total gain of the RF front end (amplifier + filter + antenna)?

Total gain $= (\text{amplifier gain})_{\text{dB}} + (\text{filter gain})_{\text{dB}} + (\text{antenna gain})_{\text{dB}}$
$= (33 - 2.2 - 2.6)\text{ dB} = 28.2\text{ dB}$

(d) What is the total power radiated by the antenna in dBm?

$P_R = P_{\text{in}}|_{\text{dBm}} + (\text{total gain})_{\text{dB}} = 30\text{ dBm} + 28.2\text{ dB} = 58.2\text{ dBm}$

(e) What is the total power radiated by the antenna in mW.

$P_R = 10^{58.2/10} = 661,245\text{ mW} = 661.2\text{ W}$

Physical Constants and Material Properties

C.1   SI Units

Symbols for SI units[1] (from the French language name Système International d'Unités) are written in upright roman font. (Source: U.S. National Institute of Standards and Technology (2006) [307], and 2002 CODATA recommended values of constants [308].)

- The unit of length is spelt 'meter' in the US and 'metre' in other countries.

- The unit designations, such as 'm' for meter, is called a symbol and not an abbreviation.

- Symbols for units are written in lower case unless the symbol is derived from the name of a person. For example the symbol for the unit of force is 'N' as it is named after Isaac Newton. An exception is the use of L for liter to avoid possible confusion with l which looks like the numeral one and the letter i.

- A space separates a name from the symbol for the unit, e.g. 5.6 kg. There is an exception for degrees with the symbol °. For example 45 degrees is written $45°$.

---

[1] The older metric systems used different fundamental units; for example, the mks metric system used meter, kilogram, and second as fundamental units; the cgs metric system used centimeter, gram, and second as fundamental units.

SI Unit	Name	Usage	In Terms of Fundamental Units
A	ampere	current	Fundamental unit
C	coulomb	charge	A·s
J	Joule	unit of energy	$kg \cdot m^2 \cdot s^{-2}$
m	meter	length	Fundamental unit
kg	kilogram	SI fundamental unit	Fundamental unit
g	gram	weight	=kg/1000
N	Newton	unit of force	$kg \cdot m \cdot s^{-2}$
Pa	Pascal	pressure	$kg \cdot m^{-1} \cdot s^{-2}$
s	second	time	Fundamental unit
W	watt	power	$J \cdot s^{-1}$
cd	candela	luminous intensity	Fundamental unit
mol	mole	amount of substance	Fundamental unit
K	Kelvin	thermodynamic temperature	Fundamental unit

C.1.1 SI Unit Combinations

When SI units are multiplied a center dot is used. For example, newton meters is written N·m. When a unit is derived from the ratio of symbols then either a solidus, '/', or a negative exponent are used. So the symbol for velocity (meters per second) is either m/s or $m \cdot s^{-1}$. The use of multiple solidus for a combination symbol is confusing and must be avoided. So the symbol for acceleration is $m \cdot s^{-2}$ and not m/s/s. The use of multiple solidi is confusing and calculations performed using symbols in this are ambiguous. For example, the thermal conductivity of aluminum at room temperature is $k = 237$ kW·m^{-1}·K^{-1} and not $k = 237$ kW/m/K or 237 kW /m·K. However 237 kW /(m·K) is sometimes used. The thermal resistance of a rod of cross-sectional area A and length ℓ is

$$R_{\mathrm{TH}} = \frac{\ell}{kA} \ . \tag{C.1}$$

If $A = 0.3$ cm^2 and $\ell = 2$ mm, the thermal resistance is

$$R_{\mathrm{TH}} = \frac{(2 \ \mathrm{mm})}{(237 \ \mathrm{kW} \cdot \mathrm{m}^{-1} \cdot \mathrm{K}^{-1}) \cdot (0.3 \ \mathrm{cm}^2)} \ . \tag{C.2}$$

$$= \frac{(2 \cdot 10^{-3} \ \mathrm{m})}{237 \cdot (10^3 \cdot \mathrm{W} \cdot \mathrm{m}^{-1} \cdot \mathrm{K}^{-1}) \cdot 0.3 \ (10^{-2} \cdot \mathrm{m})^2} \ . \tag{C.3}$$

$$= \frac{2 \cdot 10^{-3}}{237 \cdot 10^3 \cdot 0.3 \cdot 10^{-4}} \cdot \frac{\mathrm{m}}{\mathrm{W} \cdot \mathrm{m}^{-1} \cdot \mathrm{K}^{-1} \cdot \mathrm{m}^2} \ . \tag{C.4}$$

$$= 2.813 \cdot 10^{-4} \ \mathrm{K} \cdot \mathrm{W}^{-1} = 281.3 \ \mu\mathrm{K}/\mathrm{W}^{-1} \ . \tag{C.5}$$

This would be an error prone calculation if the thermal conductivity was taken as 237 kW/m/K.

C.1.2 SI Prefixes

A prefix before a unit indicates a multiple of a unit, for example, 1 pA is 10^{-12} amps. (Source: US National Institute of Standards and Technology (2006) [307], and 2002 CODATA [308].)

SI Prefixes		
Symbol	Factor	Name
10^{-24}	y	yocto
10^{-21}	z	zepto
10^{-18}	a	atto
10^{-15}	f	femto
10^{-12}	p	pico
10^{-9}	n	nano
10^{-6}	μ	micro
10^{-3}	m	milli
10^{-2}	c	centi
10^{-1}	d	deci

SI Prefixes		
Symbol	Factor	Name
10^{1}	da	deca
10^{2}	h	hecto
10^{3}	k	kilo
10^{6}	M	mega
10^{9}	G	giga
10^{12}	T	tera
10^{15}	P	peta
$10^{18|}$	E	exa
10^{21}	Z	zetta
10^{24}	Y	yotta

C.1.3 Physical and Mathematical Constants

Physical and mathematical constants in SI Units. Source: US National Institute of Standards and Technology (2006) [307], and 2002 CODATA recommended values of constants [308].

Parameter	Value	Description
c	299792458 m $\cdot$ s^{-1}	Speed of light in a vacuum (free space)
e	1.6021765310^{-19} C	Elementary charge (negative of the charge of an electron)
e	2.718281828459045	Natural log base
γ	0.577215664901532	Euler's ratio
ϕ	1.618033988749894	Golden ratio
ε_0	$8.854187817 \times 10^{-12}$ F $\cdot$ m^{-1}	Permittivity of a vacuum (free space)
h	$6.6260693 \times 10^{-34}$ J$\cdot$s	Planck constant (alt. $\hbar = h/(2\pi)$)
k	$1.3806505 \times 10^{-23}$ J $\cdot$ K^{-1}	the Boltzmann constant
m_e	$9.1093826 \times 10^{-31}$ kg	Electron mass
μ_0	$12.566370614 \times 10^{-7}$ N $\cdot$ A^{-2}	Permeability of free space $= 4\pi \times 10^{-7}$ N $\cdot$ A^{-2}
π	$3.14159265358979323846264$	Pi, ratio of circumference to diameter of a circle
	101325 Pa	Standard atmosphere (pressure)
η	376.730313461 Ω	Characteristic impedance of vacuum (free space)

C.1.4   Conversions and Standard Values

Description	Value	In Terms of Fundamental Units
Absolute zero temperature	0 K	Fundamental unit $= -273.15°$C
Room temperature		25°C, 298 K, 77°F, generally used as an imprecise measurement implying that properties are unchanged over a few degrees variation. Sometimes 22°C = 72°F is used as room temperature.

C.2   Conductors, Dielectrics, and Magnetic Materials

Electrical and thermal conductivities, and relative permittivities and permeabilities of common materials encountered in RF and microwave circuits are given in the tables below. Electrical conductivity is the inverse of resistivity, $\sigma = 1/\rho$. Where the material parameter is listed as variable, the physical parameters depend on the formulation of the alloy and typical values are not meaningful. Manufacturer's data sheets for the specific alloy must be consulted. $\perp$ indicates perpendicular to crystal axis, // indicates parallel to crystal axis. The properties of graphite and sapphire depend strongly on crystal orientation.

The resistivity values of metal conductors are those of single crystal. The resistivity values of the best fabricated materials(electroplated and deposited) used in RF and microwave applications is generally a factor of 2 higher.

Material data from several sources including the *Standard Reference Data* database of the U.S. National Institute of Standards and Technology [307], the CODATA databases of the International Council for Science: Committee on Data for Science and Technology [308], and References [309–312]. Electrical and thermal properties are functions of temperature, with thermal properties generally being stronger functions. Properties at temperatures other than 300 K should be researched.

RELATIVE PERMEABILITY	
Material	Relative Permeability μ_r
Aluminum	1.00000065
Cobalt	60
Ferrite (NiZn)	1 000
Iron, cast	60
Iron, powder	100
Mumetal	20 000
Nickel	50
Permalloy 45	2 500
Silver	0.99999981
Supermalloy	100 000
Wood (dry)	0.99999942

Dielectrics and Nonconductors				
Material	Resistivity $M\Omega \cdot m$ ρ, at 300 K	Thermal Conductivity $W \cdot m^{-1} \cdot K^{-1}$ k, at 300 K	Relative Permittivity ε_r @ 1 GHz	Loss Tangent $\tan \delta$ @ 1 GHz
Air (dry, sea level)	4×10^7	25	1.000 5	0.000
Alumina				
99.5%	$> 10^6$	30	10.1	0.000 1–0.000 2
96%	$> 10^6$	24.7	9.6	0.000 6
85%	$> 10^6$	16	8.5	0.001 5
Bakelite	1–100	16.7	4.74	0.022
Diamond	10^{10}	895		
Ferrite (MnZn)	0.1–$10\ \Omega \cdot m$	1–5	13–16	0.000 4
Ferrite (NiZn)	0.1–12.4	1–5	13–16	0.000 4
FR-4 circuit board	8×10^5	0.3	4.3–4.5	0.01
GaAs	1.0	59.1	12.85	0.000 6
InP	Up to 0.001	35	12.4	0.001
Glass	2×10^8	0.8–1.2	4–7	0.002
Mica	2×10^5	260–750	5.4	0.000 6
Mylar	10^{10}	0.1	3.2	0.005
Paper, white	3.5×10^6	40–90	3	0.008
Polyethylene	$> 10^7$	0.42–0.51	2.26	0.000 2
Polyimide	10^{10}	100–350	3.2	0.005
Polypropylene	$> 10^7$	100–220	2.25	0.0003
Quartz (fused)	7.5×10^{11}	1.44	3.8	0.00075
Sapphire				
//	$> 10^6$	35	11.6	0.00004–0.000 07
$\perp$	$> 10^6$	32	9.4	0.00004–0.00007
Polycrystalline	$> 10^6$	33	10.13	0.00004–0.00007
Silicon				
Normal for CMOS	$50\ \mu\Omega{\cdot}m$	149	11.68	0.005
High resistivity	$300\ m\Omega{\cdot}m$	149	11.68	0.005
Dioxide (SiO_2)	5.813×10^7	1.4	3.7–4.1	0.001
Poly (undoped)	—	13.8–22	—	—
Poly (doped)	—	125	—	—
Teflon (PTFE)	10^{10}	250	2.1	0.0003
Vacuum	∞	0	1	0
Water				
Distilled	182	580	80	0.1
Ice (273 K)	1	220	4.2	0.05
Wood (dry oak)	3×10^{11}	170	1.5–4	0.01
Zirconia (variable)	10^4	2.0	28	0.0009

Conductors		
Material	Electrical Resistivity ρ n$\Omega \cdot$ m at 20 $^\circ$ C	Thermal Conductivity, k, kW $\cdot$ m^{-1}K^{-1} at 300 K
Aluminum	26.50	237
Brass (a copper alloy)	Variable	120
Bronze (a copper alloy)	Variable	110
Chromium	125	93.9
Copper	16.78	401
Gold	22.14	318
Graphite // c-axis	1 200	1 950
Graphite $\perp$ c-axis	41 000	5.7
Iridium	47.1	147
Iron	96.1	80.2
Lead	208	35.3
Mercury	961	8.34
Nickel	69.3	90.9
NiChrome	1 100	0.011
Palladium	105.4	71.8
Platinum	105	71.6
Silver	15.87	429
Solder		
tin-lead 50% Pb, 50% Sn	17.2	34
lead-free 77.2% Sn, 2.8% Ag, 20% In	170	53.5
Steel, stainless	720	16
Steel, carbon (standard)	208	46
Tin	115	66.8
Titanium	4 200	21.9
Tungsten	52.8	173
Zinc	59.0	116

C.3   Greek Alphabet

Name	Upper case	Lower case
alpha	A	α
beta	B	β
gamma	Γ	γ
delta	Δ	δ
epsilon	E	ϵ
zeta	Z	ζ
eta	H	η
theta	Θ	θ
iota	I	ι
kappa	K	κ
lambda	Λ	λ
mu	M	μ
nu	N	ν
xi	Ξ	ξ
omicron	O	o
pi	Π	π
rho	P	ρ
sigma	Σ	σ
tau	T	τ
upsilon	Υ	υ
phi	Φ	ϕ
chi	X	χ
psi	Ψ	ψ
omega	Ω	ω
Additional characters		
varepsilon		ε
varphi		φ
varpi		ϖ
varrho		ϱ
varsigma		ς
vartheta		ϑ
aleph		$\aleph$

APPENDIX D

Maxwell's Equations

In this appendix, EM theory is presented in a form that aids in the understanding of distributed effects such as propagation on transmission lines, coupling of transmission lines, and how transmission line effects can be used to realize components with unique functionality. While this is a review of material that most readers have previously learned, it is presented in a slightly different form than is usual. The treatment begins with Maxwell's equations, which is not the way EM theory is presented initially. Generally the basic laws of electric and magnetic fields are presented: Ampere's circuital law relating current and magnetic field; the Biot-Savart law also relating current and magnetic field; Gauss' law relating charge and electric field; and Faraday's law relating electric field and a time-varying magnetic field. Following this treatment Maxwell's equations are presented, but Maxwell's equations cannot be derived from the early field laws. However, the order from basic laws to Maxwell's equations is what works best when EM theory is first presented. Maxwell's equations are a remarkable insight and the early field laws can be derived from them. Most importantly, Maxwell's equations describe the propagation of an EM field. This appendix presents Maxwell's equations in point form in Section D.1 and in integral form in Section D.3. From these, the early electric and magnetic field laws are derived. The effect of boundary conditions are introduced in Section D.5 to arrive at implications for multimoding on transmission lines. Multimoding is almost always undesirable, and in

designing transmission line structures so that multimoding is avoided, it is necessary to have rules that establish when multimoding can occur. These rules are derived from Maxwell's equations together with boundary conditions. Some of these boundary conditions are established by electric walls at conductors. Other boundary conditions come from the interface of dielectric regions. If these regions have significantly different permittivities then the boundary can be considered to be a magnetic wall, the dual of electric walls. This is an important conceptual insight, but implies that there are magnetic charges and magnetic currents. Of course, these do not exist, but imagining that they do leads to intuitive insight.

D.1 Point Form of Maxwell's Equations

The characteristics of EM fields are described by Maxwell's equations:

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} - \mathbf{M} \tag{D.1}$$

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J} \tag{D.2}$$

$$\nabla \cdot \mathbf{D} = \rho_V \tag{D.3}$$

$$\nabla \cdot \mathbf{B} = \rho_{mV}, \tag{D.4}$$

where μ is called the permeability of a medium and ε is called the permittivity of the medium. They are the property of a medium and describe the ability to store magnetic energy and electric energy. The other quantities in Equations (D.1)–(D.4) are

- $\mathbf{E}$, the **electric field**, with units of volts per meter (V/m)

- $\mathbf{D}$, the **electric flux** density, with units of volts per meter (V/m)

- $\mathbf{H}$, the three-dimensional **magnetic field**, with units of amperes per meter (A/m)

- $\mathbf{B}$, the **magnetic flux** density, with units of volts per meter (V/m)

- $\mathbf{J}$, the **electric current** density, with units of amperes per square meter (A/m^2).

- ρ_V, the **electric charge** density, with units of coulombs per cubic meter (C/m^3).

- ρ_{mV}, the magnetic charge density, with units of weber per cubic meter (Wb/m^3).

- $\mathbf{M}$ is the magnetic current density, with units of volts per square meter (V/m^2).

Magnetic charges do not exist, but their introduction in Maxwell's equations through the **magnetic charge** density, ρ_{mV}, and the **magnetic current** density, $\mathbf{M}$, introduce an aesthetically appealing symmetry to Maxwell's equations. Maxwell's equations are differential equations, and as with most differential equations, their solution is obtained with particular boundary conditions which here are imposed by conductors. Electric conductors (i.e., electric walls) support electric charges and hence electric current. By analogy, magnetic walls, support magnetic charges and magnetic currents. Magnetic walls also provide boundary conditions to be used in the solution of Maxwell's equations. The notion of magnetic walls is important in RF and microwave engineering, as they are approximated by the boundary between two dielectrics of different permittivity. The greater the difference in permittivity the more closely the boundary approximates a magnetic wall. As a result, the result of many structures with different dielectrics can be simplified, aiding in intuitive understanding.

The boldface quantities in Equations (D.1)–(D.4) are three-dimensional fields, for example,

$$\mathbf{E} = E_x \hat{x} + E_y \hat{y} + E_z \hat{z} \tag{D.5}$$

where $\hat{x}$, $\hat{y}$, and $\hat{z}$ are the unit vectors (having a magnitude of 1) in the x, y and z directions, respectively. $E_x \hat{x}$, $E_y \hat{y}$, and $E_z \hat{z}$ are the magnitudes of the electric field components in the x, y, and z directions, respectively.

The symbols and units used with the various field quantities and some of the other symbols to be introduced soon are given in Table D-1. $\mathbf{B}$ and $\mathbf{H}$, and $\mathbf{D}$ and $\mathbf{E}$ are related to each other by the properties of the medium embodied in μ and ε:

$$\mathbf{B} = \mu \mathbf{H} \tag{D.6}$$

$$\mathbf{D} = \varepsilon \mathbf{E}. \tag{D.7}$$

The quantity μ is called the **permeability** and describes the ability to store **magnetic energy** in a region. The permeability in free space (or vacuum) is denoted μ_0 and the magnetic flux and magnetic field are related as

$$B = \mu_0 H \tag{D.8}$$

where $\mu_0 = 4\pi \times 10^{-7}$ H/m. Permeability, μ, is generally a scalar, but in magnetic materials it can be a 3×3 matrix called a **dyadic** tensor. Then

$$\begin{bmatrix} B_x \\ B_y \\ B_z \end{bmatrix} = \begin{bmatrix} \mu_{xx} & \mu_{xy} & \mu_{xz} \\ \mu_{yx} & \mu_{yy} & \mu_{yz} \\ \mu_{zx} & \mu_{zy} & \mu_{zz} \end{bmatrix} \begin{bmatrix} H_x \\ H_y \\ H_z \end{bmatrix}. \tag{D.9}$$

The property indicated by the dyadic results from alignment of electron spins in a material. However in most materials there is no alignment of spins

Table D-1   Quantities used in Maxwell's equations. Magnetic charge and current are introduced in establishing boundary conditions, especially at dielectric interfaces.

Symbol	SI Unit	SI Unit	Name and Base Units
E	volt per meter	$\mathrm{V \cdot m^{-1}}$	electric field intensity base unit: $\mathrm{kg \cdot m^2 \cdot s^{-3} \cdot A^{-1}}$
H	amp per meter	$\mathrm{A \cdot m^{-1}}$	magnetic field intensity
D	coulomb per square meter	$\mathrm{C \cdot m^{-2}}$	$D = \varepsilon E$, electric flux density base unit: $\mathrm{A \cdot s \cdot m^{-2}}$
B	tesla, weber per square meter	T	$B = \mu H$, magnetic flux density base unit: $\mathrm{kg \cdot s^{-2} \cdot A^{-1}}$
I	amp	A	electric current
M	weber per second	$\mathrm{Wb \cdot s^{-1}}$	magnetic current base unit: $\mathrm{kg \cdot m^2 \cdot s^{-3} \cdot A^{-1}}$
q_e	coulomb	C	electric charge base unit: $\mathrm{A \cdot s}$
q_m	weber	Wb	magnetic charge base unit: $\mathrm{kg \cdot m^2 \cdot s^{-2} \cdot A^{-1}}$
ψ_e	coulomb	C	electric flux base unit: $\mathrm{A \cdot s}$
ψ_m	weber	Wb	magnetic flux base unit: $\mathrm{kg \cdot m^2 \cdot s^{-2} \cdot A^{-1}}$
ρ_V	coulomb per cubic meter	$\mathrm{C \cdot m^{-3}}$	charge density base unit: $\mathrm{A \cdot s \cdot m^{-3}}$
ρ_S	coulomb per square meter	$\mathrm{C \cdot m^{-2}}$	surface charge density base unit: $\mathrm{A \cdot s \cdot m^{-2}}$
ρ_{mV}	weber per cubic meter	$\mathrm{Wb \cdot m^{-3}}$	magnetic charge density base unit: $\mathrm{kg \cdot m^{-1} \cdot s^{-2} \cdot A^{-1}}$
ρ_{mS}	weber per square meter	$\mathrm{Wb \cdot m^{-2}}$	surface magnetic charge density base unit: $\mathrm{kg \cdot s^{-2} \cdot A^{-1}}$
J	amp per square meter	$\mathrm{A \cdot m^{-2}}$	electric current density
J_S	amp per meter	$\mathrm{A \cdot m^{-1}}$	surface electric current density
J_m	weber per second per square meter	$\mathrm{Wb \cdot s^{-1} \cdot m^{-2}}$	magnetic current density base unit: $\mathrm{kg \cdot s^{-3} \cdot A^{-1}}$
J_{mS}	weber per second per meter	$\mathrm{Wb \cdot s^{-1} \cdot m^{-1}}$	surface magnetic current density base unit: $\mathrm{kg \cdot m^{-1} \cdot s^{-3} \cdot A^{-1}}$
S	square meters	$\mathrm{m^2}$	surface
V	cubic meters	$\mathrm{m^3}$	volume
ε	farads per meter	$\mathrm{F \cdot m^{-1}}$	permittivity base unit: $\mathrm{kg^{-1} \cdot m^{-3} \cdot A^2 \cdot s^4}$
μ	henry per meter	$\mathrm{H \cdot m^{-1}}$	permeability base unit: $\mathrm{kg \cdot m \cdot s^{-2} \cdot A^{-2}}$
dS	square meter	$\mathrm{m^2}$	incremental area
$d\ell$	meter	m	incremental length
dV	cubic meter	$\mathrm{m^3}$	incremental volume
$\oint_C$			integral around a closed contour
$\oint_S$			integral over a closed surface
$\int_V$			volume integral

an $\mu = \mu_0$. The **relative permeability**, μ_r, refers to the ratio of permeability of a material to its value in a vacuum:

$$\mu_r = \frac{\mu}{\mu_0}. \qquad (D.10)$$

So $\mu_r > 0$ indicates that a material can store more magnetic energy than can a vacuum in a given volume.

The other material quantity is the **permittivity** ε, which is denoted ε_0, and in a vacuum

$$D = \varepsilon_0 E, \qquad (D.11)$$

where $\varepsilon_0 = 8.854 \times 10^{-12}$ F·m^{-1}. The **relative permittivity**, ε_r, refers to the ratio of permeability of a material to its value in a vacuum:

$$\varepsilon_r = \frac{\varepsilon}{\varepsilon_0}. \qquad (D.12)$$

So $\varepsilon_r > 0$ indicates that a material can stored more electric energy in a volume than can be stored in a vacuum. In some calculations it is useful to introduce an **electric polarization**, P_e. Then

$$D = \varepsilon E = \varepsilon_0 E + P_e \qquad (D.13)$$

and so the polarization vector is

$$P_e = (\varepsilon - \varepsilon_0)\, E = \varepsilon_0 \chi_e E, \qquad (D.14)$$

where χ_e is called the **electric susceptibility**.

Some materials require a dyadic form of ε. This usually indicates a dependence on crystal symmetry, and the relative movement of charge centers in different directions when an E field is applied. Some commonly used microwave substrates, such as sapphire, have a permittivity that it direction dependent rather than having full **dyadic permittivity**. A material in which the permittivity is a function of direction is called an anisotropic material or is said to have **dielectric anisotropy**. In an isotropic material the permittivity is the same in all directions; the material has **dielectric isotropy**. Most materials are isotropic.

Maxwell's original equations were put in the form of Equations (D.1)–(D.4) by the mathematician Oliver Heaviside from the less convenient form Maxwell originally used. The above equations are called the point form of Maxwell's equations, relating the field components to each other and to charge and current density at a point.

Maxwell's equations are fundamental properties and there is no underlying theory, so they must be accepted "as is" but they have been verified in countless experiments. Maxwell's equations have three types of derivatives. First, there is the time derivative, $\partial/\partial t$. Then there are two

spatial derivatives: $\nabla\times$, called curl, capturing the way a field circulates spatially (or the amount that it curls up on itself); and $\nabla\cdot$, called the div operator, describing the spreading out of a field. **Curl** and **div** have different forms in different coordinate systems, and in the rectangular system can be expanded as

$$\nabla \times \mathbf{A} = \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} \right) \hat{x} + \left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x} \right) \hat{y} + \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} \right) \hat{z} \quad \text{(D.15)}$$

and

$$\nabla \cdot \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}. \quad \text{(D.16)}$$

Curl and div in cylindrical and spherical coordinates are given in Equations (A.19), (A.20), (A.28), and (A.29). Studying these equations you will see that curl, $\nabla\times$, describes how much a field circles around the x, y, and z axes. That is, the curl describes how a field circulates on itself. So Equation (D.1) relates the amount an electric field circulates on itself to changes of the B field in time (and modified by the magnetic current). A spatial derivative of electric fields is related to a time derivative of the magnetic field. Also in Equation (D.2), the spatial derivative of the magnetic field is related to the time derivative of the electric field (and modified by the electric current). These are the key elements that result in self-sustaining propagation.

Div, $\nabla\cdot$, describes how a field spreads out from a point. So the presence of net electric charge (say, on a conductor) will result in the electric field spreading out from a point (see Equation (D.3)). The magnetic field (Equation (D.4)) can never diverge from a point, which is a result of magnetic charges not existing. A magnetic wall describes an open circuit and then effective magnetic charges terminate the magnetic field. What actually happens in free space or on a transmission line depends on boundary conditions, and in the case of transmission lines, the dimensions involved.

How fast a field varies with time, $\partial\mathbf{B}/\partial t$ and $\partial\mathbf{D}/\partial t$, depends on frequency. The more interesting property is how fast a field can change spatially, $\nabla \times \mathbf{E}$ and $\nabla \times \mathbf{H}$—this depends on wavelength relative to geometry. So if the cross-sectional dimensions of a transmission line are much less than a wavelength (say, less than $\lambda/4$) then it will be impossible for the fields to curl up on themselves and so, perhaps, there will be only one or, in some cases, no solutions to Maxwell's equations.

EXAMPLE D. 1   Energy Storage

Consider a material with a relative permittivity of 65 and a relative permeability of 1000. There is a static electric field (E) of 1 kV/m. How much energy is stored in the (E) field in a 10 cm^3 volume of the material?

Solution:

Energy stored in a static electric field $= \int_V \vec{D} \cdot \vec{E} \cdot dv$:

$$D = \varepsilon E \text{ and field is constant (uniform)}$$

$$\text{Energy stored} = \varepsilon E^2 \times (\text{volume}) = 65\varepsilon_0 \cdot \left(10^3\right)^2 \cdot \left(10 \cdot 10^{-6}\right) = 5.75 \text{ nJ}.$$

EXAMPLE D. 2   Polarization Vector

A time-varying electric field in the x direction has a strength of 100 V/m and a frequency of 10 GHz. The medium has a relative permittivity of 10. What is the polarization vector? Express this vector in the time domain?

Solution:

$D = \varepsilon E = \varepsilon_0 E + P_e$.
The polarization vector is

$$P_e = (\varepsilon - \varepsilon_0) E = (10 - 1) \varepsilon_0 E = 9\varepsilon_0 E.$$

Now $\bar{E} = 100\cos\left(2 \cdot \pi \cdot 10^{10} t\right)$
and $\varepsilon_0 = 8 \cdot 85 \cdot 10^{12}$ and so

$$
\begin{aligned}
P_e &= 9\left(8.85 \times 10^{-12}\right) \cdot (100) \cos\left(6.28 \times 10^{10} t\right) \\
&= 79.7 \times 10^{-10} \cdot \cos\left(6.28 \times 10^{10} t\right) \\
&= 7.79\cos\left(6.28 \times 10^{10} t\right) \text{ nC/m}^2.
\end{aligned}
\tag{D.17}
$$

D.2   Maxwell's Equations in Phasor Form

Phasors reduce the dimensionality of Maxwell's equations by replacing a time derivative by a scalar. The phasor form is used with a cosinusoidally varying quantity so, for example, an x-directed electric field in rectangular coordinates,

$$\mathbf{E} = |E_x| \cos\left(\omega t + \phi\right)\hat{x}, \tag{D.18}$$
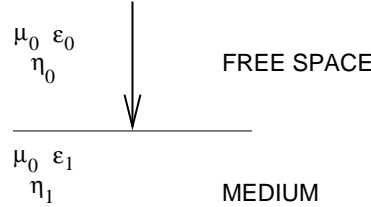
Figure D-1 Two regions used in Example D.2.

is represented as the phasor

$$E_x = |E_x|e^{\jmath\phi}\hat{x}. \tag{D.19}$$

The full three-dimensional electric field is

$$\mathbf{E} = |E_x|\cos\left(\omega t + \phi_x\right)\hat{x} + |E_y|\cos\left(\omega t + \phi_y\right)\hat{y} + |E_z|\cos\left(\omega t + \phi_z\right)\hat{z} \tag{D.20}$$

and has the phasor form

$$E = |E_x|e^{\jmath\phi_x}\hat{x} + |E_y|e^{\jmath\phi_y}\hat{y} + |E_z|e^{\jmath\phi_z}\hat{z}. \tag{D.21}$$

To express Maxwell's equations in phasor form, the full complex form of Equation (D.20) must be considered. Using Equation (A.40), Equation (D.20) becomes

$$\begin{aligned}\mathbf{E} =& \frac{1}{2}\left(|E_x|e^{\jmath\phi_x}\hat{x} + |E_y|e^{\jmath\phi_y}\hat{y} + |E_z|e^{\jmath\phi_z}\hat{z}\right)e^{\jmath\omega t} \\ &+ \frac{1}{2}\left(|E_x|e^{-\jmath\phi_x}\hat{x} + |E_y|e^{-\jmath\phi_y}\hat{y} + |E_z|e^{-\jmath\phi_z}\hat{z}\right)e^{-\jmath\omega t}\end{aligned} \tag{D.22}$$

with the time derivative

$$\begin{aligned}\frac{\partial \mathbf{E}}{\partial t} =& \frac{1}{2}\jmath\omega\left(|E_x|e^{\jmath\phi_x}\hat{x} + |E_y|e^{\jmath\phi_y}\hat{y} + |E_z|e^{\jmath\phi_z}\hat{z}\right)e^{\jmath\omega t} \\ &- \frac{1}{2}\jmath\omega\left(|E_x|e^{-\jmath\phi_x}\hat{x} + |E_y|e^{-\jmath\phi_y}\hat{y} + |E_z|e^{-\jmath\phi_z}\hat{z}\right)e^{-\jmath\omega t}.\end{aligned} \tag{D.23}$$

Equation (D.23) has the phasor form

$$\Re\left\{\frac{2}{e^{\jmath\omega t}}\frac{\partial \mathbf{E}}{\partial t}\right\} = \jmath\omega\left(|E_x|e^{\jmath\phi_x}\hat{x} + |E_y|e^{\jmath\phi_y}\hat{y} + |E_z|e^{\jmath\phi_z}\hat{z}\right). \tag{D.24}$$

Noting this, **Maxwell's equations in phasor form**, from Equations (D.1)–(D.4), are

$$\begin{aligned}\nabla \times E &= -\jmath\omega B - M \tag{D.25}\\ \nabla \times H &= \jmath\omega D + J \tag{D.26}\\ \nabla \cdot D &= \rho_V \tag{D.27}\\ \nabla \cdot B &= \rho_{mV}, \tag{D.28}\end{aligned}$$

D.3   Integral Form of Maxwell's Equations

It is sometimes more convenient to use the integral forms of Maxwell's equations, and Equations (D.1)–(D.4) become

$$\oint_S \nabla \times \mathbf{E} \cdot d\mathbf{S} = \oint_C \mathbf{E} \cdot d\ell = - \oint_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{S} - \oint_S \mathbf{M} \cdot d\mathbf{S} \qquad (D.29)$$

$$\oint_S \nabla \times \mathbf{H} \cdot d\mathbf{S} = \oint_C \mathbf{H} \cdot d\ell = \int_S \frac{\partial \mathbf{D}}{\partial t} \cdot d\mathbf{S} + \mathbf{I}I \qquad (D.30)$$

$$\int_V \nabla \cdot \mathbf{D} \, dV = \oint_S \mathbf{D} \cdot d\mathbf{s} = \int_V \rho_V \, dV = Q_{\text{enclosed}} \qquad (D.31)$$

$$\int_V \nabla \cdot \mathbf{B} \, dV = \oint_S \mathbf{B} \cdot d\mathbf{S} = 0, \qquad (D.32)$$

where Q_{enclosed} is the total charge in the volume enclosed by the surface, S. The subscript S on the integral indicates a surface integral and the circle on the integral sign indicates that the integral is over a closed surface. Two mathematical identities were used in developing Equations (D.29)–(D.32). The first identity is **Stoke's theorem**, which states that for any vector field **X**,

$$\oint_\ell \mathbf{X} \cdot d\ell \equiv \oint_S (\nabla \times \mathbf{X}) \cdot d\mathbf{S}. \qquad (D.33)$$

So the contour integral around a closed contour, C (with incremental length vector $d\ell$) of X, is the integral of $\nabla \times X$ over the surface enclosed by the contour and $d\mathbf{S}$ is the incremental area multiplied by a unit vector normal to the surface. This identity is used in Equations (D.29) and (D.30). The **divergence theorem** is the other identity and is used in Equations (D.31) and (D.32). For any vector field **X**, the divergence theorem states that

$$\oint_S \mathbf{X} \cdot d\mathbf{S} = \int_V \nabla \cdot \mathbf{X} \, dV. \qquad (D.34)$$

That is, the volume integral of $\nabla \cdot \mathbf{X}$ is equal to the closed surface integral of **X**.

D.4   Electric and Magnetic Field Laws

Before Maxwell's equations were postulated, five laws of electromagnetics were known. These are the Biot-Savart law, Ampere's circuital law, Gauss' law, and Faraday's law. The laws are for specific circumstances and mostly for static fields. Taken together, they cannot describe the propagation of EM signals, but they can be derived from Maxwell's equations. So Maxwell's equations cannot be derived from the electric and magnetic field laws. Maxwell's equations embody additional insight relating spatial derivatives to time derivatives which leads to propagating fields.
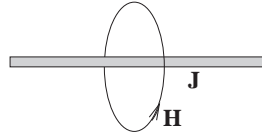
Figure D-2   Diagram illustrating Ampere's law. Ampere's law relates the current on a wire to the magnetic field around it.
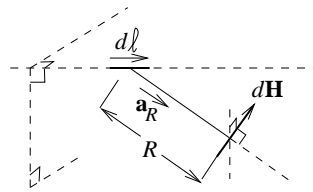
Figure D-3   Diagram illustrating the Biot-Savart law. The Biot-Savart law relates current to magnetic field.

D.4.1   Ampere's Circuital Law

Ampere's circuital law, often called just Ampere's law, relates direct current and the static magnetic field. The relationship is based on Figure D-2. In the static situation, $\partial \mathbf{D}/\partial t = 0$, and so one of Maxwell's equations, Equation (D.30), reduces to

$$\oint_\ell \mathbf{H} \cdot d\ell = I_{\text{enclosed}}. \tag{D.35}$$

This is Ampere's circuital law.

D.4.2   Biot-Savart Law

The Biot-Savart law relates current to magnetic field. With $\partial \mathbf{D}/\partial t = 0$, Equation (D.2) becomes

$$\nabla \times \mathbf{H} = \mathbf{J}. \tag{D.36}$$

In Cartesian coordinates,

$$\nabla f = \frac{\partial f}{\partial x}\hat{x} + \frac{\partial f}{\partial y}\hat{y} + \frac{\partial f}{\partial z}\hat{z}, \tag{D.37}$$

and if the current is confined to a thin wire as shown in Figure D-3, $\nabla \times H$ reduces to a line integral along the wire, but also with vector information. So the first step in the development of the Biot-Savart law (Equation (D.36)) is (ignoring vector information for now) simplified to

$$\frac{\partial H}{\partial \ell} = \frac{I}{4\pi R^2}. \tag{D.38}$$
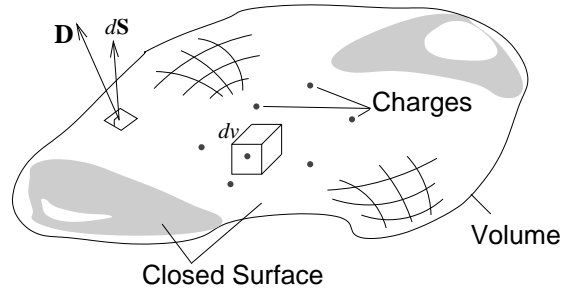
Figure D-4 Diagram illustrating Gauss' law. Charges are distributed in the volume enclosed by the closed surface. An incremental area is described by the vector $d\mathbf{S}$ which is normal to the surface and whose magnitude is the area of the incremental area.

With vector information and moving the infinitesimal length of wire, $d\ell$, from the left side of the equation to the right, the final form of the Biot-Savart law is obtained:

$$d\mathbf{H} = \frac{I d\ell \times \hat{a}_R}{4\pi R^2},\qquad\qquad\text{(D.39)}$$

which has the units of amps per meter in the SI system. In Equation (D.39), I is current, $d\ell$ is a length increment of a filament of current I, $\hat{a}_R$ is the unit vector in the direction from the current filament to the magnetic field, and R is the distance between the filament and the magnetic field increment.

D.4.3 Gauss' Law

The third law is Gauss' law, which relates the static electric flux density vector, $\mathbf{D}$, to charge. With reference to Figure D-4, Gauss' law in integral form is

$$\oint_s \mathbf{D} = \int_v \rho_v . dv = Q_{\text{enclosed}},\qquad\qquad\text{(D.40)}$$

which states that the integral of the electric flux vector, $\mathbf{D}$, is equal to the total charge enclosed by the surface, Q_{enclosed}. This is just the integral form of the third Maxwell equation (Equation (D.31)).

D.4.4 Faraday's Law

Faraday's law relates a time-varying magnetic field to an induced electromotive force, emf, which is now understood to be $\oint_\ell E \cdot d\ell$, that is, the closed contour integral of the electric field,

$$\text{emf} = \oint_\ell E \cdot d\ell = -\oint_s \frac{\partial B}{\partial t} \cdot d\mathbf{s},\qquad\qquad\text{(D.41)}$$

and this has the units of volts in the SI unit system. This is just the first of Maxwell's equations in integral form (see Equation (D.29)). The operation
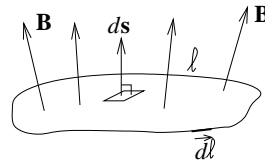
Figure D-5   Diagram illustrating Faraday's law.
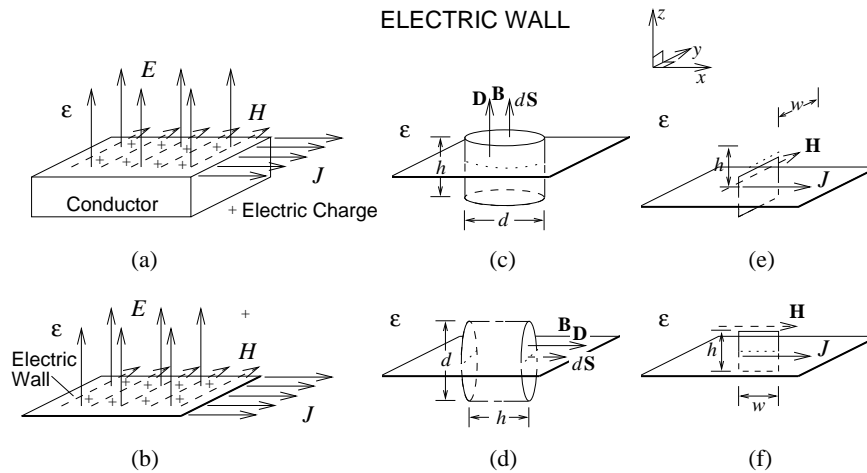
Figure D-6   Properties of an electric wall: (a) the electrical field is perpendicular to a conductor and the magnetic field is parallel to it and is orthogonal to the current; (b) a conductor can be approximated by an electric wall; (c) cylinder used in deriving the normal electric field; (d) cylinder used in showing that the tangential electric field is zero; (e) contour used in deriving the tangential magnetic field in the y direction; and (f) contour used in deriving the z- and x-directed magnetic fields (they are zero).

described in Equation (D.41) is illustrated in Figure D-5.

D.5   Electric and Magnetic Walls

The boundaries imposed by conductors and the interface between dielectrics define the mode structure supported by a transmission line. Understanding the fields at the interfaces provides the intuition required to understand the frequency limits of operation of distributed structures.

D.5.1   Electric Wall

An electric wall is formed by a perfect conductor, as shown in Figure D-6, where there are electric charges at the surface of the conductor. Consider that the conductor in Figure D-6(a) carries a current of density J and charge of density ρ_V. As the conductivity of the conductor goes to infinity, the

conductor becomes a perfect conductor and forms an electric wall, as shown in Figure D-6(b). The current and charge now occupy the very skin of the conductor at the interface to the region above which there is either free space or a dielectric. The behavior of the fields at the interface can be understood by considering the integration surfaces and volumes shown in Figures D-6(c)–(d). The cylinder in Figure D-6(c) with height h and diameter d enables Gauss' law (Equation (D.40)) to be applied. As $h \longrightarrow 0$, Equation (D.40) yields

$$\oint_s \mathbf{D} = D_z \pi d^2 = \int_v \rho_v.dv = Q_{\text{enclosed}} = \rho_S \pi d^2, \qquad \text{(D.42)}$$

where ρ_S is the surface charge density (in SI units of C/m^2) and so

$$D_z = \rho_S. \qquad \text{(D.43)}$$

Now considering the cylinder in Figure D-6(d), again as $h \longrightarrow 0$,

$$\oint_s \mathbf{D} = D_x \pi d^2 = \int_v \rho_v.dv = Q_{\text{enclosed}} = \rho_S \pi dh \longrightarrow 0.$$

That is,

$$D_x = 0. \qquad \text{(D.44)}$$

Changing the orientation of the cylinder, it can be seen that

$$D_y = 0. \qquad \text{(D.45)}$$

So the electric field (both $\mathbf{D}$ and $\mathbf{E} = \varepsilon \mathbf{D}$) is normal to the surface of the conductor.

The properties of the magnetic field at the electric wall can be deduced from Faraday's circuital law and the Biot-Savart law (Equation (D.39)). Consider the rectangular contour shown in Figure D-6(e) with height h and width w. With $h \longrightarrow 0$, only the y-directed component of the magnetic field will contribute to the magnetic field integral of Faraday's circuital law, so from Equation (D.35)

$$\oint_\ell \mathbf{H} \cdot d\ell = H_y w = I_{\text{enclosed}} = J_S w, \qquad \text{(D.46)}$$

where J_S is the surface current density (with SI units of A/m) and

$$H_y = J_S. \qquad \text{(D.47)}$$

For the field oriented in the x direction, the rectangular contour shown in Figure D-6(f) can be used, but now no current is enclosed by the contour so that with $h \longrightarrow 0$,

$$\oint_\ell \mathbf{H} \cdot d\ell = H_x w = I_{\text{enclosed}} = 0 \qquad \text{(D.48)}$$

MAGNETIC WALL

(a)   (b)   (c)
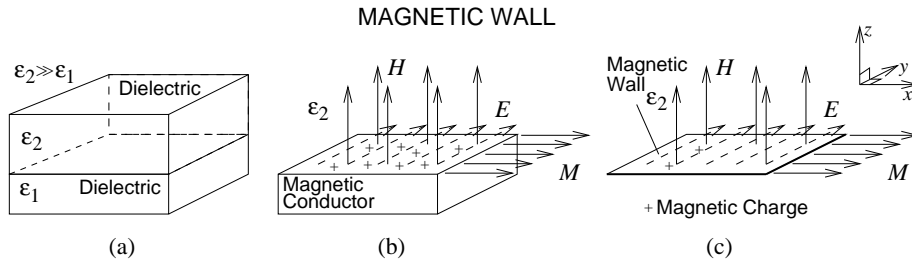
Figure D-7 Properties of the EM field at a magnetic wall: (a) the interface of two dielectrics of contrasting permittivities approximates a magnetic wall with the magnetic field perpendicular to the interface and the electric field parallel to it; (b) the dielectric with lower permittivity can be approximated as a magnetic conductor; and (c) the magnetic conductor forms a magnetic wall at the interface with the material with higher permittivity.

and

$$H_x = 0. \tag{D.49}$$

H_z has not yet been considered. Again, using the rectangular in Figure D-6(f) and with $w \longrightarrow 0$,

$$\oint_\ell \mathbf{H} \cdot d\ell = H_z h = I_{\text{enclosed}} = 0 \tag{D.50}$$

and

$$H_z = 0. \tag{D.51}$$

This could also have been derived from the Biot-Savart law (Equation (D.39)). Immediately above the electric wall $R \longrightarrow 0$ and only the filament of current immediately below the magnetic field contributes. Thus a_R is normal to the surface so that **H** will be perpendicular to the current, but in the plane of the current.

Gathering the results together, the EM fields immediately above the electric wall have the following properties:

The E field is perpendicular to the electric wall.	The H field is parallel to the electric wall.
$D_{\text{normal}} = \rho_S$, the surface charge density.	$H_{\text{parallel}} = J_S$, the surface current density.

D.5.2 Magnetic Wall

A true magnetic wall does not exist, but it is approximated by the interface between two dielectrics (as shown in Figure D-7) when the permittivity of the dielectrics differ so that $\varepsilon_1 \gg \varepsilon_2$. This is the situation with a microstrip line where the substrate has permittivity ε_1 and the air above the substrate

Table D-2   Properties of electric and magnetic walls.

	Electric Field	Magnetic Field
Electric Wall	Normal	Parallel
Magnetic Wall	Parallel	Normal

has permittivity ε_2. Considering the top dielectric region, the magnetic wall model can then be introduced at the interface between the two dielectrics complete with magnetic charges and magnetic current density, M. The situation is analogous to the electric wall situation with the roles of the electric and magnetic fields interchanged. The EM fields immediately above the magnetic wall have the following properties:

The H field is perpendicular to the magnetic wall.	The E field is parallel to the magnetic wall.
$H_{\text{normal}} = \rho_{mS}$, the surface magnetic charge density.	$E_{\text{parallel}} = M_S$, the surface magnetic current density.

What constitutes a good magnetic wall is subject to debate, as the contrast between dielectrics is not as great as the contrast between the conductivity of a good conductor and that of a dielectric such as air. However, it is an essential concept in understanding the coupling of signals and approximating transmission lines to yield intuitive understanding.

The summary that is used in the text to understand multimoding is given in Table D-2.

D.6   Fields in Lossy Media

Lossy media result in power loss of EM fields. In RF and microwave systems materials are ideally either perfect conductors or perfect dielectrics. In practice, conductors and dielectrics have finite conductivity and dielectrics have loss called dielectric relaxation loss. Dielectric relaxation loss results from loss, generally to heat, in moving charge centers in a material.

D.6.1   Lossy Dielectrics

In a lossy dielectric there can be both current flow or relaxation loss. Current flow results from applied electric field and is described by the conductivity, σ, of the material. Relaxation loss results from polarization of the material and the transfer of energy to the lattice of the material as the electric field changes direction. This is also referred to as dielectric damping. The polarization vector describes the additional storage capacity of a material to

applied electric field. The capacity is described by the electric susceptibility, χ_e, in Equation (D.14). In phasor form, loss is captured by a complex χ_e or a complex permittivity:

$$\varepsilon = \varepsilon' - \jmath\varepsilon'' = \varepsilon(1 + \chi_e). \tag{D.52}$$

Both current and relaxation respond to applied electric field. In phasor form, Maxwell's curl equation for magnetic field, Equation (D.26), becomes

$$\begin{aligned}
\nabla \times H &= \jmath\omega D + J \\
&= \jmath\omega\varepsilon E + \sigma E \\
&= \jmath\omega(\varepsilon' - \jmath\varepsilon'')E + \sigma E \\
&= \jmath\omega(\varepsilon' - \jmath\varepsilon')E + (\sigma + \omega\varepsilon'')E \\
&= \jmath\omega\left(\varepsilon' - \jmath\varepsilon'' - \jmath\frac{\sigma}{\omega}\right)E.
\end{aligned} \tag{D.53}$$

In the above, $\omega\varepsilon'$ describes the ability to store electric energy, $\omega\varepsilon'$ describes dielectric damping loss, and σ describes conductivity loss. At a single frequency, damping and conductivity losses are indistinguishable and so the term $(\omega\varepsilon'' + \sigma)$ is taken as the effective conductivity. The effective (complex) relative permittivity is

$$\varepsilon_e = \left(\varepsilon' - \jmath\varepsilon'' - \jmath\frac{\sigma}{\omega}\right)\frac{1}{\varepsilon_0}. \tag{D.54}$$

Another quantity that captures loss is the **loss tangent**, $\tan\delta$, which describes the loss to energy storage:

$$\tan\delta = \frac{\omega\varepsilon'' + \sigma}{\varepsilon'}. \tag{D.55}$$

For many materials, $\tan\delta$ is approximately independent of frequency, so it is commonly used to characterize the dielectric loss of a material.

EXAMPLE D. 3   Electromagnetic Wave Propagation in a Lossy Dielectric

Consider a plane EM wave propagating in a medium with permittivity ε and permeability μ. This problem relates to the discussion of dielectric loss in Section 4.2.2 on Page 173. The complex permittivity measured at two frequencies is experimentally characterized by the relative permittivity $(\varepsilon/\varepsilon_0)$ of the real, $\Re\{\varepsilon/\varepsilon_0\}$, and imaginary, $\Im\{\varepsilon/\varepsilon_0\}$, parts as follows:

Measured Relative Permittivity:

Frequency	Real Part	Imaginary Part
1 GHz	1.9	−0.0490
2 GHz	1.9	−0.0314

Measured Relative Permeability:

Frequency	Real Part	Imaginary Part
1 GHz	1.9	−0.001
2 GHz	1.9	−0.001

Since there is an imaginary part of the dielectric constant, there could be loss due to either dielectric damping or finite material conductivity, or both. When characterizing materials, it is only possible to distinguish the contribution to the measured imaginary part of the permittivity by considering the experimentally derived imaginary permittivity at two different frequencies.

(a) Determine the dielectric loss tangent at 1 GHz.

(b) Determine the relative dielectric damping factor at 1 GHz (the part of the permittivity due to dielectric damping).

(c) What is the conductivity of the dielectric?

(d) What is the magnetic conductivity of the medium?

Solution:

(a) At 1 GHz, $\varepsilon = \varepsilon_r \varepsilon_0$, and the measured relative permittivity is

$$\varepsilon_r = 1.9 - j0.049 = \varepsilon'_{r,e} - j\varepsilon''_{r,e}, \qquad\qquad \text{(D.56)}$$

where $\varepsilon'_{r,e}$ and $\varepsilon''_{r,e}$ are the effective (measured) values of ε'_r and ε''_r obtained from measurement. Note that dielectric conductivity $\varepsilon''_{r,e} \neq \varepsilon''_r$ (see Equation (4.9)). Now $\varepsilon'_e = \varepsilon'$, but the imaginary permittivity includes the effect of dielectric conductivity, σ, as well as dielectric damping, which is described by ε'' alone. That is,

$$\varepsilon''_e = \varepsilon'' + \frac{\sigma}{\omega} \cdot \varepsilon_r = \varepsilon''_{r,e} = 0.0480$$
$$\tan\delta = 0.0480/1.9 = 0.0258 \ .$$

(b) The imaginary relative permittivity is found by noting that $\varepsilon''_{r,e} = \varepsilon''_r + \sigma/\omega$ and $\omega\varepsilon''_r + \sigma = \omega\varepsilon''_{r,e}$.
Let $\omega_1 = 2\pi\,(1\text{ GHz})$. Then

$$\omega_2 = 2\omega_1$$
$$\omega_1\varepsilon''_{r,e}\,(1\text{ GHz}) = \omega_1\varepsilon''_r + \sigma = \omega_1\,0.0490 \quad \text{(A)}$$
$$\omega_2\varepsilon''_{r,e}\,(2\text{ GHz}) = \omega_2\varepsilon''_r + \sigma = \omega_2\,0.0314$$
$$2\omega_1\varepsilon''_e + \sigma = 2\omega_1\,(0.0314) \quad \text{(B)}$$
$$(\text{B}) - (\text{A}) \rightarrow \omega_1\varepsilon''_r = \omega_1\,[0.0628 - 0.0490]$$
$$\varepsilon''_r = 0.0138 \ \text{(at all frequencies)}\,.$$

(c) Recall that $\varepsilon''_{r,e} = \varepsilon''_r + \sigma/\omega$.
At 1 GHz, $\omega = 2\pi \times 10^9$, $\varepsilon''_{r,e} = 0.0490$, $\varepsilon''_r = 0.0138$, so

$$\sigma = \omega\,\left(\varepsilon''_{r,e} - \varepsilon''_r\right) = 2\cdot\pi\cdot10^9\,(0.0490 - 0.0138) = 0.221\cdot10^9\ \text{S/m} = 221\ \text{MS/m}$$

(d) Magnetic conductivity is nonexistent, thus magnetic conductivity $= 0.00$.

D.6.2 Lossy Conductors

Perfect conductors would have infinite conductivity, but since the conductivity of conductors is finite, EM fields penetrate to the interior of a conductor. Still the energy stored in the current is much greater than the electric energy storage capability, $\sigma \gg \omega\varepsilon'$, and the losses from conductivity are much higher than dielectric loss, $\sigma \gg \omega\varepsilon''$, so that the effective relative permittivity from Equation (D.54) is

$$
\begin{aligned}
\varepsilon_e &= \left(\varepsilon' - \jmath\varepsilon'' - \jmath\frac{\sigma}{\omega}\right)\frac{1}{\varepsilon_0} \approx \left(-\jmath\frac{\sigma}{\omega}\right)\frac{1}{\varepsilon_0} \\
&\approx \left(\frac{\sigma}{\jmath\omega}\right)\frac{1}{\varepsilon_0}.
\end{aligned}
\tag{D.57}
$$

For most conductors the permeability is μ_0 and the working permittivity is

$$
\varepsilon = \varepsilon_e\varepsilon_0 = \frac{\sigma}{\jmath\omega}.
\tag{D.58}
$$

The propagation constant of the field in a conductor is

$$
\begin{aligned}
\gamma = \alpha + \jmath\beta &\approx \jmath\omega\sqrt{\mu\varepsilon} \approx \jmath\omega\sqrt{\mu}\sqrt{\frac{\sigma}{\jmath\omega}} \\
&\approx (1+\jmath)\sqrt{\frac{\omega\mu\sigma}{2}}.
\end{aligned}
\tag{D.59}
$$

So $\alpha = \sqrt{\omega\mu\sigma/2}$. The EM field in the conductor reduces in amplitude as $e^{-\alpha x}$ after a distance x. Thus the field reduces to $1/e$ of its value after a distance called the skin depth, δ_s:

$$
\delta_s = \sqrt{\frac{2}{\omega\mu\sigma}}.
\tag{D.60}
$$

If the amplitude of the electric field at the surface of a conductor is $A(0)$, after a distance δ_s the amplitude of the electric field will have the amplitude

$$
\begin{aligned}
A(\delta_s) = e^{-\alpha x} &= e^{-\left(\sqrt{\omega\mu\sigma/2}\sqrt{2/\omega\mu\sigma}\right)} \\
&= e^{-1}.
\end{aligned}
\tag{D.61}
$$

   Table D-3 lists the skin depth and phase velocities of conductors commonly used in microwave circuits. The ideal metal conductivities come from the resistivity table on Page 844. The resistivity values here are those of a single crystal. The resistivity values of the best fabricated materials used in microwave applications are generally a factor of 2 higher. The phase velocity, v_p, is calculated from the propagation constant β and normalized to c, the speed of light in a vacuum. From Equation (D.59),

$$
\beta = \sqrt{\omega\mu\sigma/2}
\tag{D.62}
$$

Table D-3 Skin depth and phase velocity at 1 GHz of several conductors commonly used in microwave circuits. Conductivities of single-crystal metals taken from the resistivity table on Page 844. The best fabricated metals used in microwave applications generally have twice the resistivity listed.

Material	Conductivity, MS.m^{-1}	Skin Depth, μm	Phase Velocity, v_p (km.s^{-1})
1 GHz			
Aluminum (crystal)	37.7	2.59	16.3
Aluminum (2× resistivity)	18.9	3.66	23.0
Copper (single crystal)	59.6	2.06	13.0
Copper (2× resistivity)	29.8	2.92	18.3
Gold (single crystal)	45.2	2.37	14.9
Gold (2× resistivity)	22.6	3.35	21.0
Silver (single crystal)	45.2	2.01	12.6
Silver (2× resistivity)	22.6	2.84	17.8
Titanium (single crystal)	0.238	32.6	200
Titanium (2× resistivity)	0.119	46.1	290

Now $v_p = \omega/\beta$, so that

$$
\begin{aligned}
v_p &= \frac{\omega}{\beta} \\
&= \frac{\omega\sqrt{2}}{\sqrt{\omega\mu\sigma}} \\
&= \sqrt{\frac{2\omega}{\mu\sigma}}.
\end{aligned}
\tag{D.63}
$$

This can be simplified further if the relative permeability of the material is 1, as then $\mu - \mu_0 = 4\pi 10^{-7}$:

$$
\begin{aligned}
v_p &= \sqrt{\frac{2\omega}{\mu_0\sigma}} \\
&= \sqrt{\frac{4\pi f}{4\pi 10^{-7}\sigma}} \\
&= \sqrt{\frac{10^7 f}{\sigma}}.
\end{aligned}
\tag{D.64}
$$

Some notes that should be extracted from Table D-3 are, first, that the skin depth at gigahertz frequencies is close to the thickness of microstrip lines which are often only a few microns thick. Also the speed of an EM wave in a conductor is very slow. The speed of EM in copper (ideal), for example, is 0.004% of the speed of light in a vacuum.

D.7 Application of Electromagnetic Theory

EXAMPLE D. 4   Electromagnetic Calculations

This example provides a review of basic EM theory. In Figure D-1, a plane wave in free space is normally incident on a lossless medium occupying a half space with a dielectric constant of 16. Calculate the electric field reflection coefficient at the interface referred to the medium.

Solution:

Since it is not explicitly stated, assume $\mu_1 = \mu_0 = \mu_2$. (This should be done in general since μ_0 is the default value for μ, as most materials have unity relative permeability.) The characteristic impedance of a medium is $\eta = \sqrt{\mu/\varepsilon}$, and here $\eta_1 = \eta_0$. Now η is used for characteristic impedance (also called wave impedance) of a propagating wave in a medium. It is used instead of Z_0 which is generally reserved for the characteristic impedance of transmission lines. The electric field reflection coefficient is

$$\Gamma^E = \frac{\eta_1 - \eta_0}{\eta_1 + \eta_0} = \frac{\eta_1/\eta_0 - 1}{\eta_1/\eta_0 + 1}$$

$$\frac{\eta_1}{\eta_0} = \sqrt{\frac{\mu_0}{\varepsilon_1}} \cdot \sqrt{\frac{\varepsilon_0}{\mu_0}} = \sqrt{\frac{\varepsilon_0}{\varepsilon_1}} = \sqrt{\frac{1}{16}} = \frac{1}{4},$$

so

$$\Gamma^E = \frac{1/4 - 1}{1/4 + 1} = \frac{1 - 4}{1 + 4} = -\frac{3}{5} = -0.6 .$$

EXAMPLE D. 5   Intuitive Solution of Electromagnetic Problems

This example provides an intuitive way of solving an EM problem. A plane wave in free space is normally incident on a lossless medium occupying a half space with a dielectric constant of 16. What is the magnetic field reflection coefficient?

Solution:

There are several ways to answer this question:

Intuitively $|\Gamma^H| = |\Gamma^E|$, just as for a transmission line, $|\Gamma^I| = |\Gamma^V|$.

With $\vec{E}$ and $\vec{H}$ being vectors, the Poynting vector $\vec{E} \times \vec{H}$ points in the direction of propagation for a plane wave, since the reflected wave is in the opposite direction to the incident wave, thus

$$\text{sign}\left(\Gamma^H\right) = -\text{sign}\left(\Gamma^E\right).$$

Therefore $\Gamma^H = +0.6$. (Γ^E was calculated in Example 4.4.)

The formula for the magnetic field reflection coefficient is

$$\Gamma_H = \frac{\eta_1 - \eta_2}{\eta_1 + \eta_2} = \frac{1 - \eta_2/\eta_1}{1 + \eta_2/\eta_1} = \frac{1 - 1/4}{1 + 1/4} = \frac{3}{5} = 0.6 \ .$$

EXAMPLE D. 6   Impedance and Propagation Constant

A 4 GHz time-varying EM field is traveling in the $+z$ direction in Region 1 and is normally incident on another material in Region 2 (see Figure D-8). The boundary between the two regions is in the $z = 0$ plane. The permittivity of Region 1 is $\varepsilon_1 = \varepsilon_0$ and that of Region 2 is $\varepsilon_2 = (4 - j0.4)\varepsilon_0$. For both regions, $\mu_1 = \mu_2 = \mu_0$. The phasor of the forward-traveling electric field (i.e., the incident field) is $\mathbf{E}^+ = 100 \,\hat{y}$, and the phase is normalized with respect to $z = 0$. What are the wave impedance and propagation constant of Region 1?

Solution:
The wave impedance is

$$\eta = \sqrt{\frac{\mu}{\varepsilon}} = \sqrt{\frac{\mu_0}{\varepsilon_0}}\sqrt{\frac{\mu_r}{\varepsilon_r}} = \eta_0 \sqrt{\frac{M_r}{\varepsilon_r}} = 377\sqrt{\frac{1}{4}} \ \Omega$$

$$= 188.5 \ \Omega.$$

The propagation constant is

$$\begin{aligned}
\gamma &= j\omega\sqrt{\mu\varepsilon} = j\omega\sqrt{\mu_0\varepsilon_0}\sqrt{\mu_r\varepsilon_r} \\
&= j4 \cdot 2\pi \cdot 10^9 \cdot \sqrt{4\pi \cdot 10^{-7} \cdot 8.854 \cdot 10^{-12}}\sqrt{4} = j167.7 \ \text{m}^{-1}.
\end{aligned} \qquad (\text{D.65})$$
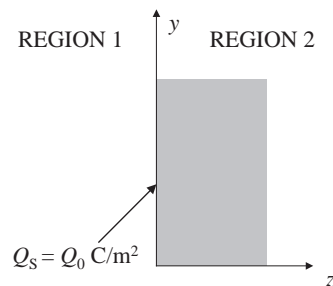
Figure D-8   Two regions used in Example D.6.

Waveguide Field Analysis

This appendix develops detailed analysis of the parallel-plate and rectangular waveguides shown in Figure E-1. It is shown that the fields have sinusoidal forms with respect to both position and time, and that multiple modes can exist if dimensions are sufficiently large. The first
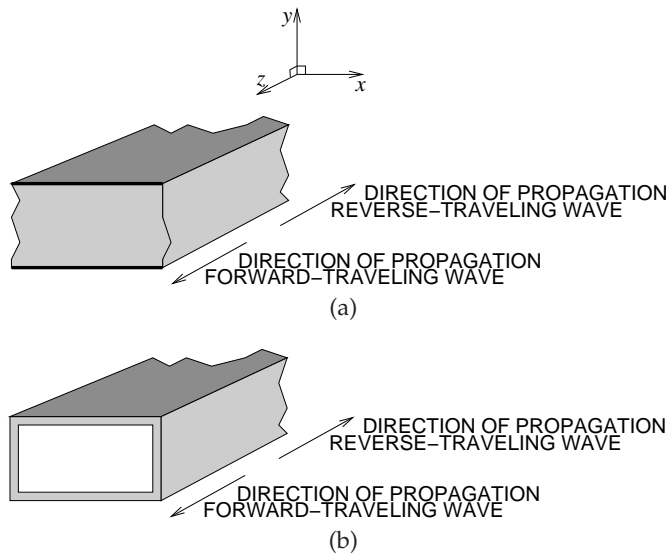
Figure E-1 Waveguides: (a) parallel-plate waveguide; and (b) rectangular waveguide.

section uses phasors and restricts propagation to one direction to simplify Maxwell's equations for subsequent use. The second section develops the wave equation and subsequent characteristics for propagation in parallel-plate waveguide. The final section undertakes a similar development for rectangular waveguides. Parallel-plate waveguide analysis provides insight into multimoding on microstrip lines and other planar interconnects. Rectangular waveguides are important transmission line structures, as they have low loss compared to transmission lines (such as microstrip and coaxial cables) that support TEM modes. They are used when handling large powers, such as in radar systems, and high frequencies, especially above 75 GHz. Rectangular waveguide achieves low loss by concentrating energy away from metal surfaces. In particular, current flow in conductors is significantly reduced. Rectangular waveguides have the special property that they do not support a TEM mode and cannot propagate an EM wave below what is called a cutoff frequency.

E.1 The Rectangular Wave Equation

In this section Maxwell's equations are put in a form that can be used in establishing the field descriptions in parallel-plate and rectangular waveguides (Figure E-1). The boundary conditions in these waveguides are established by the metal walls, and these walls match the Cartesian coordinate system. So Maxwell's equations are put in Cartesian coordinates. Simplifications about the fields can be made that relate to the positions of the metal walls. Another simplification is made by assuming that there can only be propagation in the z direction. When the wave propagates in the $+z$ direction it is called the forward-traveling wave, and when it propagates in the $-z$ direction it is called the reverse-traveling wave.

The development begins with Maxwell's equations (Equations (D.1)–(D.4) on Page 848) in a source-free region ($\rho = 0$ and $J = 0$). A further simplification comes from assuming a linear, isotropic, and homogeneous medium so that ε and μ are independent of signal level and are independent of the field direction and of position, thus

$$\nabla \times \mathbf{E} \quad = \quad \frac{\partial \mathbf{B}}{\partial t} = \mu \frac{\partial \mathbf{H}}{\partial t} \tag{E.1}$$

$$\nabla \times \mathbf{H} \quad = \quad \frac{\partial \mathbf{D}}{\partial t} = \varepsilon \frac{\partial \mathbf{E}}{\partial t} \tag{E.2}$$

$$\nabla \cdot \mathbf{D} \quad = \quad 0 = \nabla \cdot \mathbf{E} \tag{E.3}$$

$$\nabla \cdot \mathbf{B} \quad = \quad 0 = \nabla \cdot \mathbf{H} \,. \tag{E.4}$$

Taking the curl of Equation (E.1) leads to

$$\nabla \times \nabla \times \mathbf{E} = \nabla \times \mu \frac{\partial \mathbf{H}}{\partial t} = \mu \frac{\partial (\nabla \times \mathbf{H})}{\partial t}. \tag{E.5}$$

Applying the identity $\nabla \times \nabla \times \mathbf{A} = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A}$ (from Equation (A.37)) to the left-hand side of Equation (E.5), and replacing $\nabla \times \mathbf{H}$ by the right-hand side of Equation (E.2), the equation above becomes

$$-\nabla^2 + \nabla(\nabla \cdot \mathbf{E}) = \mu \frac{\partial}{\partial t} \left(\varepsilon \frac{\partial \mathbf{E}}{\partial t} \right) = -\mu \varepsilon \frac{\partial^2 \mathbf{E}}{\partial t^2}. \tag{E.6}$$

Using Equation (E.3) and assuming that ε is independent of position, this reduces to

$$-\nabla^2 \mathbf{E} = -\mu \varepsilon \frac{\partial^2 (\mathbf{E})}{\partial t^2}, \tag{E.7}$$

where

$$\nabla^2 \mathbf{E} = \frac{\partial^2 \mathbf{E}}{\partial x^2} + \frac{\partial^2 \mathbf{E}}{\partial y^2} + \frac{\partial^2 \mathbf{E}}{\partial z^2} = \nabla_t^2 \mathbf{E} + \frac{\partial^2 \mathbf{E}}{\partial z^2} \tag{E.8}$$

$$\nabla_t^2 \mathbf{E} = \frac{\partial^2 \mathbf{E}}{\partial x^2} + \frac{\partial^2 \mathbf{E}}{\partial y^2} \tag{E.9}$$

is used for fields propagating in the z direction and the subscript t indicates the transverse plane. Equation (E.9) can be put into the form of its components which will be used latter. Since

$$\mathbf{E} = E_x \hat{x} + E_y \hat{y} + E_z \hat{z}, \tag{E.10}$$

then

$$
\begin{aligned}
\nabla^2 \mathbf{E} &= \left(\frac{\partial^2 E_x}{\partial x^2} \hat{x} + \frac{\partial^2 E_y}{\partial x^2} \hat{y} + \frac{\partial^2 E_z}{\partial x^2} \hat{z} \right) + \left(\frac{\partial^2 E_x}{\partial y^2} \hat{x} + \frac{\partial^2 E_y}{\partial y^2} \hat{y} + \frac{\partial^2 E_z}{\partial y^2} \hat{z} \right) \\
&\quad + \left(\frac{\partial^2 E_x}{\partial z^2} \hat{x} + \frac{\partial^2 E_y}{\partial z^2} \hat{y} + \frac{\partial^2 E_z}{\partial z^2} \hat{z} \right)
\end{aligned} \tag{E.11}
$$

$$
\begin{aligned}
\nabla^2 \mathbf{E} &= \left(\frac{\partial^2 E_x}{\partial x^2} + \frac{\partial^2 E_x}{\partial y^2} + \frac{\partial^2 E_x}{\partial z^2} \right) \hat{x} + \left(\frac{\partial^2 E_y}{\partial x^2} + \frac{\partial^2 E_y}{\partial y^2} + \frac{\partial^2 E_y}{\partial z^2} \right) \hat{y} \\
&\quad + \left(\frac{\partial^2 E_z}{\partial x^2} + \frac{\partial^2 E_z}{\partial y^2} + \frac{\partial^2 E_z}{\partial z^2} \right) \hat{z},
\end{aligned} \tag{E.12}
$$

and

$$
\begin{aligned}
\nabla_t^2 \mathbf{E} &= \left(\frac{\partial^2 E_x}{\partial x^2} + \frac{\partial^2 E_x}{\partial y^2} \right) \hat{x} + \left(\frac{\partial^2 E_y}{\partial x^2} + \frac{\partial^2 E_y}{\partial y^2} \right) \hat{y} \\
&\quad + \left(\frac{\partial^2 E_z}{\partial x^2} + \frac{\partial^2 E_z}{\partial y^2} \right) \hat{z}.
\end{aligned} \tag{E.13}
$$

Invoking the phasor form, $\partial / \partial t$ is replaced by $\jmath \omega$, and assuming propagation only in the z direction (so that there is an assumed $e^{(\jmath \omega t - \gamma z)}$ dependence of the fields), Equation (E.7) further reduces to

$$-\nabla^2 \mathbf{E} = - \left(\nabla_t^2 \mathbf{E} + \frac{\partial^2 \mathbf{E}}{\partial z^2} \right) = \nabla_t^2 \mathbf{E} + \gamma^2 \mathbf{E} = -(\jmath \omega)^2 \mu \varepsilon \mathbf{E} = -k^2 \mathbf{E} , \tag{E.14}$$

where $k = \jmath\omega\sqrt{\mu\varepsilon}$ is known as the wavenumber, that is,

$$\nabla_t^2 \mathbf{E} = -(\gamma^2 + k^2)\mathbf{E}. \tag{E.15}$$

A similar expression can be derived for the magnetic field:

$$\nabla_t^2 \mathbf{H} = -(\gamma^2 + k^2)\mathbf{H}. \tag{E.16}$$

Equations (E.15) and (E.16) are called wave equations, or Helmholtz equations, for phasor fields propagating in the z direction. Equations (E.15) and (E.16) are usually written as

$$\nabla_t^2 \mathbf{E} = -k_c^2 \mathbf{E} \tag{E.17}$$
$$\nabla_t^2 \mathbf{H} = -k_c^2 \mathbf{H}, \tag{E.18}$$

where

$$k_c^2 = \gamma^2 + k^2 \tag{E.19}$$

is called the cutoff wavenumber. For propagating waves in a lossless medium, $\gamma = \jmath\beta$, where β is the phase constant and is real, and so

$$\beta = \pm\sqrt{k^2 - k_c^2}\,. \tag{E.20}$$

Equations (E.17) and (E.18) describe the transverse fields between the conducting plates of the parallel-plate and rectangular waveguide having a $e^{(\jmath\omega t - \gamma z)}$ dependence. The general form of the solution of these equations is a sinusoidal wave moving in the z direction. Boundary conditions, resulting from the charges and current supported by the plates, further constrain the solutions. Equation (E.1) becomes

$$\nabla \times \mathbf{E} = \jmath\omega\mu\mathbf{H}. \tag{E.21}$$

In rectangular coordinates, which are the appropriate coordinate system to use with a parallel-plate waveguide, $\mathbf{E} = E_x\hat{x} + E_y\hat{y} + E_z\hat{z}$ and $\mathbf{H} = H_x\hat{x} + H_y\hat{y} + H_z\hat{z}$, and Equation (E.21) becomes

$$\frac{\partial E_z}{\partial y} + \gamma E_y = -\jmath\omega\mu H_x \tag{E.22}$$

$$-\frac{\partial E_z}{\partial x} - \gamma E_x = -\jmath\omega\mu H_y \tag{E.23}$$

$$\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = -\jmath\omega\mu H_z. \tag{E.24}$$

Similarly

$$\frac{\partial H_z}{\partial y} + \gamma H_y = \jmath\omega\varepsilon E_x \tag{E.25}$$

$$-\frac{\partial H_z}{\partial x} - \gamma H_x = \jmath\omega\varepsilon E_y \tag{E.26}$$

$$\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = \jmath\omega\varepsilon E_z. \tag{E.27}$$

Solving Equations (E.22) and (E.22) results in

$$E_x = \frac{-1}{\gamma^2 + k^2}\left(\gamma\frac{\partial E_z}{\partial x} + \jmath\omega\mu\frac{\partial H_z}{\partial y}\right) \tag{E.28}$$

$$E_y = \frac{1}{\gamma^2 + k^2}\left(-\gamma\frac{\partial E_z}{\partial y} + \jmath\omega\mu\frac{\partial H_z}{\partial x}\right) \tag{E.29}$$

$$H_x = \frac{1}{\gamma^2 + k^2}\left(-\gamma\frac{\partial H_z}{\partial x} + \jmath\omega\varepsilon\frac{\partial E_z}{\partial y}\right) \tag{E.30}$$

$$H_y = \frac{-1}{\gamma^2 + k^2}\left(\gamma\frac{\partial H_z}{\partial y} + \jmath\omega\varepsilon\frac{\partial E_z}{\partial x}\right). \tag{E.31}$$

Now that the fields are in the appropriate forms, classification of possible solutions or modes can be developed for the parallel-plate and rectangular waveguides. At this stage the following simplifications have been made to Maxwell's equations to get them into the form of Equations (E.28)–(E.31):

- Restriction of propagation to the $+z$ and $-z$ directions

- Assuming that ε and μ are constants

- Putting the wave equations in rectangular form so that boundary conditions established by metal walls can be easily applied

The solutions of Equations (E.28)–(E.31) will result in equations for the E and H components that vary sinusoidally with x, y, and z positions. A major expression to be developed is for the cutoff wavenumber, k_c.

E.2 Parallel-Plate Waveguide

In this appendix the propagating electromagnetic fields between the plates of a parallel-plate waveguide (Figure E-2) are derived. Parallel plate waveguides are not used intentionally as a transmission line but occur in many situations resulting in undesirable effects. Examining them, however, provides insight into how metallic walls can guide EM fields without the explicit use of two conductors (e.g., in a coaxial line or in a microstrip line) to provide current and voltage support. Parallel plates also appear in printed circuit boards, and many undesirable effects that occur at higher frequencies can be traced to wave guiding.

E.2.1 TEM Mode

In the TEM mode, all of the E and H field components are in the plane transverse to the direction of propagation, that is, $E_z = 0 = H_z$. Thus Equations (E.28)–(E.31) imply that $\gamma^2 + k^2 = 0$ if E_x, E_y, H_x, and H_y are to possibly have a finite value. If $\gamma^2 + k^2 = 0$ (so that $\gamma = \pm\jmath k$) then Equations (E.15) and (E.16) imply that E_x, E_y, H_x, and H_y cannot vary with position
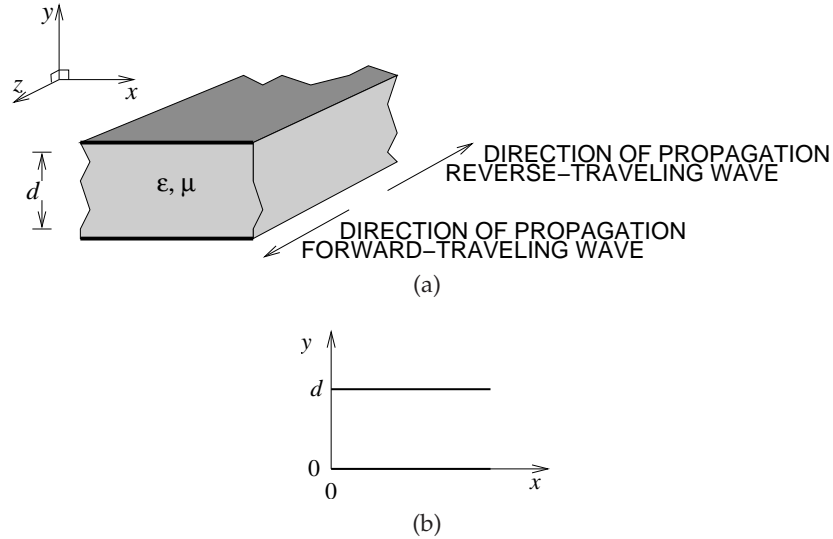
Figure E-2 Parallel-plate waveguide: (a) three-dimensional view; and (b) cross-sectional (transverse) view.

in the transverse plane (i.e., with respect to x and y). Thus E_x, E_y, H_x, and H_y must be constant between the plates. Furthermore, boundary conditions require that $H_y = 0$ and $E_x = 0$ at the conductor. So in the TEM parallel-plate mode only E_y and H_x exist, and they are constant. Equation (E.22) leads to

$$H_x = \frac{\gamma E_y}{\jmath \omega \mu} = \pm \frac{\jmath \omega \sqrt{\mu \varepsilon}}{\jmath \omega \mu} E_y = \pm \sqrt{\frac{\varepsilon}{\mu}} E_y = \eta E_y, \qquad (E.32)$$

where the plus sign describes forward-traveling fields (propagating in the $+z$ direction) and the minus sign describes backward-traveling fields (propagating in the $-z$ direction). The quantity $\eta = \sqrt{\varepsilon/\mu}$ is called the wave impedance. This field variation is shown in Figure 5-15(a) on Page 268.

E.2.2 TM Mode

The **Transverse Magnetic Mode** (**TM**) is characterized by $H_z = 0$ and $E_z \neq 0$. Another restriction that will be used here is that there is no variation of the fields in the x direction. Thus Equation (E.15) becomes

$$\frac{d^2 E_z}{dy^2} = -k_c^2 E_z. \qquad (E.33)$$

The solution to Equation (E.33) is

$$E_z = E_0 \sin(k_c y) + E_1 \cos(k_c y). \qquad (E.34)$$

To find the coefficients E_0 and E_1, boundary conditions are applied so that E_z is zero at $y = 0$ and $y = d$; that is,

$$E_z|_{y=0} = 0 = E_1 \tag{E.35}$$

and

$$E_z|_{y=d} = 0 = E_0 \sin{(k_c d)}. \tag{E.36}$$

This requires that $\sin{(k_c d)} = 0$, since the assumption is that there is an E_z field, and so $E_0 \neq 0$. The condition that $\sin{(k_c d)} = 0$ implies that there are discrete values of k_c:

$$k_c = m\pi/d \quad m = 1, 2, 3, \dots . \tag{E.37}$$

Each value of k_c corresponds to a different mode and m is the mode index so that the mth mode is the TM_m mode and m indicates the number of half-sinusoidal variations of the fields in the y direction. There must be at least one half-sinusoidal variation for a TM mode to exist, so the frequency of the field must be sufficiently high that it is possible for the field to have such a spatial variation. The fastest the field can curl is related to the wavelength, λ, of the EM signal in the medium (without walls). So the TM_m mode cannot propagate if the wavelength of the signal is such that $\lambda > \lambda_c$, where $\lambda_c (= m\pi/d)$ is called the critical wavelength of the mode.

Substituting the above results and assumptions (e.g., $\partial/\partial x = 0$) in Equations (E.28)–(E.31), and repeating Equation (E.36),

$$H_z = 0 \quad E_x = 0 \quad H_y = 0 \tag{E.38}$$

$$E_z = E_0 \sin{(k_c y)} \tag{E.39}$$

$$E_y = -\frac{\gamma}{k_c^2}\frac{dE_z}{dy} = -\frac{\gamma}{k_c}E_0 \cos{(k_c y)} \tag{E.40}$$

$$H_x = -\frac{\jmath\omega\varepsilon}{k_c^2}\frac{dE_z}{dy} = -\frac{\jmath\varepsilon}{k_c}E_0 \cos{(k_c y)}. \tag{E.41}$$

There are an infinite number of TE modes identified by the index m, which determines the cutoff wavenumber, k_c, of the particular mode. The propagation constant of the mth mode is

$$\gamma = \sqrt{k_c^2 - k^2} = \sqrt{(m\pi/d)^2 - \omega^2\mu\varepsilon}. \tag{E.42}$$

Propagation is only possible if γ has an imaginary component (or in lossless media, $\gamma = \jmath\beta$). Thus the modes have a cutoff frequency, f_c, below which propagation is not possible:

$$f_c = \frac{1}{2\pi}\frac{k_c}{\sqrt{\mu\varepsilon}} = \frac{1}{2\pi}\frac{m\pi}{d\sqrt{\mu\varepsilon}} = \frac{m\nu}{2d}, \tag{E.43}$$

where $\nu = 1/\sqrt{\mu\varepsilon}$ is the velocity of a TEM mode in the medium. A cutoff wavelength can also be defined:

$$\lambda_c = \frac{\nu}{f_c} = \frac{2d}{m}. \tag{E.44}$$

The wavelength of the mode, called the guide wavelength, λ_g, is given by

$$\lambda_g = \frac{2\pi}{\beta} = \frac{\lambda}{\sqrt{1 - (f_c/f)^2}}, \tag{E.45}$$

where λ is the wavelength of a TEM mode in the medium: $\lambda = \nu/f$. The phase velocity of the modes is also dependent on the mode index m through the cutoff frequency,

$$v_p = \frac{\omega}{\beta} = \frac{\nu}{\sqrt{1 - (f_c/f)^2}} \tag{E.46}$$

and the group velocity is

$$v_g = \frac{d\omega}{d\beta} = \nu\sqrt{1 - (f_c/f)^2}. \tag{E.47}$$

The phase velocity, v_p, of a TM mode is greater than ν, the speed of light in the medium, while the group velocity, v_g, is slower than ν. The group velocity is the velocity at which energy is transmitted and can never be faster than the speed of light, c. The phase velocity can easily be greater than c. The TM mode field variation is shown in Figure 5-15(a) on Page 268.

E.2.3 TE Mode

The **Transverse Magnetic Mode** (**TE**) is characterized by $E_z = 0$ and $H_z \neq 0$. Following the same development as for the TM mode, the TE mode fields are

$$\begin{align}
E_z &= 0, \quad E_y = 0, \quad H_x = 0 \tag{E.48}\\
H_z &= H_0 \cos(k_c y) \tag{E.49}\\
E_x &= -\frac{\jmath\omega\mu}{k_c} H_0 \sin(k_c y) \tag{E.50}\\
H_y &= \frac{\jmath\beta}{k_c} H_0 \sin(k_c y). \tag{E.51}
\end{align}$$

The equations for k_c, v_p, v_g, f_c, λ_c, and λ_g are the same as for the TM mode considered in the previous section. The TE mode field variation is shown in Figure 5-15(a) on Page 268.
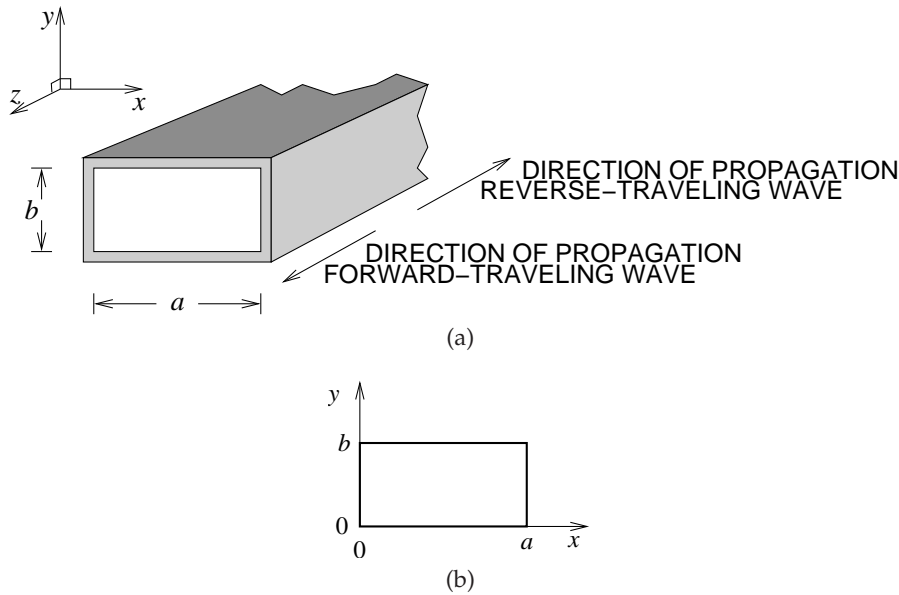
Figure E-3 Rectangular waveguide: (a) three-dimensional view; and (b) transverse crosssection with internal dimensions designated a and b.

E.3 Rectangular Waveguide

A rectangular waveguide is shown in Figure E-3. Rectangular waveguides guide EM energy between four connected electrical walls, and there is little current created on the walls. As a result resistive losses are quite low, much lower than can be achieved using coaxial lines. One of the major uses of rectangular waveguide is when losses must be kept to a minimum, so that rectangular waveguide is used in very high-power situations such as radar at a few gigahertz and above. At higher frequencies the loss of coaxial lines becomes large, and it also becomes difficult to build small-diameter coaxial lines at 100 GHz and above. As a result, rectangular waveguide is nearly always used above 100 GHz. There are many low to medium-power legacy systems that use rectangular waveguide above 1 GHz.

Rectangular waveguide supports many different modes, but it does not support the TEM mode. The modes are categorized as being either TM or TE, denoting whether all of the magnetic fields are perpendicular to the direction of propagation (i.e., transverse magnetic fields) or whether all of the electric fields are perpendicular to the direction of propagation (i.e., they are transverse electric fields). As with other transmission line structures, dimensions of the waveguide can be chosen so that only one mode can propagate. With more than one mode propagating, the different components of a signal would travel at different speeds and thus combine at a load

incoherently since the ratio of the energy in the modes would vary (usually) randomly.

The TE and TM field descriptions are derived from the solution of differential equations—Maxwell's equations—subject to boundary conditions. The general solutions for rectangular systems are sinewaves and there are possibly many discrete solutions. The nomenclature that has developed over the years to classify modes references the number of variations in the x direction, using the index m, and the number of variations in the y direction, using the index n. So there are TE_{mn} and TM_{mn} modes, and dimensions are usually selected small enough so that only the TE_{10} can propagate.

E.3.1   TM Waves

The development of the field descriptions for the TM modes begins with the rectangular wave equations derived in Section E.1. Transverse magnetic waves have zero H_z, but nonzero E_z. The differential equation governing E_z is, in rectangular coordinates, (from Equations (E.13) and (E.13))

$$\nabla_t^2 E_z \quad = \quad \frac{\partial^2 E_z}{\partial x^2} + \frac{\partial^2 E_z}{\partial y^2} = -k_c^2 E_z. \tag{E.52}$$

Using a separation of variables procedure, this equation has the solution

$$E_z = (A' \sin k_x x + B' \cos k_x x)(C' \sin k_y y + D' \cos k_y y), \tag{E.53}$$

where

$$k_x^2 + k_y^2 = k_c^2. \tag{E.54}$$

The perfectly conducting boundary at $x = 0$ requires $B' = 0$ to produce $E_z = 0$ there. Similarly the ideal boundary at $y = 0$ requires $D' = 0$. Replace $A'C'$ by a new constant A and

$$E_z = A \sin k_x x \sin k_y y . \tag{E.55}$$

The axial electric field, E_z, must also be zero at $x = a$ and $y = b$. This can only be so (except for the trivial solution $A = 0$) if $k_x a$ is an integral multiple of π:

$$k_x a = m\pi, \quad m = 1, 2, 3, \dots . \tag{E.56}$$

Similarly, for E_z to be zero at $y = b$, $k_y b$ must also be a multiple of π:

$$k_y b = n\pi, \quad n = 1, 2, 3, \dots . \tag{E.57}$$

So the cutoff condition of the TM wave with m variations in x and in y (designated TM_{mn}) is found from Equation (E.54):

$$\omega_{c_{m,n}} \quad = \quad \frac{k_{c_{m,n}}}{\sqrt{\mu\varepsilon}} = \frac{1}{\sqrt{\mu\varepsilon}} \left[\left(\frac{m\pi}{a} \right)^2 + \left(\frac{n\pi}{b} \right)^2 \right]^{1/2}. \tag{E.58}$$

Since k_c^2 is $k^2 - \beta^2$, attenuation for frequencies below the cutoff frequency of a given mode, and the phase constant for frequencies above the cutoff frequency, have the same forms as for the parallel-plate guiding system:

$$\alpha = k_{c_{m,n}} \sqrt{1 - \left(\frac{\omega}{\omega_{c_{m,n}}}\right)^2}, \quad \omega < \omega_{c_{m,n}} \tag{E.59}$$

$$\beta = k\sqrt{1 - \left(\frac{\omega_{c_{m,n}}}{\omega}\right)^2}, \quad \omega > \omega_{c_{m,n}}. \tag{E.60}$$

Phase and group velocities also have the same forms as before.

The remaining field components of the TM_{mn} wave are found with $H_z = 0$ and E_z from Equation (E.55) and Equations (E.28)–(E.31):

$$E_x = -\frac{\jmath\beta k_x}{k_{c_{m,n}}^2} A \cos k_x x \sin k_y y \tag{E.61}$$

$$E_y = -\frac{\jmath\beta k_y}{k_{c_{m,n}}^2} A \sin k_x x \cos k_y y \tag{E.62}$$

$$H_x = \frac{\jmath\omega\varepsilon k_y}{k_{c_{m,n}}^2} A \sin k_x x \cos k_y y \tag{E.63}$$

$$H_y = -\frac{\jmath\omega\varepsilon k_x}{k_{c_{m,n}}^2} A \cos k_x x \sin k_y y. \tag{E.64}$$

The spatial field variations depend on the x and y cutoff wavenumbers, k_x and k_y, which in turn depend on the mode index and the dimensions of the cross-sectional waveguide.

The lowest-order TM mode is the TM_{11} mode, with $m = 1$ and $n = 1$, and this has the minimum variation of the fields; these are shown in Figure E-4.

E.3.2 TE Waves

Transverse electric waves have zero E_z and nonzero H_z so that, in rectangular coordinates,

$$\nabla_t^2 H_z = \frac{\partial^2 H_z}{\partial x^2} + \frac{\partial^2 H_z}{\partial y^2} = -k_c^2 H_z. \tag{E.65}$$

Solution using the separation of variables technique gives

$$H_z = (A'' \sin k_x x + B'' \cos k_x x)(C'' \sin k_y y + D'' \cos k_y y), \tag{E.66}$$

where

$$k_c^2 = k_x^2 + k_y^2. \tag{E.67}$$
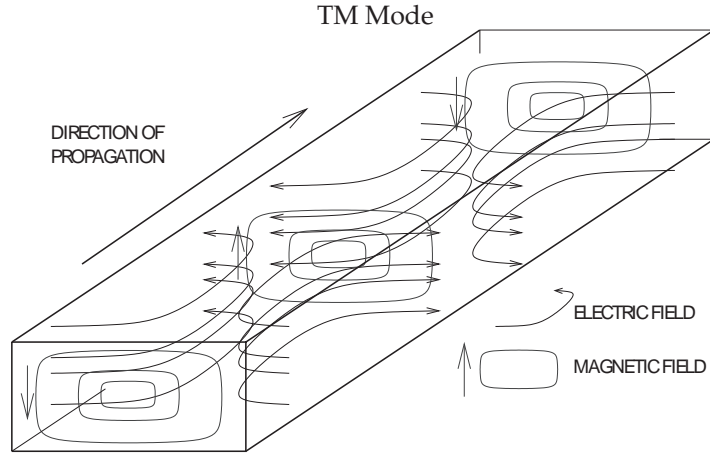
TM Mode

Figure E-4 Electric and magnetic field distribution for the lowest-order TM mode.

Imposition of boundary condition in this case is a little less direct, but the electric field components are

$$E_x = -\frac{\jmath\omega\mu}{k_c^2}\frac{\partial H_z}{\partial y} \tag{E.68}$$

$$= -\frac{\jmath\omega\mu k_y}{k_c^2}\left(A''\,\sin\,k_x x + B''\,\cos\,k_x x\right)\left(C''\,\cos\,k_y y - D''\,\sin\,k_y y\right)$$

$$E_y = \frac{\jmath\omega\mu}{k_c^2}\frac{\partial H_z}{\partial x} \tag{E.69}$$

$$= \frac{\jmath\omega\mu k_x}{k_c^2}\left(A''\,\cos\,k_x x - B''\,\sin\,k_x x\right)\left(C''\,\sin\,k_y y + D''\,\cos\,k_y y\right).$$

For E_x to be zero at $y = 0$ for all x, $C'' = 0$, and for $E_y = 0$ at $x = 0$ for all y, $A'' = 0$. Defining $B''D'' = B$, then

$$H_z = B\,\cos\,k_x x\,\cos\,k_y y. \tag{E.70}$$

E_y is zero at $x = a$ so that $k_x a$ must be a multiple of π:

$$k_x a = m\pi, \quad m = 1, 2, 3, \ldots . \tag{E.71}$$

Also, E_x is zero at $y = b$, so that $k_y b$ must zero (so that E_x is always zero) or that it is a multiple of π:

$$k_y b = n\pi \quad n = 0, 1, 2, 3, \ldots . \tag{E.72}$$

The forms of the transverse electric field with the derived simplifications
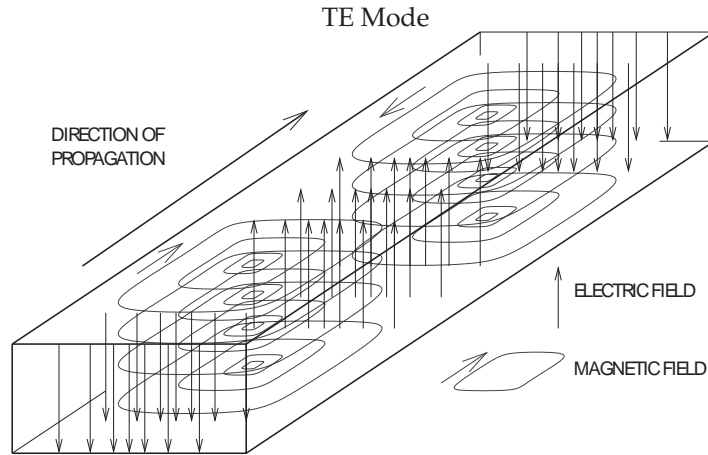
TE Mode

Figure E-5 Electric and magnetic field distribution for the lowest-order TE mode.

are

$$E_x = \frac{\jmath\omega\mu k_y}{k_{c_{m,n}}^2} B \ \cos \ k_x x \ \sin \ k_y y \qquad\qquad \text{(E.73)}$$

$$E_y = \frac{\jmath\omega\mu k_x}{k_{c_{m,n}}^2} B \ \sin \ k_x x \ \cos \ k_y y. \qquad\qquad \text{(E.74)}$$

Corresponding transverse magnetic field components are

$$H_x = \frac{\jmath\beta k_x}{k_{c_{m,n}}^2} B \ \sin \ k_x x \ \cos \ k_y y \qquad\qquad \text{(E.75)}$$

$$H_y = \frac{\jmath\beta k_y}{k_{c_{m,n}}^2} B \ \cos \ k_x x \ \sin \ k_y y. \qquad\qquad \text{(E.76)}$$

The lowest-order TE mode is the TE_{10} mode (with $m = 1$ and $n = 0$) and this has the minimum variation of the fields; these are shown in Figure E-5.

So the lowest-order TE mode, the TE_{10} mode, has one variation of the fields. This is the mode used in rectangular waveguide. The mode has a cutoff frequency which is when the wavelength (in the medium, free-space wavelength if the guide is air-filled) is twice the a dimension of the waveguide (see Figure E-3). The next higher-order mode appears when it is possible for a variation in the y direction. This occurs at a frequency corresponding to the b dimension being one-half wavelength in the medium. Consequently, with $b = 2a$, there is approximately one-half octave of available bandwidth over which the waveguide can be used. It is not practical to operate immediately above the cutoff frequency as wave impedances are too high. Neither is it desirable to operate at a frequency

(a)   (b)   (c)
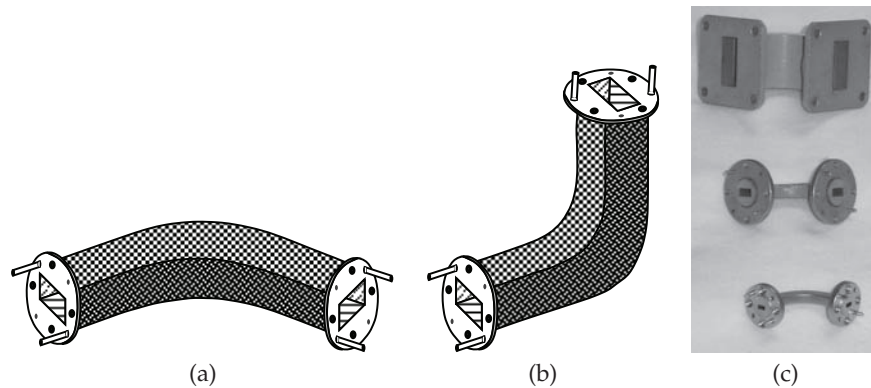
Figure E-6   Rectangular waveguide bends: (a) E-plane bend; (b) H-plane bend; and (c) photograph of various bends: top is X-band (8–12 GHz), middle is Ku-band (12–18 GHz), bottom is Ka-band (26–40 GHz).

immediately below the frequency at which the next-higher-order mode will propagate. This is because small discontinuities can launch the higher-order mode. Since rectangular waveguide is useful over a relatively narrow frequency range, standard dimensions have been developed, as listed in Table E-1. Waveguide is referred to by its waveguide standard number (its WR designation), however, the old letter designations of bands are still used.

E.3.3   Rectangular Waveguide Components

Rectangular waveguide components require considerable machining, but the equivalents of many of the components that are available in microstrip can be realized. Invariably the lowest-order TE mode is used. This is the TE_{10} mode, and the field configuration is as shown in Figure E-5. The characteristic of this mode is that the E-field is transverse to the direction of propagation. Many components have particular orientations to the planes of the E and H fields. Consider the rectangular waveguide bends shown in Figure E-6. The bend in Figure E-6(a) is called an E-plane bend, or E-bend, as the axis of the waveguide (which is in the direction of propagation) always remains parallel to the E field. With the H-plane bend, or H-bend, in Figure E-6(b), the axis of the waveguide remains parallel to the H field.

In building circuits using rectangular waveguide, it is frequently necessary to rotate and twist the waveguide so that sections can be joined. Bends enable this, but **twists** (as shown in Figure E-7) are also used. Sometimes it is necessary to use **flexible waveguide** sections (see Figure E-8). Waveguide **tees** are used to split and combine signals. There are both E-plane and H-plane versions as there were for **bends** (see Figure E-9).

There are a wide variety of waveguide components. The components are developed from EM field considerations and not derived from current and

Table E-1 Waveguide bands, operating frequencies, and internal dimensions. Waveguide dimensions specified in inches (multiply by 25.4 mm/inch to convert to millimeters). The number in the WR designation is the waveguide long internal dimension in hundreds of an inch. The TE_{10} mode cutoff frequency is when the long dimension is one-half wavelength long (the free space wavelength if vacuum- or air-filled, or modified by the square-root of the permittivity if the waveguide is dielectric filled).

Band	EIA Waveguide Band	Frequency Limits (GHz)	Internal Dimensions ($a \times b$, inches)	TE_{10} Cutoff (GHz)
	WR–2300	0.32–0.49	23.000×11.500	0.257
	WR–2100	0.35–0.53	21.000×10.500	0.281
	WR–1800	0.43–0.62	18.000×9.000	0.328
	WR–1500	0.49–0.74	15.000×7.500	0.393
	WR–1150	0.64–0.96	11.500×5.750	0.513
	WR–1000	0.75–1.1	9.975×4.875	0.592
	WR–770	0.96–1.5	7.700×3.385	0.766
	WR-650	1.12–1.70	6.500×3.250	0.908
R band	WR-430	1.70–2.60	4.300×2.150	1.37
D band	WR-340	2.20–3.30	3.400×1.700	1.74
S band	WR-284	2.60–3.95	2.840×1.340	2.08
E band	WR-229	3.30–4.90	2.290×1.150	2.58
G band	WR-187	3.95–5.85	1.872×0.872	3.15
F band	WR-159	4.90–7.05	1.590×0.795	3.71
C band	WR-137	5.85–8.20	1.372×0.622	4.30
H band	WR-112	7.05–10.00	1.122×0.497	5.26
X band	WR-90	8.2–12.4	0.900×0.400	6.56
Ku band	WR-62	12.4–18.0	0.622×0.311	9.49
K band	WR-51	15.0–22.0	0.510×0.255	11.6
K band	WR-42	18.0–26.5	0.420×0.170	14.1
Ka band	WR-28	26.5–40.0	0.280×0.140	21.1
Q band	WR-22	33–50	0.224×0.112	26.3
U band	WR-19	40–60	0.188×0.094	31.4
V band	WR-15	50–75	0.148×0.074	39.9
E band	WR-12	60–90	0.122×0.061	48.4
W band	WR-10	75–110	0.100×0.050	59.0
F band	WR-8	90–140	0.080×0.040	73.8
D band	WR-6	110–170	0.0650×0.0325	90.8
G band	WR-5	140–220	0.0510×0.0255	116
	WR-4	170–260	0.0430×0.0215	137
	WR-3	220–325	0.0340×0.0170	174
Y band	WR-2	325–500	0.0200×0.0100	295
	WR-1.5	500–750	0.0150×0.0075	393
	WR-1	750–1100	0.0100×0.0050	590

(a)   (b)
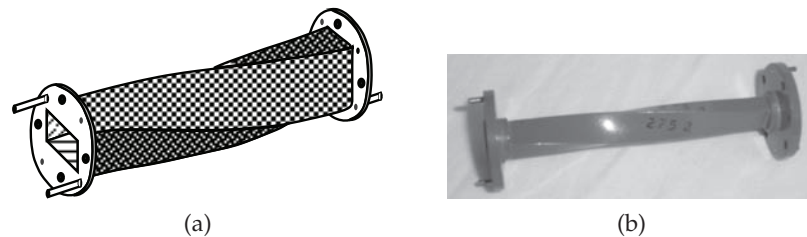
Figure E-7 Rectangular waveguide twist: (a) diagrammatic view; and (b) photograph.

Figure E-8   Photograph of a flexible rectangular waveguide section.

voltage circuits. For example, a **termination** in a rectangular waveguide is realized using a resistive wedge of material as shown in Figure E-10(a). This provides a termination with a lower reactive component than would be obtained with a lumped resistor placed at the end of the line. The matched load absorbs all of the power in the traveling wave incident on it. The functional component is lossy material, often shaped as a wedge or tall pyramid, that absorbs power over a distance corresponding, perhaps, to one-half wavelength or longer. So while the characteristic impedance of a wave on the rectangular waveguide varies with frequency, the termination is always matched to this impedance. A high-power waveguide matched load is shown in Figure E-10(b). This component uses the structure illustrated in Figure E-10(a) and has fins to aid the radiation of dissipated heat. A **waveguide attenuator** is realized by introducing resistive material, as shown in Figure E-10(c). This introduces a section of line with a high attenuation coefficient. By controlling the depth of the resistive vane, as shown in Figure E-10(d), a **variable attenuator** is obtained.

A **circulator** uses a special property of magnetized ferrites (having anisotropic permeability) which provides a preferred direction of EM propagation. A rectangular waveguide circulator is shown in Figure E-11. Other components commonly encountered are the waveguide switch (see Figure E-12), the **coaxial-to-waveguide adaptor** (see Figure E-13), and the waveguide horn antenna (see Figure E-15).

Many of the microstrip components introduced in the text rely on transmission line effects, and the functionality is not specific to microstrip.
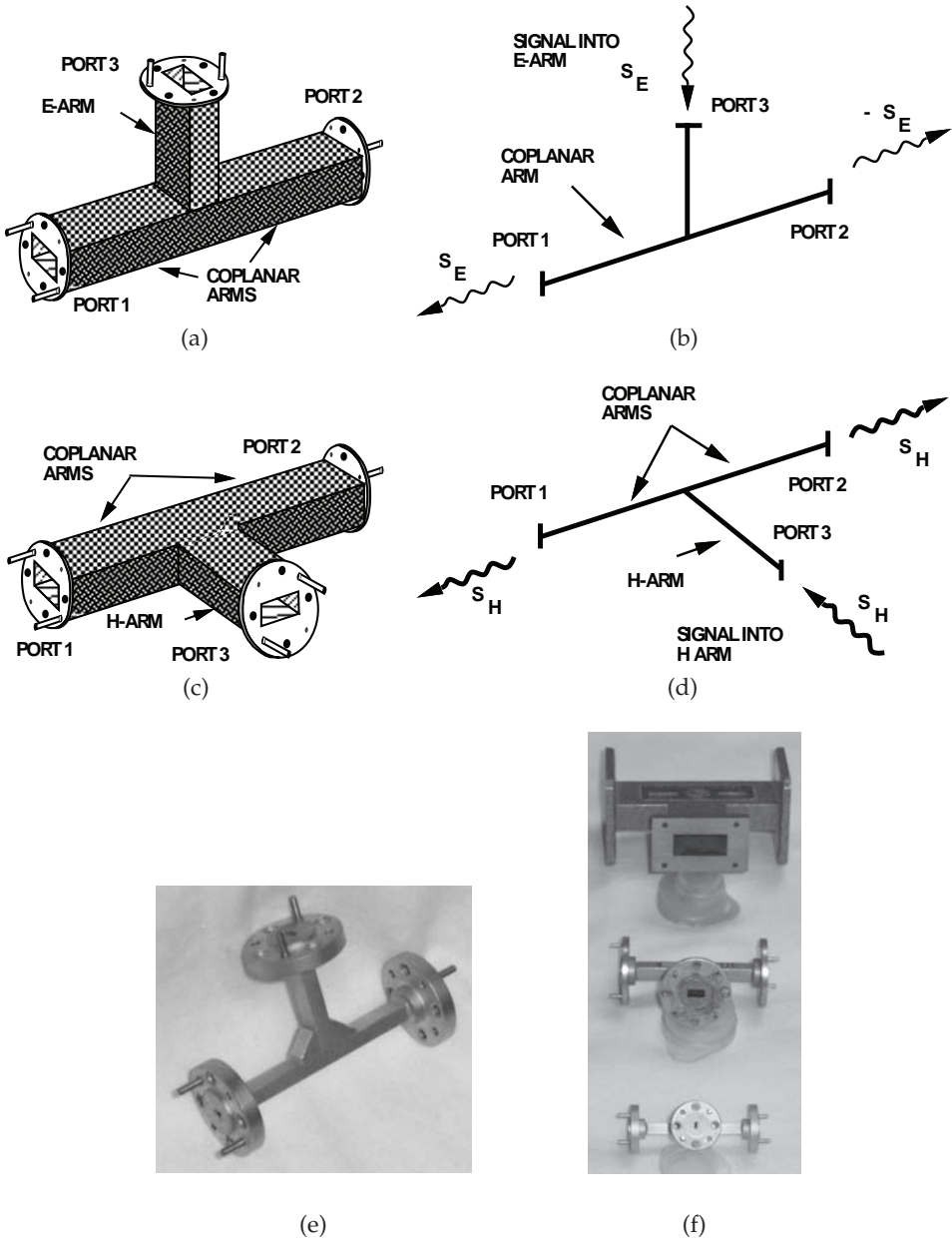
Figure E-9 Rectangular waveguide tees: (a) three-dimensional representation of an E-plane tee; (b) description of the signal flow in an E-plane Tee; (c) three-dimensional representation of an H-plane tee; (d) description of the signal flow in an H-plane tee; (e) photograph of an E-plane tee; and (f) photograph of waveguide tees for different waveguide bands (top, X-band H-plane tee; middle, Ku-band H-plane tee; bottom Ka-band E-plane tee).
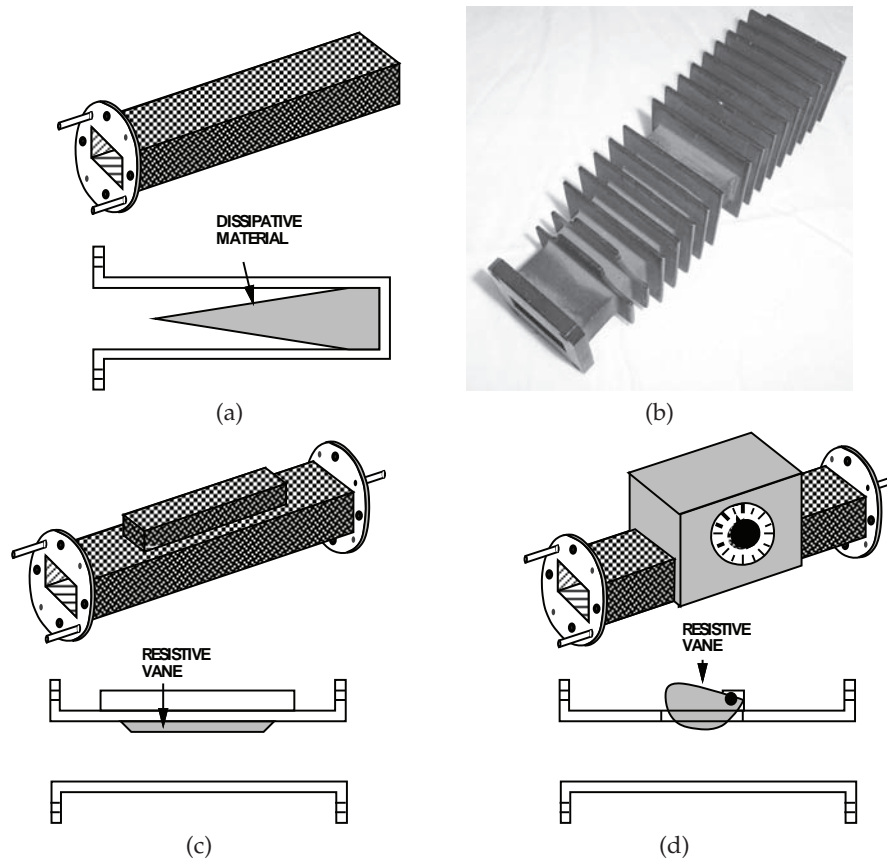
Figure E-10   Terminations and attenuators in rectangular waveguide: (a) coaxial line termination; (b) microstrip termination; (c) coaxial line attenuator; and (d) microstrip attenuator

Microstrip directional couplers were introduced in Section 9.7 on Page 483. These were based on two closely spaced parallel microstrip lines. However, it is only necessary to couple two transmission lines to create a **directional coupler**. A rectangular waveguide directional coupler is shown in Figure E-16. Here the two transmission lines, the rectangular waveguides, are coupled by slots in the common wall of the guides. In Figure E-16(b) the EM wave from the bottom waveguide leaks into the top waveguide through the coupling slots. A quick check on phasing indicates that the coupled wave in the reverse direction is canceled. Meanwhile, in the forward-traveling direction there is constructive interference of the coupled EM wave.

Variable elements available in rectangular waveguide include the micrometer **tuner**, shown in Figure E-17. The tuner shown in Figure E-17(a) typically moves a reactive element along the waveguide. One example
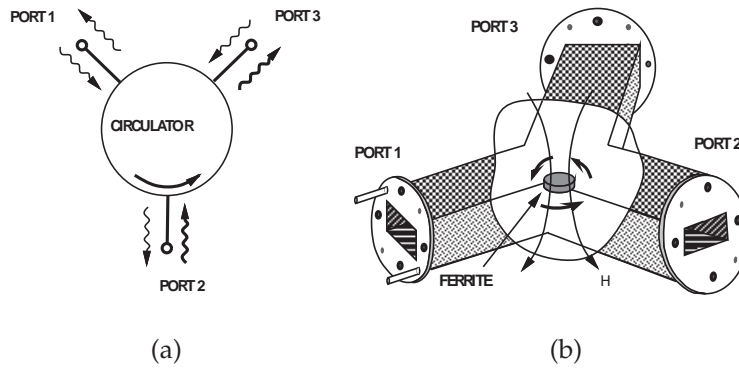
Figure E-11 Waveguide circulator: (a) schematic; and (b) three-dimensional representation showing H field lines magnetizing a ferrite disk.

Figure E-12 Waveguide switch.
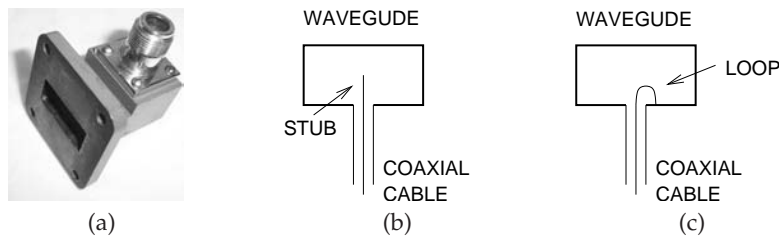
Figure E-13 Coaxial transmission line to rectangular waveguide adaptor: (a) photograph; (b) adaptor using a coupling stub; and (c) adaptor using a coupling loop.

is the movable short circuit shown in Figure E-17(b). Another variable element used in tuning is the waveguide slide tuner, shown in Figure E-17(c). Here a slot is cut in the long wall of the waveguide and a metal probe
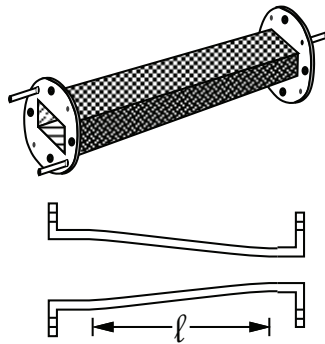
Figure E-14 Rectangular waveguide taper connecting one waveguide series to another.
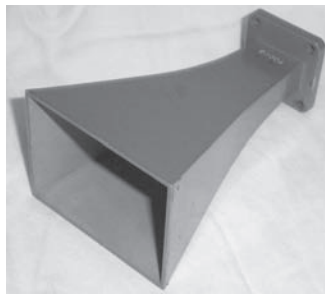
Figure E-15 Waveguide horn antenna.

is inserted. The slot is in a region where the currents in the waveguide wall are minimum, so little discontinuity is introduced. The probe introduces a reactive discontinuity, and the reactance can be varied by changing the depth of penetration of the probe using the knob shown. The probe can move up-and-down along the slot to further increase the impedance range that can be presented.

Several **discontinuities** in rectangular waveguide are shown in Figure E-18. These illustrate most clearly the use of E and H field disturbances to realize capacitive and inductive components. An E-plane discontinuity (Figure E-18(a)) is modeled approximately by a frequency-dependent capacitor. H-plane discontinuities (Figures E-18(b) and E-18(c)) resemble inductors, as does the circular iris of Figure E-18(d). The resonant waveguide iris of Figure E-18(e) disturbs the E and H fields and can be modeled by a parallel LC resonant circuit near the frequency of resonance. Posts in the waveguide are used as reactive elements (Figure E-18(f)) and to mount active devices (Figure E-18(g)). The equivalent circuits of
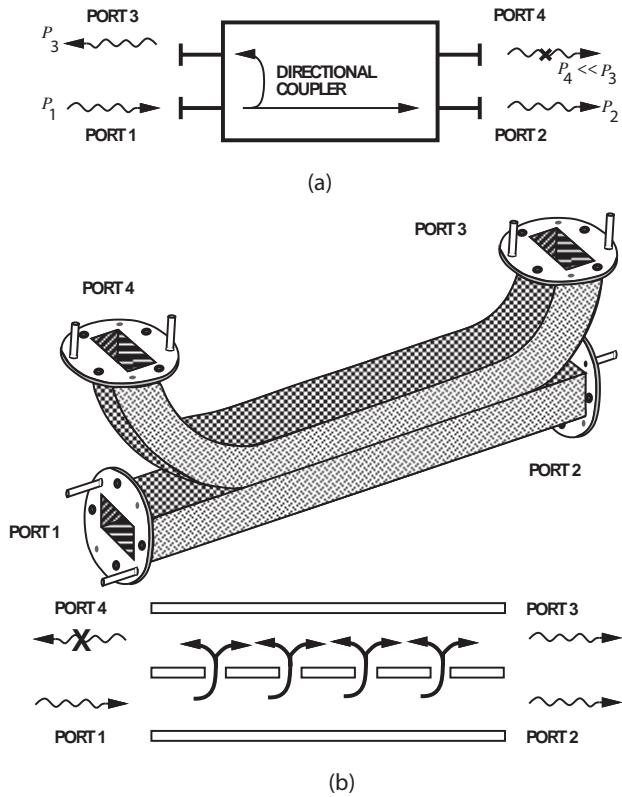
Figure E-16 Directional coupler in rectangular waveguide: (a) schematic of a directional coupler; (b) waveguide directional coupler showing coupling slots; and (c) and (d) photograph of three directional couplers with the fourth port terminated in an integral matched load. In (c) and (d), from top-to-bottom, W-band and 10 cm long, Ka-band and 15 cm long, and X-band 35 cm long.

waveguide discontinuities are modeled by capacitive elements if the E field is interrupted and by inductive elements if the H field (or current) is disturbed.

A one-quarter wavelength long impedance transformer is also available

(a)

(b)

(c)
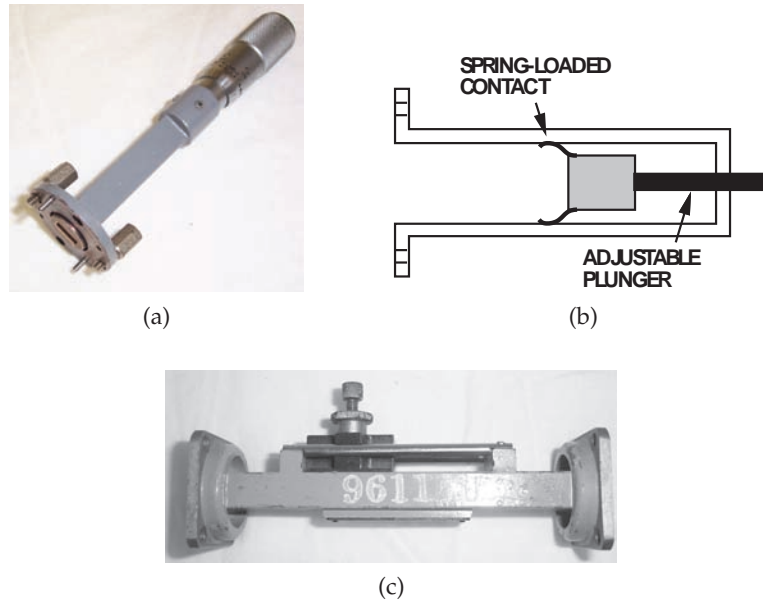
Figure E-17 Waveguide tuners (a) micrometer-driven variable short-circuit; internal details of a variable short-circuit; and (c) waveguide slide tuner.
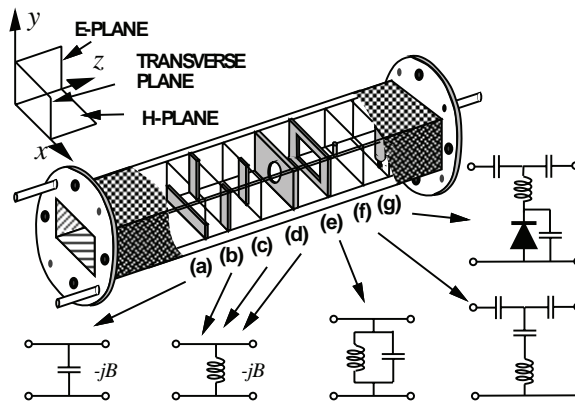
Figure E-18 Rectangular waveguide discontinuities and their lumped equivalent circuits: (a) capacitive E-plane discontinuity; (b) inductive H-plane discontinuity; (c) symmetrical inductive H-plane discontinuity; (d) inductive post discontinuity; (e) resonant window discontinuity; (f) capacitive post discontinuity; and (g) diode post mount.
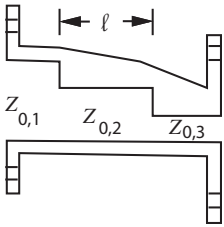
Figure E-19   Rectangular waveguide impedance transformer.
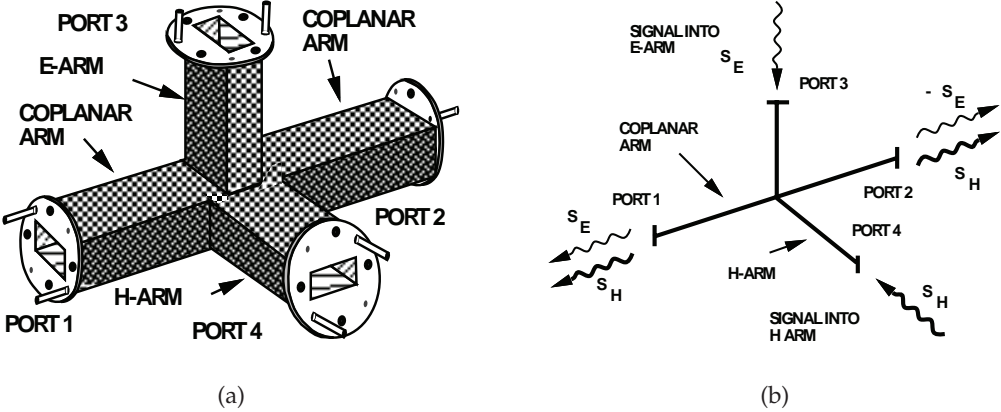
Figure E-20   Rectangular waveguide hybrid: (a) three-dimensional representation; and (b) description of signal flow.

in microstrip, as shown in Figure E-19. This section can be shorter than the tapered waveguide section which, however, had higher bandwidth. Hybrids in waveguides (Figure E-20) do not look anything like their microstrip equivalents.

RF and Microwave Circuit Schematic Symbols

This appendix lists the symbols commonly used with RF and microwave circuits. Symbols are from IEEE Standard 315-1975, "Graphic Symbols for Electrical and Electronics Diagrams (Including Reference Designation Letters)," IEEE, USA, Adopted Sept. 1975, Reaffirmed Dec. 1993 [117].

F.1   Element and Circuit Symbols

Table F-1: IEEE standard qualifying properties added to schematic symbols to identify a particular property.
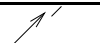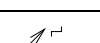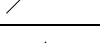
Qualifying Property	Symbol
Adjustable	
Adjustable, continuously adjustable	
Adjustable, stepped	
Linear	
Nonlinear	
Positive	+
Negative	_

Table F-2: IEEE standard schematic symbols of RF and microwave components.
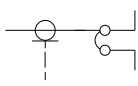
Component	Symbol	Alternate
Attenuator, fixed		
Attenuator, balanced		
Attenuator, unbalanced		
Attenuator, variable		
Attenuator, variable, continuously		
Attenuator, variable, stepped		
Amplifier		
Antenna, general		
Antenna, balanced		
Antenna, dipole		
Antenna, loop		
Balun		
Balun with coaxial line and dipole antenna		
Capacitor, general		
Capacitor, polarized		

Table F-2: (continued)

Component	Symbol	Alternate
Capacitor, variable		
Capacitor, nonlinear		
Capacitor, shielded		
Circulator		
Coaxial cable		
Conductive path		
Connector, female		
Connector, male		
Contact, fixed		
Contact, closed		
Contact, open		
Delay		
Element, linear (* to be replaced by designation)		
Ground, general		
Ground, chassis		
Coupler		
Filter, bandpass filter (BPF)		
Filter, lowpass filter (LPF)		
Filter, highpass filter (HPF)		

Table F-2: (continued)
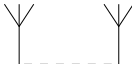
Component	Symbol	Alternate
Filter, bandstop filter (BSF)		
Isolator		
Inductor, general		
Inductor with magnetic core		
Junction	•	
Junction of paths		
Network, linear (* to be replaced by designation)	*	
Open		
Phase shifter	ϕ	β
Piezoelectric resonator		
Radio link	- - - -	
Radio link with antennas		
Rectifier		
Resistor, general		
Resistor, variable		
Resistor, nonlinear		
Shield		- - - -
Short, movable		
Source, ac		
Source, dc		

Table F-2: (continued)

Component	Symbol	Alternate
Switch, multiposition		
Test, point	●	——●
Transformer		
Transformer with magnetic core		
Transformer, center tapped		
Triax		
Twinax		
Twinax with shield showing connection		
Twinax with shield grounded		
Resistor, series	—⋀⋀⋀—●	
Resistor with open path	—⋀⋀⋀—●	
Resistor with short	—⋀⋀⋀—	
Short		
Wire	——	
Wires, connected, crossing		
Wires, unconnected, crossing		
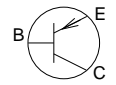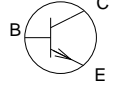
F.2 Diodes

Table F-3 IEEE standard symbols for diodes and a rectifier [117]. †Symbol is not an IEEE standard symbol but is commonly used.

Component	Symbol
Diode, general	
Gunn diode	
PIN diode	
Light emitting diode (LED)	
Rectifier	
Schottky diode†	
Varactor diode	
Zener diode	

F.3 Bipolar Junction Transistor — BJT

Table F-4 IEEE standard schematic symbols for bipolar junction transistors (BJT and HBT) [117] and commonly used symbols in layouts [174]. The letters indicate terminals: B (base), C (collector), E (emitter). These symbols are used for silicon BJTs and compound semiconductor HBTs.

Transistor	IEEE Symbol	Commonly Used Symbol
BJT, pnp		
BJT, npn		

F.4   Junction Field Effect Transistor — JFET

Table F-5   IEEE standard schematic symbols for junction field effect transistors (MESFET, HEMT, JFET) [117] and symbols more commonly used in schematics. The letters indicate terminals: G (gate), D (drain), S (source).

Transistor	IEEE Symbol	Commonly Used Symbol
FET, pJFET		
FET, nJFET, MESFET, HEMT		

F.5   Metal-Oxide-Semiconductor Field Effect Transistor — MOSFET

Table F-6   IEEE standard schematic symbols for MOSFET transistors [117] and symbols more commonly used in schematics [174]. The MOSFET symbols are for enhancement- and depletion- mode transistors.The letters indicate terminals: G (gate), D (drain), S (source), U (bulk). Four-terminal and three-terminal common symbols are shown. The three-terminal common symbol is most often used when the bulk is connected to the most negative connection in the circuit, and the pMOSFET symbol is used when the bulk is tied to V_{DD} (the most positive connection). The bulk connection is often not shown, as it is assumed to be connected to the most negative voltage point.

Transistor	IEEE Symbol	Commonly Used Symbol (3 terminal)	Commonly Used Symbol (4 terminal)
FET, nMOS, depletion			
FET, pMOS, depletion			
FET, nMOS, enhancement			
FET, pMOS, enhancement			

APPENDIX G

Active Device Models

This appendix presents the model parameters of the three most commonly used transistors in RF and microwave design. These models are available in nearly all circuit simulators. Transistor models implement device equations that have been developed from physical insight with necessary simplifications required for implementation in a simulator. The purpose of presenting these models is so that the basic physical description of operation can be examined.

G.1   Level 3 MOSFET Model

The Level 3 MOSFET model is the model of a silicon MOSFET transistor and is one of a large number of different MOSFET models that are used [314, 315, 317, 318]. MOSFETs are the most complicated transistor to model, as their operation relies on attracting carriers into the channel under the gate in a process called inversion. The MOS level 3 model here uses the charge-conserving Yang-Chatterjee model [316] for modeling charge and capacitance. For many years the Level 3 MOSFET model was implemented in circuit simulators but did not conserve charge. An example of errors that can exist in device models.

The model parameters listed in Table G-1 can be specified by the circuit designer.
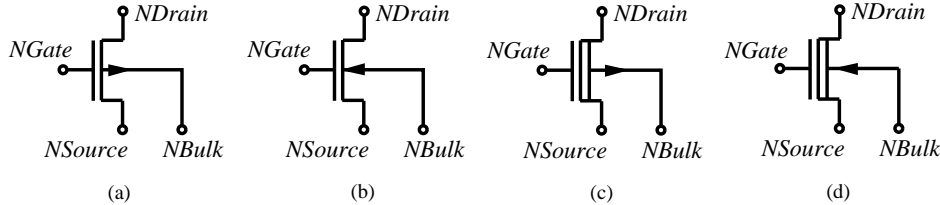
Figure G-1   MOSFET types: (a) enhancement-mode p type; (b) enhancement-mode p type; (a) depletion-mode n type; (b) depletion-mode n type;

Table G-1: Level 3 MOSFET Model Parameters.

Name	Description	Units	Default
gamma	Bulk threshold parameter	$V^{0.5}$	0
kp	Transconductance parameter	A/V^2	0.000021
l	Device length	m	0.000002
w	Device width	m	0.00005
ld	Lateral diffusion length	m	0
wd	Lateral diffusion width	m	0
nsub	Substrate doping	cm^{-3}	0
phi	Surface inversion potential	V	0.6
tox	Oxide thickness	m	1×10^{-7}
u0	Surface mobility	$cm^2/V\text{-}s$	600
vt0	Zero bias threshold voltage	V	0
kappa	Saturation field factor	m	0.2
t	Device temperature	degrees	300.15
tnom	Nominal temperature	degrees	300.15
nfs	Fast surface state density	cm^{-2}	0
eta	Static feedback on threshold voltage		0
theta	Mobility modulation	1/V	0
tpg	Gate material type		0
nss	Surface state density	cm^{-2}	0
vmax	Maximum carrier drift velocity	m/sec	0
xj	Metallurgical junction depth		0
delta	Width effect on threshold voltage		0

G.1.1 Device Equations

The device equations here are specifically for a p-type MOSFET. There are sign changes required to get the appropriate current directions for an n-type MOSFET. The subscript D refers to the drain, S refers to the source, and G refers to the gate. The constants used are

$$q = 1.6021918 \times 10^{-19} \text{ (As)} \tag{G.1}$$

$$k = 1.3806226 \times 10^{-23} \text{ (J/K)} \tag{G.2}$$

$$\epsilon_0 = 8.85421487 \times 10^{-12} \text{ (F/m)} \tag{G.3}$$

$$\epsilon_s = 11.7 \, \epsilon_0 \tag{G.4}$$

$$E_g = 1.16 - \frac{7.02 \times 10^{-4} \, \text{T}^2}{\text{T} + 1108} \text{ (V)} \tag{G.5}$$

$$C_{ox} = \frac{\epsilon_0 \, 3.9}{\text{TOX}} \text{ (F)}. \tag{G.6}$$

All parameters used are indicated in `this` font.

$$L_{\text{eff}} = \text{L} - 2\,\text{LD} \tag{G.7}$$

$$W_{\text{eff}} = \text{W} - 2\,\text{WD} \tag{G.8}$$

Depletion layer width coefficient:

$$X_d = \sqrt{\frac{2\,\epsilon_s}{q\,\text{NSUB}\,10^6}} \tag{G.9}$$

Built in voltage:

$$V_{bi} = \text{VT0} - \text{GAMMA}\,\sqrt{\text{PHI}} \tag{G.10}$$

Square root of substrate voltage:

$$V_{BS} \leq 0 \implies SqV_{BS} = \sqrt{\text{PHI} - V_{BS}}$$

$$V_{BS} > 0 \implies SqV_{BS} = \frac{\sqrt{\text{PHI}}}{1 + \frac{0.5}{\text{PHI}}\,V_{BS}(1 + \frac{0.75}{\text{PHI}}V_{BS})} \tag{G.11}$$

Short-channel effect correction factor:
In a short-channel device, the device threshold voltage tends to be lower since part of the depletion charge in the bulk terminates the electric fields at the source and drain. The value of this correction factor is determined by the metallurgical depth, `XJ`.

$$c_0 = 0.0631353 \tag{G.12}$$

$$c_1 = 0.8013292 \tag{G.13}$$

$$c_2 = -0.01110777 \tag{G.14}$$

$$T_1 = \text{XJ}\,(c_0 + c_1\,X_d\,SqV_{BS} + c_2\,(X_d\,SqV_{BS})^2) \tag{G.15}$$

$$F_s = 1 - \frac{\text{LD} + T_1}{L_{\text{eff}}}\sqrt{1 - (\frac{X_d\,SqV_{BS}}{\text{XJ} + X_d\,SqV_{BS}})^2} \tag{G.16}$$

Narrow channel effect correlation factor:
The edge effects in a narrow channel cause the depletion charge to extend beyond the width of the channel. This has the effect of increasing the threshold voltage:

$$F_n = \frac{\pi\,\epsilon_s\,\text{DELTA}}{2\,C_{ox}\,W_{\text{eff}}}. \tag{G.17}$$

Static feedback coefficient:
The threshold voltage lowers because the charge under the gate terminal depleted by the drain junction field increases with V_{DS}. This effect is drain induced barrier lowering (DIBL).

$$\sigma = \frac{8.14 \times 10^{-22}\,\text{ETA}}{L_{\text{eff}}{}^3\,C_{ox}}. \tag{G.18}$$

Threshold voltage:

$$V_{th} = V_{bi} - \sigma\,V_{DS} + \text{GAMMA}\,SqV_{BS}\,F_s + F_n\,SqV_{BS}{}^2 \tag{G.19}$$

Subthreshold operation:
This variable is invoked depending on the value of the parameter NFS and is used only during subthreshold mode:

$$X_n = 1 + \frac{q\,\text{NFS}\,10^4}{C_{ox}} + \frac{F_n}{2} + \frac{\text{GAMMA}}{2}\,\frac{F_s}{SqV_{BS}}. \tag{G.20}$$

Modified threshold voltage:
This variable defines the limit between weak and strong inversion:

$$\text{NFS} > 0 \quad \Longrightarrow \quad V_{on} = V_{th} + \frac{k\,T}{q}\,X_n \tag{G.21}$$

$$\text{NFS} \le 0 \quad \Longrightarrow \quad V_{on} = V_{th}. \tag{G.22}$$

Subthreshold gate voltage:

$$V_{gsx} = \text{MAX}(V_{GS}, V_{on}). \tag{G.23}$$

Surface mobility:

$$\mu_s = \frac{\text{U0}\,10^{-4}}{1 + \text{THETA}\,(V_{gsx} - V_{th})}. \tag{G.24}$$

Saturation voltage:
Calculation of this voltage requires many steps. The effective mobility is calculated as

$$\mu_{\text{eff}} = \mu_s\,F_{\text{drain}}, \tag{G.25}$$

where

$$F_{\text{drain}} = \frac{1.0}{1 + \frac{\mu_s\,V_{DS}}{\text{VMAX}\,L_{\text{eff}}}} \tag{G.26}$$

$$\beta = \frac{W_{\text{eff}}}{L_{\text{eff}}}\,\mu_{\text{eff}}\,C_{ox}. \tag{G.27}$$

The Taylor expansion of bulk charge is

$$F_B = \frac{\text{GAMMA}}{4}\,\frac{F_s}{SqV_{BS}} + 2\,F_n. \tag{G.28}$$

The standard value of saturation voltage is calculated as

$$V_{\text{sat}} = \frac{V_{gsx} - V_{th}}{1 + F_B}. \tag{G.29}$$

The final value of the saturation voltage depends on the parameter VMAX:

$$\text{VMAX} = 0 \quad \Longrightarrow \quad V_{d\text{sat}} = V_{\text{sat}} \tag{G.30}$$

$$\text{VMAX} > 0 \quad \Longrightarrow \quad V_c = \frac{\text{VMAX}\,L_{\text{eff}}}{\mu_s} \tag{G.31}$$

$$V_{d\text{sat}} = V_{\text{sat}} + V_c - \sqrt{V_{\text{sat}}^2 + V_c^2}. \tag{G.32}$$

Velocity saturation drain voltage:
This ensures that the drain voltage does not exceed the saturation voltage:

$$V_{dsx} = \text{MIN}(V_{DS}, V_{d\text{sat}}).$$

(G.33)

Drain current:
Linear region:

$$I_{DS} = \beta \, \frac{\mu_s}{\text{U0} \, 10^{-4}} \, F_{\text{drain}} \left(V_{gsx} - V_{th} - \frac{1 + F_B}{2} \, V_{dsx} \right) V_{dsx}.$$

(G.34)

Saturation region:

$$I_{DS} = \beta \left[(V_{GS} - V_{th}) - \frac{1 + F_B}{2} \, V_{d\text{sat}} \right] V_{d\text{sat}}.$$

(G.35)

Using Equation (G.27), this becomes

$$I_{DS} = \frac{W_{\text{eff}}}{L_{\text{eff}}} \, \mu_{\text{eff}} \, C_{ox} \left[(V_{GS} - V_{th}) - \frac{1 + F_B}{2} \, V_{d\text{sat}} \right] V_{d\text{sat}}.$$

(G.36)

Cutoff region:

$$I_{DS} = 0.$$

(G.37)

Channel length modulation:
As V_{DS} increases beyond $V_{d\text{sat}}$, the point where the carrier velocity begins to saturate moving toward the source. This is modeled by the term $\triangle_l$:

$$\triangle_l = X_d \sqrt{\frac{X_d^2 \, E_p^2}{4} + \text{KAPPA} \, (V_{DS} - V_{d\text{sat}})} - \frac{E_p \, X_d^2}{2},$$

(G.38)

where E_p is the lateral field at pinch-off and is given by

$$E_p = \frac{\text{VMAX}}{\mu_s \, (1 - F_{\text{drain}})}$$

(G.39)

The drain current is multiplied by a correction factor, l_{fact}. This factor prevents the denominator $L_{\text{eff}} - \triangle_l$ from going to zero.

$$\triangle_l \leq 0.5 \, L_{\text{eff}} \quad \Longrightarrow \quad l_{\text{fact}} = \frac{L_{\text{eff}}}{L_{\text{eff}} - \triangle_l}$$

(G.40)

$$\triangle_l > 0.5 \, L_{\text{eff}} \quad \Longrightarrow \quad l_{\text{fact}} = \frac{4 \, \triangle_l}{L_{\text{eff}}}.$$

(G.41)

The corrected value of drain-source current is

$$I_{DS\text{new}} = I_{DS} \, l_{\text{fact}}.$$

(G.42)

Subthreshold operation:
For subthreshold operation, if the fast surface density parameter NFS is specified and $V_{GS} \leq V_{\text{on}}$, then the final value of drain-source current is given by

$$I_{DS\text{final}} = I_{DSnew} \, e^{\frac{kt}{q} \frac{V_{GS} - V_{\text{on}}}{X_n}}.$$

(G.43)

Yang-Chatterjee Charge Model [316]
This model ensures continuity of the charges and capacitances throughout different regions of operation. The intermediate quantities are

$$V_{FB} = V_{to} - \mathtt{GAMMA}\sqrt{\mathtt{PHI}} - \mathtt{PHI} \tag{G.44}$$

and

$$C_o = C_{ox}\,W_{\text{eff}}\,L_{\text{eff}}. \tag{G.45}$$

Accumulation region, $V_{GS} \leq V_{FB} + V_{BS}$:

$$Q_d = 0 \tag{G.46}$$
$$Q_s = 0 \tag{G.47}$$
$$Q_b = -C_o\,(V_{GS} - V_{FB} - V_{BS}). \tag{G.48}$$

Cut-off region, $V_{FB} + V_{BS} < V_{GS} \leq V_{th}$:

$$Q_d = 0 \tag{G.49}$$
$$Q_s = 0 \tag{G.50}$$
$$Q_b = -C_o\,\frac{\mathtt{GAMMA}^2}{2}\left\{-1 + \sqrt{1 + \frac{4\,(V_{GS} - V_{FB} - V_{BS})}{\mathtt{GAMMA}^2}}\right\}. \tag{G.51}$$

Saturation region, $V_{th} < V_{GS} \leq V_{DS} + V_{th}$:

$$Q_d = 0 \tag{G.52}$$
$$Q_s = -\frac{2}{3}\,C_o\,(V_{GS} - V_{th}) \tag{G.53}$$
$$Q_b = C_o\,(V_{FB}\,\mathtt{PHI} - V_{th}) \tag{G.54}$$

Linear region, $V_{GS} > V_{DS} + V_{th}$:

$$Q_d = -C_o\left[\frac{V_{DS}^2}{8\,(V_{GS} - V_{th} - \frac{V_{DS}}{2})} + \frac{V_{GS} - V_{th}}{2} - \frac{3}{4}\,V_{DS}\right] \tag{G.55}$$

$$Q_s = -C_o\left[\frac{V_{DS}^2}{24\,(V_{GS} - V_{th} - \frac{V_{DS}}{2})} + \frac{V_{GS} - V_{th}}{2} + \frac{1}{4}\,V_{DS}\right] \tag{G.56}$$

$$Q_b = C_o\,(V_{FB}\,\mathtt{PHI} - V_{th}). \tag{G.57}$$

The final currents at the transistor nodes are given by

$$I_d = I_{DS\text{final}} + \frac{dQ_d}{dt} \tag{G.58}$$

$$I_g = \frac{dQ_g}{dt} \tag{G.59}$$

$$I_s = -I_{DS\text{final}} + \frac{dQ_s}{dt}. \tag{G.60}$$
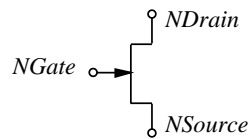
Figure G-2 MESFET element.

G.2 Materka-Kacprzak MESFET Model

The Materka-Kacprzak transistor model was developed for GaAs MESFET transistors [176] but is used to model silicon JFETs and compound semiconductor HEMT transistors as well. It is based on physical interpretation of a transistor with a junction-based gate. There are a number of other models [319–321], but the Materka-Kacprzak model is representative of JFETs.

Table G-2: Materka-Kacprzak Model Parameters

Name	Description	Units	Default
Name	Description	Units	Default
AFAB	Slope factor of breakdown current ($AFAB$)	$1/V$	0.0
AFAG	Slope factor of gate conduction current ($AFAG$)	$1/V$	38.696
AREA	Area multiplier ($AREA$)	-	1.0
C10	Gate source Schottky barrier capacitance for (C_{10})	F	0.0
CFO	Gate drain feedback capacitance for (C_{F0})	F	0.0
CLS	Constant parasitic component of gate-source capacitance (C_{LS})	F	0.0
E	Constant part of power law parameter (E)	-	2.0
GAMA	Voltage slope parameter of pinch-off voltage (γ)	$1/V$	0.0
IDSS	Drain saturation current for (I_{DSS})	A	0.1
IG0	Saturation current of gate-source Schottky barrier (I_{G0})	A	0.0
K1	Slope parameter of gate-source capacitance (K_1)	$1/V$	1.25
KE	Dependence of power law on V_{GS} , (K_E)	$1/V$	0.0
KF	Slope parameter of gate-drain feedback capacitance (K_F)	$1/V$	1.25
KG	Drain dependence on V_{GS} in the linear region, (K_G)	$1/V$	0.0
KR	Slope factor of intrinsic channel resistance (K_R)	$1/V$	0.0
RI	Intrinsic channel resistance for (R_I)	Ω	0.0
SL	Slope of the drain characteristic in the saturated region, (S_L)	S	0.15
SS	Slope of the drain characteristic in the saturated region (S_S)	S	0.0
T	Channel transit-time delay (τ)	s	0.0
VBC	Breakdown voltage (V_{BC})	V	10^{10}
VP0	Pinch-off voltage for (V_{P0})	V	-2.0

The physical constants used in the model evaluation are

k	the Boltzmann constant	$1.3806226\ 10^{-23}$ J/K
q	electronic charge	$1.6021918\ 10^{-19}$ C

Standard Calculations

$$V_{TH} = (kT)/q \tag{G.61}$$

where T is the analysis temperature. Also

$$
\begin{array}{ll}
V_{DS} & \text{Intrinsic drain source voltage} \\
V_{GS} & \text{Intrinsic gate source voltage} \\
V_{GD} & \text{Intrinsic gate drain voltage}
\end{array}
$$

G.2.1 Device Equations:

Current Characteristics

$$
\begin{aligned}
I_{DS} &= \text{Area}I_{DSS}\left[1 + S_S\frac{V_{DS}}{I_{DSS}}\right]\left[1 - \frac{V_{GS}(t-\tau)}{V_{P0} + \gamma V_{DS}}\right]^{(E + K_E V_{GS}(t-\tau))} \\
&\qquad \times \tanh\left[\frac{S_L V_{DS}}{I_{DSS}(1 - K_G V_{GS}(t-\tau))}\right] \tag{G.62}
\end{aligned}
$$

$$I_{GS} = \text{Area}I_{G0}\left[e^{A_{FAG}V_{GS}} - 1\right] - I_{B0}\left[e^{-A_{FAB}(V_{GS} + V_{BC})}\right] \tag{G.63}$$

$$I_{GD} = \text{Area}I_{G0}\left[e^{A_{FAG}V_{GD}} - 1\right] - I_{B0}\left[e^{-A_{FAB}(V_{GD} + V_{BC})}\right] \tag{G.64}$$

$$R_I = \begin{cases} R_{10}(1 - K_R V_{GS})/\text{Area} & K_R V_{GS} < 1.0 \\ 0 & K_R V_{GS} \geq 1.0\,. \end{cases} \tag{G.65}$$

Capacitance

$C_{LVL} = 1$ (default) for the standard Materka-Kacprzak capacitance model described below is used. The Materka-Kacprzak capacitances are

$$C'_{DS} = C_{DS} \tag{G.66}$$

$$C'_{GS} = \begin{cases} \left[C_{10}(1 - K_1 V_{GS})^{M_{GS}} + C_{1S}\right] & K_1 V_{GS} < F_{CC} \\ \left[C_{10}(1 - F_{CC})^{M_{GS}} + C_{1S}\right] & K_1 V_{GS} \geq F_{CC} \end{cases} \tag{G.67}$$

$$C'_{GD} = \begin{cases} \text{Area}\left[C_{F0}(1 - K_1 V_1)^{M_{GD}}\right] & K_1 V_1 < F_{CC} \\ \text{Area}\left[C_{F0}(1 - F_{CC})^{M_{GD}}\right] & K_1 V_1 \geq F_{CC} \end{cases}. \tag{G.68}$$

G.3   Gummel-Poon: Bipolar Junction Transistor Model
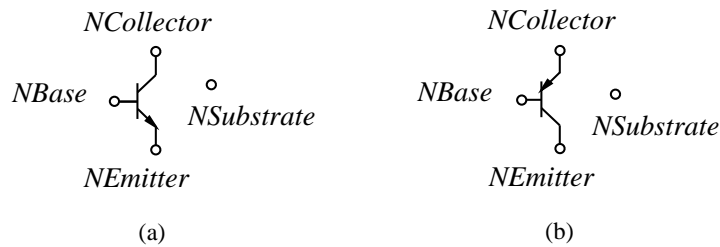
Figure G-3   Q — bipolar junction transistor: (a) npn transistor; (b) pnp transistor.

Bipolar transistor models are based on the Gummel-Poon model [171] described here. The key feature of the model is that it captures the dependence of the forward and reverse current gain on current. In essence, the BJT model is a current-controlled current source. The Gummel-Poon model and its derivatives are used to model silicon BJTs and compound semiconductor HBTs [322, 323].

Table G-3: Gummel-Poon BJT Model Parameters

Name	Description	Units	Default
AREA	Current multiplier		1.0
BF	Ideal maximum forward beta (B_F)		100.0
BR	Ideal maximum reverse beta (B_R)		1.0
C2	Base-emitter leakage saturation coefficient		I_{SE}/I_S
C4	Base-collector leakage saturation coefficient		(I_{SC}/I_S)
CJC	Base collector zero bias p-n capacitance (C_{JC})	F	0.0
CJE	Base emitter zero bias p-n capacitance (C_{JE})	F	0.0
EG	Bandgap voltage (E_G)	eV	1.11
FC	Forward bias depletion capacitor coefficient (F_C)		0.5
IKF	Corner of forward beta high-current roll-off (I_{KF})	A	10^{-10}
IKR	Corner for reverse-beta high current roll off (I_{KR})		10^{-10}
IS	Transport saturation current (I_S)	A	10^{-16}
ISC	Base collector leakage saturation current (I_{SC})	A	0.0
ISE	Base-emitter leakage saturation current (I_{SE})	A	0.0
IRB	Current at which RB falls to half of R_{BM} (I_{RB})	A	10^{-10}
ITF	Transit time dependency on IC (I_{TF})	A	0.0
MJC	Base collector p-n grading factor (M_{JC})		0.33
MJE	Base emitter p-n grading factor (M_{JE})		0.33
NC	Base-collector leakage emission coefficient (N_C)		2.0
NE	Base-emitter leakage emission coefficient (N_E)		1.5
NF	Forward current emission coefficient (N_F)		1.0

Name	Description	Units	Default
NR	Reverse current emission coefficient (N_R)		1.0
RB	Zero bias base resistance (R_B)	Ω	0.0
RBM	Minimum base resistance (R_{BM})	Ω	R_B
RE	Emitter ohmic resistance (R_E)	Ω	0.0
RC	Collector ohmic resistance (R_C)	Ω	0.0
T	Operating Temperature T	K	300
TF	Ideal forward transit time (T_S)	secs	0.0
TNOM	Nominal temperature (T_{NOM})	K	300
TR	Ideal reverse transit time (T_R)	S	0.0
TRB1	RB temperature coefficient (linear) (T_{RB1})		0.0
TRB2	RB temperature coefficient (quadratic) (T_{RB2})		0.0
TRC1	RC temperature coefficient (linear) (T_{RC1})		0.0
TRC2	RC temperature coefficient (linear) (T_{RC2})		0.0
TRE1	RE temperature coefficient (linear) (T_{RE1})		0.0
TRE2	RE temperature coefficient (quadratic) (T_{RE2})		0.0
TRM1	RBM temperature coefficient (linear) (T_{RM1})		0.0
TRM2	RBM temperature coefficient (quadratic) (T_{RM2})		0.0
VA	alternative keyword for VAF (V_A)	V	10^{-10}
VAF	Forward early voltage (V_{AF})	V	10^{-10}
VAR	Reverse early voltage (V_{AR})		10^{-10}
VB	alternative keyword for VAR (V_B)		10^{-10}
VJC	Base collector built in potential (V_{JC})	V	0.75
VJE	Base emitter built in potential (V_{JE})	V	0.75
VTF	Transit time dependency on VBC (V_{TF})	V	10^{-10}
XCJC	Fraction of CBC connected internal to RB (X_{CJC})		1.0
XTB	Forward and reverse beta temperature coefficient (X_{TB})		0.0
XTF	Transit time bias dependence coefficient (X_{TF})		0.0
XTI	IS temperature effect exponent (X_{TI})		3.0
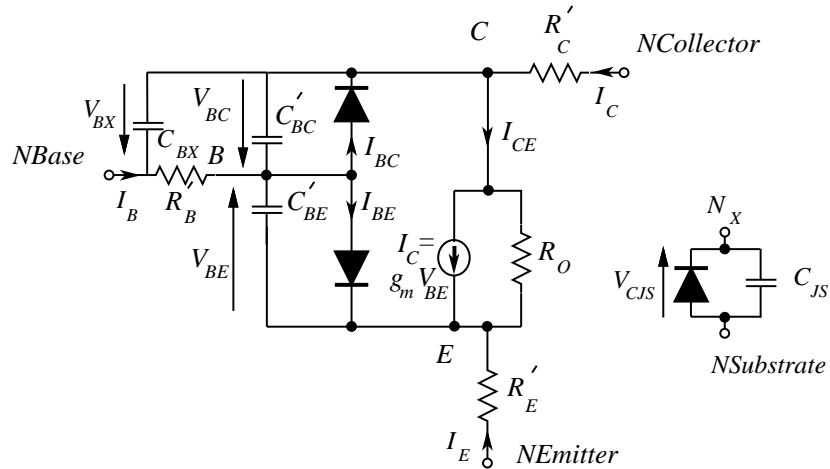
Schematic of the BJT Model

G.3.1   Standard Calculations

The physical constants used in the model evaluation are

k	the Boltzmann constant	$1.3806226\ 10^{-23}$ J/K
q	electronic charge	$1.6021918\ 10^{-19}$ C

Absolute temperatures (in kelvin, K) are used. The thermal voltage is

$$V_{TH}(T_{NOM}) = k\,T_{NOM}/q. \tag{G.69}$$

Current Characteristics
The base-emitter current is

$$I_{BE} = I_{BF}/\beta_F + I_{LE} \tag{G.70}$$

the base-collector current is

$$I_{BC} = I_{BR}/\beta_R + I_{LC}, \tag{G.71}$$

and the collector-emitter current is

$$I_{CE} = I_{BF} - I_{BR}/K_{QB}, \tag{G.72}$$

where the forward diffusion current is

$$I_{BF} = I_S\left(e^{V_{BE}/(N_F V_{TH})} - 1\right), \tag{G.73}$$

the nonideal base-emitter current is

$$I_{LE} = I_{SE}\left(e^{V_{BE}/(N_E V_{TH})} - 1\right), \tag{G.74}$$

the reverse diffusion current,

$$I_{BR} = I_S \left(e^{V_{BC}/(N_R V_{TH})} - 1 \right) \tag{G.75}$$

the non-ideal base-collector current is

$$I_{LC} = I_{SC} \left(e^{V_{BC}/(N_C V_{TH})} - 1 \right), \tag{G.76}$$

and the base charge factor is

$$K_{QB} = 1/2 \left[1 - V_{BC}/V_{AF} - V_{BE}/V_{AB} \right]^{-1} \left(1 + \sqrt{1 + 4 \left(I_{BF}/I_{KF} + I_{BR}/I_{KR} \right)} \right). \tag{G.77}$$

Thus the conductive current flowing into the base is

$$I_B = I_{BE} + I_{BC} \tag{G.78}$$

the conductive current flowing into the collector is

$$I_C = I_{CE} - I_{BC} \tag{G.79}$$

and the conductive current flowing into the emitter is

$$I_E = I_{BE} + I_{CE}. \tag{G.80}$$

Capacitances

$C_{BE} = \text{Area}(C_{BE\tau} + C_{BEJ})$, where the base-emitter transit time or diffusion capacitance is

$$C_{BE\tau} = \tau_{F,EFF} \partial I_{BF}/\partial V_{BE} \tag{G.81}$$

and the effective base transit time is empirically modified to account for base punchout, space-charge limited current flow, quasi-saturation, and lateral spreading, which tend to increase τ_F

$$\tau_{F,EFF} = \tau_F \left[1 + X_{TF}(3x^2 - 2x^3)e^{(V_{BC}/(1.44 V_{TF})} \right], \tag{G.82}$$

and $x = I_{BF}/(I_{BF} + \text{Area} I_{TF})$.

The base-emitter junction (depletion) capacitance is

$$C_{BEJ} = \begin{cases} C_{JE} \left(1 - V_{BE}/V_{JE} \right)^{-M_{JE}} & V_{BE} \le F_C V_{JE} \\ C_{JE} \left(1 - F_C \right)^{-(1 + M_{JE})} \left(1 - F_C(1 + M_{JE}) + M_{JE} V_{BE}/V_{JE} \right) & V_{BE} > F_C V_{JE}. \end{cases} \tag{G.83}$$

The base-collector capacitance is $C_{BC} = \text{Area}(C_{BC\tau} + X_{CJC} C_{BCJ})$, where the base-collector transit time or diffusion capacitance is

$$C_{BC\tau} = \tau_R \partial I_{BR}/\partial V_{BC}. \tag{G.84}$$

The base-collector junction (depletion) capacitance is

$$
C_{BCJ} = \begin{cases} C_{JC} \left(1 - V_{BC}/V_{JC}\right)^{-M_{JC}} & V_{BC} \leq F_C V_{JC} \\ C_{JC} \left(1 - F_C\right)^{-(1 + M_{JC})} \left(1 - F_C(1 + M_{JC}) + M_{JC} V_{BC}/V_{JC}\right) & V_{BC} > F_C V_{JC}. \end{cases}
$$
(G.85)

The capacitance between the extrinsic base and the intrinsic collector is

$$
C_{BX} = \begin{cases} \mathrm{Area}(1 - X_{CJC}) C_{JC} \left(1 - V_{BX}/V_{JC}\right)^{-M_{JC}} & V_{BX} \leq F_C V_{JC} \\ \\ (1 - X_{CJC}) C_{JC} \left(1 - F_C\right)^{-(1 + M_{JC})} & V_{BX} > F_C V_{JC} \\ \quad \times \ (1 - F_C(1 + M_{JC}) + M_{JC} V_{BX}/V_{JC}.) \end{cases}
$$
(G.86)

The substrate junction capacitance is

$$
C_{JS} = \begin{cases} \mathrm{Area} C_{JS} \left(1 - V_{CJS}/V_{JS}\right)^{-M_{JS}} & V_{CJS} \leq 0 \\ \mathrm{Area} C_{JS} \left(1 + M_{JS} V_{CJS}/V_{JS}\right) & V_{CJS} > 0. \end{cases}
$$
(G.87)

Index

Glossary

3GPP	3rd Generation Partnership Project; coordinating body for the production of globally applicable technical specifications and reports for a 3G system, 138
$3\pi/8$-8PSK	3π on 8, 8-state phase shift keying, 5
γ	propagation constant, the logarithmic rate of change, with respect to distance of a time-varying waveform, 163
Γ^V	voltage reflection coefficient, 195
Γ^I	current reflection coefficient, 195
ε_r	relative permittivity, the permittivity of a material relative to that of free space (ε_0), 167
ε'	real part of the permittivity, 173
ϵ_e	effective relative permittivity, 223
ε	permittivity, 166
ε_0	permittivity of free space, $\varepsilon_0 = 8.854 \times 10^{-12}$ F·m^{-1}, 166
η	characteristic or wave impedance of a propagating wave in a medium, used instead of Z_0, which is reserved for transmission lines, 866
η	efficiency, 607
η_{PAE}	power added efficiency, 607
η_{TOTAL}	total power added efficiency, 607
λ	wavelength, distance between corresponding points of a wave at one instant in time, 164
λ_0	wavelength in free space, 167
μ	permeability, 166
μ_0	permeability of free space, $\mu_0 = 4\pi \times 10^{-7}$ H·m^{-1}, 166
μ_r	relative permeability, the permeability of a material relative to that of free space (μ_0), 167
π	pi, ratio of circumference to diameter of a circle, 841
$\pi/4$-DQPSK	π on 4 differential encoded quadrature phase shift keying, 5
QPSK	quadrature phase shift keying; also known as quarternary PSK, quadraphase PSK, and quadra PSK. This is a modulation scheme that represents a signal using four distinct phase states. It can also be viewed as the superposition of two quadrature carriers and this reflects the most common hardware implementation, 5
σ	conductivity, 173

$\phi(t)$	phase waveform of baseband modulating signal, 45
Ω	ohms, 163
ω	$= 2\pi f$, radian frequency, 182
0G	zero-generation (analog) radio, 112
1G	first-generation radio, 131
2G	second-generation radio, 131
3G	third-generation radio, 138
4G	fourth-generation radio, 141
∇	nabla, a spatial derivative operator, 851
$\nabla\cdot$	div operator, captures the spreading out of a field, 851
$\nabla\times$	curl operator, captures the circulation of a field, 851
ν	form of wavenumber used by particle physicists, 182
3D	three-dimensional, 168
8PSK	8-state phase shift keying, 5
16PSK	16-state phase shift keying, 5
$A(t)$	amplitude waveform of baseband modulating signal, 45
A	ampere, SI unit of current, 839
ACI	adjacent channel interference, 23
ACPR	adjacent channel power ratio, the ratio of the power introduced into the adjacent channel to the power in the main channel, 23
ADC	analog-to-digital converter, 711
AM	amplitude modulation, 5
AMPS	advanced mobile phone system, first generation analog cellular radio, 131
AMTS	advanced mobile telephone system, 131
ANA	automatic network analyzer; alternative acronym for a VNA, vector network analyzer, 346
AP	access point, 145
APDF	amplitude probability density function, 692
ARP	autoradiopuhelin (car radio phone), 131
ASK	amplitude shift keying, 15
AuC	authentication center, 139
Autotel	public automated land mobile radiotelephone service, 131
B	bandwidth, 675
B	bels, logarithm of power ratio, 837
B_b	baseband bandwidth, 119
BER	bit error rate, 110
BFSK	binary frequency shift keying, 5
BISDN	broadband ISDN (integrated services digital network), 134
bit rate	rate of information carrying digital data., 103
B_m	modulation bandwidth, 119
bps	bits per second, 103
BPSK	binary phase shift keying, 5
BS	base station, 111

C	capacitor, describes the capacity of a structure to store electric energy, 161
$\hat{C}$	theoretical maximum channel bit rate, 131
C	coulomb, unit of charge, 839
c	speed of light, 2.99792458 m·s^{-1}, symbol derived from the Latin word *celeritas* meaning "swiftness", 164
C	coupling factor of a directional coupler, the ratio of voltage of input to voltage on coupled output, 484
CapEx	capital expenditure, 148
cd	candela, unit of luminous intensity, 839
CDMA	code division multiple access, 127
CMRR	common-mode rejection ratio, ratio of the differential-mode gain to the common-mode gain of a differential amplifier, 647
CODATA	International Council for Science: Committee on Data for Science and Technology, 839
CSD	circuit switched data, 131
CSMA	carrier-sense multiple access, 129
CW	continuous wave, 110
δ	skin depth, 251
D	directivity of a directional coupler, measure of the undesired coupling, 484
DAC	digital-to-analog converter, 1
DAMPS	digital advanced mobile phone system, 134
DataTac	point-to-point wireless data communication standard, 131
dB	decibels, 837
dBm	power in decibels with respect to 1 milliwatt, 837
DECT	digital European cordless telephone; digital cordless telephone standard incorporating features of cellular technology including picocells and call hand-off from cell-to-cell, 31
de-multiplexer	also demultiplexing—process of separating a multiplexed signal into its individual components, 125
DEMUX	see demultiplexer, 125
DMT	discrete multitone, 143
downlink	communication link from a base station to a mobile unit. Sometimes also called the forward path. The terminology comes from communication from an airborne platform such as an aircraft or satellite to a ground-based radio. The expensive part of the communication link is on the airborne platform and this one unit communicates with many ground-based units. The usage in cellular systems follows from this concept of a central unit communicating with an individual unit, 112
DR	dynamic range, 685
DS	direct sequence; most common spread-spectrum technique (used in CDMA) to spread the signal bandwidth by modulating the signal with a pseudo-noise sequence, 127
DS-CDMA	direct sequence code division multiple access, 127

FHSS	frequency hopping spread spectrum; a spread-spectrum technology in which the transmitted signal hops from one frequency to the next in discrete steps, 128
FM	frequency modulation, 5
FOMA	freedom of mobile multimedia access, 131
forward link	communication link from a base station to a mobile unit. Also called the down link, 112
forward path	alternative designation of a downlink, 112
FQPSK	Feher quadrature phase shift keying, 29
FRS	family radio service; a UHF FM walkie-talkie system used since 1996. Early walkie-talkies operated in the 27 MHz and 49 MHz band and suffered interference from citizen band radio (27 MHz) and remote-controlled devices (49 MHz), 113
FSK	frequency shift keying, 5
full duplex	two-way communication in which both units can transmit and receive simultaneously, 113

g	gram, unit of weight, 839
G_A	antenna gain, 76
GaAs	gallium arsenide, 171
GAN	generic access network, 131
GaN	gallium nitride, 42
Gbps	billion bits per second, 141
GGSN,	gateway GPRS support node, 139
GMSC	gateway MSC, 139
GMSK	minimum shift keying using Gaussian filtered data, 5
G_P	processing gain, 119
GPRS	general packet radio system, 131
GSM	Global System for Mobile Communications; formerly known as Groupe Spécial Mobile, 125

half duplex	two-way communication scheme in which only one unit transmits at a time while the other receives until the role is reversed, 113
Hicap	NTT's mobile radiotelephone service, 131
HLR	home location register, 139
HSCSD	high-speed circuit-switched data, 131
HSDPA	high-speed downlink packet access, 131
HSUPA	high-speed uplink packet access, 131
Hz	hertz, unit of frequency, cycles per second, 106

$i(t)$	in-phase baseband waveform used in quadrature modulation, 43
I	isolation of a directional coupler, degree of isolation, 484
IDC	interdigitated capacitor, 373
IF	intermediate frequency; generally the frequency that results from mixing a radio frequency signal with a local oscillator, 109
IFFT	inverse fast Fourier transform, 143

IIP3	input-referred third-order intercept point, 686
$\Im\{\}$	imaginary part of a complex number, 180
IM3	third-order intermodulation product, 680
IMD	intermodulation distortion, 110
IMP	intermodulation product, 680
IMTS	improved mobile telephone service, 131
ISDN	integrated services digital network, 134
ISM	industrial scientific, and medical bands; North American industrial, scientific, and medical bands originally set aside for heating devices but now used for unlicensed wireless communications, 109
J	joule, unit of energy, 839
K	inverter, 657
$K(S)$	characteristic polynomial, 774
K	characteristic function, 509
K	kelvin, SI unit of temperature, 839
k	symbol used for permittivity but ε preferred in electrical engineering. k is used more commonly in materials engineering, 167
k	wavenumber used in electromagnetics and in dealing with waves, 182
k	wavenumber, $k = 2\pi/\lambda = \omega\sqrt{\mu\varepsilon}$, 65
k_c	cutoff wavenumber, 266
kbps	thousand bits per second, 103
kg	kilogram, SI unit of weight, 839
kHz	kilohertz, 1000 Hz, 106
L	inductor, describes the capacity of a structure to store magnetic energy, 161
link	the RF path between the output of a transmitter and the input of a receiver, 61
LNA	low noise amplifier, 3
LO	local oscillator; generally the name given to a reference oscillator or the oscillator used with a mixer, 1
LTE	long-term evolution, cellular radio beyond 3G, 141
m	meter, SI unit of length, 839
Mbps	million bits per second, 103
MDS	minimum discernible signal, 123
MER	modulation error ratio; the ratio of the average signal power to the average error power, 39
MIM	metal-insulator-metal capacitor, 372
MIMO	multiple input, multiple output; technology using multiple antennas to transmit and receive signals, 146
MIS	metal-insulator-semiconductor; also MOS, 275
MMIC	microwave monolithic integrated circui;, an integrated circuit operating at microwave frequencies generally fabricated on a compound semiconductor substrate such as GaAs, 643

Mobitex   national public access wireless data network, 131
modulation   process of impressing information on a carrier wave, 107
mol   mole, unit of amount of substance, 839
MOS   metal-oxide-semiconductor; also MIS, 275
MPSK   M-ary (M-state) phase shift keying, 18
MSC   mobile services switching center; main switching center supporting multiple base stations, 123
MSK   minimum shift keying a form of FSK, 5
MTS   mobile telephone service, 131
multiplexer also multiplexing; process of combining multiple signals into a single signal that can be transmitted, 125
MUX   see multiplexer, 125
mW   milliwatts, 837

N   newton, unit of force, 839
N_o noise power in the time interval of a bit, 23
NADC   North American Digital Cellular, 131
NAMPS   Narrowband Advanced Mobile Phone Service; narrowband AMPS service with 10 kHz wide frequency modulated channels, 125
NextGen   next generation of cellular radio beyond 3G, 141
NF   noise figure, 675
NMT   Nordic Mobile Telephone; European analog cellular telephone standard derived for the U.S. AMPS standard, 131
N_o noise energy corresponding to a bit, 120
Node B   base station, 139
Np   Neper, unit of attenuation, to convert from dB to Np multiply by 0.115129254, 186
NTACS   narrowband version of TACS, doubles the capacity of TACS by splitting the 25 kHz TACS channel into two 12.5 kHz channels, 125

OFDM   orthogonal frequency division multiplexing, 142
OFDMA   orthogonal frequency division multiple access, 145
OIP3   output-referred third-order intercept point, 686
Ω solid angle, 75
OpEx   operating expenditure, 148
OQPSK   offset quadrature phase shift keying, 5
OTA   operational transconductance amplifier, 588

P_{avg} average power, 10
P_{peak} peak power, 9
p.u.l.   per unit length, 179
PA   power amplifier, 46
Pa   pascal, unit of pressure, 839
P_A power available usually from a source, 205
PAE   power-added efficiency, 607
PALM   Public Automated Land Mobile radiotelephone service, 131
PAR   peak-to-average ratio, 8

PBR   passband ripple, 518
PCL   parallel coupled line; most often refers to the type of element used in
                realizing a microwave network or filter, 747
PCS   personal communications service (or system); broad range of
                communications services that enable communications with persons,
                119
P_D power delivered usually to a load, 205
PDA   pseudo-differential amplifier, 644
PDC   personal digital cellular, 131
phasor   a complex number that combines amplitude and phase of a sinewave,
                164
PHS   personal handyphone system, 131
PLL   phase locked loop, 20
PM   phase modulation, 5
PSTN   public switched telephone network, 123

Q quality factor, 40
q filling factor, 229
$q(t)$ out-of-phase (quadrature) baseband waveform used in quadrature
                modulation, 43
QAM   quadrature amplitude modulation, 5
QoS   quality of service, 145

R resistor, describes the capacity of a structure to convert electric energy
                into other forms whereby the energy is dissipated from the electronic
                circuit or EM field, 161
R_s skin resistance, 252
R_s sheet resistance, 229
rad   radians, 180
R_b actual channel bit rate minus overhead (coding, etc.), 131
R_c actual channel bit rate, 131
$\Re\{\ \}$ real part of a complex number, 180
reverse   alternative designation of an uplink., 112
path
RF   radio frequency, 101
RF link   RF path between the output of a transmitter and the input of a
                receiver, 61
RFIC   radio frequency integrated circuit, 177
RL   reverse link; the communication link from a mobile unit to the base
                station. Also called the uplink., 112
RNC   radio network controller, 139
RSSI   received signal strength indicator, 123
R_{TH} Thevenin equivalent resistance, 205

s   second, unit of time, 839
SBPSK   shaped binary phase shift keying, 29
SCCS   suppressed carrier single sideband, 43

SGSN serving GPRS support, 139
SI abbreviation of the French Systme International d'Units, the most commonly used international system of units and is the modern specification of the metric system, 163
Si silicon, 171
Si_2 silicon dioxide, 171
SIM subscriber identification module; a card the size of a credit card or smaller inserted in a GSM cellular phone to identify an individual subscriber, 119
simplex one-way communication link in which only one unit in a pair transmits while the other receives, 112
SIR signal-to-interference ratio, 118
SNR_{MIN} minimum signal-to-noise ratio required for a communication or radar system to detect a signal, 685
SNR signal-to-noise ratio; ratio of desired signal power to noise power over the same frequency range as the signal, 23
SOQPSK shaped-offset quadrature phase shift keying, 29
SQPSK staggered quadrature phase shift keying, 29
sr steradian, unit of solid angle, 75
SS spread spectrum, 127
steradian unit of solid angle, 75

T transmission coefficient, 509
T transmission factor of a directional coupler, ratio of voltage at direct output to voltage at input, 484
T_0 standard temperature, 675
TACS total access communications system; European analog cellular telephone standard derived from the AMPS standard, 125
$\tan \delta$ dielectric loss tangent, 173
TD-SCDMA time division synchronous CDMA, 131
TDD time division duplex, 112
TDMA time division multiple access, 125
TEM mode whose electric and magnetic fields are in a plane perpendicular to the direction of propagation, 168
TL through-line two-tier calibration technique, 363
TPR transducer power ratio, 509
transceiver RF front architecture combining a transmitter and a receiver, 57
TRF tuned radio frequency, 40
TRL through-reflect-line calibration technique, 360
TRM through-reflect-match calibration technique, 361
TSD through-short-delay calibration technique, 359

UE unit element, alternative name for an impedance or admittance inverter, 532
UE user equipment, 139
UHF ultra high frequency, 151

UL	uplink; the communication link from a mobile unit to the base station. Also called the reverse path or reverse link., 112
UMA	unlicensed mobile access, 131
UMTS	universal mobile telecommunication service of system, 139
uplink	communication link from a mobile unit to the base station. Sometimes also called the reverse path. The terminology comes from communication from the ground to an airborne platform such as an aircraft or satellite. The expensive part of the communication link is in on the airborne platform and this one unit communicates with many ground based units. The usage in cellular systems follows from this concept of an individual unit communicating with a central unit, 112
urban canyon (or waveguide) effect	an environment including city blocks with a virtually unbroken carpet of buildings several stories tall, with the streets forming canyons below the rooftops. Signal propagation into the streets is mainly by diffraction over tops of buildings and reflection from buildings across the streets, with channeling of energy along the canyons formed by streets. Propagation in this environment is captured statistically in the COST231 Walfisch-Ikegami model [26] which is based on models described in references [27–29], 116
USDC	U.S. Digital Cellular, 131
USIM	universal subscriber identity module, 139
UWB	ultra-wideband system, 149
v_p	phase velocity, the velocity at which the a constant phase point on a sinewave appears to move, 165
v_g	group velocity, 182
VCO	voltage-controlled oscillator, 43
VHF	very high frequency, 151
VLR	visitor location register, 139
VNA	vector network analyzer, instrument for measuring scattering parameters, 346
W	watt, unit of power, 839
WARC	World Administrative Radio Conference; conference organized by CCIR to obtain international agreement on spectrum allocation, 116
wave-number	$k = 2\pi/\lambda = \omega\sqrt{\mu\varepsilon}$, 65
WCDMA	wideband CDMA, 127
WiDEN	wideband integrated dispatch enhanced network, 131
WLAN	wireless local area network, 104
WMAN	wireless metropolitan area network, 104

A comprehensive on-line glossary of telecommunications terms is available as the ATIS Telecom Glossary (see http://www.atis.org/glossary).