High Frequency Techniques

An Introduction to RF and Microwave Engineering

JOSEPH F. WHITE

HIGH FREQUENCY TECHNIQUES

HIGH FREQUENCY TECHNIQUES

An Introduction to RF and Microwave Engineering

Joseph F. White
JFW Technology, Inc.

IEEE
IEEE PRESS

WILEY-INTERSCIENCE

A John Wiley & Sons, Inc. publication

to Christopher

◼ CONTENTS

Appendices

This book is written for the undergraduate or graduate student who wishes to pursue a career in radio-frequency (RF) and microwave engineering. Today's engineer must use the computer as a design tool to be competitive. This text presumes that the student has access to a computer and network simulation software, but the book can be used without them. In either event, this text will prepare the student for the modern engineering environment in which the computer is a tool of daily use.

The computer is used in two ways. First, it performs laborious calculations based on a defined procedure and a set of circuit element values. This is the major use of network simulation, and it is employed throughout this book to show how each network that is described performs over a frequency range. The second way is like the first except that the computer varies the element values either to approach a desired performance goal (optimization) or to show the variation in performance when a quantity of circuits is built using parts whose values vary from piece to piece (yield prediction).

In the second use, the computer is like a thousand monkeys who, it was once postulated, if taught to type, would eventually type all of the world's great literature including an index to the work. But, it was further postulated, they also would type every possible wrong version, incorrect indices included. Today, the engineer's task is to obtain the useful outputs of the computer based on a fundamental understanding of the underlying principles. Within this text, the computer is used as a tool of, not a substitute for, design. This book emphasizes fundamental concepts, engineering techniques, and the regular and intelligent use of the computer as a computational aid.

Within this presentation of theoretical material, computer-generated examples provide insight into the basic performance, bandwidth, and manufacturing yield of RF and microwave networks. This facilitates the evaluation of classical circuit designs and their limitations. However, in modern engineering, rarely is a classical circuit design used in its standard form, although that was necessarily the practice before the availability of personal computers and simulation software. Rather, today the classical design is a point of embarkation from which a specific design is tailored to immediate design needs. The presence of the classical design remains important because it serves as a starting point to define what specifications might reasonably be expected as optimizer goals for the simulation. Effectively, it "gets the thousand monkeys started on the right page."

This book contains a review of wireless history and engineering fundamentals including complex numbers, alternating-current theory, and the logarithmic basis of decibels. All of the text is written in a simple and informal manner so that the presentation of concepts is easy to follow. Many derivations show intermediate steps not usually included in textbooks because the intent is to enlighten, not test, more than need be, the mathematical prowess of the reader. This book also contains exercises that do not have a black or white answer. Exercise questions are asked that require consideration beyond what is covered in the text. This is intentional. It is done to introduce the reader to what happens in the practical realm of engineering.

The reader is cautioned not to interpret the review material and easy readability of this text for a lack of conceptual rigor or thoroughness. As the reader will soon determine, the chapters of this book actually are more encompassing of theoretical concepts and advanced engineering techniques than those of most introductory microwave texts. But the emphasis is on practical technique. For example, the reader will be surprised that, based merely on Q and the complex number conversion between impedance and admittance, a technique called Q *matching* is developed that is familiar to few engineering professionals.

The emphasis of this book is how design challenges would be attacked in a real engineering environment. Some designs, such as distributed filters, are best performed either with proprietary software programs or with the thousand-monkey approach (optimization), but the emphasis is in providing the monkeys with a promising start.

The style of this textbook is derived from a hands-on industrial course that the author has been teaching for some time. In it the student builds on the computer the circuits that are presented, designing them to specifications and verifying how they perform with frequency. This approach quickly builds design confidence in the student. The exercises presented draw from this experience, and they employ the network simulator to reveal both circuit performance and the student's mastery of it. The following paragraphs summarize the major subjects covered.

Chapter 1 contains a review of the origins of wireless transmission. The early and persistent efforts of Guglielmo Marconi in developing radio is an inspiration to engineers today.

Chapter 2 is an engineering review of alternating-current analysis using complex notation (in Appendix B), impedances and decibel, dBm, and dBW measures with the aim of solidifying these basic concepts. Intuitive level proficiency in these fundamentals is as important to microwave and RF engineering as touch typewriting is to efficient writing. Practical realizations of circuit elements are described, including resistors, inductors, and capacitors and their equivalent circuits with parasitic elements. The parasitic reactances of these elements seriously limit their use at high frequencies, and the engineer does well to know these limits and how they come about.

Chapter 3 treats resonators and how their bandwidth is influenced by Q. Based upon the Q ratio of reactance to resistance and the conversion between

series and shunt impedances, the scheme called *Q matching* is derived. This enables the engineer to design a *LC* matching network in a few, simply remembered steps.

Chapter 4 introduces distributed circuits based on transmission lines and their properties. This is the beginning of microwave design theory. Important ideas such as wavelength, voltage standing-wave ratio (VSWR), reflections, return loss, mismatch loss, and mismatch error are presented. These are followed by slotted line measurements and the derivation of the telegrapher and transmission line equations. Phase and group velocity concepts and reflection coefficient related to impedance and distributed matching are introduced. The transmission line impedance transformation equation is derived and applied to special cases of easy applicability. Fano's limit is presented. It is an important restriction on the capacity for matching over a frequency band and was derived in terms of reflection coefficient.

Chapter 5 is devoted to the basis and use of the Smith chart, the *sine qua non* for microwave engineers. The Smith chart affords a window into the workings of transmission lines, rendering their very complex impedance transformation behavior clearly understandable with a single diagram. This presentation reveals how the function of the Smith chart in handling impedance transformation arises out of the constant magnitude of the reflection coefficient along a lossless line, that the chart is merely the reflection coefficient plane, a principle often overlooked. Navigating the chart using impedance, admittance, reflection coefficient, and *Q* circles is presented. Matching to complex load impedances, estimating VSWR bandwidth, and developing equivalent circuits are among the illustrated techniques.

Chapter 6 is a presentation of matrix algebra and definitions for the *Z*, *Y*, *ABCD*, *S*, and *T* matrices. Matrix use underlies most circuit derivations and measurement techniques. This chapter demonstrates how and when to use the different matrices and their limitations. For example, it shows how to employ the *ABCD* matrix to derive remarkably general equivalent circuits in just a few steps, such as the lumped equivalent circuit of a transmission line and a perfectly matched, variable attenuator.

Chapter 7 is a very broad presentation of electromagnetic (EM) field theory tailored to the needs of the microwave and RF engineer. It begins with the physics and the defining experiments that led to the formulation of Maxwell's equations, which are then used to derive fundamental results throughout the chapter. This includes the famous wave equation, from which Maxwell was first led to conclude that light and electromagnetic fields were one and the same.

Throughout this book, techniques are introduced as needed. This is particularly true in this chapter. Vector mathematics are presented including the gradient, dot product, cross product, divergence, curl, and Laplacian as they are required to describe EM field properties and relationships. This direct applicability of the vector operations helps to promote a physical understanding of them as well as the electromagnetic field relationships they are used to describe.

The depth of Chapter 7 is unusual for an introductory text. It extends from the most basic of concepts to quite advanced applications. Skin effect, intrinsic impedance of conductors, Poynting's theorem, wave polarization, the derivation of coaxial transmission line and rectangular waveguide propagating fields, Fourier series and Green's functions, higher order modes in circuits, vector potential, antennas, and radio system path loss are developed in mathematical detail.

Under the best of circumstances, field theory is difficult to master. To accommodate this wide range of electromagnetic topics, the mathematical derivations are uncommonly complete, including many intermediate steps often omitted but necessary for efficient reading and more rapid understanding of the principles.

Chapter 7 concludes with an important use of the computer to perform EM field simulation of distributed circuits. This is shown to provide greater design accuracy than can be obtained with conventional, ideal distributed models.

Chapter 8 treats directional couplers, an important ingredient of microwave measurements and systems. This chapter shows how couplers are analyzed and used. It introduces the *even- and odd-mode analysis* method, which is demonstrated by an analysis of the backward wave coupler. The results, rarely found so thoroughly described in any reference, describe an astounding device. The backward wave coupler has perfect match, infinite isolation, and exactly 90° phase split *at all frequencies*. Cohn's reentrant geometry, used to achieve a 3-dB backward wave, 5-to-1 bandwidth coupler is presented. The uses of couplers as power dividers, reflection phase shifter networks, and as impedance measuring elements in network analyzers are also discussed.

Chapter 9 shows the reader how to design filters beginning with low-pass prototypes having maximally flat (Butterworth), equal-ripple (Chebyshev), and near constant delay (Bessel) characteristics. The classic techniques for scaling these filters to high-pass, bandpass, and bandstop filters are provided. The effect of filter Q on insertion loss is demonstrated. The elliptic filter, having equal stopband ripple, is introduced. Identical resonator filters using top coupling are described as a means to extend the practical frequency range of lumped-element designs.

Half-wave transmission line resonators are used to introduce distributed filters. The Richards transformation and Kuroda's identities are presented as a means of translating lumped-element designs to distributed filters. Mumford's quarter-wave stub filters are presented and shown to be a suitable basis for simulation software optimization of equal-ripple and other passband filters. Kuroda's identities are presented in terms of transmission lines rather than the customary, but vague, "unit elements," simplifying their adoption. This permits students to understand and use Kuroda's identities immediately, even proving their validity as one of the exercises.

Chapter 9 is concluded with a treatment of manufacturing yield illustrated using a filter circuit. A special method of integrating the Gaussian, or normal curve, is presented showing how the "one-sigma" specification is used to de-

termine component and circuit yield. The evaluation of the yield of a practical filter circuit using the network simulator is presented. In this process specifications are applied to the circuit and its performance analyzed assuming it is fabricated using a random sample (*Monte Carlo analysis*) of normally distributed components. The resulting yield from 500 circuits so ''built'' is determined, showing how the effects of component tolerances and specifications on production yields can be determined even before any materials are procured or assembled.

Chapter 10 is applied to transistor amplifiers. The key to amplifier design is the stabilizing and matching of the transistor to its source and load environment, but this must be performed by taking the whole frequency range over which the device has gain into account, a massive calculation task if performed manually. This is handled using S parameters and the network simulator as a design tool. Constant gain and noise figure circles on the Smith chart are described and their design use demonstrated with actual transistor parameters.

The principal design methods including unilateral gain, operating gain, available gain, simultaneously matched, and low noise amplifier techniques are described and demonstrated with available transistor S parameters. Special amplifier topics are presented, including unilateral figure of merit, nonlinear effects, gain saturation, third-order intercept, spurious free dynamic range, and noise limits. The effects of VSWR interaction with cascaded amplifier stages are demonstrated and the use of negative feedback to reduce the VSWR interaction and to design well-matched, broadband amplifiers is shown.

The intent of including so much theoretical and practical material in this text is to provide an immediate familiarity with a wide variety of circuits, their capabilities and limitations, and the means to design them. This permits the engineer to proceed directly to a practical circuit design without the daunting task of researching the material in multiple library references. These topics are illustrated with recommendations on how to use computer optimizations intelligently to direct the computer to search for *circuits whose performance is realistically achievable*.

One could spend years in the microwave engineering practice and not gain experience with this broad a spectrum of topics. The student who reads this book and completes its exercises, in my experience, will be unusually well qualified to embark on a microwave and RF engineering career.

Comments and corrections from readers are welcome.

JOSEPH F. WHITE
jfwhite@ieee.org

◼◼◼ ACKNOWLEDGMENTS

THE AUTHOR

Joseph White is an instructor and consultant in the RF and microwave community, also known as the "wireless" industry.

He received the BS EE degree from Case Institute of Technology, the MS EE degree from Northeastern University and the Ph.D. degree from the Electrical Engineering Department of Rensselaer Polytechnic Institute with specialty in electrophysics and engaged in semiconductor engineering at M/A-

Introduction

1.1 BEGINNING OF WIRELESS

> WIRELESS TELEGRAPHY—At a time when relations are strained between Spain and this country, nothing could be more welcome than a practical method of carrying on electrical communication between distant points on land, and between ships at sea, without any prearranged connection between the two points. During the last year Guglielmo Marconi, an Italian student, developed a system of wireless telegraphy able to transmit intelligible Morse signals to a distance of over ten miles. It has been left, however, for an American inventor to design an apparatus suitable to the requirements of wireless telegraphy in this country. After months of experimenting, Mr. W. J. Clarke, of the United States Electrical Supply Company, has designed a complete wireless telegraphy apparatus that will probably come rapidly into use.
>
> —Scientific American April, 1898

This announcement appeared near the beginning of radio technology. Webster's dictionary [1] lists over 150 definitions that begin with the word *radio*, the first being:

> 1a. ... the transmission and reception of electric impulses or signals by means of electromagnetic waves without a connecting wire (includes wireless, television and radar).

This remains today the real definition of *wireless* and, equivalently, *radio*. Today the uses of radio communication include not only the broadcast of sound through amplitude modulation (AM) and frequency modulation (FM) radio and video through television, but also a broad collection of radio applications, cordless telephones, cell phones, TV, and VCR remotes, automobile remote door locks, garage door openers, and so on.

There is some question about who actually invented radio as a communica-

tive method. Mahlon Loomis, a dentist, experimented with wireless telegraphy using wires supported by kites and a galvanometer to sense the changes in current flow in a second wire when the ground connection of the first was interrupted. He received a patent in 1873 for this system [2].

James Clerk Maxwell [3], more about Maxwell's equations later, predicted the propagation of electromagnetic waves through a vacuum in about 1862. Nathan Stubblefield, a Kentucky farmer and sometimes telephone repairman, demonstrated wireless telephony as early as 1892, but to only one man, and in 1902 to a group [2].

Alexander Popov is said to have "utilized his equipment to obtain information for a study of atmospheric electricity . . . On 7 May 1895, in a lecture before the Russian Physicist Society of St. Petersburg, he stated he had transmitted and received signals at an intervening distance of 600 yards" [4]. In 1888 Heinrich Hertz conducted an experimental demonstration in a classroom at Karlsruhe Polytechnic in Berlin of the generation and detection of the propagating electromagnetic waves predicted by Maxwell [2].

Sir Oliver Lodge, a professor at Liverpool University was experimenting with wireless telegraphy in 1888, and he patented a system in 1897. Marconi purchased his patent in 1911 [2].

In the public mind Guglielmo Marconi enjoys the most credit for "inventing" radio. He was awarded patents for it; therefore, the Patent Office believed that he had made radio-related inventions. However, the U.S. Navy report [4] states

> Marconi can scarcely be called an inventor. His contribution was more in the fields of applied research and engineering development. He possessed a very practical business acumen, and he was not hampered by the same driving urge to do fundamental research, which had caused Lodge and Popoff to procrastinate in the development of a commercial radio system.

This is perhaps the most accurate description of Marconi's role in developing radio technology, a new communication medium. Nikola Tesla had earlier patents, although the focus of his work appears to have been directed to the transmission of power rather than to communication via radio waves. Tesla, well known for his *Tesla coil* that generated high voltages, actually detected signals consisting of noise bursts, resulting from the large atmospheric electrical discharges he originated, that had traveled completely around the earth. In 1943 the U.S. Supreme Court ruled that Marconi's patents were invalid due to Tesla's prior descriptions, but by that time both Marconi and Tesla were deceased [2].

From its beginnings around 1900, radio moved out to fill many communicative voids. In 1962 George Southworth, a well-known researcher in the field of microwaves, wrote a book about his 40 years of experience in the field [5, p. 1]. He begins:

One of the more spectacular technical developments of our age has been radio. Beginning about the turn of the century with ship-to-shore telegraphy, radio has been extended through the years to intercontinental telegraphy, to broadcasting, to radio astronomy and to satellite communications.

Today, after an additional 40 years, Southworth could make a much longer list of radio applications. It would include garage door openers, global positioning satellites, cellular telephones, wireless computer networks, and radar applications such as speed measurement, ship and aircraft guidance, military surveillance, weapon directing, air traffic control, and automobile anticollision systems. The frequency spectrum for practical wireless devices has expanded as well. Amplitude modulated radio begins at 535 kHz and television remote controls extend into the infrared.

The advance of wireless applications is not complete and probably never will be. Certainly the last decade has seen an explosive growth in applications. And the quantities of systems has been extraordinary, too. Witness the adoption of the cellular telephone, which today rivals the wired telephone in numbers of applications.

Sending signals over telegraph wires formed the basis for the early wireless technology to follow. Using the Current International Morse code characters for the early Morse code message transmitted over the first telegraph wires, the first message inaugurating service between Baltimore and Washington, D.C., in 1843, would have looked like

.-- .... .- -   .... .- - ....   --. --- -.. .-- .-. --- ..- --. .... -   ..--..
W h a t   h a t h   G o d w r o u g h t   ?

Most of the full code cipher is shown in Figure 1.1-1. Morse code remains useful, although fewer individuals can interpret it on the fly. A distress signal using the code in Figure 1.1-1 can be sent using a transmitting radio or even a flashlight. Marconi's early wireless transmissions used pulse code modulation,

A .-	*K* -.-	*U* ..-	*5*	,	*(COMMA)*	--..--
B -...	*L* .-..	*V* ...-	*6* -....	.	*(PERIOD)*	.-.-.-
C -.-.	*M* --	*W* .--	*7* --...	?		..--..
D -..	*N* -.	*X* -..-	*8* ---..	;		-.-.-.
E .	*O* ---	*Y* -.--	*9* ----.	:		---...
F ..-.	*P* .--.	*Z* --..	*0* -----	'	*(APOSTROPHE)*	.----.
G --.	*Q* --.-	*1* .----		-	*(HYPHEN)*	-....-
H	*R* .-.	*2* ..---		/	*(slash)*	-..-.
I ..	*S* ...	*3* ...--	(*or*)		*PARENTHESIS*	-.--.-
J .---	*T* -	*4* ....-	‾‾‾‾		*UNDERLINE*	..--.-

Figure 1.1-1 International Morse Code remains a standard for distress signals, S.O.S. is (... --- ...) *(English Characters, [1]). Derived from the work of Samuel Morse (1791–1872).*

Radiated Signal

Information
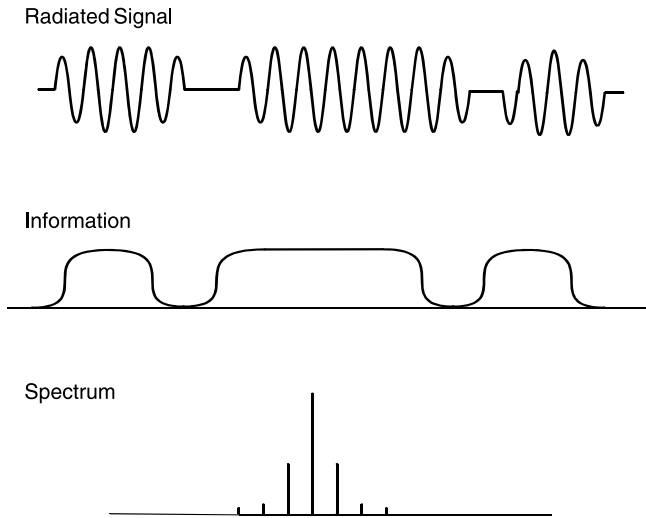
Spectrum

Figure 1.1-2   Modulation format for Morse code, illustrated for letter *R*. Today, pulse shaping, as suggested above, would be employed to reduce transmission spectrum, but Marconi's spark gap transmitter doubtless spanned an enormously wide bandwidth.

dots and dashes achieved by keying the transmitter on and off. Some nautical buoys are identifiable by the Morse letter that their lights flash.

Today, Marconi would need a transmitting license, and were he to continue with his prior transmission technique, his license almost certainly would be suspended due to the broad spectrum of his transmissions (Fig. 1.1-2). His RF source was a spark gap oscillator (Fig. 1.1-3), likely occupying a very broad transmission bandwidth. Powered by a several horsepower generator, the operating transmitter was audible without a radio receiver for several miles.

Marconi had his pivotal triumph in December, 1901, when the Morse character "s" was received at St. John's, Newfoundland (Figs. 1.1-4 and 1.1-5). It was transmitted from Poldhu, Cornwall England, 1800 miles across the Atlantic Ocean [5, p. 13; 6, p. 4]. From the South Wellfleet station, Marconi, himself, transmitted the first trans-Atlantic message on January 17, 1903, a communication from the president of the United States to the king of England.

1.2   CURRENT RADIO SPECTRUM

Today's radio spectrum is very crowded. Obtaining a commercial license to radiate carries the obligation to use bandwidth efficiently, using as little bandwidth as practical to convey the information to be transmitted (Tables 1.2-1 and 1.2-2).

Just the frequency allocations for the United States alone cannot be placed in a table of reasonable size. They occupy numerous pages of the *Rules and*
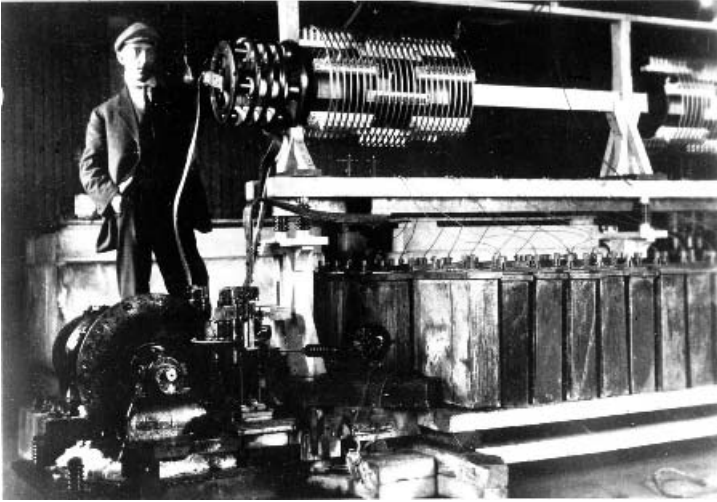
Figure 1.1-3   Joel Earl Hudson standing by Marconi's spark gap transmitter in 1907. (*Photo courtesy of Cape Cod National Seashore.*)

Regulations of the Federal Communications Commission, and have hundreds of footnotes. Since frequent changes are made in the rules and regulations, the latest issue always should be consulted [7, p. 1.8; 8].

As can be seen from Table 1.2-3, radio amateurs today enjoy many frequency allocations. This is due to the history of their pioneering efforts, partic-
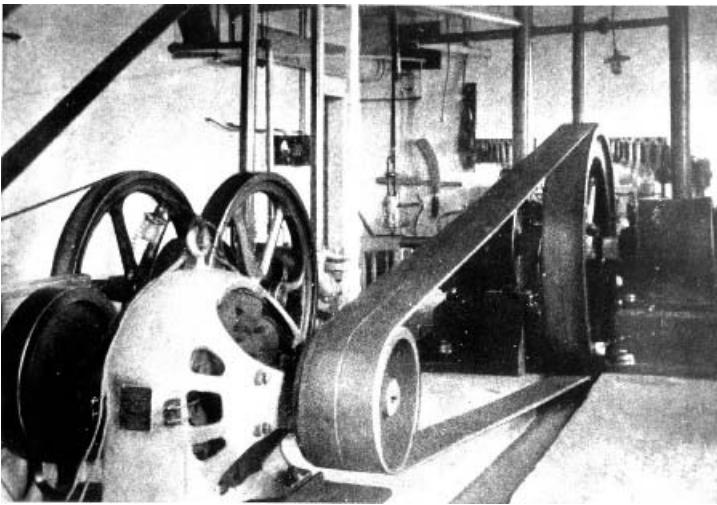
Figure 1.1-4   Prime power for Marconi's South Wellfleet transmitter. (*Photo courtesy of Cape Cod National Seashore.*)
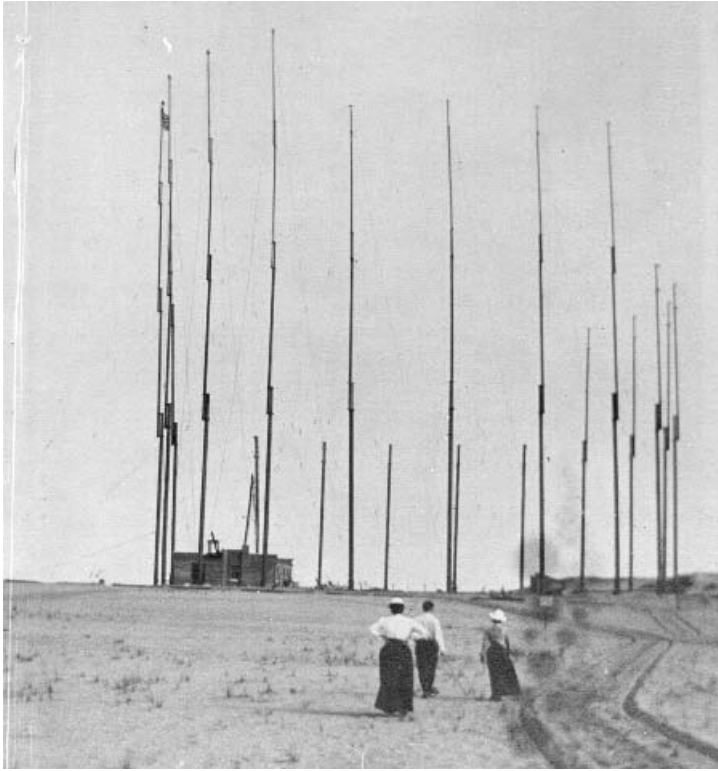
Figure 1.1-5 Marconi's first wireless station in South Wellfleet, Cape Cod, Massachusetts. Local residents predicted that antennas would blow down in first good storm. They did, and he rebuilt them. (*Photo courtesy of Cape Cod National Seashore.*)

ularly at the higher frequencies. We owe much of the rapid development of short-wave radio to the experimental enterprise of amateur radio operators. George Southworth [5, p. 83] pointed out that, in about 1930:

It is interesting that while the telephone people [researchers at the Bell Telephone Laboratories] were conducting intensive research on the lower frequencies ... much was happening in the outside world at higher frequencies.... It is said that the advantages of short waves were first discovered by an amateur who had built for himself a short-wave receiver and upon listening had found that he could hear the harmonics of distant broadcasting stations ... at distances far beyond those at which the fundamentals could be heard. Amateurs later built for themselves short-wave transmitters and soon thereafter carried on two-way communication.

Today, the electromagnetic spectrum is like a superhighway. There are only so many lanes and only so much traffic that it can sustain if everyone is to enjoy rapid and efficient transport.

Figure 1.1-6 Guglielmo Marconi (left) received the Nobel Prize for his wireless communication work. He is shown in a 1901 photo with assistant George Kemp shortly after a successful wireless transmission test. (*Photo courtesy of Marconi, Ltd.*, UK.)

The simultaneous functioning of the intricate grid of radiation allocations, only a part of which are shown in Table 1.2-3, depend upon each user occupying his or her precise frequency, modulation format, bandwidth, and effective radiated power and, furthermore, not intruding on other frequency bands by generating spurious signals with his or her equipment. This is the task and challenge of today's high frequency engineering.

TABLE 1.2-1   General Frequency Band Designations

f	λ	Band	Description
30–300 Hz	10^4–10^3 km	ELF	Extremely low frequency
300–3000 Hz	10^3–10^2 km	VF	Voice frequency
3–30 kHz	100–10 km	VLF	Very low frequency
30–300 kHz	10–1 km	LF	Low frequency
0.3–3 MHz	1–0.1 km	MF	Medium frequency
3–30 MHz	100–10 m	HF	High frequency
30–300 MHz	10–1 m	VHF	Very high frequency
300–3000 MHz	100–10 cm	UHF	Ultra-high frequency
3–30 GHz	10–1 cm	SHF	Superhigh frequency
30–300 GHz	10–1 mm	EHF	Extremely high frequency (millimeter waves)

Source: From Reference [7, Section 1].

TABLE 1.2-2   Microwave Letter Bands

f (GHz)	Letter Band Designation
1–2	L band
2–4	S band
4–8	C band
8–12.4	X band
12.4–18	Ku band
18–26.5	K band
26.5–40	Ka band

Source: From Reference [9, p. 123].

1.3   CONVENTIONS USED IN THIS TEXT

This section lists the notational conventions used throughout this text.

Sections

Sections use a decimal number. To the left of the decimal is the chapter number and to the right is the section number. Thus, 7.10 refers to the tenth section in Chapter 7.

Equations

Equations have a number sequence that restarts in each section. Therefore, a reference to (7.15-4) is directed to the fourth equation in Section 7.15.

Figures

Figure and table numbering also restarts in each section. Therefore, a reference to Figure 7.24-2 relates to the second figure in Section 7.24.

Exercises

The exercises at the end of each chapter are numbered according to the section to which they most closely relate. For example, the exercise numbered E3.5-1 is the first exercise relating to the material in Section 3.5. Material contained in prior sections also may be needed to complete the exercise.

Symbols

The principal symbols used in this text and the quantities that they represent are listed in Appendix A. For example, c refers to the velocity of electromagnetic propagation in free space, while v refers to the velocity of propagation in

TABLE 1.2-3 Selected U.S. Radio Frequency Allocations

Frequencies in kHz	Allocated Purposes
490–510	Distress (telegraph)
510–535	Government
535–1605	AM radio
1605–1750	Land/mobile public safety
1800–2000	Amateur radio

Frequencies in MHz	Allocated Purposes
26.96–27.23, 462.525–467.475	Citizen band radios
30.56–32, 33–34, 35–38, 39–40, 40.02–40.98, 41.015–46.6, 47–49.6, 72–73, 74.6–74.8, 75.2–76, 150.05–156.2475, 157.1875–161.575, 162.0125–173.4	Private mobil radio (taxis, trucks, buses, railroads)
220–222, 421–430, 451–454, 456–459, 460–512	
746–824, 851–869, 896–901, 935–940	
74.8–75.2, 108–137, 328.6–335.4, 960–1215, 1427–1525, 220–2290, 2310–2320, 2345–2390	Aviation (communication and radar)
162.0125–173.2	Vehicle recovery (LoJack)
50–54, 144–148, 216–220, 222–225, 420–450, 902–928, 1240–1300, 2300–2305, 2390–2450	Amateur radio
72–73, 75.2–76, 218–219	Radio control (personal)
54–72, 76–88, 174–216, 470–608	Television broadcasting VHF and UHF
88–99, 100–108	FM radio broadcasting
824–849	Cellular telephones
1850–1990	Personal communications
1910–1930, 2390–2400	Personal comm. (unlicensed)
1215–1240, 1350–1400, 1559–1610	Global Positioning Systems (GPS)

Frequencies in GHz	Allocated Purposes
0.216–0.220, 0.235–0.267, 0.4061–0.45, 0.902–0.928, 0.960–1.215, 1.215–2.229, 2.320–2.345, 2.360–2.390, 2.7–3.1, 3.1–3.7, 5.0–5.47, 5.6–5.925, 8.5–10, 10.0–10.45, 10.5–10.55, 13.25–13.75, 14–14.2, 15.4–16.6, 17.2–17.7, 24.05–24.45, 33.4–36, 45–46.9, 59–64, 66–71, 76–77, 92–100	Radar, all types
2.390–2.400	LANs (unlicensed)
2.40–2.4835	Microwave ovens
45.5–46.9, 76–77, 95–100, 134–142	Vehicle, anticollision, navigation
10.5–10.55, 24.05–24.25	Police speed radar
0.902–0.928, 2.4–2.5, 5.85–5.925	Radio frequency identification (RFID)
3.7–4.2, 11.7–12.2, 14.2–14.5, 17.7–18.8, 27.5–29.1, 29.25–30, 40.5–41.5, 49.2–50.2	Geostationary satellites with fixed earth receivers

TABLE 1.2-3 *(Continued)*

Frequencies in GHz	Allocated Purposes
1.610–1626.5, 2.4835–2.5, 5.091–5.25, 6.7–7.075, 15.43–15.63	Nongeostationary satellites, mobile receivers (big LEO, global phones)
0.04066–0.0407, 902–928, 2450–2500, 5.725–5.875, 24–24.25, 59–59.9, 60–64, 71.5–72, 103.5–104, 116.5–117, 122–123, 126.5–127, 152.5–153, 244–246	Unlicensed industrial, scientific, and medical communication devices
3.3–3.5, 5.65–5.925, 10–10.5, 24–24.25, 47–47.2	Amateur radio
6.425–6.525, 12.7–13.25, 19.26–19.7, 31–31.3	Cable television relay
27.5–29.5	Local multipoint TV distribution
12.2–12.7, 24.75–25.05, 25.05–25.25	Direct broadcast TV (from satellites)
0.928–0.929, 0.932–0.935, 0.941–0.960, 1.850–1.990, 2.11–2.20, 2.450–2.690, 3.7–4.2, 5.925–6.875, 10.55–10.68, 10.7–13.25, 14.2–14.4, 17.7–19.7, 21.2–23.6, 27.55–29.5, 31–31.3, 38.6–40	Fixed microwave (public and private)

a medium for which the relative dielectric and permeability constants may be greater than unity.

Prefixes

Except where noted otherwise, this text uses the International System of Units (SI). Standard prefixes are listed in Table 1.3-1.

Fonts

The font types used throughout this text to connote variable types are listed in Table 1.3-2. Combinations of these representational styles are used to convey the dual nature of some variables. For example, in Maxwell's equation

$$\nabla \cdot \vec{\mathbf{D}} = \rho$$

$\vec{\mathbf{D}}$ is written in regular type because the equation applies to all time waveforms, not just sinusoidal variations, and $\vec{\mathbf{D}}$ is also a vector quantity. On the other hand, the Helmholtz equation is written

$$\nabla^2 \vec{E} + k^2 \vec{E} = 0$$

TABLE 1.3-1   Standard Prefixes

Prefix	Abbreviation	Factor
tera	T	10^{12}
giga	G	10^{9}
mega	M	10^{6}
kilo	k	10^{3}
hecto	h	10^{2}
deka	da	10
deci	d	10^{-1}
centi	c	10^{-2}
milli	m	10^{-3}
micro	μ	10^{-6}
nano	n	10^{-9}
pico	p	10^{-12}
femto	f	10^{-15}
atto		10^{-18}

using italic type for the variable $\vec{E}$ because *this equation only applies for sinusoidal time variations*, and therefore the components of the vector $\vec{E}$ are *phasor quantities*.

Throughout this text, except where otherwise noted, the magnitudes of sinusoidal waveforms (V, I, E, D, H, B) are peak values. To obtain root-mean-square (rms) values, divide these values by $\sqrt{2}$.

1.4   VECTORS AND COORDINATES

General vector representations are three dimensional. They can be described by any three-dimensional, *orthogonal coordinate system* in which each coordinate direction is at right angles to the other two. Unless otherwise specified, *rectan-*

TABLE 1.3-2   Fonts Used in This Text to Identify Variable Types

Variable Type	Font	Examples
DC or general time-varying function (not sinusoidal)	Regular type	V, I, H, E, B, D
Explicit general time variation	Regular type, lower-case	v(t), i(t)
Explicit sinusoidal time variation	Italic type, lowercase	$v(t), i(t)$
Phasors, impedance, admittance, general functions, and variables, unit vectors	Italic type	V, I, H, E, B, D, Z, Y $f(x), g(y), x, y, z, \vec{x}, \vec{y}, \vec{z}$
Vectors	Arrow above	$\vec{E}, \vec{H}, \vec{B}, \vec{D}, \vec{E}, \vec{H}, \vec{B}, \vec{D}$
Normalized parameters	Lowercase	$z = Z/Z_0, \ y = Y/Y_0$

gular (Cartesian) coordinates are implied. Certain circular and spherical symmetries of a case can make its analysis and solution more convenient if the geometry is described in *cylindrical coordinates* or *spherical coordinates*.

In this text all coordinate systems are *right-handed orthogonal coordinate systems*. That is,

> In a right-hand orthogonal coordinate system, rotating a vector in the direction of any coordinate into the direction of the next named coordinate causes a rotational sense that would advance a right-hand screw in the positive direction of the third respective coordinate.

We define that *unit vectors are vectors having unity amplitude and directions in the directions of the increasing value of the respective variables that they represent.*

In rectangular coordinates (Fig. 1.4-1) the order is (x, y, z) and an arbitrary point is written as $P(x, y, z)$. The unit vectors in these respective directions are $\vec{x}$, $\vec{y}$, and $\vec{z}$. Thus, a three-dimensional vector field $\vec{E}$ can be written

$$\vec{E} = \vec{E}_x + \vec{E}_y + \vec{E}_z \tag{1.4-1a}$$

or

$$\vec{E} = E_x\vec{x} + E_y\vec{y} + E_z\vec{z} \tag{1.4-1b}$$

or

$$\vec{E} = \vec{x}E_x + \vec{y}E_y + \vec{z}E_z \tag{1.4-1c}$$
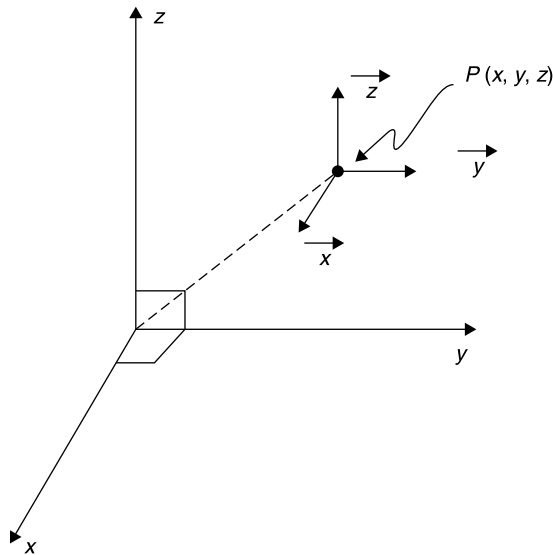
Figure 1.4-1   Rectangular (Cartesian) right-hand coordinate system.

Generally, the format of (1.4-1c) is used in this text. In the language of vector mathematics, rotating a unit vector $\vec{x}$ in the direction of another unit vector $\vec{y}$ is called *crossing* $\vec{x}$ into $\vec{y}$, and this is written as $\vec{x} \times \vec{y}$. This is a specific example of the *vector cross product*. The vector cross product can be applied to any two vectors having any magnitudes and relative orientations; but, in general, we must take into account the product of their magnitudes and the angle between them, as will be shown more specifically for the vector cross product in Chapter 7. For present purposes, since $\vec{x}, \vec{y}$, and $\vec{z}$ form a right-hand orthogonal set of unit vectors, we can express the right-handedness of their coordinate system by requiring that the following cross product relations apply:

$$\vec{x} \times \vec{y} = \vec{z} \tag{1.4-2a}$$

$$\vec{y} \times \vec{z} = \vec{x} \tag{1.4-2b}$$

$$\vec{z} \times \vec{x} = \vec{y} \tag{1.4-2c}$$

Notice that *the vector cross product yields a new vector that is orthogonal to the plane of the crossed vectors and in a direction that would be taken by the advance of a right-hand screw when the first vector is crossed into the second.*

Also notice that *for a right-hand coordinate system any coordinate unit vector can be crossed into the next named coordinate vector to yield the direction of positive increase of the remaining coordinate, beginning with any coordinate.* For example (x, y, z), (y, z, x), or (z, x, y) all satisfy the right-hand advancing rule, as specified by (1.4-2a) to (1.4-2c).

The cylindrical coordinate system is shown in Figure 1.4-2. The order of coordinate listing is (r, φ, z) and the unit vectors are $\vec{r}, \vec{\varphi}$, and $\vec{z}$, which satisfy
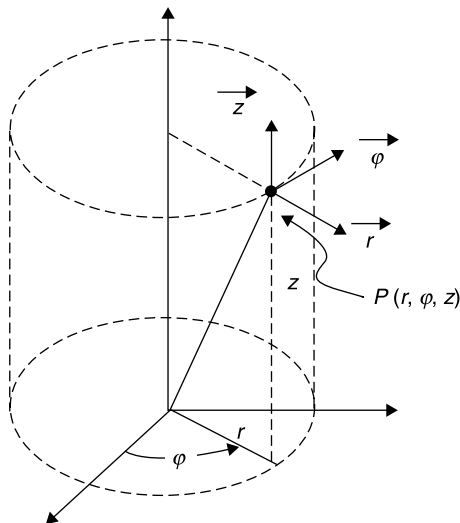
Figure 1.4-2   Cylindrical right-hand coordinate system.
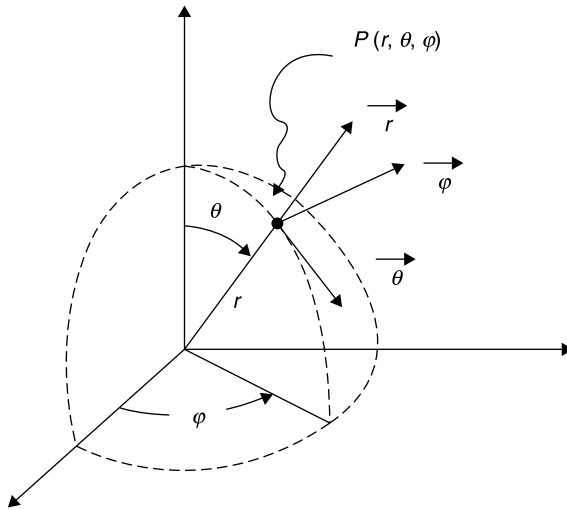
Figure 1.4-3   Spherical right-hand coordinate system.

the same sequential cross-product rules as do rectangular coordinates, namely $\vec{r} \times \vec{\phi} = \vec{z}$, $\vec{\phi} \times \vec{z} = \vec{r}$, and $\vec{z} \times \vec{r} = \vec{\phi}$.

The spherical coordinate system is shown in Figure 1.4-3. The order of coordinate listing is (r, θ, ϕ) and the unit vectors are $\vec{r}, \vec{\theta}$, and $\vec{\phi}$, which satisfy the sequential cross-product rules $\vec{r} \times \vec{\theta} = \vec{\phi}$, $\vec{\theta} \times \vec{\phi} = \vec{r}$, and $\vec{\phi} \times \vec{r} = \vec{\theta}$. *Note that this r is not the same as the r used in cylindrical coordinates.*

1.5   GENERAL CONSTANTS AND USEFUL CONVERSIONS

There are several values of physical constants, conversion factors, and identities useful to the practice of microwave engineering. For ready reference, a selection of them is printed on the inside covers of this text.

REFERENCES

1. *Webster's Third New International Dictionary*, G. & C. Meriam Co Springfield, Massachusetts, 1976. *Copy of the International Morse Code, including special characters. See* Morse code.
2. Don Bishop, "Who invented radio?" *RF Design*, February, 2002, p. 10.
3. James Clerk Maxwell, *Electricity and Magnetism*, 3rd ed., Oxford, 1892, Part II.
4. United States Navy, *History of Communications—Electronics in the United States Navy*, U.S. Government Printing Office, Washington, DC, 1963.
5. George C. Southworth, *Forty Years of Radio Research*, Gordon and Breach, New York, 1962.

6. Deryck Henley, *Radio Receiver History and Instructions*, Flights of Fancy, Leamington Spa, Warks England, 2000. *This reference is the set of instructions provided with a modern crystal radio kit*.

7. *Reference Data for Radio Engineers*, 5th ed., Howard W. Sams, New York, 1974. *New editions are available*.

8. Bennet Z. Kobb, RF Design Delivers for Design Engineers . . . from 30 MHz to 300 GHz, March 2000. Published in *RF Design* magazine.

9. George W. Stimson, *Introduction to Airborne Radar*, Hughes Aircraft Company, El Segundo, CA, 1983.

Review of AC Analysis and Network Simulation

Alternating current (AC) circuit analysis is the basis for the high frequency techniques that are covered in subsequent chapters and the subject of this text. It is assumed that all readers already have been introduced to AC analysis. However, it has been the author's teaching experience that a review is usually appreciated because it provides an opportunity to put into perspective the fundamentals needed for the fluent application of AC analysis, a skill essential to the high frequency engineer.

The AC analysis makes use of complex numbers to calculate and keep track of the relative magnitudes and phases of voltages and currents. For this reason, a summary of complex mathematics is included in Appendix B. One who comfortably reads this chapter and Appendix B can follow the remainder of this book. The properties of complex equations, such as the bilinear transformation that defines the Smith chart, were developed by mathematicians probably without any practical application considerations [1].

2.1 BASIC CIRCUIT ELEMENTS

The basic building blocks of electric circuits are the resistor R, the inductor L, and the capacitor C (Fig. 2.1-1). At high frequencies these elements do not behave as pure R, L, and C components but have additional resistances and reactances called *parasitics*. More about parasitics later in the chapter.

The Resistor

The resistor passes a current I equal to the applied voltage divided by its resistance. This can be considered a definition of the resistor. Mathematically, this is written
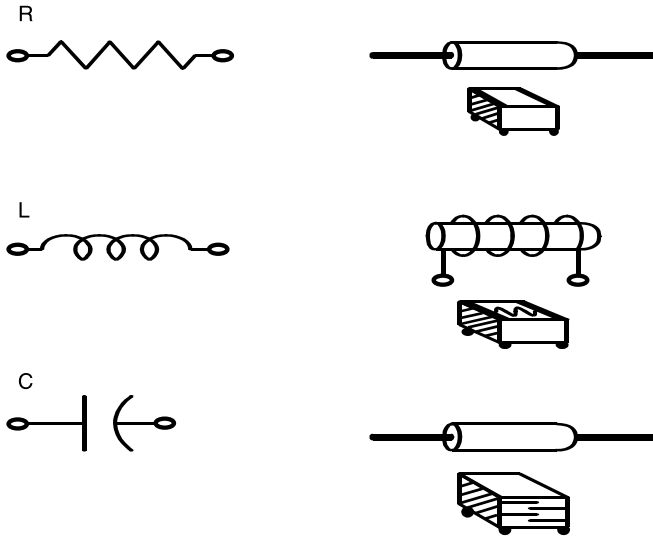
Figure 2.1-1 Practical resistor, inductor, and capacitor, the basic passive, lumped elements.

$$I \equiv \frac{V}{R} \qquad (2.1\text{-}1)$$

Notice that regular type (not italic) is used because (2.1-1) applies for direct currents (DC) as well as every time-varying waveform. For AC signals the current through and voltage across an ideal resistor are in phase (Fig. 2.1-2).

Regardless of the time variation of voltage and current, the instantaneous power dissipated in a resistor is

$$P(t) = v(t)i(t) \qquad (2.1\text{-}2)$$

where $v(t)$ is the instantaneous voltage across the element and $i(t)$ is the current through it at the same instant. With sinusoidal excitation the *average power* dissipated, P_{AVG} in any two-terminal element is

$$P_{AVG} = \tfrac{1}{2} VI \cos \theta \qquad (2.1\text{-}3)$$

where V and I are the peak values of voltage and current and θ is the phase angle between V and I. Frequently, *root-mean-squared (rms)* values are used to describe the voltage and current magnitudes. The rms value provides the same average power as a DC voltage or current of the same amplitude. For sinusoidal variations, the rms value is related to the peak value by

$$V_{rms} = \frac{V_{peak}}{\sqrt{2}} \quad \text{and} \quad I_{rms} = \frac{I_{peak}}{\sqrt{2}} \qquad (2.1\text{-}4a, b)$$
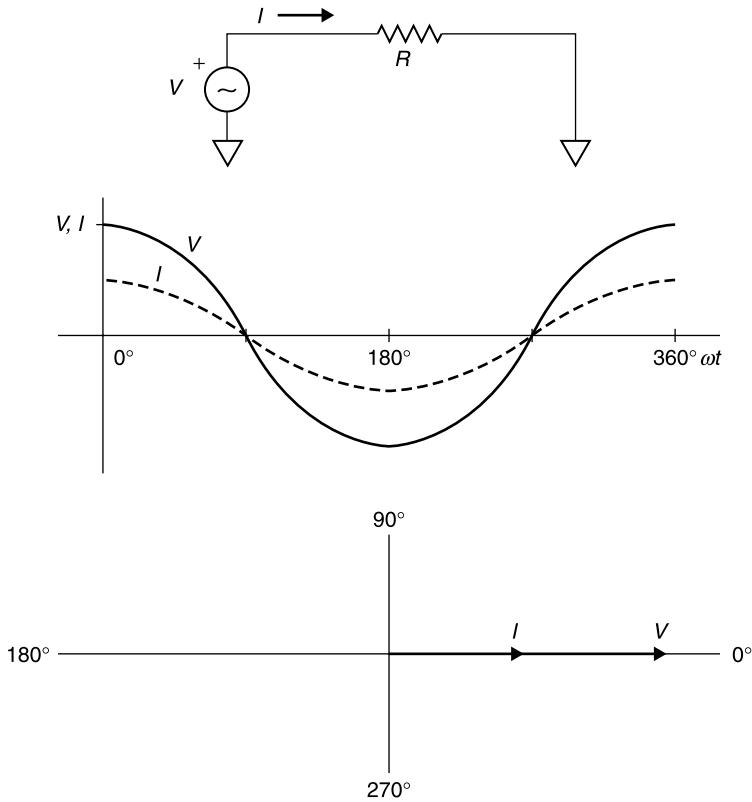
Figure 2.1-2   Ideal resistor R has AC voltage V and current I in phase.

and

$$P_{\text{AVG}} = V_{\text{rms}} I_{\text{rms}} \cos \theta \qquad (2.1\text{-}5)$$

Note that, according to our convention, the variables are in italic type to indicate that the time waveform is sinusoidal. For an ideal resistor, $\theta = 0°$ and the instantaneous power dissipated in the resistor is equal to the product v(t)i(t). *Throughout this text, except where otherwise specified, peak values are used for voltage, current, and field amplitudes.*

Ohm's Law

Ohm's law applies to all voltage waveforms across a resistor, and states that *current through a resistor is directly proportional to the applied voltage and inversely proportional to its resistance:*

$$i = \frac{v}{R} \qquad (2.1\text{-}6)$$

where v is in volts, R in ohms, and i in amperes.

Ohm's law is a *linear relationship* (current is proportional only to the first power of voltage) and is valid for all voltage and current levels that do not change the value of resistance. It does not apply, for example, at very high voltages that cause an arc over of the resistor and/or high currents that cause the resistor to change its value due to overheating.

The Inductor

In contrast to the resistor, the ideal inductor L, cannot dissipate power. The general relationship between voltage across and current through it is

$$v(t) = L\frac{di(t)}{dt} \tag{2.1-7}$$

The instantaneous current through the inductor is obtained by integrating

$$i(t) = \int_0^v v(t)\, dt \tag{2.1-8}$$

The amount of energy, U_L, stored in an inductor is equal to the time integral of the power applied to it, $v(t) \times i(t)$, to establish a current i in it from an initial condition at which $i(t = 0) = 0$:

$$U_L = \int_0^t v(t) \cdot i(t)\, dt = L \int_0^t \frac{d(i)}{dt} \cdot i(t)\, dt \tag{2.1-9}$$

which, on integrating, gives the instantaneous energy stored as

$$U_L = \tfrac{1}{2}L[i(t)]^2 \tag{2.1-10}$$

Current through the inductor *for all waveforms* is proportional to the integral of the applied voltage. With a sinusoidally applied voltage, $v(t) = V_0 \cos \omega t$, the current I is obtained by integration, noting that the constant of integration corresponds to a DC term that can be neglected for an AC solution (Fig. 2.1-3). Thus,

$$I = \frac{1}{L}\int V_0 \cos \omega t\, dt = \frac{V_0 \sin \omega t}{\omega L} \tag{2.1-11}$$

From Ohm's law current divided by voltage has the dimensions of ohms, therefore ωL must have the dimensions of "ohms." This will be used shortly in the definition of complex impedance. There is no power dissipated in an inductor. For AC excitation, the phase angle, θ, between voltage and current is $-90°$ and the power dissipated, P_{Diss}, is given by

$$P_{\text{Diss}} = \tfrac{1}{2}|V|\,|I| \cos \theta = 0 \tag{2.1-12}$$
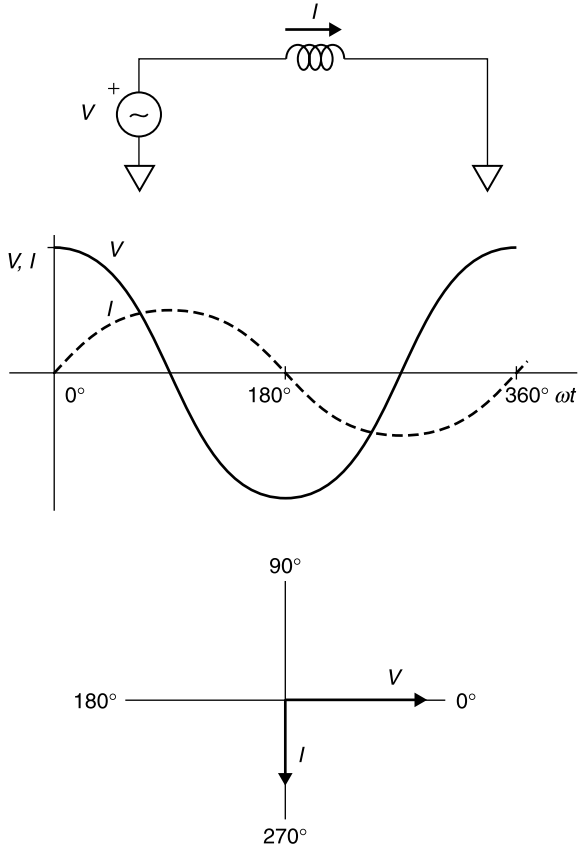
Figure 2.1-3   Sinusoidal current I through inductor lags voltage V by 90°.

The Capacitor

Like the inductor, the capacitor, C, cannot dissipate power. Rather it stores charge and, in so doing, stores energy. By definition, the capacitance C is defined as the ratio of instantaneous charge q to the instantaneous voltage $v(t)$ at which the charge is stored:

$$C \equiv \frac{q}{v(t)} \qquad (2.1\text{-}13)$$

With a current into the capacitor the stored charge increases. The time rate of change of q is equal to this current. Multiplying both sides of (2.1-13) by v and then differentiating with respect to t give

$$\frac{\partial q}{\partial t} = i(t) = C\frac{\partial v(t)}{\partial t} \qquad (2.1\text{-}14)$$
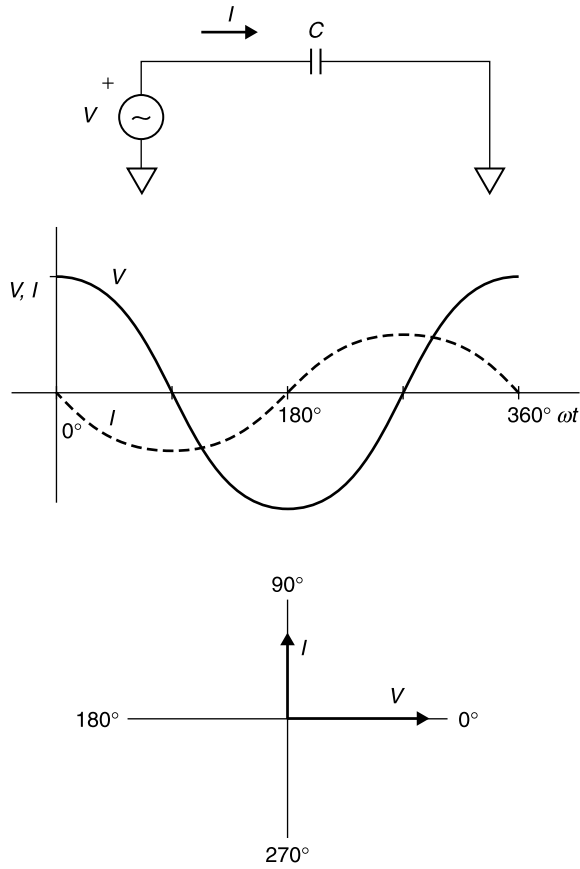
Figure 2.1-4 Ideal capacitor C is also a dissipationless component. For sinusoidal excitation, current *I* leads applied voltage *V* by 90°.

Integrating with respect to *t* gives

$$v(t) = \frac{1}{C} \int i \, dt \tag{2.1-15}$$

When a direct current is passed into a capacitor, the voltage across the capacitor's terminals "integrates" the direct current flow from the time when the capacitor had zero volts (Fig. 2.1-4). The capacitor does not dissipate power, but rather stores energy. The energy storage can be considered the presence of charge in a potential field or the establishment of an electric field between the capacitor plates. For an initially uncharged capacitor, $v(0) = 0$, the integral with respect to time of the instantaneous power delivered to the capacitor, $v(t) \times i(t)$, is the stored energy U_C in the capacitor when it is charged to a

voltage V:

$$U_C = \int_0^v v(t) i(t) \, dt = C \int_0^v v(t) \frac{dv}{dt} \, dt \qquad (2.1\text{-}16)$$

which, upon integrating, gives the instantaneous stored energy as

$$U_C = \tfrac{1}{2} C [v(t)]^2 \qquad (2.1\text{-}17)$$

This result does not depend upon the voltage or current waveforms used to store the charge. When a sinusoidal voltage is applied, the current waveform is also sinusoidal and advanced by 90°:

$$v(t) = V_0 \cos \omega t \qquad (2.1\text{-}18)$$

$$i(t) = C \frac{\partial v(t)}{\partial t} = -V_0 \omega C \sin \omega t \qquad (2.1\text{-}19)$$

From (2.1-19) it follows that $1/\omega C$ has the dimensions of ohms, as did ωL. These facts prompt the definition of *complex impedance*, to be discussed shortly.

2.2   KIRCHHOFF'S LAWS

1. *Kirchhoff's voltage law: The sum of the voltage drops about a closed circuit path is zero* (Fig. 2.2-1):

$$v_0(t) = v_R(t) + v_L(t) + v_C(t) \qquad (2.2\text{-}1)$$

2. *Kirchhoff's current law: The sum of the currents into a circuit node is zero* (Fig. 2.2-2):

$$i_0(t) = i_R(t) + i_L(t) + i_C(t) \qquad (2.2\text{-}2)$$
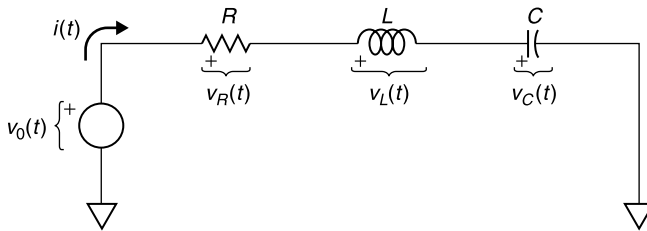
Figure 2.2-1   Kirchhoff's voltage law applied to *RLC* element loop for general time-varying applied voltage $v_0(t)$.
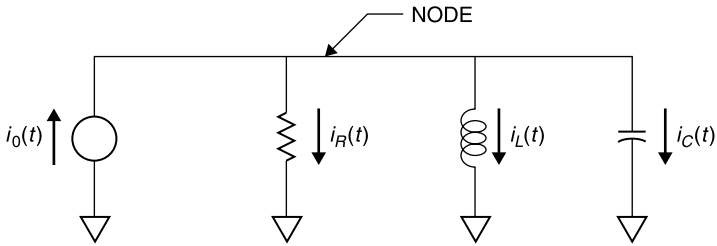
Figure 2.2-2   Kirchhoff's current law applied to node for general time-varying current $i_0(t)$.

Kirchhoff's laws apply instantaneously for all waveforms. For the series circuit of Figure 2.2-1, the applied voltage is equal to the voltage drops across the three element types. The current is continuous in the loop. Using the voltage–current relations of (2.1-6), (2.1-7), and (2.1-15) gives

$$v_0(t) = Ri + L\frac{di}{dt} + \frac{1}{C}\int i\,dt \qquad (2.2\text{-}3)$$

2.3   ALTERNATING CURRENT (AC) ANALYSIS

In electrical engineering it would be nice to have the complete solution in time for $v(t)$ and $i(t)$ whenever a circuit is analyzed in order that both the transient and steady-state behavior would be available. However, we usually find that the steady-state behavior of circuits with sinusoidal excitation is adequate, particularly since it can be obtained with much greater computational economy, as will be seen shortly. If a sinusoidal voltage or current excitation is applied to a network consisting of linear R, L, and C elements the resulting currents and voltages usually approach steady-state sinusoidal waveforms within a few RF cycles. Circuits having very high Q, to be discussed in the next chapter, require longer times, so some judgment is necessary regarding the transient effects in AC networks.

Nevertheless, ignoring transient effects and accepting a steady-state solution for an AC network is usually sufficient. Referring to (2.2-3), we notice that integral and differential expressions occur in the network equation due to the presence of L and C elements. However, the steady-state voltage and current solutions of the network are comprised solely of sinusoidal functions at a common frequency because *all integrals and derivatives of sinusoidal functions are also sinusoidal functions at the same frequency (but displaced in phase by* $\pm 90°$).

For example, if we apply a voltage

$$v(t) = V_0 \cos \omega t \qquad (2.3\text{-}1)$$

to the network in Figure 2.2-1, the resulting current *eventually will approach the steady-state waveform*

$$i(t) = I_0 \cos(\omega t - \varphi) \tag{2.3-2}$$

The steady-state solution that we seek is to solve for I_0 and φ in terms of V_0. Substituting this assumed solution into (2.2-3) and performing the indicated differentiation and integration gives

$$V_0 \cos \omega t = I_0 \left[R \cos(\omega t - \varphi) - \omega L \sin(\omega t - \varphi) + \frac{1}{\omega C} \sin(\omega t - \varphi) \right] \tag{2.3-3}$$

This equation applies for all time t after sufficient time has passed to allow transients to die out (since we ignored the constant of integration associated with the third term). In particular, consider the time for which $\omega t = \pi/2 = 90°$. Then, recognizing that $\cos(90° - \varphi) = \sin \varphi$ and $\sin(90° - \varphi) = \cos \varphi$, (2.3-3) becomes

$$0 = I_0 \left\{ R \sin \varphi - \left[\omega L - \frac{1}{\omega C} \right] \cos \varphi \right\} \tag{2.3-4a}$$

and solving for φ,

$$\tan \varphi = \left[\frac{\omega L - 1/\omega C}{R} \right] \tag{2.3-4b}$$

$$\varphi = \tan^{-1} \left[\frac{\omega L - 1/\omega C}{R} \right] \tag{2.3-4c}$$

If we substitute $t = 0$ into (2.3-3) and recognize that the value for φ in (2.3-4c) allows determination of $\cos \varphi$ and $\sin \varphi$, the result is

$$V_0 = I_0 \left\{ R \cos \varphi + \left[\omega L - \frac{1}{\omega C} \right] \sin \varphi \right\}$$

$$I_0 = \frac{V_0}{R \cos \varphi + [\omega L - 1/\omega C] \sin \varphi} \tag{2.3-5}$$

$$= \frac{V_0}{\sqrt{R^2 + [\omega L - 1/\omega C]^2}}$$

The expression in the denominator has the value of the hypotenuse of a right triangle, as shown graphically in Figure 2.3-1.
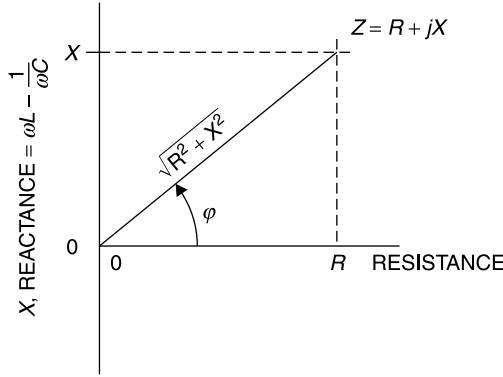
Figure 2.3-1   Orthogonal relationship between resistance R and reactance X in AC circuit. Note that if X is positive, φ is positive, which means that I lags V, consistent with (2.3-2).

The final expression of (2.3-5) appears to be in the form of Ohm's law with resistance replaced by a quantity that includes the "impeding" effects of L and C on current flow. Also, (2.3-4c) shows how L and C affect the phase relation between v and i. Both of these effects can be accounted for by defining a *complex impedance Z*.

The complex impedance Z of a series RLC combination has a real part equal to the resistance R and an imaginary part equal to j times the net reactance, $X = (\omega L - 1/\omega C)$. Thus,

$$Z = R + jX = R + j\left(\omega L - \frac{1}{\omega C}\right) = |Z|\angle\varphi \qquad (2.3\text{-}6a)$$

where

$$|Z| = \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2} \qquad (2.3\text{-}6b)$$

$$\varphi = \tan^{-1}\frac{\omega L - 1/\omega C}{R} \qquad (2.3\text{-}6c)$$

and $\omega = 2\pi f$ where f is the operating frequency in hertz.

The foregoing solution to (2.3-6) required considerable mathematical manipulation, particularly in view of the fact that only the magnitude I_0 and phase φ were unknown. We already knew that the form of $i(t)$ would be a sinusoid with time variation ωt. Clearly, an extension of Ohm's law to render the same results directly would offer a profound improvement in the efficiency of *RLC* circuit analysis.
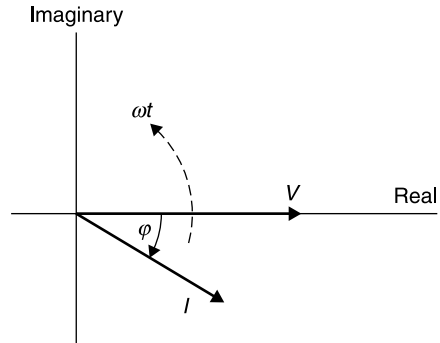
Figure 2.3-2   Voltage and current phasors for solution to network in Figure 2.2-1 where $v(t) = V_0 \cos \omega t$.

Ohm's Law in Complex Form

Ohm's law for AC circuits can be written in complex form as

$$I = \frac{V}{Z} \tag{2.3-7}$$

where V and I are complex quantities called *phasors* and Z is the AC impedance defined in (2.3-6a).

   As can be seen in Figure 2.3-2, the voltage and current are now represented by complex quantities, having real and imaginary parts. They are vectors in the complex plane. However, since electric and magnetic fields are three-dimensional vectors in space, to avoid confusing voltage and current quantities with fields in space, *the Institute of Electrical and Electronic Engineers (IEEE) recommends using the term* phasor *for the complex representations of V and I, as well as for the complex representations of sinusoidally varying field magnitudes.* This convention will be assumed throughout this text, except where specifically noted otherwise. A review of complex mathematics is given in Appendix B.

2.4   VOLTAGE AND CURRENT PHASORS

The application of complex numbers to *RLC* circuit analysis is to represent the sinusoidal voltage $v(t)$ by a phasor voltage V and the resulting current $i(t)$ by a phasor current I. These phasor quantities are complex numbers, having real and imaginary parts. *Complex voltage and current phasors are mathematical artifices that do not exist in reality*. They are vectors in the complex plane that are useful in analyzing AC circuits. Phasors do not rotate; they are fixed position vectors whose purpose is to indicate the magnitude and phase of the sinusoidal waveforms that they represent.

   However, if they were rotated counterclockwise in the complex plane at the rate of ω radians per second (while maintaining their angular separation, φ)

their projections on the real axis would be proportional to the instantaneous time-varying voltage and current waveforms that they represent. The horizontal axis projection represents the instantaneous wave amplitudes because we chose as a reference for this analysis $v(t) = V_0 \cos \omega t$, which has its maximum value at t = 0.

For the network of Figure 2.2-1 the phasors are shown diagrammatically in Figure 2.3-2. In this drawing it is presumed that the inductive reactance, ωL, has a greater magnitude than that of the capacitive reactance, $1/\omega C$. From (2.3-6c) this means that the angle φ is positive, reflecting the fact that I lags V, as shown in Figure 2.3-2.

At a given frequency, in the time domain two values are needed to specify each sinusoidal variable, its peak magnitude and value at $t = 0$. The same information is contained in complex phasors within their real and imaginary parts. Given that $v(t) = V_0 \cos \omega t$, to convert the complex phasors V and I to their respective time-domain variables, we interpret their projections onto the real axis as the instantaneous time value. Thus,

$$v(t) = \mathrm{Re}[Ve^{j\omega t}] \qquad (2.4\text{-}1)$$

$$i(t) = \mathrm{Re}[Ie^{j\omega t}] \qquad (2.4\text{-}2)$$

where V and I are the respective phasor values of voltage and current. For example, if $V = 20\angle 30°$, then $V = 20\angle 30° \equiv 20e^{j30°}$ and

$$v(t) = \mathrm{Re}\{20e^{j30°}e^{j\omega t}\} = \mathrm{Re}\{20e^{j(\omega t+30°)}\} = 20 \cos(\omega t + 30°) \quad (2.4\text{-}3)$$

The peak value of the phasor is the same as the peak value of the sinusoidal waveform that it represents. Similarly, had rms values been used, the rms magnitude of the phasor would be the same as the rms value of the sinusoid it represents.

Conventionally, the reactances of L and C elements are positive real values with the dimensions of ohms as

$$X_L = \omega L \quad \text{and} \quad X_C = \frac{1}{\omega C} \qquad \text{(in ohms)} \qquad (2.4\text{-}4a,b)$$

To obtain AC impedances

$$Z_L = j\omega L \qquad (2.4\text{-}5)$$

since it involves d/dt, which produces j, and

$$Z_C = -j\frac{1}{\omega C} \qquad (2.4\text{-}6)$$

since it involves $\int dt$, which produces $-j$. Accordingly, the impedance of the

series RLC circuit in Figure 2.2-1 is written

$$Z = R + j(X_L - X_C) = R + j\left(\omega L - \frac{1}{\omega C}\right) \qquad (2.4\text{-}7)$$

Kirchhoff's voltage and current laws also apply to the phasor forms of V and I.

    With these definitions and the rules for complex number manipulation, circuits of arbitrary complexity can be analyzed in the phasor domain to solve for the relationships between voltage and currents at any given frequency in a network of linear elements. Then (2.4-1) and (2.4-2) can be used if the instantaneous time functions are required.

2.5 IMPEDANCE

Estimating Reactance

The *reactance of an inductor, L, is ωL, and the reactance of a capacitor, C, is $1/\omega C$. Reactances have the dimensions of ohms.*
    A practicing RF and microwave engineer should be able to estimate the reactances of inductors and capacitors quickly. This need arises, for example, in order to assess their effects on a circuit, the realizability of a proposed tuning method or for a variety of other design and analysis purposes.
    Memorization of two reactance values and a simple scaling method permits these estimates to be made mentally, yielding a close approximation, even without a calculator. The reactance of an inductor, L, is given by

$$X_L = \omega L \quad \text{where} \quad \omega = 2\pi f \qquad (2.5\text{-}1)$$

At 1 GH_z the reactance magnitude of a 1-nH inductor is 6.28 Ω. Therefore, the reactance of any other inductor at any other frequency is given by

$$X_L = 6.28 fL \qquad (f \text{ in gigahertz, } L \text{ in nanohenries}) \qquad (2.5\text{-}2)$$

By remembering the 6.28 Ω/(nH-GHz) scale factor, other inductive reactance values are quickly estimated. For example, a 3-nH inductor at 500 MHz has a reactance

$$X_L = (6.28)(0.5)(3) = 9.42 \ \Omega \qquad (2.5\text{-}3)$$

Similarly, the reactance of a capacitor, C, is given by

$$X_C = \frac{1}{\omega C} = \frac{159}{fC} \qquad (f \text{ in gigahertz, } C \text{ in picofarads}) \qquad (2.5\text{-}4)$$

Thus, remembering that 1 pF yields 159 Ω at 1 GHz allows other capacitive

reactance values to be estimated. For example, a 2-picofarad (pF) capacitor has a reactance at 3 GHz of

$$X_C = \frac{159 \ \Omega}{(2)(3)} = 26.5 \ \Omega \tag{2.5-5}$$

Of course, one must also remember that *inductive reactance is directly proportional to f and L* while *capacitive reactance is inversely proportional to f and C*. In this way, memorizing the 1-GHz reactances of a 1-nH inductor and 1-pF capacitor allows simple scaling of other values.

$$X_L = 6.28fL \ (\Omega) \tag{2.5-6}$$

$$X_C = 159/fC \ (\Omega) \tag{2.5-7}$$

where f is in GHz, L in nH, and C in pF. Note that the *reactances* require a preceding j factor to become *impedances*. Thus, ωL is the reactance of an inductor, L, while $j\omega L$ is its impedance. Similarly, $1/\omega C$ is the reactance of a capacitor, while $-j/\omega C$ is its impedance.

Addition of Series Impedances

Practical circuits may have very complex interconnections. To analyze them, it is necessary to be able to combine multiple impedance and admittance values to find equivalent, overall values.

The total impedance of elements in series is obtained by adding their real and imaginary parts, respectively.

Referring to Figure 2.5-1, if

$$Z_1 = a + jb \quad \text{and} \quad Z_2 = c + jd \tag{2.5-8}$$

then,

$$Z_T = Z_1 + Z_2 = (a + c) + j(b + d) \tag{2.5-9}$$

For example, if

$$Z_1 = 5 + j9 \ \Omega \quad \text{and} \quad Z_2 = 3 - j15 \ \Omega$$
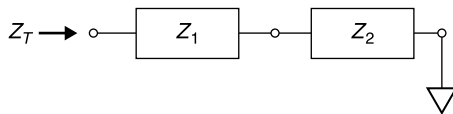
Figure 2.5-1   Series addition of impedances, $Z_T = Z_1 + Z_2$.

then

$$Z_1 + Z_2 = (5 + 3)\ \Omega + j(9 - 15)\ \Omega = 8 - j6\ \Omega$$

2.6   ADMITTANCE

Admittance Definition

The admittance Y is the complex reciprocal of impedance Z. Its unit is expressed either as the *siemen* or the *mho* (which is "ohm" spelled backward) and its symbol is the Greek capital omega written upside down (℧):

$$Y = 1/Z = G + jB \tag{2.6-1}$$

where *G* is the *conductance* and *B* is the *susceptance* of *Y*.

The *susceptance of an inductor* is

$$B_L = \frac{1}{X_L} = \frac{1}{\omega L} \tag{2.6-2}$$

and *the admittance of an inductor* is

$$-jB_L = \frac{1}{jX_L} = \frac{1}{j\omega L} = \frac{-j}{\omega L} \tag{2.6-3}$$

Similarly, the *susceptance of a capacitor* is

$$B_C = \frac{1}{X_C} = \frac{1}{1/\omega C} = \omega C \tag{2.6-4}$$

and the *admittance of a capacitor* is

$$jB_C = \frac{1}{-jX_C} = \frac{1}{-j/\omega C} = j\omega C \tag{2.6-5}$$

Addition of Parallel Admittances

The total admittance of elements in parallel is obtained by adding the real and imaginary parts of their admittances, respectively.

Frequently, one is given the impedances of elements that are in parallel. To determine the total admittance, find the equivalent admittances of each element and add them together. The reader is cautioned that one cannot form the admittance simply by adding together the reciprocals of the resistance and reactance parts of an impedance. Rather, it is necessary to find the complex
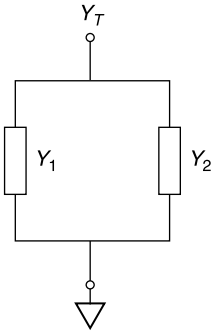
Figure 2.6-1 Parallel addition of admittances, $Y_T = Y_1 + Y_2$.

reciprocal of the impedance, which is the equivalent admittance. Thus, if the element values are expressed initially in impedance, first convert to admittance, then perform the addition of real and imaginary parts (Fig. 2.6-1).

$$Z_T = \frac{1}{Y_T} = \frac{1}{Y_1 + Y_2} \tag{2.6-6}$$

Ohm's law for AC circuits can also be written in terms of admittance,

$$I = VY \tag{2.6-7}$$

where, as before, V and I are phasor quantities and Y is admittance.

For example, suppose we wish to find the total equivalent impedance, Z_T, of a pair of elements in parallel given their individual impedances, Z_1 and Z_2 (Fig. 2.6-2).

Assume that $Z_1 = 5 + j8 \ \Omega$ and $Z_2 = 3 - j5 \ \Omega$. The first step is to convert these to polar form:

$$Z_1 = 9.43\angle 58° \ \Omega \quad \text{and} \quad Z_2 = 5.83\angle{-59.0°} \ \Omega$$

Then find their equivalent admitances (the complex reciprocals):

$$Y_1 = 1/Z_1 = 0.106\angle{-58°} \ \text{℧} \quad \text{and} \quad Y_2 = 1/Z_2 = 0.172\angle 59° \ \text{℧}$$

Next convert to rectangular form.
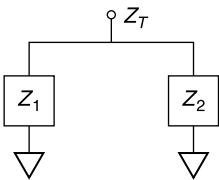
Figure 2.6-2 Parallel admittance combination example.

$$Y_1 = 0.056 - j0.090 \text{ ℧} \qquad Y_2 = 0.089 + j0.147 \text{ ℧}$$

Add the real and imaginary parts, respectively:

$$Y_T = Y_1 + Y_2 = 0.145 + j0.057 \text{ ℧}$$

To find Z_T, convert Y_T to polar form:

$$Z_T = 1/Y_T = 1/(0.156\angle 21.46°) \text{ Ω}$$
$$= 6.41\angle -21.46° \text{ Ω}$$
$$= 5.97 - j2.35 \text{ Ω}$$

The Product over the Sum

Generally, to combine impedances in parallel one must first convert them to admittances, add the admittances together to form the total admittance, then convert this admittance to an impedance value, as was just demonstrated. However, when only two parallel elements are to be combined (or two admittances in series) a short cut results by forming the *product over the sum. The total impedance of two parallel impedances equals the product of the individual impedances divided by their sum.*

The validity of this rule can be shown by performing the combination in general terms and observing the result:

$$Y_T = Y_1 + Y_2$$
$$\frac{1}{Z_T} = \frac{1}{Z_1} + \frac{1}{Z_2}$$
$$Z_1 Z_2 = Z_T(Z_1 + Z_2) \tag{2.6-8}$$
$$Z_T = \frac{Z_1 Z_2}{Z_1 + Z_2}$$

Using the values in the previous example, $Z_1 = 5 + j8$ Ω and $Z_2 = 3 - j5$ Ω,

$$Z_T = \frac{(9.43\angle 58°)(5.83\angle -59°)}{8 + j3} \text{ Ω}$$
$$= \frac{55\angle -1°}{8.54\angle 20.56°} = 6.44\angle -21.56° \text{ Ω}$$
$$= 5.99 - j2.37 \text{ Ω}$$

The same procedure applies for finding the total admittance of two admittances in series. Specifically, form their product and divide by their sum.

2.7   LLFPB NETWORKS

Up until now, the networks that we have described are composed of *LLFPB elements* to which AC analysis can be applied directly. The acronym for these networks derives from the following list of their properties. Namely, they

  L: are small compared to a wavelength (lumped).
  L: respond linearly to excitations (linear).
  F: are of finite value (finite).
  P: do not generate power (passive).
  B: have the same behavior for currents in either direction (bilateral).

Thus, they are called *lumped, linear, passive, finite, bilateral (LLFPB) elements*. We shall describe additional means to evaluate networks that do not satisfy all of these criteria. For example, circuits whose dimensions are large compared to a wavelength do not satisfy the lumped criterion. Circuits containing transistors are neither passive nor bilateral.

2.8   DECIBELS, dBW, AND dBm

Logarithms (Logs)

In wireless engineering, radio signals may be transmitted over miles of distance with substantial diminution of their signal strengths. Then they may be amplified by factors of thousands or millions of times so that their information can be heard in an audio speaker, viewed on a television screen, or employed in a digital processor. To handle the large signal ratios involved, it is easier to represent them as powers of 10 (logarithms). This measure employs the *decibel scale*. It is insightful to review some mathematical background when applying the *decibel* measures that will be described shortly.

   The logarithm of Y to the base X is the power L to which X must be raised to give Y. Thus,

$$Y = X^L \tag{2.8-1}$$

For example, if we choose base 10, then $X = 10$ and

$$Y = 10^L$$

Table 2.8-1 lists sample values of base 10 logarithms.

TABLE 2.8-1   Selected Values of *Y* and *L* for Base 10 Logarithms

Y	1000	100	20	10	2	1	1/2	1/10	1/20	0.01	0.001
Y	10^3	10^2	$10^{1.3}$	10^1	$10^{0.3}$	10^0	$10^{-0.3}$	10^{-1}	$10^{-1.3}$	10^{-2}	10^{-3}
L	3	2	1.3	1	0.3	0	−0.3	−1	−1.3	−2	−3

Multiplying by Adding Logs

To multiply numbers having the same base, write down the base and add their logs. For example,

$$10 \times 1000 = (10^1)(10^3) = 10^{1+3} = 10^4 = 10,000$$

and

$$100 \times 2 \times 5 = (10^2)(10^{0.3})(10^{0.7}) = 10^3 = 1000$$

Dividing by Subtracting Logs

To divide two numbers having the same base, write down the base and subtract the exponent of the denominator from the exponent of the numerator. For example, $10^3/10^2 = 10^{3-2} = 10^1 = 10$.

Zero Powers

Any number divided by itself, except zero, must equal unity. Regardless of the base, use of logarithms must produce this same result. As a consequence it follows that: *The zero power of any number, except zero, is 1.* For example,

$$10/10 = 10^{1-1} = 10^0 = 1$$
$$17/17 = 17^{1-1} = 17^0 = 1$$
$$459/459 = 459^{1-1} = 459^0 = 1$$

Bel Scale

The convenience of multiplying by adding logarithms as well as representing very large numbers by their logarithms first prompted the use of the *Bel scale* (named after its originator), which was simply the logarithm to the base 10 of the number. This proper name remains embedded in the present *decibel* nomenclature.

$$Y \text{ (in bels)} = \log_{10}(Y) \qquad\qquad (2.8\text{-}2)$$

Accordingly,

$$100 \text{ (in bels)} = 2 \text{ bel}$$
$$10 \text{ (in bels)} = 1 \text{ bel}$$
$$2 \text{ (in bels)} = 0.3 \text{ bel}$$
$$1 \text{ (in bels)} = 0 \text{ bel}$$

$$\tfrac{1}{2} \text{ (in bels)} = -0.3 \text{ bel}$$

$$\tfrac{1}{10} \text{ (in bels)} = -1 \text{ bel}$$

Decibel Scale

The Bel scale was quickly recognized as a useful innovation, but its steps were inconveniently large, an increase of 1 bel being a factor of 10. This objection was accommodated by a transition to the *decibel (dB) scale*.

Electrical engineers took a further step. Inherently, bels and decibels are simply means of expressing ratios between similar quantities. However, in electrical engineering, *decibels are defined to be the ratio of two values of electrical power:*

$$P/P_0 \text{ (in decibels)} = 10 \log(P/P_0) \qquad (2.8\text{-}3)$$

where, henceforth in this text, "log" shall represent the logarithm to the base 10 and "ln" shall represent the logarithm to the base e. Because power is proportional to the square of voltage (whether rms or peak), *the ratio of two voltages or currents referenced to the same impedance level* is defined as

$$V/V_0 \text{ (in decibels)} = 20 \log(V/V_0) \qquad (2.8\text{-}4)$$

Obviously, the decibel value is just 10 times the Bel scale. Thus, a factor of 10 in power is 10 dB, a factor of 2 is 3 dB (more precisely 3.01 dB), and the value of unity remains 0 dB.

In using decibels, one need only memorize the logarithms of numbers between 1 and 10. In fact, for most purposes, knowing the values corresponding to 0.5, 1, 2, 3, and 10 dB is usually sufficient for most estimating purposes (Table 2.8-2).

Often one can deduce the decibel value of a number from the decibel values known for other quantities. For example, since $2 \times 5 = 10$, and since the factor 2 corresponds to 3 dB, and the factor 10 is 10 dB, it follows that the factor 5 corresponds to 10 dB $-$ 3 dB $=$ 7 dB. Similarly, if one remembers that the factor 1.2 corresponds to about 0.8 dB, then the factor $1.2 \times 1.2 = 1.2^2 = 1.44$ corresponds to about 0.8 dB $+$ 0.8 dB $=$ 1.6 dB.

Decibels—Relative Measures

For the system example in Figure 2.8-1, using numeric multiplication we calculate

$$P_{\text{OUT}}/P_{\text{IN}} = 0.8 \times 20 \times 0.5 = 8$$

With decibel multiplication (adding logs), the same result is obtained as

TABLE 2.8-2   Selected Logarithms and Decibel Values

P/P_0	V/V_0	$\log(P/P_0)$	dB[a]
0.01	0.1	−2	−20
0.1	0.316	−1	−10
0.5	0.707	−0.3	−3
1	1	0	0
1.05	1.025	0.021	0.21
1.1	1.05	0.041	0.41
1.12	1.06	0.05	0.5
1.2	1.10	0.08	0.8
1.26	1.12	0.10	1
1.58	1.26	0.2	2
2	1.414	0.3	3
2.51	1.58	0.4	4
3.16	1.78	0.5	5
4	2	0.6	6
5	2.24	0.7	7
6.3	2.5	0.8	8
8	2.82	0.9	9
10	3.16	1	10

[a] dB = $10 \log(P/P_0)$ and $20 \log(V/V_0)$.

follows:

$$P_{OUT}/P_{IN} = -1 \text{ dB} + 13 \text{ dB} - 3 \text{ dB} = 9 \text{ dB}$$

$$P_{OUT}/P_{IN} = 9 \text{ dB}$$

$$P_{OUT}/P_{IN} = \text{antilog}(0.9) = 8$$

Due to rounding off the values of the logarithms, we may obtain a slight difference in the answers by the numeric and decibel calculation methods. However, sufficient accuracy always can be obtained by using more decimal places for the log values, but for most planning purposes carrying values to within 0.1 dB provides enough accuracy for estimating.

   Decibels are dimensionless because each is proportional to the logarithm of the ratio of two numbers. When decibels are used with dimensioned quantities, such as watts or milliwatts, *both quantities in the ratio must have the same units*.
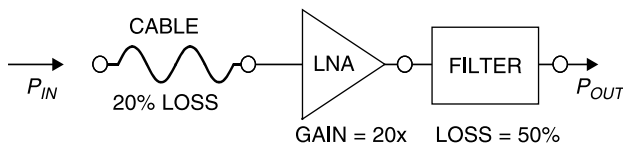
Figure 2.8-1   Subsystem design calculation using numeric and decibel methods.

Thus, for the gain of a power amplifier we might write:

$$\text{Gain} = P_{\text{OUT}} \text{ (watts)}/P_{\text{IN}} \text{ (watts)}$$

$$= 8 \text{ W}/1 \text{ W}$$

$$= 8$$

$$\text{Gain (dB)} = 10 \log(8) = 9 \text{ dB}$$

Absolute Power Levels—dBm and dBW

Because the decibel scale is so useful, it has become common practice to use it to represent *absolute power levels*. This is done by the simple step of referencing to a standard power level, either 1 W (dBW) or 1 mW (dBm) (Table 2.8-3).

These absolute power references are defined as:

$$\text{dBW} = 10 \log[P \text{ (watts)}/1 \text{ W}] \qquad \text{(referenced to 1 W)} \qquad (2.8\text{-}5)$$

$$\text{dBm} = 10 \log[P \text{ (watts)}/(0.001 \text{ W})] \qquad \text{(referenced to 1 mW)} \quad (2.8\text{-}6)$$

$$\text{dBm} = 10 \log[P \text{ (mW)}/(1 \text{ mW})]$$

Thus, 10 W (10,000 mW) can be expressed as +10 dBW or +40 dBm. One-tenth of a watt can be expressed as −10 dBW or +20 dBm.

TABLE 2.8-3   Power Levels in dBW and dBm

dBm	dBW	Power	Power (W)
−120	−150	1 fW	10^{-15}
−90	−120	1 pW	10^{-12}
−60	−90	1 nW	10^{-9}
−30	−60	1 μW	10^{-6}
−3	−33	0.5 mW	0.5×10^{-3}
0	−30	1 mW	10^{-3}
+10	−20	10 mW	10^{-2}
+20	−10	100 mW	10^{-1}
+30	0	1 W	1
+33	+3	2 W	2
+37	+7	5 W	5
+40	+10	10 W	10
+50	+20	100 W	10^2
+60	+30	1 kW	10^3
+70	+40	10 kW	10^4
+80	+50	100 kW	10^5
+90	+60	1 MW	10^6
+100	+70	10 MW	10^7
+110	+80	100 MW	10^8
+120	+90	1 GW	10^9

Customarily the sign is stated explicitly in dBW and dBm specifications to minimize the chance of misinterpretation. For example, one would say that 2 W is "plus 3 dBW."

Decibel Power Scales

Notice that this enormous power range can be represented in decibel notation simply as 240 dB, but to represent it numerically the ratio of the highest to the lowest power in the table would be expressed as the ratio 1,000,000,000,000,000,000,000,000 to 1!

2.9   POWER TRANSFER

Calculating Power Transfer

We saw in (2.1-2) that the instantaneous flow of power is given by

$$P_{\text{instantaneous}} = vi \qquad (2.9\text{-}1)$$

where v and i are the instantaneous voltage and current. However, this power flow may represent the power that is dissipated, often considered the *real power flow*, and power that flows to store energy in an inductor or capacitor, which we call the *imaginary power*. Real power flow is dissipated or, if v and i apply to antenna terminals, is radiated into space. The imaginary power flow in an AC circuit flows back and forth as the inductors and capacitors cycle from peak-to-zero energy storage conditions. When applied to an antenna, imaginary power flow goes into energy storage in the near fields of the antenna.

Usually we are more interested in real power flow in a system design. In the AC case, the real power is given by (2.1-3). This is rewritten below as the peak and average power flow that will be understood to mean dissipated or radiated power in the remainder of this text. Thus

$$P_{\text{Peak}} = VI \cos \theta \qquad (2.9\text{-}2)$$

$$P_{\text{Av}} = \tfrac{1}{2} VI \cos \theta \qquad (2.9\text{-}3)$$

The expressions can be used throughout this text, since voltage, current, and field values usually are specified in terms of their AC zero-to-peak values as, for example, an AC voltage is written $v(t) = V_0 \sin \omega t$.

It is useful to be able to express the real power values directly in terms of the complex phasors related to them without having to determine θ explicitly. To do this consider voltage and current phasors V and I shown in Figure 2.9-1.

To find the peak real power transfer, we wish to calculate

$$P_{\text{Peak}} = |V| |I| \cos \theta \qquad (2.9\text{-}4)$$
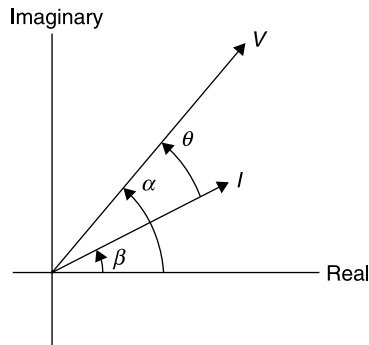
Figure 2.9-1   Voltage and current phasors used for power calculation.

where $\theta = \alpha - \beta$. By taking the complex product of V and the complex conjugate of I, I^*, the subtraction of the angles $\alpha - \beta$ produces the difference θ. Then taking the real part of the complex product yields the necessary multiplication by $\cos \theta$. Thus,

$$P_{\text{Peak}} = \text{Re}(VI^*) = |V|\,|I|\cos\theta \qquad (2.9\text{-}5)$$

$$P_{\text{Av}} = \tfrac{1}{2}\,\text{Re}(VI^*) = \tfrac{1}{2}|V|\,|I|\cos\theta \qquad (2.9\text{-}6)$$

These expressions, and their extensions to $\vec{E}$ and $\vec{H}$ fields, will prove very useful in deriving expressions for propagating power in waveguides and free space.

Maximum Power Transfer

The schematic diagram in Figure 2.9-2 represents an AC voltage source with internal impedance Z_G connected to a load impedance Z_L. In general, both Z_G and Z_L can be complex (each can have nonzero real and imaginary parts).

To determine the relationship between generator and load impedances for maximum power transfer, we first note that the maximum power will be delivered to the load when $jX_G = -jX_L$, since any net reactive impedance in the
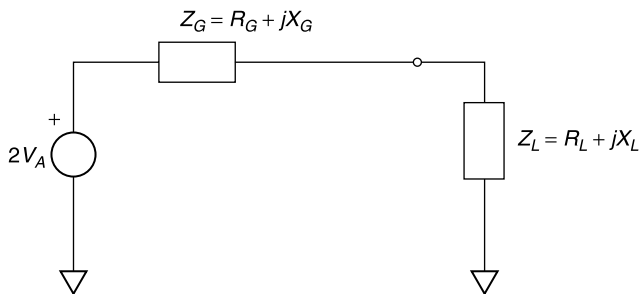
Figure 2.9-2   Maximum power transfer from generator to load.

loop would reduce the magnitude of I and with it the power delivered to the load.

The peak power (since, in our convention, V_A is a peak voltage amplitude), P_L, delivered to the load is then

$$P_L = I^2 R_L = \frac{(2V_A)^2}{(R_G + R_L)^2} R_L \tag{2.9-7}$$

and therefore

$$\frac{\partial P_L}{\partial R_L} = 4V_A^2 \frac{(R_G + R_L)^2 - R_L(2R_G + 2R_L)}{(R_G + R_L)^4} \tag{2.9-8}$$

Now set the numerator equal to zero to establish the condition for the maximum value of P_L (the zero slope of P_L versus R_L):

$$0 = R_G^2 + 2R_G R_L + R_L^2 - 2R_G R_L - 2R_L^2$$
$$R_G^2 = R_L^2 \qquad R_G = R_L \tag{2.9-9}$$

Thus, the maximum power transfer occurs when the load impedance is set equal to the complex conjugate of the generator impedance. That is,

$$Z_L = Z_G^* \tag{2.9-10}$$

The maximum peak power available to be delivered to the load is

$$P_A = V_A^2 / R_G \tag{2.9-11}$$

Usually, at microwave frequencies a generator with real impedance, Z_0, is assumed. Then the maximum available peak power becomes

$$P_A = V_A^2 / Z_0 \tag{2.9-12}$$

and this will be delivered to the Z_0 load. The symbol V_A is called the *available voltage* from the generator, not because it is the largest voltage available to the load (which would occur if Z_0 is replaced with an open circuit) but because *V_A is the voltage available to the load under the condition of maximum power transfer*.

2.10   SPECIFYING LOSS

Insertion Loss

We have seen in the preceding section that maximum power transfer occurs between generator and load when their impedances are the complex conjugates of one another. In practice, one rarely knows the equivalent impedance of a
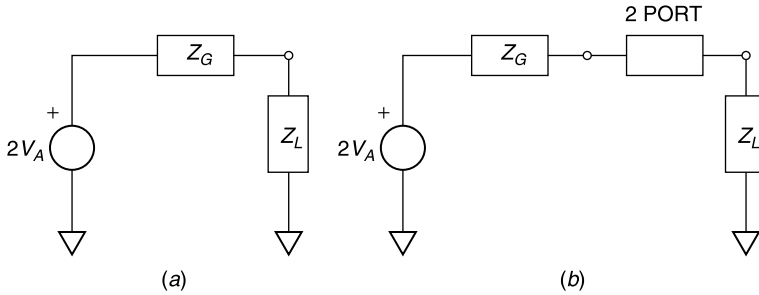
Figure 2.10-1 Insertion loss measurement.

microwave test source. Yet one of the most common measurements is that of *insertion loss*, consisting of the following procedure.

A generator with source impedance Z_G is connected to a load Z_L (Fig. 2.10-1a), and the power delivered to the load is found to be P_{L1}. Next, a two-port network is interposed between generator and load (Fig. 2.10-1b) and the power delivered to the load found to be P_{L2}. The insertion loss (IL) is defined as

$$\mathrm{IL} = \frac{P_{L2}}{P_{L1}}$$

$$\mathrm{IL} = 10 \log \frac{P_{L2}}{P_{L1}} \quad \text{(in decibels)}$$

(2.10-1)

The reader can see the problem with this procedure. The value of IL is strongly dependent upon the values of Z_G and Z_L. Without knowing these values the effect of inserting a two-port network cannot be predicted accurately.

If the two port is a passive network, one might expect that the insertion loss could range from 0 dB (no loss) to some finite loss (a positive decibel value). But this is not necessarily so. For example, if the load is 25 Ω and the generator impedance is 50 Ω, installing a two port that is a low-loss transformer might increase the power delivered to the load, resulting in a negative decibel loss value, or power gain, and this would be obtained with a passive two-port network.

Transducer Loss

The prospect of "gain" and the lack of defined source and load impedances of the insertion loss method led to the specification of an alternate measurement called *transducer loss* (TL). It is defined as

$$\mathrm{TL} = \frac{P_A}{P_L}$$

$$\mathrm{TL} = 10 \log \frac{P_A}{P_L} \quad \text{(in decibels)}$$

(2.10-2)

where P_A is the available power from the generator. Since the maximum power that can be delivered to the load, with or without a passive two port, is P_A, the transducer loss can never be less than unity (always a positive value when expressed in decibels).

   When the generator and load impedances are complex conjugates of each other, the insertion loss and the transducer loss are equal. Commonly, the generator and load are matched to the Z_0 of the test cables interconnecting them, satisfying the complex conjugate requirement.

   Within the industry, it is customary to use the insertion loss measurement. This is because of the ease of making a substitution measurement, first attaching generator and load using a "thru" connection to obtain P_{L1} and then substituting the two port to obtain P_{L2}. Fortunately, when performed with a well-padded generator (a matched, high attenuation pad at the output of the generator cable) and a well-matched power detecting load, the effect of this common practice, in effect, is actually to perform a proper transducer loss measurement, since generator and load essentially are matched to a common Z_0. Similarly, when performing "insertion loss" measurements with a network analyzer, the source and load are usually resistive and equal to the Z_0 of the test system, ensuring that a transducer loss measurement is also made. In practice, however, these techniques are commonly referred to as "insertion loss measurements."

Loss Due to a Series Impedance

The transducer loss caused by a series impedance Z placed between matched generator and load is a commonly encountered condition, often in switching or attenuating circuits (Fig. 2.10-2).

   The transducer loss may be termed *isolation* when Z is made large enough to block most of the power from reaching the load, as in the "off" state of a switch. The transducer loss (or isolation), as a function of Z and Z_0 is calculated as follows. Before Z is installed, under the matched generator and load conditions, the voltage at the load is the available or *line voltage* V_A and the current through the load, called *line current*, is $I_A = V_A/Z_0$. Then, with the series Z in place, the current through the load is reduced. The power delivered
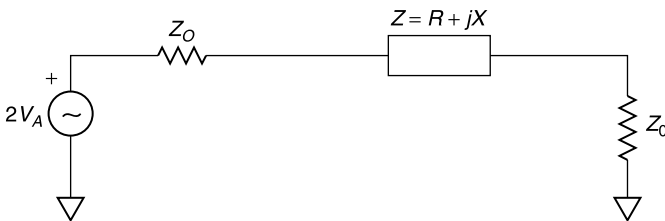
Figure 2.10-2   Equivalent circuit of impedance installed between matched generator and load.

to the load is proportional to the square of the load current. Thus,

$$\text{TL} = \frac{P_A}{P_L} = \frac{V_A^2/Z_0}{I^2 Z_0} = \frac{V_A^2/Z_0}{[(2V_A)^2/|2Z_0 + Z|^2]Z_0} = \frac{|2Z_0 + Z|^2}{(2Z_0)^2}$$

$$= \left| 1 + \frac{Z}{2Z_0} \right|^2 \tag{2.10-3}$$

Since Z/Z_0 is the *normalized impedance*, $z = r + jx$, the transducer loss (or isolation) can be written

$$\text{TL} = \left| 1 + \frac{z}{2} \right|^2 = 1 + r + \frac{r^2}{4} + \frac{|x|^2}{4} \tag{2.10-4}$$

For example, if 5 Ω resistance is placed between a 50-Ω source and a 50-Ω load, the transducer loss is

$$\text{TL} = 1 + \tfrac{5}{50} + \left(\tfrac{1}{4} \right) \left(\tfrac{5}{50} \right)^2 = 1.10 = 0.42 \text{ dB}$$

This might represent the "on state" (forward bias) of a series PIN diode installed in the transmission line to provide switching. When the diode's bias state is switched (to reverse bias), it might be represented as 0.2 pF capacitance in the "off state." The transducer isolation at, say, 500 MHz, for which the diode's reactance is 1590 Ω, is

$$\text{TL} = 1 + \left(\tfrac{1}{4} \right) \left(\tfrac{1590}{50} \right)^2 = 254 = 24 \text{ dB}$$

Loss Due to a Shunt Admittance

Another common transducer loss example is that of a shunt-connected admittance between matched generator and load (Fig. 2.10-3). An analysis similar to that used for the series Z gives

$$\text{TL} = \left| 1 + \frac{y}{2} \right|^2 = 1 + g + \frac{g^2}{4} + \frac{|b|^2}{4} \tag{2.10-5}$$
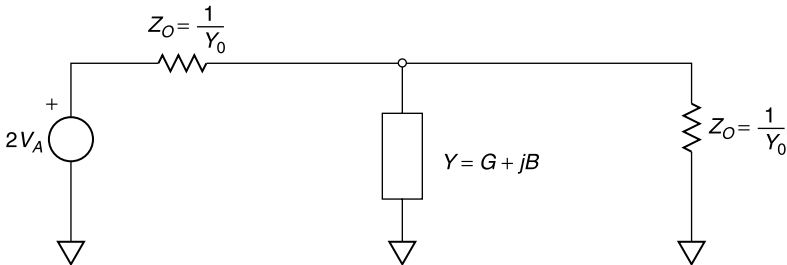
Figure 2.10-3   Equivalent circuit used with admittance transducer loss formula.

in which the normalized admittance is

$$y = \frac{Y}{Y_0} = \frac{G + jB}{Y_0} = g + jb \qquad (2.10\text{-}6)$$

Loss in Terms of Scattering Parameters

When using a circuit simulator, transducer loss is obtained by calculating the square of the magnitude of the scattering parameter, S_{21}, since for passive networks $|S_{21}| \leq 1$ the transducer loss (or isolation) as a function of S_{21} is

$$\mathrm{TL} = |S_{21}|^{-2}$$
$$= -20 \log|S_{21}| \quad \text{(in decibels)} \qquad (2.10\text{-}7)$$

where the negative signs has been added so that the loss ratio is equal to or greater than unity and a positive number when expressed in decibels. The S parameters are described in Chapter 5.

2.11   REAL RLC MODELS

Resistor with Parasitics

Real resistors have series inductance L and shunt capacitance C. These unintended elements are called *parasitic elements* or simply *parasitics*. For a standard carbon resistor these elements (typically $L = 5$ nH and $C = 1$ pF) have negligible effects at 10 MHz, but at 1 GHz their presence causes the resistance value to be changed profoundly. Figure 2.11-1 shows their effects on a 50-Ω resistor with frequency. At resonance, a little over 1.5 GHz, the resistance has doubled in value.

Inductor with Parasitics

A real inductor L has some series resistance R and some shunt capacitance C. Consider a 20-nH inductor with $Q = 60$ at 500 MHz ($R = 1\ \Omega$). Also assume a shunt capacitance of 0.5 pF. The behavior with frequency is as shown in Figure 2.11-2. The inductor goes through parallel resonance, f_R, at about 1.6 GHz. It approximates a linear inductor only up to about 600 MHz (about 40% of f_R). Above f_R its net reactance actually is capacitive!

Capacitor with Parasitics

Similarly, capacitors also have parasitic inductance, capacitance, and resistance. For the capacitor, a four-element model is required to provide an accu-
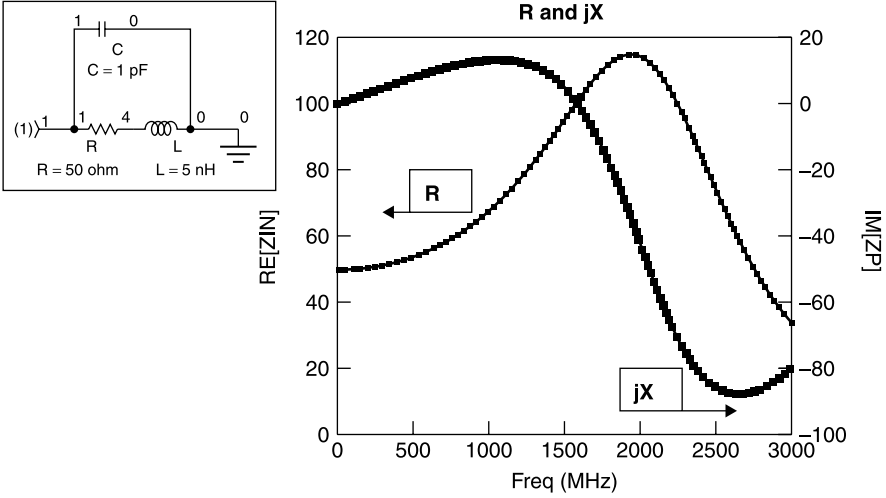
Figure 2.11-1   Frequency behavior of 50-Ω resistor with 5-nH-series-inductance and 1-pF-parallel-capacitance parasitics.
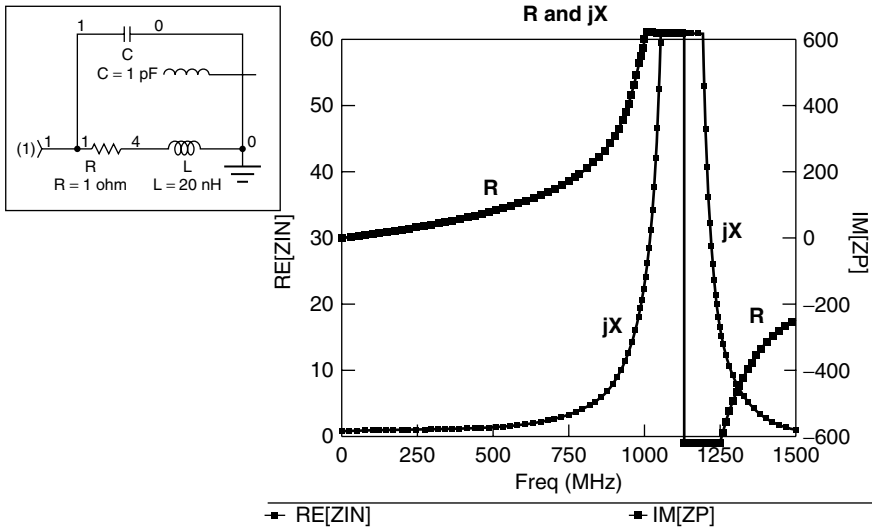
Figure 2.11-2   Frequency behavior of 20-nH inductor having parasitic shunt capacitance of 1 pF.
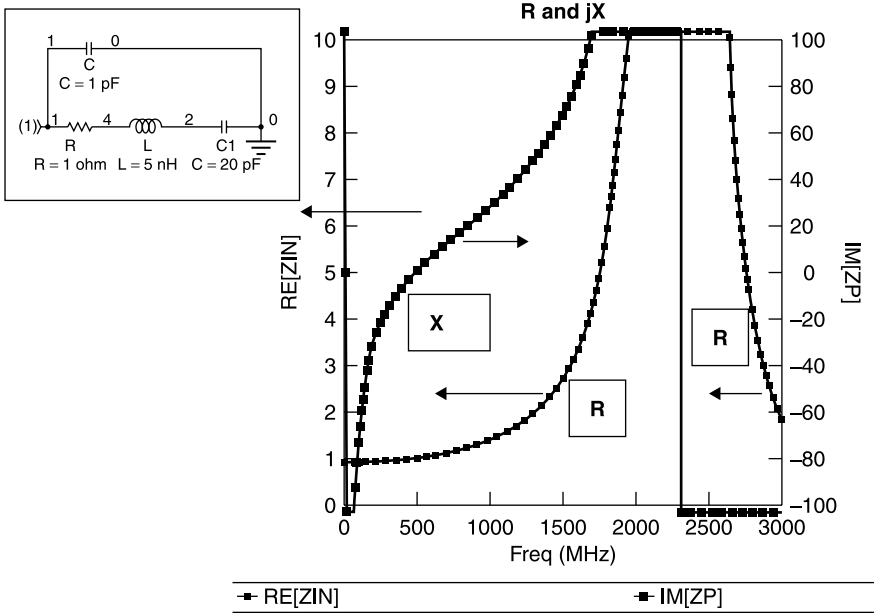
Figure 2.11-3 Reactance and impedance magnitude of 20-pF capacitor with 5-nH-series inductance and 1-pF parallel capacitance parasitics.

rate simulation over a wide frequency range (Fig. 2.11-3). Notice that the circuit goes through series resonance ($jX = 0$) at about 500 MHz and through parallel resonance ($jX = \infty$) at about 2.3 GHz.

2.12   DESIGNING *LC* ELEMENTS

Lumped Coils

The inductance of a single-layer, air core coil (Fig. 2.12-1), which can be wound on a cardboard or wood cylindrical mandrel, is given by Wheeler [2]:

$$L = n^2 \left[\frac{r^2}{9r + 10l} \right] \quad (\mu H) \tag{2.12-1}$$

where n is the number of turns, $r = D/2$ is the radius of the coil, and l is its length, both in inches. In this formula the wire thickness from which the coil is formed is assumed to be small compared to the turns' spacing.

For example, suppose that it is desired to make an RF choke for an integrated circuit module. There is enough space for a 150-mil-long coil having 10 turns of 36-gauge wire (5 mil diameter) about a 25-mil-diameter form. What
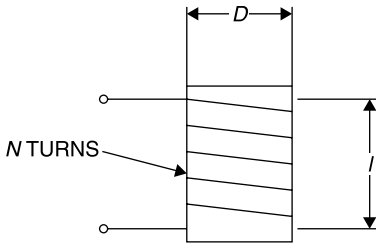
Figure 2.12-1 An "air core" coil wound on a nonmetallic form.

is L?

$$L = (10)^2 \left[\frac{0.0125^2}{9(0.0125) + 10(0.150)} \right] = 0.0097 \ \mu H = 9.7 \ nH$$

Highest Q generally results when the coil diameter $D = 2r$ is approximately equal to its length. Then (2.12-1) can be written

$$L = \frac{n^2 D}{58} \quad (\mu H) \tag{2.12-2}$$

Solving for n,

$$n = \sqrt{\frac{58L}{D}} \tag{2.12-3}$$

For example, suppose we wish to make a high-Q RF choke for use in a radio transmitter having 200 nH (0.20 μH) of inductance and wound as a $\frac{3}{8}$-in.-diameter coil on a nonmagnetic form. How many turns are required if we choose $l = D$?

$$n = \sqrt{\frac{(58)(0.2)}{0.375}} = 5.56 \ \text{turns}$$

High μ Inductor Cores—the Hysteresis Curve

If a coil or transformer is wound on a high-permeability material core ($\mu_R \gg 1$), such as ferrite material, the magnetic spins of the atoms of the material can be aligned with the applied H field, yielding a much higher flux density, B. This can produce a much higher inductance for a given coil size.

   However, the effect is complicated by the factors illustrated in Figure 2.12-2. Starting with an initially unmagnetized core, we experience the behavior that originates at the $\vec{B}$–$\vec{H}$ axes' origin. The $\vec{B}$ increases very rapidly, having a slope of $\mu_{EFF} = \mu_0 \mu_R$. Then, as all of the core's atoms align with the applied $\vec{H}$ field, saturation occurs at $\vec{B}_{SAT}$ and thereafter $\vec{B}$ increases with applied $\vec{H}$ at μ_0 slope.

   When the $\vec{H}$ field is removed, many of the core's atoms remain aligned, leaving $\vec{B}_{REM}$. As $\vec{H}$ is cycled positively and negatively, a *hysteresis curve* (a
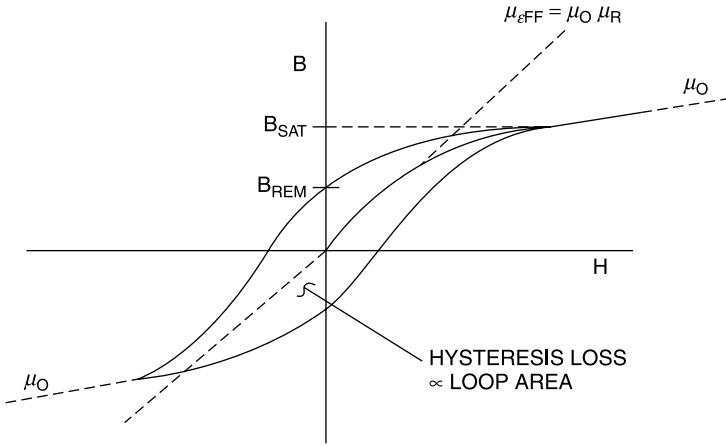
Figure 2.12-2   Use of magnetic core material can greatly increase flux density for given magnetic field ($\vec{B} = \mu_0\mu_R\vec{H}$).

curve having memory) is traced. The area within the curve is proportional to the AC losses when the coil is sinusoidally excited. Suppliers of inductors having high-permeability cores specify a maximum current for the coil's use. This is not an indication of the coil's power-handling capability, but rather the maximum current for which the coil provides its small signal inductance (operates below B_{SAT}). The hysteresis curve is also a function of temperature, and the coil's specification includes an indication of the highest temperature at which the coil's inductance specification is met. Above a critical temperature T_C the spins of the material's atoms do not align with the applied H field, and the effective permeability is only μ_0.

Estimating Wire Inductance

There are numerous formulas given for the inductance of an isolated straight wire. These are approximations, having an implicit assumption about the nature of the return current path (Fig. 2.12-3). *An inductance determination requires a closed path* just as the current in the coil requires a closed path. The inductance of a truly isolated straight wire is as undefined as the sound of one hand clapping! An exact relation (7.25-17) exists for the high frequency inductance per unit length of a coaxial cable. The high frequency approximation assumes that the $\vec{E}$ and $\vec{H}$ fields are contained between the conductors and do not penetrate the conductors appreciably.

$$L = \frac{\mu_0\mu_R}{2\pi} \ln\frac{b}{a} \ \text{H/m} \tag{2.12-4}$$

$$L = 5.08 \ln\frac{b}{a} \ \text{nH/in.} \tag{2.12-5}$$
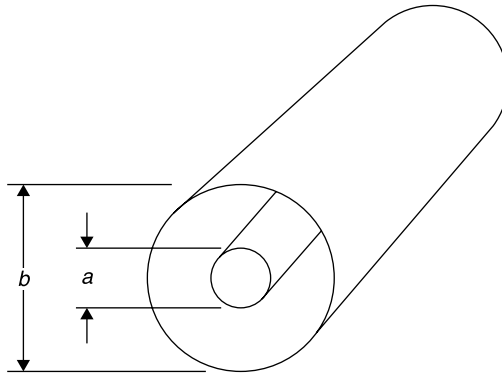
Figure 2.12-3   Inductance of isolated wire can be approached by that of coaxial line whose outer conductor is made very large.

where $\mu_R = 1$ for all nonmagnetic insulators. If we assume b is very large compared to a, then the calculated value for L approximates that of a straight wire with far removed return path.

For example, if $b/a = 7$

$$L \approx 10 \text{ nH/in.} \tag{2.12-6}$$

This is a reasonable approximation for the inductance of thin wires elevated above ground planes in microstrip circuits.

Note, however, that if we assume a value of $b = 50a$, the value for $\ln(b/a)$ doubles, and the inductance becomes 20 nH/in., also a value in the vicinity of those measured for real circuits in which the thin wire is well above the ground plane. The point to observe from the formula for the inductance of coaxial line is that the *inductance continues to increase as the outer conductor diameter is increased*, that is, as the return path for the inductive current is further removed from the center conductor "wire." It is a common but incorrect practice *to speak of the inductance of an "isolated" straight wire*. The better procedure is to approximate the inductance of a wire by the known inductance per unit length of an appropriately similar transmission line.

Parallel Plate Capacitors

The capacitance C of a parallel plate capacitor (Fig. 2.12-4) having plate area A separation d, and relative dielectric constant ε_R is given by

$$C = \varepsilon_0 \varepsilon_R \frac{A}{d} \tag{2.12-7}$$
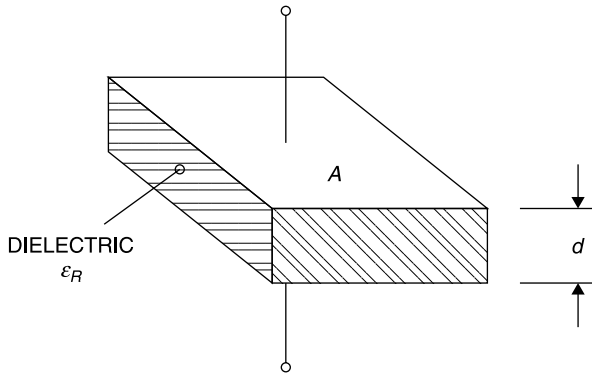
Figure 2.12-4   Parallel plate capacitor.

where

$$\varepsilon_0 = \frac{10^{-9}}{36\pi} \text{ F/m} \qquad (2.12\text{-}8)$$

is the permittivity of free space. These expressions neglect the E field fringing at the edges of the capacitor. With fringing the capacitance is larger depending on the plate area-to-separation aspect ratio.

For engineering work, a more convenient format to estimate the parallel plate capacitance is [3, Appendix A]

$$C = \varepsilon_0 \varepsilon_R \frac{A}{d}$$

$$= \varepsilon_R (0.0885) \frac{A}{d} \text{ pF/cm}$$

$$= \varepsilon_R (0.225) \frac{A}{d} \text{ pF/in.} \qquad (2.12\text{-}9a,b,c)$$

For example, suppose that we wish to construct an RF blocking capacitor. We have space for a 0.5-in.-diameter capacitor and will use 1-mil Mylar tape having a relative dielectric constant of 3 [3, Appendix B] as the plate separation dielectric. What will be the value of C?

$$C = (3)(0.225) \frac{(3.14)(0.25^2)}{0.001} = 132 \text{ pF}$$

This capacitor would have a reactance of only 1.2 Ω at 1 GHz, a reasonably small impedance in a 50 Ω system.

2.13 SKIN EFFECT

An important consideration in the design of circuits is that the effective resistance of conductors increases considerably at high frequencies. This is due to *skin effect*, which is a result of the crowding of current densities near conductor surfaces. This is analyzed in Section 7.18.

At any point, current flows in accordance with the incremental (or "point") form of Ohm's law, namely

$$\vec{J} = \sigma\vec{E} \tag{2.13-1}$$

where $\vec{J}$ (A/m^2) is the current density, $\vec{E}$ (V/m) is the electric field at the point, and σ (℧/m) is the conductivity of the material. This equation applies for any time variations. At high frequencies (sinusoidal excitations) Maxwell's equations play a special role in analyzing the effects in a conductor. The sinusoidally varying electric field $\vec{E}$ produces a sinusoidally varying current density $\vec{J}$, which in turn produces a sinusoidally varying $\vec{H}$ field, which in turn produces an opposing electric field. More about Maxwell's equations in Chapter 7. The net result is that the $\vec{E}$ field in the conductor falls off rapidly with depth from the surface, and with it so does the current density [4, p. 234; 5, p. 43] (Fig. 2.13-1).

The current density in the conductor is

$$\vec{J} = \vec{J}_0 e^{-z/\delta_s} \tag{2.13-2}$$

where the *skin depth* is

$$\delta_S = \frac{1}{\sqrt{\pi f \mu \sigma}} \tag{2.13-3}$$
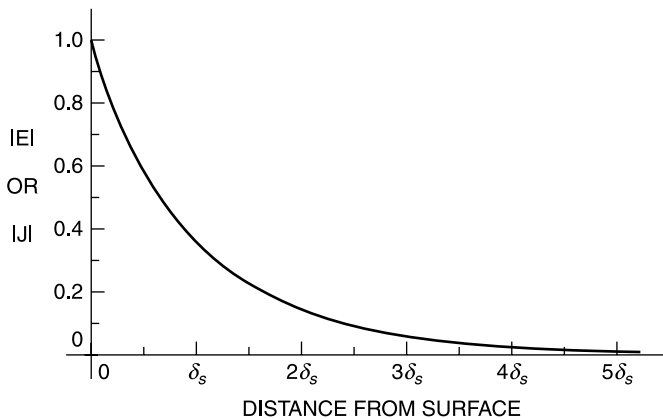
Figure 2.13-1   Fall-off of electric field $\vec{E}$ and current density $\vec{J}$ with distance from surface in skin depth units.

and f = frequency (Hz), $\mu = 4\pi \times 10^{-7}$ (H/m) for nonmagnetic materials, and $\sigma = 6.17 \times 10^7$ ℧/m for silver.

At a frequency of 1 GHz, the skin depth in silver is calculated as

$$\delta_S = \frac{1}{\sqrt{(3.14)(10^9 \text{ Hz})(4)(3.14)(10^{-7} \text{ H/m})(6.17)(10^7 \text{ ℧/m})}}$$

$$= 2.03 \times 10^{-6} \text{ m}$$

Thus, a silver conductor has a skin depth at 1 GHz of only 2 μm or about 0.1 mil! Since copper has about 90% the conductivity of silver, its skin depth is only 5% greater than that of silver, or about 2.1 μm at 1 GHz.

The term *skin depth* should not be misinterpreted. *Current does flow below the skin depth*, but its density at this depth is reduced to 1/e (about 37%) of its surface value. At 10 skin depths (about 1 mil in silver or copper) the current density would be reduced to 1/22,026 of its surface value!

For calculation purposes, it turns out that *when the skin depth is small, the AC resistance of a conductor has the same value that it would have if all the current were carried in a material depth equal to the skin depth* [5, p. 45].

For round wire the AC resistance, R_{AC}, is determined by the cross section of a hollow tube having a wire diameter D and a thickness equal to the skin depth, δ_S (Fig. 2.13-2). This can be compared with the direct current (DC) resistance R_{DC} for which the entire cross section of the wire is effective:

$$R_{AC} = \frac{l}{\pi D \delta_S \sigma} \tag{2.13-4}$$

$$R_{DC} = \frac{l}{\pi (D^2/4) \sigma} \tag{2.13-5}$$
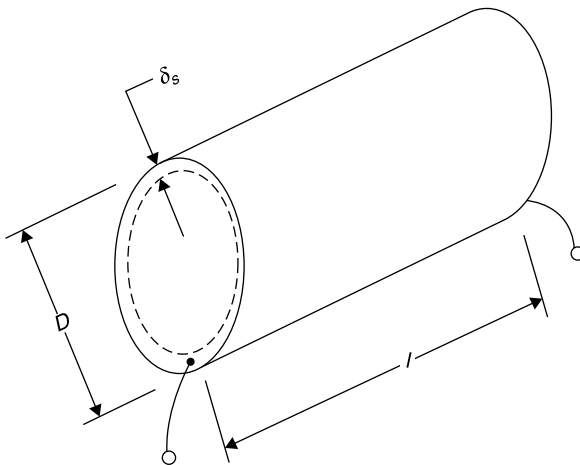
Figure 2.13-2   Skin depth region for solid round wire.

The ratio of the AC-to-DC resistance is

$$\frac{R_{\mathrm{AC}}}{R_{\mathrm{DC}}} = \frac{D}{4\delta_S} \tag{2.13-6}$$

Thus, a 30-gauge (10-mil-diameter) hookup wire used in a circuit at 1 GHz has

$$\frac{R_{\mathrm{AC}}}{R_{\mathrm{DC}}} = \frac{D}{4\delta_S} = \frac{(0.010 \text{ in.} \times 0.0254 \text{ m/in.})}{4(2 \times 10^{-6} \text{ m})} = 32$$

The DC resistance of this wire is only 0.1 Ω/ft, but its resistance at 1 GHz is 3.2 Ω/ft, a factor of 32 greater.

2.14   NETWORK SIMULATION

Using complex circuit analysis, it is possible to analyze any network to determine its performance as, for example, how much of an applied signal passes through the network to a specified load and what input impedance the network presents to a specified source. However, the application of complex mathematics is difficult for all but the simplest of networks, much more so if the analysis is to be conducted over numerous frequencies. Over the years, methods of simplifying the analysis were devised. These included filter theory (Chapter 9) and the Smith chart (Chapter 5). However, use of such techniques is generally limited to networks having specific circuit element values and relationships. For example, filter theory defines precisely the relationship of the circuit values to one another. The Smith chart is useful when the transmission line has uniform characteristic impedance.

Beginning in the 1960s computer programs became available for which a *variable topology* network could be specified and the network's performance calculated over a band of frequencies. These programs have been refined to a high degree, usually operate on a personal computer, and are generally termed *network simulators*. A sample network simulator screen is shown in Figure 2.14-1 using the Genesys software. The initial screen shown allows specification of a network. One selects components, such as resistors, inductors, capacitors, transmission lines, and so forth, from the toolbar above the workspace and *drags* them to the desired location in the schematic. Then *double clicking* on the component gives a submenu in which the component name and value can be specified. The sample network in Figure 2.14-1 was created in this manner. The program provides input and outputs whose impedances can be specified. These default to 50 Ω if no other values are specified.

Once the circuit is specified, selection of the *outputs* (a left menu option in Fig. 2.14-1) provides options for different kinds of outputs, such as rectangular or polar plots, a Smith chart format, an antenna plot, a three-dimensional plot, and so forth. It also allows specification of the functional output such as S_{11} or
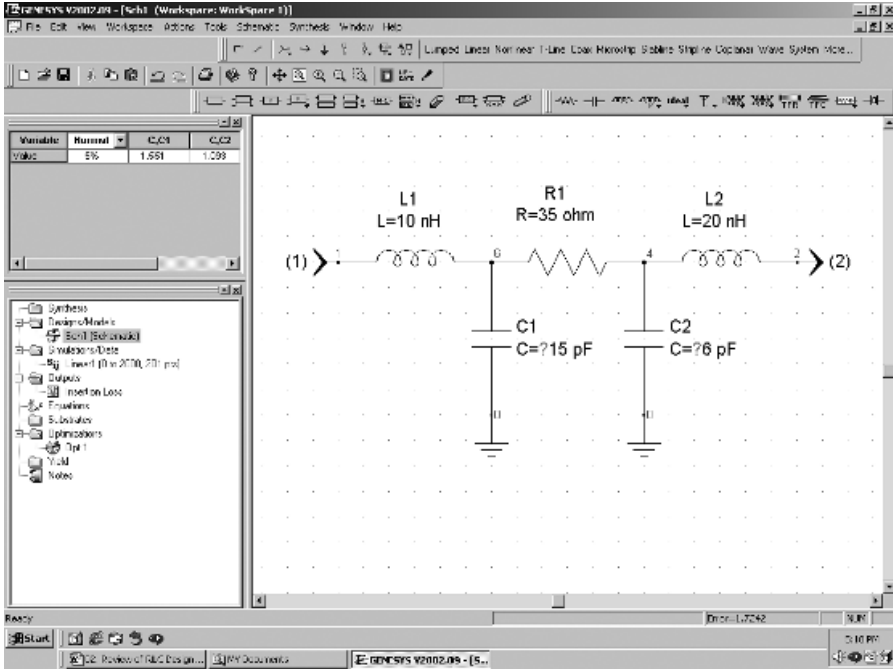
Figure 2.14-1 Inputting circuit to network simulator.

voltage standing wave ratio (VSWR) (to show reflection properties), the magnitude of S_{21} (for insertion loss) and phase of S_{21} to show transmission phase, and other functions. S parameters are described in Section 6.5. The insertion loss $|S_{21}|$ of the circuit in Figure 2.14-1 over the bandwidth 0 to 2 GHz is shown in Figure 2.14-2.

Even this simple graph would take hours or days to calculate manually. In this particular execution the insertion loss was calculated at 201 different frequencies, and each dot in the plot represents a separate calculation. However, the generation of the graph appeared almost instantaneously, since the personal computer on which it was performed can perform millions of calculations per second.

A *tuning* feature of modern network simulators causes the performance to be recalculated when the value of a given component is varied. This can be done manually or automatically. The automatic mode is termed *optimization*. For example, when the program varies the values of C1 and C2 and is directed to achieve a value of $|S_{21}| = 0$ at 1 GHz, the result is that shown in Figure 2.14-3. In this program, the variable nature of these components is denoted by the leading question mark before their values in Figure 2.14-1.

To achieve this optimization, the software varies each component whose value is allowed to be changed in a random manner and calculates the resulting
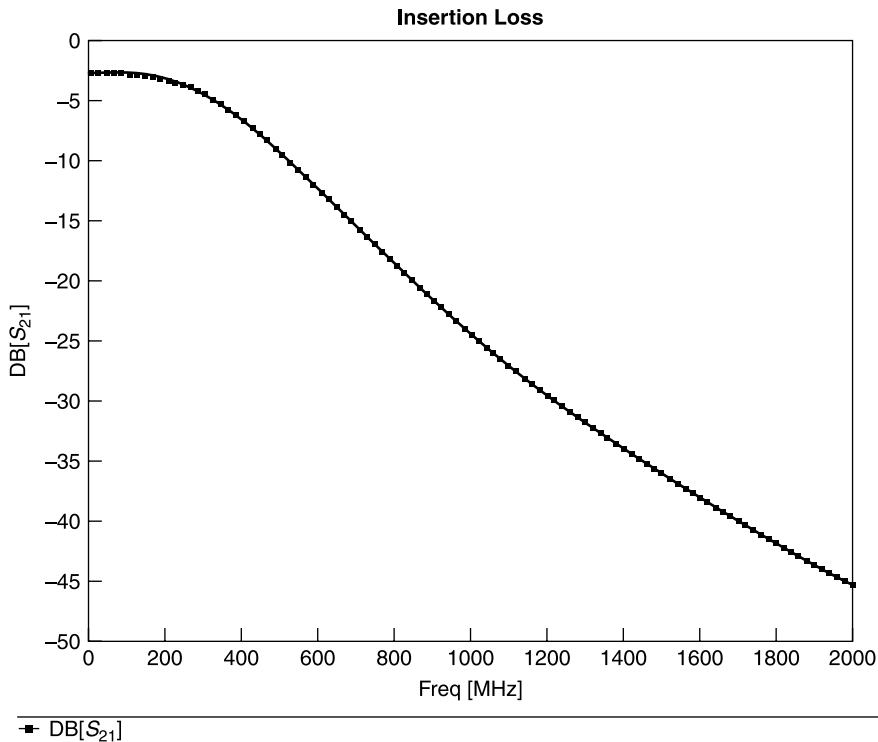
Figure 2.14-2   Insertion loss versus frequency for circuit shown in Figure 2.14-1.

performance, in this case the magnitude of S_{21}. If the change results in perfor-
mance closer to the desired value, the change is retained and another adjustable
component's value is changed. This process is repeated, usually millions of
times, until the desired result is obtained. In this example, the optimization was
terminated after a few seconds because zero insertion loss cannot be obtained
due to the finite value of R1, whose value was not allowed to vary. Usually,
such an optimization procedure would be wholly impractical to perform by
manual calculation methods.

Simulation software has become an indispensable engineering tool. How-
ever, the termination just described exemplifies the need for human interaction
with the computer. The computer does not appreciate that zero loss is un-
obtainable in a network having finite resistance, and the optimization routine
would have continued indefinitely if left unattended.

Another application of varying element values is to perform a *yield analysis*
of a circuit by which circuit element values are allowed to vary in accordance
with a given distribution, such as a Gaussian distribution. The program *builds* a
specified number of circuits to determine what the yield of circuits would be to
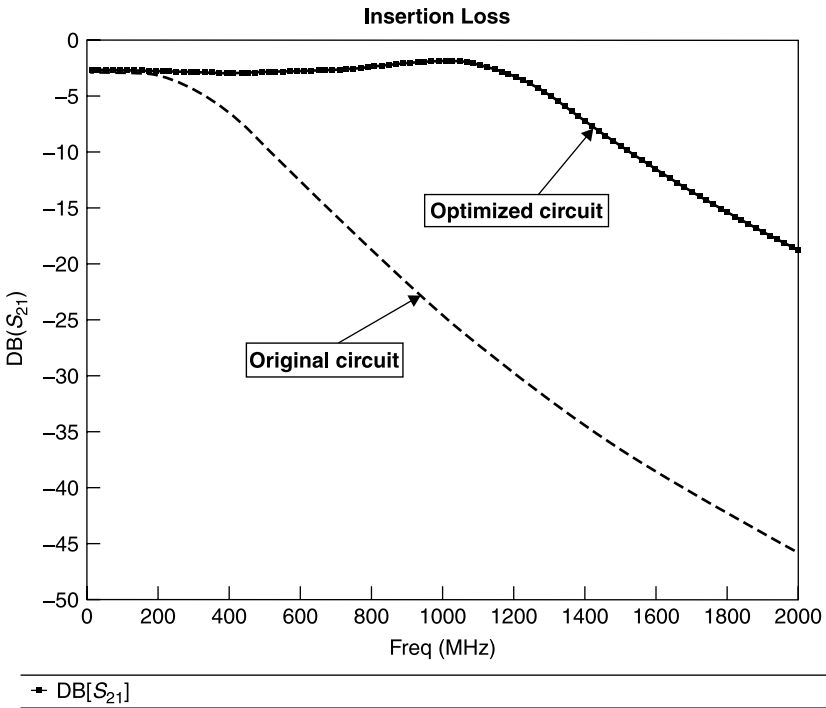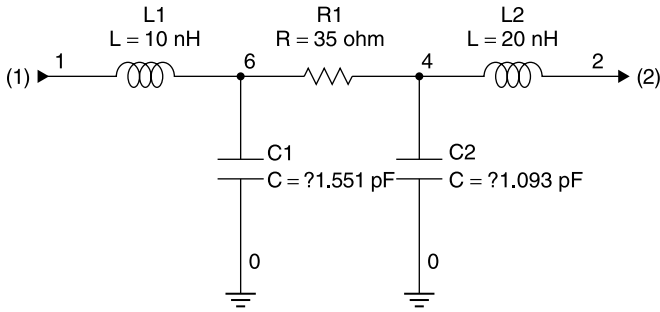
Figure 2.14-3   Result of optimizing values of C1 and C2 to minimize insertion loss at 1 GHz.

some required specification when the component parts have values with that statistical distribution. This analysis is demonstrated in Section 9.19.

For the most rigorous evaluation of networks *electromagnetic simulation* (called *EM simulation*) is performed, by which the electric and magnetic fields surrounding a circuit are computed and from them the circuit performance. This is demonstrated in Section 7.34.

The use of network simulation is especially important in transistor amplifier design. Most simulation programs contain a library of available transistors. The program can input a transistor's *S parameter table* and analyze the performance of a circuit containing the transistor (see Chapter 10 for examples). This permits stability and gain analyses at all frequencies over which the transistor has gain, a process that is essential to modern amplifier design.

These features of network simulation software have become essential to the conduct of efficient RF and microwave design. Each software package has its own format and features, but most simulation software available today performs at least the functions described. Throughout this text, network simulation is used to demonstrate the principles discussed, and network simulation exercises are provided to permit the reader to gain proficiency in these techniques.

REFERENCES

1. Kenneth S. Miller, *Advanced Complex Calculus*, Harper & Brothers, New York, 1960. *An old reference, much too advanced for our purposes, but its development of the bilinear transformation is fundamental to the Smith chart*.
2. H. A. Wheeler, "Inductance formulas for circular and square coils," *Proc. IEEE*, vol. 75, no. 2, 1982, pp. 256–257.
3. Joseph F. White, *Microwave Semiconductor Engineering*, Noble Publishing, Norcross GA 1995. *Extensive treatment of PIN diodes and their switching and phase shifting applications*.
4. Simon Ramo and John R. Whinnery, *Fields and Waves in Modern Radio*, Wiley, New York, 1944, 1953 (later revised with a third author, Van Duzer). *This is a classic introductory text describing fields and Maxwell's equations*.
5. Peter A. Rizzi, *Microwave Engineering, Passive Circuits*, Prentice-Hall, Englewood Cliffs, NJ, 1988. *Excellent microwave engineering textbook covering theory and design of transmission lines, couplers, filters, and numerous other passive devices*.

EXERCISES

E2.4-1   Four alternating currents at the same frequency have their reference directions into a common node. They are represented by:

$$i_1(t) = 10\cos(\omega t + 36.9°) \text{ A}$$

$$I_2 = 14.1\angle 315° \text{ A}$$

$$I_3 = 20\angle 126.9° \text{ A}$$

$$I_4 = 13.4\angle 243.4° \text{ A}$$

What is the total current (amplitude and phase) into the node?

E2.5-1   Realizing that you understand AC circuits, your friend asks for your help. He wishes to install a central air conditioner in his home. The unit requires $V = 220$ *V AC rms at 60 Hz* with a steady-state running current of $I = 50$ *A rms*. The maximum running current his wiring can provide is *40 A rms*. You note that the machine develops 10 horsepower (1 hp $= 746$ W) and that it is nearly 100% efficient.

   a. What is the power factor of the machine (*power factor* $= \cos \theta$, where θ is the phase angle between V and I)?

   b. Assuming the machine's motor presents an inductive load, what impedance Z does this machine present to the power line?

   c. What is the capacitance C of a capacitor connected in parallel with this machine that would make the power factor unity? What steady-state rms current must it sustain?

   d. What total steady-state, rms current would the machine draw with this capacitor installed in parallel?

E2.9-1   A 500-Ω load is to be connected to a 50-Ω transmission line to match load terminate it at 500 MHz. However, the hookup wire has 5 nH of series inductance. How can this be tuned such that the net reactance in series with the 50-Ω load is zero at 500 MHz?

E2.10-1 Show that the insertion loss (or isolation) of a normalized admittance, $y = g + jb$, mounted in shunt with a matched generator (Figure 2.10-3) and load of $Z_0 = 1$ is given by

$$\mathrm{IL} = \left| 1 + \frac{y}{2} \right|^2 = 1 + g + \frac{g^2}{4} + \frac{|b|^2}{4}$$

E2.12-1 It is proposed to realize the tuning capacitor of Exercise 2.9-1 as a pair of circular parallel plates separated by 1 mil of double-sided tape having a dielectric constant of 3. What should the diameter D of the plates be made to realize the required capacitance?

E2.12-2 A circuit designer requires a high Q small volume coil of $L = 50$ nH to be formed on a wood mandrel of 0.125-in. diameter using #30 gauge copper wire having a diameter of 0.010 in. (10 mil).

   a. How many turns of wire are required?

   b. What is the length of the wire required (not counting the ends used for hook-up)?

E2.13-1 Derive an expression for the Q of the coil in E2.12-2 in terms of its dimensions, wire conductivity, and frequency, taking skin effect into account. Then, using this expression, solve for the Q at 10, 100, and 1000 MHz.

LC Resonance and Matching Networks

3.1 *LC* RESONANCE

When the reactance magnitudes of L and C are equal, the pair resonates. At resonance the net reactance of a series-connected *LC* circuit is zero (a short circuit), and the net susceptance of a parallel-connected *LC* circuit is zero and the net reactance is infinite (an open circuit).

The resonant frequency is obtained by equating the magnitudes of the reactances of the *L* and *C* elements. Since their impedances are of opposite signs, the net reactance is zero for a series circuit or infinity for a parallel circuit, resulting in *resonance* at ω_0:

$$\omega_0 L = \frac{1}{\omega_0 C} \qquad (3.1\text{-}1)$$

$$\omega_0 = \frac{1}{\sqrt{LC}} \qquad (3.1\text{-}2)$$

$$\omega = 2\pi f \qquad (3.1\text{-}3)$$

$$f_0 = \frac{1}{2\pi\sqrt{LC}} \qquad (3.1\text{-}4)$$

The resonance solution can be obtained graphically. Given an 80-nH inductor, the reactance magnitude is 5 Ω at 10 MHz. Draw a line through this value having an upward slope and passing through successive 5's on each decade of log–log graph paper (Fig. 3.1-1). Next, suppose the capacitor to resonate when it has 0.32 pF, or a 5000-Ω-reactance magnitude at 100 MHz. Similarly, draw a line through this point having a downward slope, and passing through successive 5's. The intersection of the two lines defines the resonant frequency, seen to be 1 GHz for this example.
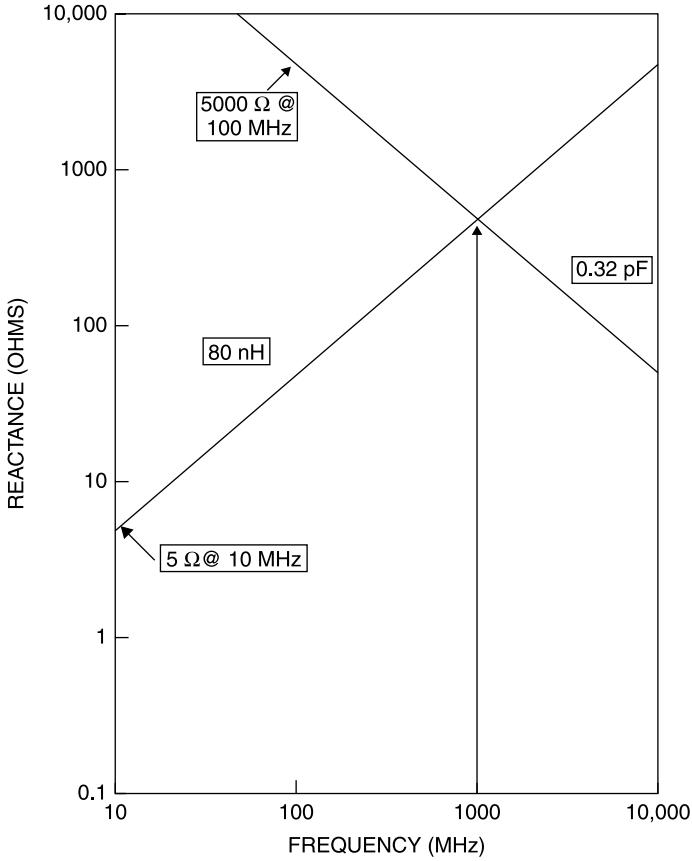
Figure 3.1-1   *L* and *C* reactance chart for determining resonance.

3.2   SERIES CIRCUIT QUALITY FACTORS

Q of Inductors and Capacitors

The *quality factor Q* of an inductor or capacitor is the ratio of its reactance magnitude to its resistance (Fig. 3.2-1). For an inductor,

$$Q = \frac{X_L}{R_L} = \frac{\omega L}{R_L} \tag{3.2-1}$$

and for a capacitor,

$$Q = \frac{X_C}{R_C} = \frac{1}{\omega C R_C} \tag{3.2-2}$$
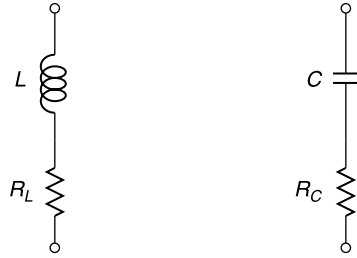
Figure 3.2-1 L and C models for Q_U definition.

When a series resonant circuit is formed from a LC pair, one can define a Q for the circuit at the resonant frequency ω_0. Called the *unloaded Q*, Q_U, it relates the energy stored in the reactive elements to the power loss in the resistive elements and, hence, is a measure of the quality of the resonator, itself. Thus,

$$Q_U = \frac{\omega_0 L}{R_L + R_C} = \frac{1}{\omega_0 C(R_L + R_C)} \tag{3.2-3}$$

Q_E, External Q

Even when a resonator is formed from an ideal, lossless LC pair, there is power dissipation when it is connected to external loads (Fig. 3.2-2). This dissipation affects the sharpness of the resonant response, which depends upon the ratio of the amount of energy stored in the capacitor or inductor relative to that dissipated in the external load or loads.

Under these conditions, *the external quality factor Q_E is the ratio of the magnitude of the reactance of either the L or C at resonance to the total external resistance:*

$$Q_E = \frac{\omega_0 L}{Z_0 + Z_0} = \frac{1}{\omega_0 C(Z_0 + Z_0)} \tag{3.2-4}$$
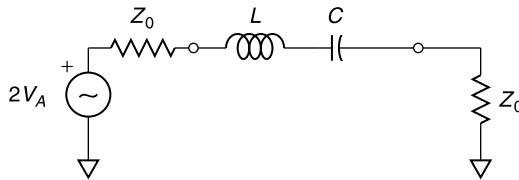
Figure 3.2-2   Series equivalent circuit of LC resonator with dissipation only in the external loads (Q_E model).
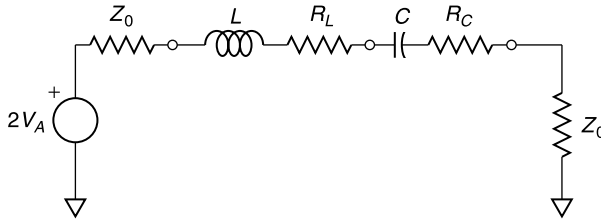
Figure 3.2-3 Q_L includes the effect of resonator resistances and external load.

Q_L, Loaded Q

The loaded quality factor Q_L is the ratio of the magnitude of the reactance of either reactor at resonance to the total circuit resistance (Fig. 3.2-3). In the case of a *doubly match-terminated resonator*, that is, one having a source resistance Z_0 and a load impedance Z_0, the total external resistance is $2Z_0$:

$$Q_L = \frac{\omega_0 L}{Z_0 + Z_0 + R_L + R_C} = \frac{1}{\omega_0 C_0(Z_0 + Z_0 + R_L + R_C)} \quad (3.2\text{-}5)$$

From (3.2-3), (3.2-4), and (3.2-5), it follows that

$$\frac{1}{Q_L} = \frac{1}{Q_E} + \frac{1}{Q_U} \quad (3.2\text{-}6)$$

Thus, it is seen that the three Q factors add reciprocally. Equivalently, the product over the sum method can be used to calculate Q_L from Q_U and Q_E:

$$Q_L = \frac{Q_U Q_E}{Q_U + Q_E} \quad (3.2\text{-}7)$$

3.3 PARALLEL CIRCUIT QUALITY FACTORS

The dual of the circuit of Figure 3.2-3 is that shown in Figure 3.3-1 in which the resonator consists of shunt-connected elements. For the shunt equivalent circuits the Q's are defined as the ratio of the susceptance of *either* the L or the C to the appropriate conductances, which are, for the inductor,

$$Q = \frac{B_L}{G_L} = \frac{1}{\omega L G_L} \quad (3.3\text{-}1)$$

and for the capacitor

$$Q = \frac{B_C}{G_L} = \frac{\omega C}{G_L} \quad (3.3\text{-}2)$$
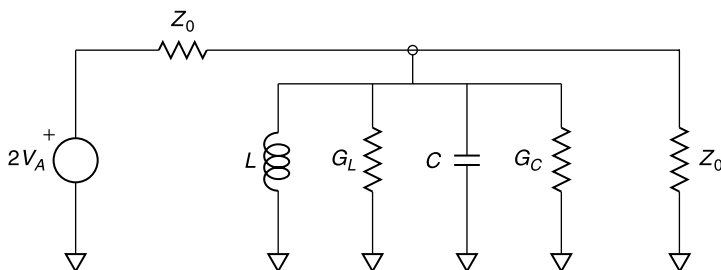
Figure 3.3-1 Equivalent circuit of parallel element resonator.

The overall Q values for the LC resonator are defined at ω_0 and given by

$$Q_U = \frac{\omega_0 C}{G_L + G_C} \quad Q_E = \frac{\omega_0 C}{2 Y_0} \quad Q_L = \frac{\omega_0 C}{G_L + G_C + 2 Y_0} \qquad (3.3\text{-}3\text{a, b, c})$$

The relationships among the Q values apply equally to the series and parallel resonant circuits.

3.4   COUPLED RESONATORS

Direct Coupled Resonators

When an LC resonator is placed in series or in shunt between matched generator and load, the result is *direct resonator coupling*. For ideal lossless resonator elements the insertion loss is *0 dB* at the resonance frequency, since all of the available generator power is delivered to the matched load.

For Q_L values above 10, the fractional 3-dB bandwidth of an LC resonator is well approximated as $1/Q_L$ [1, p. 315]. This is shown in Figure 3.4-1 for the case of a direct coupled series resonator in a 50-Ω system. Note that R_L and R_C are zero in this case.

A similar result is obtained when a parallel LC combination is directly coupled between generator and load, as shown in Figure 3.4-2. Note that in this case Q_L is calculated from (3.3-3c) with $G_L = G_C = 0$.

Lightly Coupled Resonators

A more general definition [1, p. 314] of Q_L for a resonator is

$$Q_L = \frac{\omega_0(\text{time-average energy stored in the system})}{\text{energy loss per second in the system}} \qquad (3.4\text{-}1)$$

where ω_0 is the resonant frequency of the resonator. This more general defini-
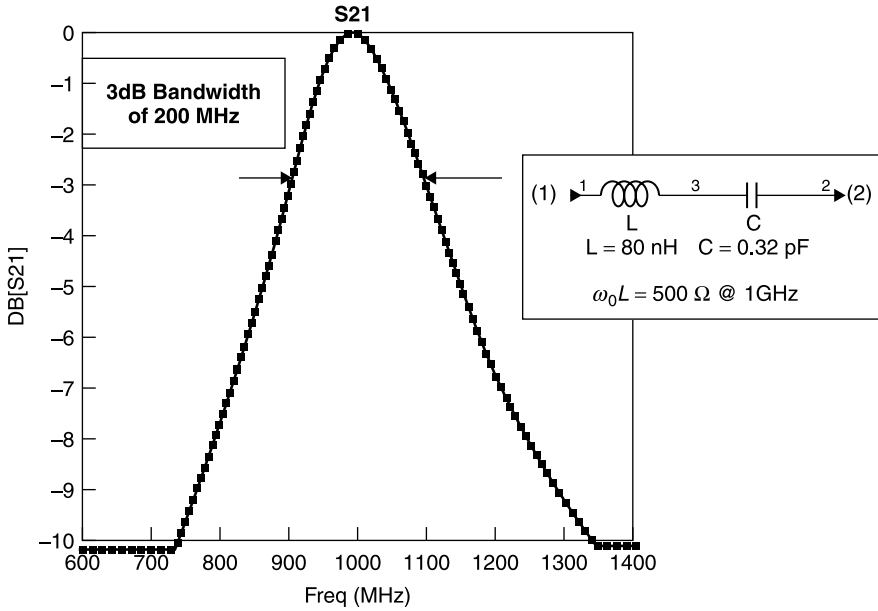
Figure 3.4-1   Resonant *LC* with $Q_L = 5$. Note that $Q_L = 500/(50 + 50) = 5$ and therefore the 3-dB bandwidth is $f_0/Q_L = 1000/5 = 200$ MHz.
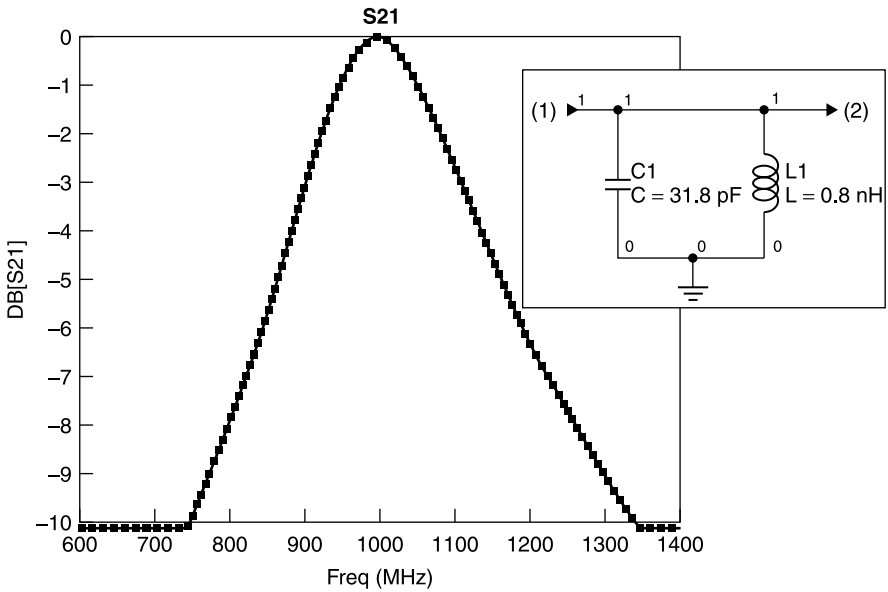
Figure 3.4-2   Shunt *LC* resonator mounted between 50-Ω source and load. The 3-dB bandwidth is 200 MHz, about 1000 MHz center frequency for a loaded $Q = 5$.

tion has the advantage that it can be applied to any type of resonator, including cavity and dielectric resonators. As with the lumped element model, the fractional 3-dB bandwidth remains a good approximation to the reciprocal of the loaded Q when $Q_L = 10$ or more.

The series and shunt resonators shown in the previous two examples are *directly coupled*. That is, either the current through the series resonator is the same as that through the load or the voltage across the parallel resonator is the same as that across the load. Consequently, to achieve a high Q_L the reactance magnitudes of the series elements at resonance must be relatively high compared to the load impedance. In a parallel circuit the magnitude of the shunt susceptance of each element at resonance must be correspondingly higher than the load conductance.

Practically, direct coupling constrains the amount of energy that can be stored in the LC resonator relative to that which is dissipated in the load because it is difficult to realize large ratios between resonator and load impedance magnitudes. For example, to obtain the modest Q_L of 5 at 1 GHz in the previous example with a load impedance of 50 Ω required that the reactance of the shunt inductance be 5 Ω, resulting in a very low inductance value of 0.8 nH.

If, instead, the resonator is only lightly coupled to the generator and load, the voltage across and the current through the L and C elements is theoretically unlimited. This can permit a much higher loaded Q, as the following example demonstrates. Of course, real resonators have finite losses, and therefore the Q_L is not unlimited in practice, but the following lossless example (Fig. 3.4-3) serves to illustrate the principles of the lightly coupled resonator.

Paradoxically, *the lighter the coupling the higher the stored energy in the resonator*. For comparison we first consider a direct coupled parallel resonator with resonant frequency of 1000 MHz, $L = 3.18$ nH and $C = 7.95$ pF. Directly coupled in a 50-Ω system, this results in a loaded Q of only 1.25, so low that the actual 3-dB bandwidth, as determined from circuit simulation, is about 800 MHz.

Next, the same resonator is coupled to generator and load through 1.16-pF capacitors, as shown in the schematic in Figure 3.4-3. The resulting 3-dB bandwidth is only 100 MHz. Some retuning of the resonator's capacitor from 7.95 to 5.911 pF was required because the capacitive coupling reactance increases the total capacitance of the circuit and would lower the resonant frequency if uncompensated. This configuration is called a *top-C coupled parallel resonator filter* [2, p. 155]. Notice that the resonator's inductor is 3.18 nH, four times as large as, and more practically realizable than, the 0.8 nH of the preceding 200-MHz direct coupled, shunt resonator example.

Physically, the action of the network can be visualized by considering what occurs when the generator is first turned on. This is a transient condition, not covered by our steady-state AC analysis. Due to the light coupling only a small amount of energy enters the resonator. The peak voltage across the capacitor does not build up as rapidly as it would with direct coupling because energy must enter through the high impedance of the coupling capacitor. Hence it may
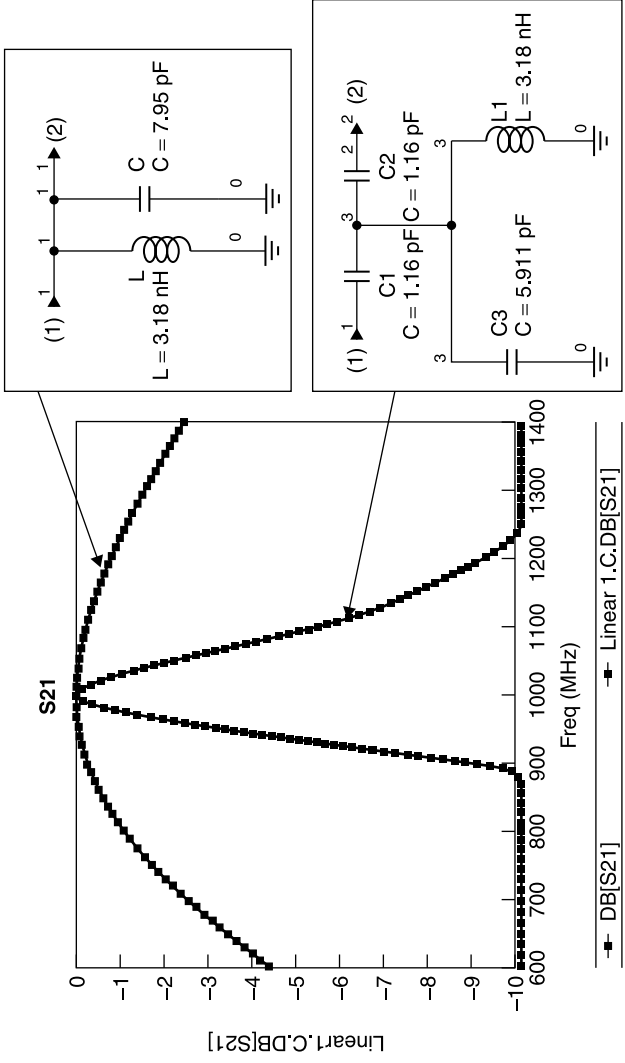
Figure 3.4-3 Performance comparison of direct and light coupling of *LC* resonator between matched generator and load.

take several RF cycles for the steady-state AC voltage to become established across the resonator.

The voltage increases, from cycle to cycle, until it is sufficient to block further energy increase from the generator. Since the network is symmetric in this case, in the steady state energy flows out of the resonator to the load at the same rate that it enters from the generator. For this ideal, lossless *LC* case there is no insertion loss at the resonant frequency. The higher the reactance of the coupling capacitors, the greater the resonator voltage (to block further power from entering) and the greater the stored energy in the resonant *LC*. This results in a higher Q_L and a narrower 3-dB bandwidth.

Why did the loaded *Q* of the lightly coupled resonator circuit increase by a factor of 8 relative to that of the directly coupled resonator having the same *LC* elements? We can infer that the energy stored in the coupled resonator also increased by a factor of about 8. This energy storage in a parallel *LC* resonator is proportional to the square of the voltage across it [1, pp. 314–315]. Clearly, the voltage across the top coupled resonator must have increased relative to the line voltage V_A of the direct coupled case.

This must be true because the same power reaches the load in both cases, yet for the top coupled circuit the load current must be driven through not only the 50 Ω load resistance but the series 1.16 pF capacitance as well, an additional $-j137\ \Omega$ at 1 GHz. The impedance of capacitor and load is $(50 - j137)\ \Omega$, and its magnitude is 146 Ω, requiring a voltage increase of 146/50 or 2.92. The corresponding energy storage increase in the *LC* resonator of the top coupled case is therefore a factor of 2.92^2 for a factor of 8.5, approximately equal to the observed 3-dB bandwidth decrease by a factor of 8.

3.5 *Q* MATCHING

A very useful method for accomplishing resistive impedance transformation at a given frequency is that of the *LC* network in which one element, either an inductor or a capacitor, is placed in series or parallel with the load to be transformed and the remaining element is made the opposite kind and placed, respectively, in parallel or series with the combination. Interestingly, the mathematical development of this matching technique arises from the formula for converting a series impedance to its parallel equivalent, or vice versa. To demonstrate the method and its basis, we review the *series-to-parallel equivalent circuit transformation* (Fig. 3.5-1).

Low to High Resistance

For this development, it is illustrative to derive the equivalence between Z and Y by first expressing Z in terms of its Q. As will be seen, this leads to a lumped circuit resistance transformation [3, p. 35]. In this exercise, we simply refer to the quality factor of Z as Q. We derive the result starting with a positive
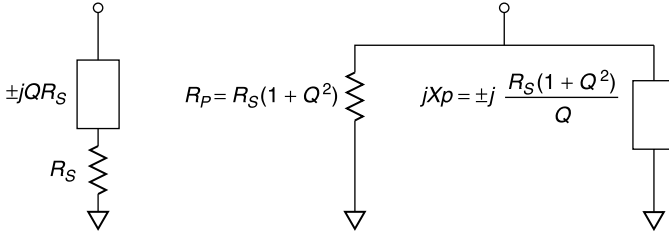
Figure 3.5-1   Series-to-parallel equivalent circuits.

imaginary value for Z, but the derivation is similar if begun with a negative imaginary value:

$$Z = R_S + jX$$
$$Z = R_S(1 + jQ) \tag{3.5-1}$$

$$Y = \frac{1}{Z} = \frac{1}{R_S(1 + jQ)} \cdot \frac{1 - jQ}{1 - jQ}$$

$$Y = \frac{1 - jQ}{R_S(1 + Q^2)} \tag{3.5-2}$$

$$Y = G - jB$$

$$G = \frac{1}{R_S(1 + Q^2)} \tag{3.5-3}$$

$$B = \frac{Q}{R_S(1 + Q^2)} \tag{3.5-4}$$

Converting the parallel admittance terms into parallel resistance R_P and parallel reactance X_P,

$$R_P = R_S(1 + Q^2) \tag{3.5-5}$$

$$X_P = \frac{R_S(1 + Q^2)}{Q} \tag{3.5-6}$$

The important point to notice is that *the parallel equivalent circuit of a series impedance has a resistance value that is higher than the series value by the factor* $(1 + Q^2)$. This suggests that resonating the reactance in the parallel equivalent circuit [4, p. 120] yields a lumped circuit, resistance transforming method. *To transform a low resistance to one of higher resistance, add sufficient series reactance to bring the parallel equivalent resistance to the value desired, and then resonate the resulting circuit by adding an appropriate parallel susceptance.*

Generally, the purpose in transforming a resistive load to a different value is to match the load to a circuit of different source impedance. This method

can be called *LC matching* (because opposite reactor types must be used), *L matching* (because of the *L* orientation of the reactors), or *Q matching* [5] (because the *Q* is chosen based on the transformation ratio).

In the remainder of this text we will call the method *Q matching*. As an example, suppose we wish to transform 5 to 50 Ω at 1 GHz. This requires that we *increase* the resistance by a factor of 10. Accordingly

$$1 + Q^2 = 10$$
$$Q = 3$$

Since we wish to increase resistance, add a series reactor, say L_1 (either *L* or *C* could be used), to the 5-Ω resistor. Examination of (3.5-5) and (3.5-6) shows that *the parallel equivalent circuit has the same Q as the series circuit*. For example, if the series circuit consisted of a resistor and an inductor, its *Q* is given by

$$Q = \frac{\omega L}{R_S} = \frac{X_L}{R_S}$$

The *Q* of a parallel *LC* circuit is

$$Q = \frac{B_L}{G}$$

By substituting the values for B_L and *G* from (3.5-5) and (3.5-6), we obtain

$$Q = \frac{B_L}{G} = \frac{R_P}{X_P} = \frac{R_S(1 + Q^2)}{R_S(1 + Q^2)/Q} = Q$$

which demonstrates that the series and parallel *LR* circuits have the same *Q*. We expect this result since the series and parallel circuits are equivalent and therefore must have the same quality factor.

Given this condition, the reactance of L_2 in the parallel equivalent circuit can be written down immediately. It must have a parallel reactance of $50/Q = 16.7$ Ω because in the parallel circuit the reactance magnitude is the resistance divided by the *Q*.

Next, add a capacitive reactance of equal magnitude to parallel resonate the circuit. The resulting net admittance is 0.02 ℧, or 50 Ω impedance at 1 GHz, the frequency at which the match is performed. Notice that *in the Q matching method, the two reactors are always of opposite type* (Fig. 3.5-2). This is because the final circuit is resonated to remove any net reactance or susceptance. In the present example, when these reactances are converted to their corresponding *L* and *C* values, the actual *Q* matched circuit is shown in Figure 3.5-3. Notice that the resistance peaks just above 1000 MHz. This circuit might be made to
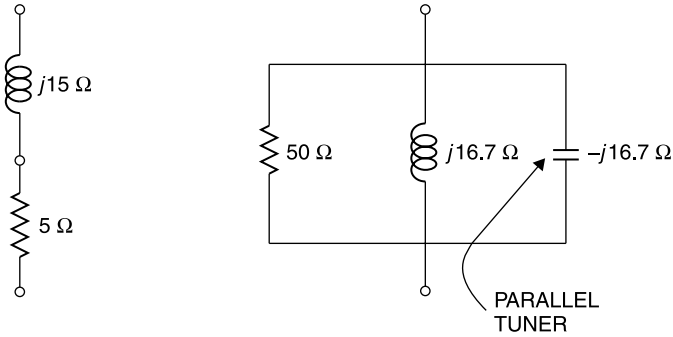
Figure 3.5-2 *Q* matching example, 5 to 50 Ω at 1 GHz.

have a broader bandwidth (closer to 50 Ω over a wider spread of frequencies around 1 GHz) by using the network simulator optimizer. Nevertheless, the transformation from 5 to 50 Ω was determined very simply using the *Q* method (Fig. 3.5-4).

Broadbanding the *Q* Matching Method

A single *LC* matching circuit has limited bandwidth. *To increase bandwidth, use multiple matching circuits, changing the resistance in smaller, equal ratio steps.*

For n-section Q matching of a resistance ratio R, make each successive section change the load impedance by the ratio $\sqrt[n]{R}$. Thus,

$$1 + Q^2 = \sqrt[n]{R} \tag{3.5-7}$$

To illustrate this multiple resistance step transformation, repeat the previous example using three sections. Note that $\sqrt[3]{10} = 2.154 = 1 + Q^2$; then $Q = 1.074$ for each matching section.

Therefore, the first section transforms the 5-Ω load to $2.154(5\ \Omega) = 10.77\ \Omega$. The second section transforms $10.77\ \Omega$ to $2.154(10.77) = 23.20\ \Omega$. The third section transforms $23.20\ \Omega$ to $2.154(23.20) = 50\ \Omega$. When the *Q* matching method is performed three times to obtain the transformations listed above, the result is that shown in Figure 3.5-5.
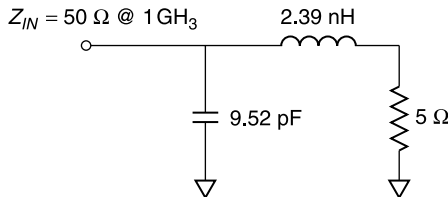
Figure 3.5-3 Final *Q* matched resistor network.
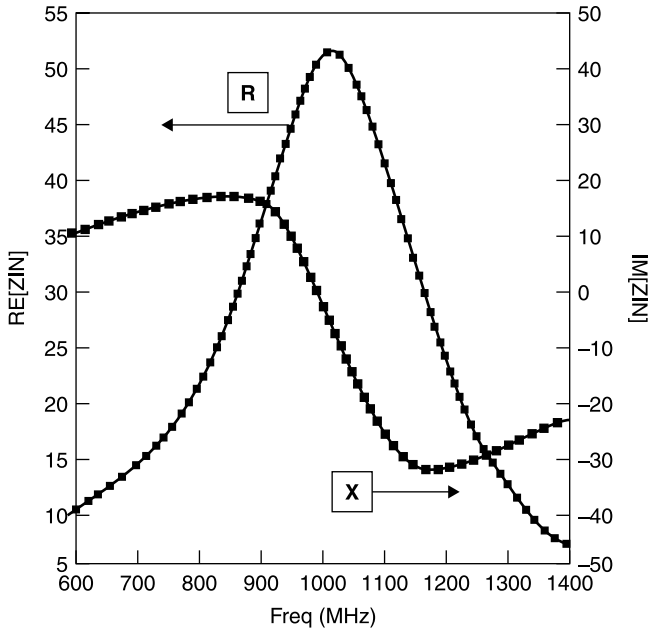
Figure 3.5-4   Frequency response of *Q* transformation network from 5 to 50 Ω.

Notice that with the three-section matching network, the bandwidth is much greater. The input resistance *R* is within 5 of 50 Ω from 800 to 1250 MHz, and the imaginary part, *jX*, is very near zero over the same range. This is a much broader bandwidth than the 970 to 1100 MHz bandwidth (Fig. 3.5-4) over which the single-stage matching yielded within 5 of 50 Ω for the real part and a much larger series reactance.

High to Low Resistance

If the resistance of a load is higher than desired, the same Q matching method can be used except that a shunt reactor is used as the first transformation element.

The same equivalent circuits and circuit element relations given in Figure 3.5-1 are employed, but in this case the transformation is from right to left, that is, from the parallel equivalent to the series equivalent circuit. We demonstrate this with a load circuit consisting of a 250-Ω resistor in series with a 0.7958-pF capacitor ($-j200$ Ω at 1 GHz). This example also demonstrates how to absorb the existing series reactance. We wish to transform this impedance to 50 Ω at 1 GHz.

The $(250 - j200)$-Ω load is shown in Figure 3.5-6a. Since the 250-Ω real part exceeds the desired 50 Ω, a parallel circuit is needed to use the *Q* transformation approach. The first step is to transform this series circuit to its parallel
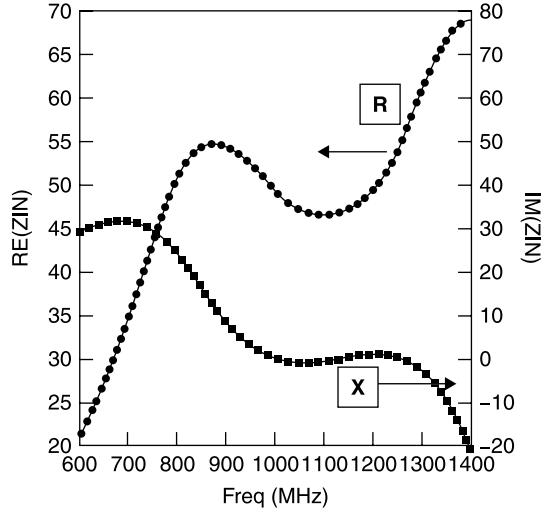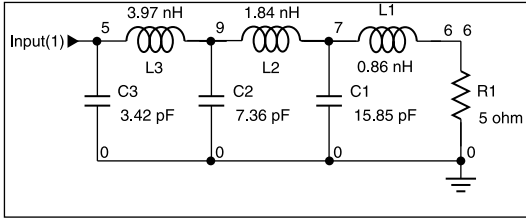
Figure 3.5-5 Three-section Q matching from 5 to 50 Ω.

equivalent. The parallel equivalent is easily obtained by noting that the Q of the load is 0.8. In the parallel equivalent circuit the shunt resistor will be larger than 250 Ω by the factor $(1 + Q^2) = 1.64$. Thus the shunt resistor is 410 Ω. The shunt equivalent circuit must have the same Q. Observing that in the parallel equivalent circuit the shunt reactance is related to the shunt resistor by the factor $(1/Q)$, we see that the shunt capacitive reactance magnitude must be 410/0.8 Ω = 512.5 Ω. At 1 GHz this corresponds equivalently to a capacitance
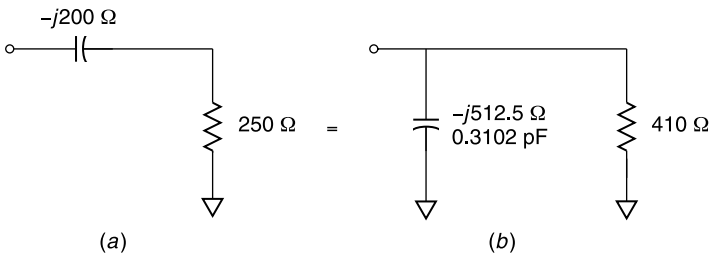
Figure 3.5-6 (*a*) The $(250 - j200)$-Ω load and (*b*) its parallel equivalent circuit.
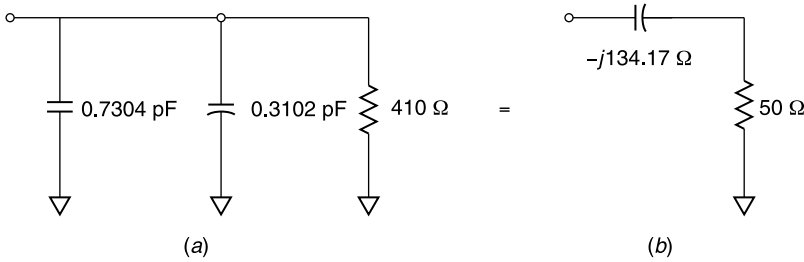
Figure 3.5-7   Equivalent circuit of load after the addition of 0.7304 pF capacitor.

of 0.3102 pF. Note that *the nature of the reactance is unchanged between the series and parallel equivalent circuits*. In this case, the reactance is capacitive.

Next, the task is to transform 410 to 50 Ω. This is a ratio of 8.2. Therefore, $1 + Q^2 = 8.2$ and $Q = 2.6833$. The large number of decimal places is carried to show that the method produces an arbitrarily accurate result if executed with sufficient precision. In order to change the *Q* of the parallel equivalent circuit in Figure 3.5-6b, we add shunt capacitance to produce a total shunt reactance of 410 Ω/2.6833 = 152.80 Ω. This corresponds to a total capacitance of 1.0406 pF. Since 0.3102 pF is already available, only an additional 0.7304 pF need be added, as shown in Figure 3.5-7a. In this way the original 200 Ω reactance has been absorbed within the *Q* matching circuit.

The series equivalent circuit (Fig. 3.5-7b) has the required 50 Ω resistance. The magnitude of the series reactance must have the same *Q* of the parallel circuit, in this case $Q = 2.6833$. The resulting capacitive reactance magnitude is 50 Ω × 2.6833 = 134.17 Ω. The reactance type is unchanged by expressing the circuit in series format; hence, the reactance is capacitive, as shown in Figure 3.5-7b.

The final step is to resonate the $-j134.7$ Ω with a series inductor (21.353 nH) of $+j134.7$ Ω to obtain the desired 50 Ω real input impedance. The resulting circuit and its performance with frequency is shown in Figure 3.5-8.

This tuning could have been performed equally well, with the same *Q* values, by using a shunt inductor as the first tuning element. The inductive susceptance would be larger than that of the capacitor actually used, since it would be required to neutralize the extant capacitive susceptance already in place and then to provide the necessary susceptance magnitude required. But this would not change the required *Q* of 2.6833. With this procedure, the second tuning element would have been a capacitor instead of an inductor.

In some circuit applications having a series capacitor could be advantageous as a "DC block" to prevent bias currents from passing to the signal terminal. In other circuits the presence of a series inductor to conduct bias current might be useful. *The availability of two different circuit topologies for Q matching is a design advantage.*

As was demonstrated, the *Q* matching method is not limited to transforming a pure load resistance. The load can be complex. This method of tuning of
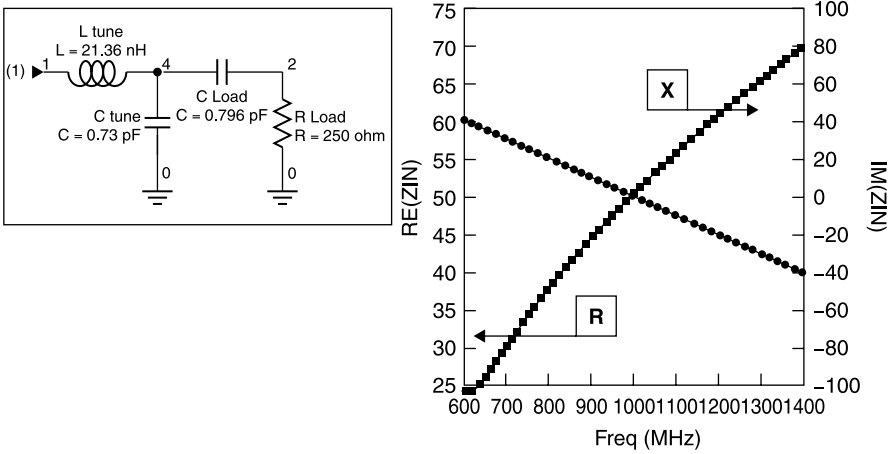
Figure 3.5-8   Circuit to tune $(250 - j200)$-Ω load to 50 Ω at 1 GHz.

using one series reactor and one shunt reactor can transform any lossy imped-
ance to any resistance. This distinction about the load being lossy is necessary
because, using lossless *L* and *C* elements, *a lossless load cannot be transformed
into a lossy one, nor vice versa.*

In practice the range of transformations is limited by the fact that the tuning
reactors do not have zero resistance and they have parasitic inductances and
capacitances. Nevertheless, the typically realizable range and efficiency of tun-
ing in practical circuits makes this *Q* matching technique a valuable design
tool.

REFERENCES

1. Robert E. Collin, *Foundations for Microwave Engineering*, McGraw-Hill, New York,
   1966. *This is an excellent and very fundamental treatment of microwave engineering
   topics.*
2. Randall W. Rhea, *HF Filter Design and Computer Simulation*, Noble Publishing,
   Atlanta, GA, 1994. *This book is must reading for the filter designer. It contains an
   excellent introduction to filter theory with numerous examples of both lumped element
   and distributed filters. It includes practical effects such as coil winding and the element
   Q's that are practically realizable.*
3. Chris Bowick, *RF Circuit Design*, Butterworth-Heinemann, Woburn, MA, 1982. *This
   is a very readable and practical design guide to filter design and matching circuits.*
4. Guillermo Gonzalez, *Microwave Transistor Amplifiers, Analysis and Design*, 2nd ed.,
   Prentice-Hall, Upper Saddle River, NJ, 1984. *An excellent engineering reference for
   transistor amplifier design.*
5. Pieter L. D. Abrie, *Q-Match: Lumped-Element Impedance Matching Software and
   User's Guide*, Artech House, Norwood, MA, 2000.

EXERCISES

Note. For exercises requiring the use of a network simulator, plot the magnitude of S_{21} in decibels for insertion loss and the magnitude of S_{11} in decibels for return loss. For a definition of S parameters, see Section 6.5 and for a return loss definition see Section 4.6.

E3.2-1   Use a network simulator to find the loaded Q of an 80-nH inductor having 5 Ω of series resistance and a 0.032-pF capacitor having 50,000 Ω of parallel resistance when the two are connected in series and used as a resonator in a doubly match-terminated 50-Ω line.

E3.2-2   Use a network simulator to design a dual resonator, top-C coupled network to provide a 3-dB bandwidth of 100 MHz and a center frequency of 1000 MHz. Start with the circuit values shown below, keep L1 and L2 fixed, and maintain symmetry with respect to the input and output ports of 50 Ω. Vary C1, C2, and C3 to obtain a 3-dB bandwidth of 100 MHz.
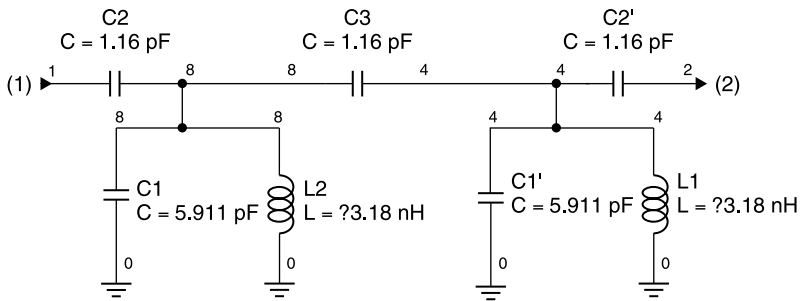
a. How does the frequency response compare with that of the single resonator, lightly coupled circuit of Figure 3.4-3?

b. What roles do the respective capacitors C1, C2, and C3 play in the frequency response of this network?

c. What points does this exercise demonstrate?

E3.2-3   If the inductors for the circuit in E3.2-2 are realized using a straight length of copper wire 0.025-in. diameter (22 gauge) and the conductivity of copper is 5.8×10^7 $\mho$/m:

a. What wire length is needed for the 3.18-nH inductors, assuming the wire is well removed from ground?

b. What is the inductor's resistance at 1000 MHz?

c. What is the inductor's Q at 1000 MHz.

d. How does this affect the resonant frequency insertion loss of the network?

e. What can you say about the practicality of this network as a radio filter?

E3.5-1 Use a network simulator to design a *LC* matching network to match a 5-Ω resistor to 50 Ω at 1 GHz. Do not use the optimizer. *Hint. Start with a series inductor L1 near the 5-Ω resistor and try to get the real part of the admittance at node 3 to be 0.02 ℧, then parallel resonate the input with C1.*
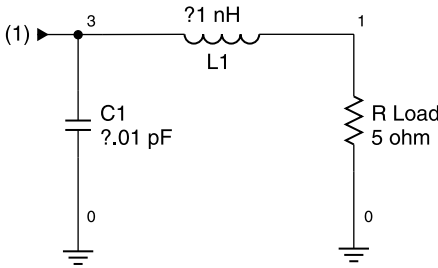
E3.5-2 **a.** Repeat E3.5-1 but this time use the *Q* matching method and transform 10 to 50 Ω.

   b. Use a circuit simulator to compare the return loss versus frequency of this circuit with that used to transform 5 to 50 Ω in E3.5-1.

   c. Was this manual calculation easier than the computer-aided solution used for E3.5-1?

   d. How do the bandwidths of the transformations from 5 to 50 and from 10 to 50 Ω compare? Why is one narrower?

E3.5-3 A semiconductor manufacturer has hired you as a consultant. He wants to build a power transistor for 2-GHz applications and expects that it will have an input equivalent circuit consisting of 4 Ω in parallel with 80 pF. He has been told that a *LC* network can be designed to tune any impedance to 50 Ω, so he believes that his proposed transistor equivalent circuit should be all right.
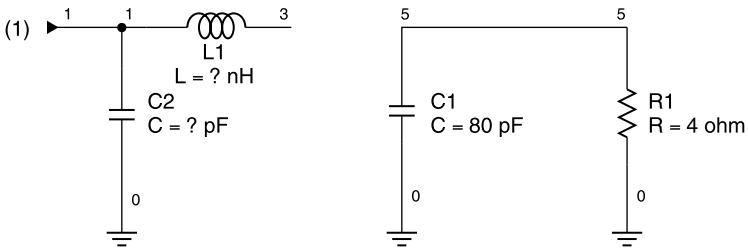
   a. What are the values for L1 and C2 of the tuner?

     b. Examine the frequency behavior of the tuned circuit using a circuit simulator.

     c. Comment on the practicality of this tuning arrangement.

E3.5-4   **a.** Use the Q matching method to design an 800-MHz LC matching circuit to match a transistor having 10 Ω input resistance in series with 0.8 nH of lead inductance to a 50-Ω source.

     b. Use a circuit simulator and determine: How wide is the 20-dB-return-loss bandwidth? How wide is the 10-dB-return-loss bandwidth?

E3.5-5   **a.** For the application of E3.5-4, design a two-stage matching network.

     b. Compare the performance of this circuit with that of the previous single-stage matching design. How much bandwidth improvement is obtained?

E3.5-6   Use the network simulator with the optimizer, varying all four elements of the two-stage tuner developed in E3.5-5. Set as an objective the achievement of a return loss of less than -20 dB (the magnitude of $S_{11} < -20$ dB) from 600 to 1000 MHz.

Distributed Circuits

4.1 TRANSMISSION LINES

A single inch of wire can have 10 nH of inductance, an impedance of $+j63\ \Omega$ at 1 GHz. Interconnecting circuit parts with uncontrolled wire lengths would lead to large, unwanted reactances. It would also lead to unpredictable circuit behavior because circuit components having transmission paths larger than about one tenth of a wavelength impart not only reactive changes but resistive transformations in impedance as well.

Today, with computer assistance, it is possible to analyze circuits having nearly arbitrary geometry, and therefore any hookup wire lengths theoretically can be handled analytically. Nevertheless, it is inadvisable to use irregular transmission line geometries except when there is a specific good reason for doing so.

In the early 1930s computers were not available to analyze circuits, and it was necessary that engineers confine transmission circuit designs to simple, regular cross sections in order that performance could be predicted. These were called *uniform transmission lines*, and they permitted more predictable behavior for distributed circuits.

Even today, for *distributed circuits*, wherein *the lengths of the transmission paths are appreciable compared to the operating wavelength*, it is generally advisable to interconnect circuits and circuit parts with uniform transmission lines. Unless a transmission line is kept uniform in cross section, its effect on the circuit becomes impractically difficult to compute by hand, and it is equally difficult to gain insight into how the circuit works.

A transmission line is a set of conductors that are long compared to a wavelength (generally considered to be $>\lambda/10$) and have a uniform cross section along their length for which a characteristic impedance Z_0 can be defined.

In this chapter we first describe without proof the behavior of signals on transmission lines in order to provide a perspective of their operation. Later these transmission line circuit behaviors are derived analytically.
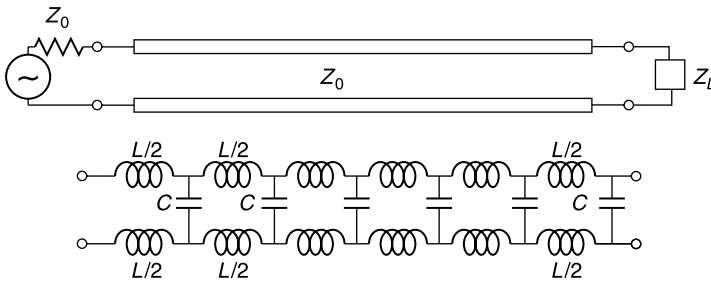
Figure 4.1-1   Balanced, two-wire, lossless transmission line with generator and load connections and the equivalent circuit.

The lumped circuit model for a lossless, balanced, uniform transmission line is shown in Figure 4.1-1. In the equivalent circuit, *L is the inductance/unit length* and *C is the capacitance/unit length* of the uniform transmission line. For the lossless case the *characteristic impedance* Z_0 of the line is given by

$$Z_0 = \sqrt{\frac{L}{C}} \; (\Omega) \tag{4.1-1}$$

where Z_0 *is defined as the ratio of voltage-to-current for a traveling wave in either direction*. For the lossless line Z_0 is independent of frequency, a fact that is most important for broadband matching. *When a transmission line is terminated in its characteristic impedance* (Z_0), *no reflections occur*. That is, all power traveling toward the load is absorbed at the load, and none is reflected back toward the generator. There are other ways to accomplish total absorption of a traveling wave by the load, and we shall examine them as matching techniques.

The transmission line shown in Figure 4.1-2 is an example of a *balanced line* because both of its conductors have the same impedance to ground. For this reason stray voltage and currents induced onto either conductor of the line from interfering sources tend to be identical on both conductors, producing *common mode* voltage on the conductors and canceling insofar as the signal is concerned. The signal voltage is the difference in potential between the two conductors, and accordingly is called the *differential mode* voltage.

A balanced line example is the twisted pair phone line. The twist is added to further ensure that the two lines have identical impedances to ground, and, consequently, noise pickup is almost entirely in the common mode. This noise is not induced onto the signals carried by the conductors since the signals are impressed on the line in the differential mode.

Another example of a two-wire balanced transmission line is the *twin lead* used between an outside roof antenna and a TV set. Its geometry yields a characteristic impedance of about 300 Ω, more closely matching that of the source impedance of television antennas. Its balanced nature provides the same noise immunity to such sources as automobile ignitions, motors, and other
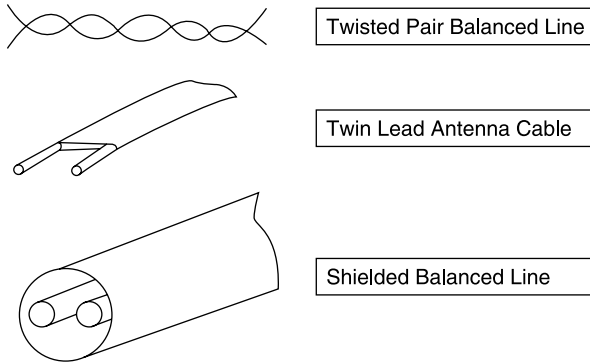
Figure 4.1-2   Examples of balanced transmission lines.

noise sources. However, it cannot provide immunity to noise that enters the system through radiation reception by the television aerial itself.

The balanced line pair can be further protected from interference by enclosing it within a shield, in which case the line is a *balanced shielded transmission line*.

For actual radio circuits *unbalanced transmission lines*, one of whose conductors is generally at ground potential, are more common (Fig. 4.1-3). The primary examples are stripline and microstrip, requiring only a "center conductor" pattern that can be photoetched for low cost and easy reproducibility.

Coaxial lines (also called *coax*) were developed to provide a flexible line with inherent shielding. They are used throughout the microwave bands, for example, up to 1 GHz for television cable systems and to 50 GHz or more in instrumentation applications.

Stripline evolved from coax in about 1950 and was considered a breakthrough due to its manufacturing ease. Its lack of mechanical flexibility was not a disadvantage since it was generally used to interconnect circuit parts installed on a common circuit board.

Microstrip evolved after stripline. It shared the reproducibility advantages of stripline but had the further advantage of providing easy access to circuit com-
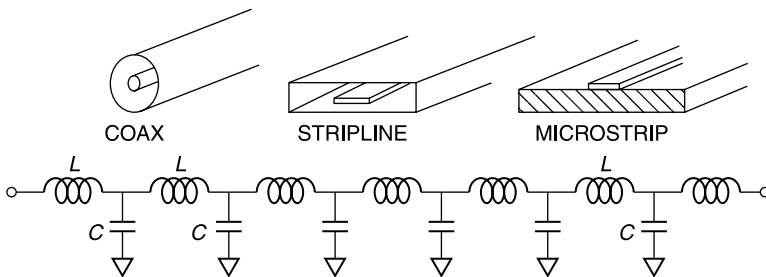
Figure 4.1-3   Unbalanced transmission line formats commonly used in wireless applications.

ponents since there was no upper ground plane to get in the way of component installation and tuning. However, microstrip's format usually results in an *inhomogeneous dielectric* medium, that is, a differing dielectric media for various portions of the transmission line's cross section. This occurs because microstrip is usually implemented as a conductor pattern on a dielectric material with air dielectric above the transmission line patterns.

As frequency increases, the capacitance between center conductor and ground causes an increasingly larger percentage of the energy to propagate within the dielectric, resulting in a greater signal delay for higher frequencies than for lower frequencies. This variation in signal delay with frequency distorts wideband signals because their various frequency components experience differing transmission delays through the inhomogeneous transmission line medium.

4.2 WAVELENGTH IN A DIELECTRIC

The *wavelength is the distance a sinusoid propagates in its period*. Wavelength is reduced in proportion to the square root of the dielectric constant ε_R of the propagating medium, since the speed of propagation is reduced in that proportion (Fig. 4.2-1). For a nonmagnetic dielectric

$$\lambda = \frac{\lambda_0}{\sqrt{\varepsilon_R}} \tag{4.2-1a}$$

$$v = \frac{c}{\sqrt{\varepsilon_R}} \tag{4.2-1b}$$

$$f\lambda = v \tag{4.2-1c}$$

$$f\lambda_0 = c \tag{4.2-1d}$$

AIR DIELECTRIC $\varepsilon_R = 1$

λ_0

COAX LINE $\qquad\qquad\qquad\qquad \varepsilon_R = 2$

$\frac{\lambda_0}{\sqrt{\varepsilon_R}}$
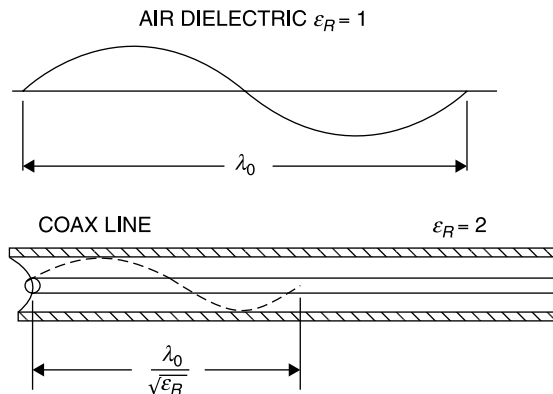
Figure 4.2-1 Effect of dielectric material is to decrease propagation speed, and with it, wavelength.

where λ_0 is the free space wavelength, c is the velocity of light in free space $(2.9979 \times 10^8$ m/s), and λ is the wavelength in the dielectric medium.

Typical transmission line dielectrics range from pure Teflon with relative dielectric constant, $\varepsilon_R = 2.03$, to alumina (Al_2O_3) with $\varepsilon_R = 10$. Sapphire is the single-crystalline form of alumina and is an ideal, albeit expensive, microwave dielectric. Ceramic materials are available with relative dielectric constants up to 100 for use as primary resonators for oscillator circuits. Pure water has a dielectric constant of 81 [1] at RF and microwave frequencies.

Teflon material has very low dissipative loss and is a primary choice for microwave circuit boards. However, in pure form, it both expands and cold flows substantially with temperature and therefore is usually reinforced with embedded glass fibers for dimensional stability. The glass often is woven into a cloth fabric for reinforcing printed circuit (PC) boards. The presence of glass fibers increases the dielectric constant, and, if the fibers have a particular average orientation, as occurs with woven fabrics, the result is a dielectric constant that is a function of direction in the dielectric. Such a material is called an *anisotropic dielectric*.

To overcome the directional variation of dielectric constant the dielectric is sometimes reinforced using small, randomly oriented glass fibers (microfibers), making it uniform in all directions. A popular microfiber circuit board material is Rogers *Duroid*.

4.3   PULSES ON TRANSMISSION LINES

The concept of reflections on transmission lines can be visualized by considering the reaction of an open-circuited transmission line when a voltage pulse is sent down it.

Consider a voltage pulse applied to a lossless transmission line. The pulse propagates at the speed of light in a transmission line with air or vacuum surrounding the conductors or propagates at reduced speed if the line is embedded within a higher dielectric constant medium. For a lossless line, the pulse propagates along the line, as depicted in sketches 1 to 4 in Figure 4.3-1, its voltage undiminished with distance since there are assumed to be no dissipative or radiative losses. Associated with the voltage wave is a current wave. The ratio of voltage to current is Z_0.

In this theoretical idealization, on reaching the open-circuited end, all of the incident energy must be reflected since it can neither be radiated nor stored there. Of course, on a real open-circuited transmission line, some dissipative and radiative losses would occur, but we ignore them in this example.

Since all of the power of the pulse must be reflected at the open end of the line, the situation is equivalent to the introduction of a reverse-going pulse that is otherwise identical to the incident pulse. At the open circuit the current must be zero. To satisfy this condition, the reflected current must be directed opposite to the incident current. This requires that the associated reflected voltage
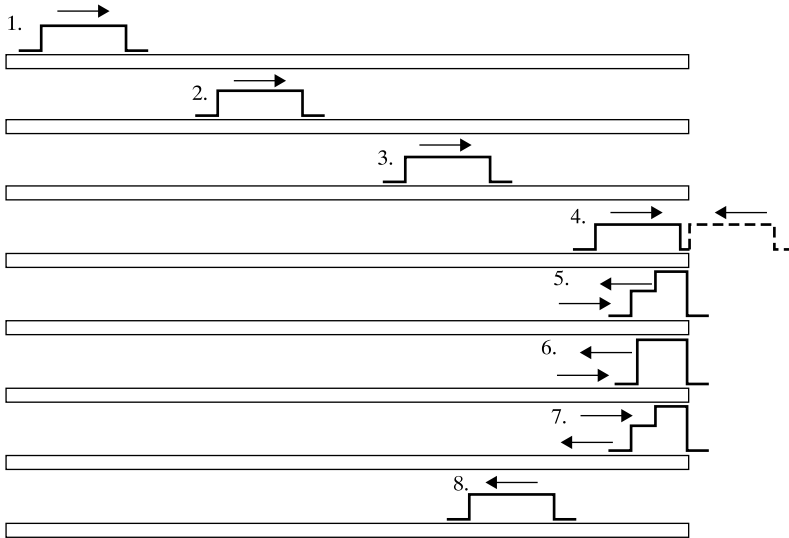
Figure 4.3-1 Pulse on lossless transmission line and its reflection from the open-circuited end.

have the same polarity as the incident voltage, which results in a doubling of the voltage at the end of the line.

The total voltage on the line is equal to the sum of the incident and reverse-going pulses (sketches 4 to 7 of Fig. 4.3-1). Since the reflected voltage wave has the same amplitude as the incident wave, *reflection from an open-circuit terminated line produces a doubling of the voltage at the end of the line as the voltages of the incident and reflected waves add together*. The current at the end of the line is the sum of an incident current pulse and an equal magnitude but oppositely directed reflected current pulse, resulting in a net current of zero. We would expect this at an open-circuit termination.

After reflection at the open-circuit end of the line, the pulse travels back toward its source. At the source the pulse could again reflect if the source impedance is not equal to the line's characteristic impedance Z_0. This pulse example is useful in demonstrating physically how the doubling of voltage occurs at the open-circuit termination of a transmission line. Similar reasoning shows that the current at a short-circuit termination of the transmission line produces a doubling of current at the line's end and a zero voltage.

4.4 INCIDENT AND REFLECTED WAVES

Just as pulses on a transmission line that is open circuited produce a doubling of the incident voltage, so also do continuous sine waves when applied to the same transmission line termination condition.
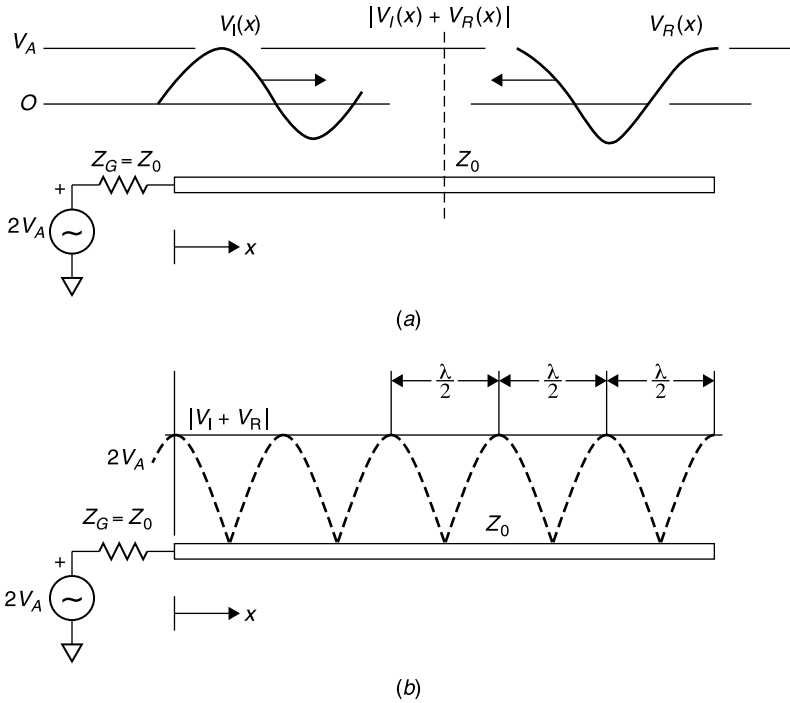
Figure 4.4-1   (*a*) Incident and reflected sine waves on open-circuit terminated line and (*b*) the peak voltage magnitude as function of position *x* on the line.

In this case, however, the sine wave, of voltage V_I, is continuously applied at the source end. The reflected voltage V_R, an identical sine wave propagating back toward the generator, combines with it to produce a *standing-wave pattern*, creating voltage maxima when V_I and V_R are in phase and minima when they are out of phase (Fig. 4.4-1). The peak voltage magnitude of the standing-wave pattern is shown in Figure 4.4-1*b*. The *maxima and minima are separated by a half wavelength due to the doubling of their relative velocities resulting from their opposite directions of travel*.

In general, the net voltage at any point on the line is the result of the phasor addition of V_I and V_R at that point. Note that the amplitude of the standing-wave pattern is not a sine wave. It is the envelope of the peak AC voltage as a function of distance along the line. Such voltage (and the associated current) standing waves are analogous to those of waves in water that reflect from a seawall or other reflecting obstacle. In this example of a lossless line terminated in an open circuit, the peak voltage is $2V_A$ and the minimum voltage is zero. At the voltage nulls the current is $2I_A$ where $I_A = V_A/Z_0$. This will be seen from the analysis of transmission lines to follow.

4.5 REFLECTION COEFFICIENT

In the open-circuit load, lossless line case, the magnitudes of incident and re-flected voltages are equal, and at the "load" end of the line they are in phase. This produces the voltage doubling there. But for arbitrary loads, V_I and V_R are neither equal in magnitude nor do they have the same phase. We define a complex *reflection coefficient* at any point, x, on the line as

$$\Gamma(x) = \frac{V_R(x)}{V_I(x)} = \rho \angle \varphi \tag{4.5-1}$$

where V_R and V_I are reflected and incident phasor voltages, respectively.

On a lossless line the respective magnitudes of V_I and V_R do not change, thus *the magnitude of the reflection coefficient, ρ, does not change either.* Only the angle φ changes as the two waves travel by each other in opposite directions; φ depends upon the relative phases of V_R and V_I at the load as well as the electrical distance from the load.

For example, returning to the open-circuited case in Figure 4.4-1, since $\varphi_{\text{LOAD}} = 0°$, $V_{\text{LOAD-INC}} + V_{\text{LOAD-REF}} = 2$, a voltage *peak* occurs at the load. On the other hand, if there is a short circuit at the load, the total voltage at the end of the line must be zero. Then $\varphi_{\text{LOAD}} = 180°$, $V_{\text{LOAD-INC}} + V_{\text{LOAD-REF}} = 0$, and there is a voltage *null* at the load position.

In general, a line is terminated in some complex load, and the incident and reflected voltages have neither the same magnitudes nor are they precisely in or out of phase at the line's terminus. The magnitudes of the voltage peak and null values are

$$V_{\text{MAX}} = V_{\text{Incident}}(1 + \rho) \tag{4.5-2a}$$

$$V_{\text{MIN}} = V_{\text{Incident}}(1 - \rho) \tag{4.5-2b}$$

where ρ is the ratio of the magnitude of the reflected wave to that of the inci-dent wave. One of the earliest microwave measurement techniques consists of cutting a narrow longitudinal slot in a transmission line and moving a rectify-ing diode along the line to determine $|V_{\text{MAX}}|$ and $|V_{\text{MIN}}|$ as well as the loca-tions of these maximum and minimum voltages relative to the load position. This is called a *slotted line measurement*. Prior to the development of the network analyzer (about 1965) this was the only practical way to measure mi-crowave impedance. Special microwave oscillators having a 1-kHz amplitude modulation were used in conjunction with *standing-wave ratio (SWR) meters*, consisting of tuned and calibrated 1-kHz amplifiers, to make the measurement more sensitive and precise. The ratio of the maximum and minimum voltage amplitudes was and continues to be called the *voltage standing-wave ratio*, or *VSWR*, sometimes just SWR:

$$\text{VSWR} = \frac{|V_{\text{MAX}}|}{|V_{\text{MIN}}|} = \frac{1 + \rho}{1 - \rho} \qquad (4.5\text{-}3a)$$

and conversely

$$\rho = \frac{\text{VSWR} - 1}{\text{VSWR} + 1} \qquad (4.5\text{-}3b)$$

4.6 RETURN LOSS

Relative to the incident power, return loss is the fraction of power returned from the load:

$$\text{Return loss} = \rho^2 \text{ (fraction)} = 20 \log \rho \text{ (dB)} \qquad (4.6\text{-}1)$$

Within the industry, VSWR and return loss are used as alternative means for specifying match, even though the slotted line is rarely if ever used for measurements. A perfect match of the load to the transmission line occurs when

$$Z_L = Z_0 \qquad (4.6\text{-}2)$$

This condition produces no reflected wave, hence $\rho = 0$ and VSWR $= 1$. Then all incident power is absorbed in the load; no power is reflected, and, when expressed in decibels, the return loss is $-\infty$.

4.7 MISMATCH LOSS

Relative to the incident power, mismatch loss is the fraction of power absorbed (not returned) from the load:

$$\text{Mismatch loss} = 1 - \rho^2 \text{ (fraction)} = 10 \log(1 - \rho^2) \text{ (dB)} \qquad (4.7\text{-}1)$$

In fractional form, return loss and mismatch loss sum to unity. Care should be taken to study these definitions. It is a common error to confuse them. Table 4.7-1 lists values of return loss and mismatch loss for various values of the reflection coefficient, ρ.

   As an example, if the VSWR of a load is 1.5, ρ *is 0.2*, return loss is 0.04, and mismatch loss is 0.96. This means that the load absorbs 96% of the incident power and 4% is reflected. In decibels, the return loss is 14 dB and the mismatch loss is 0.18 dB. Notice that the minus sign for the losses in decibels is dropped conversationally because it is implied in the term "loss." However, *when using a network simulator, losses in decibels must be entered as negative quantities.*

TABLE 4.7-1   Relationships Among Reflection Coefficient Magnitude, ρ, VSWR, Return Loss and Mismatch Loss

ρ	VSWR $\dfrac{1+\rho}{1-\rho}$	Return Loss Fraction ρ^2	Return Loss (dB) $20\log\rho$	Mismatch Loss Fraction $1-\rho^2$	Mismatch Loss (dB) $10\log(1-\rho^2)$
0	1.0	0	$-\infty$	1.0	0
0.01	1.02	0.0001	-40.0	0.9999	-0.0004
0.02	1.04	0.0004	-34.0	0.9996	-0.0017
0.05	1.11	0.0025	-26.0	0.9975	-0.011
0.10	1.22	0.01	-20.0	0.99	-0.044
0.15	1.35	0.0225	-16.5	0.9775	-0.099
0.2	1.5	0.04	-14.0	0.96	-0.18
0.3	1.86	0.09	-10.5	0.91	-0.41
0.33	2	0.11	-9.6	0.89	-0.50
0.4	2.3	0.16	-8.0	0.84	-0.76
0.5	3	0.25	-6.0	0.75	-1.25
0.6	4	0.36	-4.4	0.64	-1.94
0.7	5.7	0.49	-3.1	0.51	-2.92
0.707	5.8	0.50	-3.01	0.50	-3.01
0.8	9	0.64	-1.94	0.36	-4.44
0.9	19	0.81	-0.92	0.19	-7.21
1.0	∞	1.00	0.0	0.0	$-\infty$

4.8   MISMATCH ERROR

In Sections 4.6 and 4.7 return loss and mismatch loss were defined in terms of reflections at a mismatched load. However, these terms also can be applied to the input of a two-port network terminated by a matched or mismatched load. For example, a reactance connected in shunt with an otherwise match-terminated transmission line produces a reflection. For this circuit return loss and mismatch loss can be calculated.

Now, if there are two or more sources of reflection on the line (Fig. 4.8-1), the resulting reflections can combine to produce an overall reflection that depends not only upon the individual reflections but their reflective interactions as
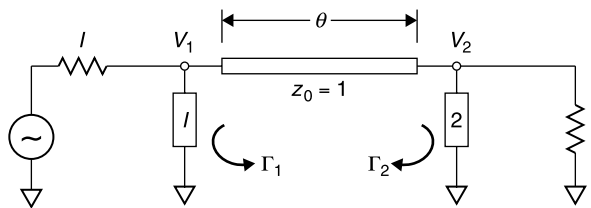
Figure 4.8-1   Interaction between two reflecting networks separated by transmission line length.

well. To evaluate this combination, suppose two circuits, having Γ_1 and Γ_2 complex reflection coefficients are interconnected by a lossless transmission line of electrical length θ. For this discussion it is assumed that generator and load are matched to each other and to the characteristic impedance of the line, in this case normalized to $Z_0 = 1$.

Due to the multiple reflections between them and their spacing, the obstacles will interact in a manner that causes their total insertion loss to vary. The loss of the combination may be higher or lower than their *simple loss*, that is, *the sum of their individual mismatch losses* that would be measured when they are separately connected between generator and load.

The amount by which the actual loss differs from the simple loss is called the *mismatch error*, usually expressed in decibels. The term *mismatch error* arises because, if we ignore their interaction and estimate their combined loss to be their *simple loss*, we would encounter an error equal to this mismatch error value. To analyze this reflection interaction it is useful to define the *transmission coefficient T*.

At the point of reflection on a transmission line, we defined an incident voltage V_I and a reflected voltage V_R. The phasor sum of these two voltages is the *total voltage* V_T. This is the voltage on the line at the point of reflection:

$$V_T = V_I + V_R \tag{4.8-1}$$

The transmission coefficient T is the ratio of V_T to V_I, just as the reflection coefficient Γ is the ratio of V_R to V_I. Then it follows that

$$T = 1 + \Gamma \tag{4.8-2}$$

where both T and Γ are complex numbers. Note that T can be greater than unity. In fact, this occurs at the end of an open-circuited transmission line at which $\Gamma = 1$ and $T = 2$.

The situation when two sources of reflection exist on a transmission line is shown in Figure 4.8-2. An incident voltage $V_1 = 1$ encounters the first obstacle and the voltage at the first obstacle is T_1. This voltage proceeds on toward the second obstacle. When T_1 reaches the second obstacle, it is further affected. In
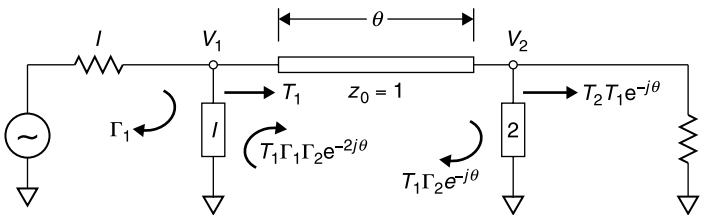
Figure 4.8-2   The reflection interaction between two circuits separated by lossless line length [1, p 407].

the absence of reflections between the obstacles, the resulting voltage at the second obstacle would be V_2 (*expected*). This voltage launches a wave traveling toward the load as

$$V_2 \text{ (expected)} = T_1 T_2 e^{-j\vartheta} \tag{4.8-3}$$

The "expected" insertion loss (IL), or simple loss, of the two obstacles, neglecting their reflective interactions, would be the sum of their individual mismatch losses or

$$\text{IL (expected)} = 20 \log|T_1|^{-1} + 20 \log|T_2|^{-1} \tag{4.8-4}$$

where the negative exponents are used in order that IL is a positive number when expressed in decibels.

However, when T_1 reaches the second obstacle, a part of its energy is reflected by it. This reflected wave travels back toward the first obstacle at which a third reflection occurs. The process continues in a manner similar to that of the multiple reflections occurring between facing, partially reflecting mirrors, creating an infinite series of reflections and re-reflections.

The initial waves and the multiple reflections are as described in Figure 4.8-2. Each time the reflections make a round trip between obstacles, they gain an additional phase change of $e^{-2j\theta}$ as well as the factor $\Gamma_1 \Gamma_2$. Each such round trip path adds another term to the expression for the V_2 wave traveling toward the load. The result is an infinite series of terms:

$$\frac{V_2}{V_1} = T_1 T_2 e^{-j\theta} [1 + \Gamma_1 \Gamma_2 e^{-2j\theta} + (\Gamma_1 \Gamma_2 e^{-2j\theta})^2 + \cdots] \tag{4.8-5}$$

The magnitude of the quantity outside the brackets, $|T_1 T_2|$, is the simple loss ratio. The quantity inside the square brackets, $1 + \Gamma_1 \Gamma_2 e^{-2j\theta} + (\Gamma_1 \Gamma_2 e^{-2j\theta})^2 + \cdots$, is the effect of the reflective interaction between the two obstacles on the transmission line. Fortunately, this infinite series has a closed form [2, pp. 2–3], namely

$$1 + x + x^2 + x^3 + \cdots = \frac{1}{1 - x} \quad \text{for } |x| < 1 \tag{4.8-6}$$

Because $|\Gamma_1 \Gamma_2 e^{-2j\theta}| < 1$ for passive obstacles, (4.8-5) reduces to

$$\frac{V_2}{V_1} = T_1 T_2 e^{-j\theta} [1 + \Gamma_1 \Gamma_2 e^{-2j\theta}]^{-1} \tag{4.8-7}$$

Since insertion loss is $\text{IL} = 20 \log|V_1/V_2|$, the actual insertion loss for the circuit in Figure 4.8-2 is

$$\text{IL (actual)} = 20 \log|T_1|^{-1} + 20 \log|T_2|^{-1} + |\text{ME}| \text{ (dB)} \tag{4.8-8}$$

where ME is called the *mismatch error* and given by

$$ME = 20 \log_{10} |1 - \Gamma_1 \Gamma_2 e^{-j2\theta}| \ (dB) \qquad (4.8\text{-}9)$$

The mismatch error can either *increase* or *decrease* the base insertion loss of the two obstacles. The extreme values for ME are given, respectively, by using the $+$ and $-$ options. That is,

$$ME \ (\text{extreme values}) = 20 \log[1 \pm \rho_1 \rho_2] \qquad (4.8\text{-}10)$$

where ρ_1 and ρ_2 are the magnitudes of the reflection coefficients.

   The larger the product of the reflection coefficients, the larger the mismatch error, and accordingly the larger the spread between maximum and minimum mismatch loss values. Customarily, *it is usually the extreme mismatch error values that are called the mismatch errors* because they are the error limits that can be experienced by ignoring the interaction of two reflections spaced along a transmission line. Notice that the *mismatch error occurs between each pair of obstacles*. If there are three obstacles spaced along a transmission line, there can be two separate mismatch errors.

   A mismatch error can occur between a practical generator (or measurement system) and the reflections of a device under test if the generator is not perfectly matched to the transmission line. Mismatch errors also occur between connectors when neither has a unity VSWR and/or when one or more imperfectly matched connectors interact with an imperfectly matched device to which they are attached. In summary *a mismatch error can arise whenever two causes of reflection are spaced along a transmission system*.

   Mismatch errors are especially large with cascaded reactive filters because these circuits accomplish their filtering action by reflecting power in the stopband. For example, suppose that two reactive filters are connected in cascade, each having 20 dB of isolation (when installed between matched source and load) at a frequency we wish to block. We would expect, neglecting reflective interaction, to obtain 40 dB of isolation in the stopband, the sum of the separate isolation values of the two filters. However, to provide 20 dB of isolation reactively, each filter must have $\rho = 0.995$, or $\rho^2 = 0.99$. For the pair of filters the mismatch error limits are

$$ME_+ = 20 \log(1 + 0.99) = +6 \ dB \qquad (4.8\text{-}11a)$$

$$ME_- = 20 \log(1 - 0.99) = -40 \ dB \qquad (4.8\text{-}11b)$$

   In the first case, ME_+ has the same sign as the base losses. Thus, under ideal spacing of the filters, we can obtain 46 dB of stopband isolation. This is 6 dB greater isolation than expected from the simple addition of the individual isolation values. However, in the second case, ME_-, which occurs with the least optimal spacing, *placing two lossless, reactive 20-dB filters in cascade may yield*

0 dB of total attenuation. Nor is this unfavorable spacing unlikely since the stopband may represent a very wide frequency bandwidth over which the least favorable electrical spacing is likely to occur at various stopband frequencies.

Of course, real filters, even if designed as reactive networks, have finite loss, and the observation of 0 dB loss for a pair of them is a limiting, but unrealizable condition. Nevertheless, the calculation shown above suggests that a substantial reduction in isolation from that expected when ignoring VSWR interaction is readily possible, dependent upon their electrical spacing at each stopband frequency.

Interestingly, there is another way of treating an interacting pair of mismatches using their VSWR values [3, Example 4.5, 8]. The overall VSWR due to a pair of dissipationless mismatches having individual VSWR values SWR_1 and SWR_2 (where $SWR_1 > SWR_2$) that are referenced to and separated by a length of lossless Z_0 line can range from SWR_1/SWR_2 to $(SWR_1)(SWR_2)$.

Applying this to the previous example, for a reactive filter providing 20 dB of isolation, $1/(1 - \rho^2) = 100$, therefore $\rho = 0.994987$ and the corresponding $SWR_1 = SWR_2 = 397.997487$. Since the VSWRs are equal, $SWR_1/SWR_2 = 1$ and their minimum insertion loss is 0 dB. The maximum $VSWR = (397.99)^2 = 158,402$. The corresponding $\rho = 0.999987$. The return loss is $1/(1 - \rho^2) = 39,601 = 45.98$ dB, which probably agrees with the prior estimates of 0 and 46 dB within the accuracy of the significant figures carried in the calculations.

4.9   THE TELEGRAPHER EQUATIONS

To establish a mathematical basis for the analysis of transmission lines, we shall use a uniform line equivalent circuit that includes line losses, as shown in Figure 4.9-1. Note that, while the series elements are shown only in the upper conductor, this equivalent circuit is applicable to both balanced and unbalanced transmission lines. Also, this analysis applies to any waveforms that propagate on the transmission line, not just sinusoidal excitations.

In this equivalent circuit $R, L, G,$ and C are, respectively, the series resistance, series inductance, shunt conductance, and shunt capacitance per unit
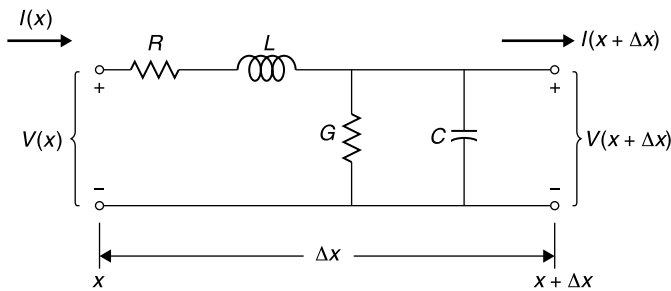
Figure 4.9-1   Equivalent circuit, including losses, for infinitesimally short length of uniform transmission line.

length of the transmission line. The small section of transmission line can be analyzed by applying Kirchhoff's laws and taking the llimit as this circuit length tends to zero to obtain the differential equations for the circuit [4, p. 23]. *Kirchhoff's voltage law* requires that the sum of the voltages about a closed loop is zero. Thus

$$v(x, t) - R\,\Delta x\, i(x, t) - L\,\Delta x \frac{\partial i(x, t)}{\partial t} - v(x + \Delta x, t) = 0 \qquad (4.9\text{-}1)$$

Similarly, applying Kirchhoff's current law, which requires that the sum of the currents into a circuit node be zero, gives

$$i(x, t) - G\,\Delta x\, v(x + \Delta x, t) - C\,\Delta x \frac{\partial v(x + \Delta x, t)}{\partial t} - i(x + \Delta x, t) = 0 \quad (4.9\text{-}2)$$

For the above two equations, divide all terms by Δx and let $\Delta x \to 0$. Note that as $\Delta x \to 0$

$$v(x + \Delta x, t) \to v(x, t)$$
$$i(x + \Delta x, t) \to i(x, t)$$

Also,

$$\frac{v(x + \Delta x, t) - v(x, t)}{\Delta x} \to \frac{\partial v(x, t)}{\partial t}$$
$$\frac{i(x + \Delta x, t) - i(x, t)}{\Delta x} \to \frac{\partial i(x, t)}{\partial t}$$

The resulting differential equations in (4.9-3) are the time-domain form of the transmission line equations and are called the *telegrapher equations*, presumably because their initial use was in the analysis and design of long telegraph lines:

$$\frac{\partial v(x, t)}{\partial x} = -Ri(x, t) - L \frac{\partial i(x, t)}{\partial t} \qquad (4.9\text{-}3\text{a})$$

$$\frac{\partial i(x, t)}{\partial x} = -Gv(x, t) - C \frac{\partial v(x, t)}{\partial t} \qquad (4.9\text{-}3\text{b})$$

4.10   TRANSMISSION LINE WAVE EQUATIONS

These telegrapher equations apply for any time-varying $v(x, t)$ and $i(x, t)$ on the transmission line. In the steady-state sinusoidally excited case, they reduce to

$$\frac{dV(x)}{dx} = -(R + j\omega L)I(x) \qquad (4.10\text{-}1a)$$

$$\frac{dI(x)}{dx} = -(G + j\omega C)V(x) \qquad (4.10\text{-}1b)$$

in which R, L, G, and C are per unit length and $V(x)$ and $I(x)$ are the phasor forms of $v(x, t)$ and $i(x, t)$, whose time variation is implicit. These two equations can be solved simultaneously to yield two equations, one a function of only V and the other a function of only I. To do so, for example, differentiate the second equation with respect to x to get

$$\frac{d^2 I(x)}{dx^2} = -(G + j\omega C)\frac{dV(x)}{dx}$$

or

$$\frac{dV(x)}{dx} = -\frac{1}{G + j\omega C}\frac{d^2 I(x)}{dx^2} \qquad (4.10\text{-}2)$$

Then substitute this value for $dV(x)/dx$ into (4.10-1a).

$$\frac{d^2 I(x)}{dx^2} - (R + j\omega L)(G + j\omega C)I(x) = 0$$

or

$$\frac{d^2 I(x)}{dx^2} - \gamma^2 I(x) = 0 \qquad (4.10\text{-}3a)$$

Using the same procedure, (4.10-1b) yields

$$\frac{d^2 V(x)}{dx^2} - \gamma^2 V(x) = 0 \qquad (4.10\text{-}3b)$$

in which we define

$$\gamma \equiv \sqrt{(R + j\omega L)(G + j\omega C)} \qquad (4.10\text{-}4)$$

Solving differential equations involves removing the differentials (integrating) by guessing or recognizing the solution from prior differentiation experience. The only functions that are unchanged after double differentiation are the exponentials, $e^{\pm\gamma x}$. (A close possibility would have been $\sin \gamma x$, however, twice differentiating $\sin \gamma x$ would give *minus* $\sin \gamma x$.) A second-order differential equation such as this should have two independent solutions. In this case the

two solutions correspond to the two values for the $\pm$ sign, they are

$$\frac{d^2 e^{\gamma x}}{dx^2} = \gamma^2 e^{\gamma x} \quad \text{and} \quad \frac{d^2 e^{-\gamma x}}{dx^2} = \gamma^2 e^{-\gamma x}$$

Accordingly, the sum of these two solutions represents the complete solution to either (4.10-3a) or (4.10-3b).

$$V(x) = V_I e^{-\gamma x} + V_R e^{+\gamma x} \tag{4.10-5a}$$

$$I(x) = I_I e^{-\gamma x} - I_R e^{+\gamma x} \tag{4.10-5b}$$

where γ is the propagation constant for the voltage and current waves. Use of the negative sign before I_R in (4.10-5b) results in the same value for the reflection coefficient for I as for V. Physically the two terms in each equation correspond to waves propagating, respectively, in the $+x$ and $-x$ directions. These two equations are called the *wave equations* for a uniform transmission line.

4.11 WAVE PROPAGATION

The terms with negative exponents (such as $V_I e^{-\gamma x}$) correspond to a wave propagating in the $+x$ direction, from left to right in Figure 4.9-1. Since the generator is attached at the left and the load at the right in Figure 4.9-1, this corresponds to an *incident wave*, hence the subscript I. Similarly, the positive exponent terms (such as $V_R e^{+\gamma x}$) correspond to a wave traveling in the $-x$ direction and represent a *reflected wave* in our convention. These use the subscript R.

The total voltage at any x location on the line is the phasor sum of the incident and reflected voltages, $V(x) = V_I(x) + V_R(x)$ and $I(x) = I_I(x) - I_R(x)$. We can cast the relationship between $V(x)$ and $I(x)$ into the format of impedance by resorting to their interrelation expressed in either (4.10-1a) or (4.10-1b). For example, invoking (4.10-1a) and substituting into it the value for $V(x)$ given by (4.10-5a) gives

$$I(x) = -\frac{1}{R + j\omega L} \frac{dV(x)}{dx} = -\frac{1}{R + j\omega L} \frac{d(V_I e^{-\gamma x} + V_R e^{+\gamma x})}{dx}$$

$$I(x) = \frac{\gamma}{R + j\omega L} (V_I e^{-\gamma x} - V_R e^{+\gamma x}) \tag{4.11-1}$$

From this we see that the current, $I(x)$ has terms that are of the same form as those of the voltage $V(x)$, but there is a negative sign between them. This is because the current convention for reflected waves reverses while that for voltages does not.

The factor $(R + j\omega L)/\gamma$ has the dimensions of impedance and is called the *characteristic impedance* Z_0 of the transmission line. It can be rewritten as

$$Z_0 = \frac{R + j\omega L}{\gamma} = \frac{R + j\omega L}{\sqrt{(R + j\omega L)(G + j\omega C)}}$$

$$Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}}$$

$$(4.11\text{-}2)$$

With this definition, the equation for $I(x)$ has exactly the form of Ohm's law, noting that the minus sign between terms accounts for the reversal of current direction for a reflected wave. In other words, Z_0 *is the ratio of the respective voltage to current of waves traveling in either direction:*

$$I(x) = I_I(x) - I_R(x) = \frac{V_I}{Z_0} e^{-\gamma x} - \frac{V_R}{Z_0} e^{+\gamma x} \qquad (4.11\text{-}3)$$

Notice that for a lossy line, for which R and G are nonzero, the characteristic impedance is a complex quantity that varies with frequency. However, *for the lossless line, the characteristic impedance is a constant, independent of frequency* and given by

$$Z_0 = \sqrt{\frac{L}{C}} \qquad (4.11\text{-}4)$$

For high frequency, low-loss transmission lines, (4.11-4) is a good approximation, since $\omega L \gg R$ and $\omega C \gg G$. For a transmission line with loss, the propagation constant γ is a complex quantity. Conventionally this is written with real and imaginary parts as

$$\gamma = \alpha + j\beta = \sqrt{(R + j\omega L)(G + j\omega C)} \qquad (4.11\text{-}5)$$

where α and β are, respectively, the attenuation and phase constants of the transmission line with units of reciprocal length. The real term, α, is responsible for loss and is expressed in *nepers/unit length*; and the imaginary quantity, β, is responsible for phase change and is expressed in *radians/unit length*.

For a transmission line having loss, both the incident and reflected waves diminish in amplitude as they travel along the line due to dissipative and/or radiative losses. For an incident wave the amplitude diminishes with x according to the factor $e^{-\alpha x}$. For example, if α were 0.1 nepers/meter, then a wave traveling 1 m would have its amplitude decreased by the factor $e^{-0.1} = 0.9$. That is, the voltage of the traveling wave would be reduced by 10%. Since power is proportional to the square of the voltage, the power loss would be about 19% or 0.8686 dB for this 1-m length of line.

Notice that this example describes the conversion between loss in nepers and decibels, namely Loss (dB) = 8.686 loss (nepers), conversely Loss (nepers) = 0.115 loss (dB). These conversions are listed in the inside cover summaries for convenient reference.

As noted, the imaginary portion of the propagation constant β produces a phase change in the wave as a function of distance. It has the dimensions of *radians/unit length* or, optionally, *degrees/unit length*. An any instant of time, with time frozen so to speak, a snapshot of the voltage amplitude along the line of the incident wave (assuming there is no reflected wave for this example) would be a sinusoid. The distance between adjacent wave crests, or between adjacent troughs, or between any other similar pair of corresponding points of the wave sinusoid is the wavelength λ. Since the sinusoid completes a cycle in 2π radians, it follows that

$$\beta = \frac{2\pi}{\lambda} \text{ radians/unit length} \qquad (4.11\text{-}6a)$$

$$\beta = \frac{360°}{\lambda} \text{ degrees/unit length} \qquad (4.11\text{-}6b)$$

For the lossless line,

$$\alpha = 0 \qquad (4.11\text{-}7a)$$

$$\beta = \omega\sqrt{LC} \qquad (4.11\text{-}7b)$$

Usually one does not know explicitly the distributed L and C values for a transmission line. Conventionally, the transmission line Z_0 is provided, along with the effective relative dielectric constant ε_R of the line. Therefore, it is more convenient to determine electrical length by first determining the wavelength on the line, finding the fraction of a wavelength of the circuit in question, and multiplying either by 2π for length in radians or by $360°$ to obtain electrical length in degrees. This is demonstrated by the following example.

Electrical length can be expressed in either radians or degrees. For example, consider a 6 in. length ($x = 6$ in.) of coaxial cable having a polyethylene insulator between its inner and outer conductors, for which the relative dielectric constant is 2.26. What is the cable's electrical length θ at 1 GHz? Since the wavelength in free space for a 1-GHz sinusoid is 11.8 in.,

$$\theta = \beta x = \frac{x}{\lambda} 360° = \frac{x}{\lambda_0/\sqrt{e_R}} 360° = \frac{6 \text{ in.}}{11.8 \text{ in.}/\sqrt{2.26}} 360° = 275° \quad (4.11\text{-}8)$$

Since 2π radians corresponds to $360°$, the electrical length in radians is

$$\theta = 275° \frac{6.28}{360°} = 4.80 \text{ rad} \qquad (4.11\text{-}9)$$

Notice that it was not necessary to determine β explicitly. However, since the wavelength in the cable at 1 GHz is 7.85 in., then

$$\beta = \frac{360°}{7.85 \text{ in.}} = 45.9°/\text{in. at 1 GHz} \qquad (4.11\text{-}10)$$

Note from (4.11-7b) that β is frequency dependent. In a nondispersive transmission media (for which electrical length is proportional to frequency), β is directly proportional to frequency. Thus, at 1.2 GHz the same cable would have a propagation constant of

$$\beta = \frac{1.2 \text{ GHz}}{1 \text{ GHz}} 45.9°/\text{in.} = 55.1°/\text{in. at 1.2 GHz} \qquad (4.11\text{-}11)$$

and the same 6-in. cable would have an electrical length of 330° or 5.76 rad at 1.2 GHz.

4.12 PHASE AND GROUP VELOCITIES

The incident wave, traveling from the generator toward the load, is, from (4.10-5a)

$$V(x)_{\text{Incident}} = V_I e^{-\gamma x} \qquad (4.12\text{-}1)$$

in which V_I is a phasor with implicit sinusoidal time variation. Written explicitly

$$v(x)_I = \text{Re}[V_I e^{-\gamma x} e^{j\omega t}] = C e^{-\alpha x} \cos(\omega t - \beta x + \varphi) \qquad (4.12\text{-}2)$$

where C is a constant amplitude factor and φ is the initial phase of the wave at $t = x = 0$. This is interpreted as a sinusoidal wave propagating in the $+x$ direction with an exponential attenuation due to the factor α. The phase constant β is given in (4.11-6), namely, $\beta = 2\pi/\lambda$. A similar equation can be written for the reflected wave that travels back toward the generator in the $-x$ direction. The voltage $v(x, t)_I$ has a constant phase for

$$\beta x = \omega t \quad \text{or} \quad x = \frac{\omega t}{\beta} \qquad (4.12\text{-}3)$$

Therefore, a constant phase point on the traveling wave moves at a *phase velocity* given by dx/dt or

$$v_P = \frac{\omega}{\beta} \qquad (4.12\text{-}4)$$

Thus, the voltage on the transmission line for either an incident or reflected wave varies as [4, pp. 45–46]

$$e^{j(\omega t - \beta x)} = e^{j\omega(t - x/v_P)} \tag{4.12-5}$$

Any periodic function of time can be represented as a sum of sinusoidal waves using a Fourier analysis (see Section 7.27). If v_P is the same for all of them, then their addition will faithfully reproduce along the line the temporal wave shape at the input, only delayed by the time of propagation x/v_P. This is the case for a lossless transmission line having $\gamma = j\beta = \omega\sqrt{LC}$ for which

$$v_P = \frac{1}{\sqrt{LC}} \tag{4.12-6}$$

For lossless transverse electromagnetic (TEM) mode transmission lines, the velocity at which a constant phase, v_P, moves on the line and the velocity at which a packet of information, such as a modulation envelope, moves, v_G, are identical and simply equal to the velocity of propagation, v.

   This produces a convenient means for determining the characteristic impedance of the TEM mode for a lossless line. Since

$$Z_0 = \sqrt{\frac{L}{C}}$$

and

$$v = \frac{1}{\sqrt{LC}}$$

Therefore,

$$Z_0 = \frac{1}{vC} \tag{4.12-7}$$

This equation provides a convenient expression for determining the characteristic impedance of a TEM mode, lossless transmission line if its distributed capacitance and velocity of propagation are known.

   However, for transmission lines having losses and for waveguides v_P does vary with frequency, dispersion occurs, and the signal waveform at the input will be distorted as its separate frequency components travel along the line at different velocities. If the dispersion is significant, it may not be meaningful to speak of a single propagation delay for a signal waveform. However, if the dispersion is moderate over the signal frequency bandwidth, an approximate *group velocity* v_G can be defined [4]. To determine its value, suppose that an

incident voltage wave, v_1, consists of two frequencies that are traveling on the line, one lower and one higher in frequency than ω_0, respectively, having a combined value at $x = 0$ given by

$$v_I(x = 0, t) = A[\sin(\omega_0 - d\omega)t + \sin(\omega_0 + d\omega)t] \qquad (4.12\text{-}8)$$

Then, at a general point x on the line

$$v_I(x, t) = A\{\sin[(\omega_0 - d\omega)t - (\beta_0 - d\beta)x] + \sin[(\omega_0 + d\omega)t - (\beta_0 + d\beta)x]\}$$

in which β is assumed to be a function of ω and hence is the reason for the differential $d\beta$. Rearranging terms in the arguments gives

$$v_I(x, t) = A\{\sin[(\omega_0 t - \beta_0 x) - (d\omega\, t - d\beta\, x)] + \sin[(\omega_0 t - \beta_0 x) + (d\omega\, t - d\beta\, x)]\}$$
$$(4.12\text{-}9)$$

Next, recognizing that the two sine terms are of the form $\sin(A - B)$ and $\sin(A + B)$, expanding them and simplifying gives

$$v_I(x, t) = 2A \cos(d\omega\, t - d\beta\, x) \sin(\omega_0 t - \beta_0 x) \qquad (4.12\text{-}10)$$

This reveals that the total voltage on the line corresponds to a high frequency sinusoid varying at ωt rate whose amplitude is modulated by

$$\cos[(d\omega)t - (d\beta)x]$$

which varies with both time and distance. The modulation, itself, has the properties of a traveling wave. A constant phase of the modulation occurs for

$$(d\omega)t = (d\beta)x \qquad (4.12\text{-}11)$$

from which a group velocity, $v_G = x/t$, can be defined as

$$v_G = \frac{d\omega}{d\beta} \qquad (4.12\text{-}12)$$

This can be related to the phase velocity $v_P = \omega/\beta$ as follows:

$$v_G = \frac{d\omega}{d\beta} = \frac{d\omega}{d(\omega/v_P)} = \frac{1}{(d/d\omega)(\omega/v_P)} = \frac{1}{[v_P - \omega(dv_P/d\omega)]/v_P^2}$$
$$v_G = \frac{v_P}{1 - (\omega/v_P)(dv_P/d\omega)} \qquad (4.12\text{-}13)$$

This result was derived for only two frequencies, but using similar reasoning it can be argued that a signal comprising many individual frequencies will travel with an envelope velocity approximately equal to v_G, if the dispersion is not

too great. However, this will not be true for lines having large dispersion, but in those cases it is inappropriate to assign any velocity to such a modulation because its shape would be changing so rapidly as it propagates. The concept of group velocity also is not meaningful when $dv_P/d\omega$ is positive, a situation called *anomalous dispersion*, [4, p. 48] because then, as can be seen from (4.12-13) group velocity would appear to be infinite or negative.

4.13   REFLECTION COEFFICIENT AND IMPEDANCE

We can evaluate the reflection coefficient in terms of the load impedance by dividing (4.10-5a) by (4.11-3). For this evaluation let us choose $x = 0$ at the load, then at the load location the exponential functions equal one and the load impedance Z_L is given by

$$Z_L = \frac{V(0)}{I(0)} = Z_0 \frac{V_I + V_R}{V_I - V_R} = Z_0 \frac{1 + V_R/V_I}{1 - V_R/V_I} \tag{4.13-1}$$

where V_I and V_R are phasor quantities. But $\Gamma(x) = V_R(x)/V_I(x)$ is the definition of the reflection coefficient and therefore

$$Z_L = Z_0 \frac{1 + \Gamma_L}{1 - \Gamma_L} \tag{4.13-2}$$

where, in general, Γ and Γ_L are complex numbers. Equivalently,

$$\Gamma_L = \frac{Z_L - Z_0}{Z_L + Z_0} = \frac{Y_0 - Y_L}{Y_0 + Y_L} \tag{4.13-3}$$

where $Y_0 = 1/Z_0$. From this we see *that there will be no reflected wave* if $Z_L = Z_0$, that is, when the load impedance is equal to the characteristic impedance of the line. Under this condition, all of the power incident on the line will be absorbed in the load.

Since the choice of $x = 0$ is arbitrary, (4.13-2) must apply for any x. It follows that

$$Z(x) = Z_0 \frac{1 + \Gamma(x)}{1 - \Gamma(x)} \tag{4.13-4}$$

and

$$\Gamma(x) = \frac{Z(x) - Z_0}{Z(x) + Z_0} = \frac{Y_0 - Y(x)}{Y_0 + Y(x)} \tag{4.13-5}$$

for all x.

4.14   IMPEDANCE TRANSFORMATION EQUATION

One of the most common tasks in microwave engineering is the determination of how a load impedance Z_L is transformed to a new input impedance Z_{IN} by a length of uniform transmission line of characteristic impedance Z_0 and electrical length θ (Fig. 4.14-1).

To simplify this derivation, we assume that the line length is lossless. With the choice of $x = 0$ at the load, the input to the line is at $x = -l$, and the input impedance there is

$$Z_{IN} = Z(x = -l) = \frac{V(x = -l)}{I(x = -l)} = Z_0 \left[\frac{e^{j\beta l} + \Gamma_L e^{-j\beta l}}{e^{j\beta l} - \Gamma_L e^{-j\beta l}}\right] \qquad (4.14\text{-}1)$$

Now substitute $\Gamma_L = (Z_L - Z_0)/(Z_L + Z_0)$, $\beta l = \theta$, the identities $e^{j\beta l} = \cos \beta l + j \sin \beta l$ and $e^{-j\beta l} = \cos \beta l - j \sin \beta l$ into (4.14-1), and remove canceling terms to get

$$Z_{IN} = Z_0 \left[\frac{2Z_L \cos \theta + j2Z_0 \sin \theta}{2Z_0 \cos \theta + j2Z_L \sin \theta}\right]$$

$$Z_{IN} = Z_0 \left[\frac{Z_L + jZ_0 \tan \theta}{Z_0 + jZ_L \tan \theta}\right] \qquad (4.14\text{-}2)$$

Similar reasoning can be used to evaluate the input impedance when the transmission line has finite losses. The result is

$$Z_{IN} = Z_0 \left[\frac{Z_L + Z_0 \tanh \gamma l}{Z_0 + Z_L \tanh \gamma l}\right] \qquad (4.14\text{-}3)$$

This is one of the most important equations in microwave engineering and is called *the impedance transformation equation*. Remarkably, (4.14-2) and (4.14-3) have exactly the same format when derived in terms of admittance. For the lossless line

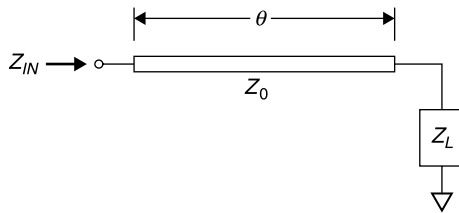$$Y_{IN} = Y_0 \left[\frac{Y_L + jY_0 \tan \theta}{Y_0 + jY_L \tan \theta}\right] \qquad (4.14\text{-}4)$$

Figure 4.14-1   Equivalent circuit of lossless uniform transmission line of electrical length θ and characteristic impedance Z_0 terminated in impedance Z_L.

and for the line with loss

$$Y_{IN} = Y_0 \left[\frac{Y_L + Y_0 \tanh \gamma l}{Y_0 + Y_L \tanh \gamma l} \right] \tag{4.14-5}$$

where $Y_0 = 1/Z_0$, $Y_L = 1/Z_L$, and $Y_{IN} = 1/Z_{IN}$. Expressions (4.14-4) and (4.14-5) can be verified by substituting these equivalences, respectively, into (4.14-2) and (4.14-3).

In general, even the lossless expressions of (4.14-2) and (4.14-4) are complex to apply, as will be seen by a subsequent example. Beside the fact that when $Z_L = Z_0$ the line is always matched and $Z_{IN} = Z_0$, there are four additional cases that are both easy to apply and worthy of note.

Case 1   *Lossless lines of zero length or multiples of a half wavelength present the load impedance unchanged at their input terminals:*

$$Z_{IN} = Z_L \tag{4.14-6}$$

If the transmission line's length is zero or an integer number of half wavelengths $(0°, 180°, 360°, 540° \ldots)$, then $\tan \theta = 0$, and (4.14-2) reduces to $Z_{IN} = Z_L$. That is, whatever impedance terminates, the line will be presented unchanged at the input.

Case 2   *A short-circuit-terminated lossless transmission line is inductive for θ up to 90°, becomes an open circuit when $\theta = 90°$, and is capacitive for $90° < \theta < 180°$. Thereafter the behavior is repeated every half wavelength. For the short-circuit termination (4.14-2) reduces to*

$$Z_{IN} = jZ_0 \tan \theta \tag{4.14-7a}$$

A plot of the input reactance $(Z_0 \tan \theta)$ for $Z_0 = 1$ is shown in Figure 4.14-2. The line length is chosen to be 90° at 1000 MHz. Notice that when the electrical length of the line is 45° (at 500 MHz) the reactance is inductive and numerically equal to Z_0, while at 135° (1500 MHz) it is capacitive and the reactance is numerically equal to $-Z_0$.

The fact that a shorted quarter wavelength line $(\theta = 90°)$ becomes an open circuit (at 1000 MHz in Fig. 4.14-2) is useful in rotary joints for antennas, bias injection and filter networks, and a variety of other applications.

Notice also that if the shorted line section is short compared to a wavelength, $\tan \theta \approx \theta$ (in radians) and (4.14-2) can be approximated as

$$Z_{IN} \approx jZ_0\theta = j\omega \left(\frac{Z_0 l}{v} \right) \tag{4.14-7b}$$

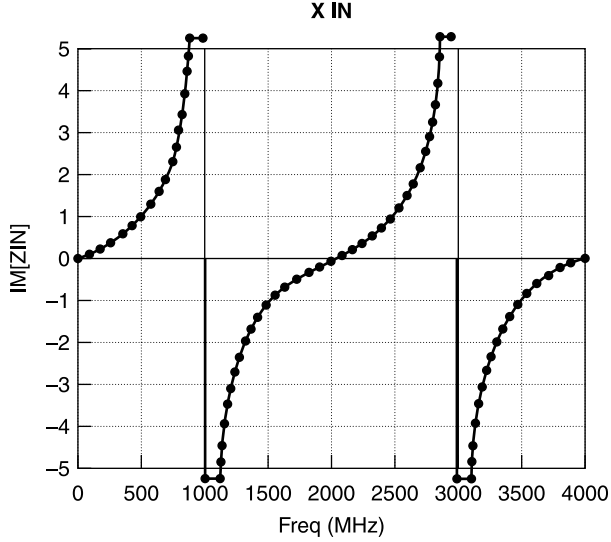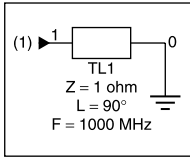since $\theta = \beta l = (2\pi/\lambda)l = 2\pi(f/v)l = \omega l/v$. For most transmission lines θ is

Figure 4.14-2 Reactance of lossless short-circuited transmission line versus frequency ($\theta = 90°$ at 1 GHz).

linearly proportional to frequency, as is the impedance of an inductor, $Z = j\omega L$. Therefore *a short length of short-circuited terminated transmission line behaves as an inductor of value*

$$L_{\text{EFF}} \approx \frac{\theta Z_0}{\omega} = \frac{Z_0 l}{v} \quad (\theta \leq 0.53 \text{ rad}) \tag{4.14-7c}$$

For example, if a 1-cm length of $Z_0 = 120 \ \Omega$, air coax line is terminated in a short:

$$L_{\text{EFF}} \approx \frac{Z_0 l}{v} = \frac{120(1 \text{ cm})}{3 \times 10^{10} \text{ cm/s}} = 4 \text{ nH}$$

Since 1 cm $\approx$ 0.4 in., the inductance of a high impedance air-dielectric line is approximately equal to 10 nH/in., the approximation presented earlier in (2.12-6).

The approximation (4.14-7c) also applies to a short length of Z_0 transmission line terminated in a load Z_L provided that its characteristic impedance $Z_0 \gg |Z_L|$, as can be verified by applying this condition and $\tan \theta \approx \theta$ in (4.14-2).

The approximation $\tan \theta \approx \theta$ occurs frequently and has an error magnitude of less than 10% for values of θ up to 0.524 rad (30°). A radian is $360°/2\pi \approx 57.3°$. Thus 0.5 rad is nearly $\frac{1}{12}$ of a wavelength. The values of θ and

TABLE 4.14-1   Approximating tan θ by Its Argument θ (in radians)

θ (deg)	0	10	20	25	30
θ (rad)	0	0.175	0.349	0.436	0.524
tan θ	0	0.176	0.364	0.466	0.577
Error in tan θ ≈ θ	0	−0.6%	−4.1%	−6.4%	−9.1%

tan θ are compared in Table 4.14-1 along with the size of the error in using this approximation, showing that the error magnitude is less than 10% at $\theta = 30°$ and diminishes rapidly for lesser values of θ.

Case 3 *An open-circuit terminated lossless transmission line is capacitive for θ up to 90°, becomes a short circuit when $\theta = 90°$, and is inductive for $90° < \theta < 180°$. Thereafter the behavior is repeated every half wavelength. Thus, from (4.14-4) with $Y_L = 0$*

$$Y_{IN} = jY_0 \tan \theta \quad \text{or} \quad Z_{IN} = -jZ_0 \cot \theta \qquad (4.14\text{-}8a)$$

This is plotted in Figure 4.14-3. Notice that the input reactance magnitude is capacitive and equal to $-Z_0$ when the line is 45° long (500 MHz). For $\theta = 90°$ (1000 MHz in Fig. 4.14-3) the open-circuit load is transformed to a short circuit at the input. It is inductive and equal to Z_0 when 135° long (1500 MHz). The input repeats every half wavelength after 180°. The facility for transforming an open circuit to a short circuit is equally useful in a variety of microwave circuits.
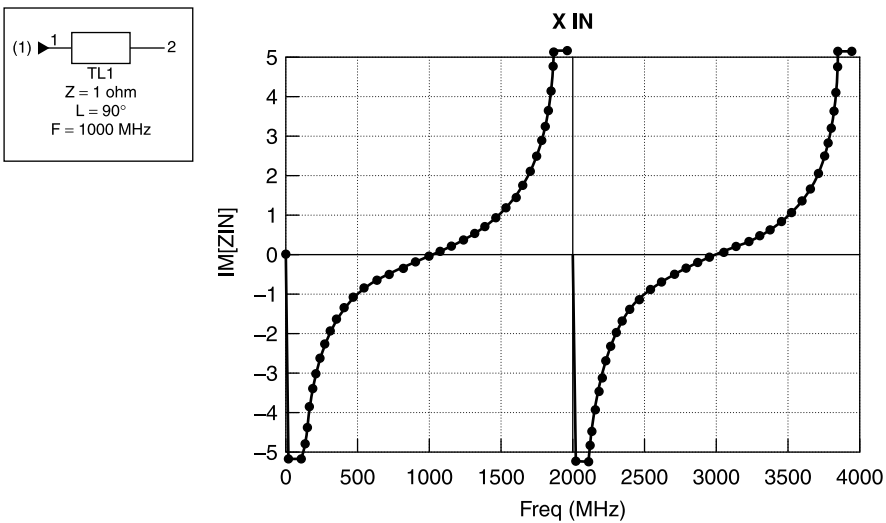

Figure 4.14-3 Normalized reactance of lossless open-circuited transmission line versus frequency (electrical length = 90° at 1 GHz).

By using similar reasoning to that applied to a short section of a shorted transmission line, and again applying the approximation $\tan\theta \approx \theta$ for small values of θ, a short section of open-circuited transmission line has a capacitive admittance given by

$$Y_{\text{IN}} \approx jY_0\theta \tag{4.14-8b}$$

Since $B_C = \omega C$ and $\theta = \omega l/v$, the equivalent capacitance is

$$C_{\text{EFF}} \approx \frac{\theta}{\omega Z_0} = \frac{l}{Z_0 v} \quad (\theta \le 0.53 \text{ rad}) \tag{4.14-8c}$$

and a reactance of

$$\frac{1}{\omega C} \approx \frac{Z_0}{\theta} \tag{4.14-8d}$$

For example, a section of 25-Ω line that is 20° long at 1 GHz has an equivalent capacitance of

$$C_{\text{EFF}} \approx \frac{0.35 \text{ rad}}{6.28 \times 1 \times 10^9 \text{ Hz} \times 25 \ \Omega} = 2.2 \text{ pF}$$

Its reactance X_C at 1 GHz is

$$\frac{1}{\omega C_{\text{EFF}}} \approx \frac{25 \ \Omega}{0.35} = 71 \ \Omega$$

If a short length of Y_0 line is connected to a load of admittance Y_L, it behaves as a shunt capacitance approximated by (4.14-8c) provided that $|Y_L| \ll Y_0$ (i.e., $|Z_L| \gg Z_0$) and $\tan\theta \approx \theta$.

Case 4 *The normalized input impedance to a quarter wavelength long transmission line is the reciprocal of the normalized load impedance.*

When the transmission line is a quarter wavelength long, $\theta = 90°$, and the input impedance to the line is

$$Z_{\text{IN}} = Z_0\frac{Z_0}{Z_L} \tag{4.14-9}$$

Normalizing to Z_0 (dividing all impedances by Z_0),

$$z_{\text{IN}} = \frac{1}{z_L} \tag{4.14-10}$$

For this reason a quarter wavelength line is often called an *impedance inverter*. Notice that the normalized input impedance is equal to the normalized load admittance, that is,

$$z_{\text{IN}} = \frac{1}{z_L} = y_L \qquad (4.14\text{-}11)$$

where the normalized admittance $y_L = Y_L/Y_0$ and $Y_0 = 1/Z_0$.

Later we will see that this property allows us to convert between admittance and impedance using the Smith chart. Notice also that this impedance inverting property occurs not only for quarter wavelength length lines but for line lengths that are any odd multiple of a quarter wavelength $(\lambda/4, 3\lambda/4, 5\lambda/4, \ldots)$.

In general, to transform any resistive load R_L to a desired input resistance R_{IN}, one need only select a quarter wave line section having a Z_0 that satisfies

$$Z_0 = \sqrt{R_{\text{IN}}R_L} \qquad (4.14\text{-}12)$$

As an example of the resistive transformation of a 90° line length, consider a 50-Ω, quarter wavelength, transmission line terminated in 25 Ω. From (4.14-9) the input resistance is

$$Z_{\text{IN}} = Z_0 \frac{Z_0}{Z_L} = 50 \frac{50}{25} = 100 \ \Omega \qquad (4.14\text{-}13)$$

Thus, the 25-Ω resistive load has been "transformed" to an input resistance of 100 Ω at 1 GHz, and the variation of the input impedance with frequency for this case is shown in Figure 4.14-4.

It is a common but ill-advised practice to refer to the quarter wavelength transmission line as a *quarter wave transformer* because the inversion property can be used to transform a resistive load of one value to an input resistance of another value *at a single frequency*. However, this practice should be avoided. It is better to visualize it as a *quarter wave impedance inverter*, recognizing that even that property is limited to a single frequency of operation.

The inversion changes the real parts of the load, but in addition the sign of the imaginary part is also reversed, turning an impedance with an inductive part into one with a capacitive part or vice versa. For example, if in the prior case, rather than a purely resistive load of 25 Ω, we had instead a load of $(25 + j25)$ Ω, the input impedance becomes

$$Z_{\text{IN}} = Z_0 \frac{Z_0}{Z_L} = 50 \frac{50}{25 + j25} = 50 \frac{50}{25 + j25} \frac{25 - j25}{25 - j25}$$

$$= \frac{(2500)(25 - j25)}{625 + 625} = (50 - j50) \ \Omega \qquad (4.14\text{-}14)$$

Notice that the transformation ratio of the real part is now 2 instead of the 4 obtained with a purely resistive load. Equally noteworthy, notice that the imaginary part of the input impedance is capacitive whereas the load's imagi-
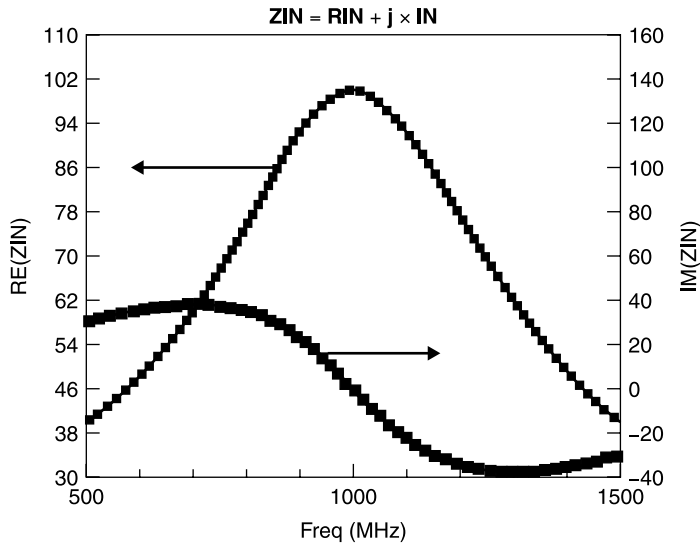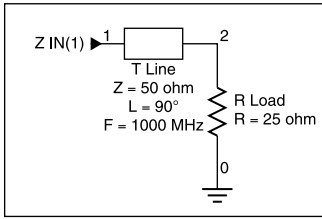
Figure 4.14-4 Variation of input impedance with frequency when 50-Ω "impedance inverter" is used to match load of 25 Ω to input resistance of 100 Ω at 1 GHz.

nary part was inductive! Thus *the action of the quarter wavelength transmission line is quite different from that of an ideal impedance transformer*, which simply multiplies any impedance by a constant factor.

Also, an *ideal transformer*, which is well approximated at low frequencies by a pair of magnetically coupled wire windings, *has a fixed impedance transformation ratio that is frequency independent*. On the other hand, the impedance transforming property of a quarter wave line section only approximates a constant transformation ratio for resistive loads and only over a narrow bandwidth centered at the design frequency as shown in Figure 4.14-4. Despite these limitations, the quarter wavelength line section has considerable practical utility for its transforming properties.

The preceding four special transmission line cases demonstrate the insight to be gained by application of the input impedance formula of (4.14-2). This formula is key to transmission line analyses. However, for complex load values and arbitrary line lengths, the formula can be quite tedious to apply. As an example, suppose that we wish to find the input impedance to a 50-Ω line that is 50° long and terminated in a load impedance of $(25 + j30)$ Ω (Fig. 4.14-5).
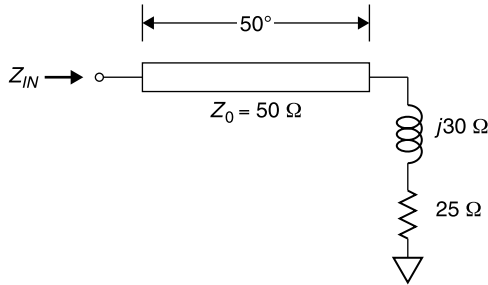
Figure 4.14-5   Example of determination of input impedance to length of 50-Ω transmission line terminated in a complex impedance of $(25 + j30)\ \Omega$.

Applying (4.14-2) gives

$$Z_{\mathrm{IN}} = Z_0 \left[\frac{Z_L + jZ_0 \tan \theta}{Z_0 + jZ_L \tan \theta} \right] = 50 \frac{25 + j30 + j50 \tan 50°}{50 + j(25 + j30) \tan 50°}$$

$$= 50 \frac{25 + j(30 + 59.6)}{50 - 35.8 + j29.8} = 50 \frac{25 + j89.6}{-14.2 + j29.8}$$

$$= 50 \frac{\sqrt{625 + 8027} \angle 74.4°}{\sqrt{202 + 888} \angle 64.5°} = 50 \frac{93.0 \angle 74.4°}{33.0 \angle 64.5°}$$

$$= 141 \angle 10° = (139 + j24)\ \Omega \qquad\qquad (4.14\text{-}15)$$

Notice that the 25-Ω real part of the load has been transformed to 139 Ω and that the $+j30$-Ω imaginary part of the load has been transformed to $j24$-Ω. From this it can be seen that, *in general, the impedance transformation properties of a simple transmission line can be quite complex*.

It is tedious to perform this calculation and very easy to make a mathematical error while doing so, thereby obtaining an erroneous result. Also, the calculation must be re-performed for each frequency of evaluation.

It would be a considerable challenge to predict, using only (4.14-2), how Z_{IN} varies with θ as it is varied over, say, 0° to 180°. This evaluation would require charting Z_{IN} for numerous θ values using the laborious calculation method described in (4.14-15). It was for this reason that Phillip Smith found an elegant graphical solution to determine input impedance, the subject of the next chapter.

4.15   IMPEDANCE MATCHING WITH ONE TRANSMISSION LINE

In the previous section it was shown that a length of transmission line can transform a real load resistance to one of a different value. Accordingly, one might ask whether any complex load impedance $Z_L = R_L + jX_L$ can be con-
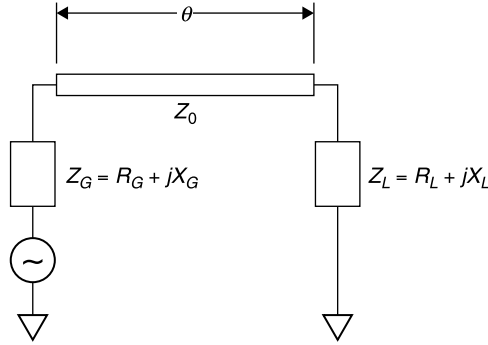
Figure 4.15-1   Conjugately matching complex load impedance Z_L to a generator impedance Z_G, using only transmission line of characteristic impedance Z_0 and electrical length θ.

jugately matched to any other complex source impedance, $Z_G = R_G + jX_G$, merely by proper selection of a line length of the appropriate Z_0 and electrical length θ. The following derivation (refer to Figure 4.15-1) yields the conditions under which this can be achieved.

From (4.14-2) with Z_{IN} set equal to the complex conjugate of Z_G,

$$Z_{\text{IN}} = R_G - jX_G = Z_0 \frac{Z_L + jZ_0 \tan \theta}{Z_0 + jZ_L \tan \theta} \tag{4.15-1}$$

Next, set the real and imaginary parts of this equation separately equal to each other and solve for Z_0 and θ. Note that $-jX_G$ was used in the equation in order that Z_{IN} will be the complex conjugate of Z_G:

$$Z_0 = \sqrt{\frac{R_L(R_G^2 + X_G^2) - R_G(R_L^2 + X_L^2)}{R_G - R_L}} \tag{4.15-2}$$

$$\theta = \tan^{-1}\left[\frac{Z_0(R_L - R_G)}{R_L X_G - R_G X_L}\right] \tag{4.15-3}$$

The above equations are always mathematically valid, however, a practical solution exists only when the solution for Z_0 is a real and finite number.

4.16   FANO'S (AND BODE'S) LIMIT

Type A Mismatched Loads

Over a given bandwidth there is a definite limit to how well one can tune a mismatched termination. The most well known are called *Fano's limits*. Fano derived these interesting and useful results within a thesis project [5].
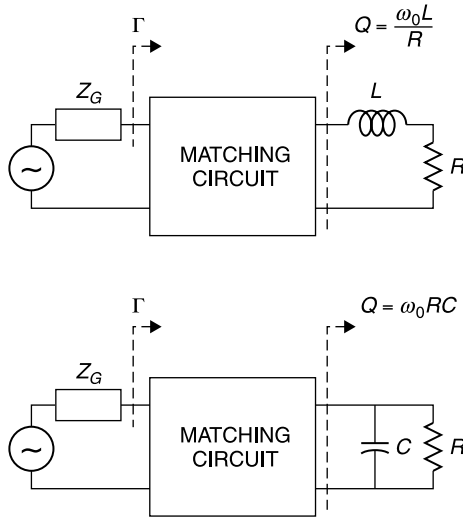
Figure 4.16-1   Fano's type A mismatched loads [5, 6].

One of Fano's two integral equation limits applies to what he described as type A mismatched loads (Fig. 4.16-1). It is

$$\int_0^\infty \ln\left|\frac{1}{\Gamma}\right| d\omega \leq \frac{\pi\omega_0}{Q} \tag{4.16-1}$$

where the reflection coefficient Γ is referenced to Z_G and the real part of Z_G is R, the same as the load resistance R in Figure 4.16-1.

Basically, (4.16-1) states that no matter what matching circuit is used, the area under the $\ln(1/\Gamma)$ curve cannot exceed the value $\pi\omega_0/Q$. Given this limit, we would usually prefer that Γ_{MIN} would be obtained over our specified bandwidth, $\omega_1 - \omega_2$, and then that Γ be unity for all other frequencies. This ideal matching requirement is shown in Figure 4.16-2.

In practice, of course, the matching behavior described in Figure 4.16-2 can only be approximated using real tuning circuits, but the ideal limit serves as an insightful restriction on how well a circuit can be tuned over a given bandwidth.

If we assume that the reflection coefficient is a constant equal to its minimum value in the passband, then the integral for the type A mismatched loads reduces to a constant multiplied by the frequency interval:

$$\int_0^\infty \ln\left|\frac{1}{\Gamma}\right| d\omega = (\omega_2 - \omega_1) \ln\left|\frac{1}{\Gamma}\right| \tag{4.16-2}$$

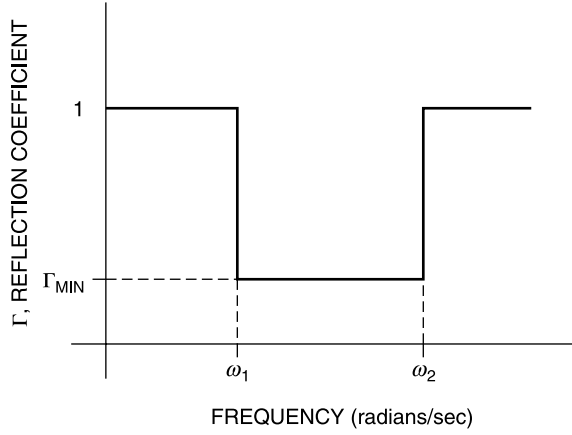where ω_1 is the frequency (in radians/second) at the lower band edge and ω_2 is

Figure 4.16-2   Idealized Γ_{MIN} match in the passband and total reflection ($\Gamma_{MIN} = 1$) at all other frequencies [5, 6].

the upper band edge. We also define

$$\text{Center frequency} = \omega_0 = \sqrt{\omega_1 \omega_2} \qquad (4.16\text{-}3)$$

and

$$\text{Fractional bandwidth} = \Delta\omega = \frac{\omega_2 - \omega_1}{\omega_0} \qquad (4.16\text{-}4)$$

Then for the type A mismatched load, assuming that a minimum and constant value of reflection coefficient, Γ_{MIN}, could be obtained throughout the passband $\omega_1 - \omega_2$

$$(\omega_2 - \omega_1) \ln\left|\frac{1}{\Gamma_{MIN}}\right| \le \frac{\pi\omega_0}{Q_L} \qquad (4.16\text{-}5)$$

Equivalently,

$$|\Gamma_{MIN}| \ge e^{-\pi\omega_0/Q(\omega_2 - \omega_1)} = e^{-(\pi/Q)(f_0/\Delta f)} \qquad (4.16\text{-}6)$$

where $f_0 = \omega_0/2\pi$ and $\Delta f = (\omega_2 - \omega_1)/2\pi$.

For example, assume there is a type A circuit consisting of a 50-Ω load resistor in parallel with a 10-pF capacitor to be matched to a 50-Ω generator over the 700- to 1100-MHz band. Notice that 10 pF presents a large susceptance in parallel with the 50-Ω load at the geometric center of this frequency band, 877.5 MHz. In reactance terms, at 877.5 MHz this is a capacitive reactance of only 18.1 Ω!

The minimum reflection coefficient magnitude, $|\Gamma_{MIN}|$, that we can expect to average over the band is found as follows.

First, calculate Q, f_0, and Δf:

$$f_0 = \sqrt{(700)(1100)} \text{ MHz} = 877.5 \text{ MHz}$$

$$X_C = \frac{159}{(10 \text{ pF})(0.8775 \text{ GHz})} = 18.1 \ \Omega$$

$$Q = \frac{R}{X_C} = \frac{50 \ \Omega}{18.1 \ \Omega} = 2.76$$

$$\Delta f = 700 - 1100 = 400 \text{ MHz}$$

Then

$$|\Gamma_{MIN}| \geq e^{(-\pi/Q)(f_0/\Delta f)} = e^{(-3.14/2.76)(877.5/400)} = e^{-2.5} = 0.082 \quad (4.16\text{-}7)$$

and the corresponding minimum average VSWR is

$$\text{VSWR}_{MIN} = \frac{1 + |\Gamma_{MIN}|}{1 - |\Gamma_{MIN}|} = \frac{1 + 0.082}{1 - 0.082} = 1.18 \quad\quad (4.16\text{-}8)$$

To examine this case, a 10-element ladder network was used as a matching network [10, Sec. 5.3]. The element values were selected arbitrarily, 10 nH for inductors and 10 pF for capacitors, and the circuit optimized using a network simulator. The simulation goal was set equal to a 1.2 maximum VSWR over the 700 to 1100 MHz bandwidth. The result, obtained once the optimization appeared to produce little further improvement, is shown in Figure 4.16-3. An average VSWR of about 1.3 was obtained over the 700 to 1100 MHz bandwidth, comparing moderately well with the theoretical limit of 1.18.

Type B Mismatched Loads

Similar reasoning can be used with Fano's type B mismatched loads (Fig. 4.16-4), for which Fano's integral limit is

$$\int_0^\infty \frac{1}{\omega^2} \ln\left|\frac{1}{\Gamma}\right| d\omega \leq \frac{Q}{\pi\omega_0} \quad\quad (4.16\text{-}9)$$

The limits for both type A and type B mismatched loads are commonly attributed to Fano [5]; however, Fano, himself, credits the limit for the type A circuits to Bode [7]. Whatever their origin, the limits give a useful insight into the problem of broadband matching.
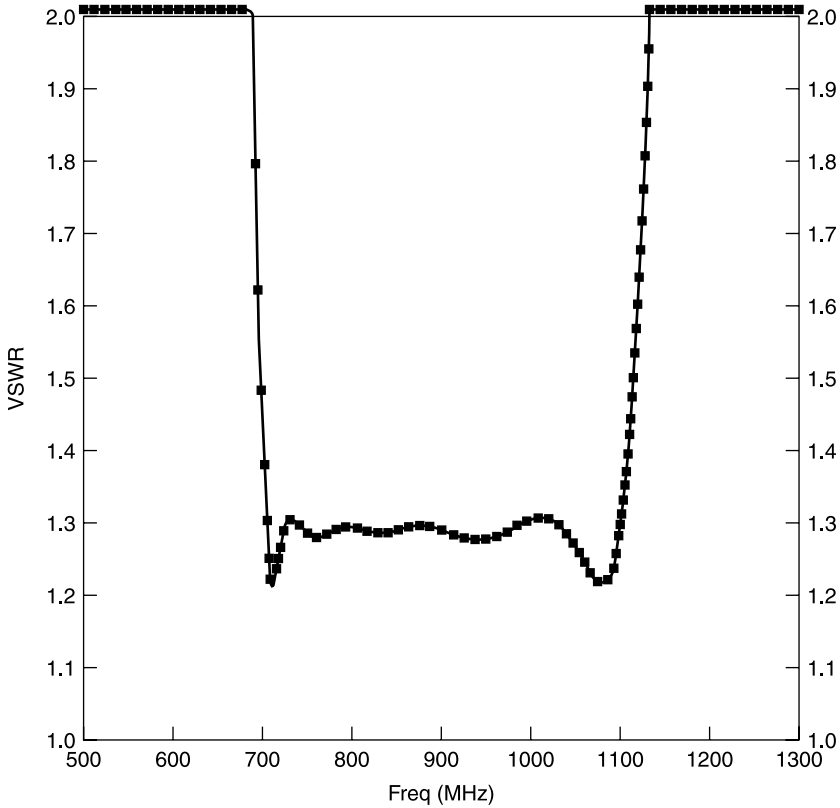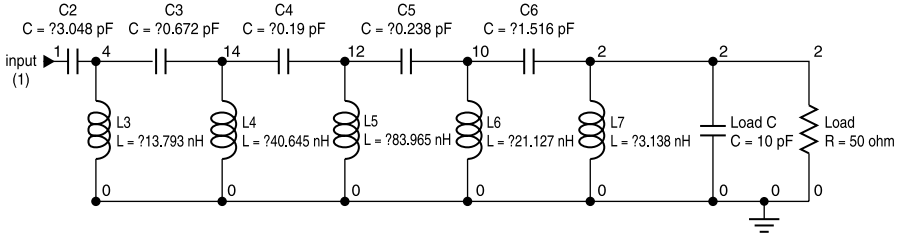
Figure 4.16-3   A 10-element matching network and its VSWR.

Impedance Transformation Not Included

When Fano and Bode performed their research, presumably for the telephone industry, the frequencies of their concern were relatively low, a few megahertz. For this reason they probably considered that a simple change in resistive level could always be accommodated using an "ideal" transformer, which has no bandwidth limitations. Probably for this reason, they only considered the reactive part of the load as a mismatch with inherent bandwidth limitations. At
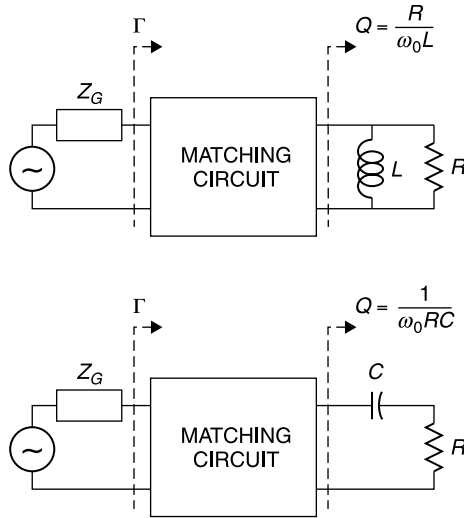
Figure 4.16-4   Fano's type B mismatched loads [5, 6].

microwave frequencies the construction of a transformer that approximates frequency-independent behavior is not as practical. Thus, the usefulness of Fano's limits at microwave frequencies are mainly confined to those cases for which the load has the correct resistive value $(R = Z_G)$ but has some series or shunt reactance that must be tuned using a matching network.

REFERENCES

1. Joseph F. White, *Microwave Semiconductor Engineering*, Noble Publishing, Norcross GA, 1995. *Extensive treatment of PIN diodes and their switching and phase shifting applications. Also includes fundamentals of microwave circuits.*

2. L. B. W. Jolley, *Summation of Series*, 2nd rev. ed., Dover, New York, 1961.

3. Peter A. Rizzi, *Microwave Engineering, Passive Circuits*, Prentice-Hall, Englewood Cliffs, NJ, 1988. *Excellent microwave engineering textbook covering theory and design of transmission lines, couplers, filters, and numerous other passive devices.*

4. Simon Ramo and John R. Whinnery, *Fields and Waves in Modern Radio*, Wiley, New York, 1944, and 1953. *This is a classic introductory text describing electromagnetic fields and Maxwell's equations.*

5. R. M. Fano, "Theoretical limitations on the broad-band matching of arbitrary impedances," *Journal of the Franklin Institute, Vol. 249, 1950*, p. 57. *Also see his doctoral thesis from the Massachusetts Institute of Technology (1947) and Technical Report No. 41, Massachusetts Institute of Technology, Research Laboratory of Electronics.*

6. Theodore Grosch, *Small Signal Microwave Amplifier Design*, Noble Publishing, Norcross, GA, 1999. *A thorough derivation of all of the important formulas for tran-*

sistor amplifier design and evaluation, based on the S parameters of the transistor and the circuit that surrounds it.

7. H. W. Bode, "A method of impedance correction," *Bell System Technical Journal*, Vol. IX, October 1930, Sections 10.7 and 10.8, pp. 794–835.

8. Theodore Saad (Ed.), *The Microwave Engineer's Handbook, Vols. I and II*, Horizon House Microwave (now Artech Books, Canton, Massachusetts) 1971. *An excellent collection of graphs and charts useful in microwave engineering.*

9. Simon Ramo, John R. Whinnery, and Theodore Van Duzer, *Fields and Waves in Communication Electronics*, 3rd ed., Wiley, New York, 1994. *This is the updated version of [4].*

10. Devendrak K. Misra, *Radio Frequency and Microwave Communication Circuits*, Wiley, New York, 2001. *This is a well-written communication text and reference.*

EXERCISES

E4.1-1 A manufacturer of PIN switching diodes proposes to make a broadband switch module consisting of a PIN diode in shunt with a 50-Ω transmission line. When the diode is reverse biased, it appears as a 0.5-pF capacitor in shunt with the transmission line and is to pass signals with minimum reflection.
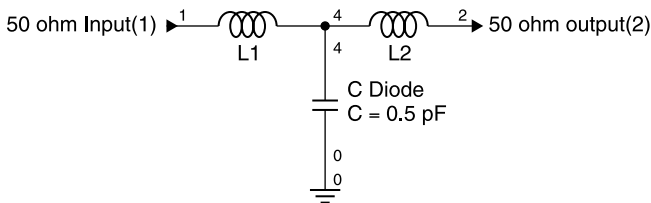
He suggests to you that if the hookup wires (from the top of the diode to the center conductor of the line and shown as L1 and L2 above) are designed to have inductance values to satisfy

$$Z_0 = \sqrt{\frac{L}{C}} = 50 \ \Omega$$

the module will be electrically indistinguishable from a length of 50-Ω line and consequently will provide matched transmission at all frequencies.

a. What should the hookup wire inductances ($L/2$ each) be made to satisfy the above relation?

b. Is it correct that this will provide an infinitely broadband transmission match?

    c. Test your answer on a network simulator over the bandwidth 0 to 10 GHz.

E4.1-2 A 50-Ω air dielectric coaxial transmission line has an inner diameter of the outer conductor equal to 0.500 in. and a center conductor diameter of 0.217 in. What is its distributed capacitance per inch?

E4.4-1 A voltage wave with 100-V peak amplitude is incident on a transmission line extending in the z direction. When it reaches the load, one quarter of its power is reflected back toward the generator which is matched to the line.

    a. What is the peak voltage on the line between generator and load?

    b. What is the minimum voltage on the line between generator and load?

    c. Sketch the manner in which the incident and reflected waves add and subtract along the line as z changes for a fixed time t.

E4.5-1 **a.** What is the magnitude of the reflection coefficient in E4.4-1?

    b. What is the VSWR?

E4.5-2 A friend has just purchased a boat containing a marine radio. The radio has a built-in VSWR meter that indicates that the antenna has a VSWR of 3 to 1.

    a. How much does this reduce the radiated power of the transmitter relative to that which would be radiated with a perfectly matched antenna?

    b. What would you advise to tune the radio to the antenna?

E4.8-1 Show that for a susceptance jB connected in shunt with an otherwise matched transmission line of characteristic impedance Z_0, the magnitude of the transmission coefficient T is less than unity for all values of B.

E4.8-2 You wish to send the output of a 500-MHz signal source to another building via a 50-Ω coaxial cable. The cable loss is 10 dB. which must be compensated by adding an amplifier at the output of the source. When connected to the 50-Ω cable, the source sees the cable as a reflection with VSWR $= 10$ (equivalently, the source output VSWR is 10), and you have a 20-dB amplifier with input VSWR of 20 to 1 referenced to 50 Ω. Neglecting other system interactions:

    a. What will be the mismatch error in connecting this amplifier to the source?

    b. Will it overcome the 10-dB cable loss in all cases?

E4.11-1 Show that the real part of the propagation constant, α, can be approximated for a low-loss transmission line having R ohms/unit length and no shunt conductance by

$$\alpha \approx \frac{R}{2Z_0}$$

E4.11-2   Use the result you derived in Exercise 4.11-1 to estimate the loss in decibels at 1 GHz for 100 ft of JAN Type RG 224/U coaxial cable having a copper inner conductor of 0.106 in. diameter (d_1) and a copper outer conductor with 0.370 in. diameter (d_2). Assume the resistivity for copper is 1.72×10^{-8} Ω-m (then conductivity, $\sigma = 5.8 \times 10^7$ ℧/m) and neglect the conduction losses through the dielectric. *Hint: Assume an equivalent current flow within one skin depth on the conductor surfaces.*

E4.14-1   Beginning with (14.4-2) verify that

$$Y_{\text{IN}} = Y_0 \left[\frac{Y_L + jY_0 \tan \theta}{Y_0 + jY_L \tan \theta} \right]$$

E4.14-2   Show that a *short length* of Z_0 transmission line terminated by a load Z_L can be approximated as a series inductance $L_{\text{EFF}} \approx Z_0 \theta / \omega$ when $Z_0 \gg |Z_L|$.

E4.14-3   A ham radio friend has purchased a 52-MHz transmitter with matched impedance antenna, both 50 Ω. He wishes to install the antenna on his roof and the transmitter in his study, a separation distance of about 25 ft. Unfortunately, he only has a 75-Ω cable to make the hookup. The relative dielectric constant of the cable's dielectric is 2.8. Is there any way that he can retain the very good match between transmitter and antenna using this cable?

E4.14-4   A resourceful engineer found a way to connect a 12.5-Ω load to a 50-Ω generator without reflection when provided only with a spool of 50-Ω cable. How did he do it?

E4.14-5   An error was made in the design of an integrated circuit narrow-band amplifier with the result that its input impedance is $(50 - j50)$ Ω instead of 50 Ω. Your company mounts the integrated chip on a "motherboard" printed circuit inside a housing with 50-Ω connectors. Vibration specifications preclude installing a lumped coil at the input to tune the amplifier. You propose to print a section of 150-Ω transmission line on the motherboard in cascade between the chip and input connector, reasoning that it will appear inductive and help to tune the mismatch. Specifications require an input VSWR of no more than 1.2.
a. What is the VSWR before this tuning?
b. How long should you make the 150-Ω line (in degrees or wavelengths)?

    c. What is the VSWR after the proposed tuning? Does it meet the objective of a VSWR = 1.2 maximum?

E4.14-6 An engineer has a coil of coaxial cable and needs to determine its dielectric material for a thermal analysis. He has been told that the dielectric material is either Teflon ($\varepsilon_R = 2.03$) or polyethelene ($\varepsilon_R = 2.26$). He forms a line section $D = 1$ m long, and he places a short circuit at its load end. Then he connects the open end to a slotted line and examines the input impedance using a variable frequency generator. He finds that it has its first resonance (a short circuit at the input) at 105.28 MHz. Can you say what the dielectric material is?

E4.15-1 Can you find a 50-Ω perfectly matched input solution to E4.14-6 using a properly chosen Z_0 and electrical length? If so, what are the values?

E4.15-2 A load impedance of $10 + j10$ Ω is to be matched to a 50-Ω generator. Can this be done using only a section of transmission line in cascade with the load? What is its Z_0 and electrical length θ?

E4.16-1 An available power amplifier design is being considered for an application requiring that the input VSWR not exceed 1.5 over the 1700 to 2100 MHz band. The amplifier, however, has an input resistance of 50 Ω that is shunted by 10 pF capacitance and presently does not meet this specification.

    a. What is f_0 for the 1700 to 2100 MHz bandwidth?

    b. What is the Q of this capacitively loaded input circuit?

    c. What is Fano's limit for the least VSWR obtainable over the 1700 to 2100 MHz bandwidth?

    d. What value of input shunting inductance will parallel resonate the 10 pF. Use a network simulator to determine over what bandwidth the 1.5 VSWR can be achieved with this single shunt L element. Does this meet the amplifier requirement?

    e. Using as many as six tuning elements and the network simulator optimizer, find a matching network that provides as low a VSWR over the 1700 to 2100 MHz bandwidth as you can achieve. Does it meet the maximum 1.5 VSWR over the 1.7 to 2.1 GHz bandwidth?

The Smith Chart

5.1 BASIS OF THE SMITH CHART

Knowledge of the basis and use of the *Smith chart*, a graphical presentation of the reflection coefficient with normalized impedance as a parameter overlay, is a sine qua non for the microwave engineer. Originally created as an aid to determining the input impedance to a transmission line, the Smith chart has become a universal aid to the design of matching circuits and to the display of measured data.

In the 1930s, when Phillip Smith [1–3] invented the graphical solution that facilitated the determination of how impedances were transformed by lengths of transmission lines, digital computers were unavailable to engineers. Engineers employed graphing strategies to gain insight into the variations produced by the independent variables of complex formulas. However, such insight was not forthcoming by attempting directly to graph the input impedance formula (4.14-2), repeated below, for a transmission line.

$$Z_{\text{IN}} = Z_0 \left[\frac{Z_L + jZ_0 \tan \theta}{Z_0 + jZ_L \tan \theta} \right] \tag{5.1-1}$$

The key to obtaining insight into this transformation was to recognize that, although impedance varies in a complex manner as one moves away from the load along a lossless transmission line, *reflection coefficient variation is quite simple! In traveling from the load to the generator along a lossless line, only the angle of the reflection coefficient changes, not its magnitude* (Fig. 5.1-1).

The reflection coefficient is the complex ratio of the reflected to the incident voltage waves. The total voltage on the line consists of the incident, V_I, and the reflected, V_R, voltages:

$$V(x) = V_I e^{-j\beta x} + V_R e^{+j\beta x}$$

$$V(x) = V_I e^{-j\theta} + V_R e^{+j\theta} \tag{5.1-2}$$

$$\Gamma(x) = \frac{V_R(x)}{V_I(x)} = \rho \angle \varphi° \tag{5.1-3}$$
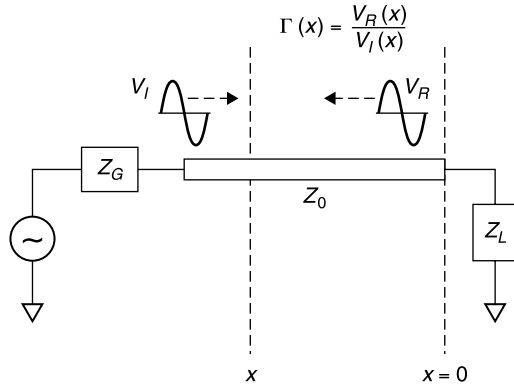
Figure 5.1-1 Moving from the load to generator on lossless transmission line, reflection coefficient changes in phase only because the amplitudes of incident voltage V_I and the reflected voltage V_R do not change.

where $V_R(x) = V_R e^{+j\theta}$, $V_I(x) = V_I e^{-j\theta}$, and $V_R(0)/V_I(0) = \Gamma_{\text{Load}}$. Choosing $x = 0$ at the load and moving an electrical distance θ toward the generator from the load causes the reflected wave V_R to have a negative change of phase, equal to θ, relative to its phase at the load. The incident wave V_I has a positive change of phase equal to θ, but the incident voltage is in the denominator, hence

$$\Gamma = \Gamma_L e^{-2j\theta} \tag{5.1-4}$$

Moving an electrical distance θ *from the load, the reflection coefficient argument is reduced by* -2θ.

This is a key relationship. We will make use of it frequently as we use reflection coefficient change to determine the input impedance of a transmission line.

If the relationships of (4.13-4) and (4.13-5) are normalized to Z_0 (all impedances divided by Z_0), $\Gamma(x)$ and $z(x) = Z(x)/Z_0$ are related by

$$z(x) = \frac{1 + \Gamma(x)}{1 - \Gamma(x)} \tag{5.1-5}$$

$$\Gamma(x) = \frac{z(x) - 1}{z(x) + 1} \tag{5.1-6}$$

Both of these equations are *bilinear transformations* [4, p. 202]. That is, they are of the form

$$w = \frac{Az + B}{Cz + D} \tag{5.1-7}$$

where w and z are complex variables and $A, B, C,$ and D are complex constants satisfying

$$\Delta = AD - BC \neq 0 \qquad (5.1\text{-}8)$$

The bilinear transformation is a *one-to-one mapping that* has been well studied in the theory of complex functions. In the case of (5.1-5) and (5.1-6) this means that: *Every point in the $z(x)$ plane corresponds to a unique point in the $\Gamma(x)$ plane, and vice versa.*

To illustrate this z and Γ correspondence, consider the example of the load impedance $(25 + j30)$ Ω terminating the 50° long, 50-Ω line described in Figure 4.14-5 and evaluated in (4.14-13). Normalized, this impedance and the corresponding reflection coefficient at the load are

$$z_L = 0.5 + j0.6 \qquad (5.1\text{-}9)$$

$$\Gamma_L = \frac{z_L - 1}{z_L + 1} = \frac{0.5 + j0.6 - 1}{0.5 + j0.6 + 1} = 0.48\angle 108° \qquad (5.1\text{-}10)$$

Let us now determine the impedance at a point 50 electrical degrees from the load. To do so we decrease the angle of the reflection coefficient by twice the electrical length (50°) of the transmission line, or 100°, to obtain the reflection coefficient at an electrical distance $\theta = 50°$ from the load:

$$\Gamma_\theta = \Gamma_L e^{-2j\theta} = 0.48\angle(108° - 100°) = 0.48\angle 8° \qquad (5.1\text{-}11)$$

The corresponding normalized impedance is

$$z_\theta = \frac{1 + \Gamma_\theta}{1 - \Gamma_\theta} = \frac{1 + 0.475 + j0.07}{1 - 0.475 - j0.07} = 2.76 + j0.48 \qquad (5.1\text{-}12)$$

$$Z_\theta = z_\theta(50\ \Omega) = (138 + j24)\ \Omega \qquad (5.1\text{-}13)$$

These results are plotted in the Γ plane diagram in Figure 5.1-2.

Our strategy has succeeded. The load impedance $25 + j30$ Ω was normalized to the 50-Ω transmission line and the corresponding reflection coefficient determined ($\Gamma_L = 0.48\angle 108°$). This reflection coefficient was transformed along the 50° long line by preserving its magnitude (0.48) and reducing its angle (108°) by twice the electrical length of the line (100°). The resulting reflection coefficient ($\Gamma_L = 0.48\angle 8°$) was used to calculate the normalized impedance ($z_\theta = 2.76 + j0.48$) at the input to the line, and this value, when unnormalized to 50 Ω, gives the input impedance $Z_\theta = (138 + j24)$ Ω. This is close to the value of $(139 + j24)$ Ω obtained for the same example in (4.14-13) by using the impedance transformation formula. The small difference in the real part is likely due to round-off error. Notice that we are able to treat the electrical dis-
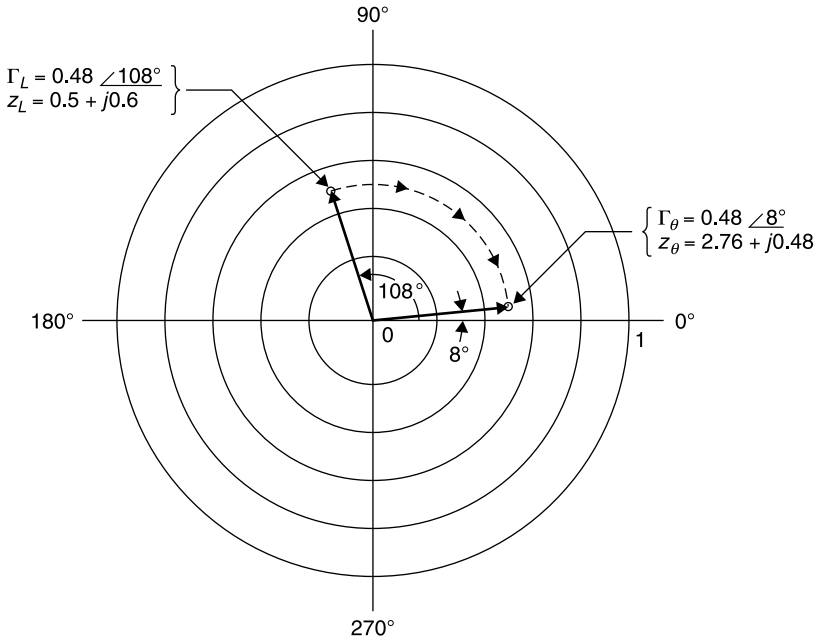
Figure 5.1-2   The Γ plane representation of load impedance z_L normalized to 50 Ω and transformed through a 50° length of 50-Ω transmission line.

tance of the line to the load, θ, as a positive quantity, since the clockwise rotation of Γ takes its sign into account.

While successful, our strategy of using the simple rotation of Γ to calculate the impedance transformation of a load by a line length has not resulted in any real economy of computational effort due to the need to perform the complex transformations from z to Γ (5.1-6) and then from Γ to z (5.1-5).

However, we note that *once a z plane point is mapped into the Γ plane it is done once and for all*, since the z and Γ points have a one-to-one correspondence (Fig. 5.1-3). Thus, plotting normalized impedance values in the Γ plane in advance would permit entering the Γ unit circle directly with a normalized load impedance value and, after performing the required 2θ clockwise rotation, reading the corresponding normalized input impedance to the line. *This is the basis of the Smith chart.*

However, performing this translation of the z to the Γ plane poses some obvious problems. First, the flags used to identify the corresponding z and Γ points take up so much room on the Γ chart that it would be impractical to show a sufficient number of them to permit accurate plotting of z. Second, the z plane is semi-infinite for passive loads, corresponding to all impedances for which $r \geq 0$; hence an unduly large sheet of paper would be required to ac-
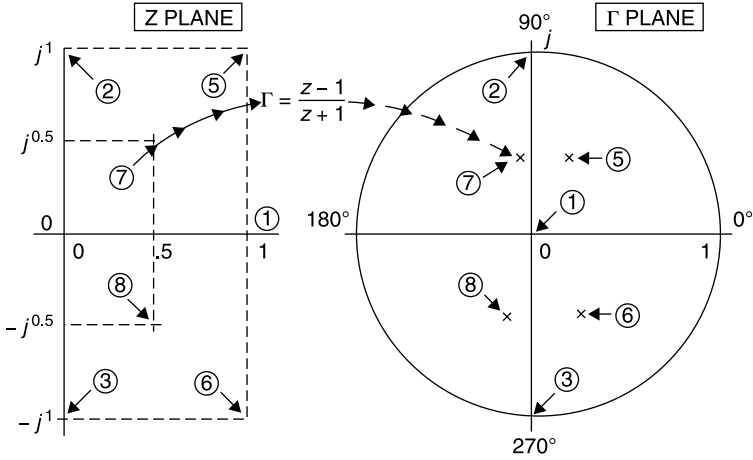
Figure 5.1-3 Mapping discrete impedance points from normalized z plane to Γ plane under the transformation $\Gamma = (z-1)/(z+1)$.

commodate large impedances in the z plane. We note, however, that no such problem exists in the Γ plane, wherein all passive loads must yield a $|\Gamma| \leq 1$, thereby fitting neatly within the unit circle.

This latter observation raises an interesting philosophical question. How can the points of a semi-infinite area (the right half of the z plane) be mapped into a finite area within the $|\Gamma| \leq 1$ circle? The answer is similar to that of the question: "How many angels can dance on the head of a pin?" The answer is: "All the angels can, because angels do not require any space." Similarly, impedance points, no matter how many of them, take up no area and therefore can be mapped from any area into any other area.

The solution to the two problems stated above lies in finding a means of *mapping* not individual points but contours from the z plane to the Γ plane. When we use the z plane, we find the intersection of the $r = $ const and $x = $ const lines for the particular impedance at hand. Once these contours are mapped onto the Γ plane (the Smith chart), we can enter the Γ plane directly, using the normalized load impedance on a line; and, after the required clockwise rotation, can read directly the corresponding normalized input impedance at that point on the line. No complex calculations would be required to perform the input impedance calculation of (5.1-1).

In summary, Philip Smith's first important idea in developing the Smith chart was recognizing that reflection coefficient rather than impedance should be used to track movement on a transmission line. The second important idea was the mapping of $r = $ const and $x = $ const contours in the z plane to the Γ plane. The Γ plane with this mapping of normalized impedance is called the *Smith chart*.

5.2 DRAWING THE SMITH CHART

The normalized, passive impedance Smith chart is the polar plot of $|\Gamma| \leq 1$ with an overlay of r = const and x = const contours that relate Γ to z. Normalizing the Smith chart allows its use with any characteristic impedance transmission line. Alternatively, one can draw a Smith chart for use with a given absolute value of Z_0, such as 50 Ω.

The normalized, passive impedance Smith chart consists of constant resistance and reactance contours mapped from the z to the Γ plane according to the function

$$\Gamma = \frac{z-1}{z+1} \quad \text{where } z = r + jx,\ r \geq 0,\ -\infty \leq x \leq +\infty \qquad (5.2\text{-}1)$$

The mapping function of (5.2-1) is a bilinear transformation, which has four properties important to the drawing of the Smith chart [4]. These are:

1. *The semi-infinite right half of the z plane (for $r \geq 0$) is mapped into the $|\Gamma| \leq 1$ unit circle.* (Note that the left half of the z plane for which $r < 0$ would map into the semi-infinite area outside of the $|\Gamma| > 1$ circle, the negative resistance domain.)

2. *The transformation maps circles into circles*, with straight lines being considered circles of infinite radius and points being circles of zero radius.

3. *The mapping is conformal (angle preserving)*. Contours in the z plane that are orthogonal to each other will map into contours in the Γ plane that likewise are orthogonal to each other.

4. *The mapping is analytic*. Therefore, continuous contours in the z plane are mapped into continuous contours in the Γ plane.

Usually the reflection coefficient Γ is expressed in polar form as $\rho e^{j\varphi}$ or $\rho\angle\varphi$; however, it is difficult to recognize the equation of a circle in this format unless the circle's center lies at the origin. Therefore, to demonstrate this mapping, we express Γ in its Cartesian format, $\Gamma = e^{j\varphi} = u + jv$, retaining the same coordinate center for both polar and Cartesian representations.

The relationships between polar and rectangular coordinates are, from polar to rectangular

$$u = \rho \cos \varphi \qquad (5.2\text{-}2a)$$

$$v = \rho \sin \varphi \qquad (5.2\text{-}2b)$$

and from rectangular to polar

$$\rho = \sqrt{u^2 + v^2} \qquad (5.2\text{-}3a)$$

$$\varphi = \tan^{-1}\left(\frac{v}{u}\right) \qquad (5.2\text{-}3b)$$

Because a complex function such as $\Gamma = u + jv$ has two parts, real and imaginary, as does its complex independent variable, $z = r + jx$, it is not possible to graph $\Gamma(z)$ versus z in a single plane as we would the real variable $f(x)$ versus x. Instead we will show how a contour in the z *plane*, such as $r = 2$, is *mapped* into a corresponding contour in the Γ *plane*.

In this mapping task, we wish to set $r = $ const for one set of contours and then $x = $ const for the other, orthogonal, set of contours to be drawn in the Γ plane. To do this it proves convenient to use the relation that gives z as a function of Γ. In this way z can be set either to a constant r or jx and the expression solved for the Γ rectangular coordinates. Thus,

$$z = \frac{1 + \Gamma}{1 - \Gamma} \tag{5.2-4}$$

where z and Γ continue to be functions of position x on the transmission line (not to be confused with normalized reactance, x).

The relation (5.2-4) can be rewritten as

$$r + jx = \left(\frac{1 + \rho e^{j\varphi}}{1 - \rho e^{j\varphi}}\right)\left(\frac{1 - \rho e^{-j\varphi}}{1 - \rho e^{-j\varphi}}\right) \tag{5.2-5}$$

Carrying out the indicated multiplication to remove the complex quantities from the denominator and noting that $e^{j\varphi} - e^{-j\varphi} = 2j \sin \varphi$ and $e^{j\varphi} + e^{-j\varphi} = 2 \cos \varphi$ we obtain

$$r + jx = \frac{1 - \rho^2 + 2j\rho \sin \varphi}{1 + \rho^2 - 2\rho \cos \varphi} = \frac{1 - u^2 - v^2 + 2jv}{1 + u^2 + v^2 - 2u} \tag{5.2-6}$$

This complex equation requires that separately both the real and imaginary parts of each side of the equation be equal to each other. Equating the real parts,

$$r = \frac{1 - u^2 - v^2}{1 + u^2 + v^2 - 2u} \tag{5.2-7}$$

We know from the properties of the bilinear transformation that (5.2-7) must be the equation of a circle in the $\Gamma = u + jv$ plane for $r = $ const. But it requires some algebraic manipulation [5, p. 75] to cast (5.2-7) into the easily recognizable expression for a circle with center at $u_1 + jv_1$ and radius a, namely the format

$$(u - u_1)^2 + (v - v_1)^2 = a^2 \tag{5.2-8}$$

To do so, rewrite (5.2-7) as

$$r(1 + u^2 + v^2 - 2u) = 1 - u^2 - v^2$$

$$u^2(r + 1) - 2ur + (r - 1) + v^2(r + 1) = 0$$

$$u^2 - \frac{2ur}{r + 1} + \frac{r - 1}{r + 1}\frac{r + 1}{r + 1} + v^2 = 0$$

$$\left(u - \frac{r}{r + 1}\right)^2 + v^2 = \left(\frac{1}{r + 1}\right)^2 \qquad (5.2\text{-}9)$$

The last form of (5.2-9) is the recognizable expression for a circle in the Γ plane with center at

$$u = \frac{r}{r + 1} \qquad jv = 0$$

and radius equal to $1/(r + 1)$. For example, the contour $r = 1$ in the z plane is mapped into a circle having radius 0.5 and center at $u = 0.5$, $jv = 0$, as shown in Figure 5.2-1.

At this point in the mapping, we note that all impedance values having $r = 1$ as a real part must lie on the $r = 1$ circle in the Γ plane. We do not yet know how to place specific impedances, on this circle, since we have not yet developed the orthogonal x contours for the Γ plane. In a limited sense, we can deduce that $\Gamma = 0$ when $z = 1 + j0$, and that accordingly this impedance point must be the one at which the $r = 1$ circle passes through the origin, since the
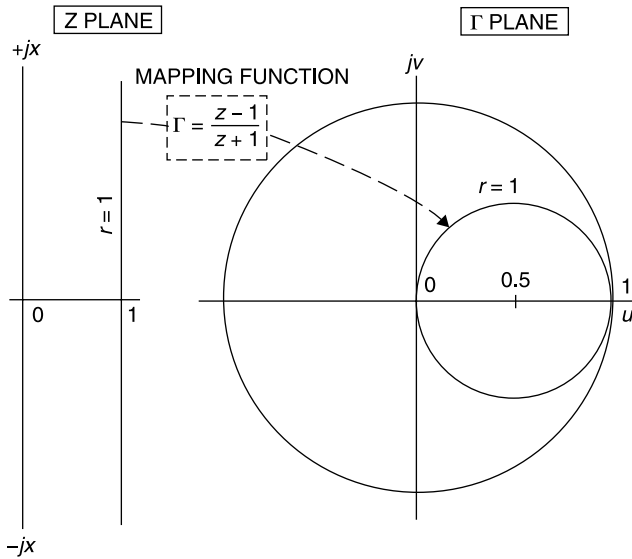
Figure 5.2-1 Contour $r = 1$ in z plane is mapped into corresponding contour in Γ plane.
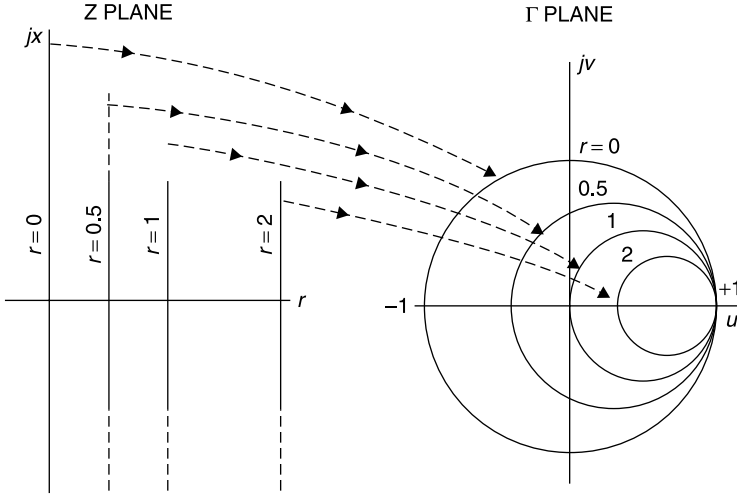
Figure 5.2-2 Constant-resistance lines of z plane map into circles tangent at $u = 1$, $jv = 0$ in Γ plane and centered on real axis.

corresponding reflection coefficient magnitude must be zero. We can also say that at $u = 1$, $jv = 0$, $\Gamma = +1$ (an open circuit), and so this must correspond to the $z = 1 \pm j\infty$ point.

As additional contours are plotted, the constant r lines in the z plane map into circles that are tangent to each other at the $u = 1$, $jv = 0$ point in the Γ plane, as shown in Figure 5.2-2.

Next, we map the orthogonal $x = \text{const}$ contours into the Γ plane. To do so, equate the imaginary parts on both sides of (5.2-6) to obtain

$$x = \frac{2v}{1 + u^2 + v^2 - 2u} \tag{5.2-10}$$

This is manipulated as

$$u^2 - 2u + 1 + v^2 - \frac{2v}{x} + \frac{1}{x^2} = \frac{1}{x^2}$$

$$(u - 1)^2 + \left(v - \frac{1}{x}\right)^2 = \left(\frac{1}{x}\right)^2 \tag{5.2-11}$$

For x equal to a constant, this is the equation of a circle with center at $u = 1$, $jv = j/x$ and radius equal to $1/x$. When contours are mapped for $x = -2$, $-1, -0.5, 0, +0.5, 1$, and 2, the results give partial circles within the $|\Gamma| \leq 1$ unity circle of the Γ plane. The circles can be extended beyond this region, but the corresponding impedances have negative real parts, and our interest for
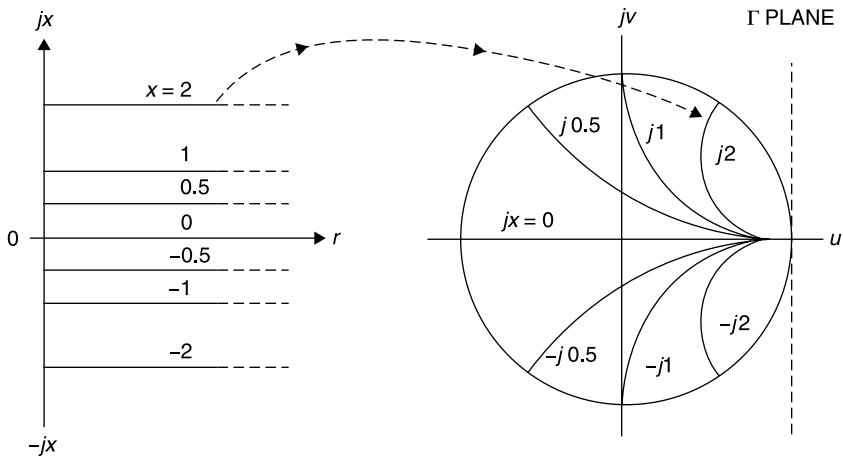
Figure 5.2-3   Constant-reactance x lines of z plane map into circles in Γ plane, but only portions of circles fall within the unit circle, corresponding to passive impedances.

now is the Smith chart for use with passive impedances only, that is, for reflection coefficients no greater than unity. More about negative resistance Smith charts later.

The contours in the Γ plane shown in Figures 5.2-2 and 5.2-3 form orthogonal sets, mapping the $z = r + jx$ values into the reflection coefficient plane, as we desired. Notice that, while the contours are circular, at their intersections their lines are at right angles (orthogonal) to each other. When both sets of contours ($r = \text{const}$ and $x = \text{const}$) are combined, the result is the skeletal form of the Smith chart shown in Figure 5.2-4.

Notice that *positive reactances (inductances) lie in the upper half and negative reactances (capacitances) in the lower half of the Smith chart* (as was true in the z plane).

The diagram of Figure 5.2-4 is very comprehensible because it shows the evolution of the Smith chart by presenting all of the necessary coordinates and scales. However, if the Smith chart is to be used graphically, as was the original intent, it must contain many more impedance contours in order that impedance can be read more precisely. For this reason the conventional format for the Smith chart is as shown in Figure 5.2-5.

Because so many r and x contours are required for a Smith chart that is to be used graphically to determine impedance transformations on transmission lines, there is insufficient space within the $|\Gamma| = 1$ circle to print explicit scales for the magnitude of reflection coefficient, even though *reflection coefficient is the basis of the Smith chart*. Also, the j is not used ahead of reactive impedances or susceptive admittances.

That *the Smith chart fundamentally is a plot of reflection coefficient* is emphasized here because it is not uncommon for individuals to become quite
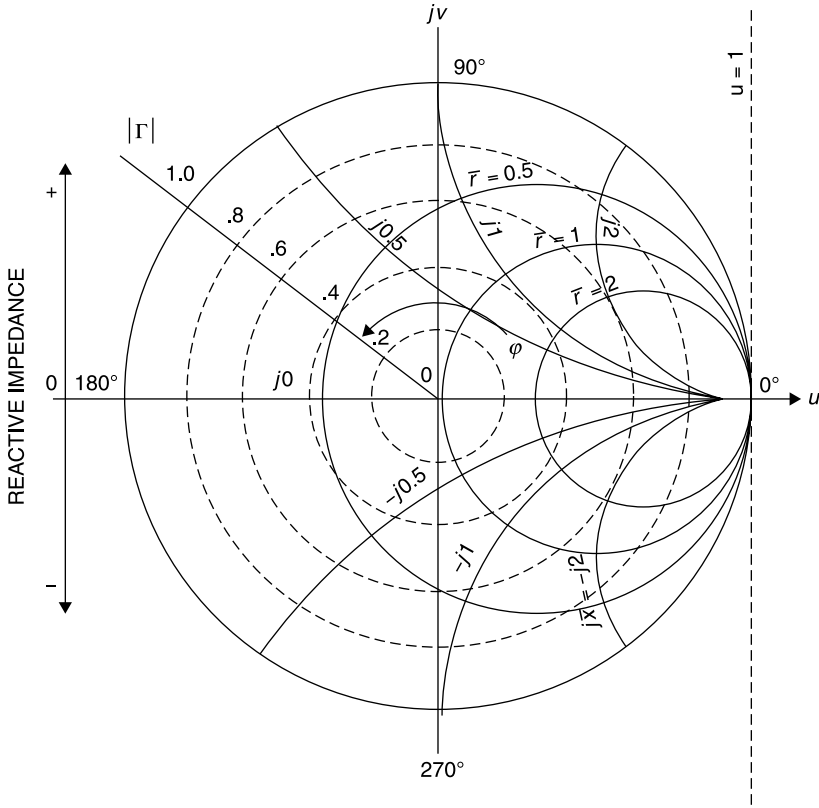
Figure 5.2-4 Smith chart showing both impedance contours r and jx as well as $|\Gamma|$ contours and Γ's angle φ in 90° segments.

adept at using the Smith chart to transform impedances yet forget, or never have been aware of, the fact that the Smith chart is just a polar plot of reflection coefficient with impedance and/or admittance overlays.

Knowing this fact, it is easy to answer many other questions about transmission lines, such as what is the voltage magnitude at a given position x on the line. It is simply $V_I |1 + \Gamma(x)|$ where Γ is readily read for any specified point on the chart.

To facilitate the determination of reflection coefficient and other important transmission line parameters, there are radial scales printed at the bottom of the chart. These are used in combination with a pair of dividers to determine the magnitude of the reflection coefficient. Only the linear reflection coefficient magnitude scale, ρ, is necessary, since the angle of Γ can be read from the degrees scale on the periphery of the Smith chart. The other parameters, return loss, mismatch loss, and other values are readily determined, given ρ; but for convenience separate radial scales for these values are also printed below the chart.
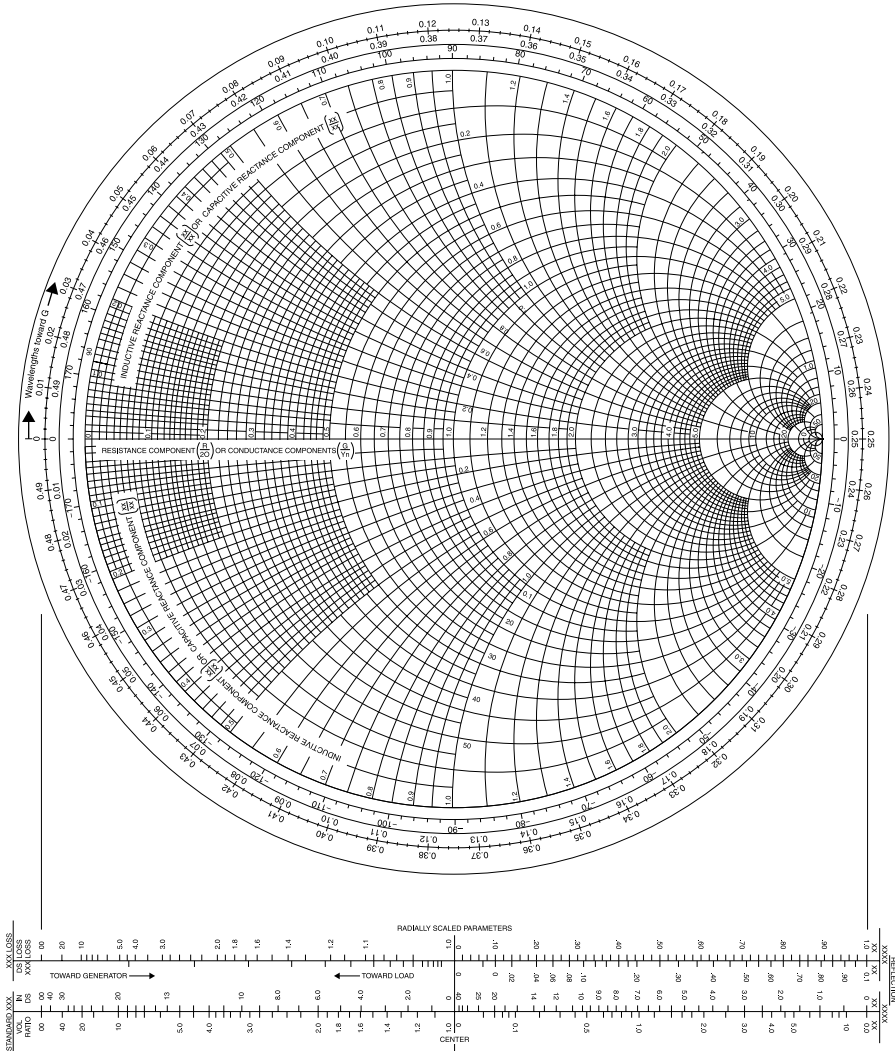
Figure 5.2-5   Smith chart as conventionally printed for use as a graphical calculation tool. (*Smith is a registered trademark of Analog Instrument Co., Box 950, New Providence, NJ 07974. Smith charts are reprinted within this book with permission.*)

5.3   ADMITTANCE ON THE SMITH CHART

At any location on the transmission line the reflection coefficient is related to the normalized impedance by

$$\Gamma = \frac{z - 1}{z + 1} \tag{5.3-1}$$

Since normalized admittance y is just the reciprocal of normalized impedance z,

$$\Gamma = \frac{1/y - 1}{1/y + 1} = \frac{1 - y}{1 + y} \qquad (5.3\text{-}2)$$

where the bilinear transformation in (5.3-2) is simply the negative of the transformation in (5.3-1). That is, *the reflection coefficient plot as a function of y is identical to that of z except for a rotation within the Γ plane of 180°*. When the admittance coordinates for constant conductance g and constant susceptance b are plotted in the Γ plane, the result is the admittance version of the Smith chart, shown in Figure 5.3-1.

This plot is identical to that of the impedance plot in Figure 5.2-4 except for a 180° rotation about the center of the chart. All of the rules for transforming impedance along a line length apply equally to admittances. With normalized admittances, *the rotation moving from the load toward the generator remains clockwise*, as it was for impedances.
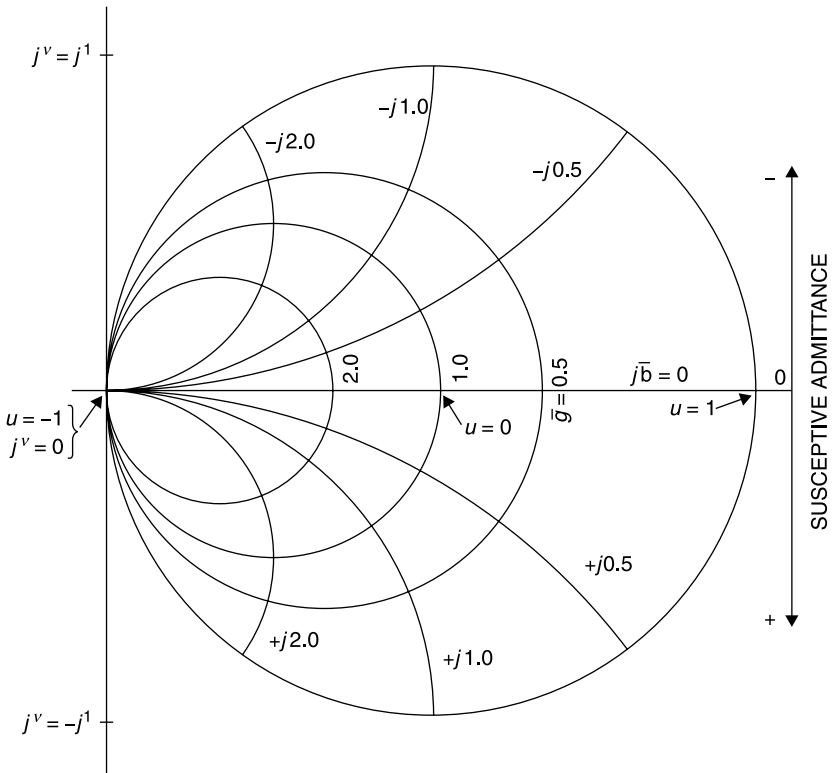
Figure 5.3-1   Smith chart with normalized admittance coordinates.

Inductors have positive reactance but negative susceptance. Capacitors have negative susceptance but positive susceptance with the result that *the upper half of the Smith chart is always inductive and the lower half is always capacitive*.

Due to the mirror symmetry of the impedance and admittance Smith charts, one could create an admittance Smith chart merely by rotating an impedance chart through 180°. In fact, for many years the practice was to consider that the normalized coordinates in Figure 5.2-5 could be defined to be either admittance or impedance, and the "standard" chart shown in Figure 5.2-5 was, and often currently is, labeled "impedance or admittance coordinates."

The problem with defining these coordinates as admittance is that the angle of the reflection coefficient is in error by 180°, and this must be corrected when interpreting the "standard" Smith chart as an admittance Smith chart. Another problem is that this practice is more prone to errors, since it is easy to forget how the coordinates are presently defined, particularly since conversions between admittance and impedance frequently must be made in the solution of a problem.

The modern and preferred procedure is to print both normalized impedance and admittance contours on the same Smith chart, overlaying the coordinates derived in Figures 5.2-4 and 5.3-1, so that one can read directly at any given point either value without resort to crossing or rotating the chart. The resulting Smith chart, having both impedance and admittance contours, is shown in Figure 5.3-2.

To aid the graphical use of this chart, it is usually printed in two colors, for example, red for normalized impedance and blue for normalized admittance. Even so, as a graphical tool it is difficult to use. Fortunately, today there are software versions of the Smith chart for use on a personal computer. Fewer contours need be drawn because the computer allows entry onto the chart at the exact value of load impedance or admittance, creates the traces of transmission line and other transformations, and displays the exact value of input impedance or admittance, as will be seen. Furthermore, either impedance or admittance contours or both can be displayed as desired. This will be demonstrated in the examples to follow.

5.4   TUNING A MISMATCHED LOAD

Originally, the Smith chart was developed to facilitate the determination of load mismatch and the requisite tuning that would match it to the characteristic impedance of the transmission line. While the chart has other useful applications, this load tuning remains a principal application of the Smith chart.

Movement along a lossless line corresponds to a contour having a constant reflection coefficient magnitude, ρ, and thus a constant radial distance from the center of the Smith chart, that is, a circle with center at the center of the chart. The angle, φ, of the reflection coefficient, $\Gamma = \rho e^{j\varphi} = \rho\angle\varphi$, changes at twice the rate of movement along the line because it is the ratio of an incident and a
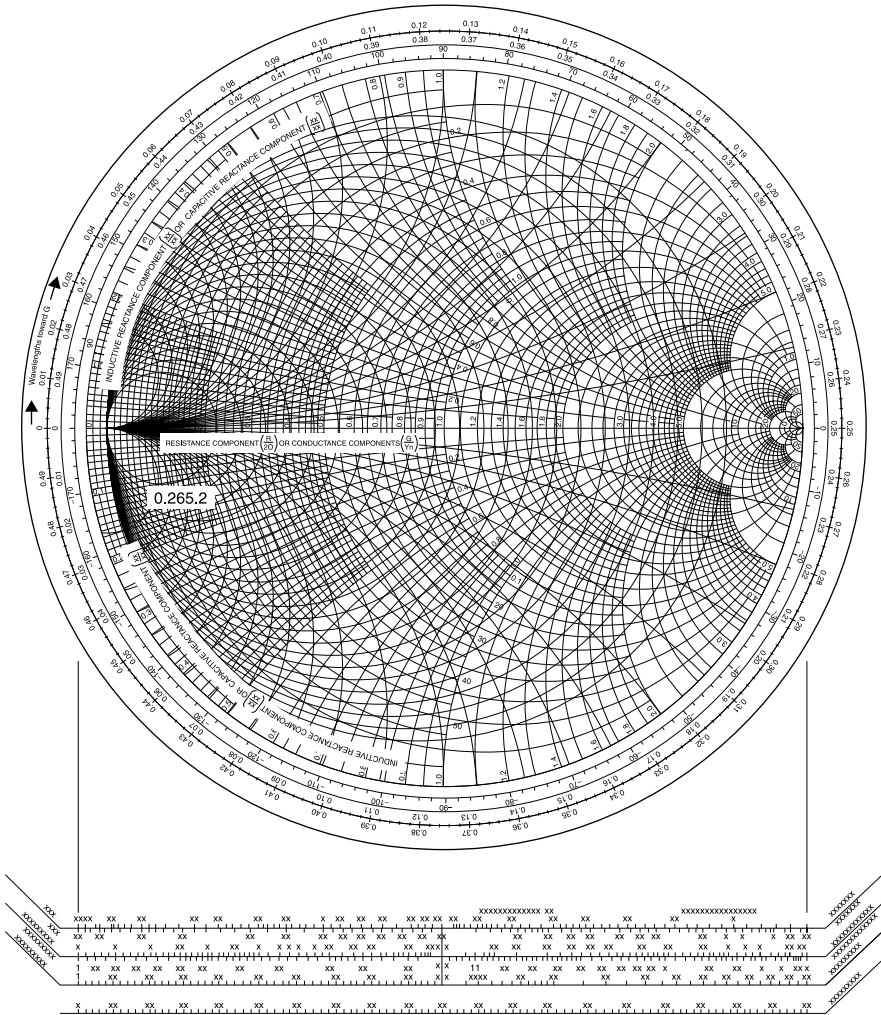
Figure 5.3-2   Smith chart with simultaneous impedance and admittance contours.

reflected wave, each of whose arguments change by an amount equal to the movement and their angles are additive in taking their ratio. *Movement from the load toward the generator on the Smith chart is clockwise.*

To facilitate the computation of movements, the Smith chart has three peripheral scales. The innermost scale gives degrees of the angle of Γ, a total of 360° for the full circle. The outer two scales are labeled "wavelengths toward the generator" and "wavelengths toward the load." These two scales go through one-half wavelength for the full circle, thereby taking into account that the reflection angle changes by twice the angle of movement along the transmission line. This movement is demonstrated in the following example.
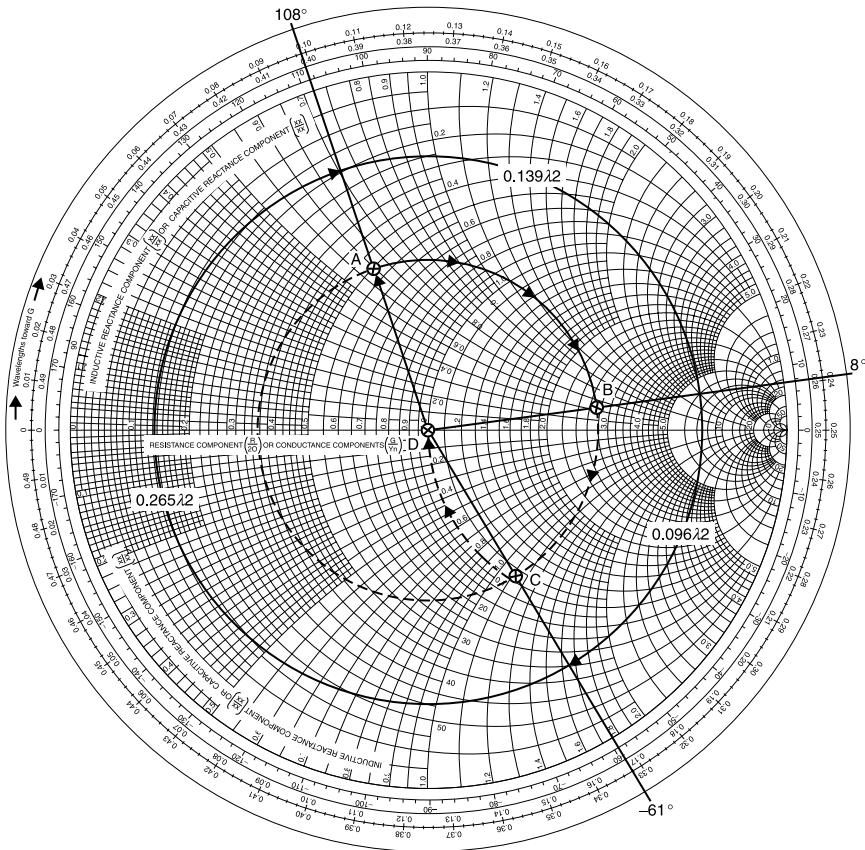
Figure 5.4-1   A $(25 + j30)$-Ω load (point A) on 50-Ω line with tuning to match it to 50 Ω (point D).

For this example, we return to the case of Figure 4.14-5 consisting of a $(25 + j30)$-Ω load at the end of a 50° long section of 50-Ω transmission line. Normalizing the load impedance to 50 Ω, it becomes $0.5 + j0.6$. This is shown as point A in Figure 5.4-1.

To determine the arc length movement on the Smith chart, we note that traveling 50° along the line *toward the generator* results in a reflection coefficient angle change of $-100°$ (*clockwise rotation*) to point B, at which we read a normalized impedance of $2.8 + j0.5$. Unnormalized (multiplying both parts by 50 Ω) this is $(140 + j25)$ Ω, reasonably close to the value of $139 + j24$ calculated in (4.14-15), given the graphical accuracy of the Smith chart, an analog instrument. Note the economy of effort with which this result is obtained relative to the solution with the input impedance formula.

As an alternative to the determination of the arc length movement on the Smith chart, we can use the "wavelengths toward generator" scale. Traveling 50° on the line toward the generator results in a clockwise movement on

the chart by $50°/360° = 0.139$ wavelengths. This dimension is shown in Figure 5.4-1 and results in the same travel from point *A* to point *B*.

From Figure 5.4-1 the value of the Smith chart is evident. The extended circular contour (shown dashed) at a constant radial distance from the center of the Smith chart reveals all of the normalized impedance values that would be encountered if the transmission line were extended to a half wavelength or longer.

Specifically, we see that, if the transmission line is lengthened until the reflection coefficient angle is $-61°$ (point *C*) the normalized input impedance will be $1.0 - j1.1$ ($50 \, \Omega - j55 \, \Omega$ when unnormalized). At point *C* the total change in reflection coefficient angle measured from point *A* is $-169°$, corresponding to a total transmission line length of $84.5°$. To reach point *C* we must add an additional $34.5°$ of 50-Ω line, corresponding to a clockwise movement on the chart of an additional 0.096 wavelengths. At point *C* the insertion of an inductive reactance of 1.1 ($55 \, \Omega$ unnormalized) in series with the line cancels the capacitive reactance and moves the normalized input impedance to 1 ($50 \, \Omega$ unnormalized) at point *D*, thereby matching the load to the transmission line.

This example demonstrates that *not only does the Smith chart facilitate the calculation of the input impedance to a line terminated in a mismatched load, it also readily reveals how the mismatch can be tuned to yield a matched input*.

This tuning method is not unique. Actually, there are an infinite number of possible tuning approaches that can be determined with the skillful use of the Smith chart. By continuing clockwise on the dotted circular arc, an additional 0.27 wavelength, between points *C* and *A*, respectively, of Figure 5.4-1, we can observe the remainder of the wide range of normalized impedance values that would have been encountered on the transmission line had the electrical length, θ, been varied from $0°$ to $180°$. The total excursion is 0.5 wavelength, after which we are returned to the starting impedance. This duplicate of the load impedance occurs every half wavelength multiple from the load along a lossless line, as was noted in Section 4.14.

In the current example, the real part of the normalized input impedance varies between 0.33 and 3, a 9-to-1 ratio. The imaginary part varies between $-j1.2$ and $+j1.2$. Using only the input impedance formula, these conclusions, obvious with the Smith chart, would have been very laborious to reach. For this reason the Smith chart is a universal presentational format for microwave measurements, usually a direct data output option for network analyzer measurements and network simulation software.

5.5 SLOTTED-LINE IMPEDANCE MEASUREMENT

From the 1930s to the 1950s, before network analyzers were available, the *slotted line* was used to measure microwave impedance. Even today, VSWR, the data value obtained with a slotted line, continues to be a common term used to describe the tolerance of impedance matching.

The slotted-line measurement was effected by cutting a nonradiating slot along the direction of propagation in a coaxial or waveguide transmission line and sliding a rectifying detector diode along the slot to sample the rectified time-average voltage amplitude as a function of position along the line. This fixture was called a *slotted line*. *The ratio of the maximum to minimum voltage values is the voltage standing-wave ratio*, or *VSWR*. The VSWR coupled with the location of the associated voltage minimum (*null*) could be used to determine the load impedance Z_L. The following example demonstrates this procedure.

First, a short circuit is connected at the end of the transmission line prior to connecting the load (Fig. 5.5-1*a*), and the positions of the resulting voltage minima, or nulls, recorded. These null locations are equivalent to the electrical position of the load plane. The rectified voltage amplitude varies sharply with position near a null (Fig. 5.5-1*a*), yielding the most precise measurement of the equivalent load positions. *The distance between two successive minima is one-half wavelength at the operating frequency*, providing a means of determining the actual operating frequency of the generator. *Each set of VSWR and null data is taken at a single microwave frequency*. To obtain data for a band of frequencies, the measurement must be re-performed at each frequency of evaluation.

For example, suppose that, with the short circuit installed, nulls are found at 2, 4.95, and 7.90 in. along the slotted line. The distance between minima is a half wavelength, therefore

$$\lambda = 2(4.95 \text{ in.} - 2.00 \text{ in.}) = 5.90 \text{ in.} \qquad (5.5\text{-}1)$$

If the slotted line has air dielectric, as is usually the case to permit unimpeded probe movement,

$$f_0 = \frac{11.8 \text{ in.}}{5.9 \text{ in.}} 1000 \text{ MHz} = 2000 \text{ MHz} \qquad (5.5\text{-}2)$$

This demonstrates that *the slotted line provides a means of frequency determination*.

The locations of the voltage minima can be considered equivalent to the electrical position of the load plane, repeating at half wavelength intervals.

Next, the load is attached and the rectified voltage measurements yield $V_{\text{MAX}} = 1.5$ units and $V_{\text{MIN}} = 0.5$ units (the actual voltage scale is unimportant) with null positions at 2.59 and 5.54 in. From these data

$$\text{VSWR} = \frac{V_{\text{MAX}}}{V_{\text{MIN}}} = \frac{1.5}{0.5} = 3 \qquad (5.5\text{-}3)$$

$$\rho = \frac{\text{VSWR} - 1}{\text{VSWR} + 1} = \frac{3 - 1}{3 + 1} = 0.5 \qquad (5.5\text{-}4)$$
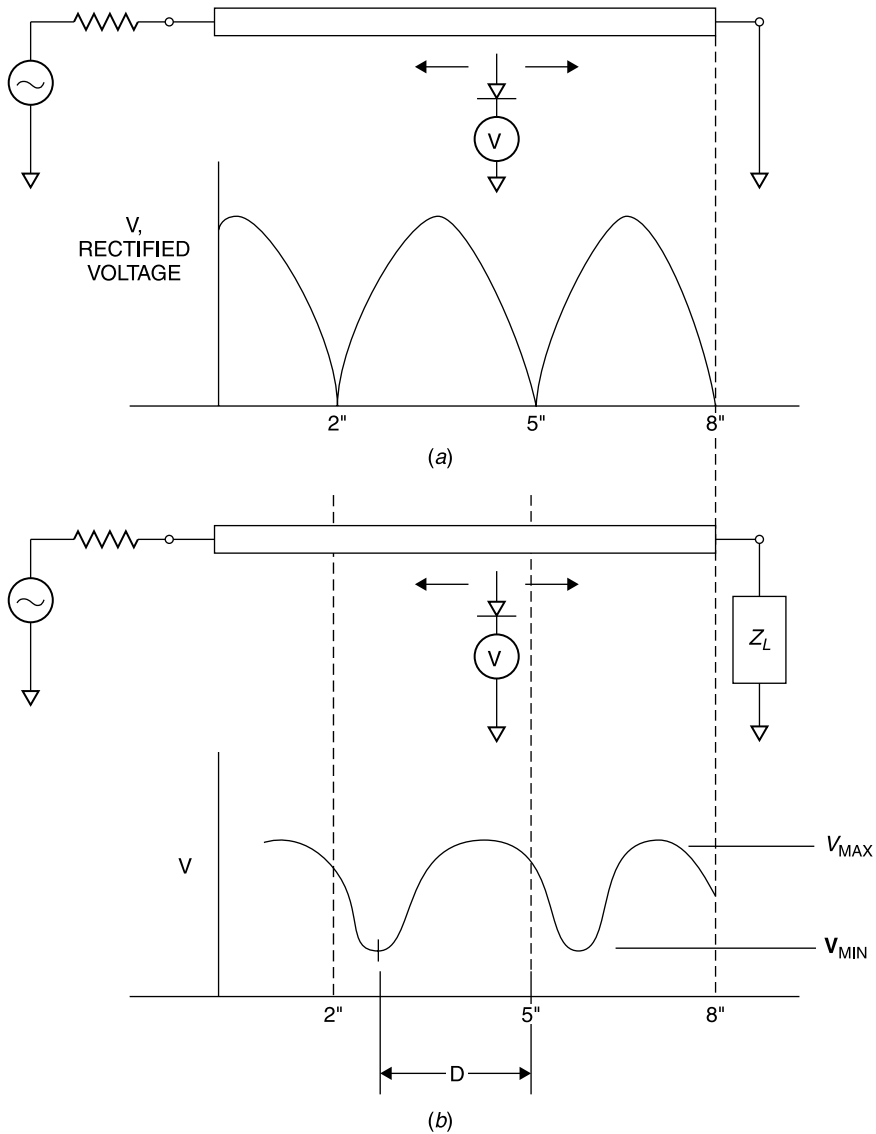
Figure 5.5-1 (*a*) Schematic diagram of slotted-line measurement with short-circuit termination and rectified standing-wave voltage with distance. (*b*) Rectified voltage with an unknown load Z_L.

The null with load in place is located a distance $D = 7.90$ in. $- 5.54$ in. $= 2.36$ in. *toward the generator from the load plane*. To locate the normalized load impedance we convert this distance to wavelengths.

$$D = \frac{2.36 \text{ in.}}{5.90 \text{ in./wavelength}} = 0.40 \text{ wavelength} \tag{5.5-5}$$

On the Smith chart a half wavelength of travel on the transmission line produces a 360° change in the argument of the reflection coefficient:

$$0.5\lambda \text{ of travel on line} \Rightarrow 360° \text{ around chart} \tag{5.5-6}$$

Therefore, in degrees

$$D = 0.4\lambda\left(\frac{360°}{0.5\lambda}\right) = 288° \tag{5.5-7}$$

The circumferential scales on the Smith chart read both in degrees of reflection coefficient angle and in wavelengths. Starting from the short-circuit point $(z = 0, \varphi = 180°)$ the distance D is shown plotted in the Smith chart of Figure 5.5-2. Note that the 288° (0.4λ) arc is drawn *counterclockwise toward the load*
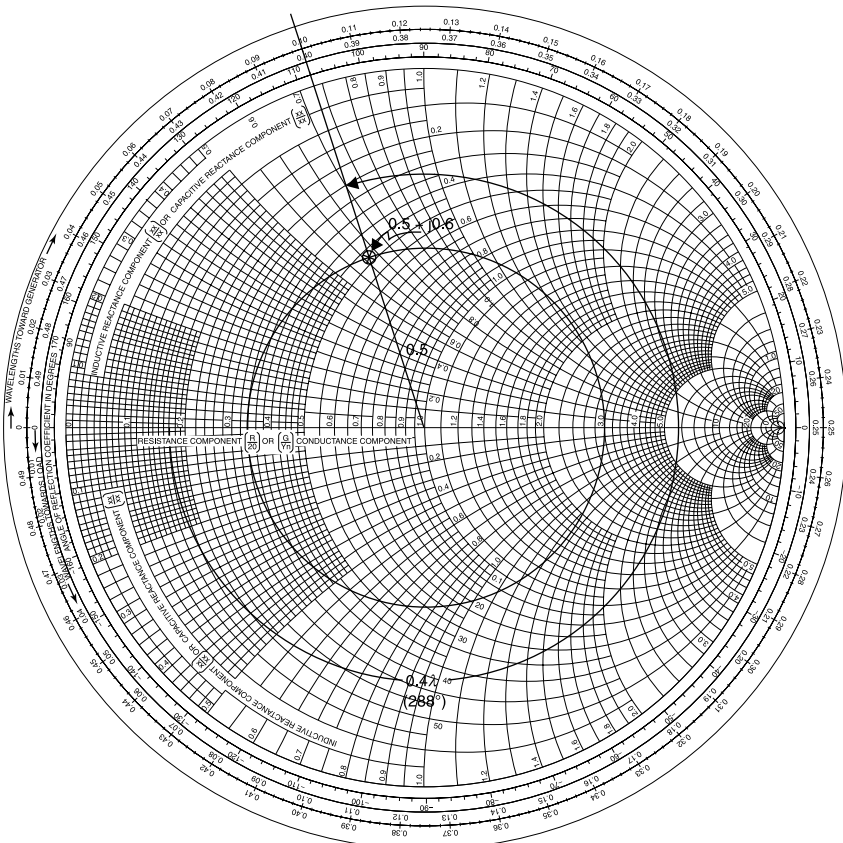
Figure 5.5-2   Smith chart used to compute normalized load impedance from slotted-line measurements. (*Smith chart reproduced through the courtesy of Analog Instrument Co., Box 950, New Providence, NJ 07974.*)

since this is the direction from the load minimum to the plane of the load in question.

The angle of the reflection coefficient at the load is read directly from the peripheral scale on the Smith chart as

$$\varphi = 108° \tag{5.5-8}$$

and hence

$$\Gamma_L = \rho\angle\varphi = 0.5\angle 108° \tag{5.5-9}$$

At this point the normalized load impedance can be read directly from the chart or calculated from

$$z_L = \frac{1 + \Gamma_L}{1 - \Gamma_L} = \frac{1 + 0.5\angle 108°}{1 - 0.5\angle 108°} = 0.5 + j0.6 \tag{5.5-10}$$

Unnormalized this is $(25 + j30)$ Ω, the value used in the example of Figure 4.14-5.

Often the difficult part of the slotted-line measurement is deciding in which direction to rotate on the Smith chart from the null position obtained with the load in place. It may help to remember that if the load impedance were known and plotted on the Smith chart, one would make a negative angle rotation (*clockwise from the load toward the generator*) to the position of the minimum impedance (the $\pm 180°$ axis of the chart). Therefore, to travel from the null *toward the load rotate counterclockwise* on the Smith chart.

The reader will notice from this example, in which the VSWR is 3, that the constant reflection coefficient circle intersects the $r = 3$ circle on the 0° axis of the Smith chart. The circle always intersects the r circle at the value of VSWR, as will be demonstrated in the next section.

5.6   VSWR = *r*

When a constant magnitude reflection coefficient circle of radius greater than zero is drawn on the Smith chart, it intersects the horizontal axis of the chart at two points, r and $1/r$. In the case of a matched load, $\rho = 0$, the circle has zero radius, and the two intersections coalesce into one point, the center of the chart at which $r = 1$, $x = 0$.

For the intersection for which $r > 1$, $z = r + j0$, the reflection coefficient has a real value given by

$$\Gamma = \frac{z - 1}{z + 1} = \rho = \frac{r - 1}{r + 1} \tag{5.6-1}$$

The corresponding VSWR is

$$\text{VSWR} = \frac{1+\rho}{1-\rho} = \frac{r+1+r-1}{r+1-r+1} = r \qquad (5.6\text{-}2)$$

For this reason *the VSWR for any normalized load impedance can be obtained by drawing the constant ρ circle through it. The VSWR is numerically equal to the resistance value r, corresponding to the circle's intercept with the horizontal Smith chart axis for which $r > 1$.*

5.7 NEGATIVE RESISTANCE SMITH CHART

The area that is outside the unit circle of the Smith chart corresponds to reflection coefficient magnitudes greater than one, the negative resistance and conductance domain. The resistance circles defined by (5.2-9) can be plotted for negative values of r, and the circles for constant reactance can be extended outside the $\rho = 1$ circle. For $\rho \le 3.16$ (a reflection gain of 10 dB) the result is that shown in Figure 5.7-1. The example of impedance Z_{IN}/Z_0 and admittance Y_{IN}/Y_0 shown on this standard chart is consistent with identifying the admittance of a point on a $180°$ diametrically opposite location. However, as was noted earlier, the chart can and should be drawn with simultaneously valid Z and Y coordinates in the format of Figure 5.3-2.

Unlike the passive impedance/admittance domain of the Smith chart, the negative resistance domain is unbounded. A new chart must be drawn for each maximum anticipated value of ρ.

The negative resistance domain of the Smith chart (where $\rho > 1$) is defined by the same equations used for the passive Smith chart. This negative resistance domain is useful, for example, to describe reflection amplifiers and to plot the stability circles for transistor amplifiers. However, for most matching designs, the area within the passive impedance Smith chart is sufficient and our matching examples are confined to this area.

5.8 NAVIGATING THE SMITH CHART

For lossless matching, two principal paths, or *contours*, on the Smith chart are employed. These are

1. Constant radius contours for movement along a transmission line and
2. Constant resistance (or conductance) contours for the addition of reactance (or susceptance)

Movement on constant reactance (or susceptance) circles is also possible but implies the addition of resistance (or conductance), consequently lossy match-
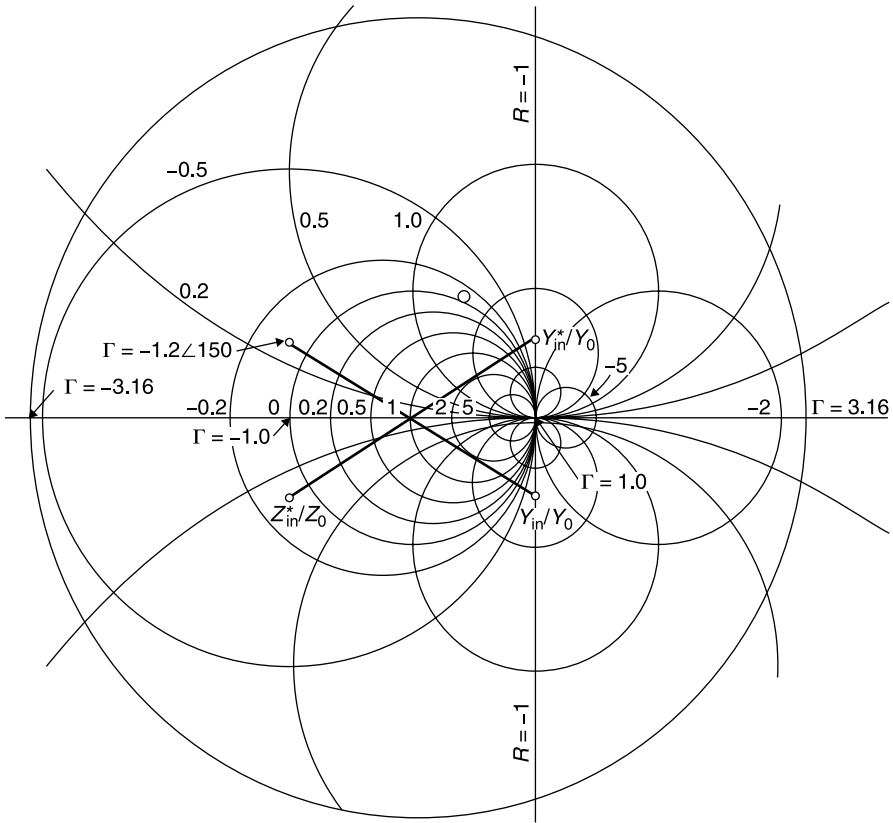
Figure 5.7-1 Smith chart extended to include the negative resistance/conductance domain out to a reflection coefficient magnitude of 3.16. This is sometimes called the "compressed Smith chart." (*Negative* impedance *Smith chart reproduced through the courtesy of Analog Instrument Co., Box 950, New Providence, NJ 07974.*)

ing. Lossy matching has its place, as for very broadband matching, but it is usually avoided in favor of lossless matching.

If the load on a transmission line contains no resistance or conductance, (1) the load can absorb no power, (2) the reflected and incident waves have the same magnitude, (3) the reflection coefficient magnitude is unity ($\rho = 1$), and (4) Γ traces the rim of the passive Smith chart, traversing every reactance (and susceptance) from $-j\infty$ to $+j\infty$ in a transmission line travel of one-half wavelength. This means that open- or short-circuited transmission lines, or stubs, can be used as reactive/susceptive tuning elements.

The next example demonstrates these Smith chart contours and stub matching. A 100-Ω load is connected in series with a 1-pF capacitor. The resulting impedance at 1 GHz of $(100 - j159)\ \Omega$ is to be connected to the end of a 50-Ω transmission line using wire leads having a total of 10 nH of inductance
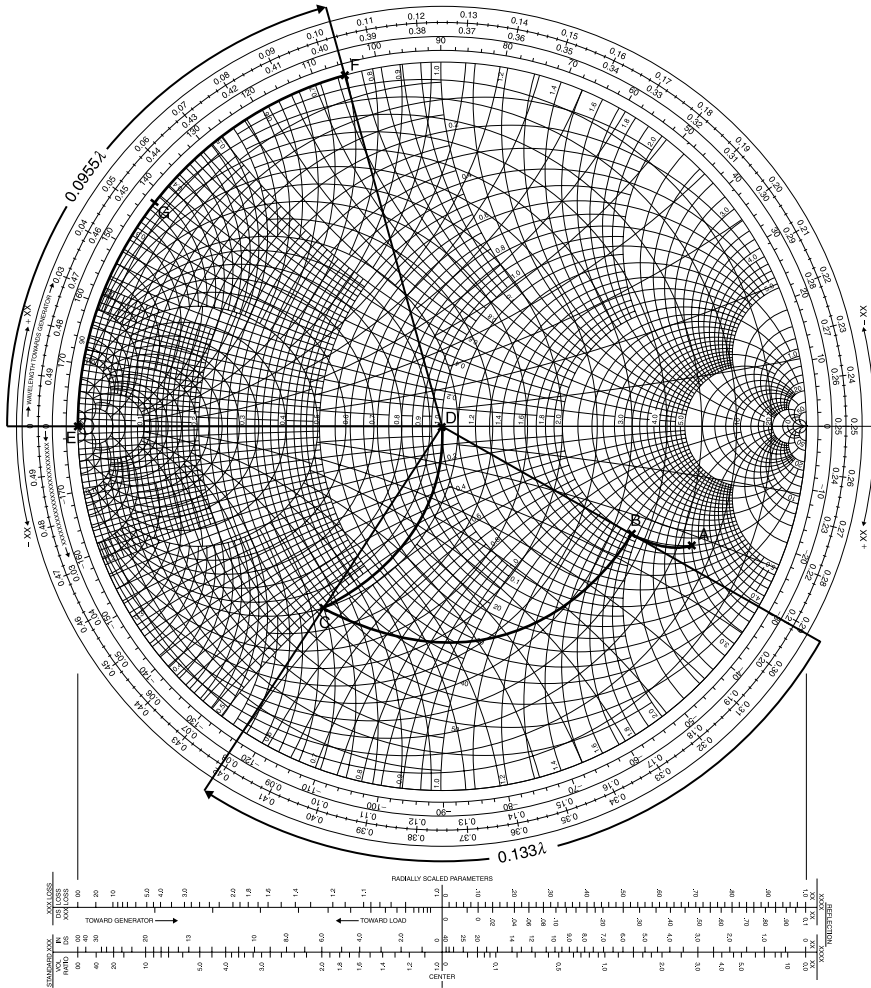
Figure 5.8-1 Procedure to match $(100 - j159)$-Ω load with a $+j63$-Ω series inductance to 50 Ω using a length of 50-Ω line and a short-circuited 50-Ω shunt stub.

($+j63$ Ω at 1 GHz). The resulting load is to be matched to 50 Ω using a shunt tuning stub. The stub may be either open or short circuit terminated. The type of stub termination, the length of the stub, and its location on the main line is to be determined to match this load to 50 Ω (Fig. 5.8-1).

The matching is carried out in the following steps:

1. The reference impedance of the Smith chart is defined to be 50 Ω (the chart is normalized to 50 Ω).

2. The load impedance $(100 - j159)$ Ω is divided by 50 Ω to normalize it, giving $2 - j3.2$ (to the accuracy consistent with the graphical process). This is point A in Figure 4.8-1.

3. The $+j63$-Ω series inductance has a normalized value of $+j1.3$. This is added to the load impedance by *traveling along the constant resistance circle* ($r = 2$) to the reactance intercept $-j1.9$. The initial load (point A) is in the lower half of the Smith chart since its reactance is capacitive. Adding the series inductance causes an upward movement, toward the inductive upper half of the chart, to point B.

4. Drawing a clockwise, constant reflection coefficient magnitude arc intersects the unity *conductance circle* at point C. This point is important. We wish to add a shunt stub. The shunt addition requires the use of admittance coordinates, hence intercept on the unity conductance circle.

   Using the wavelength scale of the chart's periphery, we determine that the starting point B is at 0.292λ and that the $g = 1$ intersection, point C, is at 0.425λ. The distance traveled is $0.425\lambda - 0.292\lambda = 0.133\lambda$, or $48°$. This is the distance from the load at which the shunt tuning stub is to be connected. The length of line between load and stub is $48°$. This is not to be confused with the angular change in Γ, which is $-2(48°) = -96°$.

5. At point C the normalized admittance is read from the chart to be $1 + j1.5$. To obtain a match, add a normalized susceptance of -1.5 thereby *traveling on the $g = 1$ circle* to the center of the chart, point D, completing the match.

6. Short-circuited or open-circuited transmission lines can be used as tuning elements of any desired reactance. We have previously seen that such stubs are described by (4.14-7a) and (4.14-8a), but the Smith chart allows their reactances or susceptances to be read directly. We are free to use any characteristic impedance stub. Suppose for this example that we choose $Z_0 = 50$ Ω.

   Then the same Smith chart normalization applies, and we can perform this calculation on the chart used to determine the matching requirement. To obtain the required $-j1.5$ susceptance (an inductance) we begin with a *shorted stub* at the $y = \infty$ location (point E) on the Smith chart and *travel clockwise along the chart's periphery* to the required inductive susceptance at point F. This requires a stub length of 0.0955λ, or about $34.4°$. It is emphasized that this is a side calculation and this travel is not part of the original contour from A to D.

   Shorted stubs are inductive when their length is less than a quarter wavelength. An open-circuited, 50-Ω stub could be used, but its length would be a quarter wavelength longer (0.3455λ instead of 0.0955λ). Consequently, its frequency sensitivity would be nearly four times greater than that of the shorted stub.

Finally, the same Smith chart can be used even if the tuning stub impedance is different from 50 Ω. For example, suppose a 100-Ω stub is used, then

$$Y_{\text{STUB}} = -j1.5(0.02 \ \text{℧}) = -j0.03 \ \text{℧}$$

$$y_{\text{STUB}} \ (\text{normalized to } 100 \ \Omega) = \frac{-j0.03}{0.01} = -j3$$

This is point G in Figure 5.8-1 and corresponds to a stub length of about 0.05λ. Notice that for this side calculation the Smith chart is normalized to 100 Ω. This is a shorter stub and would take up much less space in the circuit layout, both because it is half the length of the 50-Ω stub and because 100-Ω stubs have much narrower width, hence can be "serpentined" more easily. Shorter stubs also can have less frequency variation, however, both the 50- and 100-Ω stubs of this example are so short that their susceptances would not vary much more rapidly than inversely with frequency.

The completed circuit with 50-Ω stub is shown in Figure 5.8-2 along with its VSWR performance, referenced to 50 Ω. Errors in round off and graphical inaccuracies make the result slightly off of a perfect 1.00 VSWR and the frequency of best match slightly above the design value of 1000 MHz. This result greatly improves the efficiency of power transmission to the load, as indicated below.

When the arc described by points B and C is extended to the $0°$ axis of the Smith chart, it intersects it at about $r = 4.2$. Therefore, the VSWR of the load,
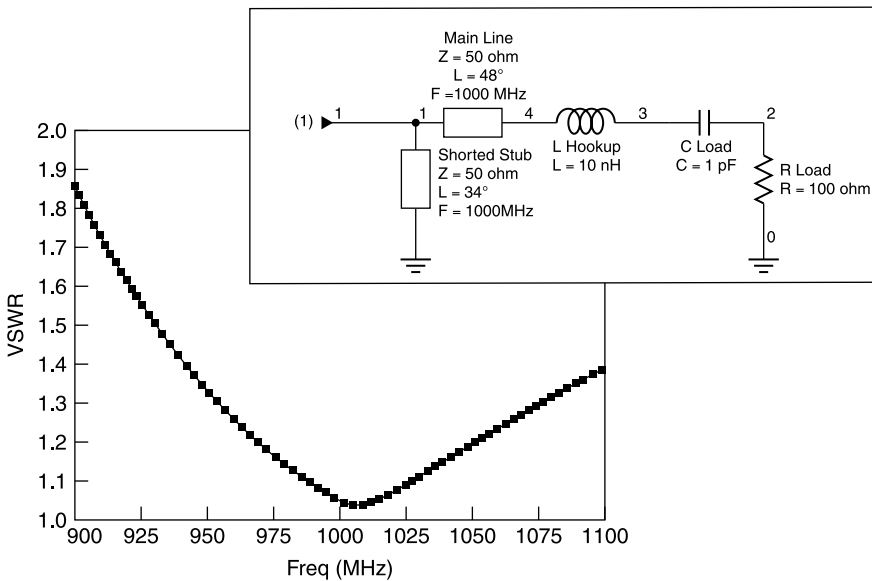
Figure 5.8-2   VSWR performance of stub tuned load.

including its series inductance, is 4.2. The reflection coefficient ρ is 0.62. This is read using the auxiliary scale (not shown in the figure) at the bottom of the standard Smith chart sheet or calculated from VSWR and (4.5-3b).

Without the stub tuning, the return loss (fraction of power returned to the generator) of the load before tuning would have been

$$RL = \rho^2 \times 100\% = 38\% \tag{5.8-1}$$

Thus the mismatch loss is 62%, which corresponds to a mismatch loss of 2.1 dB. With the stub tuning, the VSWR is below 1.1 and ρ is below 0.05 from $(990 - 1020)$ MHz. The corresponding return loss is less than 0.3%. This improvement in efficiency, easily designed using the Smith chart, is obtained merely by printing a side stub on the transmission line and grounding it.

5.9 SMITH CHART SOFTWARE

Although the procedures for using the Smith chart as a graphical tool are fairly simple, it is easy to err by rotating in the wrong direction, misreading the coordinate values or trying to match using the incorrect circle. To minimize errors, remember to: *Rotate to the* $r = 1$ *circle when adding a series reactance* and *rotate to the* $g = 1$ *circle when adding a shunt susceptance.*

Furthermore, the accuracy of the procedure is limited by the graphical process itself. Finally, the Smith chart with both impedance and admittance coordinates is a crowded format that is difficult to read.

These problems are considerably lessened through the use of *Smith chart software* written for the personal computer. The following examples are illustrated using one such program called winSmith [6].

For example, given a load of $(25 + j30)$ Ω at 2000 MHz terminating a 50-Ω transmission line, it is desired to match this load to 50 Ω using a tuning element connected in shunt with the transmission line at an appropriate distance from the load on the generator side. What is the electrical spacing, element type, and value for this tuning element?

This load point is entered as $(25 + j30)$ Ω directly into the program along with the reference impedance for the chart. The program automatically locates impedances on the chart given the reference impedance so there is no need for normalization by the user. The parameters of the cursor-selected point (x in Fig. 5.9-1) are listed in the upper right corner of the display. The units of Z and Y are ohms and mhos, respectively. The S term is S_{11} with magnitude in decibels and angle in degrees (S parameters are discussed in 6.4). The term G stands for gamma (Γ), the reflection coefficient, listed here as $0.49\angle 108°$. This is a more precise value than the $0.5\angle 108°$ value that we read graphically from the Smith chart in the earlier example. The corresponding VSWR, listed as "V" in the display, is 2.89, again a more accurate value than the 3.0 value previously read graphically from the chart.
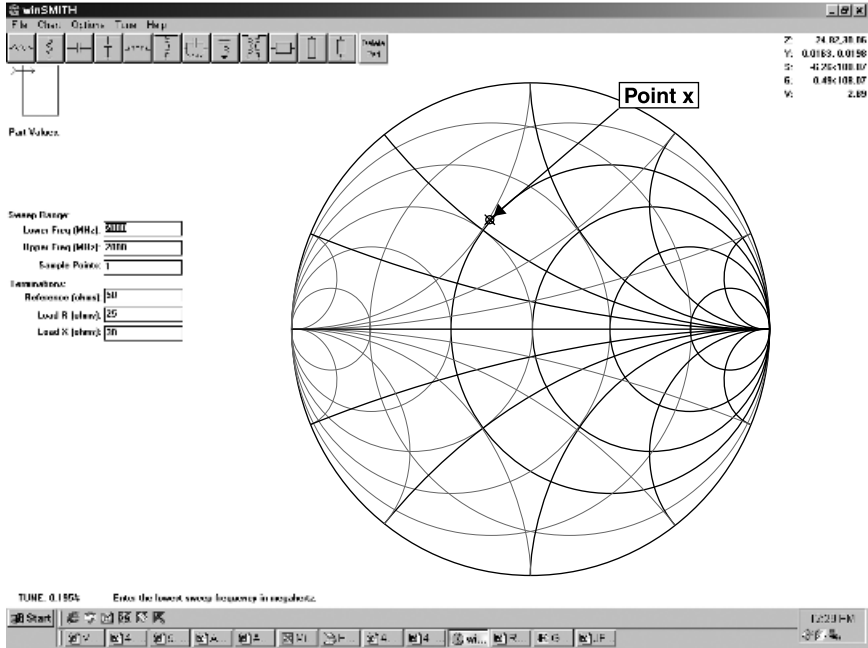
Figure 5.9-1   Personal computer display of Smith chart showing $(25 + j30)$-Ω load on 50-Ω transmission line. (*Produced using winSmith software* [6].)

A selection of pull-down components in the upper left of the display allows the construction of a cascade circuit layout of various topologies. The resulting Smith chart contour is automatically computed and displayed for the circuit chosen and the variable values entered.

Both impedance (dark lines) and admittance (light lines) coordinates are shown. Recall that the location of the load is the same using this Smith chart format, whether the load is expressed in impedance or admittance coordinates. Unlike the graphical Smith chart, the software displayed Smith chart is not normalized, since the value of Z_0 is also entered. The impedance and admittance are read directly in ohms and mhos, respectively.

The tuning procedure is simplified further by selecting for display only the desired coordinates, impedance, or admittance appropriate to the current movement being made on the chart. We wish to move along the 50-Ω line, a constant reflection coefficient magnitude, from the load (now relabeled point A in Fig. 5.9-2) to the intersection with the 0.02-℧ conductance circle (point B). Here the addition of an appropriate shunt susceptance will match the input to the line to 50 Ω (0.02 ℧). The admittance at point B lies in the lower half of the Smith chart, and therefore its imaginary part corresponds to a capacitor (positive susceptance). It is essential to keep in mind that *the lower half of the Smith*
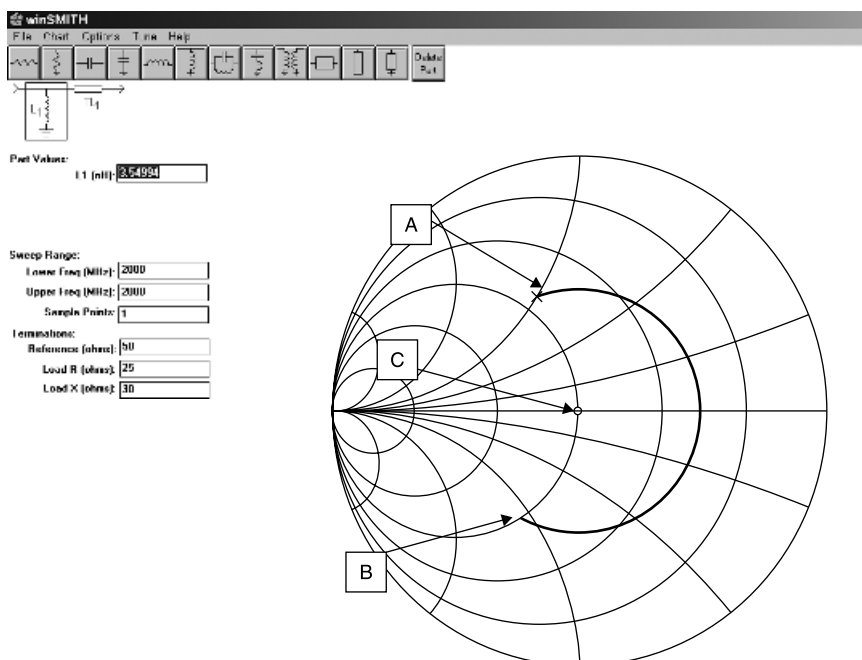
Figure 5.9-2   Smith chart used to determine shunt tuning of the $(25 + j30)$-Ω load on a $Z_0 = 50$ Ω transmission line.

chart is capacitive, and the upper half is inductive, whether expressed as impedance or admittance.

From the Smith chart, the 50-Ω line length required to move from point A to the unity conductance circle (point B) is $114°$ (0.32λ). Notice that this results in a clockwise rotation of $-228°$ on the Smith chart. From the cursor reading (not visible in this view), the admittance at point B was read to be $(0.02 + j0.022)$ $\mho$. The addition of a 3.54-nH shunt inductor (having susceptance of $-j0.022$ $\mho$ at 2 GHz) moves the contour to point C, tuning the load to 50 Ω, as shown in Figure 5.9-2.

5.10   ESTIMATING BANDWIDTH ON THE SMITH CHART

Each contour on the Smith chart applies at a single frequency. However, a measure of the bandwidth of a particular tuning procedure can be obtained by adding a contour at each band edge. Suppose that we are interested in the bandwidth 1800 to 2200 MHz for the previous example, shown plotted in Figure 5.9-2. Adding these band edge contours yields the results in Figure 5.10-1. The load has been modeled as a 25-Ω resistor in series with a 2.39-nH inductor $(+j30$ Ω at 2000 MHz) in order to simulate the frequency variation of the load.
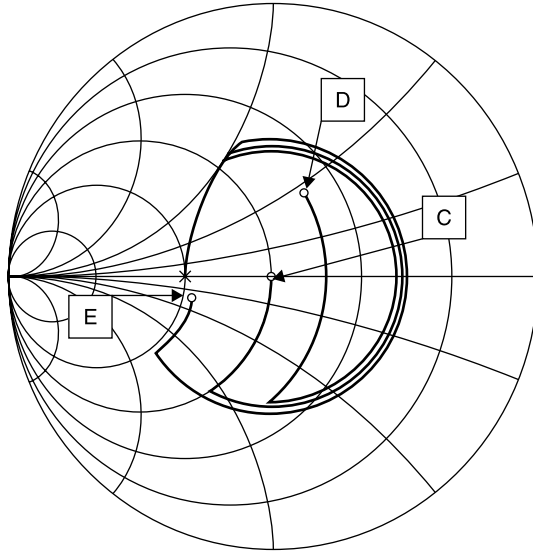
Figure 5.10-1   Smith chart contours for 1800 MHz (point *D*), 2000 (point *C*), and 2200 MHz (point *E*) for the previous tuning example.

Notice that the contours for 1800 MHz (terminating at point *D*) and 2200 MHz (terminating at point *E*) show large variations in the angle for the transmission line section. This is because the line length changes by $\pm11.4°$ at these band edges. This change is further amplified because on the Smith chart the angular change is doubled to $\pm22.8°$. Using the cursor, the VSWR values at the end of these excursions were 1.94 at 1800 MHz and 1.96 at 2200 MHz. The resulting mismatch loss is about 0.5 dB.

5.11   APPROXIMATE TUNING MAY BE BETTER

The engineer who first learns matching with the Smith chart is inclined to try for a "perfect match" at the design center frequency, on the assumption that this will also give the best performance over a given bandwidth. But this assumption is not necessarily valid. In some cases an approximate solution at the center frequency may give better overall performance. In addition, settling for a reasonably "close match" at the center frequency might be obtainable using more economical tuning elements, while actually providing a better overall match throughout the required frequency bandwidth. The following example illustrates such a case.

Referring to Figure 5.9-1, we see that the $(25 + j30)\ \Omega$ starting impedance (point *A*) lies close to but is not on the unity conductance circle. At this point
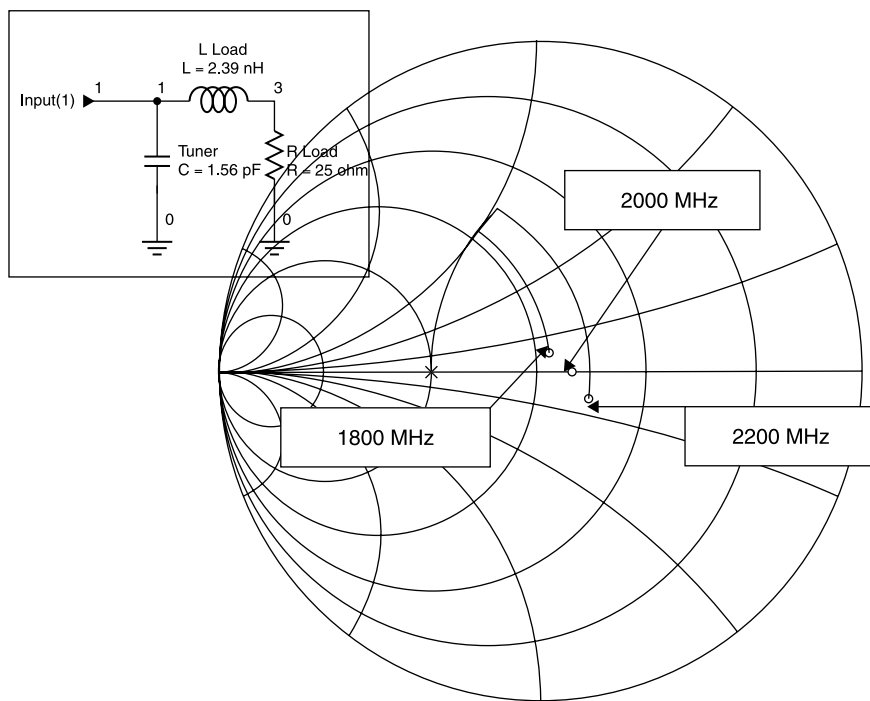
Figure 5.11-1 Result of tuning the $(25 + j30)$-Ω load in an approximate manner at f_0 using a parallel capacitor.

the admittance is $(0.0164 - j0.197)$ ℧. If a capacitor is added in shunt with the load (point A) having a susceptance of $+j0.0197$ ℧ $(-j50.8\ \Omega)$ at 2000 MHz, it must have a capacitance of 1.56 pF. The resulting total admittance at point A would be 0.0164 ℧ (61 Ω real), corresponding to a 1.22 VSWR in a 50-Ω system. With this approximate tuning at the center frequency, the three frequency contours are as shown in Figure 5.11-1.

With this tuning the VSWR values are 1.15, 1.22, and 1.42 at 1800, 2000, and 2200 MHz, respectively. The highest mismatch loss in the band, at 2200 MHz, is only 0.13 dB, much better than was obtained when the load was perfectly matched at 2000 MHz using the transmission line and shunt inductor.

This example further exemplifies the advantage of the Smith chart. The opportunity for approximate tuning was evident when the load was shown with admitttance coordinates on the Smith chart. The example demonstrates two points about matching.

First, *a broader band match usually can be obtained when the tuning is performed close to the load.* Second, *accepting an approximate match at the center frequency may result in a better average match over the operating band.*

5.12   FREQUENCY CONTOURS ON THE SMITH CHART

If the points in Figures 5.10-1 and 5.11-1 that describe low to high frequency impedances for the same location in the circuit are connected with a line, they can be seen to describe a *clockwise arc*. This is always the case, a consequence of Foster's reactance theorem.

   Foster's reactance theorem states that the reactive portion of impedances and the susceptive portion of admittances increase positively as frequency increases.

   Impedances and admittances on the Smith chart trace clockwise arcs as frequency is increased.

   Often the swept frequency response of a network analyzer is presented on a Smith chart display without frequency labels. The clockwise increase with frequency is a ready means of identifying the low and high frequency ends of the display trace.

5.13   USING THE SMITH CHART WITHOUT TRANSMISSION LINES

Notice that in the preceding example there were no transformations along transmission lines (constant reflection coefficient magnitude arcs). The VSWR values were calculated as those which would be incurred *if* the load and its tuner were connected to a 50-Ω line.

   The Smith chart can be used without transmission lines, even though it was developed as an aid to transforming impedances along them. It is an orthogonal set of resistance and reactance contours (or conductance and susceptance contours). It is especially useful for performing matching calculations because every passive impedance or admittance, including those with infinite magnitudes, can be plotted within the unit circle enclosed by the $|\Gamma| \leq 1$ Smith chart. This is not true of a rectangular impedance or admittance chart for any given scale factor, whose dimensions increase without bound with the magnitudes of the impedance and admittance components to be plotted.

   When using the Smith chart without an associated transmission line, we are free to select arbitrarily its reference impedance, the Z_0 impedance to which all others are normalized. For the best graphical accuracy it is desirable to select the reference impedance near the geometric mean between the largest and smallest impedance values to be plotted.

   For example, consider the Q matching example of Figure 3.5-2. We desired to Q match a 5-Ω resistor to 50 Ω at 1 GHz. To demonstrate how this matching could have been performed using the Smith chart instead of the Q matching method, we start by selecting the reference impedance of the chart arbitrarily to be 25 Ω. The 5 Ω resistance is first plotted on the Smith chart (point A) and a value of series inductance added to bring the contour to the intersection with the $G = 0.02$ ℧ conductance circle (point B), as shown in Figure 5.13-1.

   At this point the cursor indicates a susceptance of $-j0.06$ ℧. In reactance this is $+j16.7$ Ω, which is resonated at 1 GHz using a capacitor of 9.54 pF.
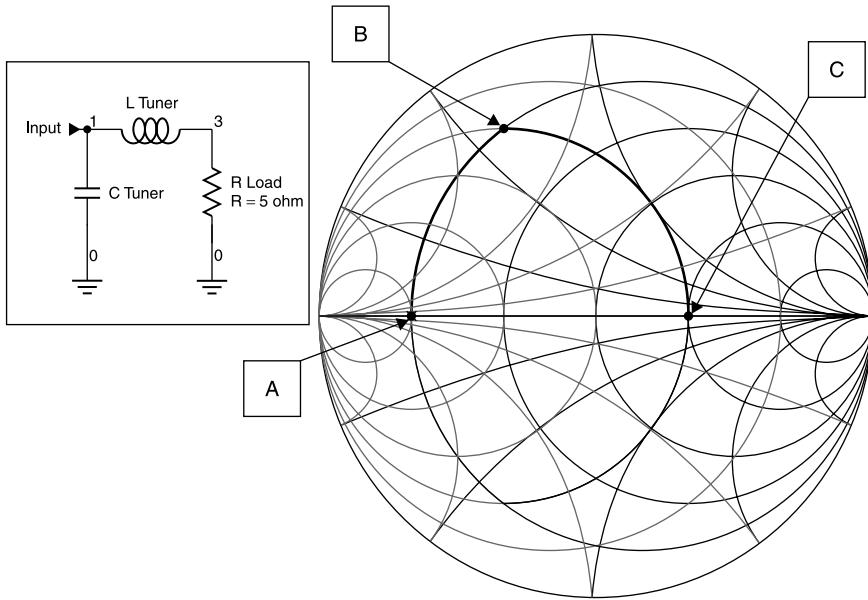
Figure 5.13-1   Smith chart used for *LC* matching without any transmission line refer-ences.

When this value is added, the contour extends to point *C*, the $R = 50\ \Omega$ value on the 25 Ω reference impedance Smith chart. These tuning elements are the same as those derived earlier using the *Q* matching procedure (see Figure 3.5-2).

5.14   CONSTANT *Q* CIRCLES

In a Cartesian-type impedance plane the contours of constant *Q* are defined by the straight lines $X = QR$, where *X* is the reactance and *R* the resistance of a series impedance. We saw that the Smith chart is produced by a bilateral transformation having the property that circles are transformed into circles, straight lines being cases of circles of infinite radius. Thus we expect that con-stant *Q* lines drawn in the Smith chart (Γ) plane will be circular or straight lines. This is the case. When the $Q = 3$ circles are drawn in the Smith chart of the previous example, the result is that shown in Figure 5.14-1.

Recall that $Q = 3$ as a condition of the *Q* matching procedure used in Figure 3.5-2. Given this *Q* value, the 5- to 50-Ω transformation could be performed on the Smith chart by finding the intersections with the $Q = 3$ circles.

When *Q* matching is used, consisting of reactances, which alternate in type (*L* or *C*) and in topology (series/parallel), adhering to a low *Q* value ensures broad bandwidth. When viewed on the Smith chart the matching contour
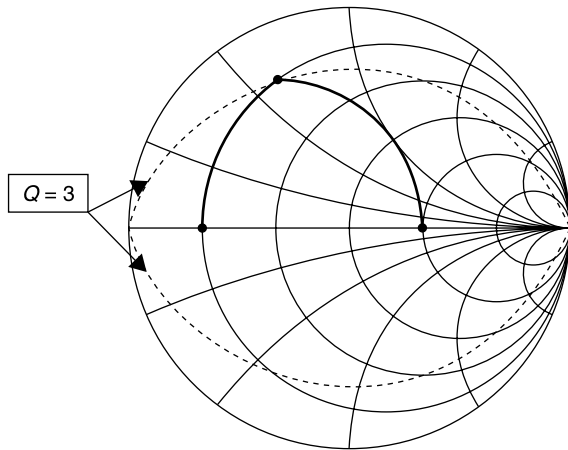
Figure 5.14-1   Smith chart of previous example with $Q = 3$ circles.

remains close to the horizontal $(X = 0)$ axis for low Q values. In the broadband matching example of Figure 3.5-5, the Q was reduced to 1.074. When this example is plotted on the Smith chart, the result is that shown in Figure 5.14-2.

Notice that the Q circles are symmetric with respect to the $x = 0$ axis of the Smith chart. This suggests that the dual of this matching circuit, consisting of a series capacitor connected to the 5-Ω load, followed by a shunt inductor and alternating C and L elements thereafter would produce the same result. This is the case, as shown in Figure 5.14-3.

The element values shown in Figure 5.14-3 were found graphically by extending the contour segment for each tuning element addition to intersect alternately the $Q = 1.074$ circle and the horizontal axis. As such, they are
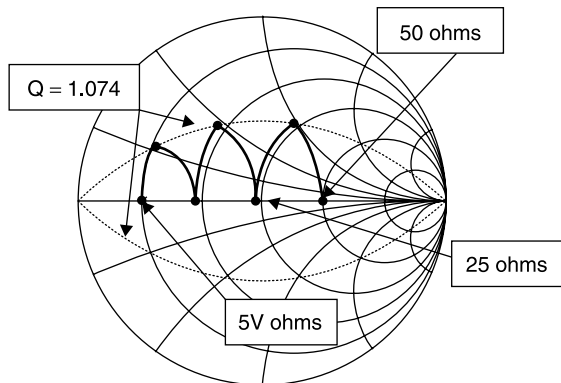
Figure 5.14-2   Broadband LC matching circuit plotted on Smith chart with 25 Ω reference impedance.
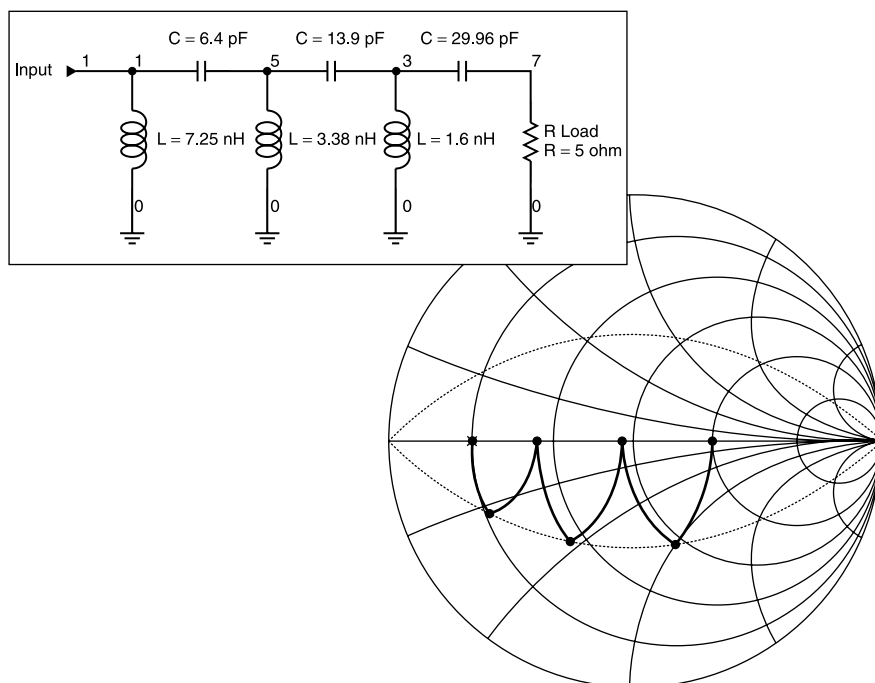
Figure 5.14-3 Dual of broadband tuning circuit derived graphically using lower $Q = 1.074$ circle.

approximate. The final circuit obtained using these approximate values transforms 5 to 49.5 Ω. With more careful graphical techniques, a closer approximation to 50 Ω could be obtained, but the associated VSWR of 1.01 of this result would be sufficient for practically all applications.

5.15 TRANSMISSION LINE LUMPED CIRCUIT EQUIVALENT

Often the properties of a uniform transmission line section are desired but space does not permit the use of an actual section of line. An example is the realization of the branch line hybrid coupler (covered in Section 8.9) using lumped elements. In such cases a lumped equivalent circuit for a transmission line section is useful. The equivalent circuit is only equivalent at one frequency, but the approximation over a modest bandwidth is adequate for many applications.

As an example, suppose that we wish to simulate a 90° length of 50-Ω line using an *LCL* tee circuit at 1 GHz (Fig. 5.15-1). What are the element values for the equivalent circuit?

(*b*) Case 1   (*c*) Case 5
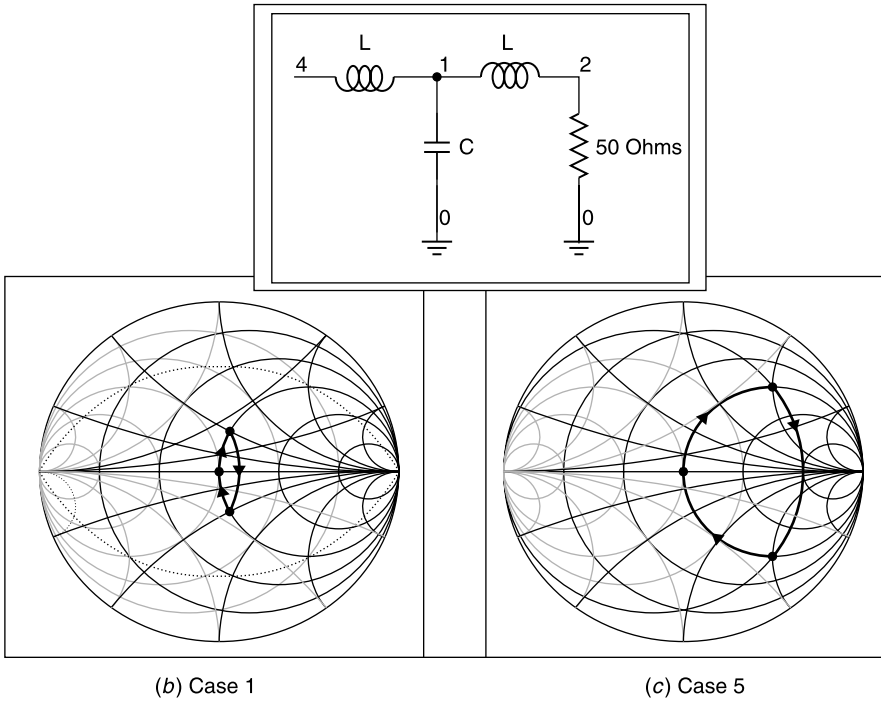
Figure 5.15-1   Combinations 1 and 5 yielding 50-Ω input for *LCL* tee circuit.

We recognize that any length of 50-Ω transmission line presents an input impedance of 50 Ω when the line is terminated in 50 Ω. After some experimentation on the Smith chart, we find that, starting with a load of 50 Ω (and normalizing the chart to this value), there are any number of *L* and *C* reactance combinations that will yield 50 Ω input impedance. A set of five combinations is shown in Table 5.15-1.

It might seem that this Smith chart application has failed since it does not provide a unique circuit solution, but this is not so. We have only required that the circuit deliver 50 Ω to the input when the load is 50 Ω, but *any 50-Ω line length does so*; and the Smith chart results properly indicate that numerous equivalent circuits exist for line sections of various lengths. To determine the

TABLE 5.15-1   Sets of *L* and *C* Values Yielding 50 Ω Input to Tee Circuit

Combination	1	2	3	4	5
X_L	25	35	50	75	100
X_C	−63	−53	−50	−54.1	−62.6
L	3.98 nH	5.57 nH	7.96 nH	11.94 nH	15.92 nH
C	2.52 pF	3.0 pF	3.18 pF	2.94 pF	2.54 pF
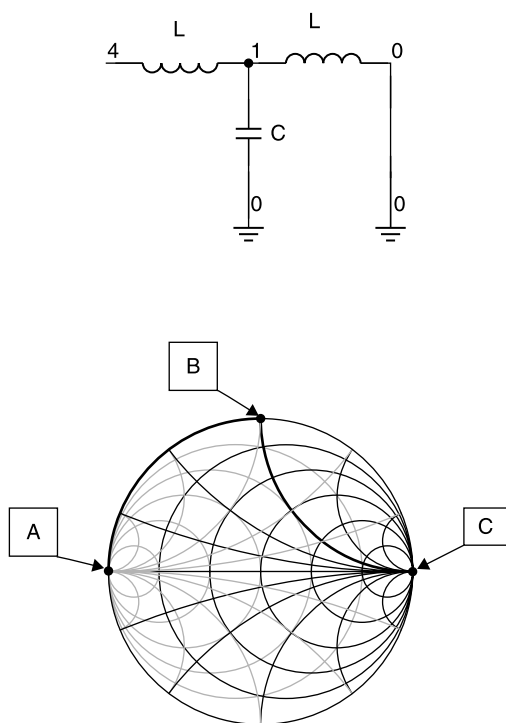
Figure 5.15-2   Smith chart plot of tee circuit with short-circuit load.

equivalent circuit specifically for a 90° long, 50-Ω line section, we must add another condition that relates input and output impedances unique to a 90° long, 50-Ω line section. One such combination is a short-circuit load, which is transformed to an open circuit by a 90° line section as indicated by (4.14-7a). Among the test cases of Table 5.15-1, we find that case 3 yields this result, as shown in Figure 5.15-2.

Initially, we locate the short circuit at point *A*. Adding a series reactance of +50 Ω takes the contour to point *B*. Note that this is the $z = 0 + j1$ point on the chart normalized to 50 Ω. Next we must use admittance since the capacitor of the tee circuit is a shunt element. However, the contour remains on the periphery of the Smith chart because both the $r = 0$ and the $g = 0$ circles lie on the $|\Gamma| = 1$ circle. To reach the horizontal axis we must add a normalized capacitive susceptance of +1. Notice that this brings us to the infinite impedance, point *C*, on the Smith chart, regardless of the value of the remaining series inductor. However, this second inductor is necessary since only the cases listed in Table 5.15-1 give a matched 50-Ω input when the circuit is loaded with *50* Ω. Given this result, it is now clear why the capacitive and inductive reactances must be equal; they must be *parallel resonant* in order that the input
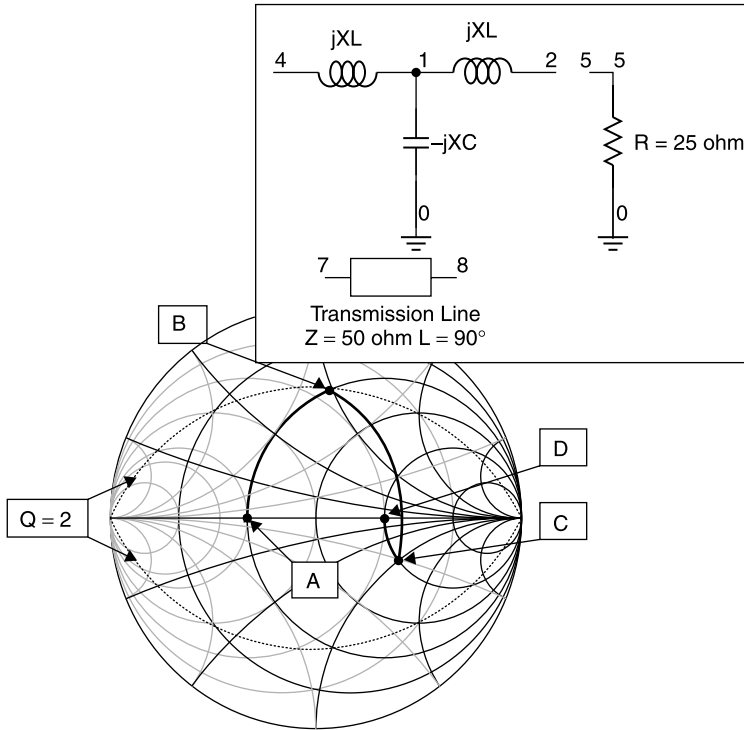
Figure 5.15-3 The circuit and Smith chart contour of the three-element, tee section equivalent circuit of a 90° long, 50-Ω transmission line at 1 GHz.

will be an open circuit when the output of the tee circuit is short circuited. While this is now obvious, the Smith chart revealed this fact graphically.

As a test of this equivalent circuit, consider another pair of load and input impedances. We noted earlier that a quarter-wave transmission line behaves as an impedance inverter. For example, when terminated in *25* Ω, the 90° section of a *50*-Ω line presents *100* Ω at its input terminals. The tee circuit must provide the same transformation if it is to be equivalent.

We begin at point *A* (Fig. 5.15-3), which represents the 25-Ω load on a Smith chart normalized to 50 Ω. Adding a series inductor follows a constant resistance circle to point *B*. Adding a shunt capacitor follows a constant conductance circle to point *C*. Then adding a series inductor of the same value as used between points *A* and *B* follows a constant resistance contour to point *D*, the 100 Ω required input resistance value.

This three-element tee equivalent of a transmission line section can be analyzed exactly using the *ABCD* matrix approach of the next chapter. However, the matrix solution is more difficult to apply for a five-element equivalent circuit, as might be employed to achieve a broader bandwidth of operation.
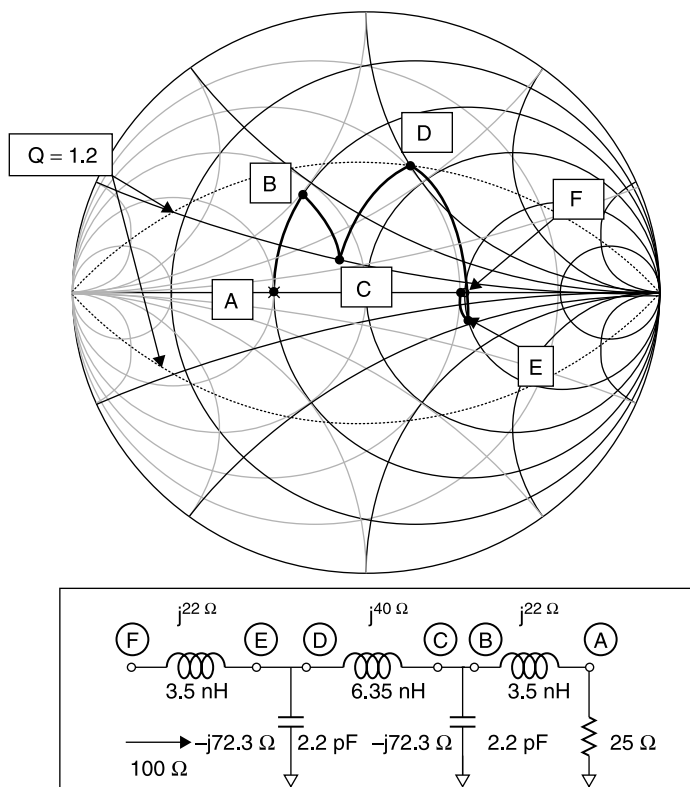
Figure 5.15-4 Smith chart determination of symmetric five-element circuit that is equivalent to a 90°, 50-Ω transmission line at 1 GHz.

The five-element model was designed using the Smith chart, maintaining circuit symmetry (end two elements are equal to each other, as are the second and fourth elements) and requiring that a *25*-Ω load result in a *100*-Ω input. After another cut-and-try process on the Smith chart, the symmetric five-element circuit and its contour shown in Figure 5.15-4 was obtained. The results are approximate, but as will be seen, the match is quite close to 1.00 at the center frequency.

Notice that the contour is just inside a $Q = 1.2$ circle drawn on the chart, while the contour for the three-element circuit is just inside the $Q = 2$ circle. Due to the lower Q of the five-element circuit we would expect broader bandwidth for it, and this is the case as can be seen in Figure 5.15-5. For these simulations the input VSWR is referenced to 100 Ω in order to indicate the deviation of Z_{IN} from the required 100-Ω value.

These examples show that the Smith chart can be employed to design matching networks having considerable complexity, which might otherwise make them more difficult to design by purely analytical means. The examples
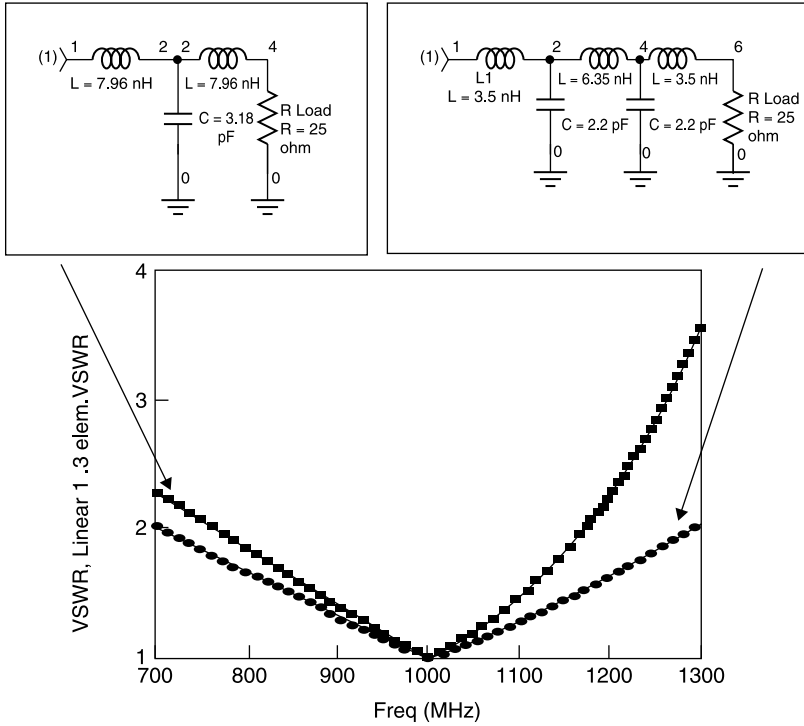
Figure 5.15-5   Performance comparison of three-element and five-element transmission line equivalent circuits designed graphically using the Smith chart.

also are intended to demonstrate that what can be obtained using the Smith chart is only limited by the imagination and skill of its user.

REFERENCES

1. Phillip H. Smith, ''A transmission line calculator,'' *Electronics*, Vol. 12, No. 1, 1939, p. 29. *Smith's original paper describing what became known as the ''Smith chart.''*

2. Phillip H. Smith, ''An improved transmission line calculator,'' *Electronics*, Vol. 17, No. 1, 1944, p. 130. *Refinements to Smith's original paper.*

3. Phillip H. Smith, *Electronic Applications of the Smith chart*, Noble Publishing, Norcross, GA, 1995. *This is Smith's original book on the Smith chart, first published by McGraw-Hill in 1969 with the assistance of Kay Electric.*

4. Kenneth S. Miller, *Advanced Complex Calculus*, Harper & Brothers, New York, 1960. *An old reference, much too advanced for our purposes, but its development of the properties of the bilinear transformation is fundamental to the realizability of the Smith chart.*

5. Chris Bowick, *RF Circuit Design*, Butterworth-Heinemann, Woburn, MA, 1982. *This is a very readable and practical design guide to filter design and matching circuits.*

6. *winSmith 2.0*, Noble Publishing, Norcross, GA, 1995, 1998. *A personal computer software program for performing Smith chart computations. Mention this textbook for a courtesy discount.*

EXERCISES

E5.1-1   Derive an analytic expression for the input admittance Y_{IN} to a lossless, Z_0 transmission line that is θ degrees long and terminated in an admittance Y_L.

E5.3-1   Derive in terms of Y_0 and electrical length, θ, expressions for the susceptance of (a) open-circuited lossless transmission lines and (b) short-circuited lossless lines.

E5.3-2   A 50-Ω transmission line has a matched generator with available average power of 100 W. What is the voltage 45° from the load if the load is $(50 + j50)$ Ω?

E5.4-1   A *50*-Ω microstrip transmission line has a load of $(50 - j50)$ Ω that is to be matched using a shunt *50*-Ω, open-circuit terminated stub. Since this method of tuning can be "printed" on the board and requires no "via" to connect it to ground, it is expected to be the most economical, reproducible, and vibration-resistant tuning method. Assume the load impedance is constant over the bandwidth.
   a. Use the Smith chart to determine the location and electrical length of the stub needed for the matching.
   b. Use the Smith chart to determine the VSWR at the edges of a $\pm 10\%$ bandwidth.

E5.4-2   Can you find a 50-Ω perfectly matched input solution to the problem in E5.4-1 using a properly chosen Z_0 and electrical length? If so, what are the values?

E5.4-3   A transistor has an input, which consists of 25 Ω shunted by 6.4 pF. Use the Smith chart to design a matching network to 50 Ω at 1 GHz using only a 50-Ω transmission line and an open-circuited 50-Ω shunt stub.

E5.4-4   Can you find a way to tune the amplifier input of E5.4-2 using only transmission lines that have a lower VSWR over the 900 to 1100 MHz bandwidth?

E5.6-1   What are the maximum and minimum resistances that will be seen along a 50-Ω transmission line having a length of at least one-half wavelength and a load VSWR of 3?

E5.11-1   A patch antenna at 850 MHz is measured to have an impedance of $(5 - j25)$ Ω. Use the Smith chart to design a matching network to 50 Ω as well as you can (using an approximate method is acceptable) *using two in-line (cascaded) transmission line sections* of appropriate characteristic impedances.

E5.11-2   Repeat Exercise 5.11-1, this time using lumped elements and the Q matching method.

E5.11-3   Repeat Exercise 5.11-2, this time using only a cascade section of a 50-Ω line and an open-circuited, shunt 50-Ω line stub. Use the Smith chart to determine the length of the shunt stub and its location on the main line.

E5.11-4   Use a network simulator to compare the bandwidths for -20-dB return loss for the solutions that you obtained for Exercises 5.11-1, 5.11-2, and 5.11-3.

E5.11-5   **a.** Can you find a way to tune the load in E5.4-1 using an approximate method that would satisfy the requirement for an all printed matching network and give a VSWR over the band no higher than that achieved in the solution of E5.4-1.

   b. Explain why this matching method gives lower VSWR over the operating bandwidth.

E5.14-1   Only two reactive elements are needed to conjugately match any complex impedance load to any complex impedance generator. Demonstrate that there is an unlimited number of matching networks possible by using 6 to 10 reactive elements (L's and C's) to transform 10 to 50 Ω. Use a "random walk" on the Smith chart that starts at 10 Ω and eventually, within 6 to 10 elements, "arrives" at 50 Ω.

E5.14-2   **a.** Estimate the Q of a single LC matching network needed to transform 7 to 50 Ω at 1000 MHz.

   b. Draw Q circles on the Smith chart and use them as a guide to design a matching network consisting of a series L and a shunt C.

   c. Change the reactances of the two tuning elements to what they become at 900 and 1100 MHz. Plot their respective contours on the Smith chart. They will not quite transform to 50 Ω but to some other impedances. What are the VSWR values at these impedances (at therefore at 900 and 1100 MHz)?

E5.14-3   Repeat E5.14-2, but this time use a three-section (six-element) tuner. Does this give a better match (lower VSWR) over the 900 to 1100 MHz bandwidth?

Matrix Analysis

6.1 MATRIX ALGEBRA

For our purposes, *a matrix is an ordered array of the coefficients of a set of linear simultaneous equations*. For example, a linear equation in two variables, x and y, describes a line in the xy plane. Two such equations describe two lines. If the lines are not parallel and one does lie on top of the other, they will intersect. The intersection corresponds to a common solution for x and y that satisfies both equations:

$$a_{11}x + a_{12}y = a_{13} \tag{6.1-1a}$$

$$a_{21}x + a_{22}y = a_{23} \tag{6.1-1b}$$

In matrix notation these equations are written

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_{13} \\ a_{23} \end{pmatrix} \tag{6.1-2}$$

The rules for matrix manipulation allow us to recover from (6.1-2) the same information contained in (6.1-1a,b). *To multiply an $n \times n$ square matrix A by an $n \times 1$ column matrix B, form the resultant column matrix C whose respective elements j are equal to the sum of the products of the jth row of A by the corresponding column elements of B.*

Explicitly this rule is demonstrated for a 2×2 matrix multiplying a 2×1 matrix as

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} (a_{11}b_1 + a_{12}b_2) \\ (a_{21}b_1 + a_{22}b_2) \end{pmatrix} = \begin{pmatrix} a_{13} \\ a_{23} \end{pmatrix} \tag{6.1-3}$$

To multiply a square matrix by another square matrix the rule is: *Multiplying an $n \times n$ matrix by another $n \times n$ matrix produces an $n \times n$ product matrix whose*

ij terms are the sum of the products obtained by multiplying each element in the *i*th row of the first matrix by the corresponding elements in the *j*th column of the second matrix.

For example, multiplying a 2×2 by another 2×2 matrix yields

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} (a_{11}b_{11} + a_{12}b_{21}) & (a_{11}b_{12} + a_{12}b_{22}) \\ (a_{21}b_{11} + a_{22}b_{21}) & (a_{21}b_{12} + a_{22}b_{22}) \end{pmatrix} \quad (6.1\text{-}4)$$

We also define that *two matrices are equal to each other if every element of the first is equal to the corresponding element of the second.*

Given these rules, (6.1-3) is equivalent to (6.1-1a,b) where in this case $b_1 = x$ and $b_2 = y$. Some additional rules are: *To add two matrices together, form the sum matrix whose elements are the sum of the respective elements of the two matrices.* Thus

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} (a_{11} + b_{11}) & (a_{12} + b_{12}) \\ (a_{21} + b_{21}) & (a_{22} + b_{22}) \end{pmatrix} \quad (6.1\text{-}5)$$

Similarly, *to subtract a matrix B from matrix A, form the resulting matrix C whose elements are $c_{ij} = a_{ij} - b_{ij}$ for all i and j.* Thus

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} - \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} (a_{11} - b_{11}) & (a_{12} - b_{12}) \\ (a_{21} - b_{21}) & (a_{22} - b_{22}) \end{pmatrix} \quad (6.1\text{-}6)$$

If two simultaneous equations have a common solution, we could find it algebraically by solving one equation in terms of one unknown and then substituting this value into the other equation to solve for the remaining unknown. Similar manipulations would allow the simultaneous solution of *n*-independent equations in *n* unknowns.

However, the use of matrix notation and its rules allows us to formalize this solution to a simple procedure called *Cramer's rule* [1]. Before describing the rule we must first define the value of the determinant, (pronounced "del").

For a 2×2 matrix, the value of del is obtained by cross multiplying terms beginning with the first product at the upper left and subtracting the second cross product as shown below:

$$\Delta = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21} \quad (6.1\text{-}7)$$

If the simultaneous equations have a common solution,

$$\Delta \neq 0$$

Given the nonzero value for Δ, Cramer's rule can be applied to the solution of

(6.1-1a,b) as

$$x = \frac{\begin{vmatrix} a_{13} & a_{12} \\ a_{23} & a_{22} \end{vmatrix}}{\Delta} \qquad (6.1\text{-}8a)$$

$$y = \frac{\begin{vmatrix} a_{12} & a_{13} \\ a_{21} & a_{23} \end{vmatrix}}{\Delta} \qquad (6.1\text{-}8b)$$

in which the determinants in the numerators of x and y are evaluated in the same manner described for the del in (6.1-6).

For matrices larger than 2×2, we must use a reduction process, successively reducing the size of the starting matrix by evaluating *minors* and *cofactors*. *The minor of an element with subscript ij in a determinant is the determinant that results when the ith row and the jth column are deleted.*

For example, the minor of the a_{11} element of the 3×3 matrix is

$$\text{Minor of } a_{11} \text{ in } \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} \qquad (6.1\text{-}9)$$

The *cofactor* is the minor with the appropriate algebraic sign. *The determinant is equal to the sum of the products of the elements of any row or column by their respective cofactors, where the cofactor of the element in the ith row and the jth column is* $(-1)^{i+j}$ *times its minor.*

Thus, the evaluation of the determinant of a 3×3 matrix first proceeds by reducing the value of del to three terms, each containing a 2×2 matrix, whose numeric value is obtained by forming the diagonal products with appropriate signs, as shown below:

$$\Delta = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

$$= a_{11}(a_{22}a_{33} - a_{32}a_{23}) - a_{12}(a_{21}a_{33} - a_{31}a_{23}) + a_{13}(a_{21}a_{32} - a_{31}a_{22}) \quad (6.1\text{-}10)$$

Notice that the negative sign of the second term results because the sum of the subscripts of a_{12} is odd, hence the negative sign.

With this basis, Cramer's rule can be stated formally as: *Given a nonzero $n \times n$ determinant Δ of the coefficients of a set of n simultaneous linear equations, the common solution for the ith unknown is equal to the fraction of two determinants having a denominator equal to Δ and a numerator determinant obtained by replacing the ith column of Δ with the $a_{i,n+1}$ values of the equation set.*

For example, consider a two-equation set of the form

$$a_{11}x + a_{12}y = a_{13} \tag{6.1-11a}$$

$$a_{21}x + a_{22}y = a_{23} \tag{6.1-11b}$$

with specific coefficients

$$x - y = 0 \tag{6.1-12a}$$

$$2x + y = 6 \tag{6.1-12b}$$

First, note that the determinant for the equation set is not equal to zero:

$$\Delta = 1 - (-2) = 3 \tag{6.1-13}$$

Therefore the equations have a common, nontrivial solution.

Next, using Cramer's rule, the simultaneous solution of the two equations is obtained as

$$x = \frac{\begin{vmatrix} 0 & -1 \\ 6 & 1 \end{vmatrix}}{\Delta} = \frac{1 \cdot 0 - (-1) \cdot 6}{3} = 2 \tag{6.1-14}$$

$$y = \frac{\begin{vmatrix} 1 & 0 \\ 2 & 6 \end{vmatrix}}{\Delta} = \frac{1 \cdot 6 - 2 \cdot 0}{3} = 2 \tag{6.1-15}$$

The reader can verify that the solution $x = 2$, $y = 2$ satisfies both equations, and that graphing the straight lines that they represent reveals that their common solution (intersection) is at $(2, 2)$ in the xy plane.

While the rules given for the evaluation of determinants and the application of Cramer's rule apply for any size determinant, most of our applications can be accommodated with no more than a 3×3 determinant.

6.2 *Z* AND *Y* MATRICES

A two-port network has two ports, each having a pair of terminals. *The behavior of a linear two-port network can be described entirely at one frequency by a pair of equations, any of whose terms may be complex.* One of several such equation descriptions employs the *Z matrix* or *impedance parameters* (Fig. 6.2-1).

The equations that define the circuit's behavior are

$$V_1 = Z_{11}I_1 + Z_{12}I_2 \tag{6.2-1}$$

$$V_2 = Z_{21}I_1 + Z_{22}I_2 \tag{6.2-2}$$
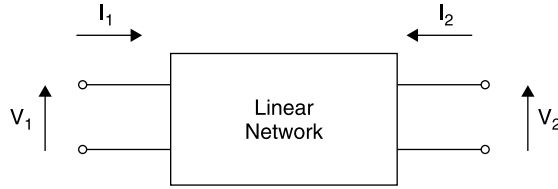
Figure 6.2-1   Voltage and current conventions for Z and Y matrices.

where V and I are phasor quantities. The Z parameter values are found by solving circuit equations for the various ratios of network voltages and currents. Thus,

$$Z_{11} = \left.\frac{V_1}{I_1}\right|_{I_2=0} \qquad Z_{12} = \left.\frac{V_1}{I_2}\right|_{I_1=0} \qquad Z_{21} = \left.\frac{V_2}{I_1}\right|_{I_2=0} \qquad Z_{22} = \left.\frac{V_2}{I_2}\right|_{I_1=0}$$

(6.2-3)

As can be seen from (6.2-3), *to measure the Z parameters requires that open circuits be placed at the input and output terminals* at various times. This may be undesirable to do with microwave networks, but if the Z parameters can be derived analytically, this poses no problem.

Expressed in matrix notation,

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

(6.2-4)

The definition of the Z matrix can be applied to a network having an arbitrary number of ports, the requirement being a separate independent equation relating the voltage at every port to the currents at all of the ports. Thus, an N-port network requires an $N \times N$ matrix representation.

One can also write an admittance matrix based on Kirchhoff's node equations. The nodal format of the Y matrix is the basis of the *nodal analysis* format used extensively in network analysis and simulation software, such as the *Genesys* program employed for the examples in this text.

The *admittance parameters*, or Y *matrix*, use the same voltage and current conventions (Fig. 6.1-1) that are employed to define the Z matrix:

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

(6.2-5)

and the elements are defined in a manner similar to that used to define the Z matrix elements, that is

$$Y_{11} = \left.\frac{I_1}{V_1}\right|_{V_2=0} \qquad Y_{12} = \left.\frac{I_1}{V_2}\right|_{V_1=0} \qquad Y_{21} = \left.\frac{I_2}{V_1}\right|_{V_2=0} \qquad Y_{22} = \left.\frac{I_2}{V_2}\right|_{V_1=0}$$

(6.2-6)
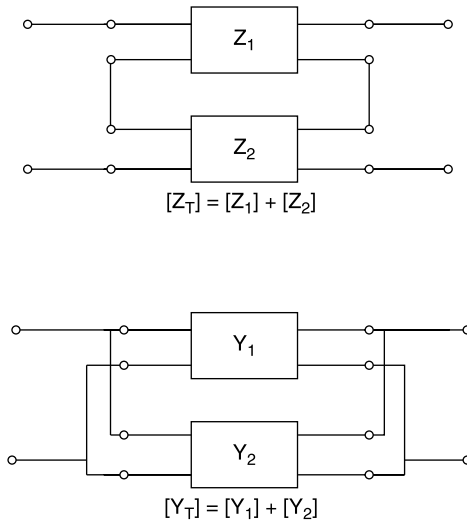
Figure 6.2-2   Series and parallel network combinations represented by Z and Y matrix additions [5, p. 190].

Notice that the Y parameters are the *dual* of the Z parameters in every manner. Impedance is replaced by admittance, currents and voltages are interchanged, and the elements are evaluated using a zero voltage at a terminal (a short circuit) rather than the zero current (open circuit) used to evaluate the Z matrix elements.

The impedance matrix, and its dual, the admittance matrix, have the interesting and very useful properties that the overall matrix of a series or parallel combination of circuits can be obtained by adding together the corresponding elements, respectively, of the Z or Y matrices. This is shown schematically below for pairs of two ports, but the result can be applied to arbitrarily extensive series hookups for Z parameter circuits or parallel hookups for Y parameter circuits.

The relationships in Figure 6.2-2 demonstrate the value of using matrix notation. It facilitates the recognition of identities that might not be so evident were only the underlying simultaneous equations employed.

6.3   RECIPROCITY

A network is reciprocal if a zero impedance source and a zero impedance ammeter can be placed at any locations in a network and their positions interchanged without changing the ammeter reading [3, p. 544]. This is a general relationship attributed to Lorentz and is derived in the general case by applying Maxwell's equations to a general isotropic medium [4, Sec. 1.9]. The conditions for the

evaluation of the Z_{ij} and Z_{ji} matrix elements is equivalent to the Lorentz experiment of interchanging generator and ammeter. All of the circuits in Figure 6.4-3 are reciprocal [3, p. 544].

Accordingly, as a consequence of reciprocity, for the Z matrix description of a reciprocal network

$$Z_{12} = Z_{21} \tag{6.3-1}$$

and for the Y matrix,

$$Y_{12} = Y_{21} \tag{6.3-2}$$

Basically, reciprocity results with network elements that are linear and bilateral, that is, they have the same behavior for currents flowing in either direction. Invoking reciprocity is a powerful tool that provides an additional relationship useful in the analysis of circuits and electromagnetic field cases. It is customarily applied to show that characteristics of transmitting antennas apply equally to receiving antennas and vice versa.

6.4   THE *ABCD* MATRIX

The *ABCD parameters, or ABCD matrix,* are similar to the Z and Y parameters but with three important differences [5, pp. 187–188; 2, Sec. 11.08]. First, *the ABCD matrix is defined only for a two-port network.* Second, the output current direction exits the network. Third, the dependent and independent variables are mixed, each containing both a voltage and a current (Fig. 6.4-1). The result is that the outputs of the network, V_2 and I_2, have the same directions and definitions as the inputs of a following network. For this reason *ABCD matrices can be multiplied together to form the overall matrix of a cascade of circuits.* Because of this property the *ABCD* matrix is also called the *chain matrix.*

The defining equations for the *ABCD* parameters are

$$V_1 = AV_2 + BI_2 \tag{6.4-1}$$
$$I_1 = CV_2 + DI_2 \tag{6.4-2}$$

in which the $A, B, C,$ and D parameters are evaluated using the relationships in
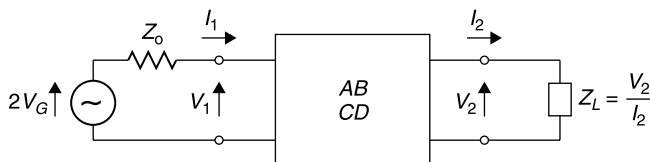
Figure 6.4-1   Defining voltage and current conventions used for the *ABCD* matrix.

(6.4-3) and noting that $I_2 = 0$ corresponds to an open-circuit termination and $V_2 = 0$ corresponds to a short-circuit termination:

$$A = \left.\frac{V_1}{V_2}\right|_{I_2=0} \qquad B = \left.\frac{V_1}{I_2}\right|_{V_2=0} \qquad C = \left.\frac{I_1}{V_2}\right|_{I_2=0} \qquad D = \left.\frac{I_1}{I_2}\right|_{V_2=0} \qquad (6.4\text{-}3)$$

As is true of the elements of the Z and Y matrices, the $A, B, C,$ and D coefficients may be complex quantities. The $ABCD$ matrices for commonly encountered two-port networks are listed in Figure 6.4-2 and can be verified using (6.4-3).
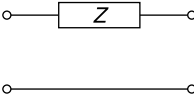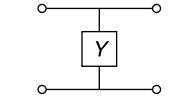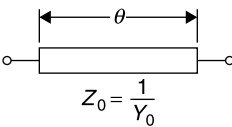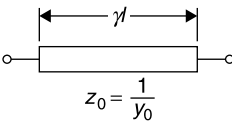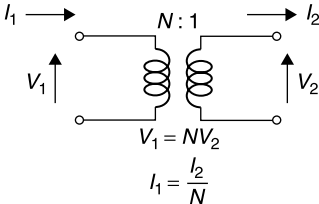
Figure 6.4-2   *ABCD* matrices for common circuit elements.
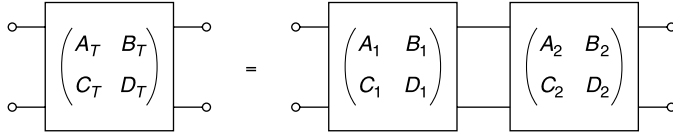
Figure 6.4-3 Cascaded circuits can be represented by a new circuit whose *ABCD* parameters are the result of matrix multiplying the *ABCD*s of the individual circuits.

Remarkably, the modest library of elements and their *ABCD* matrices shown in Figure 6.4-2 suffices to permit the calculation of numerous practical RF and microwave cascaded networks. The principal advantage of the *ABCD* parameters is the fact that the overall matrix representation of a cascade of circuits, $ABCD_T$, can be formed by matrix multiplying their individual *ABCD* matrices. This is shown schematically in Figure 6.4-3.

The required matrix multiplication is

$$\begin{pmatrix} A_T & B_T \\ C_T & D_T \end{pmatrix} = \begin{pmatrix} A_1 & B_1 \\ C_1 & D_1 \end{pmatrix} \cdot \begin{pmatrix} A_2 & B_2 \\ C_2 & D_2 \end{pmatrix}$$

$$= \begin{pmatrix} (A_1 A_2 + B_1 C_2) & (A_1 B_2 + B_1 D_2) \\ (C_1 A_2 + D_1 C_2) & (C_1 B_2 + D_1 D_2) \end{pmatrix} \qquad (6.4\text{-}4)$$

The matrix multiplication defined by (6.4-4) is not *commutative*. That is, $[ABCD]_1 \times [ABCD]_2$ is not necessarily equal to $[ABCD]_2 \times [ABCD]_1$. Therefore, care must be exercised to preserve the order of the matrix multiplications of the cascaded circuits shown in Figure 6.4-3 and defined by (6.4-4). When a cascade of more than two networks is encountered, *a good procedure to find the overall ABCD matrix is to begin by combining the two networks nearest to the load*. Then that result is combined with the next nearest network to the load. The process can be repeated until the *ABCD* matrix for an arbitrarily long cascade is obtained.

By dividing (6.4-1) by (6.4-2) we obtain an expression for the input impedance, Z_{IN}, to the network in terms of the load impedance, Z_L, as follows:

$$Z_{IN} = \frac{AV_2 + BI_2}{CV_2 + DI_2} = \frac{A(V_2/I_2) + B}{C(V_2/I_2) + D}$$

$$= \frac{AZ_L + B}{CZ_L + D} \qquad (6.4\text{-}5)$$

The insertion loss and insertion phase of a two-port network are readily determined from its *ABCD* matrix. Since these values are referenced to the voltage of the generator, V_G, rather than V_1, it is useful to find the total matrix, $ABCD_T$, of the network that comprises both the two port as well as the impedance of the generator, Z_G. This is diagrammed in Figure 6.4-3. This overall
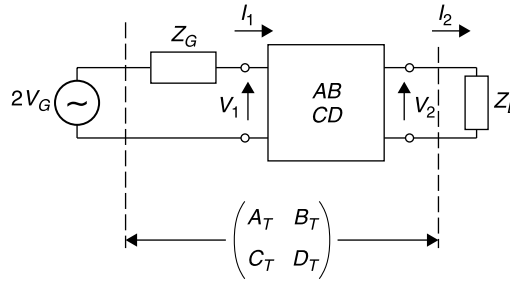
Figure 6.4-4 The *ABCD* matrix of a two-port and the generator impedance is used to compute insertion loss and insertion phase.

matrix is shown in Figure 6.4-4 and is calculated as follows:

$$\begin{pmatrix} A_T & B_T \\ C_T & D_T \end{pmatrix} = \begin{pmatrix} 1 & Z_G \\ 0 & 1 \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} (A + CZ_G) & (B + DZ_G) \\ C & D \end{pmatrix} \quad (6.4\text{-}6)$$

Applying (6.4-1),

$$2V_G = A_T V_2 + B_T I_2 = A_T V_2 + \frac{B_T V_2}{Z_L}$$

$$\frac{2V_G}{V_2} = A_T + \frac{B_T}{Z_L}$$

$$\frac{2V_G}{V_2} = A + CZ_G + \frac{B}{Z_L} + \frac{DZ_G}{Z_L}$$

This expression can be used to find the insertion loss and insertion phase resulting when a network described by the matrix, *ABCD*, is inserted between a generator, Z_G, and load Z_L. When $Z_G = Z_L = Z_0 =$ a real quantity, and since the maximum power deliverable to the load is

$$P_A = V_G^2 / Z_0$$

then the transducer loss and insertion loss (IL) of a two-port network defined by its *ABCD* matrix are the same and equal to [2, p. 187]

$$\text{IL} = \left| \frac{V_G}{V_2} \right|^2 = \frac{1}{4} \left| A + \frac{B}{Z_0} + CZ_0 + D \right|^2 \quad (Z_L = Z_0 = \text{real}) \quad (6.4\text{-}7)$$

Under these conditions the initial phases of V_G and V_L (more precisely the respective arguments of their phasor representations) are the same and can be defined as the reference phase, $0°$. The change in the argument of the voltage at

the load incurred by inserting the *ABCD* network is that of the ratio of V_G/V_2. It is called the *insertion phase*, φ_I, of the *ABCD* network and given by

$$\varphi_I = \tan^{-1}\left\{\frac{\text{Im}(A + B/Z_0 + CZ_0 + D)}{\text{Re}(A + B/Z_0 + CZ_0 + D)}\right\} \qquad (6.4\text{-}8)$$

Note that a positive value of φ_I indicates a phase delay through the two-port network while a negative value indicates a phase advance.

All of the circuits shown in Figure 6.4-2 are reciprocal. If the network is passive and reciprocal [3, p. 534],

$$AD - BC = 1 \qquad (6.4\text{-}9)$$

and if the network is symmetrical with respect to its input and output ports [3, p. 534]

$$A = D \qquad (6.4\text{-}10)$$

As a further example in the use of the *ABCD* matrix, suppose that an integrated circuit requires a quarter wavelength transmission line. If there is insufficient space for the line itself, its behavior can be *mimicked at one frequency* by an equivalent lumped element Pi circuit containing two capacitors and an inductor (Fig. 6.4-5).

The series impedance jX_L of the inductance L is $j\omega L$ and the shunt admittance jB_C of each capacitance C is $j\omega C$. To find the overall matrix for this circuit, $ABCD_{\text{Pi}}$ find the respective *ABCD* matrices for the elements (Fig. 6.4-2) and perform successive multiplication of them *beginning at the load end* [5, p. 189]:

$$\begin{pmatrix} A_{\text{Pi}} & B_{\text{Pi}} \\ C_{\text{Pi}} & D_{\text{Pi}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ jB_C & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & jX_L \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ jB_C & 1 \end{pmatrix}$$

$$\begin{pmatrix} A_{\text{Pi}} & B_{\text{Pi}} \\ C_{\text{Pi}} & D_{\text{Pi}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ jB_C & 1 \end{pmatrix} \cdot \begin{pmatrix} (1 - B_C X_L) & jX_L \\ jB_C & 1 \end{pmatrix} \qquad (6.4\text{-}11)$$

$$\begin{pmatrix} A_{\text{Pi}} & B_{\text{Pi}} \\ C_{\text{Pi}} & D_{\text{Pi}} \end{pmatrix} = \begin{pmatrix} (1 - B_C X_L) & jX_L \\ jB_C(2 - B_C X_L) & (1 - B_C X_L) \end{pmatrix}$$

Comparing $ABCD_{\text{Pi}}$ matrix with that of a lossless transmission line of Z_0 im-
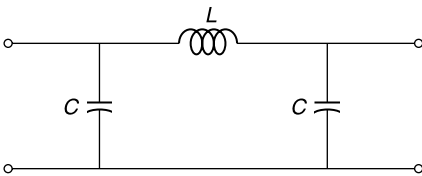
Figure 6.4-5 Pi circuit used to simulate a 90° transmission line.

pedance and θ electrical length (Fig. 6.4-2) equivalence requires that

$$A_{\text{Pi}} = \cos \theta = 1 - B_C X_L \qquad (6.4\text{-}12)$$

However, for a 90° line section, $\cos \theta = 0$, and therefore

$$X_L = \frac{1}{B_C} \qquad (6.4\text{-}13)$$

Next solve for B/C for the lossless line length and equate it to the same quotient for the Pi circuit, yielding

$$\frac{B_{\text{Pi}}}{C_{\text{Pi}}} = \frac{Z_0}{Y_0} = Z_0^2 = \frac{X_L}{2B_C - B_C^2 X_L} \qquad (6.4\text{-}14)$$

and for the case of the 90° line length, for which $X_L = 1/B_C$, (6.4-14) reduces to

$$X_L = \frac{1}{B_C} = Z_0 \qquad (6.4\text{-}15)$$

For example, a 50-Ω 90° long line section can be simulated as a pair of shunt 50-Ω capacitive reactances and a series 50-Ω inductor. The broad utility of this derivation is that it allows the synthesis of any line length and any impedance. The usefulness of the *ABCD* matrix representation can be appreciated by considering the additional computational labor that would have been required to derive this very general result by other means.

6.5 THE SCATTERING MATRIX

The Z, Y, and *ABCD* matrices are useful at low frequencies and for proving general analytical results. However, they can be inconvenient to apply to microwave measurements of devices because they require application of short- and open-circuit terminations for evaluation. In addition, it is not practical to measure discrete voltages and currents at high frequencies. Instead, it is relatively easy to measure traveling waves through the use of directional couplers.

For these reasons, the common practice to evaluate the behavior of a multiport network is to use *incident waves* as excitations at each port and to measure the resulting exiting waves [5, p. 541]. The exiting waves are called *reflected waves*. In actuality, the wave exiting a given port may be the result of an incident wave at another port, and therefore more properly might be called a *transmitted wave*, but this is not the practice. The waves that impinge on the network are called *incident waves* and those that exit are called *reflected waves*, regardless of how they come about.
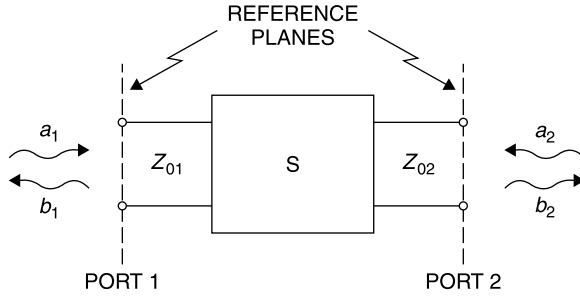
Figure 6.5-1   Defining conventions for the incident, a, and reflected, b, waves.

The incident waves are designated as a_i and the reflected waves are designated as b_i where i is the number of the port. Both the *a and b waves are phasors, having both magnitude and phase at the specified terminals of the network port*. The wave directions and terminal planes are shown schematically in Figure 6.5-1 for a two-port network; however, the convention can be applied to a network with any number of ports, and *the ports need not all be of the same type!* For example, a network might have a coaxial line port, a waveguide port, and a balanced transmission line port.

For each port the incident and reflected waves are defined in terms of the characteristic impedance of the transmission line connected to that port. The wave dimensions are such that when an a or b wave is multiplied by its respective complex conjugate the resulting product has the dimensions of power:

$$a_i = \frac{V_{iI}}{\sqrt{Z_{0i}}} = I_{iI}\sqrt{Z_{0i}} \qquad (6.5\text{-}1)$$

$$b_i = \frac{V_{iR}}{\sqrt{Z_{0i}}} = I_{iR}\sqrt{Z_{0i}} \qquad (6.5\text{-}2)$$

The I subscript stands for incident wave, and the R for reflected wave, and Z_{0i} is the characteristic impedance of the transmission line connected to the ith port of the network. Note that a, b, V, and I are phasors and their phases are specified relative to their respective *network ports*.

Often the a and b waves are loosely called *power waves*, but this practice is misleading and should be avoided because *the power associated with each wave is equal to the magnitude of the square of its amplitude*. It is more precise to recognize that they are traveling *waves having amplitudes and phases* just as do incident voltages and currents.

By using a and b waves, a *linear* network can be characterized by a set of simultaneous equations describing the reflected waves from each port in terms of the incident waves at all of the ports. The constants that characterize the network under these conditions are called *S parameters*.

For example, for a two-port network, the b wave leaving port 1 is the phasor sum of a wave reflected from the input port $(S_{11}a_1)$ plus a wave that passed through the two-port from port 2 $(S_{12}a_2)$:

$$b_1 = S_{11}a_1 + S_{12}a_2 \qquad (6.5\text{-}3)$$

Similarly

$$b_2 = S_{21}a_1 + S_{22}a_2 \qquad (6.5\text{-}4)$$

This descriptive approach can be applied to a network with an arbitrary number of ports. An independent equation is needed for each port. The S parameters can be expressed in matrix format, yielding a square, $n \times n$ matrix, where n is the number of network ports. For the two ports the matrix format is

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \qquad (6.5\text{-}5)$$

Each of the S parameters has its own physical significance. For example, suppose that the S parameters apply to a transistor that is embedded in a matched 50-Ω system. Then S_{11} corresponds to the reflection coefficient at the input port when the output is match terminated $(a_2 = 0)$. Similarly, S_{22} is the reflection coefficient at the output port when the input is match terminated $(a_1 = 0)$. Parameter S_{12} is the feedback to the input from any wave entering port 2, such as a wave reflection from the load; and S_{21} is the amplification (or loss if $|S_{21}| < 1$) of the input wave amplitude as it travels through the transistor to port 2. Accordingly, $|S_{21}|^2 = S_{21}S_{21}^*$ is the power gain of the transistor (or loss if $|S_{21}| < 1$). These interpretations apply not only to this transistor example but to any linear two-port network to which the S parameters are applied. This separation of the functions of the parameters is especially useful. For example, consider the S parameter file for the 2N6679 transistor shown in Table 6.5-1.

Merely by examining this file, most of the device performance characteristics can be read directly. For example, at 100 MHz the voltage gain $|S_{21}|$ is 38.6. At this frequency the power gain with 50-Ω source and load is $38.6^2 = 1490$, or 32 dB. With 50-Ω source and load, the gain of the transistor drops to unity, 0 dB, at 6.5 GHz. However, with proper matching networks the device can provide more than unity gain at 6.5 GHz, since some of the power is being reflected at the input and output ports.

For example, at 100 MHz, the input reflection coefficient magnitude is 0.6. This means that the input return loss is 36%, which corresponds to a mismatch loss of 1.9 dB. Similarly, the output return loss is 69%, which corresponds to a mismatch loss of 5.1 dB. This means that with proper tuning, the transistor could provide $1.9 + 38.6 + 5.1 = 45.6$ dB total gain at 100 MHz. We also note that the feedback term, S_{12}, is low at 100 MHz, which, as we will see later, has encouraging implications for stability at this frequency.

TABLE 6.5-1 *S* **Parameter File for the Motorola 2N667A Bipolar Transistor**[a]

! 2N6679A.S2P
! 2N6679
! VCE = 15 V; IC = 25 mA

GHZ S MA R 50

! S-PARAMETER DATA

0.1	0.60	−76	38.6	141	0.01	55	0.83	−20
0.5	0.67	−158	12.7	95	0.02	40	0.50	−27
1	0.68	−178	6.6	77	0.03	53	0.46	−32
1.5	0.68	170	4.4	64	0.04	54	0.47	−41
2	0.69	162	3.4	54	0.05	54	0.47	−50
2.5	0.69	154	2.7	42	0.06	55	0.49	−59
3	0.69	146	1.3	31	0.07	55	0.53	−70
3.5	0.69	138	1.9	21	0.08	54	0.55	−79
4	0.69	131	1.7	11	0.09	51	0.57	−89
4.5	0.69	123	1.5	1	0.10	49	0.59	−97
5	0.69	114	1.4	−9	0.12	44	0.62	−106
5.5	0.69	106	1.2	−19	0.14	39	0.64	−113
6	0.69	98	1.1	−28	0.15	33	0.68	−122
6.5	0.69	90	1.0	−37	0.17	31	0.69	−130

[a]By convention the columns contain left to right: $f(\text{GHz})$, $|S_{11}|$, $\arg(S_{11})$, $|S_{21}|$, $\arg(S_{21})$, $|S_{12}|$, $\arg(S_{12})$, $|S_{22}|$, $\arg(S_{22})$ [6].

The appeal of *S* parameters is apparent in that all of this information is directly manifest from the *S* parameter data table. For microwave applications, especially transistor characterization, the *S* parameters are particularly desirable because their evaluation *does not require short- or open-circuit terminations*. Instead *matched loads are placed at the input or output* ports and either the transmission or reflection coefficients is measured. The individual *S* parameters are evaluated using these relationships:

$$S_{11} = \frac{b_1}{a_1}\bigg|_{a_2=0} \qquad S_{12} = \frac{b_1}{a_2}\bigg|_{a_1=0} \qquad S_{21} = \frac{b_2}{a_1}\bigg|_{a_2=0} \qquad S_{22} = \frac{b_2}{a_2}\bigg|_{a_1=0} \qquad (6.5\text{-}6)$$

where $a_i = 0$ denotes a matched load at port i.

All excitations, *a waves*, are incident traveling waves, easily measured using directional couplers. Similarly, all exiting waves, *b waves*, are traveling waves in the opposite direction, also readily measured using directional couplers. More about directional couplers and measurements in Chapter 8.

If the network is reciprocal [3, p. 544] (which a transistor is not), then

$$S_{ij} = S_{ji} \tag{6.5-7}$$

The S parameters are complex quantities, having phase angles that correspond to the respective network port reference planes used in their measurement. Yet another convenience of S parameters is that a change of reference planes *farther away from the network by an electrical distance* θ *degrees results in only a negative argument change in the S parameter. For movement closer to the network the argument change is positive.* For diagonal elements, a movement away from the network gives

$$s_{nn}^{\text{new}} = s_{nn} e^{-j2\vartheta_n} \tag{6.5-8}$$

and for off-diagonal elements a movement away from the network gives

$$s_{mn}^{\text{new}} = s_{mn} e^{-j(\theta_m + \theta_n)} \tag{6.5-9}$$

where the m and n subscripts correspond to the respective m and n reference plane movements. This feature of S parameters allows convenient network *deembedding* in which the S parameters of a network can be easily inferred, even when the measurements were made using reference planes that were remote from the actual network ports as, for example, when a device is measured in a test fixture. That is, the *S parameters are most convenient for measurements, since a change of reference planes is easily calculated by adjusting only the S parameter phase angles.* This reference plane shifting is an option built into most network analyzers.

For example, suppose we wish to extend the reference planes an additional 10° from the input port and an extra 15° from the output. The new matrix, in terms of the original S parameters, becomes

$$\begin{pmatrix} s_{11}e^{-j20°} & s_{12}e^{-j25°} \\ s_{21}e^{-j25°} & s_{22}e^{-j30°} \end{pmatrix}_{\text{New}} \tag{6.5-10}$$

Similarly, by reversing the signs in the exponential factors, we can remove line lengths that may have separated an actual device, such as a transistor from the test ports of a test fixture, a process called *deembedding*.

Since the net input power to a lossless network must equal the net output power, it is possible to show that certain relationships hold for the S parameters of a lossless network [3, p. 544]. For example, *for a lossless network, the sum of the squares of the magnitudes of the S parameters of any column is unity:*

$$\sum_{n=1}^{N} |S_{ni}|^2 = 1 \tag{6.5-11}$$

where i is any column from 1 to N for a $N \times N$ matrix.

This is easily proved. Consider, for example, a lossless two-port network that we excite at port 1 with one unit of power and place a matched load at port 2. The total power into the lossless network must equal the total power that emerges, or

$$|a_1^2| = 1 = |b_1^2| + |b_2^2| = |S_{11}^2| + |S_{21}^2| \tag{6.5-12}$$

This argument can be used by successive application of power to any port with matched loads at the other ports. It applies to a lossless network having any number of ports. For a lossless two-port network this means that

$$|S_{11}|^2 + |S_{21}|^2 = 1 \tag{6.5-13}$$

$$|S_{12}|^2 + |S_{22}|^2 = 1 \tag{6.5-14}$$

It is also true that *for a lossless network, the sum of the products of the elements of any column by the complex conjugates of the corresponding elements of another column is zero:*

$$\sum_{n=1}^{N} S_{na} S_{nb}^* = 0 \qquad a \neq b \tag{6.5-15}$$

A matrix whose elements satisfy (6.5-11) and (6.5-15) is called a *unitary matrix* [7, p. 175]. If the network is reciprocal, then $S_{ij} = S_{ji}$ and (6.5-11) and (6.5-15) also apply to rows. Applying this to a lossless two-port network,

$$S_{11} S_{12}^* + S_{21} S_{22}^* = 0 \tag{6.5-16}$$

Because of their numerous advantages, S parameters are standard output format options for nearly all microwave measurements.

However, the advantages of S parameters come at a price. Unlike Z, Y, and *ABCD* parameters, whose values are independent of the measurement system, the evaluation of S parameters is dependent upon the characteristic impedances (port reference impedances) of the transmission lines connected to their ports. *The S parameters are only valid when employed with the reference impedances used for their measurement.* Usually this is 50 Ω. For example, given 50-Ω S parameters for a transistor, they cannot be used directly for an amplifier design intended for 75 Ω characteristic impedance. While a mathematical conversion could be performed, common practice is to remeasure the transistor using a 75-Ω characteristic impedance network analyzer.

6.6   THE TRANSMISSION MATRIX

Since the a waves used with the S parameters are directed inward and the b waves outward, the overall S matrix of a cascade of two ports cannot be

formed by matrix multiplication of the S matrices of the elements of the cascade. However, a and b waves can be rearranged in the defining equations to overcome this limitation. The resultant matrix is known as the *wave transmission matrix*, or *T matrix*. The T matrix uses the same definitions for the a and b waves shown in Figure 6.5-1.

As was true of the *ABCD* matrix, the T matrix achieves its cascading attribute merely by redefining the input and output equation format so that it is compatible with cascade multiplication. One such T matrix definition is

$$\begin{pmatrix} b_1 \\ a_1 \end{pmatrix} = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \begin{pmatrix} a_2 \\ b_2 \end{pmatrix} \tag{6.6-1}$$

in which the a and b waves have the same definitions as previously. The necessary changes of direction of the waves is accommodated in the signs of the respective T parameters. The T matrices can be cascade multiplied in the same manner used for *ABCD* matrices.

Since measurements are commonly made using S parameters only, conversions between T and S matrices are useful. The equivalences, assuming $Z_{01} = Z_{02}$, are [3, p. 548]:

$$\begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} = \begin{pmatrix} \dfrac{s_{12}s_{21} - s_{11}s_{22}}{s_{21}} & \dfrac{s_{11}}{s_{21}} \\ \dfrac{-s_{22}}{s_{21}} & \dfrac{1}{s_{21}} \end{pmatrix} \qquad (Z_{01} = Z_{02}) \tag{6.6-2}$$

and

$$\begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} = \begin{pmatrix} \dfrac{T_{12}}{T_{22}} & \dfrac{T_{11}T_{22} - T_{12}T_{21}}{T_{22}} \\ \dfrac{1}{T_{22}} & \dfrac{-T_{21}}{T_{22}} \end{pmatrix} \qquad (Z_{01} = Z_{02}) \tag{6.6-3}$$

Reciprocity requires that [5]

$$S_{12} = S_{21} \tag{6.6-4}$$

and equivalently,

$$T_{11}T_{22} - T_{12}T_{21} = 1 \tag{6.6-5}$$

Care must be taken in the use of the conversions in (6.6-2) and (6.6-3) since they are based on the defining relations in (6.6-1). There are other defining relationships used for the transmission matrix for which different conversion formulas must be used.

REFERENCES

1. George B. Thomas, Jr., *Calculus and Analytic Geometry*, Addison-Wesley, Cambridge, MA, 1953. *This is a classic calculus textbook.*
2. Simon Ramo and John R. Whinnery, *Fields and Waves in Modern Radio*, Wiley, New York, 1944, 1953 (later revised with a third author, Van Duzer). *This is a classic introductory text describing fields and Maxwell's equations.*
3. Peter A. Rizzi, *Microwave Engineering, Passive Circuits*, Prentice-Hall, Englewood Cliffs, NJ, 1988. *Excellent microwave engineering textbook covering theory and design of transmission lines, couplers, filters, and numerous other passive devices.*
4. Robert E. Collin, *Field Theory of Guided Waves*, McGraw-Hill, New York, 1960.
5. Joseph F. White, *Microwave Semiconductor Engineering*, Noble Publishing, Norcross GA, 1995. *Extensive treatment of PIN diodes and their switching and phase shifting applications. Also includes many fundamental treatments of microwave circuits.*
6. S Parameter Directory, within the *Genesys* network simulator software, Eagleware, 630 Pinnacle Court, Norcross, GA 30071 circa 2000.
7. Robert E. Collin, *Foundations of Microwave Engineering*, McGraw-Hill, New York, 1966. *Collin is a superb theoretician. This textbook gives a fundamental and precise introduction to microwave engineering.*

EXERCISES

E6.2-1 Show that the Z parameters for a two-port network consisting of a length of lossless transmission line of characteristic impedance Z_0 and electrical length θ are

$$\begin{pmatrix} Z_{11} & Z_{12} \\ Z_{211} & Z_{22} \end{pmatrix} = \begin{pmatrix} -jZ_0 \cot \theta & jZ_0 \csc \theta \\ jZ_0 \csc \theta & -jZ_0 \cot \theta \end{pmatrix}$$

E6.2-2 Derive an expression for the input impedance Z_{IN} to a two-port network in terms of its Z parameters. How does this derivation compare in difficulty to that of (6.4-5) for use with the *ABCD* matrix?

E6.2-3 Use the Z parameters you verified in E6.2-1 along with the input impedance formula you derived in E6.2-2 to show that the input impedance Z_{IN} to the network is always Z_0 when the network consists of a Z_0 transmission line of any length, θ, terminated in Z_0.

E6.4-1 **a.** Design a Pi circuit of the form shown in E6.4-2 so that it is equivalent at 1 GHz to a 90° length of 50-Ω transmission line. Then use a network simulator to answer.
   b. What is the bandwidth of a maximum insertion loss of 0.1 dB?
   c. Over what bandwidth is the insertion phase shift, φ, within 10° for the two circuits?

d. The group delay for the line section is 0.25 ns (a quarter period at 1 GHz). Over what bandwidth does the Pi circuit have a group delay $(-\partial\varphi/\partial\omega)$ that is within 10% of this value (within 0.025 ns of 0.25 ns)?

E6.4-2 Design a Pi circuit variable attenuator in the format shown in the figure below.
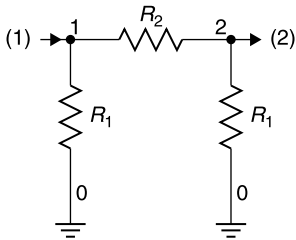
a. Solve for the relationship of R_2 in terms of R_1 such that the attenuator is matched to a Z_0 load.

b. From the results of your analysis, are there any restrictions on the values of R_1 or R_2?

E6.4-3 **a.** Derive an expression for the attenuation (insertion loss) of the Pi attenuator circuit of E6.4-2. *Hint:* Use the formula for insertion loss in terms of the *ABCD* parameters.

b. Find the numeric values for R_1/Z_0 and R_2/Z_0 to obtain 10 dB of insertion loss.

E6.4-4 A lumped element phase shifter consists of switching between the *low-pass* and *high-pass* networks shown in the figure below. Note that the values are normalized to $Z_0 = Y_0 = 1$.

a. Calculate the overall matrix, $ABCD_T$, for the three-element network treating the shunt elements as jb and the series element as jx, where the signs of x and b can be plus or minus (same matrix will apply to both networks).

b. Find the relation between x and b to obtain a match ($Z_{IN} = Z_0 = 1$) in general jb and jx terms applicable to either network.

c. Find the values of x and b for each network that provide a match.

d. Find the values of x and b that give $-90°$ of transmission phase through the low-pass network and $+90°$ through the high-pass network. (Then switching between these two circuits will provide a phase shift of $180°$.)

e. Use a network simulator to determine the bandwidth over which each circuit has an input VSWR < 1.2 (or a return loss < 20 dB).

Lowpass Network   Highpass Network
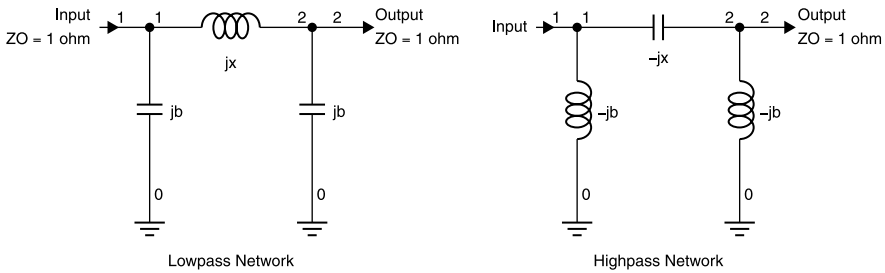
E6.4-5 **a.** Find the overall matrix for the three-element, loaded-line network, $ABCD_T$, shown in the figure below.

   b. Use the result to find an equivalent length of transmission line of Y_E characteristic admittance and θ_E electrical length which represents the three-element network. *Hint:* To do this, compare the $ABCD_T$ with the $ABCD$ matrix of a uniform line of Y_E admittance and θ_E electrical length.
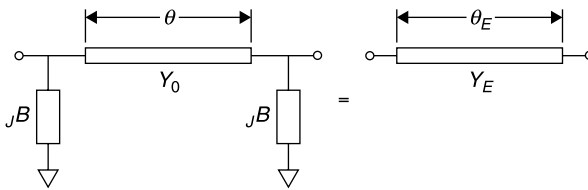
   c. Use the result in part (b) to find the electrical spacing, θ, that matches two similar lossless discontinuities (each represented as a shunt susceptance, $\pm jB$) on a transmission line ($Z_{IN} = Z_0$ when $Z_L = Z_0$). Assume the line, source, and load impedances are Z_0 (characteristic admittance $Y_0 = 1/Z_0$). Express the result in terms of normalized susceptance, $\pm b$ (normalized to Y_0).

E6.4-6 It is proposed to make a phase trimmer for a 50-Ω cable that consists of a pair of capacitive screws spaced a quarter wavelength apart. This is called a *transmission phase shifter*.

   a. Use the equivalent line length formulation of E6.4-5 to determine the maximum amount of capacitance each screw must have to provide an insertion phase tuning range of 0° to 11° at 1 GHz.

   b. Use a network simulator to determine the bandwidth over which the circuit is well matched (VSWR < 1.10) if the design center frequency is 1 GHz and the tuning capacitors are adjusted to provide the maximum "line lengthening" of +11°.

   c. Suppose that a long cascade of these circuits is formed, each loaded-line network beginning at the output of the preceding one.

What would be the maximum VSWR of the cascade? *Hint:* Find the Y_E of the cascade.

E6.5-1 **a.** Determine the S parameters for a lossless 50-Ω transmission line having an electrical length, θ, and located between Z_0 load and Z_0 source.

b. How does this solution compare with that obtained for the Z parameters of a lossless line in E6.2-2? What differences would there be in using these two different network representations for a length of transmission line?

E6.5-2 Find the S matrix for a series inductor of 7.96 nH at 1 GHz and referenced to 50-Ω.

E6.5-3 Verify the S matrix that you obtained in E6.5-2 is unitary, as it should be for a lossless circuit.

E6.5-4 How does the insertion loss of the series inductor network calculated in E6.5-2 using the S matrix compare with that for a series impedance given by (2.10-4)? Assume matched generator and load $(Z_G = Z_L = Z_0)$.

Electromagnetic Fields and Waves

7.1 VECTOR FORCE FIELDS

Phenomena by which noncontact forces can be exerted from a distance are called *vector force fields*. From our earliest experiences, we are familiar with the noncontacting force of gravity. On the Earth's surface it is essentially a constant downward force. More precisely, the gravitational vector force field can be described as *the force of gravity F between any two masses m_1 and m_2 is attractive, directly proportional to their product and inversely proportional to the square of the distance between them and directed along the radius line r connecting their centers of mass:*

$$\vec{\mathbf{F}} = \frac{m_1 m_2 G}{r^2}\vec{r} \tag{7.1-1}$$

where $\vec{r}$ is a unit vector in the *r* direction. For this expression a positive force $\vec{\mathbf{F}}$ is attractive.

Based upon Earth-borne and celestial observations, it is believed that (7.1-1) is applicable to all masses throughout the universe. It can be applied on the moon or any other location. Of course, gravitational forces from every star, planet, moon, asteroid, and every other mass in the universe exert a force component on us; but the Earth exerts so much larger a force that we can neglect the others with only small errors. The gravity field and its mathematical expression in (7.1-1) is an empirically developed *primary concept*. It cannot be derived from more fundamental principles (Fig. 7.1-1).

We cannot see the force field, $\vec{F}$, due to gravity: but since its effect is a daily experience, the gravity field seems intuitively obvious to most individuals. Few persons claim not to understand gravity because it is an invisible force.

In the same manner the presence and behavior of equally invisible *electric and magnetic fields*, or simply *EH fields*, must be intuitively familiar to the insightful RF and microwave engineer. Their presence is experienced in varying
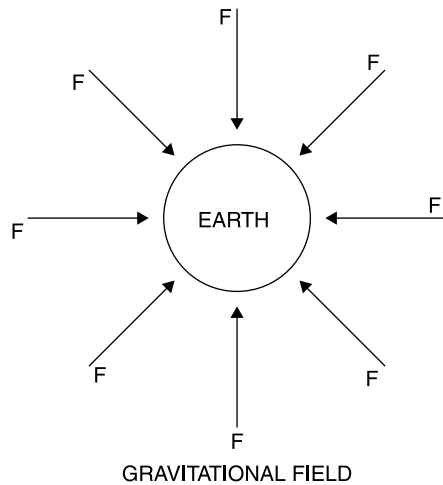
Figure 7.1-1   Earth's noncontacting force field, gravity.

GRAVITATIONAL FIELD

degrees within all circuits and throughout all spaces, even extending throughout the entire known universe. Gravity and electromagnetic waves have in common that they are the only known wave phenomena that can propagate through a vacuum. However, gravity "waves," that is, a time-varying change in the gravitational force have not yet been observed, the difficulty being that a change of mass is required to produce them. Since this requires the destruction or creation of matter, gravity waves have not been and probably never will become practical for routine communicative purposes.

The $\vec{E}$ and $\vec{H}$ fields, however, are eminently practical for communication, since their creation requires merely the separation of electrical charges or the creation of electric currents, both being stimuli that are easy to arrange. Like gravity, behaviors of $\vec{E}$ and $\vec{H}$ fields conform to empirically developed laws.

These are summarized by *Maxwell's equations, which* he formulated from various empirical observations made by others [1, Chapters 1 and 2]. These equations were integrated into a unified theory of electromagnetism to give precise mathematical expressions to Gauss's laws for electric and magnetic fields, Faraday's law (with Lenz's law) of induction, and Ampere's law, with displacement current added by Maxwell.

Using his equations James Clerk Maxwell proved mathematically that electromagnetic waves could propagate through a nonconducting medium at the speed of light. He formulated his unified theory in about 1862 [2, p. 73], but his equations were more numerous at that time. It remained for Oliver Heaviside [3] to reduce them to the four equations, as we now know them. Maxwell had stated that *"we have strong reason to conclude that light itself is an electromagnetic disturbance in the form of waves."* Later, Heinrich Hertz verified Maxwell's theory of electromagnetic waves, but the ideas were not widely accepted.

It took time to displace the erroneous but commonly held view that the

transmission of visible light required a "luminiferous aether" that pervaded all space and material. It had no weight but elastic properties to permit it to propagate the disturbance that light waves represented (as the elasticity of a drum's head allows the disturbance caused by striking it with a drumstick to propagate across the membrane).

> Faraday had speculated upon this possibility [that light was a form of electric wave] several years earlier [than Maxwell's similar conclusion], and had pointed out that whereas one infinite and all-pervasive imponderable aether is a severe strain upon one's imagination, belief in two co-existent, infinite, all-pervasive, and imponderable aethers, one for light and one for electricity, is simply beyond the limits of credulity. Therefore he suspected, on this basis alone, that light is an electric phenomenon [4, p. 112].

Usually Maxwell's equations are introduced on a mathematical basis. Since they describe three-dimensional field quantities, this requires a knowledge of vector mathematics which, when studied separately, can be complex and may seem to be arbitrary in definition. In our treatment EH fields will be introduced through the definitions, applications, and experiments by which they were first discovered. Within this context we shall introduce, as needed, the vector mathematics required to describe them. The author believes that this approach provides a more intuitive understanding, not only of the behavior of EH fields and waves but also of the fundamentals of vector mathematics needed to describe and analyze them.

7.2   E AND H FIELDS

Wireless transmission occurs by means of propagating $\vec{E}$ and $\vec{H}$ fields, or electromagnetic waves. *Electromagnetic waves—comprising radio, microwave, millimeter waves, infrared, visible light, ultraviolet and X rays—are the only practical means of propagating energy and thereby communicating through a vacuum.*
Electromagnetic waves are as fundamental to radio as the Big Bang is to the origin of the known universe. *The $\vec{E}$ (electric) and $\vec{H}$ (magnetic) fields are vector force fields, having both magnitude and direction at every point. They can be static (non-time-varying), but for propagating applications they must vary in time, and when employed for communication this variation usually is sinusoidal or can be approximated by a sum of sinusoidal components.*

7.3   ELECTRIC FIELD E

The electric field $\vec{E}$ at a given position is the magnitude and direction of the force $\vec{F}$ which is experienced by a unit positive charge q when it is exposed to the $\vec{E}$ field at that position:

$$\vec{E} = \frac{\vec{F}}{q} \tag{7.3-1}$$

In the MKS system, $\vec{F}$ is in newtons, q is in coulombs, and $\vec{E}$ is in newtons/coulomb (or equivalently, volts/meter).

Electrostatically, the force between two charges is expressed by *Coulomb's law*. It is analogous to the gravitational force between two masses. Coulomb's law states that the force exerted on a point charge q due to another point charge q_1 is directly proportional to the product qq_1 and inversely proportional to the square of the distance r between them. That is,

$$\vec{F} = \frac{qq_1}{4\pi\varepsilon r^2}\vec{r} \tag{7.3-2}$$

where $\varepsilon = \varepsilon_0\varepsilon_R$, ε_0 is a constant (see inside cover for value), ε_R is a function of the dielectric material, and $\vec{r}$ is a unit vector along the line between the two charges and directed away from q. Thus, a positive value of $\vec{F}$ is a force of repulsion while a negative value indicates an attractive force. It follows from (7.3-1) that the electric field due to a point charge q_1 is given by

$$\vec{E} = \frac{q_1}{4\pi\varepsilon r^2}\vec{r} \tag{7.3-3}$$

For example, suppose a 9-V battery is connected to a vertically oriented capacitor with parallel plates 1 cm apart in a vacuum. Since the $\vec{E}$ field can be presumed uniform between the plates, its integral with respect to distance, $E \times d = V$, is potential energy or voltage. The $\vec{E}$ field will be 9 V/cm or 900 V/m. Then suppose that a single electron is placed on the lower plate and released [5, Vol. 2, pp. 420–421].

What is the ratio of the gravitational force on the electron compared to the force of the electric field?

$$\text{Electronic charge} = e = 1.60 \times 10^{-19} \text{ C}$$

$$\text{Electronic mass} = 9.1 \times 10^{-31} \text{ kg}$$

$$F_{\text{electric}} = eE$$

$$= (1.6 \times 10^{-19} \text{ C}) \times (0.9 \times 10^3 \text{ N/C}) = 1.4 \times 10^{-16} \text{ N} \tag{7.3-4}$$

$$F_{\text{gravity}} = mg$$

$$= (9.1 \times 10^{-31} \text{ kg}) \times (9.8 \text{ N/kg}) = 8.9 \times 10^{-30} \text{ N} \tag{7.3-5}$$

The ratio of the electric to gravitational forces is

$$\frac{F_e}{F_g} = \frac{1.4 \times 10^{-16}}{8.9 \times 10^{-30}} = 1.6 \times 10^{13} \tag{7.3-6}$$

Thus, the gravitational force, even in this fairly mild electric field, is negligible compared to the electric force.

Starting from rest, to what vertical speed will the electron accelerate before it strikes the opposite plate?

Acceleration $= a = \dfrac{F}{m} = \dfrac{eE}{m}$

$$= \dfrac{(1.60 \times 10^{-19} \text{ C}) \times (0.9 \times 10^3 \text{ N/C})}{9.1 \times 10^{-31} \text{ kg}} = 1.6 \times 10^{14} \text{ m/s}^2 \quad (7.3\text{-}7)$$

The velocity after being accelerated through a distance x is

Velocity $= \sqrt{2ax}$

$$= \sqrt{(2 \times 1.6 \times 10^{14} \text{ m/s}^2) \times 10^{-2} \text{ m}} = 1.8 \times 10^6 \text{ m/s}$$

$$= 6 \times 10^6 \text{ ft/s} = 8{,}530{,}000 \text{ mph!} \quad (7.3\text{-}8)$$

How long will it take for the electron to traverse the 1 cm distance?

$$t = \dfrac{v}{a} = \dfrac{1.8 \times 10^6 \text{ m/s}}{1.6 \times 10^{14} \text{ m/s}^2} = 1.1 \times 10^{-8} \text{ s} = 11 \text{ ns} \quad (7.3\text{-}9)$$

7.4 MAGNETIC FLUX DENSITY

For magnetic force problems we use the flux density vector $\vec{\mathbf{B}}$ to calculate the magnitude and direction of the force. For homogenous media $\vec{\mathbf{B}}$ and $\vec{\mathbf{H}}$ are related by a constant. In the case of magnetic forces acting upon a charge, the situation is more complex than that of electric field forces. This is because *the magnetic force on a charge only acts when the charge is moving relative to the field and then the direction of the force is at right angles to the plane that contains the field and the charge's velocity vectors* (Fig. 7.4-1).

The magnetic induction ($\vec{\mathbf{B}}$ field) is one weber/m^2 (one Tesla) when one coulomb of charge, moving with a perpendicular component of velocity of one m/s ($\vec{v} \sin \varphi$) experiences a force of one newton.

The condition of maximum force on the moving charge occurs when the charge's velocity $\vec{v}$ is oriented exactly 90° relative to the $\vec{\mathbf{B}}$ field. However, in general, the charge's velocity is at an angle φ relative to the $\vec{\mathbf{B}}$ field and the magnitude of the force $\vec{\mathbf{F}}$ is given by

$$F = qvB \sin \varphi \quad (7.4\text{-}1)$$

where $\vec{\mathbf{F}}$ is a force in the plane orthogonal to the plane of $\vec{v}$ and $\vec{\mathbf{B}}$ and in the
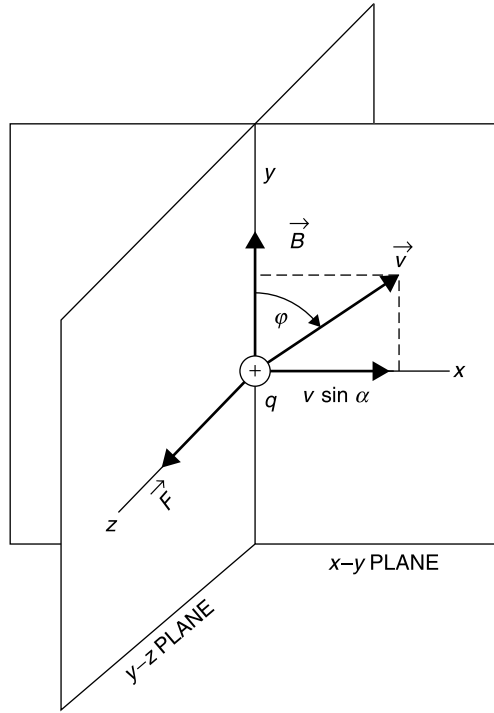
Figure 7.4-1 Vectors $\vec{B}$ and $\vec{v}$ lie in a plane that is orthogonal to the resulting force $\vec{F}$ on the moving charge q.

direction of a right-hand screw's advance when $\vec{v}$ is rotated into $\vec{B}$ through the smaller angle φ between $\vec{v}$ and $\vec{B}$.

The expression in (7.4-1) relates the magnitudes of the vectors $\vec{F}, \vec{v}$, and $\vec{B}$ but further explanation is necessary to convey their relative directions. For this situation the definition of the *vector cross product* proves useful.

7.5 VECTOR CROSS PRODUCT

The vector cross product, (pronounced "$\vec{A}$ cross $\vec{B}$") of two vectors $\vec{A}$ and $\vec{B}$ is equal to a third vector $\vec{C}$ whose magnitude is equal to the product of the magnitudes of $\vec{A}, \vec{B}$ and the sine of the angle φ between them and whose direction is perpendicular to the plane of $\vec{A}$ and $\vec{B}$ with a positive sense defined as that of the advance of a right-hand screw were $\vec{A}$ to be rotated into $\vec{B}$ through the smaller angle between them:

$$\vec{A} \times \vec{B} \equiv \vec{C} \tag{7.5-1}$$
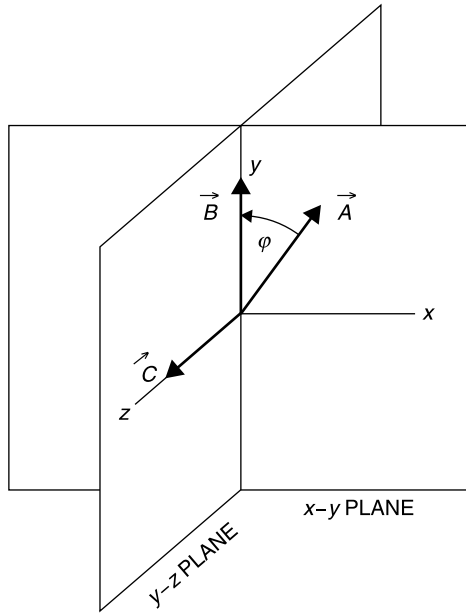
Figure 7.5-1   Vector cross product.

This definition, although lengthy in words, completely describes the three-dimensional geometry of Figure 7.5-1. Once learned, only two rules are required to remember and apply the vector cross product. *First*, $\vec{A}$ is rotated into $\vec{B}$ to produce a vector $\vec{C}$ that advances in the direction of a right-hand screw. *Second*, the magnitude of $\vec{C}$ is the product of the magnitudes of $\vec{A}$, $\vec{B}$, and $\sin \varphi$, where φ is the angle (less than 90°) between them.

Keep in mind that a positive or negative sign applies to the curl according to the sense of rotating $\vec{A}$ into $\vec{B}$. For example, in Figure 7.5-2, the rotation of $\vec{A}_1$ into $\vec{B}$ through φ_1 produces a resultant, $\vec{C}_1$ in the $+\vec{z}$ direction. However, the rotation of $\vec{A}_2$ through φ_2 produces the opposite rotational sense and, accordingly, a $\vec{C}_2$ in the $-\vec{z}$ direction. In either case it is necessary to *use the smaller angle, φ, between $\vec{A}$ and $\vec{B}$* to determine the rotational sense of $\vec{A}$ into $\vec{B}$ and from it the corresponding sign of $\vec{C}$.

Because the sense of the advance of a right-hand screw changes when the order of the vectors of the cross product is reversed, it follows that

$$\vec{A} \times \vec{B} = -\vec{B} \times \vec{A} \tag{7.5-2}$$

Notice that Figure 7.4-1 displays a *right-handed coordinate system* wherein crossing $\vec{x}$ into $\vec{y}$ yields a vector in the $+\vec{z}$ direction, $\vec{y}$ into $\vec{z}$ yields $+\vec{x}$, and $\vec{z}$ into $\vec{x}$ yields $\vec{y}$. In other words, crossing any two successive positive directional vectors in the xyz coordinate order of a right-hand coordinate system yields a

x–y PLANE
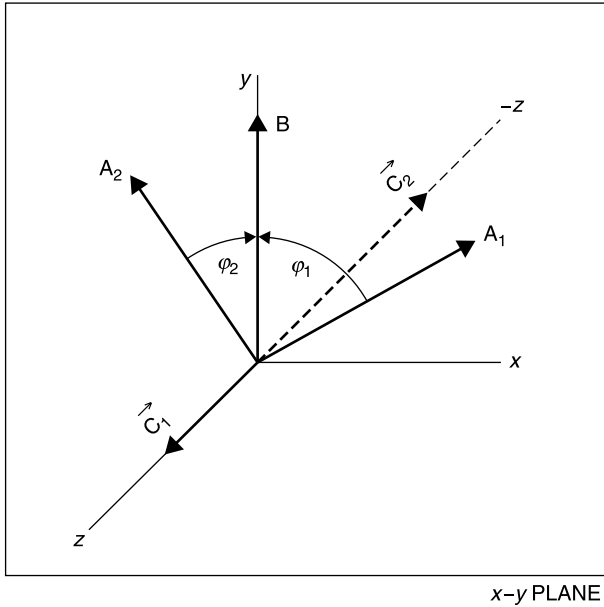
Figure 7.5-2   Direction (sign) of $\vec{C}$ changes with the sense of rotation of $\vec{A}$ into $\vec{B}$.

cross product in the positive direction of the third. It is best to use a right-hand coordinate system to avoid errors in calculating the cross product. Stated in vector notation [6, pp. 88–89],

$$\vec{x} \times \vec{y} = \vec{z} = -\vec{y} \times \vec{x} \qquad (7.5\text{-}3a)$$

$$\vec{y} \times \vec{z} = \vec{x} = -\vec{z} \times \vec{y} \qquad (7.5\text{-}3b)$$

$$\vec{z} \times \vec{x} = \vec{y} = -\vec{x} \times \vec{z} \qquad (7.5\text{-}3c)$$

$$\vec{x} \times \vec{x} = \vec{y} \times \vec{y} = \vec{z} \times \vec{z} = 0 \qquad (7.5\text{-}3d)$$

where $\vec{x}$, $\vec{y}$, and $\vec{z}$ are unit vectors in the respective positive directions of the x, y, and z coordinate directions. If the vectors $\vec{A}$ and $\vec{B}$ are expressed in terms of their rectangular coordinates in this right-hand system,

$$\vec{A} \times \vec{B} = (A_x\vec{x} + A_y\vec{y} + A_z\vec{z}) \times (B_x\vec{x} + B_y\vec{y} + B_z\vec{z}) \qquad (7.5\text{-}4)$$

where A_x, A_y, and A_z are the components of $\vec{A}$ in the $\vec{x}$, $\vec{y}$, and $\vec{z}$ directions, and similarly with B_x, B_y, and B_z. The cross product of (7.5-4) would be expected to produce nine terms, since each component of $\vec{A}$ is crossed separately into each component of $\vec{B}$. However, vector components of $\vec{A}$ which have the same corresponding directions as components of $\vec{B}$ produce a zero cross product according to (7.5-3d), so only six terms are nonzero. By observing the cross products

in (7.5-3a,b, and c), (7.5-4) can be expanded explicitly as

$$\vec{A} \times \vec{B} = (A_y B_z - A_z B_y)\vec{x} + (A_z B_x - A_x B_z)\vec{y} + (A_x B_y - A_y B_x)\vec{z} \quad (7.5\text{-}5)$$

A short-hand method for writing and remembering (7.5-5) is the solution to the determinant

$$\vec{A} \times \vec{B} = \begin{vmatrix} \vec{x} & \vec{y} & \vec{z} \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix} \quad (7.5\text{-}6)$$

Expressions (7.5-5) and (7.5-6) produce the same result. Given the definition of the vector cross product, the force on a moving charge can be written very tersely yet in a manner than conveys all of the relationships between the magnitudes and directions of $\vec{F}, \vec{v}$, and $\vec{B}$.

If a charge q has velocity v *that is at right angles to the magnetic field* $\vec{B}$*, the force on the moving charge q is given as*

$$\vec{F}_{\text{magnetic}} = q\vec{v} \times \vec{B} \quad (7.5\text{-}7)$$

Some useful conversions for the units of magnetic flux, $\vec{B}$, are

$$1 \text{ T (Tesla)} = 10{,}000 \text{ G (Gauss)} = 1\frac{\text{Wb (Weber)}}{\text{m}^2}$$

$$= 1\frac{\text{N (Newton)}}{\text{C (Coulomb)} \times \text{m/s}} = 1\frac{\text{N (Newton)}}{\text{Am (Ampere)}} \quad (7.5\text{-}8)$$

For example, an electron moving at 1×10^6 ms/s at right angles to a 1-T magnetic field would experience a force given by

$$F_{\text{magnetic}} = 1.6 \times 10^{-19} \text{ C} \times 1 \times 10^6 \text{ m/s} \times 1\frac{\text{N}}{\text{C} \times \text{m/s}} = 1.6 \times 10^{-13} \text{ N} \quad (7.5\text{-}9)$$

The force of gravity on the same electron is

$$F_{\text{gravity}} = mg = 9.1 \times 10^{-31} \text{ kg} \times 9.8 \text{ N/kg} = 8.9 \times 10^{-30} \text{ N} \quad (7.5\text{-}10)$$

Again, the gravitational force is negligible. Notice that the magnetic force is 1000 times as great as the electric force calculated for the electric field example, given this 1-T field strength. One tesla is a strong magnetic field.

When a charged particle enters at right angles to a magnetic field at a constant velocity, its orbit is a circle [5, Vol. 2, p. 561]. The circular electron orbit is the basis of the magnetron microwave tube [7, p. 41; 8, p. 321]. Even in the 1940s magnetrons were capable of 100 kW of peak power for short pulses of a microsecond or less. These high pulsed power levels were orders of magnitude greater than those achievable with prior tube types and made practical the development of tactical radars during the latter part of World War II.
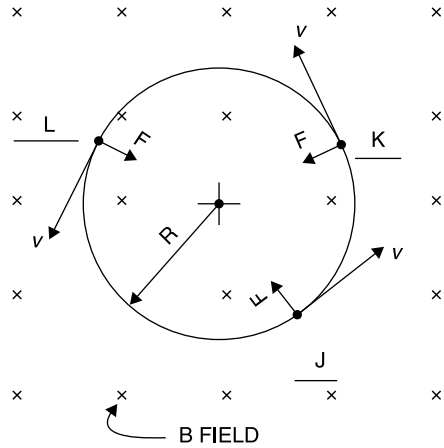
Figure 7.5-3 Positive charge with v speed enters a $\vec{B}$ field (crosses indicate $\vec{B}$ field direction is "into the paper") at point J and has its path turned into a circular orbit (curving to points K and L) with a radius R. The magnetic force $\vec{F}$ is always directed toward the circular orbit's center and is equal and opposite to the particle's centrifugal force.

The centrifugal force of a particle traveling in a circular orbit of radius R (Fig. 7.5-3) is

$$F_{\text{centrifugal}} = \frac{mv^2}{R} \tag{7.5-11}$$

Equating this to the magnetic force gives [5, Vol. 2, p. 561]

$$q v B = m \frac{v^2}{R} \tag{7.5-12}$$

and thus,

$$R = \frac{mv}{Bq} \tag{7.5-13}$$

In addition to the magnetron tube, this method of orbiting a charged particle also is employed in the cyclotron particle accelerator and in mass spectrographs to analyze ionized samples—the radii of whose orbits are directly proportional to the particle's mass and hence its atomic element number.

As an example, suppose that an electron enters a magnetic field of flux density $\vec{B} = 2 \times 10^{-4}$ T (2 G) traveling in a direction that is at right angles to $\vec{B}$ with a velocity of 1×10^6 m/s:

$$R = \frac{mv}{Bq} = \frac{(9.1 \times 10^{-31} \text{ kg}) \times (1 \times 10^6 \text{ m/s})}{(2 \times 10^{-4} \text{ T}) \times (1.6 \times 10^{-19} \text{ C})} = 2.85 \text{ cm} = 1.12 \text{ in.} \tag{7.5-14}$$
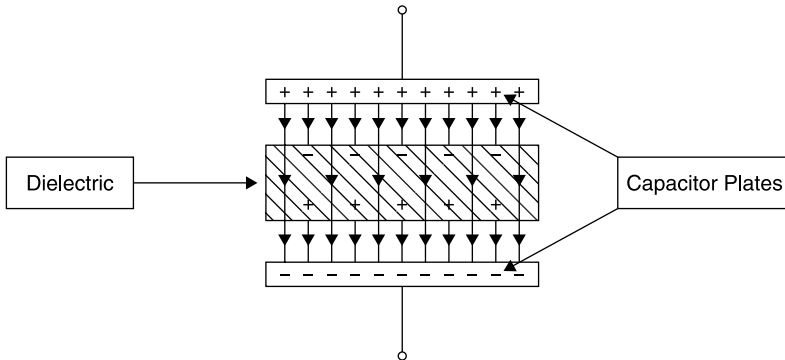
Figure 7.6-1   Induced charges on a dielectric inserted between the plates of a capacitor reduce the electric field within the dielectric.

7.6   ELECTROSTATICS AND GAUSS'S LAW

The magnitude of the electric field $\vec{E}$ depends upon the medium in which it is immersed [5, Vol. 2, p. 478]. For example, when a dielectric material is placed between the charged plates of a capacitor, induced charges in the dielectric cause the $\vec{E}$ field lines to terminate on them, reducing the $\vec{E}$ field within the dielectric if the capacitor is open circuited (Fig. 7.6-1). If a constant voltage is applied the voltage source will supply additional charge to the capacitor to build up to the original $\vec{E}$ field and voltage between the plates.

The electric field at a distance r from a fixed point charge q is given by (7.3-3) and repeated here (Fig. 7.6-2):

$$\vec{E} = \frac{q}{4\pi\varepsilon_0\varepsilon_R r^2}\vec{r} \tag{7.6-1}$$
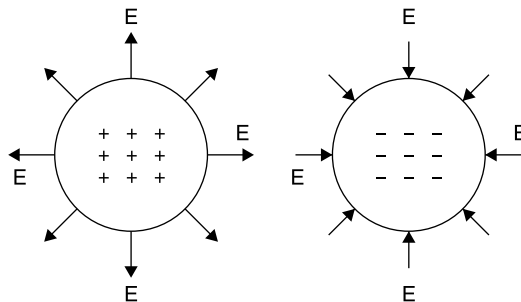
Figure 7.6-2   Coulomb observed that the electric field at a radial distance r from a point charge q is proportional to q and inversely proportional to the square of r. The field direction points away from a positive charge (to repel a like positive charge) and inward toward a negative charge (to attract an opposite, positive charge).
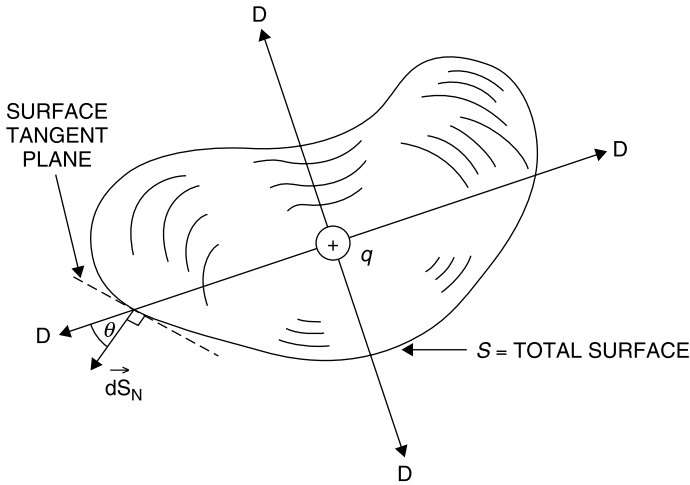
Figure 7.6-3   Arbitrary volume encloses a charge, q. The resulting $\vec{D}$ vector field exits the volume radially over the closed surface S. Incrementally, the surface is divided into dS areas having a surface normal direction $\vec{dS}$ at angle θ to the vector field $\vec{D}$.

This observation can be written without specifying the dielectric environment in which the charge is found by defining a companion vector, *electric flux density* $\vec{D}$ as

$$\vec{D} = \varepsilon\vec{E} = \varepsilon_0\varepsilon_R\vec{E} = \frac{q}{4\pi r^2}\vec{r} \tag{7.6-2}$$

The reason for defining $\vec{D}$ is that the effect of induced charge in the dielectric is absorbed in ε_R and therefore the right-hand side (7.6-2) is only required to account for free charge. A more general observation that is suggested by the spherical geometry of Figure 7.6-2 is known as *Gauss's law*.

Gauss's law states that for any closed surface containing a total fixed charge q, the product of the emerging $\vec{D}$ field magnitude normal to each infinitesimally small area $\vec{dS}_N$, summed over the entire containing surface, S, is equal to the total charge q contained within the surface (Fig. 7.6-3). A mathematical expression of this statement is provided in the next section.

7.7   VECTOR DOT PRODUCT AND DIVERGENCE

To describe the mathematical process that Gauss's law requires, we assign a vector, $\vec{dS}$, to the surface that is always normal to and has the magnitude dS of a small area of the surface at each point on the surface. It has a direction defined as positive when directed toward the outside of the volume contained

by the surface. The vector field $\vec{\mathbf{D}}$ (Fig. 7.6-3) resulting from charge q is a function of the x, y, z position on the surface S. Then, everywhere on the surface S we form the product

$$\text{Vector product} = |\vec{\mathbf{D}}|\,|\vec{\mathrm{dS}}|\cos\theta \qquad (7.7\text{-}1)$$

where θ is the angle between $\vec{\mathbf{D}}$ and $\vec{\mathrm{dS}}$. Note that when $\vec{\mathbf{D}}$ and $\vec{\mathrm{dS}}$ point in opposite directions the $\cos\theta$ is negative. In this way, the vector product in (7.7-1) produces both magnitude and sign. The product described by (7.7-1) occurs regularly in physical situations and has a formal vector mathematical definition called the *dot product*.

   The vector dot product, $\vec{\mathbf{A}}\cdot\vec{\mathbf{B}}$ *(read as "A dot B") of two vectors* $\vec{\mathbf{A}}$ *and* $\vec{\mathbf{B}}$ *separated by an angle* θ *is formed by multiplying the magnitude of* $\vec{\mathbf{A}}$ *times the magnitude of* $\vec{\mathbf{B}}$ *times the cosine of the angle between them to produce a scalar quantity having the sign of* $\cos\theta$:

$$\vec{\mathbf{A}}\cdot\vec{\mathbf{B}} = (\mathrm{A}_x\vec{x} + \mathrm{A}_y\vec{y} + \mathrm{A}_z\vec{z})\cdot(\mathrm{B}_x\vec{x} + \mathrm{B}_y\vec{y} + \mathrm{B}_z\vec{z}) \qquad (7.7\text{-}2)$$

in which, as previously, $\vec{x}, \vec{y},$ and $\vec{z}$ are unit vectors in the positive directions of the $x, y,$ and z axes, respectively, of a right-hand coordinate system. As was true of the cross product, some of the multiplication terms described by (7.7-2) are zero. In particular, when each component of $\vec{\mathbf{A}}$ is dot multiplied by the three components of $\vec{\mathbf{B}}$, only like-directed components produce a nonzero product. Furthermore, since the A_i and B_i components (for $i = x, y, z$) are in exactly the same direction, the cosine of the angle between them is unity. Accordingly, (7.7-2) reduces to

$$\vec{\mathbf{A}}\cdot\vec{\mathbf{B}} = \mathrm{A}_x\mathrm{B}_x + \mathrm{A}_y\mathrm{B}_y + \mathrm{A}_z\mathrm{B}_z \qquad (7.7\text{-}3)$$

Note that, unlike the vector cross product which produces a new vector, *the dot product produces a scalar quantity*.

   In terms of the dot product, Gauss's law can be stated in integral form as

$$\oint_S \vec{\mathbf{D}}\cdot\vec{\mathrm{dS}} = q \qquad (7.7\text{-}4)$$

Taking the limit as the volume enclosed by S tends to zero, this equation can be used to define the *divergence of the vector* $\vec{\mathbf{D}}$.

   The divergence $\nabla\cdot\vec{\mathbf{A}}$ *(pronounced "del dot* $\vec{\mathbf{A}}$*") of a vector field* $\vec{\mathbf{A}}$ *from a closed surface S is equal to the dot product of* $\vec{\mathbf{A}}$ *with a vector* $\vec{\mathrm{dS}}$ *at the surface having a magnitude equal to the surface area at the location of evaluation, a direction that is orthogonal to the surface, and a sign that is positive when the surface vector points outside the volume enclosed by the surface.*

Written in differential form this is

$$\nabla \cdot \vec{A} \equiv \lim_{\Delta V \to 0} \frac{\oint_S \vec{A} \cdot \overrightarrow{dS}}{\Delta V} \tag{7.7-5}$$

The divergence can be calculated for any vector field. Using the divergence, Gauss's law can be written in differential form as

$$\nabla \cdot \vec{D} = \rho \tag{7.7-6}$$

where ρ is the volume *charge density* at the location at which $\vec{D}$ is evaluated. If (7.7-6) is integrated over any given volume [6, p. 71]

$$\int_V \nabla \cdot \vec{D} \, dV = \int_V \rho \, dV \tag{7.7-7}$$

and then, replacing the right side by its equivalent from Gauss's law

$$\int_V \nabla \cdot \vec{D} \, dV = \int_S \vec{D} \cdot \overrightarrow{dS} \tag{7.7-8}$$

This result has been obtained from a consideration of the $\vec{D}$ field; however, it is a general result that applies to any vector $\vec{A}$, and is called the *divergence theorem* (and sometimes *Gauss's theorem*).

   The divergence theorem: For any continuous vector function of space, $\vec{A}$, the integral of the divergence of $\vec{A}$ over any volume completely enclosed by surface S *is equal to the integral of $\vec{A} \cdot \overrightarrow{dS}$ over the whole surface* S.

$$\int_V \nabla \cdot \vec{A} \, dV = \int_S \vec{A} \cdot \overrightarrow{dS} \tag{7.7-9}$$

Like the dot product the *divergence produces a scalar result*. The divergence of a vector $\vec{A} = A_x\vec{x} + A_y\vec{y} + A_z\vec{z}$ can be evaluated using

$$\nabla \cdot \vec{A} = \frac{\partial}{\partial x}A_x + \frac{\partial}{\partial y}A_y + \frac{\partial}{\partial z}A_z \tag{7.7-10}$$

7.8   STATIC POTENTIAL FUNCTION AND THE GRADIENT

In an electrostatic region, in which charges are fixed, each coordinate point (x, y, z) has an associated *static potential energy*, $\Phi(x, y, z)$ for a point charge relative to some reference. The potential energy of a test charge in a region of electrostatic potential energy is analogous to a skier who has potential energy

for each elevated location on a mountain relative to the base of the mountain or to some other location defined as a reference at which his potential energy is defined to be zero.

The energy which must be expended to move a unit positive charge from location 1 to some different location 2 is given by [6, p. 73]

$$\Phi_2 - \Phi_1 = -\int_1^2 \vec{E} \cdot \overrightarrow{dl} \tag{7.8-1}$$

The negative sign occurs because a positive $\vec{E}$ field produces a force that tends to move a positive charge from the higher potential to the lower one. To move in the direction opposite to the force field ($\vec{E}$ field), we must do work (provide energy) rather than recover it. *Only a difference in potential is important*; the zero potential reference can be selected arbitrarily.

In an electrostatic system, having no moving charges (currents) or changing magnetic fields, the potential is the same as voltage. However, we reserve the term *voltage* for later to describe the combined effects of electrostatic charge distributions as well as changing magnetic fields.

The potential energy in a region due to the presence of electrostatic (fixed) charges and their resulting $\vec{E}$ fields has a single value at every location independent of how that location is approached. Thus, it follows that the integral of $\vec{E}$ about any closed path in the region must be zero. The same is true of a gravitational field around which any closed path (in which you return to the starting point) produces no change in potential energy. Similarly, there is no net energy change in moving a given charge about a closed path in an electrostatic $\vec{E}$ field. This means that *the $\vec{E}$ field is conservative*. Practically, there may be some energy lost to friction (resistance) in moving charge about a system, but *no difference in potential energy results in completing a closed path within a conservative field.*

Mathematically this is stated

$$\oint \vec{E} \cdot \overrightarrow{dl} = 0 \tag{7.8-2}$$

Suppose that we wish to determine the potential at point P due to a charge q_1 located at a point Q_1 (Fig. 7.8-1).

Setting the zero potential at infinity, the work done to move a unit test charge from infinity to the point P is equal to the work integral

$$\Phi_1 = -\int_\infty^P \vec{E} \cos \theta \, dl = -\int_\infty^P \frac{q_1 \, dr}{4\pi\varepsilon r^2} = \frac{q_1}{4\pi\varepsilon r_1} \tag{7.8-3}$$

in which the equivalence $\cos \theta \, dl = dr$ has been recognized, the value for $\vec{E}$ has been substituted from (7.8-1), and r_1 is the distance from point Q_1 to P. The
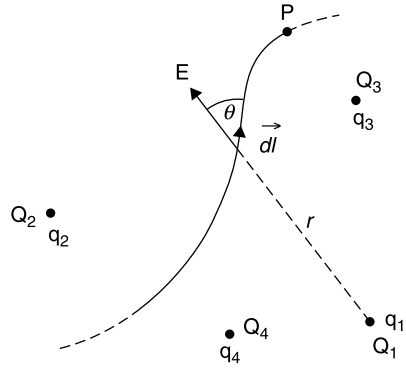
Figure 7.8-1 A work integration path to point P in a field of discrete charges q_1 at position Q_1, q_2 at Q_2, q_3 at Q_3 and q_4 at Q_4 [6, p. 72, with permission].

total potential at P due to a collection of n charges is

$$\Phi = \frac{q_1}{4\pi\varepsilon r_1} + \frac{q_2}{4\pi\varepsilon r_2} + \cdots = \sum_{i=1}^{n} \frac{q_i}{4\pi\varepsilon r_i} \qquad (7.8\text{-}4)$$

If, instead of discrete charges, we have a charge density distribution,

$$\Phi = \frac{1}{4\pi} \int_V \frac{\rho \, dV}{\varepsilon_r \varepsilon_0 r} \qquad (7.8\text{-}5)$$

which defines the static potential function Φ. The volume charge density ρ is a function of position. The constant of integration in (7.8-5) is used to set the zero potential point. If this point is made infinitely far away, the constant is zero.

The effect of individual charges or a charge distribution can be taken into account using Φ. This involves only scalar calculations, the advantage gained by defining the scalar potential function. From Φ we will be able to calculate $\vec{E}$ very easily. On the other hand, determining $\vec{E}$ directly requires that we sum the three directional components of the $\vec{E}$ field due to each charge source, or use a three-dimensional integration for a charge density. This direct approach would result in a much more complex calculation.

Once Φ is determined, the $\vec{E}$ field at any location is given by the *gradient* of Φ, $\nabla\Phi$. *The gradient of a function is the rate of change of the function per unit distance in three mutually orthogonal directions*. In rectangular coordinates it is

$$\nabla\Phi = \vec{x}\frac{\partial\Phi}{\partial x} + \vec{y}\frac{\partial\Phi}{\partial y} + \vec{z}\frac{\partial\Phi}{\partial z} \qquad (7.8\text{-}6)$$

$$\vec{E} = -\nabla\Phi \qquad (7.8\text{-}7)$$

The gradient operates on a scalar function and produces a vector function (Fig. 7.8-2). Note that $\vec{E}$ *is the negative of the gradient of* Φ [the reason is given below
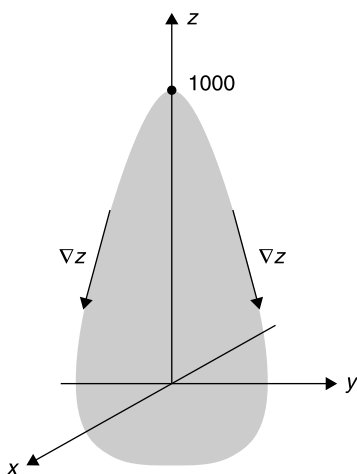
Figure 7.8-2 The ski mountain, with gradient (slope), ∇z.

(7.8-1)]. The gradient can be used with any scalar function; it is not limited to the static potential function defined in this section. For example, consider the function representing an idealized ski mountain:

$$z = 1000 - 100x^2 - 100y^2 \qquad (7.8\text{-}8)$$

where x, y, and z are in meters, z is the height of the mountain, $\pm x$ are the north–south directions, and $\pm y$ are the east–west directions.

If the mountain in Figure 7.8-2 is truncated in any xy plane [$z =$ constant, the section formed is a circle $x^2 + y^2 = (1000 - z)/100$], and therefore the function of (7.8-7) has cylindrical symmetry with respect to z. As a consequence, the fall line of the mountain is the same in any direction (north, south, east, west, or any intermediate direction). At the very top, 1000 m, the slope of the mountain is zero. A skier might just balance there without being drawn downward. However, a movement of 1 m in any direction would result in a drop of 100 m, and at 1 m the slope is 200 m/m. The slope becomes ever steeper with increasing values of x and/or y, as one descends the mountain. The exact slope and its vector direction at any x, y, z point on the mountain is given by

$$\nabla z = (-200x)\vec{x} + (-200y)\vec{y} \qquad (7.8\text{-}9)$$

where x and y must satisfy (7.8-7) to be points on the surface of the mountain. Notice that the gradient is independent of the height of the mountain, depending only on the slope at the location (x, y, z) on the mountain's surface.

The ski mountain example, $z = f(x, y)$, was chosen in order to depict the gradient graphically. However, a static potential function, Φ, in general would

be a function of x, y, and z. That is, $\Phi = \Phi(x, y, z)$, and the resultant gradient is three dimensional, as indicated by (7.8-6).

7.9 DIVERGENCE OF THE B FIELD

Previously, we related the electric flux density and the permittivity (dielectric constant) ε, to the electric field $\vec{E}$ such that $\vec{D} = \varepsilon\vec{E}$. Similarly, it is useful to define the magnetic field $\vec{H}$ that is related to the *magnetic flux density* $\vec{B}$ by a property of the magnetic medium, the *magnetic permeability* μ. The reason for defining two variables, $\vec{B}$ and $\vec{H}$, is that $\vec{B}$ has zero divergence, while $\vec{H}$ does not. The two variables allow the "Amperian" currents (spins) of the molecules of magnetic materials to be accounted for by μ_R, and then the right-hand side of Ampere's law does not include these Amperian currents. Ampere's law is discussed in the next section. Finally, the normal component of the $\vec{B}$ field does not change across a magnetic material boundary, while that of the $\vec{H}$ field may:

$$\vec{B} = \mu\vec{H} \tag{7.9-1}$$

Thus the *electric and magnetic flux densities are $\vec{D}$ and $\vec{B}$*, respectively, and the *electric and magnetic fields are $\vec{E}$ and $\vec{H}$*, respectively. For homogeneous (uniform) and isotropic media (wherein the electric and magnetic properties are the same for all directions and locations) both ε and μ are constants.

Unlike the electric flux density $\vec{D}$, which has as its vector sources discrete charges, q, there are no point sources for the magnetic flux density $\vec{B}$ (Fig. 7.9-1). Stated another way, *the divergence of $\vec{B}$ is zero:*
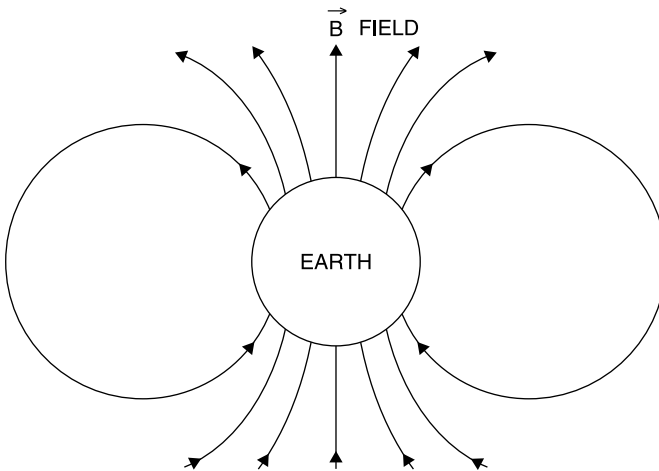
Figure 7.9-1   Divergence of the $\vec{B}$ field is zero, illustrated using an idealized sketch of the Earth's magnetic field.

$$\nabla \cdot \mathbf{B} = 0 \qquad (7.9\text{-}2)$$

Under static conditions the $\vec{\mathbf{B}}$ field lines, when extended sufficiently, form closed contours.

7.10 AMPERE'S LAW

Just as Gauss's law related electric flux density to contained charge, similarly, Ampere's law relates magnetic field to current.

Ampere's law states that the line integral of the $\vec{\mathbf{H}}$ *field parallel to the line defining a closed path is equal to the current I enclosed by the path:*

$$\oint \vec{\mathbf{H}} \cdot \vec{\mathrm{dl}} = \mathrm{I} \qquad (7.10\text{-}1)$$

For the geometry shown in Figure 7.10-1, consisting of a circle coaxial with a conductor carrying current I, symmetry requires that the H_φ is independent of φ. Since $\mathrm{dl} = \mathrm{r}\,\mathrm{d}\varphi$, integration around the closed circular path, from $\varphi = 0$ to $\varphi = 2\pi$, leads to

$$2\pi \mathrm{r} H_\varphi = \mathrm{I} \quad \text{or} \quad H_\varphi = \frac{\mathrm{I}}{2\pi \mathrm{r}} \qquad (7.10\text{-}2)$$
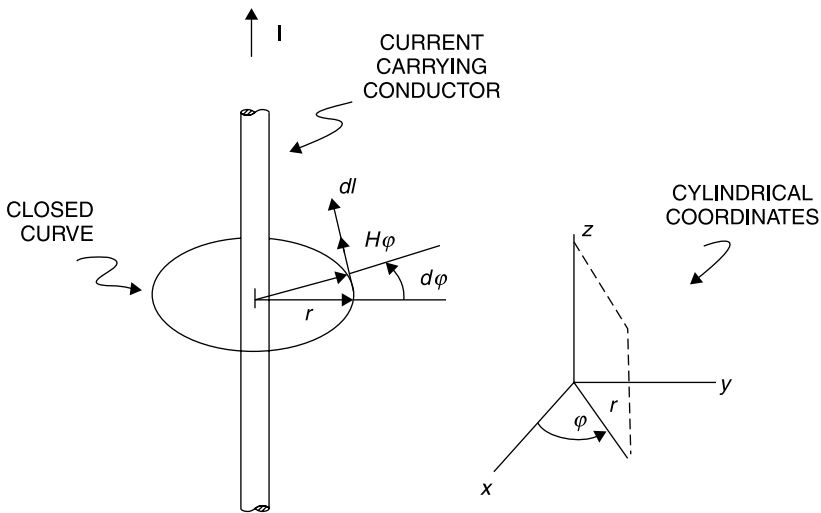
Figure 7.10-1 Closed curve defined about a current-carrying conductor (current I is in the +z direction).

This is a static relation. Both H_φ and I are constant (DC) values that are assumed to have been in place for "an infinite period of time." The effects of time-varying fields and currents will be discussed later.

7.11 VECTOR CURL

Figure 7.10-1 describes a situation wherein the integration of the force field ($\vec{H}$ in this case) around a closed path *is not zero*. This phenomenon occurs often in the physical world, for example, in the rotational force of a tornado. Thus the integral around the closed path represents the rotational component of the field. This concept can be defined at a point by taking the limit as the path and its enclosed area tend to zero (become infinitesimal). For a given vector field $\vec{A}$, this mathematical operation is known as the curl of $\vec{A}$ (*pronounced "del cross $\vec{A}$" or the "curl of $\vec{A}$"*) and is written as $\nabla \times \vec{A}$.

The curl of a vector $\vec{A}$ at a point i is equal to the limit as $\Delta S_i \to 0$ of the line integral of $\vec{A} \cdot \overrightarrow{dl}$ about a path enclosing surface S_i:

$$[\text{curl } \vec{A}]_i \equiv \nabla \times \vec{A} \equiv \lim_{\Delta S_i \to 0} \frac{\oint \vec{A} \cdot \overrightarrow{dl}}{\Delta S_i} \tag{7.11-1}$$

This curl is a general vector identity and can be applied to any vector field. *The curl of a vector field is also a vector field.* The curl can be envisioned by imagining that the $\vec{A}$ field represents a wind or water force on a paddle wheel placed in the field [4, Chapter 2]. For example, if a paddle wheel with its rotational axis oriented in the z direction is placed in the xy plane in Figure 7.11-1 and if the line integral of $\vec{A} \cdot \overrightarrow{dl}$ is positive for the closed path shown, then the paddle wheel would rotate counterclockwise when viewed from above on the +z axis, as shown.

The paddle wheel test is not as whimsical as it first appears. It might seem that the curl of a vector is simply a measure of its circulation, but this is not so. It is possible to have a field that forms circles, yet whose curl is zero, as will be shown by the coaxial line example later in this section.

The components of the curl can be calculated separately. For example, in rectangular coordinates the z component of the curl is calculated using the geometry of Figure 7.11-1 [6, pp. 96–97] for a surface with sides dx and dy:

$$\oint \vec{A} \cdot \overrightarrow{dl} = dy\, A_y|_{x+dx} - dx\, A_x|_{y+dy} - dy\, A_y|_x + dx\, A_x|_y \tag{7.11-2}$$

In the limit as the surface defined by dx and dy tends to zero

$$A_x|_{y+dy} = A_x|_y + dy\, \frac{\partial A_x}{\partial y}\bigg|_y \quad \text{and} \quad A_y|_{x+dx} = A_y|_x + dx\, \frac{\partial A_y}{\partial x}\bigg|_x \tag{7.11-3a,b}$$

Then
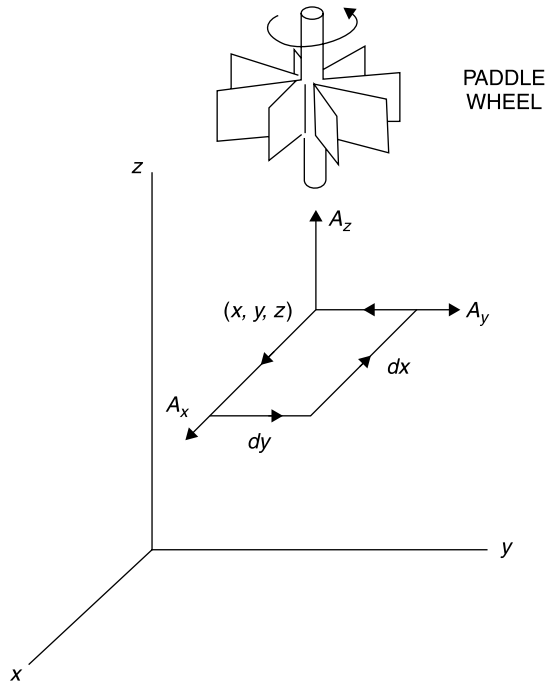
Figure 7.11-1 Determination of the z component of the curl of a vector $\vec{A}$ [6, p. 96] with paddle wheel model [4, p. 24, reprinted with permissions].

$$\oint \vec{A} \cdot \vec{dl} = \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right) dx\, dy \qquad (7.11\text{-}4)$$

and applying the definition of the curl gives the z-directed component as

$$(\nabla \times \vec{A})_{z\,\text{direction}} = \left(\frac{\oint \vec{A} \cdot \vec{dl}}{dx\, dy}\right)\vec{z} = \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right)\vec{z} \qquad (7.11\text{-}5)$$

In like fashion the x- and y-directed components of the curl can be evaluated to produce

$$\nabla \times \vec{A} = \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right)\vec{x} + \left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right)\vec{y} + \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right)\vec{z} \quad (7.11\text{-}6)$$

The curl can be written using the shorthand format of the determinant as

$$\nabla \times \vec{A} = \begin{vmatrix} \vec{x} & \vec{y} & \vec{z} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ A_x & A_y & A_z \end{vmatrix} \qquad (7.11\text{-}7)$$

Expressions (7.11-6) and (7.11-7) produce the same result. Using the curl the differential form of Ampere's law is written:

$$\nabla \times \vec{H} = \vec{J}_C + \frac{\partial \vec{D}}{\partial t} \tag{7.11-8}$$

where $\vec{J}_C$ is the *conduction current* density and $\partial \vec{D}/\partial t$ is the *displacement current density* at the location at which the curl is taken. The displacement current is due to a changing electric field, as, for example, the current that flows between the plates of a capacitor when it is charging or discharging. The displacement current is present with any time-varying $\vec{D}$ field. When a sinusoidally varying voltage is applied to a capacitor, the changing $\vec{D}$ field between the capacitor plates produces the AC displacement current density between the plates.

While the definition of the curl has some similarities to the divergence, the two operations are very different in that *taking the divergence of a vector field produces a scalar value* while *taking the curl of a vector field produces another vector field*. Also, the *curl is a differential relationship that applies at a point in space*. The significance of this point relationship can be appreciated from the following example.

Suppose that it is desired to find the $\vec{H}$ field within the coaxial pair of conductors shown in Figure 7.11-2 [6, p. 94]. Further assume that I is a DC current. Because of the cylindrical symmetry, H_φ is independent of φ; and this allows easy application of Ampere's law. We have two forms for Ampere's law, the line integral relation of (7.10-1) and the point relation of (7.11-8). The line
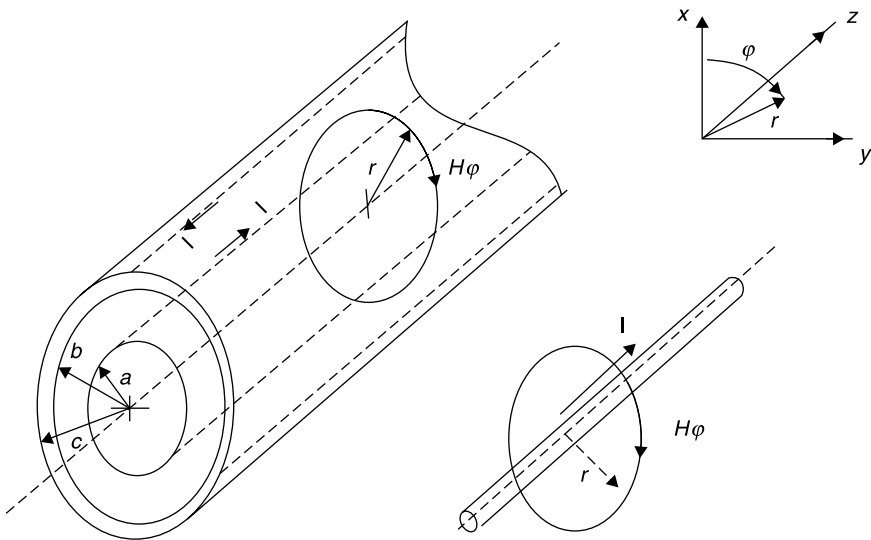
Figure 7.11-2 Magnetic field H_φ about a current-carrying wire, and the result extended to the space within a coaxial transmission line.
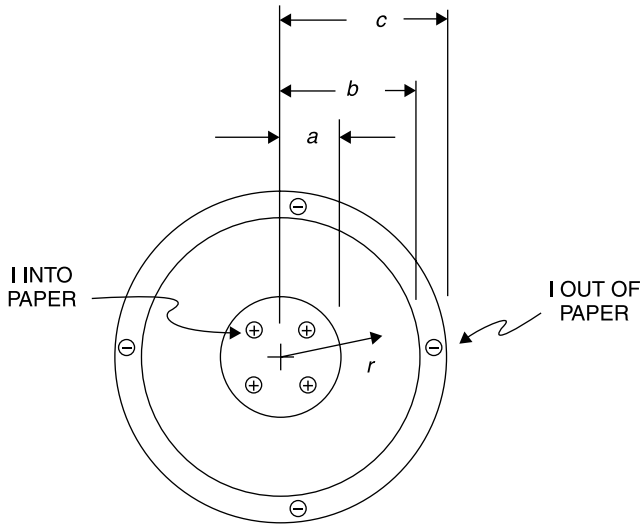
Figure 7.11-3   Cross section of the coaxial line.

integral format requires specification of the *total current within the closed contour* while the point format requires the current density *at the point* at which $\vec{H}$ is to be computed. Suppose the point format is used to calculate H_φ within the center conductor, that is, for $r \le a$ (Fig. 7.11-3).

Under DC current excitation, I is uniformly distributed over the area of the center conductor, and the current density has only a conduction component given by

$$J_C = \frac{I}{\pi a^2} \tag{7.11-9}$$

Applying (7.11-8),

$$\nabla \times \vec{H} = \frac{\vec{I}}{\pi a^2} \tag{7.11-10}$$

Expressing (7.11-10) in cylindrical coordinates (vector operators in various coordinates are provided in Section 7.14) gives

$$\nabla \times \vec{H} = \vec{r}\left[\frac{1}{r}\frac{\partial H_z}{\partial \varphi} - \frac{\partial H_\varphi}{\partial z}\right] + \vec{\varphi}\left[\frac{\partial H_r}{\partial z} - \frac{\partial H_z}{\partial r}\right] + \vec{z}\left[\frac{1}{r}\frac{\partial(rH_\varphi)}{\partial r} - \frac{1}{r}\frac{\partial H_r}{\partial \varphi}\right] = \frac{\vec{I}}{\pi a^2} \tag{7.11-11}$$

In this example, I has only a z component. Thus, equating the z-directed terms in (7.11-14) and noting that, due to symmetry, there can only be a φ component

of $\vec{H}$ gives

$$\frac{1}{r}\frac{\partial(rH_\varphi)}{\partial r} = \frac{I}{\pi a^2}$$

$$\frac{\partial(rH_\varphi)}{\partial r} = \frac{I}{\pi a^2}r \qquad (7.11\text{-}12)$$

and integrating gives

$$rH_\varphi = \frac{I}{\pi a^2}\frac{r^2}{2}$$

$$H_\varphi = \frac{I}{2\pi a^2}r \quad \text{for } 0 \le r \le a \qquad (7.11\text{-}13)$$

To calculate the value of $\vec{H}$ in the space between the conductors we cannot apply the curl expression because *the current in this region is zero*. This would seem to be a contradiction to what has been studied thus far. Specifically, we know that there is an H_φ between the conductors and this field closes on itself, forming circles, suggesting a finite value of curl. Resolving these two dilemmas is important to understanding the curl and the point form of Ampere's law as expressed by (7.11-8).

First, the value of H_φ was evaluated previously from the line integral form of Ampere's law for an axially directed current in (7.11-2). The result repeated here is

$$H_\varphi = \frac{I}{2\pi r} \quad \text{for } a < r < b \qquad (7.11\text{-}14)$$

Although there is a finite value for H_φ between the conductors, the value of the curl in this region must be zero, since there is no current density there. To see that the curl is zero, consider that only the fifth term on the left side of (7.11-11) could be nonzero since there is only an H_φ component and it has only an r variation. But this term is

$$\nabla \times \vec{H} = \vec{r}\frac{1}{r}\frac{\partial(rH_\varphi)}{\partial r} \qquad (7.11\text{-}15)$$

and substituting in the value for H_φ gives

$$\frac{1}{r}\frac{\partial(rH_\varphi)}{\partial r} = \frac{1}{r}\frac{\partial(r(I/2\pi r))}{\partial r} = 0 \qquad (7.11\text{-}16)$$

Thus, even though there is an $\vec{H}$ field between the conductors, its value cannot be found using the differential form of Ampere's law.
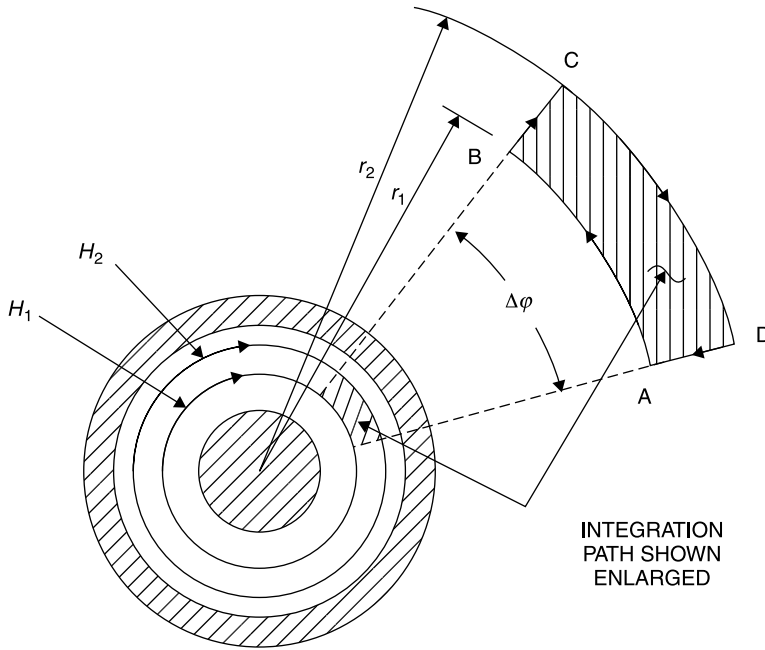
Figure 7.11-4 Closed-path integration to evaluate the curl of H between coaxial conductors.

Next, consider the second *apparent* dilemma: the $\vec{H}$ field as sketched in Figure 7.11-2 closes on itself and *appears* to have a finite value of curl. Actually, the curl is zero, as was shown by (7.11-16). To arrive at this conclusion in another way, use the line integral method of evaluating the curl about the closed path ABCD shown in Figure 7.11-4. The radial sides of the path, BC and DA produce no contribution to the integral of $\vec{H} \cdot dl$ since there is only a circumferential $\vec{H}$ field component H_φ. The contributions to the circumferential portions of the path, AB and CD, produce

$$H_\varphi \cdot dl = H_1 r_1 \, \Delta\varphi - H_2 r_2 \, \Delta\varphi = \Delta\varphi \left[\frac{I}{2\pi r_1} r_1 - \frac{I}{2\pi r_2} r_2 \right] = 0 \quad (7.11\text{-}17)$$

Thus, it can be seen that the *curvature of the H field and closure on itself does not always imply a finite value of curl*. In this case, the variation of H_φ with r is just sufficient to cancel the curl. Analogously, if a test paddle wheel with rotational axis in the z direction is placed between the conductors, it would not rotate were the $\vec{H}$ field replaced with a wind current pressure having the same r variation. It might move with the field following a circular orbit about the center conductor, but *it would not rotate about its own axis*.

To compete the analysis of the H_φ in the coaxial line cross section, consider the range r > c. Applying (7.10-1),
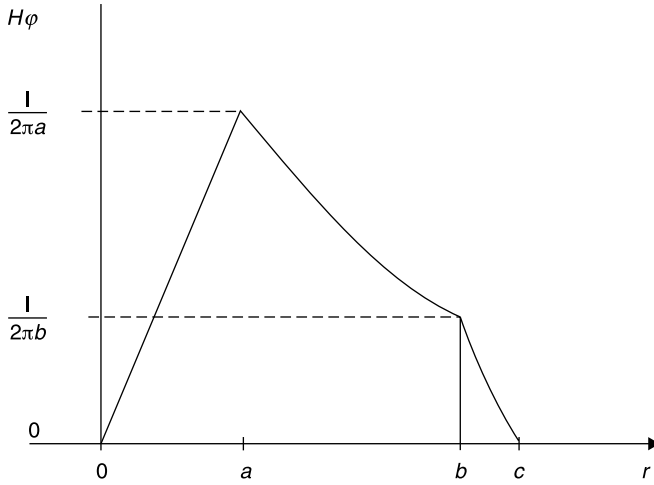
Figure 7.11-5 Magnitude H field of the coaxial conductor set as a function of the radius.

$$\oint \vec{H}_\varphi \cdot \overrightarrow{dl} = I - I = 0 \qquad (7.11\text{-}18)$$

$$H_\varphi = 0 \quad \text{for } r > c \qquad (7.11\text{-}19)$$

The field outside the outer conductor is zero because the net current within is zero, a positive current on the center conductor and an equal and opposite current returning on the outer conductor. The field as a function of r is shown in Figure 7.11-5.

7.12 FARADAY'S LAW OF INDUCTION

Faraday's law of induction states that a time-varying magnetic flux passing through a surface produces a voltage around the perimeter of the surface equal to the surface integral of the changing flux through the surface. The polarity of the induced voltage is such as to establish a current to resist the change of magnetic flux:

$$\oint \vec{E} \cdot \overrightarrow{dl} = -\int_S \frac{\partial \vec{B}}{\partial t} \cdot \overrightarrow{dS} \qquad (7.12\text{-}1)$$

Independently, Lenz duplicated many of Faraday's and Ampere's discoveries. Lenz is credited with the law for determining the polarity of the voltage induced by a changing magnetic field. When Faraday's law is written in differ-
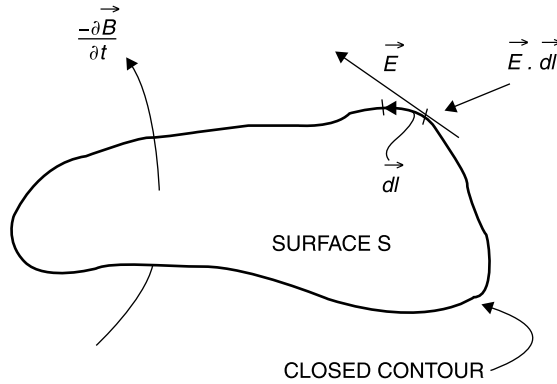
Figure 7.12-1   Diagram for the right-hand sense application of Faraday's law.

ential form using the curl definition, we get the third of Maxwell's equations:

$$\nabla \times \vec{\mathbf{E}} = -\frac{\partial \vec{\mathbf{B}}}{\partial t} \qquad (7.12\text{-}2)$$

As defined in (7.12-1) and (7.12-2), the sense of Faraday's law can be re-membered by the use of the right-hand rule. If the fingers of the right hand curl in the direction of the path that gives a positive value for the line integral of $\vec{\mathbf{E}} \cdot \vec{dl}$, then the thumb gives the direction of $-\partial\vec{\mathbf{B}}/\partial t$, as shown in Figure 7.12-1. It is important to note that in Figure 7.12-1 it is not the direction of the $\vec{\mathbf{B}}$ field that is shown but rather the *direction of the negative rate of change of the* $\vec{\mathbf{B}}$ *field*.

The negative sign in (7.12-2) is the result of *Lenz's law which states that the induced electric field polarity is such as to create a current that opposes a change in the magnetic field.*

7.13   MAXWELL'S EQUATIONS

Maxwell's Four Equations

Understanding propagation of electromagnetic signals begins with understanding physically how electric and magnetic fields interact with one another. The empiric observations just presented that describe the behavior of $\vec{\mathbf{E}}$ and $\vec{\mathbf{H}}$ fields and their auxiliary $\vec{\mathbf{D}}$ and $\vec{\mathbf{B}}$ fields were interrelated eloquently in a unified theory in about 1882 by James Clerk Maxwell. The theory contained numerous equations and concepts then. However, some time later Oliver Heaviside reduced them to four principal equations together with auxiliary relationships. Maxwell's four equations in both differential and integral formats are summarized as follows:

1. *The divergence of the $\vec{\mathbf{D}}$ field equals the volume charge density.* This is Gauss's law:

$$\nabla \cdot \vec{\mathbf{D}} = \rho \qquad\qquad (7.13\text{-}1a)$$

$$\oint_S \vec{\mathbf{D}} \cdot \vec{dS} = \int_V \rho \, dv \qquad\qquad (7.13\text{-}1b)$$

2. *The divergence of the $\vec{\mathbf{B}}$ field is zero:*

$$\nabla \cdot \vec{\mathbf{B}} = 0 \qquad\qquad (7.13\text{-}2a)$$

$$\oint_S \vec{\mathbf{B}} \cdot \vec{dS} = 0 \qquad\qquad (7.13\text{-}2b)$$

3. *The curl of the $\vec{\mathbf{E}}$ field is equal to minus the time rate of change of the magnetic flux density:*

$$\nabla \times \vec{\mathbf{E}} = -\frac{\partial \vec{\mathbf{B}}}{\partial t} \qquad\qquad (7.13\text{-}3a)$$

$$\oint \vec{\mathbf{E}} \cdot \vec{dl} = -\int_S \frac{\partial \vec{\mathbf{B}}}{\partial t} \cdot \vec{dS} \qquad\qquad (7.13\text{-}3b)$$

4. *The curl of the $\vec{\mathbf{H}}$ field equals the combined conduction and displacement current densities:*

$$\nabla \times \vec{\mathbf{H}} = \vec{\mathbf{J}}_C + \frac{\partial \vec{\mathbf{D}}}{\partial t} \qquad\qquad (7.13\text{-}4a)$$

$$\oint \vec{\mathbf{H}} \cdot \vec{dl} = \oint_S \vec{\mathbf{J}_C} \cdot \vec{dS} + \oint_S \frac{\partial \vec{\mathbf{D}}}{\partial t} \cdot \vec{dS} \qquad\qquad (7.13\text{-}4b)$$

Auxiliary Relations and Definitions

To Maxwell's famous four equations must be added definitions and auxiliary relations to complete the experimental evidence unified by Maxwell. These are:

The Force Law   This applies to a charge q moving at velocity $\vec{v}$ in a region having both electric field $\vec{\mathbf{E}}$ and magnetic flux density $\vec{\mathbf{B}}$:

$$\vec{f} = q[\vec{\mathbf{E}} + \vec{v} \times \vec{\mathbf{B}}] \qquad\qquad (7.13\text{-}5)$$

Ohm's Law and the Definition of Conduction Current Density   The point form of *Ohm's law* is

$$\vec{\mathbf{J}}_C = \sigma \vec{\mathbf{E}} \qquad\qquad (7.13\text{-}6)$$

where $\vec{\mathbf{J}}_C$ is the current density (in amperes/square meter), σ is the conductivity of the material (in mhos/meter), and E is the electric field (in volts/meter).

Convection Current Density   The definition of *convection current density* is

$$\vec{\mathbf{J}}_{CV} \equiv \rho\vec{\mathbf{v}}_\rho \qquad (7.13\text{-}7)$$

where $\vec{\mathbf{v}}_\rho$ is the velocity at which the volume charge density ρ is moving [6, Sec. 4.06].

Permittivity (Dielectric Constant)   The definition of *permittivity* (*dielectric constant*) is defined by

$$\vec{\mathbf{D}} \equiv \varepsilon\vec{\mathbf{E}} = \varepsilon_0\varepsilon_R\vec{\mathbf{E}} \qquad (7.13\text{-}8)$$

where $\varepsilon_0 = 8.854186 \times 10^{-12}$ ($\approx 10^{-9}/36\pi$) farads per meter is the permittivity of free space and ε_R is the *relative dielectric constant* used to account for the effects of atomic and molecular dipoles of various materials.

Permeability   The definition of *permeability* is defined within

$$\vec{\mathbf{B}} = \mu\vec{\mathbf{H}} = \mu_0\mu_R\vec{\mathbf{H}} \qquad (7.13\text{-}9)$$

where $\mu_0 = 4\pi \times 10^{-7}$ H/m and μ_R is the *relative permeability* used to account for the magnetic dipole moments of atoms within various materials.

In general μ and ε can be *tensor quantities*, whose values vary with position and direction. For example, a specialty of microwave engineering is the design of artificial (anisotropic) dielectrics [9, Chapter 12], in which metal strips are embedded in an otherwise homogeneous dielectric to produce a composite dielectric having very different relative dielectric constants for different applied electric field directions. Similarly, microwave ferrite materials [10, Sec. 6.6] exhibit quite different relative permeabilities with direction when immersed within a strong biasing magnetic field. However, for many engineering purposes μ and ε can be treated as simple scalar constants.

Visualizing Maxwell's Equations

To review, consider the first of Maxwell's equations:

$$\nabla \cdot \vec{\mathbf{D}} = \rho \quad \text{or} \quad \nabla \cdot \varepsilon\vec{\mathbf{E}} = \rho \qquad (7.13\text{-}10)$$

This equation states that the "divergence" of the *flux density* $\vec{\mathbf{D}}$ is equal to the charge density in the volume from which the $\vec{\mathbf{D}}$ field is emanating. When the $\vec{\mathbf{D}}$ field lines exhibit a net departure (*divergence*) from any volume (even a microscopic one), there is a net positive electric charge inside. Vice versa, if the lines
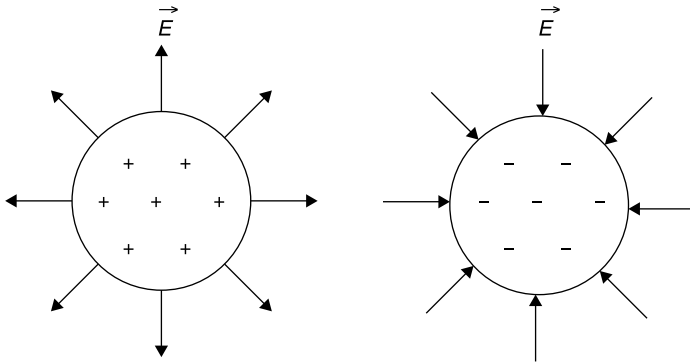
Figure 7.13-1   Divergence of the $\vec{E}$ field is proportional to the contained charge density, and its direction is related to the sign of the net charge, as shown.

enter the volume, there is a net negative charge inside. Departing lines mean that a *positive charge* outside the volume will experience a force that repels it from the volume. That makes sense since *like charges repel*. The $\vec{E}$ field is proportional to the $\vec{D}$ field and the $\vec{E}$ field emerges or enters a volume in proportion to the net positive or negative charge within, respectively, as shown in Figure 7.13-1.

Next consider Maxwell's second equation:

$$\nabla \cdot \vec{B} = 0 \qquad (7.13\text{-}11)$$

This equation simply indicates that *the B field has no divergence* (Fig. 7.13-2). This means that the $\vec{B}$ field lines are continuous throughout space, they do not originate anywhere, and they always close on themselves. For homogeneous regions, the $\vec{H}$ field lines also close on themselves, being proportional to $\vec{B}$. That is, $\vec{B} = \mu\vec{H}$.

Because the $\vec{B}$ field is continuous, even through a boundary dividing regions of different permeability, it follows that the $\vec{H}$ field cannot be continuous through the permeability boundary. Its field strength must reflect the ratio of permeabilities in the two regions, as shown in Figure 7.13-3:

$$\vec{H}_2 = \frac{\mu_1}{\mu_2}\vec{H}_1 \qquad (7.13\text{-}12)$$

The third of Maxwell's equations introduces the curl concept, the amount of circulation of a vector field:

$$\nabla \times \vec{E} = -\frac{\partial\vec{B}}{\partial t} \qquad (7.13\text{-}13)$$

This says that a changing $\vec{B}$ field induces an $\vec{E}$ field at right angles. The $\vec{E}$ field closes on itself in a closed loop or curl. This is the incremental form of Far-
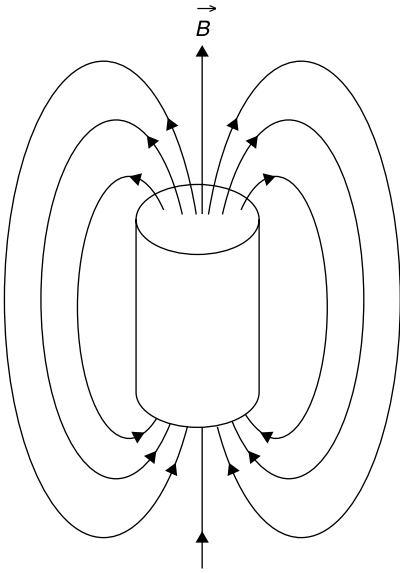
Figure 7.13-2 Bar magnet, illustrating that since the divergence of the $\vec{B}$ field is zero, the field lines, when sufficiently extended, always close on themselves.

aday's law, which says that the voltage induced in a loop is proportional to the time rate of change of the magnetic field within the loop and (Lenz's law) of such a polarity as to oppose the change in the $\vec{B}$ field.

The applications of this phenomenon are widespread, including transformers for power distribution and the coils used to generate high voltages for the spark plugs of gasoline engines. We shall see that this and the next equation are responsible for the propagation of radio waves, the essence of wireless signal propagation.

The fourth, and last, of Maxwell's equations, states that the curl of the $\vec{H}$ field is proportional to the enclosed current:

$$\nabla \times \vec{H} = \vec{J}_C + \frac{\partial \vec{D}}{\partial t} \qquad (7.13\text{-}14)$$
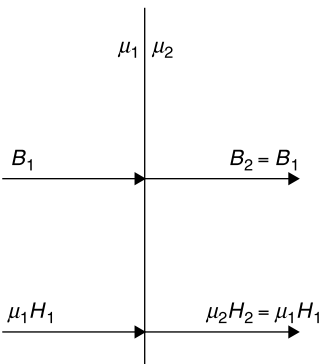
Figure 7.13-3 While the $\vec{B}$ lines have zero divergence (are always continuous), the $\vec{H}$ lines change across a boundary to account for a change in permeability from one region to the next.

which has two components: (1) the conduction current density and (2) the time rate of change of the electric flux density $\partial\vec{D}/\partial t$ (the displacement current density).

The two current forms in this equation have many important practical applications. For example, a conduction current in a vertical wire (such as a monopole antenna) provides a means of exciting a time-varying $\vec{H}$ field in the space about the antenna. In turn, the time-varying $\vec{H}$ field produces a time-varying $\vec{E}$ field perpendicular to the $\vec{H}$ field (from Maxwell's third equation). We shall see that this combined action generates a *propagating electromagnetic waveform* in which each field is induced by the time derivative of the other. Once one field is created the other is induced by the time rate of change of the first, and the resulting waveforms propagate away from the source as a *radio wave*.

7.14   PRIMARY VECTOR OPERATIONS

In the course of describing the behavior of electromagnetic fields, we have found it useful to define and apply basic vector operations such as the vector dot product, cross product, gradient, divergence, and curl. These vary in form. Some are performed on vectors to yield new vectors, some are performed on vectors to yield scalar quantities, and some are performed on scalar quantities to yield vectors. The fundamental vector operations that we have introduced are summarized in Table 7.14-1.

TABLE 7.14-1   Basic Vector Operators in Rectangular Coordinates (1a,b,c,d,e)[a]

Name	Symbol	Operates on	Yields	Formula
Scalar product (also dot product)	$\vec{A}\cdot\vec{B}$	2 vectors	Scalar function	$\vec{A}\cdot\vec{B} = A_xB_x + A_yB_y + A_zB_z$
Vector product (also cross product)	$\vec{A}\times\vec{B}$	2 vectors	1 vector	$\vec{A}\times\vec{B} = \begin{vmatrix} \vec{x} & \vec{y} & \vec{z} \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix}$
Divergence	$\nabla\cdot\vec{A}$	1 vector	Scalar function	$\nabla\cdot\vec{A} = \dfrac{\partial}{\partial x}A_x + \dfrac{\partial}{\partial y}A_y + \dfrac{\partial}{\partial z}A_z$
Curl	$\nabla\times\vec{A}$	1 vector	1 vector	$\nabla\times\vec{A} = \begin{vmatrix} \vec{x} & \vec{y} & \vec{z} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ A_x & A_y & A_z \end{vmatrix}$
Gradient	$\nabla\Phi$	Scalar function	1 vector	$\nabla\Phi = \left(\vec{x}\dfrac{\partial\Phi}{\partial x} + \vec{y}\dfrac{\partial\Phi}{\partial y} + \vec{z}\dfrac{\partial\Phi}{\partial z}\right)$

[a]Where $\vec{A}$ and $\vec{B}$ are vectors $\vec{A} = A_x\vec{x} + A_y\vec{y} + A_z\vec{z}$ and $\vec{B} = B_x\vec{x} + B_y\vec{y} + B_z\vec{z}$, and $\Phi(x, y, z)$ is a scalar function.

The use of vector operations is simplified by using coordinate systems that exploit the symmetry of the application. In addition to rectangular coordinates, cylindrical and spherical coordinates are often useful. Rather than convert the application to rectangular coordinates, it is more convenient to carry out the vector operations directly in the coordinate system of choice. The vector operations are listed below in cylindrical and spherical coordinates.

Cylindrical Coordinates

$$\nabla\Phi = \vec{r}\frac{\partial\Phi}{\partial r} + \vec{\varphi}\frac{1}{r}\frac{\partial\Phi}{\partial\varphi} + \vec{z}\frac{\partial\Phi}{\partial z} \tag{7.14-1a}$$

$$\nabla\cdot\vec{A} = \frac{1}{r}\frac{\partial(rA_r)}{\partial\vec{r}} + \frac{1}{r}\frac{\partial A_\varphi}{\partial\varphi} + \frac{\partial A_z}{\partial z} \tag{7.14-1b}$$

$$\nabla\times\vec{A} = \vec{r}\left[\frac{1}{r}\frac{\partial A_z}{\partial\varphi} - \frac{\partial A_\varphi}{\partial z}\right] + \vec{\varphi}\left[\frac{\partial A_r}{\partial z} - \frac{\partial A_z}{\partial r}\right] + \vec{z}\left[\frac{1}{r}\frac{\partial(rA_\varphi)}{\partial r} - \frac{1}{r}\frac{\partial A_r}{\partial\varphi}\right] \tag{7.14-1c}$$

Spherical Coordinates

$$\nabla\Phi = \vec{r}\frac{\partial\Phi}{\partial r} + \vec{\theta}\frac{1}{r}\frac{\partial\Phi}{\partial\theta} + \vec{\varphi}\frac{1}{r\sin\theta}\frac{\partial\Phi}{\partial\theta} \tag{7.14-2a}$$

$$\nabla\cdot A = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2A_r) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta A_\theta) + \frac{1}{r\sin\theta}\frac{\partial A_\varphi}{\partial\varphi} \tag{7.14-2b}$$

$$\nabla\times\vec{A} = \vec{r}\frac{1}{r\sin\theta}\left[\frac{\partial}{\partial\theta}(A_\varphi\sin\theta) - \frac{\partial A_\theta}{\partial\varphi}\right] + \vec{\theta}\frac{1}{r}\left[\frac{1}{\sin\theta}\frac{\partial A_r}{\partial\varphi} - \frac{\partial}{\partial r}(rA_\varphi)\right]$$

$$+ \vec{\varphi}\frac{1}{r}\left[\frac{\partial}{\partial r}(rA_\theta) - \frac{\partial A_r}{\partial\theta}\right] \tag{7.14-2c}$$

7.15   THE LAPLACIAN

After applying an operation from Table 7.14-1, it is often useful to apply the same or another operation to the result. One of the most important is the *Laplacian*. It occurs so often that it is considered a separate operator. This operator has two definitions, one for operation on a scalar function and one for operation on a vector field. *When applied to a scalar function* $\Phi(x, y, z)$, *the Laplacian is equal to the divergence of the gradient*. Its value is the sum of the second partial derivatives with respect to x, y, and z of Φ. This can be evaluated by first evaluating the gradient of Φ and then finding the divergence. The result [4, p. 30] is:

Rectangular Coordinates

$$\nabla^2\Phi = \nabla\cdot\nabla\Phi = \frac{\partial^2\Phi}{\partial x^2} + \frac{\partial^2\Phi}{\partial y^2} + \frac{\partial^2\Phi}{\partial z^2} \tag{7.15-1a}$$

Cylindrical Coordinates

$$\nabla^2\Phi = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\Phi}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2\Phi}{\partial\varphi^2} + \frac{\partial\Phi^2}{\partial z^2} \qquad (7.15\text{-}1b)$$

Spherical Coordinates

$$\nabla^2\Phi = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\Phi}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Phi}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2\Phi}{\partial\varphi^2} \quad (7.15\text{-}1c)$$

Recall that the electric field $\vec{E}$ is the negative of the gradient of the potential function ($\vec{E} = -\nabla\Phi$) and that the divergence of the static electric field (in the absence of changing $\vec{B}$ field) is equal to the charge density divided by the dielectric constant ($\nabla \cdot \vec{E} = \rho/\varepsilon$), thus

$$\nabla^2\Phi = -\frac{\rho}{\varepsilon} \qquad (7.15\text{-}2)$$

This relationship is called *Poisson's equation*. It applies to a region of space containing a charge density, ρ. In a region that is free of charge:

$$\nabla^2\Phi = 0 \qquad (7.15\text{-}3)$$

This relationship is called Laplace's equation. One might ask: If there is no charge, how can there be an electrostatic potential, and accordingly, a need for Laplace's equation? Consider the diagram in Figure 7.15-1.

   In region 1 the contained charge results in a finite value of divergence of the $\vec{E}$ field from the region, and Poisson's equation applies. In region 2 there is no
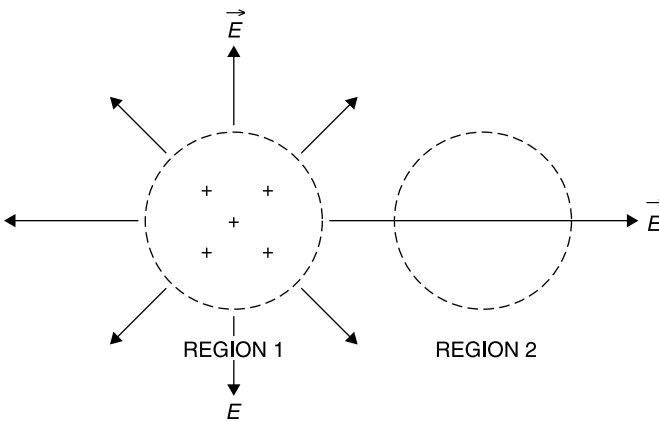
Figure 7.15-1 Poisson's equation applies in a region containing a charge density (region 1), while Laplace's equation is defined in a region without charge (region 2).

charge; the $\vec{E}$ field divergence is zero. However, there is an $\vec{E}$ field due to the charge in the adjacent region 1, and Poisson's equation for region 2 therefore reduces to that of Laplace.

When the Laplacian is applied to a vector field, the result is a vector [4, p. 30]:

$$\nabla^2 \vec{A} = \nabla^2(\vec{x}A_x + \vec{y}A_y + \vec{z}A_z) \tag{7.15-4}$$

in which the ∇^2 operator is defined in rectangular coordinates to be

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{7.15-5}$$

Care must be taken that *the entire ∇^2 operator is applied fully to each component of $\vec{A}$*. Explicitly,

Rectangular Coordinates

$$\nabla^2 \vec{A} = \vec{x}\left(\frac{\partial^2 A_x}{\partial x^2} + \frac{\partial^2 A_x}{\partial y^2} + \frac{\partial^2 A_x}{\partial z^2}\right) + \vec{y}\left(\frac{\partial^2 A_y}{\partial x^2} + \frac{\partial^2 A_y}{\partial y^2} + \frac{\partial^2 A_y}{\partial z^2}\right)$$

$$+ \vec{z}\left(\frac{\partial^2 A_z}{\partial x^2} + \frac{\partial^2 A_z}{\partial y^2} + \frac{\partial^2 A_z}{\partial z^2}\right) \tag{7.15-6}$$

Cylindrical Coordinates

$$\nabla^2 \vec{A} = \vec{r}\left[\nabla^2 A_r - \frac{2}{r^2}\frac{\partial A_\varphi}{\partial \varphi} - \frac{A_r}{r^2}\right] + \vec{\varphi}\left[\nabla^2 A_\varphi + \frac{2}{r^2}\frac{\partial A_r}{\partial \varphi} - \frac{A_\varphi}{r^2}\right] + \vec{z}[\nabla^2 A_z] \tag{7.15-7}$$

where

$$\nabla^2 = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2}{\partial \varphi^2} + \frac{\partial^2}{\partial z^2}$$

Spherical Coordinates

$$\nabla^2 \vec{A} = \vec{r}\left[\nabla^2 A_r - \frac{2}{r^2}\left(A_r + \cot\theta A_\theta + \csc\theta\frac{\partial A_\varphi}{\partial \varphi} + \frac{\partial A_\theta}{\partial \theta}\right)\right]$$

$$+ \vec{\theta}\left[\nabla^2 A_\theta - \frac{1}{r^2}\left(\csc^2\theta A_\theta - 2\frac{\partial A_r}{\partial \theta} + 2\cot\theta\csc\theta\frac{\partial A_\varphi}{\partial \varphi}\right)\right]$$

$$+ \vec{\varphi}\left[\nabla^2 A_\varphi - \frac{1}{r^2}\left(\csc^2\theta A_\varphi - 2\csc\theta\frac{\partial A_r}{\partial \varphi} - 2\cot\theta\csc\theta\frac{\partial A_\theta}{\partial \varphi}\right)\right] \tag{7.15-8}$$

where

$$\nabla^2 = \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2}{\partial \varphi^2}$$

7.16   VECTOR AND SCALAR IDENTITIES

The Laplacian operators of the previous section result from the successive application of operators. The scalar and vector forms of the Laplacian are considered operators, themselves. In addition to the Laplacian, there are numerous other such multiple operations that can be defined and found to be useful. When applied in general form, they become *operational identities*. Many of them may appear to have no obvious physical significance, but these identities nevertheless prove useful in deriving certain electromagnetic results, just as the trigonometric identity $\sin^2 \theta + \cos^2 \theta = 1$ finds numerous applications.

An example of a multiple operation that does have clear physical significance is obtained in first taking the gradient of a scalar function and then taking the curl of the result. Previously, we noted that the value of a scalar function $\Phi(x, y, z)$ is dependent only on the coordinates (x, y, z) and not on how that point is reached. It follows that *the gradient always describes a conservative vector function*, that is to say, the line integral of the gradient about any closed path is zero. Since the gradient is a vector function, we can operate on it to find its curl. But since the curl is the line integral of a vector about a closed path, the curl of the gradient will always be zero. Put simply, *"a field that is the gradient of something has no curl"* [4, p. 29]:

$$\text{Curl of gradient of } \Phi = \nabla \times \text{gradient } \Phi = \nabla \times \nabla \Phi = 0 \qquad (7.16\text{-}1)$$

for any scalar function Φ from which the gradient is derived. This identity has the physical significance of describing how closed-path integrals of conservative functions are zero. The validity of (7.16-1) can be verified by evaluating the gradient of a function and then finding its curl, which leads to equal and opposite partial derivatives. This equality will prove useful in the derivation of the *wave equation*.

A second equality obtained by successive vector operations is that *"a field that is the curl of something has no divergence"* [4, p. 29]:

$$\text{Divergence of curl} = \nabla \cdot \nabla \times \vec{A} = 0 \qquad (7.16\text{-}2)$$

where $\vec{A}$ is any vector field $\vec{A} = A_x \vec{x} + A_y \vec{y} + A_z \vec{z}$. The basic vector operations previously defined were independent of the coordinate system employed. For example, the divergence was defined as the limit of the integral of a vector emerging orthogonal to the surface of a volume as the volume contained by the surface was shrunk to zero. In the previous section, the vector Laplacian was

TABLE 7.16-1   Vector Identities

$\nabla(\Phi + \Psi) = \nabla\Phi + \nabla\Psi$	(7.16-4)
$\nabla \cdot (\vec{A} + \vec{B}) = \nabla \cdot \vec{A} + \nabla \cdot \vec{B}$	(7.16-5)
$\nabla \times (\vec{A} + \vec{B}) = \nabla \times \vec{A} + \nabla \times \vec{B}$	(7.16-6)
$\nabla(\Phi\Psi) = \Phi\nabla\Psi + \Psi\nabla\Phi$	(7.16-7)
$\nabla \cdot (\Psi\vec{A}) = \vec{A} \cdot \nabla\Psi + \Psi\nabla \cdot \vec{A}$	(7.16-8)
$\nabla \cdot (\vec{A} \times \vec{B}) = \vec{B} \cdot \nabla \times \vec{A} - \vec{A} \cdot \nabla \times \vec{B}$	(7.16-9)
$\nabla \times (\Phi\vec{A}) = \nabla\Phi \times \vec{A} + \Phi\nabla \times \vec{A}$	(7.16-10)
$\nabla \cdot \nabla\Phi = \nabla^2\Phi$	(7.16-11)
$\nabla \cdot \nabla \times \vec{A} = 0$	(7.16-12)
$\nabla \times \nabla\Phi = 0$	(7.16-13)
$\nabla \times \nabla \times \vec{A} = \nabla(\nabla \cdot \vec{A}) - \nabla^2\vec{A}$	(7.16-14)
$\vec{A} \times (\vec{B} \times \vec{C}) = \vec{B}(\vec{A} \cdot \vec{C}) - \vec{C}(\vec{A} \cdot \vec{B})$	(7.16-15)

Source: From Ramo and Whinnery [6, p. 114]; reprinted with permission.

defined in three coordinate systems: rectangular, cylindrical, and spherical. Mathematicians do not like to have a fundamental definition dependent upon a coordinate system. To circumvent this impasse, an alternate definition for the vector Laplacian that can be applied to any coordinate system is [6, p. 111, and 4, p. 30]:

$$\nabla^2\vec{A} \equiv \nabla(\nabla \cdot A) - \nabla \times \nabla \times A \qquad (7.16-3)$$

This is an identity. Even so, it would appear to be totally esoteric. Remarkably, however, it is employed in several important proofs, as will be seen in the succeeding sections. Its validity can be demonstrated by showing that the left and right sides of (7.16-3) both give the same result when expressed in rectangular coordinates. Since both the divergence and the curl of the curl are both expressible in any orthogonal coordinate system, *the vector Laplacian can be considered to be defined by (7.16-3)* making the definition independent of the coordinate system.

Several vector identities resulting from successive vector operations are listed in Table 7.16-1. They can be verified by performing the indicated operations that their equations represent.

7.17   FREE CHARGE WITHIN A CONDUCTOR

The derivation of the charge distribution in a conductor is a practical example of the application of Maxwell's equations and Ohm's law [6, Sec. 6.03]. Suppose that a quantity of free charge having a volume density ρ is somehow introduced into a conductor. We solve for its distribution by beginning with Ohm's law for the current density:

$$\vec{J} = \sigma\vec{E} \tag{7.17-1}$$

where σ is the conductivity of the conductor, and the regular type font indicates that (7.17-1) applies to all time variations. When this value of $\vec{J}$ is substituted into Maxwell's fourth equation the result is

$$\nabla \times \vec{H} = \sigma\vec{E} + \frac{\partial\vec{D}}{\partial t} \tag{7.17-2}$$

Now take the divergence of both sides of (7.17-2) and note that the divergence of the curl of any vector field is zero:

$$0 = \nabla \cdot \nabla \times \vec{H} = \frac{\sigma}{\varepsilon}\nabla \cdot \vec{D} + \frac{\partial(\nabla \cdot \vec{D})}{\partial t} \tag{7.17-3}$$

From Maxwell's first equation $\nabla \cdot \vec{D} = \rho$, where ρ in this case is the free charge density, allowing (7.17-3) to be written as

$$\frac{\sigma}{\varepsilon}\rho + \frac{\partial\rho}{\partial t} = 0 \tag{7.17-4}$$

The function that is unchanged by differentiation is the exponential, and therefore the solution to (7.17-4) is

$$\rho(t) = \rho_0 e^{-(\sigma/\varepsilon)t} \tag{7.17-5}$$

This means that any free charge density within a conductor decays exponentially in time with a time constant σ/ε. Since we have not provided for a means by which the charge can be annihilated, it follows that it must flow to the surface of the conductor. The time constant for copper, assuming the dielectric constant of free space within copper, would be

$$\frac{\sigma}{\varepsilon} = \frac{5.8 \times 10^7 \text{ \textmho/m}}{8.6 \times 10^{-12} \text{ F/m}} = 7 \text{ μs} \tag{7.17-6}$$

It is not practical to measure the relative dielectric constant of materials having high conductivity, such as copper. However, to the extent that the dielectric constant of copper exceeds that of free space, the time constant would be even shorter. Thus, we can conclude that free charge does not linger within a conductor. This makes sense physically. If there were a cloud of charge within a conductor, all of the individual charges would repel one another. Since the conductivity is high, they would move as far apart and as rapidly as possible, taking up positions on the surface when they could go no further. On the surface we could expect that they would distribute themselves in a manner that maximizes their average separation distances from one another.

Another way of estimating the quality of a conductor for sinusoidal currents is to write Maxwell's fourth equation in its phasor form

$$\nabla \times \vec{H} = \vec{J}_C + \frac{\partial \vec{D}}{\partial t} = \vec{E}(\sigma + j\omega\varepsilon) \tag{7.17-7}$$

In a good conductor the displacement current is negligible compared to the conduction current, and this requires that

$$\sigma \gg \omega\varepsilon \tag{7.17-8}$$

Again, since it is impractical to measure the dielectric constant of conductors, let us suppose that for most conductors ε_R is no greater than 10. Further suppose that we wish (7.17-8) to be satisfied by a factor of 100 before we will consider the material a "good conductor." Then a conductor is "good" up to a maximum frequency, f_{max}, given by

$$f_{max} \leq \frac{\sigma}{2\pi(100)(10)\varepsilon_0} \tag{7.17-9}$$

Applying this criterion to copper, having $\sigma = 5.8 \times 10^7$ ℧/m gives

$$f_{max} \leq \frac{5.8 \times 10^7}{2\pi(10^3)8.854 \times 10^{-12}} = 1 \times 10^{15} \text{ Hz}$$

$$\leq 1000 \text{ GHz} \quad \text{(for copper)} \tag{7.17-10}$$

Conductors satisfying (7.17-9) are called good conductors because for a given applied electric field or voltage the conduction current is much greater than the displacement current.

7.18 SKIN EFFECT

The phenomenon of *skin effect*, by which the current is crowded toward the surface of conductors at high frequencies, was introduced in Section 2.13. Its mathematical derivation [6, Sec. 6.04] provides another example of the application of Maxwell's equations, and the resultant conclusion is an insight into the current distribution *within a conductor*. Having shown that the displacement current is negligible compared to the conduction current in a good conductor, Maxwell's fourth equation can be written for the region within a conductor as

$$\nabla \times \vec{H} = \sigma\vec{E} \tag{7.18-1}$$

Taking the curl of both sides, and applying the identity (7.16-3),

$$\nabla \times \nabla \times \vec{H} = \nabla(\nabla \cdot \vec{H}) - \nabla^2\vec{H} = \sigma(\nabla \times \vec{E}) \qquad (7.18\text{-}2)$$

Recognizing that the divergence of $\vec{H}$ is zero and that the curl of $\vec{E}$ is related to $\vec{H}$ through Maxwell's third equation gives

$$\nabla^2\vec{H} = \sigma\mu\frac{\partial\vec{H}}{\partial t} \qquad (7.18\text{-}3)$$

That this same relationship applies to the $\vec{E}$ field can be shown by beginning with Maxwell's third equation, taking the curl of both sides and noting that the displacement current is negligible to get

$$\nabla \times \nabla \times \vec{E} = \nabla(\nabla \cdot \vec{E}) - \nabla^2\vec{E} = -\mu\frac{\partial}{\partial t}(\nabla \times \vec{H}) = -\mu\sigma\frac{\partial\vec{E}}{\partial t} \quad (7.18\text{-}4)$$

We showed in the previous section that free charge does not remain in a good conductor; hence the divergence of $\vec{E}$ is zero, and then (7.18-4) can be written as

$$\nabla^2\vec{E} = \mu\sigma\frac{\partial\vec{E}}{\partial t} \qquad (7.18\text{-}5)$$

Since $\vec{J} = \sigma\vec{E}$, it follows that

$$\nabla^2\vec{J} = \mu\sigma\frac{\partial\vec{J}}{\partial t} \qquad (7.18\text{-}6)$$

Equations (7.18-1) to (7.18-6) apply for all time variations; however, we are usually interested in the sinusoidal case. Then (7.18-3), (7.18-5), and (7.18-6) can be rewritten in phasor form, respectively, as

$$\nabla^2\vec{H} = j\omega\sigma\mu\vec{H} \qquad (7.18\text{-}7)$$

$$\nabla^2\vec{E} = j\omega\sigma\mu\vec{E} \qquad (7.18\text{-}8)$$

$$\nabla^2\vec{J} = j\omega\sigma\mu\vec{J} \qquad (7.18\text{-}9)$$

We find the solution to this equation format by solving for the current distribution in an infinitely thick conductor in the $x > 0$ direction, although it will be seen shortly that all of the interesting distribution of current occurs in a very thin layer near the surface of this conductor. Referring to Figure 7.18-1, suppose that there is a current sheet directed in the $+z$ direction and extending
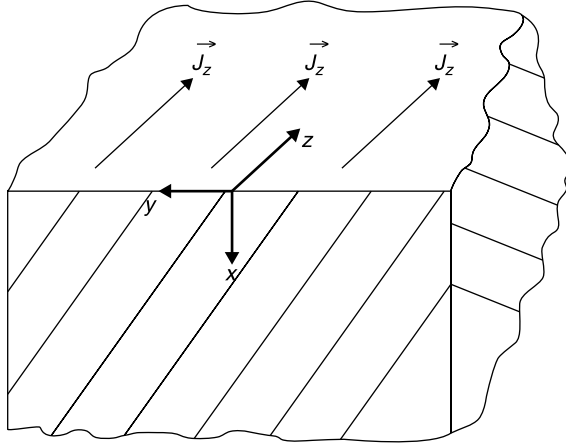
Figure 7.18-1 Infinitely thick conductor in the $x > 0$ direction having a uniform current sheet $\vec{J_z}$ in the yz plane moving in the $+z$ direction. (After Ramo and Whinnery, 6, p. 237, with permission.)

uniformly in the $+y$ and $-y$ directions. In other words, there is a uniform, infinite sheet of current in the yz plane that is moving in the $+z$ direction.

Since the current has no y or z variations by assumption, its Laplacian is simply

$$\frac{d^2 J_z}{dx^2} = j\omega\mu\sigma J_z = K^2 J_z \tag{7.18-10}$$

where J_z is the magnitude of $\vec{J}$ in the z direction and

$$K = \sqrt{j}\sqrt{\omega\mu\sigma} = \sqrt{j}\sqrt{2\pi f \mu\sigma} \tag{7.18-11}$$

Note that

$$\sqrt{j} = (1\angle 90°)^{1/2} = 1\angle 45° = \frac{1+j}{\sqrt{2}} \tag{7.18-12}$$

and therefore

$$K = (1+j)\sqrt{\pi f \mu\sigma} \quad (\text{meters}^{-1}) \tag{7.18-13}$$

Since the constant K has the dimensions of reciprocal meters, it is preferable to cast it in reciprocal form, resulting in

$$K = \frac{1+j}{\delta_S} \quad (\text{meters}^{-1}) \tag{7.18-14}$$

where δ_S is defined as the *skin depth* and given by

$$\delta_S = \frac{1}{\sqrt{\pi f \mu \sigma}} \quad \text{(meters)} \tag{7.18-15}$$

In general, the differential equation (7.18-10) has two independent exponential solutions, namely

$$\vec{J}_z = A e^{-Kx} + B e^{+Kx} \tag{7.18-16}$$

However, the positive exponential term cannot apply practically, else current density would increase to infinity with $+x$. Therefore B must equal zero for this situation. Also, A can be set equal to the current density, J_0, at the surface of the conductor at which $x = 0$, then

$$J_z = J_0 e^{-x/\delta_S} e^{-j(x/\delta_S)} \tag{7.18-17}$$

This result indicates that the magnitude of the current density falls off exponentially with distance from the surface, reaching $1/e$ of its surface value when $x = \delta_S$. For good conductors δ_S is an extremely short distance at RF and microwave frequencies. For example, when evaluated for copper at 1 GHz (2.13-4) δ_S is only 2.1 μm. At 25 μm/mil, this is less than 0.1 mil.

The term *skin depth* can be misleading, implying that all current flows within the skin depth. This, of course, is not true. In a conductor with finite conductivity, current penetrates to arbitrary depths. However, practically speaking, in a depth of only three times the skin depth, the current density is only 5% of its surface value.

7.19 CONDUCTOR INTERNAL IMPEDANCE

From the expression of 7.18-17 it can be seen that not only is there a fall off in current density from the conductor surface, there is also a lagging phase in the current relative to the surface value. Since the electric field and current are in phase at the surface, this means that the current density below the surface has an inductive reactance effect. It also means that the total current flow lags the applied electric surface field on the conductor, and that accordingly the conductor has an effective *internal impedance*. We can examine this effect by evaluating the total complex current J_z as follows [6, Sec. 6.06]. The conductor's internal impedance for a unit length and unit width is equal to the ratio of the voltage at the surface divided by the total current per unit width, I_W:

$$I_W = \int_0^\infty J_z \, dx = \int_0^\infty J_0 e^{-(1+j)(x/\delta_S)} \, dx = -J_0 \frac{\delta}{1+j} e^{-(1+j)(x/\delta_S)} \Big|_0^\infty$$

$$= J_0 \frac{\delta_S}{1+j} \tag{7.19-1}$$

The electric field at the surface is related to the current density by

$$E_z(x = 0) = \frac{J_0}{\sigma} \qquad (7.19\text{-}2)$$

and for a unit length of the conductor the voltage V_z has the same numeric value. The internal impedance of the conductor is then

$$Z_S = \frac{V_z(x = 0)}{I_W} = \frac{1}{\sigma \delta_S} + j \frac{1}{\sigma \delta_S} \qquad (7.19\text{-}3)$$

This impedance can be seen to have the form of a resistance in series with an inductance:

$$Z_s = R_s + j\omega L_i \qquad (7.19\text{-}4)$$

where

$$R_s = \frac{1}{\sigma \delta_S} \qquad (7.19\text{-}5)$$

and

$$\omega L_i = \frac{1}{\sigma \delta_S} = R_s \qquad (7.19\text{-}6)$$

The value Z_s is called the *internal impedance* of the conductor to distinguish it from impedances calculated from fields external to the conductors. For many high frequency applications, the internal impedance is neglected, as, for example, when computing the characteristic impedance of a transmission line wherein a lossless model is usually used for the transmission line conductors. However, a more exact model would include the internal impedance.

The results of (7.19-4), (7.19-5), and (7.19-6) indicate that:

1. The surface resistance per square (per unit length and unit width, where any unit may be used), R_s, of a conductor is that which would be found from the material's conductivity and *by assuming that all current flows uniformly in a depth of material equal to the skin depth*. This interesting fact may be what leads the inexperienced to believe incorrectly that no current flows below the skin depth.
2. The internal reactance of a conductor equals the surface resistance R_s at *all frequencies*.

It is interesting to compare the internal conductor impedance (Fig. 7.19-1) with the distributed reactance of, say, a 50-Ω coaxial cable. The inductance per
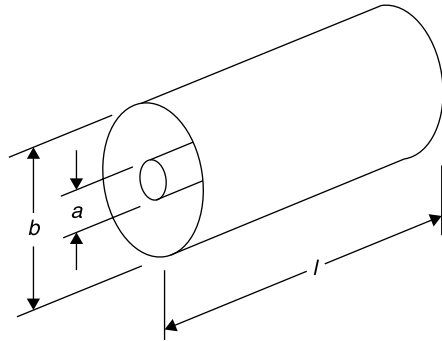
Figure 7.19-1   Coaxial transmission line used to estimate effect of conductor internal impedance.

unit length L of coaxial transmission line will be shown in Section 7.25 to be

$$L = \frac{\mu}{2\pi} \ln \frac{b}{a} = \frac{4\pi \times 10^{-7}}{2\pi} \ln\left(\frac{1 \text{ cm}}{0.43 \text{ cm}}\right) = 166 \text{ nH/m} \qquad (7.19\text{-}7)$$

where the b/a ratio of 2.3 for a 50-Ω air line has been used.

The external inductive reactance, based on the magnetic field between the conductors, at 1 GHz is then

$$\omega L = (6.28 \ \Omega/\text{GHz}/\text{nH})(1 \text{ GHz})166 \text{ nH} = 1042 \ \Omega/\text{m} \qquad (7.19\text{-}8)$$

From (7.19-5) the internal resistance of the line is

$$R_S = \frac{1}{\sigma \delta_S} = \frac{1}{(5.8 \times 10^7 \ \mho/\text{m})2 \times 10^{-6} \text{ m}} = 0.0086 \ \Omega/\text{square} \qquad (7.19\text{-}9)$$

The center conductor has a circumference of $\pi d = (0.43 \text{ cm})\pi = 1.37$ cm. We will consider that this is a flat surface, since its radius of curvature is very large compared to the skin depth. This is the side of a square having the resistance R_s. Then in 1 m, the center conductor has a distributed resistance of

$$R(\text{per meter}) = 0.0086(\Omega/\text{square}) \times \frac{100 \text{ cm}}{1.37 \text{ cm}} = 0.63 \ \Omega/\text{m} \qquad (7.19\text{-}10)$$

The outer conductor also has internal resistance, but it is less by the factor of $(1/2.3)$, or $0.27 \ \Omega/\text{m}$ because of its larger diameter. The total internal conductor resistance is then $0.9 \ \Omega/\text{m}$.

The internal inductive reactance is numerically equal to the internal resistance, hence it is equal to $0.9 \ \Omega/\text{m}$, and for most purposes this is negligibly small compared to the external reactance of $1042 \ \Omega/\text{m}$ given by (7.19-8). However, its presence results in the coaxial line having a longer electrical length than an equal length air path of a propagating plane wave (see Exercise 7.19-1).

In general, the average power loss per unit area of a conductor, P_{CL}, is equal to the product of the surface resistivity R_s and the square of the rms current per unit width, I_W. Thus,

$$P_{CL} = R_s |I_W|^2 \qquad (7.19\text{-}11)$$

This is a more general expression that can be used to find the conductor losses in cables, waveguides, and cavities.

7.20 THE WAVE EQUATION

To demonstrate that $\vec{E}$ and $\vec{H}$ fields can propagate in free space (more precisely in a charge-free, nonconducting region) we begin with Maxwell's third equation:

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \qquad (7.20\text{-}1)$$

To eliminate $\vec{B}$, take the curl of both sides, noting that $\vec{B} = \mu \vec{H}$ and that the curl and time derivative can be taken in any order:

$$\nabla \times \nabla \times \vec{E} = -\nabla \times \frac{\partial \vec{B}}{\partial t} = -\mu \frac{\partial (\nabla \times \vec{H})}{\partial t} \qquad (7.20\text{-}2)$$

Next, Maxwell's fourth equation applied to any region devoid of conduction currents is

$$\nabla \times \vec{H} = \frac{\partial \vec{D}}{\partial t} = \varepsilon \frac{\partial \vec{E}}{\partial t} \qquad (7.20\text{-}3)$$

Substituting this into (7.20-2),

$$\nabla \times \nabla \times \vec{E} + \mu \varepsilon \frac{\partial^2 \vec{E}}{\partial t^2} = 0 \qquad (7.20\text{-}4)$$

Next, apply the vector identity (7.16-3)

$$\nabla \times \nabla \times \vec{A} = \nabla(\nabla \cdot \vec{A}) - \nabla^2 \vec{A}$$

and since for any charge-free region, $\nabla \cdot \vec{E} = 0$, it follows that

$$\nabla^2 \vec{E} - \mu \varepsilon \frac{\partial^2 \vec{E}}{\partial t^2} = 0 \qquad (7.20\text{-}5)$$
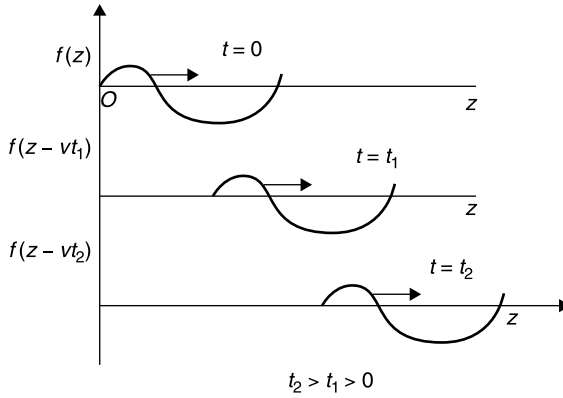
Figure 7.20-1   Electrical disturbance propagating in the $+z$ direction.

This is the *wave equation*. As expressed, it applies in three dimensions. In order to examine this result, let us consider the case for which $\vec{E}$ has only an $\vec{x}$-directed component and only a z variation. For this example (7.20-5) becomes

$$\frac{\partial^2 E_x}{\partial z^2} - \mu\varepsilon\frac{\partial^2 E_x}{\partial t^2} = 0 \tag{7.20-6}$$

This equation describes a propagating wave. It is satisfied by any function of the form [6, p. 273, also 10, p. 26]

$$E_x^+ = f(z - vt) \tag{7.20-7}$$

which describes a disturbance propagating in the $+z$ direction (Fig. 7.20-1), since at the velocity v the quantity $z - vt$ is a constant when z and t both increase together, resulting in a constant value of E_x^+ moving at velocity v in the $+z$ direction with increasing t.

Since (7.20-6) is a second-order differential equation, it has two independent solutions. The second solution is a function $f(z + vt)$ that propagates in the $-z$ direction:

$$E_x^- = f(z + vt) \tag{7.20-8}$$

With either solution, when the indicated differentiations are applied to (7.20-6), the result is

$$f'' + v^2\mu\varepsilon f'' = 0 \tag{7.20-9}$$

which requires that the velocity of propagation is

$$v = \frac{1}{\sqrt{\mu\varepsilon}} = \frac{1}{\sqrt{\mu_r\varepsilon_r}\sqrt{\mu_0\varepsilon_0}} \tag{7.20-10}$$

For free space, μ_r and ε_r are unity. Therefore

$$v = c = \frac{1}{\sqrt{4\pi \times 10^{-7}\frac{\text{H}}{\text{m}}\frac{1}{36\pi} \times 10^{-9}\frac{\text{F}}{\text{m}}}} \approx 3 \times 10^8 \frac{\text{m}}{\text{s}}$$

One can imagine Maxwell's excitement on concluding that electromagnetic waves propagated at 300,000 km/s, which he understood to be the approximate speed of light. This probably contributed to his surmise *"we have strong reason to conclude that light itself is an electromagnetic disturbance in the form of waves." (James Clerk Maxwell circa 1863)* [1, p. 4].

Using the same approach, the electric field can be eliminated from Maxwell's third and fourth equations with the result that the $\vec{\text{H}}$ field satisfies the same format as (7.20-5), namely

$$\nabla^2\vec{\text{H}} - \mu\varepsilon\frac{\partial^2\vec{\text{H}}}{\partial t^2} = 0 \qquad (7.20\text{-}11)$$

Rather than re-solve the wave equation for the $\vec{\text{H}}$ field, it is usually more convenient to solve for either the $\vec{\text{E}}$ or $\vec{\text{H}}$ field and find the other using the appropriate Maxwell third or fourth curl equation.

7.21   THE HELMHOLTZ EQUATIONS

Thus far, the wave equation expressed in terms of $\vec{\text{E}}$ (7.20-5) or $\vec{\text{H}}$ (7.20-11) applies to any time-varying "electrical disturbance," as the generalized waveform sketch of Fig. 7.20-1 describes. However, usually we are interested in sinusoidally varying signals having the implicit time variation $e^{j\omega t}$. Since each differentiation with respect to t produces a factor $j\omega$, the phasor form of the wave equation can be written immediately as

$$\nabla^2\vec{E} + \omega^2\mu\varepsilon\vec{E} = 0 \qquad (7.21\text{-}1)$$

or

$$\nabla^2\vec{E} + k^2\vec{E} = 0 \qquad (7.21\text{-}2)$$

and

$$\nabla^2\vec{H} + k^2\vec{H} = 0 \qquad (7.21\text{-}3)$$

where

$$k \equiv \omega\sqrt{\mu\varepsilon} = \frac{2\pi f}{v} = \frac{2\pi}{\lambda} \qquad (7.21\text{-}4)$$

$$v = \frac{1}{\sqrt{\mu_R\mu_0\varepsilon_R\varepsilon_0}} = \frac{c}{\sqrt{\mu_r\varepsilon_r}} \qquad (7.21\text{-}5)$$

Equations (7.21-2) and (7.21-3) are called the *Helmholtz equations*. Note that for uniform plane waves in an unbounded propagating medium k is the *propagation constant*, v is the *velocity of propagation*, and λ is the *wavelength*. For free space propagation, in which $\mu = \mu_0$ and $\varepsilon = \varepsilon_0$, the propagation constant, propagation velocity, and wavelength are denoted k_0, c, and λ_0, respectively. The Helmholtz equations are more convenient to use when the fields have sinusoidal variation.

7.22   PLANE PROPAGATING WAVES

Returning to waves having arbitrary time dependence, suppose we have plane waves that are propagating in the $+z$ and/or $-z$ directions, but have no variations in the x or y directions (Fig. 7.22-1). Then $\partial/\partial x = \partial/\partial y = 0$; and applying Maxwell's third equation produces

$$\nabla \times \vec{E} = -\mu \frac{\partial \vec{H}}{\partial t} = \begin{vmatrix} \vec{x} & \vec{y} & \vec{z} \\ 0 & 0 & \dfrac{\partial}{\partial z} \\ E_x & E_y & E_z \end{vmatrix} \tag{7.22-1}$$

Note that for this plane wave case, *although the $\vec{E}$ and $\vec{H}$ fields have no variation with respect to x or y, there can be* E_x, E_y, H_x, and H_y *field components*. Equating the separate vector components on each side of (7.22-1),
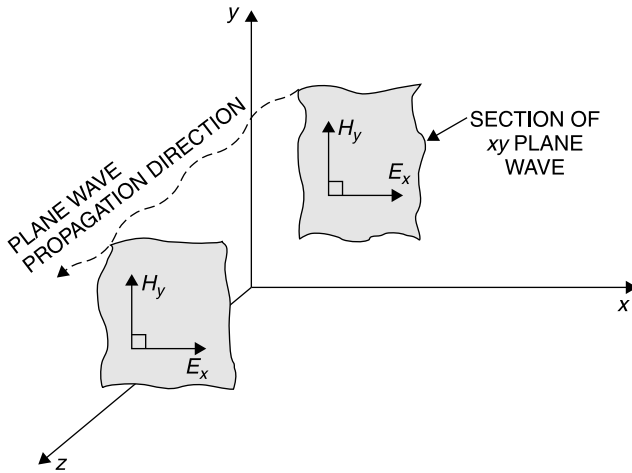
Figure 7.22-1   Plane wave having $\vec{E}$ and $\vec{H}$ fields only in the xy plane and propagation in the $+z$ direction.

$$-\frac{\partial E_y}{\partial z} = -\mu\frac{\partial H_x}{\partial t} \qquad (7.22\text{-}2)$$

$$\frac{\partial E_x}{\partial z} = -\mu\frac{\partial H_y}{\partial t} \qquad (7.22\text{-}3)$$

$$0 = -\mu\frac{\partial H_z}{\partial t} \qquad (7.22\text{-}4)$$

Similarly, from Maxwell's fourth equation

$$\nabla \times H = \varepsilon\frac{\partial E}{\partial t} = \begin{vmatrix} \vec{x} & \vec{y} & \vec{z} \\ 0 & 0 & \dfrac{\partial}{\partial z} \\ H_x & H_y & H_z \end{vmatrix} \qquad (7.22\text{-}5)$$

Again, equating vector components on each side of the equation,

$$-\frac{\partial H_y}{\partial z} = \varepsilon\frac{\partial E_x}{\partial t} \qquad (7.22\text{-}6)$$

$$\frac{\partial H_x}{\partial z} = \varepsilon\frac{\partial E_y}{\partial t} \qquad (7.22\text{-}7)$$

$$0 = \varepsilon\frac{\partial E_z}{\partial t} \qquad (7.22\text{-}8)$$

We first notice, from (7.22-4) and (7.22-8), that the z components of $\vec{E}$ and $\vec{H}$ must be zero, except for a possible time invariant term, but that is of no interest in the case of propagation. Therefore, the time-varying electric and magnetic fields are transverse to the directions ($+z$ and $-z$) of propagation; and the plane wave is said to be a *transverse electromagnetic (TEM) wave*.

We also notice that differentiating (7.22-3) with respect to z and (7.22-6) with respect to t, and substituting the value for $\partial^2 H_y/\partial t\,\partial z$ into (7.22-3) yields the one-dimensional wave equation previously derived:

$$\frac{\partial^2 E_x}{\partial z^2} - \mu\varepsilon\frac{\partial^2 E_x}{\partial t^2} = 0 \qquad (7.22\text{-}9)$$

We denote the fields that propagate in the $+z$ direction as H^+ and E^+. Specifically, we can relate H_y^+ and E_x^+ to each other [6, p. 278] using (7.22-3). From (7.22-3) we conclude that H_y^+ and E_x^+ have the same functional relationship and differ at most by a constant factor (as well as integration constants which do not vary with z or t, and can be ignored for propagating waves).

We have noted that $E_x^+ = f(z - vt)$. This can be recast as $E_x^+ = g(t - z/v)$. Then on differentiating E_x^+ with respect to z and noting that $v = 1/\sqrt{\mu\varepsilon}$,

(7.22-3) can be written

$$\frac{\partial H_y^+}{\partial t} = -\frac{1}{\mu}\frac{\partial E_x^+}{\partial z} = \frac{1}{\mu}\frac{1}{v}g'\left(t - \frac{z}{v}\right) = \sqrt{\frac{\varepsilon}{\mu}}g'\left(t - \frac{z}{v}\right) \qquad (7.22\text{-}10)$$

Since H_y and E_x have the same functional form, we can integrate (7.22-10) to get

$$H_y^+ = \sqrt{\frac{\mu}{\varepsilon}}E_x^+ \qquad (7.22\text{-}11)$$

$$E_x^+ = \eta H_y^+ \qquad (7.22\text{-}12)$$

where

$$\eta = \sqrt{\frac{\mu}{\varepsilon}} = \sqrt{\frac{\mu_0\mu_r}{\varepsilon_0\varepsilon_r}} \qquad (7.22\text{-}13)$$

Since $\vec{E}$ is in volts/meter and $\vec{H}$ is expressible in amperes/meter, η has the dimension of *ohms*. The factor η occurs frequently as a proportionality constant between $\vec{E}$ and $\vec{H}$ propagating fields and is called the *intrinsic impedance* of the propagating medium. Its reciprocal $1/\eta$ is called the *intrinsic admittance*. In free space

$$\eta_0 = \sqrt{\frac{\mu_0}{\varepsilon_0}} \approx \sqrt{\frac{4\pi \times 10^{-7} \text{ H/m}}{\dfrac{10^{-9} \text{ F/m}}{36\pi}}} \approx 120\pi \ \Omega \approx 377 \ \Omega \qquad (7.22\text{-}14)$$

A plane wave propagating in either the $+z$ or $-z$ direction can have an E_x component, an E_y component, or both. We determined the relationship of (7.22-14) by beginning with (7.22-3). Using similar reasoning and beginning with (7.22-2) yields [6, p. 279]

$$E_y^+ = -\eta H_x^+ \qquad (7.22\text{-}15)$$

Thus, for plane waves traveling in the $+z$ direction

$$\frac{E_x^+}{H_y^+} = -\frac{E_y^+}{H_x^+} = \eta \qquad (7.22\text{-}16)$$

Similar reasoning for waves propagating in the $-z$ direction shows that

$$\frac{E_x^-}{H_y^-} = -\frac{E_y^-}{H_x^-} = -\eta \qquad (7.22\text{-}17)$$

The relations (7.22-16) and (7.22-17) reveal that

1. $\vec{E}$ *and* $\vec{H}$ *are orthogonal to one another in plane propagating waves.*
2. *The instantaneous value of* $\vec{E}$ *is always* η *times the instantaneous value of* $\vec{H}$ *for each orthogonal pair of wave components (when propagating in unbounded media).*
3. *The cross product* $\vec{E} \times \vec{H}$ *always points in the direction of propagation.* Later we will formalize this result as Poynting's theorem.
4. *The relationships between the transverse* $\vec{E}$ *and* $\vec{H}$ *fields of the TEM plane wave apply for any electrical disturbance, regardless of its functional relationship with time.*

For example, the electric field might be an electromagnetic pulse created by a nuclear detonation. Applying a plane wave approximation to the resulting wave when it is sufficiently far from the source of the disturbance, the $\vec{E}$ and $\vec{H}$ fields propagate with their amplitudes related by the intrinsic impedance.

Thus far, this treatment of plane waves applies for any time variation. To determine the solution for $\vec{E}$ and $\vec{H}$ in terms of propagation in the $+z$ or $-z$ directions, it would be necessary to substitute the actual form of the time variation into the wave equations and solve for the fields. However, usually we are interested in a sinusoidal time variation, in which case the Helmholtz equations can be applied. Thus the fields for the plane wave case can be written in vector phasor form, namely

$$\vec{E} = \vec{E}_T^+ e^{-jkz} + \vec{E}_T^- e^{+jkz} \qquad (7.22\text{-}18)$$

$$\vec{H} = \vec{H}_T^+ e^{-jkz} + \vec{H}_T^- e^{+jkz} \qquad (7.22\text{-}19)$$

where

$$H_T^+ = \frac{E_T^+}{\eta} \quad \text{and} \quad H_T^- = -\frac{E_T^-}{\eta}$$

where E_T and H_T are the description of the electric and magnetic fields, respectively, in the plane transverse to the direction of propagation. The terms in (7.22-18) and (7.22-19) with superscript "+" represent waves propagating in the $+z$ direction, and those with the superscript "−" represent waves propagating in the $-z$ direction.

7.23 POYNTING'S THEOREM

We now return to the general time variation case to derive Poynting's theorem. As the $\vec{E}$ and $\vec{H}$ fields propagate through space, they carry power. Actually, it
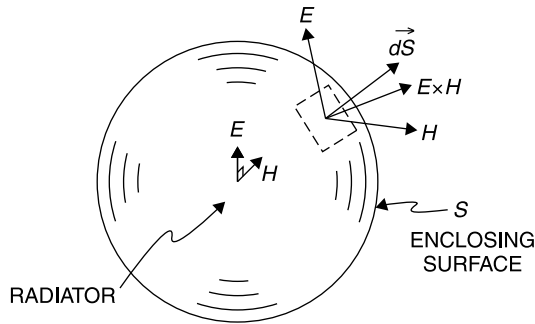
Figure 7.23-1   Poynting's vector represents the power density/unit area radiated out of a closed surface containing a source of propagation.

is impossible to launch a perfectly plane wave because this would require a field that was uniform in, say, the xy plane and extending infinitely far in the $+x, -x, +y$, and $-y$ directions. However, a *directional antenna approximates a plane wave over some aperture*, the larger the aperture dimensions compared to the transmitted wavelength, the more directive (plane-wave-like) the beam and the more gradually it diverges.

Although a practically launched wave that is propagating far from its source can be approximated as a plane wave for the purpose of relating the $\vec{E}$ and $\vec{H}$ field components, its wave front actually tends to assume a spherical shape, spreading out so that the amount of power crossing a given area, say, one square meter, becomes less and less with increasing distance from the transmitter. The power density, in watts per square meter, approaches an inverse square law function of distance when measured far from the source that launched the wave (Fig. 7.23-1).

We shall demonstrate that the amount of power passing through a given surface area S of space is given by *Poynting's vector*. This is the cross product of the $\vec{E}$ and $\vec{H}$ fields. The cross product is the magnitudes of the fields multiplied by the sine of the angle between them. But for a wave propagating in space, the $\vec{E}$ and $\vec{H}$ fields are already orthogonal to one another, so their power cross product is simply the product of their amplitudes [6, p. 275].

In this analysis the $\vec{E}$ and $\vec{H}$ fields are their actual instantaneous values, and the power that they convey is the instantaneous power flow regardless of their functional time variation. Later, when applying Poynting's theorem to sinusoidal fields, rms values can be used to determine average power flow.

Poynting's theorem states that the power flow propagating out of a closed surface S is equal to the integral over the surface of the real part of Poynting's vector, $\vec{P} = \vec{E} \times \vec{H}$, where P, E, and H are time-varying functions:

$$\text{Power flow}|_S = \int_S \vec{P} \cdot \vec{dS} \qquad (7.23\text{-}1)$$

To prove this theorem, we employ the vector identity (7.16-9) into which we have substituted the $\vec{E}$ and $\vec{H}$ fields:

$$\nabla \cdot (\vec{E} \times \vec{H}) = \vec{H} \cdot \nabla \times \vec{E} - \vec{E} \cdot \nabla \times \vec{H} \qquad (7.23\text{-}2)$$

Next, we substitute values for $\nabla \times \vec{E}$ and $\nabla \times \vec{H}$ from Maxwell's third and fourth equations:

$$\nabla \cdot (\vec{E} \times \vec{H}) = -\vec{H} \cdot \frac{\partial \vec{B}}{\partial t} - \vec{E} \cdot \frac{\partial \vec{D}}{\partial t} - \vec{E} \cdot \vec{J}_c \qquad (7.23\text{-}3)$$

If ε and μ are constant over the volume of interest,

$$\vec{E} \cdot \frac{\partial \vec{D}}{\partial t} = \frac{1}{2} \frac{\partial(\varepsilon \vec{E}^2)}{\partial t} \qquad (7.23\text{-}4)$$

and

$$\vec{H} \cdot \frac{\partial \vec{B}}{\partial t} = \frac{1}{2} \frac{\partial(\mu \vec{H}^2)}{\partial t} \qquad (7.23\text{-}5)$$

Substituting these values into (7.23-3) and integrating over the volume contained by the closed surface S,

$$-\int_V \nabla \cdot (\vec{E} \times \vec{H}) \, dV = \int_V \left[\frac{1}{2} \cdot \frac{\partial(\mu \vec{H}^2)}{\partial t} + \frac{1}{2} \cdot \frac{\partial(\varepsilon \vec{E}^2)}{\partial t} + \vec{E} \cdot \vec{J}_c \right] dV \quad (7.23\text{-}6)$$

Applying the divergence theorem,

$$-\int_S (\vec{E} \times \vec{H}) \cdot \vec{dS} = \int_V \left[\frac{1}{2} \cdot \frac{\partial(\mu \vec{H}^2)}{\partial t} + \frac{1}{2} \cdot \frac{\partial(\varepsilon \vec{E}^2)}{\partial t} + \vec{E} \cdot \vec{J}_c \right] dV \quad (7.23\text{-}7)$$

The negative sign on the left side means that instead of diverging from the volume V, the energy is *entering* V. This is consistent with the reasoning that follows.

The first term on the right side has the dimensions of energy per unit volume, henrys $\times$ amperes^2/m^3. When integrated with respect to volume, this is the same form, even to the factor of $\frac{1}{2}$, of the energy stored in an inductor, specifically $LI^2/2$. As such, this term represents, after integration, the increase in stored magnetic energy per unit time (power inflow) in volume V resulting from the inflow (negative divergence) of Poynting's vector. Similarly, the second term having the dimensions farads $\times$ voltage^2/m^3 also has the dimensions of energy and is equivalent to the energy storage of a capacitor, $CV^2/2$, having unit volume. This term represents the increase in stored electric energy per unit time (power inflow).

The third term has the dimensions of power dissipated voltage $\times$ current/m^3 and represents, after integration, the amount of power being dissipated within the volume V. Alternatively, if there are mobile charges in V, this term represents the power which increases their motion or, if $\vec{E} \cdot \vec{J}$ is negative, it represents the power contributed by their deceleration [6, p. 275].

All of the net power inflow and dissipation represented on the right side of (7.23-7) comes from the inflow of Poynting's vector. Reversing the sign of this power flow gives the rate at which *energy is leaving* V, that is the power flow out of V.

$$\text{Power flow}|_S = \int_S \vec{P} \cdot \vec{dS} \quad \text{where } \vec{P} = \vec{E} \times \vec{H} \qquad (7.23\text{-}8)$$

This is *Poynting's theorem*—the result that we wished to show. Note that P is a power density (watts/meter^2).

Poynting's theorem applies for any time-varying $\vec{E}$ and $\vec{H}$, regardless of the form of time variation (it need not be sinusoidal). *Poynting's theorem gives the instantaneous power flow when instantaneous values are used for the $\vec{E}$ and $\vec{H}$ fields.*

For sinusoidally varying fields, the average power flow density can be determined using the phasor forms for $\vec{E}$ and $\vec{H}$, specifically

$$\vec{P}_{\text{Av}} = \tfrac{1}{2} \operatorname{Re}(\vec{E} \times \vec{H}^*) \qquad (7.23\text{-}9)$$

where * represents the complex conjugate. The imaginary part of $(\vec{E} \times \vec{H}^*)$ represents the ebb and flow of reactive power density through the surface S.

7.24   WAVE POLARIZATION

The electromagnetic wave propagation described in Section 7.22 (Fig. 7.22-3) is known as a linear or plane-polarized wave, since the $\vec{E}$ and $\vec{H}$ fields each lie along an unchanging direction in the plane transverse to the direction of propagation. For example, an electric field oriented in the x direction and propagating in the +z direction has a full period of variation in one wavelength at a given instant of time (Fig. 7.24-1).

The orthogonal magnetic field, in this example in the y direction, has the same pattern along the z axis (Fig. 7.24-2).

The combination of any two or more linearly polarized waves of the same frequency that are in phase and propagating in the same direction results in a linearly polarized wave. Figure 7.24-3 depicts the case of two orthogonal waves, represented by $\vec{E}_x$ and $\vec{E}_y$, that are in phase and of equal amplitude. The resultant wave, $\vec{E}_R$, is oriented at 45° to both $\vec{E}_x$ and $\vec{E}_y$.
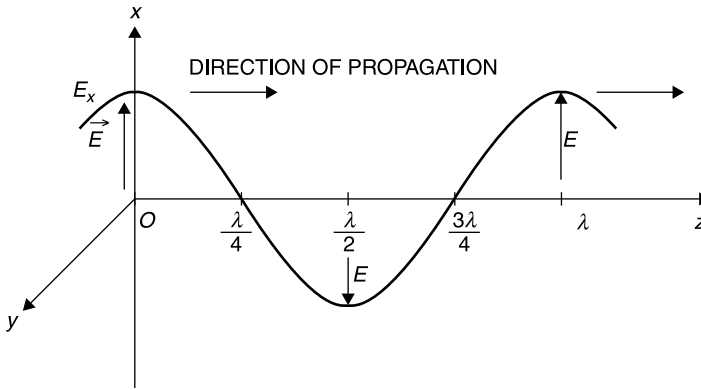
Figure 7.24-1   Amplitude of E_x at an instant in time for a wave propagating in the $+z$ direction. A sinusoid is shown for illustration, but the wave equation applies for all time variations [1, p. 32].

Thus, if

$$\vec{E}_x = \vec{x}E_0 \cos(\omega t - kz) \quad \text{and} \quad \vec{E}_y = \vec{y}E_0 \cos(\omega t - kz) \qquad (7.24\text{-}1)$$

where $k = 2\pi/\lambda$ is the propagation constant of waves traveling in the $+z$ direction. Then, defining $\vec{E}_R \equiv \vec{E}_x + \vec{E}_y$, the resultant electric field is given by

$$\vec{E}_R(\vec{x} + \vec{y})E_0 \cos(\omega t - kz) \qquad (7.24\text{-}2)$$

which is simply a vector having the peak magnitude $\sqrt{2}E_0$ and a *direction that is always at an angle of 45° from the x or y axes* (Fig. 7.24-4).

Suppose that instead of two in-phase field components, we excite two space orthogonal components of electric field that have equal amplitudes but are *90° out of phase* with each other.
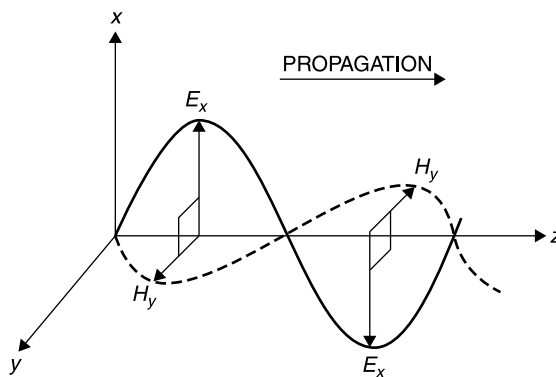
Figure 7.24-2   Orthogonal E and H fields of the wave in Figure 7.24-1 [1, p. 32].
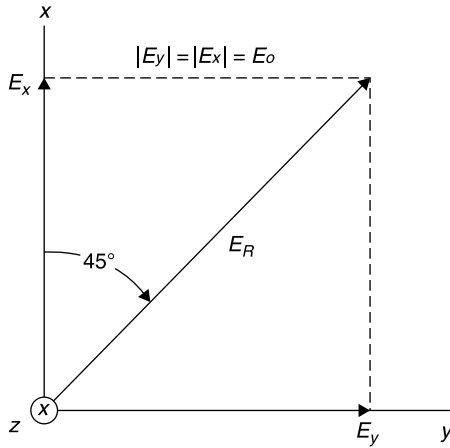
Figure 7.24-3 Sum of two in-phase, linearly polarized waves is also a linearly polarized wave.

In this case

$$\vec{E}_x = \vec{x} E_0 \cos(\omega t - kz) \quad \text{and} \quad \vec{E}_y = \vec{y} E_0 \sin(\omega t - kz) \qquad (7.24\text{-}3)$$

At $z = 0$ for instance,

$$\vec{E}_R = \vec{x} E_0 \cos \omega t + \vec{y} E_0 \sin \omega t \qquad (7.24\text{-}4)$$
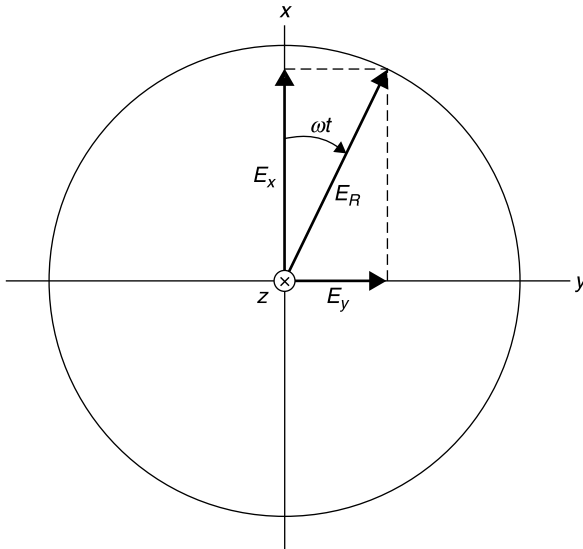
Figure 7.24-4 Circularly polarized plane wave has a *constant amplitude* when projected onto the plane normal to its direction of propagation.
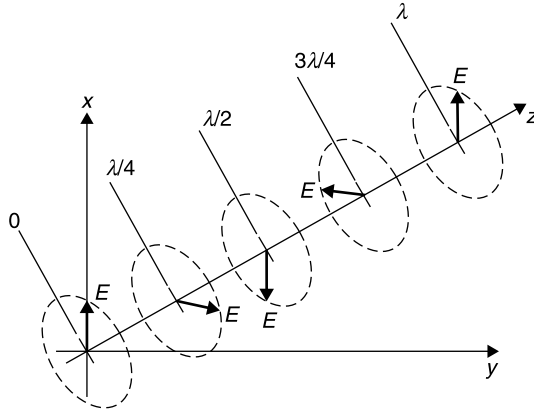
Figure 7.24-5 Left-hand circularly polarized (LHCP) wave propagating in the $+z$ direction. This diagram depicts the steady-state wave amplitudes at quarter wavelength intervals in space at an instant in time $(t = 0)$ [1, p. 37].

Since $\cos^2 \omega t + \sin^2 \omega t = 1$, the resultant field, $\vec{E}_R$, has a constant magnitude, E_0, and rotates in the xy plane at an ωt rate. The sense of rotation can be deduced by considering (7.24-4) and Figure 7.24-4. When $\omega t = 0$, $\vec{E}_R$ is oriented in the $+x$ direction. A quarter period later, when $\omega t = \pi/2$, $\vec{E}_R$ is oriented in the $+y$ direction. Thus, for the relative phases of $\vec{E}_x$ and $\vec{E}_y$ in (7.24-3) the rotation shown in Figure 7.24-4 is clockwise. Reversing the 90° phase difference between $\vec{E}_x$ and $\vec{E}_y$ changes the direction of rotation. Thus, for counterclockwise rotation at $z = 0$,

$$\vec{E}_R = \vec{x}E_0 \cos \omega t - \vec{y}E_0 \sin \omega t \tag{7.24-5}$$

If the wave rotates clockwise in a fixed transverse plane when viewed so that it propagates away from you, the polarization is called right-hand circular polarization (RHCP). That is, the advance is the same as that of a right-hand screw.

If it rotates counterclockwise while propagating away from you, it is called left-hand circular polarization (LHCP). Figure 7.24-5 depicts a LHCP wave propagating in the $+z$ direction. Note that for the wave in this figure, *for any fixed transverse plane*, the constant amplitude electric field rotates counterclockwise when viewed along the propagation direction (see Exercise 7.24-3).

If the magnitudes of E_x and E_y are not equal and/or their phase difference is not precisely 90°, then the polarization is called *elliptical polarization* (right or left hand) because the tip of the resultant vector E_R traces an elliptical pattern in the xy plane, rather than a perfect circle.

Circular polarization (CP) transmission is useful in communicating with a linearly polarized receiving antenna (such as a monopole or dipole). The signal will be received no matter how the receiving antenna is tilted. However, a linearly polarized antenna will intercept only one half of the available CP power.

All of the power can be recovered only by using a CP antenna of the same sense as is received. Since the directional sense of propagation is reversed on viewing a wave leaving the transmitter and the same wave arriving at the receiver, if a RHCP antenna was used for transmission, a LHCP antenna must be used for reception.

7.25   EH FIELDS ON TRANSMISSION LINES

Once an electromagnetic wave is launched into space, it diverges and its field strength diminishes due to the spreading out of its power density. This effect can be diminished by using a guiding structure which sets up boundary conditions that reduce the wave's spreading, such as *two conductors* forming a *transmission line*. Because it reduces a propagating wave's spreading, a two-wire transmission line could be considered a form of waveguide. However, the term *waveguide* is generally used to describe hollow pipes or solid dielectrics used for electromagnetic transmission. These terms are not used consistently in the industry, and there are exceptions, such as *coplanar waveguide*, a variation of stripline or microstrip.

For the two-wire line (Fig. 7.25-1) the fields and currents are considerably constrained to travel axially along the line. However, some energy is lost due to
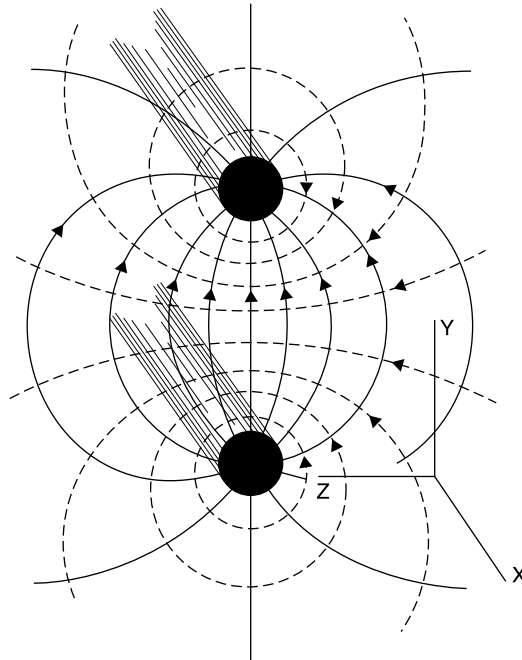
Figure 7.25-1   A two-wire transmission line and its $\vec{E}$ (solid lines) and $\vec{H}$ (dashed lines) fields. (After Skilling, 4, p. 160, with permission.)
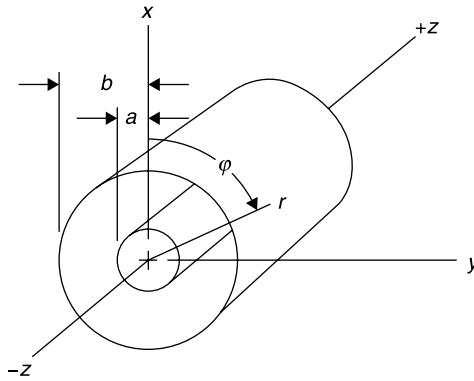
Figure 7.25-2 Coaxial transmission line with cylindrical coordinates r, φ, and z. Note that a is the radius of the inner conductor and b is the radius of the outer conductor.

radiation or leakage of the fields surrounding the line. The principal mode is TEM (transverse electromagnetic), which means that both the electric and magnetic fields are always and everywhere transverse to the direction of propagation. Note that the $\vec{E}$ field lines terminate on the conductors, and the $\vec{H}$ field lines are orthogonal to the $\vec{E}$ lines, as shown in Figure 7.25-2. Both fields fringe outward without limit in the plane transverse to the direction of propagation.

The radiation that leaks from the two-wire line can be reduced to negligible values by causing one conductor to surround the other, as is done with the *coaxial transmission line*. The shielding would be total if the outer conductor could have zero resistivity; but, even with finite resistivity, for all practical purposes the shielding is complete when the outer conductor is a solid metal cylinder of any practical thickness. For analysis we will assume perfect conductors, that is, having infinite conductivity. The following analysis determines the $\vec{E}$ and $\vec{H}$ fields in the insulating region of the coaxial transmission line (between the conductors). From this can be derived expressions for the distributed capacitance C and inductance L of the line as well as its characteristic impedance Z_0.

For this transmission line, as for most lines, there are different field patterns by which energy can be propagated. Each is called a *mode* of the transmission line or waveguide. In general, the mode which can propagate at the lowest frequency is called the *dominant mode*. Transmission lines having two separate conductors can propagate all frequencies, theoretically from direct current to light frequencies in the lowest order, or dominant, mode. However, at some frequency, propagation in more than one mode becomes possible, and this is undesirable. Therefore, it is good practice to operate the transmission line at frequencies below the lowest frequency at which the first higher-order mode above the dominant mode can propagate. This is known as the *cutoff frequency of the mode*.

Thus far, we have presented fields and field relationships in rectangular coordinates. However, various symmetries favor the use of other coordinate systems. The most common orthogonal systems are rectangular (also called Cartesian), cylindrical, and spherical. The coaxial transmission line is most easily analyzed using cylindrical coordinates.

To evaluate the fields of the dominant TEM mode we note that by definition both the $\vec{E}$ and $\vec{H}$ fields have components only in the plane transverse to the propagation directions, $+z$ and $-z$. This means that their z-directed components are zero. Since this mode "propagates" at DC (zero frequency), it is possible to evaluate the fields using a static analysis. To solve for the electric field [1, p. 21], apply Gauss's law assuming that there is a charge, $+Q$, on the center conductor and an equal and opposite charge, $-Q$, on the outer conductor of a coaxial line of unit length. Since this is a per-unit-length analysis, we may ignore any fringing fields at the ends of the portion of the line used for analysis.

For any cylindrical surface of radius r defined between the conductors, $a < r < b$,

$$Q = \int_0^L \int_0^{2\pi} \varepsilon |E| r \, d\varphi \, dz \tag{7.25-1}$$

where L is unity, since the analysis is on a per-unit length of transmission line basis. Noting that $\vec{E}$ is independent of z and, due to symmetry, independent of φ, the integration produces

$$Q = \varepsilon E_r(2\pi r) \quad \text{or} \quad E_r = \frac{Q}{2\pi\varepsilon r} \tag{7.25-2}$$

which is the magnitude of the radial $\vec{E}$ field between the conductors (Fig. 7.25-3). The $\vec{E}$ field is the gradient of a scalar function that we will identify as the voltage V having zero potential (voltage) on the outer conductor.
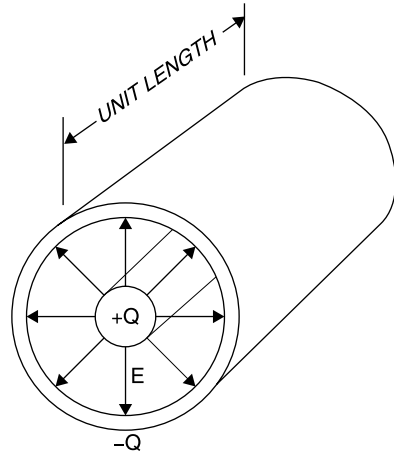
Figure 7.25-3   The $\vec{E}$ field of the TEM mode in coaxial line.

The gradient in cylindrical coordinates is

$$\nabla V = \vec{r}\frac{\partial V}{\partial r} + \vec{\phi}\frac{1}{r}\frac{\partial V}{\partial \phi} + \vec{z}\frac{\partial V}{\partial z} \tag{7.25-3}$$

where $\vec{r}, \vec{\phi}$, and $\vec{z}$ are unit vectors in the r, ϕ, and z directions. Since the electric field and with it the voltage V vary only in the r direction, only the first term of (7.23-3) is nonzero; thus

$$\vec{E}_r = -\nabla V = -\vec{r}\frac{\partial V}{\partial r}$$

$$V(r) = -\int \vec{E}\cdot\vec{r}\,dr = \frac{-Q}{2\pi\varepsilon}\int\frac{dr}{r} = \frac{-Q}{2\pi\varepsilon}\ln r + K \tag{7.25-4}$$

If we define the value of the voltage to be V_0 on the inner conductor and zero on the outer conductor, the imposition of these boundary conditions requires that

$$K = \frac{Q}{2\pi\varepsilon}\ln b$$

$$V_0 = \frac{Q}{2\pi\varepsilon}\ln\left(\frac{b}{a}\right) \tag{7.25-5}$$

$$\frac{Q}{2\pi\varepsilon} = \frac{V_0}{\ln(b/a)} \tag{7.25-6}$$

Since the distributed capacitance C of the coaxial transmission line is Q/V_0, an expression for C is obtained from (7.25-6) as

$$C = \frac{2\pi\varepsilon}{\ln(b/a)} \tag{7.25-7}$$

Substituting the value for K and the value for $Q/2\pi\varepsilon$ into (7.25-4), the expression for the transverse variation of voltage is

$$V(r) = \frac{V_0}{\ln(b/a)}[-\ln(r) + \ln(b)] = \frac{V_0\ln(b/r)}{\ln(b/a)}$$

$$V(r) = \frac{V_0\ln(r/b)}{\ln(a/b)} \tag{7.25-8}$$

Then E_r can be obtained by differentiating $V(r)$ with respect to r:

$$E_r = -\frac{\partial V}{\partial r} = \frac{-V_0}{\ln(a/b)}\frac{1}{r} = \frac{V_0}{r\ln(b/a)} \tag{7.25-9}$$
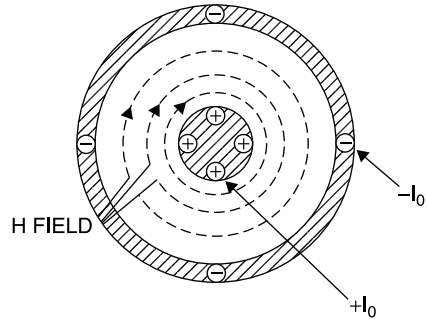
Figure 7.25-4 Magnetic field lines H and the current distribution I_0 for the TEM mode in coaxial line.

where the negative sign in (7.25-9) was eliminated by interchanging the numerator and denominator of the ln expression. Up to this point the expressions apply for any time variation. Usually coaxial lines are used to transmit sinusoidal signals, in which case $\vec{E}$ must also satisfy the Helmholtz equation, and therefore have a variation with z of the form e^{-jkz}. The phasor form of E_r becomes

$$E_r = \frac{V_0}{r \ln(b/a)} e^{-jkz} \tag{7.25-10}$$

Next, we will obtain an expression for the magnetic field between the conductors. The magnetic field lines are concentric with the center conductor as shown in Figure 7.25-4. Its value is determined by using the integral form of Maxwell's fourth equation:

$$\oint \vec{H} \cdot \overrightarrow{dl} = \int_S \vec{J} \cdot \overrightarrow{dS} + \int_S \frac{\partial \vec{D}}{\partial t} \cdot \overrightarrow{dS}$$

The axial $\partial D/\partial t$ term is zero since E and D have no z component. The integral of the current density is just the total current I_0 that flows in the $+z$ direction on the inner conductor and the $-z$ direction on the outer conductor (as is required for the $\vec{H}$ field to be zero outside the coaxial line). Then, since the line integral of $\vec{H}$ is just its magnitude times the circumferential path, the magnetic field for $a \leq r \leq b$ is obtained from

$$2\pi r H_\varphi = I_0$$

Thus

$$H_\varphi = \frac{I_0}{2\pi r} \tag{7.25-11}$$

Equation (7.25-11) is valid for both DC and AC fields. For sinusoidal varia-
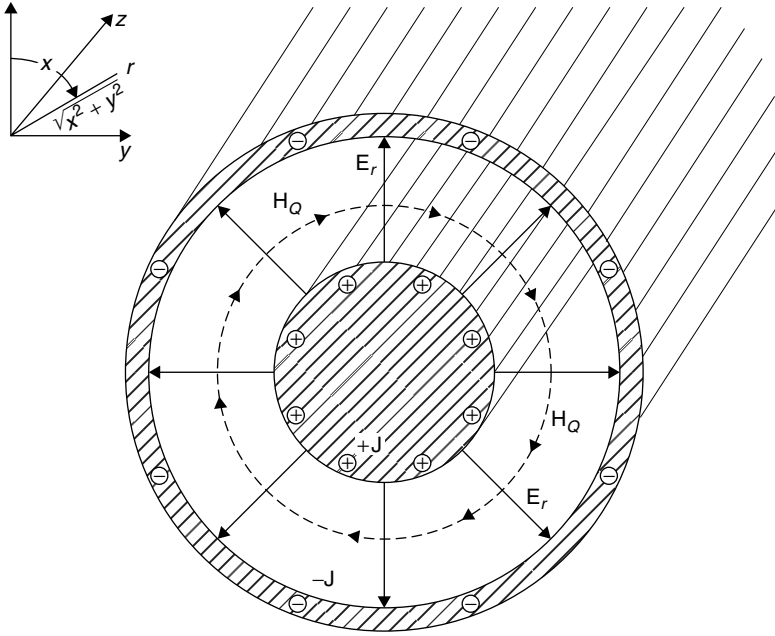
Figure 7.25-5 Coaxial line with transverse E and H fields and axial current density J.

tions, the phasor form of H_φ includes the propagation factor e^{-jkz}. Thus,

$$H_\varphi = \frac{I_0}{2\pi r} e^{-jkz} \qquad (7.25\text{-}12)$$

For purposes of illustration the current lines in Figure 7.25-5 are shown as filaments into the paper ($+z$ direction) on the center conductor and out of the paper on the inside of the outer conductor. In actuality, however, the currents are distributed uniformly throughout cross sections of the conductors for DC currents and in uniform sheets near the inner surfaces (due to skin effect) of the conductors for high frequency AC currents.

The ratio $V_0/I_0 = Z_0$ is defined as the *characteristic impedance* of the transmission line. This can be evaluated using (7.25-9) and (7.25-11) with the result

$$Z_0 = \frac{V_0}{I_0} = \frac{E_r}{H_\varphi} \frac{\ln(b/a)}{2\pi} \qquad (7.25\text{-}13)$$

Next, if the value for the intrinsic impedance $\eta = \sqrt{\mu/\varepsilon}$ is substituted for the ratio of the orthogonal fields E_r and H_φ,

$$Z_0 = \frac{1}{2\pi} \sqrt{\frac{\mu}{\varepsilon}} \ln \frac{b}{a} \qquad (7.25\text{-}14)$$

Substituting the values $\mu = \mu_r \mu_0$, $\varepsilon = \varepsilon_r \varepsilon_0$, and $\sqrt{\mu_0/\varepsilon_0} \approx 120\pi$ gives

$$Z_0 \approx 60\sqrt{\frac{\mu_r}{\varepsilon_r}} \ln\left(\frac{b}{a}\right) \qquad (7.25\text{-}15)$$

Previously in (7.25-7) we determined the distributed capacitance C of the transmission line. From (4.11-4) the corresponding distributed inductance is related to Z_0 and C by

$$Z_0 = \sqrt{\frac{L}{C}} \qquad (7.25\text{-}16)$$

Substituting C from (7.25-7) and Z_0 from (7.25-14) into (7.25-16), L is determined to be

$$L = \frac{\mu}{2\pi} \ln \frac{b}{a} \qquad (7.25\text{-}17)$$

It is interesting to note that we only proved that the ratio of the orthogonal $\vec{E}$ and $\vec{H}$ fields of a propagating wave was equal to η under the conditions that these fields had no variation in the plane transverse to propagation. Although the E_r and H_φ fields for the coaxial line are functions of r, their ratio is equal to the intrinsic impedance η of the insulating region between the conductors. This can be verified by the execution of Exercise 7.25-3.

These derivations demonstrate how Maxwell's equations can be used to solve for the properties of transmission lines. While the principles are the same for all transmission lines, closed-form analytic expressions for the fields about the conductors of various lines such as stripline, microstrip, and coplanar waveguide are much more difficult to formulate that those for coaxial line, one of the few transmission line formats permitting a simple and exact solution. In many cases, a complete analytic approximation is not practical, and electromagnetic (EM) simulation is required to obtain characteristic impedance and other transmission line parameters of sufficient precision for engineering purposes.

7.26   WAVEGUIDES

General Waveguide Solution

The solution for the fields in waveguides (conducting hollow pipes) is different than the solutions for a free space plane wave or a multiconductor transmission line. This is because a TEM wave cannot propagate within a hollow pipe that lacks an axial conductor to carry current, as is available, for example, in a coaxial line. The argument is that a transverse electric field that has a nonzero

curl requires an axial magnetic field. Alternatively, a transverse magnetic field would require either an axial conduction current (which cannot exist since there is no center conductor) or a changing axial electric field. The outcome is that waveguides support *transverse electric (TE)* and *transverse magnetic (TM)* modes of propagation but not a *TEM* mode.

We will solve for the fields of the dominant mode of rectangular waveguide since this is the type most commonly used. The method, described by Ramo and Whinnery [6, Chapters 8 and 9] and attributed by them to Schelkunoff [11], is sufficiently general that it allows solution of the various modes in all types of waveguides. The derivation of the $\vec{E}$ and $\vec{H}$ fields is begun by expanding Maxwell's third and fourth equations, as was done for the plane wave case, but in this case we do not assume that the $\vec{E}$ and $\vec{H}$ fields have no x or y variations, as was assumed for the propagating plane wave case.

On the other hand, we will assume a sinusoidal time variation of the fields since this is the way in which waveguides are used. Indeed, as will be shown, waveguides exhibit a cutoff frequency, below which propagation cannot occur. It should also be noted that even above the cutoff frequency, waveguides are dispersive. That is, they do not have constant propagation delay versus frequency. Consequently, *waveguides cannot propagate absolutely distortion-free representations of arbitrary electrical disturbances*, as can a free space plane wave or the TEM wave in a two or more conductor transmission line.

Waveguides have only a single conductor, and for this reason the transverse electrical field cannot be derived from a static potential function, as was done for the coaxial line. This is because no static potential difference can exist within a charge-free, completely enclosed conductor. Unlike the TEM mode for multiconductor transmission lines, the currents in the waveguide walls have transverse as well as longitudinal components. Waveguides do not have uniquely identifiable voltages and currents associated with propagation, and therefore *waveguides have no uniquely defined characteristic impedance*. As an alternative, a characteristic wave impedance can be defined for waveguide as the ratio of the orthogonal $\vec{E}$ to $\vec{H}$ fields in the transverse plane. However, as we shall see in the study of Green's functions, an absolute impedance for waveguide can be defined under certain conditions.

It will be shown that the characteristic wave impedance of waveguide is approximately equal to the intrinsic impedance of free space, about 377 Ω, for frequencies sufficiently above the cutoff frequency. The fact that waveguides have broad conducting walls results in relatively low ohmic losses. Furthermore, waveguides require no dielectric to support their structure, hence air dielectric is practical, with negligible dielectric loss. Finally, since waveguide propagation is totally enclosed, there is no radiative loss. As a result, waveguides usually provide the lowest loss practical transmission media for guided waves. Their low loss as well as their high power capacity are principal reasons for their use. In addition to being an efficient transmission medium, they find considerable use in the construction of high Q resonator elements, couplers, and power dividers.

We will assume lossless conditions and begin the analysis of waveguides by expanding Maxwell's third and fourth equations, assuming a z dependence of $e^{-\gamma z}$ for waves traveling in the $+z$ direction and $e^{+\gamma z}$ for those traveling in the $-z$ direction. A real value for γ means wave attenuation, while an imaginary value for γ means lossless propagation.

The following analysis assumes $e^{j\omega t}$ time dependence. Taking the curl of $\vec{E}$ and of $\vec{H}$, noting that $\overrightarrow{J_C} = 0$ in a source-free region, performing differentiation with respect to time by multiplying by $j\omega$, and differentiation with respect to z by multiplying by $-\gamma$ (for $+z$-directed waves) gives

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} = \begin{vmatrix} \vec{x} & \vec{y} & \vec{z} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ E_x & E_y & E_z \end{vmatrix} \qquad \nabla \times \vec{H} = \frac{\partial \vec{D}}{\partial t} = \begin{vmatrix} \vec{x} & \vec{y} & \vec{z} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ E_x & E_y & E_z \end{vmatrix}$$

For $+z$-directed propagation, the phasor forms of the above equations become

$$-j\omega\mu H_x = \frac{\partial E_z}{\partial y} + \gamma E_y \qquad\qquad (7.26\text{-}1\text{a})$$

$$-j\omega\mu H_y = -\frac{\partial E_z}{\partial x} - \gamma E_x \qquad\qquad (7.26\text{-}1\text{b})$$

$$-j\omega\mu H_z = \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} \qquad\qquad (7.26\text{-}1\text{c})$$

$$j\omega\varepsilon E_x = \frac{\partial H_z}{\partial y} + \gamma H_y \qquad\qquad (7.26\text{-}1\text{d})$$

$$j\omega\varepsilon E_y = -\frac{\partial H_z}{\partial x} - \gamma H_x \qquad\qquad (7.26\text{-}1\text{e})$$

$$j\omega\varepsilon E_z = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} \qquad\qquad (7.26\text{-}1\text{f})$$

[Equations (7.26-1a) to (7.26-1b) from Ramo and Whinnery, 6, p. 316, reprinted with permission.]

For all of the equations in (7.26-1) and in the following derivation, $E_x, E_y, E_z, H_x, H_y,$ and H_z are functions only of x and y. To recover the time and z dependence, multiply by $e^{(j\omega t - \gamma z)}$ for waves traveling in the $+z$ direction and take the real part thereof. For waves traveling in the $-z$ direction replace $-\gamma z$ with $+\gamma z$.

With diligent algebraic application, the relations given in (7.26-1) can be manipulated to give $H_x, H_y, E_x,$ and E_y in terms of E_z and H_z. For example, the first equation in (7.26-2) is obtained by solving (7.26-1a) for E_y and substituting this value into (7.26-1e). Also note that $k^2 = \omega^2\mu\varepsilon$, as was previously defined. Then, for $+z$-directed propagation:

$$H_x = \frac{1}{\gamma^2 + k^2} \left[j\omega\varepsilon \frac{\partial E_z}{\partial y} - \gamma \frac{\partial H_z}{\partial x} \right] \qquad (7.26\text{-}2a)$$

$$H_y = -\frac{1}{\gamma^2 + k^2} \left[j\omega\varepsilon \frac{\partial E_z}{\partial x} + \gamma \frac{\partial H_z}{\partial y} \right] \qquad (7.26\text{-}2b)$$

$$E_x = -\frac{1}{\gamma^2 + k^2} \left[\gamma \frac{\partial E_z}{\partial x} + j\omega\mu \frac{\partial H_z}{\partial y} \right] \qquad (7.26\text{-}2c)$$

$$E_y = \frac{1}{\gamma^2 + k^2} \left[-\gamma \frac{\partial E_z}{\partial y} + j\omega\mu \frac{\partial H_z}{\partial x} \right] \qquad (7.26\text{-}2d)$$

where $k = \omega\sqrt{\mu\varepsilon}$. [Equations (7.26-2a) to (7.26-2d) from Ramo and Whinnery, 6, p. 317, reprinted with permission.] Note that in this derivation, *k is not the propagation constant*. The equations of (7.26-2) apply to all transmission lines and waveguides since they have been derived directly from Maxwell's equations without any limiting assumptions, other than a sinusoidal time variation. In addition, the electric and magnetic fields must also satisfy the Helmholtz equations. Therefore,

$$\nabla^2 \vec{E} = -k^2 \vec{E} \quad \text{and} \quad \nabla^2 \vec{H} = -k^2 \vec{H} \qquad (7.26\text{-}3a,b)$$

The operator ∇^2 is three dimensional and can be separated dimensionally. Thus

$$\nabla^2 \vec{E} = \nabla^2_{xy} \vec{E} + \frac{\partial^2 \vec{E}}{\partial z^2} \qquad (7.26\text{-}4)$$

$$\nabla^2 \vec{H} = \nabla^2_{xy} \vec{H} + \frac{\partial^2 \vec{H}}{\partial z^2} \qquad (7.26\text{-}5)$$

The second term is due to the axial field variation along the waveguide, and it has the assumed functional description $e^{-\gamma z}$ in the $+z$ direction. Thus

$$\frac{\partial^2 \vec{E}}{\partial z^2} = \gamma^2 \vec{E} \qquad (7.26\text{-}6)$$

$$\frac{\partial^2 \vec{H}}{\partial z^2} = \gamma^2 \vec{H} \qquad (7.26\text{-}7)$$

Substituting this format into (7.26-3a) and (7.26-3b) gives

$$\nabla^2_{xy} \vec{E} = -(\gamma^2 + k^2)\vec{E} \qquad (7.26\text{-}8)$$

$$\nabla^2_{xy} \vec{H} = -(\gamma^2 + k^2)\vec{H} \qquad (7.26\text{-}9)$$

Some various waveguide types are shown in Figure 7.26-1 along with the $\vec{E}$ field of their *dominant mode*, the mode that propagates at the lowest frequency.
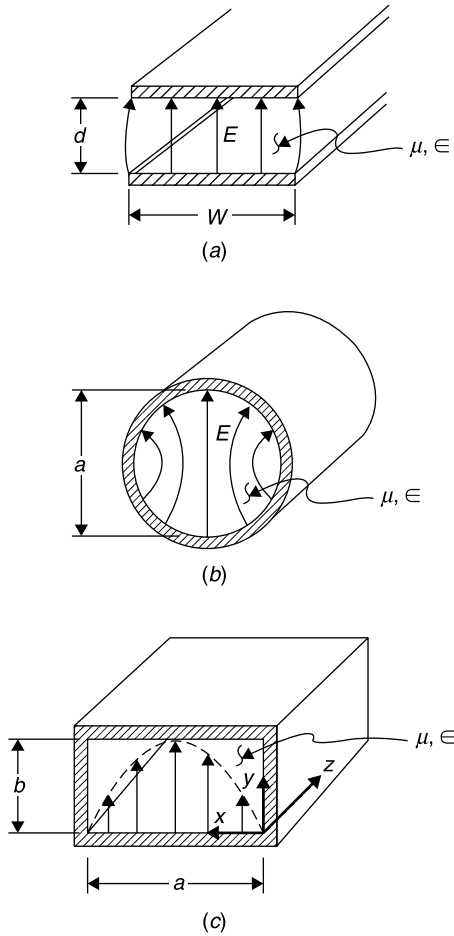
Figure 7.26-1 Waveguide types and the cross-sectional E field pattern of the dominant mode: (*a*) parallel-plate waveguide, (*b*) circular waveguide, and (*c*) rectangular waveguide.

The above equations will be used to determine the propagation characteristics of rectangular waveguide.

Waveguide Types

The parallel-plate waveguide is simple in form and propagates at all frequencies, even direct current. Since it is a two-conductor line, it also can be classed as a transmission line. In principle, it does not radiate because $\vec{E} \times \vec{H}$ is only in the z direction. However, any imperfections or discontinuities in its construction can result in energy leakage. Thus, in practice, it can have radiation loss.

Also, some means must be provided to support the conductors for air dielectric operation since there are no sides.

The round waveguide shown in (Fig. 7.26-1*b*) also can be fabricated simply. In fact, copper water pipe of suitable diameter would serve nicely. The disadvantage of this waveguide is that the polarization of the electric field, shown vertically in Figure 7.26-1*b*, can rotate as the wave propagates due to irregularities in the waveguide. As a result, when the wave reaches the end of the waveguide, a vertically polarized receptor may not be able to intercept all, or even any, of its energy. The rectangular waveguide overcomes these disadvantages and therefore is the most commonly used. The method of solution follows that of Ramo and Whinnery [6].

Rectangular Waveguide Fields

Rectangular waveguide is commonly used in microwave applications and therefore will be the subject of our analysis. Equations (7.26-8) and (7.26-9) are to be satisfied by all of the field components in the waveguide. We have already noted that either TE or TM modes can exist in the waveguide but not TEM modes. Furthermore, we have related the transverse $\vec{E}$ and $\vec{H}$ fields to the axial $\vec{E}$ and $\vec{H}$ fields by the equations in (7.26-2). Thus, it is only necessary to solve for the axial field component for a mode class, TE or TM, and then the transverse fields can be determined from the axial field. Let us consider the TE modes that have an axial $\vec{H}$ field but no axial $\vec{E}$ field. Applying (7.26-9) to the H_z component of the field,

$$\nabla^2_{xy} H_z = \frac{\partial^2 H_z}{\partial x^2} + \frac{\partial^2 H_z}{\partial y^2} = -k_c^2 H_z \qquad (7.26\text{-}10)$$

where $k_c^2 = \gamma^2 + k^2$ and $k^2 = \omega^2 \mu \varepsilon$. Equation (7.26-10) is a partial differential equation that can be solved by the method of separation of variables. To do so we assume it has a solution of the form

$$H_z = f(x)g(y) \qquad (7.26\text{-}11)$$

The functions $f(x)$ and $g(y)$ describe the field dependency in the plane transverse to the direction of propagation, where $f(x)$ is only a function of x and $g(y)$ is only a function of y. In the transverse (xy) plane the solution is of the form

$$H_z = fg \qquad (7.26\text{-}12)$$

in which the z variation is implicit. Substituting this into (7.26-10) and performing the indicated differentiations give

$$f''g + fg'' = -k_c^2 fg$$

$$\frac{f''}{f} + \frac{g''}{g} = -k_c^2 \qquad (7.26\text{-}13)$$

where the double prime indicates the second derivative. Since f and g can be varied independently, each of the terms on the left side of (7.26-13) must equal a constant. There are various forms for the solution of (7.26-13) depending upon whether the ratios are negative, positive, or one negative and one positive [6, p. 355]. If both ratios are assumed to be negative, then we can assign them the values k_x^2 and k_y^2 such that

$$\frac{f''}{f} = -k_x^2 \qquad (7.26\text{-}14)$$

$$\frac{g''}{g} = -k_y^2 \qquad (7.26\text{-}15)$$

The solutions for positive values of k_x^2 and k_y^2 are sinusoids:

$$f(x) = C_1 \cos k_x x + C_2 \sin k_x x \qquad (7.26\text{-}16)$$

$$g(y) = C_3 \cos k_y y + C_4 \sin k_y y \qquad (7.26\text{-}17)$$

and from (7.26-13)

$$k_x^2 + k_y^2 = k_c^2 = \gamma^2 + \omega^2 \mu \varepsilon \qquad (7.26\text{-}18)$$

There are also solutions in terms of sinh and cosh functions if the ratios are positive, but for present purposes the sinusoidal solutions suffice.

Applying Boundary Conditions

For the rectangular waveguide the electric field tangential to a conducting waveguide wall must be zero. Therefore, $E_x = 0$ at $x = 0$ *and a* and $E_y = 0$ at $y = 0$ *and b*. Since, for the TE waves, $E_z = 0$, and, referring to (7.26-2c) and (7.26-2d), since E_y and E_x are proportional to $\partial H_z / \partial y$ and $\partial H_z / \partial x$, respectively, it follows that

$$\frac{\partial H_z}{\partial x} = 0 \text{ at } x = 0, a \quad \text{and} \quad \frac{\partial H_z}{\partial y} = 0 \text{ at } y = 0, b \qquad (7.26\text{-}19a,b)$$

The boundary conditions at $x = 0$ and $y = 0$ require that only the cosine terms in (7.26-16) and (7.26-17) can apply. We will use the remaining boundary conditions at $x = a$ and $y = b$ shortly. Thus

$$H_z = H_0 \cos k_x x \cos k_y y \qquad (7.26\text{-}20)$$

where H_0 is an amplitude constant. Applying (7.26-2a) thru (7.26-2d) and recognizing that $E_z = 0$,

$$H_x = \frac{\gamma k_x}{k_c^2} H_0 \sin k_x x \cos k_y y \qquad (7.26\text{-}21)$$

$$H_y = \frac{\gamma k_y}{k_c^2} H_0 \cos k_x x \sin k_y y \qquad (7.26\text{-}22)$$

$$E_x = \frac{j\omega\mu k_y}{k_c^2} H_0 \cos k_x x \sin k_y y \qquad (7.26\text{-}23)$$

$$E_y = \frac{-j\omega\mu k_x}{k_c^2} H_0 \sin k_x x \cos k_y y \qquad (7.26\text{-}24)$$

where $k_c^2 = \gamma^2 + k^2$ and $k^2 = \omega^2 \mu\varepsilon$.

Propagation Constants and Waveguide Modes

The boundary conditions that $E_y = 0$ at $x = a$ and $E_x = 0$ at $y = b$ in (7.26-24) and (7.26-23) require, respectively, that

$$k_x = \frac{m\pi}{a} \quad \text{and} \quad k_y = \frac{n\pi}{b} \qquad (7.26\text{-}25a,b)$$

where m and n are integers $(0, 1, 2, 3 \ldots)$, both of which cannot be zero simultaneously, that denote the mode number. More about modes shortly. Then

$$(k_c)_{m,n} = \sqrt{k_x^2 + k_y^2} = \sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2} \qquad (7.26\text{-}26)$$

For lossless propagation

$$\gamma = \alpha + j\beta \qquad (7.26\text{-}27a)$$

must be imaginary. Thus, since

$$\gamma^2 = k_c^2 - k^2 \qquad (7.26\text{-}27b)$$

$$\gamma = j\beta = \sqrt{k_c^2 - k^2} = \sqrt{k_c^2 - (2\pi f)^2 \mu\varepsilon} \qquad (7.26\text{-}27c)$$

For propagation to occur the quantity under the square root must be negative, or

$$f > \frac{k_c}{2\pi\sqrt{\mu\varepsilon}} = \frac{1}{2\pi\sqrt{\mu\varepsilon}} \sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2} \qquad (7.26\text{-}28)$$

Since the operating wavelength at frequency f is $\lambda = 1/f\sqrt{\mu\varepsilon}$, the condition at which (7.26-28) is just satisfied corresponds to the *cutoff wavelength* and is given by

$$\lambda_C = \frac{2\pi}{k_c} = \frac{1}{\sqrt{(m/2a)^2 + (n/2b)^2}} \qquad (7.26\text{-}29)$$

Because $f_c\lambda_c = v = 1/\sqrt{\mu\varepsilon}$, a *cutoff frequency* can be defined as

$$f_c = \frac{v}{\lambda_c} = \frac{c}{\lambda_c}\frac{1}{\sqrt{\mu_r\varepsilon_r}} \qquad (7.26\text{-}30)$$

Thus when $f < f_c$ (or $\lambda > \lambda_c$), the waveguide does not propagate in that mode. Since rectangular waveguide is made with $a > b$, and since m and n cannot both be zero simultaneously, the longest free space wavelength for propagation in rectangular waveguide applies to the TE$_{10}$ mode ($m = 1$ and $n = 0$) given by

$$(\lambda_C)_{10} = 2a \qquad (7.26\text{-}31)$$

Substituting $k_x = m\pi/a$, $k_y = n\pi/b$, $\gamma^2 + k^2 = k_c^2$, and $\gamma = j\beta$ for propagating waves into (7.26-21) through (7.26-24) gives the transverse field components as

$$E_x = \frac{j\omega\mu}{k_c^2}\frac{n\pi}{b}H_0\cos\frac{m\pi x}{a}\sin\frac{n\pi y}{b} \qquad (7.26\text{-}32)$$

$$E_y = \frac{-j\omega\mu}{k_c^2}\frac{m\pi}{a}H_0\sin\frac{m\pi x}{a}\cos\frac{n\pi y}{b} \qquad (7.26\text{-}33)$$

$$H_x = \frac{j\beta}{k_c^2}\frac{m\pi}{a}H_0\sin\frac{m\pi x}{a}\cos\frac{n\pi y}{b} \qquad (7.26\text{-}34)$$

$$H_y = \frac{j\beta}{k_c^2}\frac{n\pi}{b}H_0\cos\frac{m\pi x}{a}\sin\frac{n\pi y}{b} \qquad (7.26\text{-}35)$$

Note that H_z is given by (7.26-20). Generally, a rectangular waveguide is operated in the *dominant mode*, the mode that propagates at the lowest frequency. This is the TE$_{10}$ mode whose cutoff wavelength is $2a$. For the TE$_{m0}$ modes $E_x = 0$ and E_y has a sinusoidal variation in amplitude across the x dimension of the waveguide. The amplitude patterns for the TE$_{10}$, TE$_{20}$, and TE$_{30}$ modes are shown in Figure 7.26-2.

Similarly, for the TE$_{0n}$ modes $E_y = 0$ and the E_x field component varies sinusoidally in the y direction as shown in Figure 7.26-3. Field patterns for additional TE and TM modes can be found in [6].

It should be noted that (7.26-20) through (7.26-24) and (7.26-32) through (7.26-35) do not contain the implicit $e^{-j(\omega t - \beta z)}$ factor. To obtain the actual field
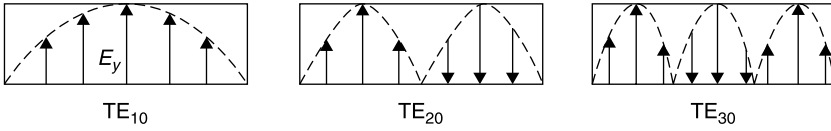
Figure 7.26-2 Electric field magnitude patterns of the TE_{10}, TE_{20}, and TE_{30} modes.

components as a function of z and t, one must reinsert this $e^{-j(\omega t - \beta z)}$ factor and then take the real part thereof. For example,

$$E_y(x, y, z, t) = \frac{\omega\mu k_x}{k_c^2} H_0 \sin \frac{m\pi x}{a} \cos \frac{n\pi y}{b} \cos(\omega t - \beta z) \qquad (7.26\text{-}36)$$

$$H_x(x, y, z, t) = \frac{-\beta k_x}{k_c^2} H_0 \sin \frac{m\pi x}{a} \cos \frac{n\pi y}{b} \cos(\omega t - \beta z) \qquad (7.26\text{-}37)$$

Note that E_y and H_x are in phase for all values of t and z and the minus sign preceding H_x. Thus $\vec{E} \times \vec{H}$ yields real power flow and it is in the $+z$ direction.

Usually, it is undesirable to have the waveguide capable of propagation in more than one mode because energy in the dominant mode might be coupled to a *higher order mode*, one having a shorter cutoff wavelength, and then that energy may not be recoverable by the mode launchers and receptors used at the ends of the transmission system. Even if the launchers could recover the higher order mode, the various modes travel at different phase velocities, as will be seen, and different modes would not usually combine in phase.

In short, the best practice is usually to operate in the frequency range for which only one mode, the *dominant mode*, can propagate. Standard waveguides have a ratio of width to height of at least 2. Thus, the next higher propagating mode, the TE_{20}, would have a cutoff wavelength of a, yielding an operating bandwidth ratio of 2 to 1 for the dominant TE_{10} mode only. However, operation very near to cutoff is impractical, and so practically the operating bandwidth is about 1.5 to 1. For example, WR-90 waveguide has a recommended operating band of 8.2 to 12.4 GHz. See Appendix E.

The modes that do not propagate are called *evanescent modes*, evanescent meaning "vanishing." Their propagation constant is real, and this corresponds
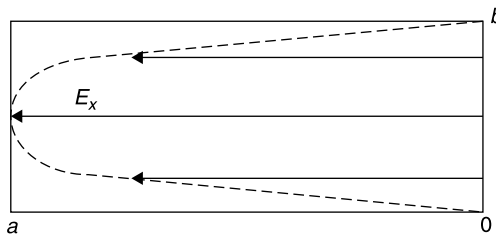
Figure 7.26-3 Electric field magnitude of the TE_{01} mode (for which $\lambda_c = 2b$).

to an attenuating wave amplitude with distance from the site at which they are excited. Irregularities and obstacles in the waveguide serve to excite the field patterns of higher order modes; and, while a waveguide operated only in the dominant mode frequency range may not propagate them, evanescent modes and their resulting fields, which decrease reactively from their excitation sites, store energy, causing them to appear as reactive circuit elements in the waveguide.

   Although we expressed the cutoff of the waveguide relative to the operating wavelength λ, the wavelength of propagation *inside the waveguide,* λ_g, is considerably longer than the operating wavelength. This is particularly true as λ approaches the cutoff wavelength λ_c. This is shown by

$$\lambda_g = \frac{2\pi}{\beta} = \frac{2\pi}{\sqrt{(2\pi f)^2 \mu\varepsilon - k_c^2}} \qquad (7.26\text{-}38)$$

Since the operating wavelength is related to frequency by $f\lambda = 1/\sqrt{\mu\varepsilon}$ and since, from (7.26-29) $\lambda_c k_c = 2\pi$, (7.26-38) can be rewritten as

$$\lambda_g = \frac{\lambda}{\sqrt{1 - (\lambda/\lambda_c)^2}} = \frac{\lambda}{\sqrt{1 - (f_c/f)^2}} \qquad (7.26\text{-}39)$$

Note that for an air-filled waveguide, $\mu_r = \varepsilon_r = 1$ and $\lambda = \lambda_0 = c/f$.

Characteristic Wave Impedance for Waveguides

The ratio of voltage to current for a traveling wave is Z_0, the characteristic impedance of the transmission line. However, a waveguide does not have unique terminals at which voltage and current can be specified and measured. On the other hand, the ratio of the transverse $\vec{E}$ and $\vec{H}$ fields is unique and is defined as the *characteristic wave impedance*, as was the case for a plane wave propagating in an unbounded medium. In that case, it was shown that the impedance was frequency independent and equal to $\eta = \sqrt{\mu/\varepsilon}$. In waveguide, this ratio approaches η for frequencies well above cutoff but changes rapidly near cutoff with the same functionality as the guide wavelength in (7.26-39).

   We find the transverse $\vec{E}$ and $\vec{H}$ fields from (7.26-21) through (7.26-24) and define the characteristic wave impedance for the TE modes as

$$Z_{\text{TE}} = \frac{E_x}{H_y} = -\frac{E_y}{H_x} = \frac{j\omega\mu}{\gamma} \qquad (7.26\text{-}40)$$

Substituting $\gamma = j\sqrt{k^2 - k_c^2}$, $k = 2\pi/\lambda$, $k_c = 2\pi/\lambda_c$, and $\omega\mu = k\eta$, (7.26-40) can be rewritten as

$$Z_{\text{TE}} = \frac{\eta}{\sqrt{1 - (\lambda/\lambda_C)^2}} = \frac{\eta}{\sqrt{1 - (f_C/f)^2}} \qquad (7.26\text{-}41a)$$

where $\eta = \sqrt{\mu/\varepsilon} \approx 120\pi\sqrt{\mu_r/\varepsilon_r} \approx 377\sqrt{\mu_r/\varepsilon_r}\ \Omega$. By a similar analysis one can obtain the characteristic wave impedance for the TM waves, namely

$$Z_{\text{TM}} = \eta\sqrt{1 - \left(\frac{\lambda}{\lambda_C}\right)^2} = \eta\sqrt{1 - \left(\frac{f_C}{f}\right)^2} \qquad \text{(7.26-41b)}$$

Note that both Z_{TE} and Z_{TM} are real only for frequencies above cutoff ($f > f_C$).

Phase and Group Velocities

Waveguide shows high dispersion near the cutoff frequency. However, pulsed signals on microwave carrier frequencies propagated through waveguides are not likely to have significant distortion since the relative bandwidths are usually fairly small and waveguide paths usually are fairly short, for example, about 100 ft to ascend a communication tower. The definitions and expressions for *phase velocity* and *group velocity* are as follows:

$$\text{Phase velocity} = v_P \equiv \frac{\omega}{\beta} = \frac{1}{\sqrt{\mu\varepsilon}\sqrt{1 - (f_C/f)^2}} \qquad \text{(7.26-42)}$$

$$\text{Group velocity} = v_g \equiv \frac{\partial\omega}{\partial\beta} = \frac{\sqrt{1 - (f_C/f)^2}}{\sqrt{\mu\varepsilon}} \qquad \text{(7.26-43)}$$

From (7.26-43) it is seen that group velocity, the speed at which information propagates, is slower than the speed of light in the medium, $v = 1/\sqrt{\mu\varepsilon} = c/\sqrt{\mu_r\varepsilon_r}$, ever more so as the operating frequency approaches the cutoff frequency of the waveguide. The phase velocity is correspondingly faster. This violates no relativity constraints, since no physical matter (or information) travels at this speed, only a *steady-state* constant-phase wave front.

TE and TM Mode Summary for Rectangular Waveguide

The same technique just used to solve for the TE modes can be applied to solve for the fields and related parameters of TM modes. The results can be cast into the format shown below using the various equivalences covered in the TE mode development [after Ramo and Whinnery, 6, Sec. 9.03, reprinted with permission].

TE$_{mn}$ Modes

$$H_z = H_0 \cos\frac{m\pi x}{a} \cos\frac{n\pi y}{b} \qquad \text{(7.26-44a)}$$

$$E_x = j\eta\frac{n\pi}{bk_c}\frac{f}{f_C}H_0 \cos\frac{m\pi x}{a} \sin\frac{n\pi y}{b} \qquad \text{(7.26-44b)}$$

$$E_y = -j\eta \frac{m\pi}{ak_c} \frac{f}{f_C} H_0 \sin \frac{m\pi x}{a} \cos \frac{n\pi y}{b} \qquad (7.26\text{-}44\text{c})$$

$$H_x = -\frac{E_y}{Z_{\text{TE}}} \qquad (7.26\text{-}44\text{d})$$

$$H_y = \frac{E_x}{Z_{\text{TE}}} \qquad (7.26\text{-}44\text{e})$$

$$\lambda_C = \frac{2\pi}{k_c} = \frac{1}{\sqrt{(m/2a)^2 + (n/2b)^2}} \qquad (7.26\text{-}44\text{f})$$

where

$$\eta = \sqrt{\frac{\mu}{\varepsilon}} \qquad k_c = \sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2} \qquad Z_{\text{TE}} = \frac{\eta}{\sqrt{1 - (\lambda/\lambda_C)^2}}$$

TM$_{mn}$ Modes

$$E_z = E_0 \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} \qquad (7.26\text{-}45\text{a})$$

$$H_x = j\frac{n\pi}{b\eta k_c} \frac{f}{f_C} E_0 \sin \frac{m\pi x}{a} \cos \frac{n\pi y}{b} \qquad (7.26\text{-}45\text{b})$$

$$H_y = -j\frac{m\pi}{a\eta k_c} \frac{f}{f_C} E_0 \cos \frac{m\pi x}{a} \sin \frac{n\pi y}{b} \qquad (7.26\text{-}45\text{c})$$

$$E_x = Z_{\text{TM}} H_y \qquad (7.26\text{-}45\text{d})$$

$$E_y = -Z_{\text{TM}} H_x \qquad (7.26\text{-}45\text{e})$$

$$\lambda_C = \frac{2\pi}{k_c} = \frac{1}{\sqrt{(m/2a)^2 + (n/2b)^2}} \qquad (7.26\text{-}45\text{f})$$

where

$$\eta = \sqrt{\frac{\mu}{\varepsilon}} \qquad k_c = \sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2} \qquad Z_{\text{TM}} = \eta\sqrt{1 - (\lambda/\lambda_C)^2}$$

TE$_{10}$ Mode   For the TE$_{10}$ mode, $m = 1$, $n = 0$, and $\lambda_C = 2a$. The resultant expressions for the $\vec{E}$ and $\vec{H}$ fields of the TE$_{10}$ mode are obtained from the general expressions of (44):

$$E_y = E_0 \sin \frac{\pi x}{a} \qquad (7.26\text{-}46)$$

$$H_x = -\left(\frac{E_0}{Z_{\text{TE}}}\right) \sin\left(\frac{\pi x}{a}\right) \qquad (7.26\text{-}47)$$

$$H_z = \frac{jE_0}{\eta} \left(\frac{\lambda}{2a} \right) \cos \frac{\pi x}{a} \qquad (7.26\text{-}48)$$

$$E_x = H_y = 0 \qquad (7.26\text{-}49)$$

$$\lambda_g = \frac{\lambda}{\sqrt{1 - (\lambda/2a)^2}} \qquad (7.26\text{-}50)$$

$$Z_{\text{TE}} = \frac{\eta}{\sqrt{1 - (\lambda/2a)^2}} \qquad (7.26\text{-}51)$$

$$v_p = \frac{v}{\sqrt{1 - (\lambda/2a)^2}} \qquad (7.26\text{-}52)$$

$$v_g = v\sqrt{1 - (\lambda/2a)^2} \qquad (7.26\text{-}53)$$

$$\lambda_c = 2a \qquad (7.26\text{-}54)$$

$$f_c = \frac{1}{2a\sqrt{\mu\varepsilon}} \qquad (7.26\text{-}55)$$

where $\eta \approx 377\sqrt{\mu_r/\varepsilon_r}\ \Omega$ and $v = 1/\sqrt{\mu\varepsilon} = c/\sqrt{\mu_r\varepsilon_r}$.

The field pattern of the TE$_{10}$ mode can be seen in Figure 7.26-4 to move as a packet, one guide wavelength long. To obtain a complete expression for the field pattern, reinsert the $e^{j(\omega t - \beta z)}$ factor and take the real part of the result using (7.26-46), (7.26-47), and (7.26-48). For example,

$$E_y(x, z, t) = E_0 \sin \frac{\pi x}{a} \cos(\omega t - \beta z) \qquad (7.26\text{-}56)$$

The pattern shown in Figure 7.26-4 is for $t = 0$. The E_y and H_x field components are in phase, as is necessary for real power flow. Their cross product, Poynting's vector, gives the power density flow in the waveguide's cross section. Integration over the cross-sectional area of the waveguide yields an expression for the total power flow. The conducting waveguide walls constrain the fields to within the hollow waveguide pipe. They do so with a current distribution that confines the fields within the waveguide walls, rather than radiating into the surrounding space. The current distribution can be obtained from the field components within the waveguide using Maxwell's fourth equation.

The conduction current distribution on the inside walls of the waveguide are depicted in Figure 7.26-5. Notice that for the instant of time shown, the currents converge at a point $\lambda_g/4$ from the left end of the guide. At this point the currents continue into the space within the waveguide as *displacement currents*. Note also that at this point and time, $\vec{E}$ is zero but $\partial\vec{E}/\partial t$ is a maximum as the
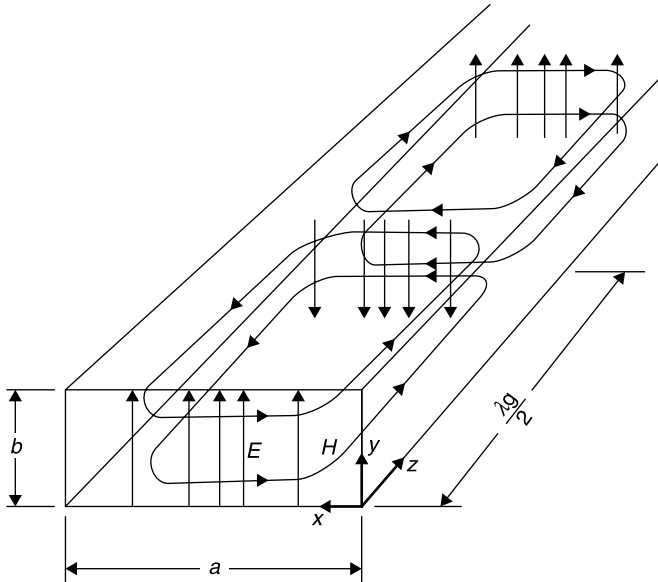
Figure 7.26-4 E and H fields of the TE_{10} mode in rectangular waveguide.

wave propagates. This is true for all sinusoids, namely that when the amplitude is zero, its derivative is maximum. A maximum $\partial \vec{E}/\partial t$ corresponds to a maximum displacement current.

Keep in mind that while the fields and currents are shown as discrete lines in Figures 7.26-4 and 7.26-5 for purposes of illustration, they actually are continuously variable sine functions.
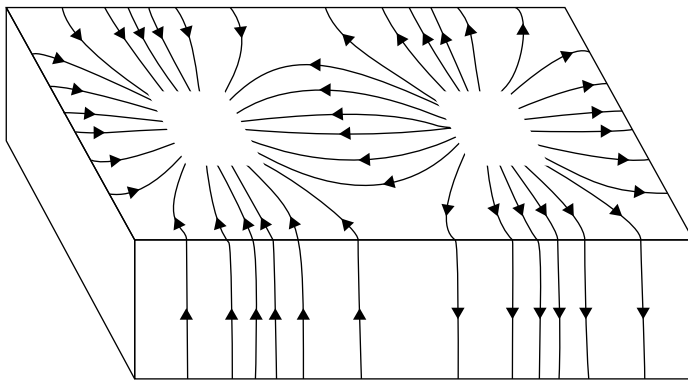
Figure 7.26-5 Current distribution on the inside walls of the waveguide for the TE_{10} mode. [After Ramo and Whinnery, 6, p. 370, with permission.]

7.27   FOURIER SERIES AND GREEN'S FUNCTIONS

Fourier Series

A function can be approximated over an interval by a constant value plus a collection of harmonically related sine and cosine terms:

$$f(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos nx + \sum_{n=1}^{\infty} b_n \sin nx \qquad (7.27\text{-}1)$$

where

$$a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\,dx \qquad (7.27\text{-}2)$$

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx\,dx \qquad (7.27\text{-}3)$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx\,dx \qquad (7.27\text{-}4)$$

This equivalence is termed a *Fourier series* representation [2, Sec. 7.11], and the statement that the equivalent Fourier series can be found is called the *Fourier theorem*. By judicious choice of the origin $x = 0$, it may be found that the function has even symmetry, $f(x) = f(-x)$, in which case the series has only cosine terms. Or, if it has odd symmetry, $f(x) = -f(-x)$, the series has only sine terms. In general, the Fourier series only represents $f(x)$ over the interval $0 \le x \le L$ unless $f(x)$ is periodic with period L *for all* x, that is, $f(x) = f(x - L)$ for all x, in which case the series represents $f(x)$ for all x.

As an example, consider the function

$$f(t) = A \qquad (7.27\text{-}5)$$

for

$$0 \le t \le T \qquad (7.27\text{-}6)$$

where A is a constant (Fig. 7.27-1). This function has a Fourier series representation:

$$f(t) = \frac{4A}{\pi} \left(\sin \frac{\pi t}{T} + \frac{1}{3} \sin \frac{3\pi t}{T} + \frac{1}{5} \sin \frac{5\pi t}{T} + \frac{1}{7} \sin \frac{7\pi t}{T} + \cdots \right) \quad (7.27\text{-}7)$$

in which the independent variable, x, in (7.27-1) has been set equal to $\pi t/T$ in order to match the half period of the sine wave with the interval T over which
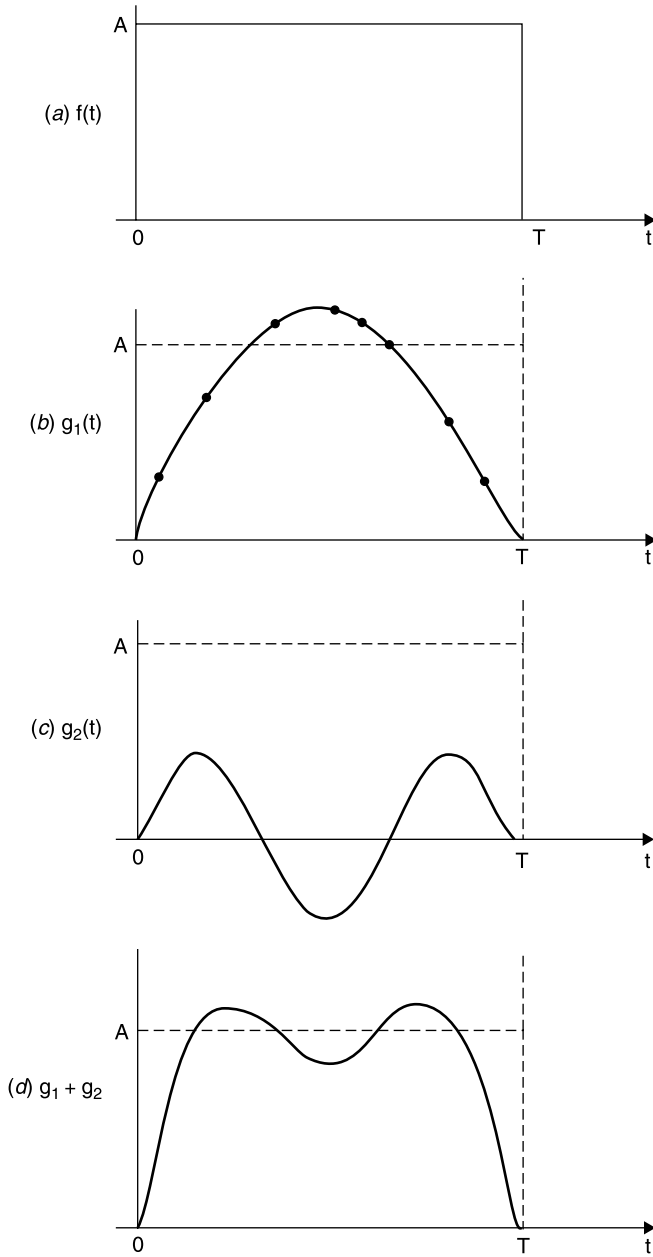
Figure 7.27-1   Representation of a rectangular waveform by the first two terms of a Fourier series.
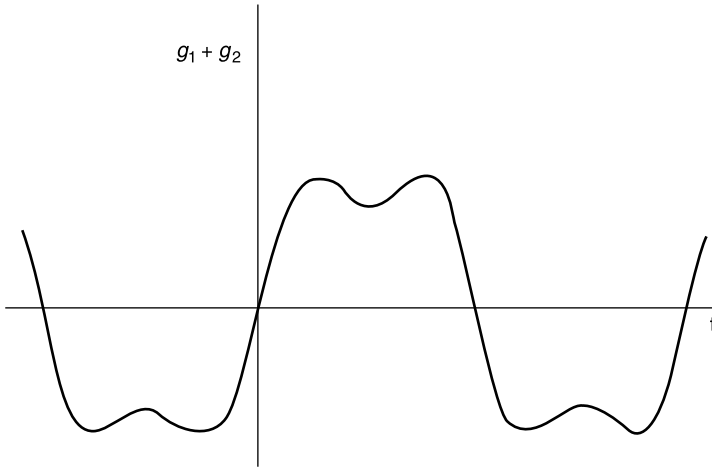
Figure 7.27-2   Periodicity of the function, $g_1 + g_2$, shown in Figure 7.27-1.

the function is defined. An approximation, $g_1 + g_2$, to the result in (7.27-7) using just the first two terms of the series is shown in Figure 7.27-1. The use of more terms would produce a better approximation. This can be seen in Figure 7.27-1. The first term, g_1 shown in Figure 7.27-1*b*, is not as good a representation of $f(t)$ as is the sum of the first two terms, $g_1 + g_2$ shown in Figure 7.27-1*d*. Notice that outside of the interval $0 \leq t \leq T$ the series representation is periodic with a period of $2T$ (Fig. 7.27-2), even though the function $f(t)$ may not be defined outside the range of $0 \leq t \leq T$. Also note that the function $g_1 + g_2$ shown in Figure 7.27-2 is an odd function, having been formed from sine terms.

Green's Functions

Green's functions perform a role similar to that of the Fourier series by describing the fields in waveguides and other structures having distinct allowed modes. Such distinct modes are called eigenmodes. The separate terms of the Green's function series are the allowed modes. With the Fourier series we were free to select frequency components that were periodic over any interval desired, although selecting the period of the fundamental sine wave can simplify the analysis. Since the Green's functions incorporate the eigenvalue mode solutions to the geometry in question, their periodicity is predetermined by the form of those modes. Also, Green's functions are written in terms of differential forms of excitation of the modes, such as a current filament.

The use of a Green's function is best demonstrated by an example. In this example the Green's function is used to relate the magnitude of the dominant TE_{10} mode in the waveguide to the current on a thin post installed between the waveguide's broad walls. To simplify the analysis we approximate the behavior
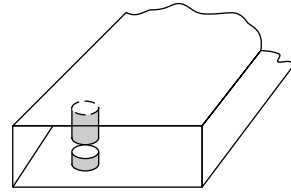
Figure 7.27-3   Thin post with a small gap into which a diode can be installed in rectangular waveguide.

of the post by a thin metal strip (Fig. 7.27-3). Often it is desired to connect a switching PIN diode or microwave source Gunn diode within a gap in the post, thereby coupling the diode to the power that propagates in the waveguide.

The following problem arises: since the waveguide has no uniquely defined characteristic impedance, it is not possible to calculate what impedance is presented to the diode when it is installed in series with the post. In this application example, we will use the Green function to relate the current on the post that is induced by an incident TE_{10} mode wave in the waveguide, and from this infer what absolute characteristic impedance Z_g should be defined for the waveguide. Given this Z_g definition, the load impedance presented to the diode can be calculated with the customary transmission line theory.

In the previous section we solved for the TE and TM modes in rectangular waveguides. These modes are the eigenvalue field solutions for the waveguide. Any electric field distribution in the rectangular waveguide must be representable as a series summation of either the TE_{mn} or the TM_{mn} modes or both since these are the only field patterns that are allowed by the waveguide's boundary conditions. For the rectangular waveguide, these fields are expressed by (7.26-44) and (7.26-45).

Green's functions are written in terms of the field modes that can exist in a region as functions of the primary excitations that can bring them into being, such as an electric field or current distribution in a waveguide. As an example, consider a filamentary unit current in a rectangular waveguide flowing in a vertical metal strip in the waveguide (Fig. 7.27-4).

Such a current filament excites an infinite number of TE_{m0} modes in the waveguide, and the resulting electric field E_y everywhere in the waveguide (at any x, z position) is related to the unit current filament at location (x', z') according to [9, Sec. 5.6]

$$E_y = G(x, z : x', z') = \frac{j\omega\mu_0}{a} \sum_{m=1}^{\infty} \frac{1}{\gamma_m} \sin\frac{m\pi x}{a} \sin\frac{m\pi x'}{a} e^{-\gamma_m|z-z'|} \quad (7.27\text{-}8)$$

where

$$\gamma_m = \sqrt{(m\pi/a)^2 - (2\pi/\lambda)^2} \quad (7.27\text{-}9)$$

and $G(x, z : x', z')$ is the Green's function for a uniform current filament at
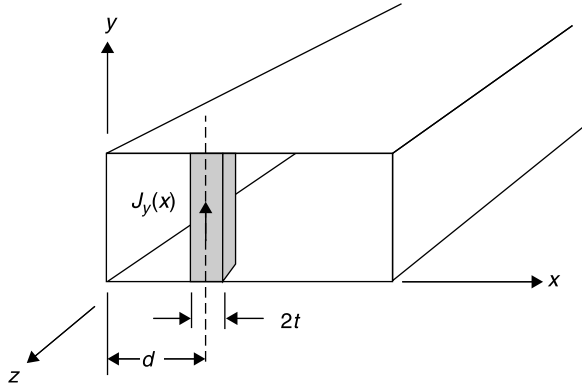
Figure 7.27-4 Vertical current filament serves as an excitation source for rectangular waveguide TE_{m0} modes.

(x', z'). It is an expression for all of the TE_{10} modes in rectangular waveguides that are excited by a unit current filament located at (x', z').

The time dependence, $e^{j\omega t}$, is assumed. Note that in (7.27-8), because of the magnitude sign in the exponential, the modes excited by the current filament propagate equally in both the $+z$ and $-z$ directions away from the filament. The expression for the peak value of E_y in (7.27-8) is for a unit current filament, having a magnitude of one ampere peak. The E_y response is linear; thus, if a 2 A current is applied, the value for E_y would be doubled. Term E_y is expressed in (7.27-8) as a summation of the TE_{m0} modes of the waveguide since these are the only modes excited by a vertical (y-directed) current filament. If there were more than one current filament, the total E_y at any location (x, z) is found by superimposing, or integrating, the current distribution of all such current filaments. Needless to say, this general formulation quickly could become too complex for solution; however, in the example to follow, we assume all of the current is located at one (x', z') location. Notice that in (7.27-8) the primed coordinates (x', z') are the locations of current filaments and the unprimed (x, z) are the locations at which the resulting E_y is evaluated.

A filamentary current in the waveguide is like a radiating antenna. A free space, vertically directed wire antenna, analyzed later in Section 7.32, can radiate power in any azimuthal direction. However, this "waveguide current antenna" located at (x', z') is constrained by the waveguide to propagate energy only in the $+z$ and $-z$ directions and only in the TE_{10} mode, since we assume that the waveguide is operated in the dominant mode frequency range. Note that a current filament having no y variation in the waveguide can only excite TE_{m0} modes, as (7.27-8) indicates.

In this example, the thin post (Fig. 7.27-3) is approximated by the strip geometry in Figure 7.27-4 because the strip exists at a single z' location. For thin posts this is a reasonable approximation. The procedure for determining Z_g is as follows [12]. Assume that there is a TE_{10} mode wave defined by (7.27-10)

and propagating in the $+z$ direction and incident on the conducting strip. The key to defining Z_g is to relate the magnitude of this current, I, on the strip to a given amount of power incident in the TE_{10} mode. The incident TE_{10} wave is described as

$$E_{IN} = E_0 \sin(\pi x/a) e^{-\gamma_1 z} \qquad (7.27\text{-}10)$$

where a is the waveguide width and γ_1 is defined in (7.27-9). When this incident wave impinges on the strip, the tangential electric field on the strip must be zero since we assume it to be perfectly conducting. As a consequence, this incident wave induces a current distribution, $J_y(x)$, on the strip. The current distribution excites an infinite number of TE_{m0} modes, the sum of whose fields exactly cancel the incident TE_{10} electric field on the surface of the strip. All of the induced TE_{m0} modes so induced, except the TE_{10}, are evanescent since we assume that the waveguide is being operated in the dominant mode frequency range. The evanescent modes affect the energy storage in the vicinity of the strip obstacle, and accordingly its normalized reactance, jx, but for now only the scattered (reflected) TE_{10} dominant mode component need be evaluated. This scattered (reflected) dominant mode, E_R, is the propagating TE_{10} wave that is reflected by the strip. It is defined in (7.27-11) and evaluated using (7.27-8). For this example we assume that the waveguide is air filled and therefore $\varepsilon = \varepsilon_0$, $\mu = \mu_0$, and $\lambda = \lambda_0$. Then

$$E_R = \Gamma E_0 \sin(\pi x/a) e^{\gamma_1 z} \qquad (7.27\text{-}11)$$

$$E_R = \frac{-j\omega\mu_0}{a\gamma_1} \int_{d-t}^{d+t} J_y(x) \sin(\pi x/a) \, dx \qquad (7.27\text{-}12)$$

where $\Gamma = E_R/E_{IN}$ is the complex reflection coefficient of the strip to the incident TE_{10} wave. For the narrow strip assumed $J_y(x) \approx J_0 = $ constant. Then $J_0 2t = I$, the peak current induced on the strip. Integrating (7.27-12) under the condition that $t \ll a$ gives

$$E_R = \frac{-j\omega\mu_0}{a\gamma_1} |I| \sin\left(\frac{\pi d}{a}\right) \qquad (7.27\text{-}13)$$

And, since for a lossless waveguide operated in the dominant propagating mode

$$\gamma_1 = j\beta_1 = j\frac{2\pi}{\lambda_g} \qquad (7.27\text{-}14)$$

where λ_g is the guide wavelength of the dominant mode, and

$$\omega\mu_0 = 2\pi f \mu_0 = 2\pi\mu_0 \frac{c}{\lambda_0} = 2\pi\mu_0 \frac{1}{\lambda_0\sqrt{\mu_0\varepsilon_0}} = 2\pi\sqrt{\frac{\mu_0}{\varepsilon_0}} \frac{1}{\lambda_0} \qquad (7.27\text{-}15)$$

Using these equivalences (7.25-13) can be rewritten as

$$E_R = \frac{-|I|}{a}\sqrt{\frac{\mu_0}{\varepsilon_0}}\frac{\lambda_g}{\lambda_0}\ \sin\left(\frac{\pi d}{a}\right) \qquad (7.27\text{-}16)$$

Notice that E_R is negative, $180°$ out of phase with E_{IN}, as must be true, since the reflected wave's polarity must be opposite that of the incident wave in order to contribute, together with higher order excited modes, to the cancellation of the incident field on the surface of the conducting strip.

When Poynting's theorem is applied to the incident TE_{10} wave, its peak propagating power P is related to its peak amplitude E_0 by

$$E_0 = \sqrt{\frac{2P}{ab}\frac{\lambda_g}{\lambda_0}}\sqrt{\frac{\mu_0}{\varepsilon_0}} \qquad (7.27\text{-}17)$$

Since the reflection coefficient Γ is just the complex ratio E_R/E_I, (7.27-17) can be written

$$|E_R| = |\Gamma|\sqrt{\frac{2P}{ab}\frac{\lambda_g}{\lambda_0}}\sqrt{\frac{\mu_0}{\varepsilon_0}} \qquad (7.27\text{-}18)$$

Equating the expressions for E_R in (7.27-16) and (7.27-18), squaring, multiplying both sides of the result by 2, and simplifying gives

$$|I|^2\left[2\frac{b}{a}\sqrt{\frac{\mu_0}{\varepsilon_0}}\frac{\lambda_g}{\lambda_0}\ \sin^2\left(\frac{\pi d}{a}\right)\right] = 4P|\Gamma|^2 \qquad (7.27\text{-}19)$$

Alternatively, from circuit theory a reactance exposed to an incident peak power P and producing the reflection coefficient Γ when connected across a Z_g transmission line (Fig. 7.27-5) has a peak current I passing through it according to

$$|I|^2 Z_g = 4P|\Gamma|^2 \qquad (7.27\text{-}20)$$

Comparing (7.27-19) and (7.27-20), it follows that the appropriate value of waveguide impedance to produce the same current on the thin strip or post for a given incident power is

$$Z_g = 2\frac{b}{a}\sqrt{\frac{\mu_0}{\varepsilon_0}}\frac{\lambda_g}{\lambda_0}\ \sin^2\left(\frac{\pi d}{a}\right) \qquad (7.27\text{-}21)$$

The resulting equivalent circuit for a device placed within a gap, *AB*, in the waveguide post is that shown in Figure 7.27-6. Note that if the post is placed in the center of the waveguide
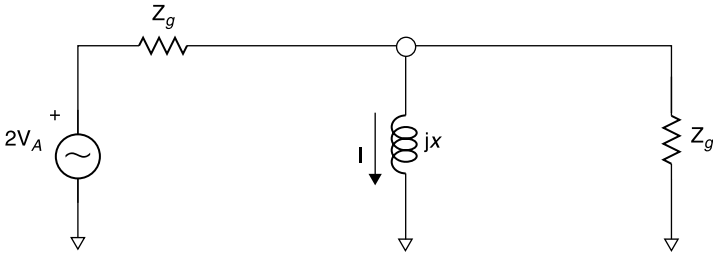
Figure 7.27-5 Equivalent circuit of the waveguide with conducting thin vertical strip (or post).

$$Z_g = 2\frac{b}{a}\sqrt{\frac{\mu_0}{\varepsilon_0}}\frac{\lambda_g}{\lambda_0} \tag{7.27-22}$$

It is interesting to note that this definition for Z_g, which is based on the post current and the TE_{10} mode incident power, is the same value that would be obtained were Z_g defined in terms of the TE_{10} voltage across the waveguide at the location of the post and the propagating power associated with the TE_{10} mode. In fact, (7.27-22) is the commonly defined *voltage-power waveguide impedance* definition. But this definition could not have been advanced a priori because, *with the post present, the voltage across the waveguide at the post location is zero*.

This application of the Green function, which required the evaluation of only one term, has produced a very useful analytical result, allowing the absolute equivalent circuit for the post in a waveguide of Figure 7.27-6. The absolute impedance of the post, $jX = jx(Z_g)$, can be found by determining x either by measurement of or from an analysis of the stored energy of the evanescent fields [13, Sec. 5.11]. The complete circuit, which usually includes a resonant cavity instead of matched load terminations, can then be analyzed using trans-
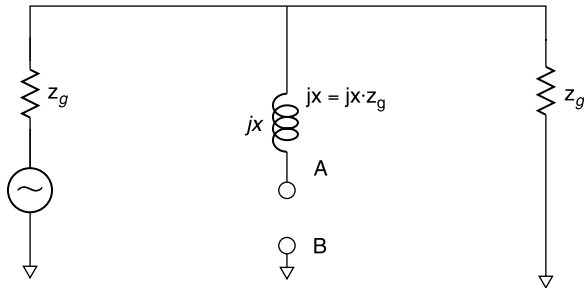
Figure 7.27-6 Equivalent circuit for a device location, *AB*, in the gap of a waveguide post with absolute impedance values.

mission line theory as described in Chapters 4 and 5. Post coupling of semi-conductor devices to the waveguide is a common application, and this equivalent circuit allows calculation of the *absolute load impedance* seen at the device terminals AB.

7.28 HIGHER ORDER MODES IN CIRCUITS

Many engineers have little direct application for waveguides. Circuits, particularly integrated circuits, are small and high frequency interconnections between them usually are made with coaxial transmission lines. Nevertheless, understanding waveguides helps to diagnose problems in circuits that otherwise might seem unexplainable. For example, consider the common situation shown in Figure 7.28-1 in which a circuit board is installed in a metal box.

   The rectangular metal box forms a waveguide at frequencies for which its width a is at least $\lambda_0/2$. If its length in the other lateral dimension is $\lambda_g/2$, it forms a resonator in the TE_{10} waveguide mode and has a high Q resonance at some frequency f_R. This is especially problematical if the circuit must operate or can oscillate (has gain) at f_R. Energy can couple to the resonator via the
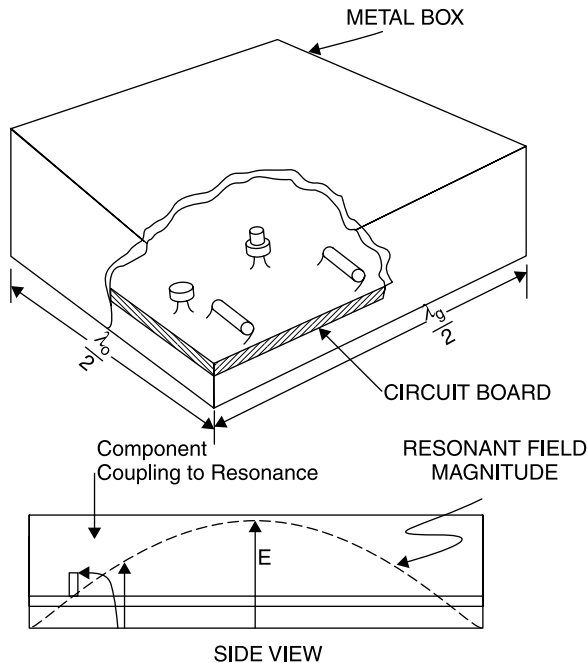
Figure 7.28-1 Circuit housed in a metal box and the TE_{10} waveguide mode resonance field magnitude (side view).

fringing electric field about a component, as shown in the side view of Figure 7.28-1, or through the magnetic field near a component. This coupling will result in the absorption of power from the circuit and dissipation at f_R since practical resonators are not lossless.

Recall from Chapter 3 that the higher the Q the lighter the coupling needed to approach or achieve critical coupling, at which any fraction up to and including all of the circuit's power could be absorbed. Alternatively the circuit may break into oscillation at f_R because the cavity can provide unintended feedback between the output and input circuits of a transistor. Such feedback need not be at the intended operational frequency of the circuit, it can occur at any frequency for which the circuit has gain, and the unintended coupling is appropriate for oscillation.

Without an understanding of the waveguide modes, this problem can be very puzzling and difficult to diagnose. The power absorption or oscillation can be quite sharply defined with frequency since waveguide cavities frequently have unloaded Qs of 1000 or more. Thus, the circuit may perform near perfectly except for a very narrow band of frequencies around f_R.

Moreover, if the box dimensions are considerably larger than $\lambda_0/2$, multiple resonant modes can exist in the box, each of which generally has a different resonant frequency. The simplest means of avoiding such resonances is to ensure that one or both lateral dimensions of the box are less than $\lambda_0/2$ at the highest frequency used in the circuit or at which the circuit has gain. This also assumes that the height of the box is also short compared to $\lambda_0/2$. When the lateral dimensions must be larger than $\lambda_0/2$, *mode suppressors* can be employed. These consist of conductive posts (screws or rivets work nicely) to interconnect the top and bottom covers (Fig. 7.28-2), limiting the free length dimensions in all directions inside the box to less than $\lambda_0/2$.

In this example, we have assumed a rectangular metal circuit enclosure, but higher order modes can exist in a round box structure or in any shape having internal dimensions that extend for a half wavelength in any two of the container's three dimensions.
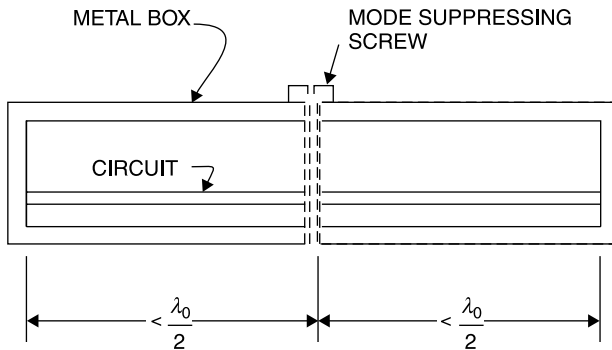
Figure 7.28-2   Use of a conductive post mode suppressor to prevent a box resonance.

7.29 VECTOR POTENTIAL

We demonstrated in Section 7.8 that the static $\vec{E}$ field could be derived from the *static scalar potential function*. This scalar potential was an intermediate calculation whose value was determined by the integration of a charge distribution over a volume. Once determined, the potential function facilitated determination of the $\vec{E}$ field using the gradient function, a derivative process. This process generally poses a simpler task than that of the vector integration that would otherwise be necessary to construct the $\vec{E}$ field directly from a charge distribution.

The scalar potential has obvious physical significance, as the ski slope mountain example demonstrated, and it greatly facilitates the determination of electric field by use of the gradient vector. The success of the scalar potential technique leads to the reasoning that another function might be defined to facilitate determination of the $\vec{B}$ field as well as the time-varying portion of the $\vec{E}$ field [6, Chapter 4]. It will be seen that this can be done by defining a *vector potential*. Unlike the scalar potential, the physical significance of the vector potential is not apparent. It is a mathematical artifice. Nevertheless, it does result in a useful method for routinizing the calculation of the $\vec{B}$ field.

Although Maxwell's equations eloquently describe the physics of electromagnetics, they are not the most convenient format for the solution of some real field situations, such as antennas. It will be seen from the following definition of the vector potential and its use in evaluating the short wire antenna in Section 7.32 that the use of the vector potential neatly derives the near and far fields of the antenna, even taking into account the propagation phase delay between the received signal and its current source.

The newly initiated to electromagnetics may feel overburdened with concepts already, and this feeling is not unjustified, given the complexity of vector mathematics and mastery of concepts necessary to reach this point. It is a little known fact that Maxwell's unified field theory originally was expressed in about 20 equations and included potential functions. Oliver Heaviside eliminated the potential functions and reduced the equations to the now well-known 4 equations [3]. Heaviside was a brilliant, self-taught mathematician and physicist. It was Heaviside who formulated the telegrapher equations that took distributed inductance of transmission lines into account for the first time. Earlier analyses of transmission lines only included resistance and capacitance. If he had not had such an acerbic personality, it is likely that we would be referring to Heaviside's equations instead of Maxwell's. It is therefore noteworthy that no less a theoretician than the brilliant Heaviside is said to have hated the notion of potential functions.

Indeed, few practicing microwave engineers are comfortable enough with the vector potential to employ it. Nevertheless, the vector potential is a concept not much more difficult to apply than the scalar potential already studied. Its format as a vector field has the same form as the distributed current density that produces the $\vec{B}$ field, so no new vector visualization is required.

We begin with a general definition of the vector potential as a three-dimensional vector field, $\vec{A}$, whose curl is equal to the $\vec{B}$ field. Since the divergence of the $\vec{B}$ field is zero in all cases, the divergence of $\vec{A}$ can be defined subsequently in a manner that relates $\vec{A}$ to the time-varying $\vec{E}$ field. In this way the suitably defined $\vec{A}$ allows the determination of both the $\vec{E}$ and $\vec{B}$ fields using only differential processes, not integrals. To this end we define $\vec{A}$ to satisfy

$$\vec{B} \equiv \nabla \times \vec{A} \tag{7.29-1}$$

Then applying Maxwell's third equation, $\nabla \times \vec{E} = -\partial\vec{B}/\partial t$, we can rewrite (7.29-1) as

$$\nabla \times \vec{E} = -\frac{\partial(\nabla \times \vec{A})}{\partial t}$$

$$\nabla \times \left[\vec{E} + \frac{\partial\vec{A}}{\partial t}\right] = 0 \tag{7.29-2}$$

since the time derivative and curl can be taken in any order. The curl of the quantity in brackets is zero, therefore the quantity can be defined as the gradient of a scalar function, Φ, more precisely, the negative of the gradient in order to be consistent with the earlier definition of the scalar potential Φ in Section 7.8. Thus,

$$\vec{E} + \frac{\partial\vec{A}}{\partial t} \equiv -\nabla\Phi \tag{7.29-3}$$

Equivalently,

$$\vec{E} = -\nabla\Phi - \frac{\partial\vec{A}}{\partial t} \tag{7.29-4}$$

The *dynamic scalar potential* Φ defined in (7.29-3) is similar to the static potential function, but they are not the same. The static scalar potential applies to fixed charges. In this section Φ is a *dynamic scalar potential* and applies to moving charges. Generally, no confusion results between the two cases since a given application usually involves either static or moving charge. For example, in propagation situations the presence of fixed charge is of no consequence since the DC fields so created do not propagate. Furthermore, in the event that both static and dynamic charge distributions are of interest, their resulting fields can be added together by superposition if the medium is linear.

Although no assumptions have been made about the medium thus far, potential functions are most useful in linear, isotropic (same behavior in all directions), and homogeneous (uniformity throughout space) media. Accord-

ingly, in using potential functions we assume that μ and ε are scalars of constant value throughout the volume of analysis.

With (7.29-1) and (7.29-4) the $\vec{B}$ and $\vec{E}$ fields are completely specified. Next, let us find expressions for Φ and $\vec{A}$ that will allow their determination from their primary sources, moving charges, or currents. To this end substitute (7.29-4) into Maxwell's first equation, $\nabla \cdot \vec{E} = \rho/\varepsilon$, to obtain

$$-\nabla^2\Phi - \frac{\partial}{\partial t}(\nabla \cdot \vec{A}) = \frac{\rho}{\varepsilon} \tag{7.29-5}$$

Applying Maxwell's fourth equation, $\nabla \times \vec{H} = \vec{J_C} + \partial\vec{D}/\partial t$, and noting that $\nabla \times \vec{A} = \mu\vec{H}$,

$$\nabla \times \nabla \times \vec{A} = \mu\vec{J_C} + \mu\varepsilon\left[-\nabla\left(\frac{\partial\Phi}{\partial t}\right) - \frac{\partial^2\vec{A}}{\partial t^2}\right] \tag{7.29-6}$$

Applying the vector identity $\nabla \times \nabla \times \vec{A} = \nabla(\nabla \cdot \vec{A}) - \nabla^2\vec{A}$ to (7.29-6) gives

$$\nabla(\nabla \cdot \vec{A}) - \nabla^2\vec{A} = \mu\vec{J_C} - \mu\varepsilon\nabla\left(\frac{\partial\Phi}{\partial t}\right) - \mu\varepsilon\frac{\partial^2\vec{A}}{\partial t^2} \tag{7.29-7}$$

We have not used the option of defining the divergence of $\vec{A}$, an action that will serve to define $\vec{A}$ uniquely. Accordingly, we define the divergence to satisfy

$$\nabla \cdot \vec{A} \equiv -\mu\varepsilon\frac{\partial\Phi}{\partial t} \tag{7.29-8}$$

This allows (7.29-5) and (7.29-7) to be written as

$$\nabla^2\Phi - \mu\varepsilon\frac{\partial^2\Phi}{\partial t^2} = -\frac{\rho}{\varepsilon} \tag{7.29-9}$$

$$\nabla^2\vec{A} - \mu\varepsilon\frac{\partial^2\vec{A}}{\partial t^2} = -\mu\vec{J_C} \tag{7.29-10}$$

The expressions of (7.29-9) and (7.29-10) are the basis for the integral expressions by which Φ and $\vec{A}$ are evaluated. Once Φ and $\vec{A}$ are evaluated, $\vec{B}$ and $\vec{E}$ can be derived from the scalar and vector potentials using derivative processes, the vector curl being a derivative operation. These results are summarized in (7.29-11) and (7.29-12):

$$\vec{B} = \mu\vec{H} = \nabla \times \vec{A} \tag{7.29-11}$$

$$\vec{E} = -\nabla\Phi - \frac{\partial\vec{A}}{\partial t} \tag{7.29-12}$$

The dynamic scalar potential is evaluated using an integral expression similar to that employed with the static scalar potential, namely

$$\Phi = \int_V \frac{\rho \, dV}{4\pi\varepsilon r} \qquad (7.29\text{-}13)$$

Using similar reasoning, the vector potential *for the static or steady-state cases* can be evaluated using a volume integral of its current density sources:

$$\vec{A} = \mu \int_V \frac{\overrightarrow{J_C} \, dV}{4\pi r} \qquad (7.29\text{-}14)$$

In both (7.29-1) and (7.29-2) the volume V includes all of the respective mobile charge density ρ and current density $\overrightarrow{J_C}$. Even so, these two equations can only be used for steady-state not transient time variation cases because they do not recognize the propagation delay between the sources and the points at which the potentials are evaluated. This limitation is accommodated using the *retarded potentials* discussed in the next section.

7.30   RETARDED POTENTIALS

The corresponding solutions for Φ and $\vec{A}$ in *transient time-varying cases* must take propagation delays into account. This is done by replacing (7.29-13) and (7.29-14) with (7.30-1) and (7.30-2), respectively [6, p. 199]:

$$\Phi = \int_V \frac{[\rho]_{t-(r/v)} \, dV}{4\pi\varepsilon r} \qquad (7.30\text{-}1)$$

$$\vec{A} = \mu \int_V \frac{[\overrightarrow{J_C}]_{t-(r/v)} \, dV}{4\pi r} \qquad (7.30\text{-}2)$$

in which $v = 1/\sqrt{\mu\varepsilon}$; $v = c \approx 3 \times 10^8$ m/s in free space. For example, in (7.30-1) the quantity Φ to be evaluated at point P that is a distance r from charge ρ is to use the value of ρ present at time $t - (r/v)$. In this way the time necessary for the effect of ρ to propagate to P is taken into account. The effect travels at velocity $v = 1/\sqrt{\mu\varepsilon}$, the velocity of the plane wave evaluated in 7.22. The same provision for propagation is made in the evaluation of $\vec{A}$ using (7.30-2).

Because of this retardation effect, the potentials evaluated by (7.30-1) and (7.30-2) are called *retarded potentials* because their arrival is retarded by the time of propagation.

7.31 POTENTIAL FUNCTIONS IN THE SINUSOIDAL CASE

For the case of sinusoidal excitations, one is usually interested in the steady-state values of the fields. Therefore, the propagation time from the source to a point a distance r away can be fully accounted for by a phase delay kr, where $k = \omega/v$ is the propagation constant. Given this assumption, the following analysis shows that the expression for $\vec{E}$ in (7.29-12) can be written exclusively in terms of $\vec{A}$. To derive this new expression take the gradient of both sides of (7.29-8) to obtain

$$\nabla(\nabla \cdot \vec{A}) = -\mu\varepsilon\frac{\partial}{\partial t}\nabla\Phi \tag{7.31-1}$$

Since time differentiation and the gradient can be taken in any order, and since $\partial/\partial t = j\omega$ for sinusoidal excitations,

$$\nabla\Phi = -\frac{\nabla(\nabla \cdot \vec{A})}{j\omega\mu\varepsilon} = \frac{j\omega}{k^2}\nabla(\nabla \cdot \vec{A}) \tag{7.31-2}$$

where $k = \omega\sqrt{\mu\varepsilon}$ is the propagation constant in the medium. Then, substituting (7.31-2) into (7.29-12), the expressions for $\vec{B}$ and $\vec{E}$ for the sinusoidal case become

$$\vec{B} = \nabla \times \vec{A} \tag{7.31-3}$$

$$\vec{E} = -j\omega\left[\frac{1}{k^2}\nabla(\nabla \cdot \vec{A}) + \vec{A}\right] \tag{7.31-4}$$

where $\vec{A}, \vec{B}$, and $\vec{E}$ are in phasor form. This result is very useful for antenna pattern analysis, as will be seen in the next section.

7.32 ANTENNAS

Short Straight Wire Antenna

Consider a short wire antenna that is excited by a sinusoidal current I_0. The current excites the base of the antenna creating a surrounding $\vec{H}$ field whose time rate of change induces an orthogonal $\vec{E}$ field. The pair of sinusoidally varying fields store energy as well as propagate it.

The simple geometry in Figure 7.32-1 belies the complexity of the resulting electromagnetic field analysis of the short wire antenna, as will be seen from the following analysis. However, the approach is straightforward, and its execution provides an example of the use of the vector potential $\vec{A}$ and the determination of $\vec{E}$ and $\vec{H}$ fields at an arbitrary distance from an antenna [6, Sec. 12.05].
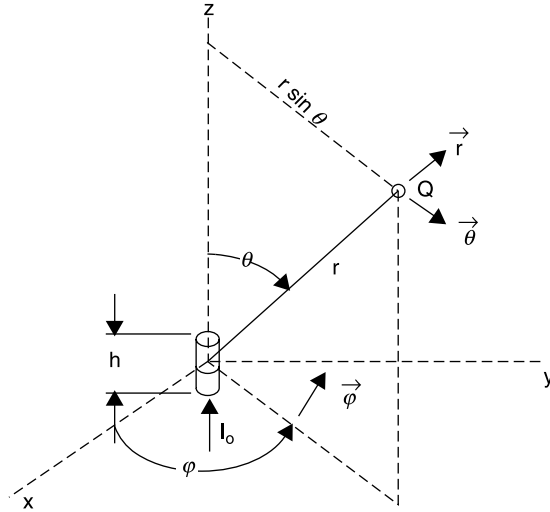
Figure 7.32-1 Geometry for analysis of the short straight wire antenna. [6, pg 497 Reprinted with permission]

The procedural convenience of using the vector potential can be appreciated in this analysis of the short straight wire antenna excited by a sinusoidal current phasor $\vec{I_0}$, for which the factor $e^{j\omega t}$ is implicit. The expression (7.30-2) for evaluating $\vec{A}$ indicates that each differential current contributes a part of $\vec{A}$ that is in the same direction as that current element. In other words, in this example of the short wire, since $\vec{I_0}$ points only in the z direction, so must $\vec{A}$. The fact that the vector potential has the same vector form as the currents that excite $\vec{E}$ and $\vec{H}$ fields underlies its usefulness.

For this analysis we assume that the length h and the diameter of the short straight wire antenna are very small compared to the operating wavelength, hence there is negligible change of phase in $\vec{I_0}$ over its length and diameter. The integration for $\vec{A}$ produces, for any point Q at radius r from the antenna, gives

$$\vec{A} = \vec{z}A_z \tag{7.32-1}$$

$$A_z = \mu \frac{hI_0}{4\pi r} e^{-j(\omega r/v)} \tag{7.32-2}$$

To evaluate $\vec{A}$ as a function of the radial distance r from the current source, it is best to express it in spherical coordinates as

$$A_r = A_z \cos\theta = \mu \frac{I_0 h}{4\pi r} e^{-jkr} \cos\theta \tag{7.32-3}$$

$$A_\theta = -A_z \sin\theta = -\mu \frac{I_0 h}{4\pi r} e^{-jkr} \sin\theta \tag{7.32-4}$$

where A_r and A_θ are the components of A_z in the $\vec{r}$ and $\vec{\theta}$ directions, respectively, and the medium propagation constant k is given by

$$k = \frac{\omega}{v} = \omega\sqrt{\mu\varepsilon} = \frac{2\pi}{\lambda} \tag{7.32-5}$$

Notice that, due to the symmetry of this case, all $\partial/\partial\varphi \equiv 0$, and there is no A_φ component. The utility of the vector potential can now be appreciated by observing that the integration needed to evaluate it results in a vector having the same direction as the current that produces it. To evaluate $\vec{H}$, note that when the curl of $\vec{A}$ is evaluated for this case in spherical coordinates only the φ component is nonzero. Thus, using (7.29-1)

$$\vec{H} = \frac{1}{\mu}\nabla \times \vec{A}$$

$$\vec{H} = \vec{H_\varphi} = \frac{1}{\mu}\frac{\vec{\varphi}}{r}\left[\frac{\partial}{\partial r}(rA_\theta) - \frac{\partial A_r}{\partial\theta}\right] = \frac{1}{\mu}\frac{\vec{\varphi}}{r}\frac{\mu h I_0}{4\pi}\left[-(-jk)e^{-jkr}\sin\theta + \frac{1}{r}e^{-jkr}\sin\theta\right]$$

$$\vec{H_\varphi} = \vec{\varphi}\frac{hI_0}{4\pi}e^{-jkr}\left[\frac{jk}{r} + \frac{1}{r^2}\right]\sin\theta \tag{7.32-6}$$

Next find $\vec{E}$ using (7.31-4) which is repeated below:

$$\bar{E} = -j\omega\left[\frac{1}{k^2}\nabla(\nabla\cdot\vec{A}) + \vec{A}\right] \tag{7.32-7}$$

To evaluate $\vec{E}$, first calculate $\nabla\cdot\vec{A}$, noting that $A_\varphi = 0$:

$$\nabla\cdot\vec{A} = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2 A_r) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta A_\theta)$$

$$\nabla\cdot\vec{A} = \frac{\mu I_0 h}{4\pi}\left[\frac{\cos\theta}{r^2}\frac{\partial}{\partial r}(re^{-jkr}) - \frac{e^{-jkr}}{r^2\sin\theta}\frac{\partial}{\partial\theta}(\sin^2\theta)\right] \tag{7.32-8}$$

$$\nabla\cdot\vec{A} = \frac{-\mu I_0 h}{4\pi}\left[\frac{jk}{r} + \frac{1}{r^2}\right]e^{-jkr}\cos\theta$$

Since $\nabla\cdot\vec{A}$ is not a function of φ,

$$\frac{1}{k^2}\nabla(\nabla\cdot\vec{A}) = \frac{1}{k^2}\left\{\vec{r}\frac{\partial(\nabla\cdot\vec{A})}{\partial r} + \vec{\theta}\frac{1}{r}\frac{\partial(\nabla\cdot\vec{A})}{\partial\theta}\right\}$$

$$= \frac{-\mu I_0 h}{k^2 4\pi}\left\{\vec{r}\cos\theta\frac{\partial}{\partial r}\left[\frac{jke^{-jkr}}{r} + \frac{e^{-jkr}}{r^2}\right]\right.$$

$$\left. + \vec{\theta}\frac{e^{-jkr}}{r}\left[\frac{jk}{r} + \frac{1}{r^2}\right]\frac{\partial(\cos\theta)}{\partial\theta}\right\} \tag{7.32-9}$$

Performing the indicated steps and simplifying gives

$$\frac{1}{k^2} \nabla(\nabla \cdot \vec{A}) = \frac{-\mu I_0 h}{k^2 4\pi} \left\{ \vec{r} \cos\theta e^{-jkr} \left[\frac{k^2}{r} - \frac{2jk}{r^2} - \frac{2}{r^3} \right] - \vec{\theta} e^{-jkr} \left[\frac{jk}{r^2} + \frac{1}{r^3} \right] \sin\theta \right\}$$

(7.32-10)

Next substitute (7.32-10), (7.32-3), and (7.32-4) into (7.32-7) to get (7.32-11), (7.32-12), and (7.32-13):

$$E_r = j\omega \frac{\mu I_0 h}{k^2 4\pi} e^{-jkr} \left\{ \cos\theta \left[\frac{k^2}{r} - \frac{2jk}{r^2} - \frac{2}{r^3} \right] - \frac{k^2 \cos\theta}{r} \right\}$$

$$= -j\omega \frac{\mu I_0 h}{k^2 4\pi} e^{-jkr} \left[\frac{2jk}{r^2} + \frac{2}{r^3} \right] \cos\theta$$

(7.32-11)

Recognizing that $k^2 = \omega^2 \mu \varepsilon$ and $\eta = \sqrt{\mu/\varepsilon}$, (7.32-11) can be written

$$E_r = \frac{I_0 h}{4\pi} e^{-jkr} \left(\frac{2\eta}{r^2} - \frac{2j}{\omega \varepsilon r^3} \right) \cos\theta$$

(7.32-12)

Similarly,

$$E_\theta = \frac{I_0 h}{4\pi} e^{-jkr} \left[\frac{\eta}{r^2} + \frac{j\omega\mu}{r} + \frac{-j}{\omega \varepsilon r^3} \right] \sin\theta$$

(7.32-13)

Again noting that $\vec{A}$ has no φ component and no φ variation, the $\vec{H}$ field is obtained as

$$\vec{H} = \frac{1}{\mu}(\nabla \times \vec{A}) = \frac{1}{\mu} \vec{\varphi} \left[\frac{1}{r} \frac{\partial}{\partial r} (rA_\theta) - \frac{1}{r} \frac{\partial A_r}{\partial \theta} \right]$$

$$H_\varphi = \frac{I_0 h}{4\pi} e^{-jkr} \left[\frac{jk}{r} + \frac{1}{r^2} \right] \sin\theta$$

(7.32-14)

Certainly, the development of expressions (7.32-12), (7.32-13), and (7.32-14), while straightforward, is mathematically tedious. But once expressed, they reveal the field behavior of the short antenna quite handily. First, the terms in $1/r^2$ and $1/r^3$ fall off rapidly for large r, leaving the $1/r$ terms.

The $1/r$ terms of E_θ and H_φ are in phase with and orthogonal to each other, hence they give peak propagating power density equal to Poynting's vector $\vec{P}_D$. Their ratio, $E_\theta/H_\varphi = \eta$, is the same as that for the plane propagating wave analyzed in Section 7.18. Taking their directions into account, $\vec{\theta} \times \vec{\varphi} = \vec{r}$, the power propagates in the $\vec{r}$ direction, away from the antenna. The peak propagating power density is given by

$$\vec{P}_D = \vec{E} \times \vec{H}^* = \vec{r} E_\theta H_\varphi = \vec{r}\eta |H_\varphi|^2 \text{ watts/meter}^2$$

(7.32-15)

where $\eta \approx 377\ \Omega$ for free space. Since peak values are used for the $\vec{E}$ and $\vec{H}$ fields, the time-average power flow is equal to one half the peak value of $\vec{P}_D$, or

$$\vec{P}_D(r) = \vec{r}\frac{1}{2}\eta\left(\frac{kI_0h}{4\pi r}\sin\theta\right)^2 \text{ watts/meter}^2 \qquad (7.32\text{-}16)$$

where h and r are in meters and k is in $(\text{meters})^{-1}$. The total propagating power across a spherical surface of radius r is

$$P = \int_S \vec{P}_D \cdot \vec{dS} = \int_0^{2\pi}\int_0^{\pi} P_D r^2 \sin\theta\, d\theta\, d\varphi$$

$$P = \pi\eta\left(\frac{kI_0h}{4\pi}\right)^2 \int_0^{\pi} \sin^3\theta\, d\theta \qquad (7.32\text{-}17)$$

Since $\int_0^{\pi} \sin^3\theta\, d\theta = \frac{4}{3}$, and for free space $\eta \approx 120\pi\ \Omega$, and $k = 2\pi/\lambda_0$, (17) reduces to

$$P = 40\pi^2 I_0^2 \left(\frac{h}{\lambda_0}\right)^2 \text{ watts} \qquad (7.32\text{-}18)$$

which is the total power radiated out of a sphere of radius r. Note that while P is independent of r, the power density P_D is inversely proportional to r^2.

Radiation Resistance

The radiation resistance of an antenna is defined as the resistance that would dissipate the same power as is radiated by the antenna when supplied with the same current. For the short, straight wire antenna radiating into free space

$$\frac{1}{2}I_0^2 R_R = 40\pi^2 I_0^2 \left(\frac{h}{\lambda_0}\right)^2$$

$$R_R = 80\pi^2 \left(\frac{h}{\lambda_0}\right)^2 \text{ ohms} \qquad (7.32\text{-}19)$$

For example, a 1-in.-long antenna at 1 GHz would have a radiation resistance of

$$R_R = 80\pi^2 \left(\frac{1\text{ in.}}{11.8\text{ in.}}\right)^2 \Omega = 5.7\ \Omega \qquad (7.32\text{-}20)$$

By itself, 5.7 Ω would be a large mismatch on a 50-Ω feed line; however, the mismatch is exacerbated further by the fact that there is a large reactance in series with the radiation resistance. Calculation of the reactance is more diffi-
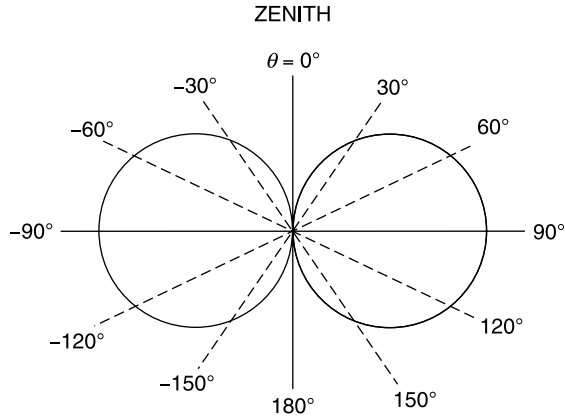
Figure 7.32-2   Radiation pattern of a dipole antenna.

cult because it is a volume integral over all space. One cannot use the integral limits $r = 0, \infty$ because the fields have a singularity at $r = 0$. Rather the integration must extend from the actual surface of the antenna to $r = \infty$. While radiation resistance is not a sensitive function of the diameter of the wire antenna, the reactance is very sensitive to the diameter because very thin wires have much stronger surrounding fields.

Radiation Pattern

The short wire antenna just analyzed (Fig. 7.32-1) is a dipole because the "horizon" plane bisects its vertical axis. Since, by assumption, there is no ground plane nearby, it radiates equally above and below the horizontal plane $(\theta = 90°)$, and the $\vec{E}_\theta$ and $\vec{H}_\varphi$ fields both vary as sin θ. Thus, the radiated power is maximum in the horizontal plane and is zero in the vertical direction, or zenith, as shown in Figure 7.32-2.

   The radiation is maximum at right angles to the wire antenna in the azimuth direction (on the horizon). The pattern is circularly symmetric about the antenna's axis. The pattern shown in Figure 7.32-2 is an *elevation plane cut*. This pattern applies when the antenna is located well above the earth, which would serve as a *ground plane*. In three dimensions the pattern resembles a bagel with a very small center hole (Fig. 7.32-3).

Half-Wavelength Dipole

Since the radiation resistance for a short monopole is very low, most wire antennas are made longer, but then the "short" approximation no longer holds, and a new analysis must be made, this time assuming a current distribution
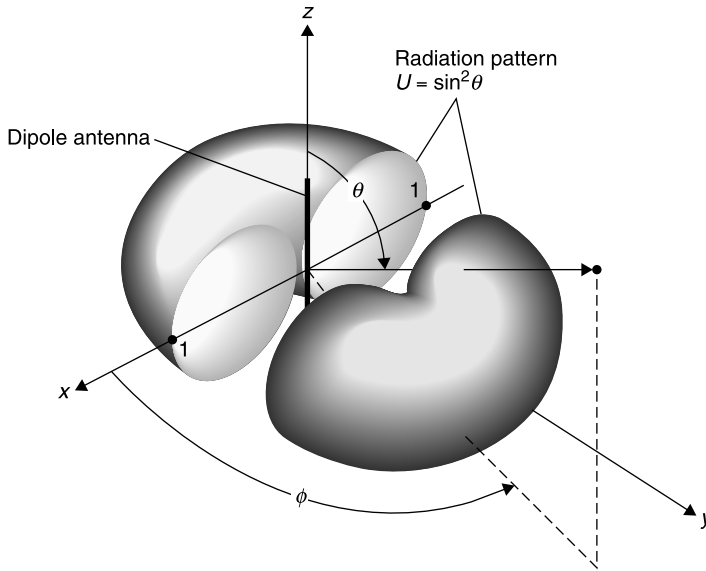
Figure 7.32-3 Sketch of three-dimensional radiation pattern of short wire dipole antenna [after Balanis, 15, Fig. 4.3j, with permission].

which is sinusoidal and goes to zero at the open ends. For the half-wave dipole, for which $l = \lambda/4$ (Fig. 7.32-4), the electric field $\vec{E}_\theta$ is found by integrating the field contributions from infinitesimal lengths of straight antennas of the type just covered and assuming a sinusoidal current distribution with the maximum I_m at the dipole's center feed point. The result of this analysis [6, Sec. 12.7] is that for the center-fed half-wavelength dipole antenna,
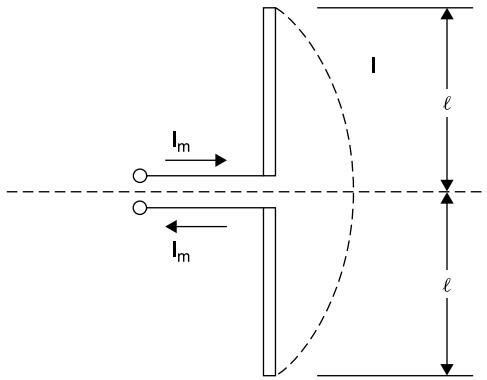
Figure 7.32-4 Center-fed dipole antenna.

$$|E_\theta| = \frac{60 I_m}{r} \left[\frac{\cos[(\pi/2) \cos \theta]}{\sin \theta} \right] \text{ volts/meter} \qquad (7.32\text{-}21)$$

$$P_D(r) = \frac{15 I_m^2}{\pi r^2} \left[\frac{\cos[(\pi/2) \cos \theta]}{\sin \theta} \right]^2 \text{ watts/meter}^2 \qquad (7.32\text{-}22)$$

$$R_R = 73.09 \ \Omega \qquad (7.32\text{-}23)$$

where I_m is the maximum current, presumed equal to the value at the center feed point. The driving point impedance is nearly equal to the radiation resistance of 73 Ω since the antenna is nearly resonant when its length is $\lambda/2$. Interestingly, the half-wavelength dipole and the short wire antenna have nearly the same radiation patterns [6, Secs. 12.05 to 12.07]. However, the short wire antenna is much more difficult to drive because of its low impedance and high reactance.

A practical problem arises when it is necessary to interconnect the half-wave dipole antenna to a transmitter whose output port is an unbalanced transmission connector or transmission line, such as coax. The antenna requires a balanced feed at terminals AB in Figure 7.32-5. That is, both of the AB terminals are to be separated from ground by the same impedance. A coaxial output connector or a coaxial cable used to interconnect transmitter and antenna is an unbalanced transmission system. However, this problem is easily corrected by using a *balun*. On transmission, the balun converts the unbalanced line from the transmitter to the balanced format of the antenna, as shown in Figure 7.32-5. It does this by creating an infinite impedance to ground at point C by virtue of the quarter wavelength collar attached to the end of the coaxial line. The balun's operation is reciprocal. On receiving a signal, the balun converts the balanced format of the antenna to the unbalanced format of the receiver. The function of the balun is limited in bandwidth by the quarter wavelength requirement of the collar.
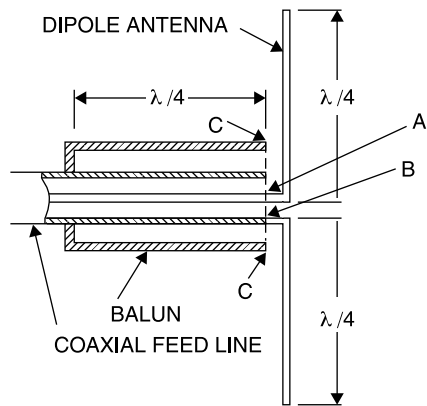
Figure 7.32-5   Balun used to connect a coaxial cable to a balanced, half-wavelength dipole antenna.

Antenna Gain

For a given radiated power an antenna that radiates more power in certain directions and less in others, as all practical antennas do, can place more power at a receiver than would an *isotropic radiator* that distributes the same input power uniformly over a spherical surface. Actually, it is impossible to build a perfectly isotropic radiator, but the concept is useful for defining the performance of practical antennas.

The power density delivered in a given direction by a practical antenna compared to what would be delivered by an isotropic radiator is called *antenna gain*. It is expressed in dBi, for decibels above (or below, if negative) the level that would be achieved with an isotropic radiator. Gain can be defined in any direction, but customarily, unless otherwise noted, it is defined for the direction in which the antenna radiates the greatest power density. For the short wire antenna the maximum radiated power density occurs on the horizon ($\sin^2 \theta = 1$) and has the value

$$P_{D,\max} = \frac{1}{2}\eta|H_\varphi|^2 = \frac{1}{2}\eta\left(\frac{kI_0 h}{4\pi r}\right)^2$$

where H_φ is the radiated component of the H field at r. Then, since $k = 2\pi/\lambda_0$ and $\eta \approx 120\pi$ Ω, this reduces to

$$P_{D,\max} \approx 15\pi\left(\frac{h}{\lambda_0}\right)^2\frac{I_0^2}{r^2} \qquad (7.32\text{-}24)$$

Since the surface area of a sphere at radius r is equal to $4\pi r^2$, an isotropic radiator produces a power density at radius r of

$$P_{D,i} = \frac{1}{4\pi r^2}\frac{I_0^2}{2}R_R = \frac{1}{8\pi r^2}I_0^2 80\pi^2\left(\frac{h}{\lambda_0}\right)^2 = 10\pi\frac{I_0^2}{r^2}\left(\frac{h}{\lambda_0}\right)^2 \quad (7.32\text{-}25)$$

Dividing (7.32-24) by (7.32-25), the gain of the short dipole antenna is [6, p. 504]

$$G_{\text{short dipole}} = \frac{P_{D,\max}}{P_{D,i}} = 1.5 = 1.76 \text{ dB} \qquad (7.32\text{-}26)$$

Using the same method, the gain of the half-wavelength dipole antenna is

$$G_{\text{half wavelength dipole}} = 1.64 = 2.14 \text{ dB} \qquad (7.32\text{-}27)$$

It should be noted, however, that although the short dipole and the half-wavelength dipole have nearly identical radiation patterns and gains, the half-

wavelength dipole is much easier to excite due to its higher radiation resistance and its near zero reactance.

Antenna Effective Area

In calculating the overall loss between transmitter and receiver, we need to know the *effective area* A_E of the receiving antenna. Aperture antennas such as dishes and waveguide horns have an effective signal capture area that is approximately equal to their physical area in the direction of the received or transmitted signal, but even the small wire antenna, whose physical area is infinitesimal, has an appreciable effective area. This effective area is related to its gain.

The fact that receiving antennas have varying gains and hence capture relatively more or less power is treated by considering that the power they capture from the radiation field is proportional to their respective effective areas [6, p. 560, and 14, Sec. 11.12]. But the received power is also proportional to the gain of the receiving antenna. Therefore, it follows that there is a proportionality constant relating gain G to effective area A_E. *Once this factor is found for any one receiving antenna, the same factor can be applied to all receiving antennas.* Furthermore, the network between transmitter and receiver terminals, including the free-space propagation medium, is linear and reciprocal. Accordingly, what applies to all transmitting antennas, by reciprocity, applies to all receiving antennas.

To determine the proportionality factor between gain and effective area, we use as a receiving antenna the short wire dipole previously evaluated. We define that the power delivered to the receiver terminals by the receiving antenna is given by

$$P_{\text{rec}} \equiv P_D A_E \tag{7.32-28}$$

where P_{rec} is the power delivered to the receiver connected to the receiving antenna, P_D is the power density (in watts/meter^2) of the wave incident on the receiving antenna and A_E is the effective area of the receiving antenna.

In the vicinity of the receiving antenna, P_D is related to the rms electric field strength E parallel to the antenna by

$$P_D = \frac{E^2}{\eta} = \frac{E^2}{120\pi} \tag{7.32-29}$$

where η is the impedance of free space. The available power absorbed in a matched load connected to the receiving antenna is given by

$$P_{\text{rec}} = \frac{V_{\text{OC}}^2}{4R_R} = \frac{(Eh)^2}{4R_R} \tag{7.32-30}$$

where R_R is the radiation resistance of the receiving antenna given by (7.32-19).

It might seem that the factor of 4 should not appear in the denominator in (7.32-30), that instead the full voltage Eh delivers power to the resistance R_R; but this is not so. The transmitter, its antenna, the free-space propagation path, and the receiving antenna taken together represent a linear network. The voltage Eh induced at the receiving antenna terminals is an "open-circuit voltage" *that exists in the absence of the receiving antenna or when the receiving antenna terminals are open circuited*, that is, when the incident radiation is not "loaded." When the R_R matched load is applied, the voltage at the receiving antenna terminals drops to half its open-circuit value, and the power delivered to the load at the receiver terminals is that given by (7.32-30). Substituting the value for R_R from (7.32-19) into (7.32-30) gives

$$P_{\text{rec}} = \frac{(Eh)^2}{4} \frac{\lambda_0^2}{80\pi^2 h^2} = E^2 \frac{\lambda_0^2}{320\pi^2} \qquad (7.32\text{-}31)$$

Solving (7.32-28) for A_E and substituting the values for P_D from (7.32-29) and P_{rec} from (7.32-31) into it,

$$A_E = \frac{P_{\text{rec}}}{P_D} = \frac{E^2 \lambda_0^2}{320\pi^2} \frac{120\pi}{E^2} = \frac{3}{8} \frac{\lambda_0^2}{\pi} = \frac{\lambda_0^2}{4\pi} \frac{3}{2} \qquad (7.32\text{-}32)$$

But from (7.32-26) the gain of the short dipole antenna is $G = 1.5 = \frac{3}{2}$, therefore (7.32-32) can be written

$$A_E = \frac{\lambda_0^2}{4\pi} G \qquad (7.32\text{-}33)$$

Monopole Antenna

The fact that the dipole antenna radiates into the lower hemisphere as well as the upper hemisphere is a disadvantage for terrestrial communications since most receptors are on a line with the horizon or above it. Consequently, the *monopole antenna* with ground (Fig. 7.32-6) has the advantage of an additional 3 dB of gain compared to the dipole because it radiates all of its power into only the upper hemisphere. The monopole antenna is often called a *whip antenna* because it was often used for very high frequency (VHF) police radios and connected to the rear bumper of an automobile. The "whip action" resulted as the car accelerated and decelerated. Due to its symmetry with the half-wavelength dipole, its driving point impedance is one half that of the dipole.

For the quarter-wave monopole antenna

$$R_r = 36.5 \ \Omega \qquad (7.32\text{-}34)$$

$$G_{\text{max}} = 3.28 = 5.16 \text{ dB} \qquad (7.32\text{-}35)$$
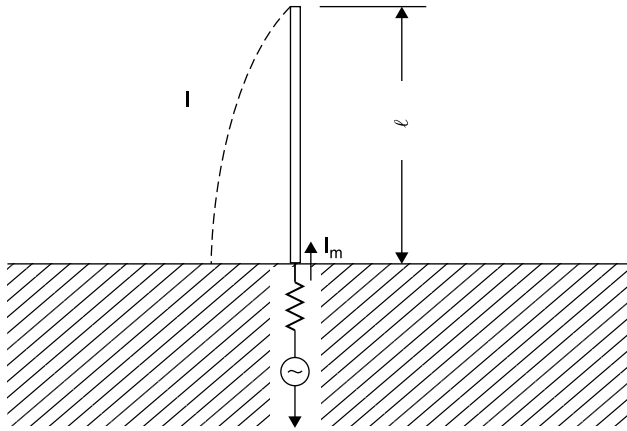
Figure 7.32-6   Monopole antenna with ground plane.

It should be noted that *the presence of a good ground plane is as important to achieving this gain improvement as the antenna*. For example, with a small hand-held monopole, the radiation pattern is likely to be that of the short wire antenna, radiating both above and below the horizon plane that contains the antenna, unless that plane coincides with a good earth ground plane at the frequency of operation. Large AM broadcast antennas are examples of monopole antenna applications. Their monopole function is usually ensured by installing *ground radials* at the base of the antenna, consisting of wires that run radially in all directions from the base of the antenna. Their lengths in the earth are comparable to the antenna's height. The conventional automobile antenna consisting of a straight wire mast also may approximate the monopole, achieving conducting ground plane by means of the sheet metal of the automobile near the aerial's base.

Aperture Antennas

The larger the area of an equiphase front created by an antenna the less the propagating beam, called the *main beam*, radiated by it will diverge, hence the greater its ''gain'' in its preferred propagation direction, and the more power it can place on a target receptor compared to an isotropic radiator transmitting the same power. By reciprocity, the same comments apply to the receiving antenna.

It was shown that the effective area of an antenna is directly proportional to its maximum gain in what is usually termed the *main beam* or *bore site* direction, the direction orthogonal to the equiphase plane of the antenna. The term *bore site* derives from the practice of installing a small telescope on high gain antennas as an aid to aiming them in the desired direction. For most aperture
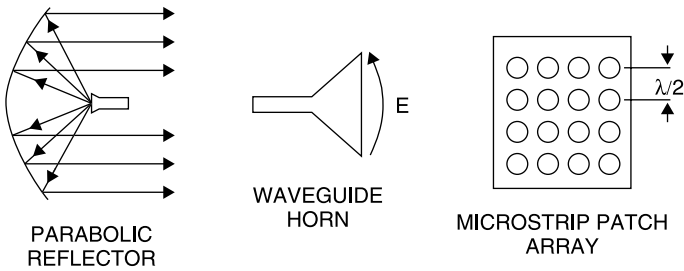
Figure 7.32-7   Aperture antennas.

antennas (Fig. 7.32-7) the area of the aperture would be approximately the effective area if the antenna were *uniformly illuminated,* that is, if it radiated a uniform power density over its entire aperture.

However, uniform illumination is often undesirable because it results in an abrupt discontinuity in the $\vec{E}$ and $\vec{H}$ fields near the edges of the antenna. This in turn results in auxiliary radiation patterns, called *sidelobes* that are separate from and radiate in different directions than that of the main beam. Sidelobes waste power and therefore correspond to signal loss. Often they are even more objectionable because they may intercept spurious signals that interfere with the desired reception. To minimize sidelobe strengths, an *illumination taper* is employed whereby the field strength is maximum near the center of the aperture and diminishes near the periphery.

Since the illumination taper requires that not all of the aperture is energized to the fullest, the result is that the effective area of the antenna is correspondingly less than its physical cross-sectional aperture, A_{CS} (the area projected in the propagation direction). In fact, any reduction in the signal strength over the aperture reduces the antenna's effective area, and with it the main beam gain. Nevertheless, this may be more desirable than encountering interfering signals through sidelobes. Other causes of aperture reduction include dielectric covers to protect the antenna, called *radomes,* surface irregularities such as roughness or actual holes made in the antenna surface to reduce its weight or wind resistance and structural supports of the feed horn.

Equation (7.32-33) can be rewritten as

$$G = \frac{4\pi}{\lambda_0^2} A_E \qquad (7.32\text{-}36)$$

To account for an effective area A_E less than the actual cross-sectional area A_{CS} of the antenna, a factor to relate them is defined according to

$$A_E = \varepsilon_{AP} A_{CS} \qquad (7.32\text{-}37)$$

where ε_{AP} is the *aperture efficiency* and $0 \leq \varepsilon_{AP} \leq 1$ [15, Sec. 12.5.3]. As already noted, the aperture efficiency can be less than unity for a variety of reasons.

Phased Arrays

Antennas used for point-to-point communications are fixed in position. These include terrestrial communication link applications and Earth to synchronous-satellite communications. Those used for scanning generally are moved mechanically to cover a given surveillance area, as for radio ranging and detection, or *radar*, applications. However, antennas that can scan electronically without need of mechanical motion have appeal in applications for which mechanical slewing is either too slow or impractical, as in some satellite-borne applications or for tracking rapidly moving radar targets such as missiles. For these applications the electronically scanned *phased array antenna* may be employed (Fig. 7.32-8).

The phased array antenna uses small-aperture radiating elements that, because of their small area, necessarily and desirably have a wide radiation pat-
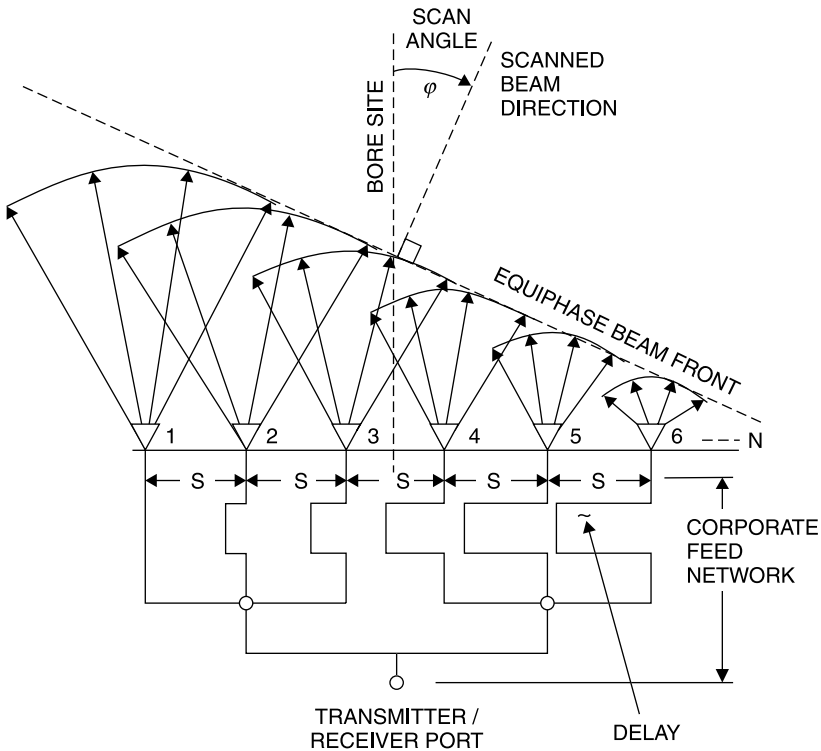
Figure 7.32-8   Schematic diagram of a phased array antenna using elements with time delay to scan the radiated beam.

tern in the bore site direction of the array. A one-dimensional, linear array is shown in Figure 7.32-8 for illustration of the principles, but phased arrays more usually are two dimensional. Steering is accomplished by adjusting the phase of the signals applied to each element so that the aggregate equiphase *wave front* representing the summation of the fields of all of the elements is in a plane at an angle skewed from the bore-site direction. This is called the phased array *scan angle*. In Figure 7.32-8 the steering shown uses *time delay* rather than *phase shift*. When time delay is used, the scan angle is independent of frequency, and the antenna can be used with arbitrarily wide bandwidth, subject to the limits of the radiating elements. The amount of delay, τ_N (in wavelengths) required for the Nth linear element is given by

$$\tau_N = NS \sin \theta \qquad (7.32\text{-}38)$$

where S is the spacing between elements (in wavelengths) and θ is the scan angle. For two-dimensional scanning of an $M \times N$ array, *in azimuth and elevation*, (7.32-38) must be applied separately for both directions and the results totaled for the delay τ_{MN} to be applied to the MNth element. The element spacing is usually about one-half wavelength since radiating elements necessarily occupy about this minimum dimension on a side. Furthermore, wider spacing would give rise to *grating lobes*, spurious radiation beams for which the wider spacing permits redundant equiphase planes at large scan angles. Note that in Figure 7.32-8 the delay is shown as a fixed line length for purposes of illustration. Array scanning requires that variable delay be selectable. Also note that one wavelength of transmission line produces one period ($\omega t = 360°$, where $t = T$) of time delay.

For large phased array antennas whose apertures are many wavelengths across, the use of time-delay steering is impractical to apply to each element. For example, consider a square antenna 10 wavelengths on a side with half-wavelength spacing between elements in azimuth and elevation directions. This antenna would have 400 elements and from (7.32-33) up to 31 dB of gain in the bore site direction, neglecting losses and sidelobes and assuming an illumination efficiency of 100%. Phased arrays usually are limited to about $\pm45°$ steering. This limit is imposed by the beamwidth of the subelements as well as the fact that for larger steering angles the array's projected area would be too small reducing its gain too much. But to accomplish *45°* steering for this example array, the time delay required of the edge elements for each scanning direction would be, from (7.32-38), equal to 14.14λ. For two-dimensional scanning this is doubled and therefore a delay of 28.28λ would be required.

Time delay is usually arranged using electronic switches that provide discrete delay steps. Thus a given required delay is managed only approximately, and there is a delay error equal to as much as one half the smallest switched step, or *delay quantization*. To provide for continuous scan angle steering of this size antenna, a delay quantization of 0.125λ might be employed corresponding to a phase error of no more than $\pm0.0625\lambda$ or $\pm22.5°$, one half the smallest delay

step. Then time-delay devices capable of delay adjustment from about 0.125λ to 28λ in 0.125λ steps (corresponding to 45° phase steps) would be required. Thus a control device having steps of 0.125λ, 0.25λ, 0.5λ, 1.0λ, 2.0λ, 4.0λ, 8.0λ, and 16λ is needed. This would be termed an *8-bit time-delay device* providing control to 31.875λ in 0.125λ steps.

Such a device is difficult to build. These delay lines, being several wavelengths long at the center frequency, have multiple resonances within the operating bandwidth. This is because diodes are characterized by a high reactance (small capacitance) in their "off state." A long line with small capacitors at each end becomes a high-Q resonator. When the line is long enough, it is certain to have resonant frequencies in the operating band of the antenna, and at these resonances high absorption of the signal power occurs. When the switches do not turn off the long line completely. Even without the resonances, high insertion loss can be expected due to the long line lengths to be switched and the losses of a cascade of switches required to select them.

For these reasons, phase shift rather than time delay is usually applied to each element. The required phase shift is obtained by computing the time delay required at the center frequency of operation and subtracting integer multiples of 2π. For example, the most remote element of the 20λ by 20λ array for a 45° scan in *both azimuth and elevation* would require 28.28λ of delay. This could simulated using a phase shift corresponding to 0.28λ or 100°. In turn, the phase shifters also would be realized in binary form. A 4-bit phase shifter would provide 0° to 337.5°. In the phase domain 360° is equivalent to 0° of control in 22.5° steps. The nearest step to 100° is 90°, and this would be selected. The fact that phase shift control is only correct at the frequency for which the time-delay-to-phase-shift conversion is made means that the radiated antenna beam will *squint* off the intended scan angle as the frequency of operation departs from this center frequency.

Unlike mechanically scanned antennas, phased array antennas have reduced effective area, consequently less gain, as they are steered off the bore site direction because their projected area in the scanned direction is correspondingly less.

Because of these considerations and other complexities not described, large-aperture phased array antennas are employed only when their motionless, high-speed scanning properties are deemed essential. However, small apertures of a few wavelengths in diameter may prove practical in some communication applications.

7.33   PATH LOSS

Although Maxwell's equations are fundamental to radio engineering, their application requires an analytic description of the $\vec{E}$ and $\vec{H}$ fields, and this can be quite complex as was seen for the short wire antenna analysis. However, some propagation problems can be solved with a much simpler approach once the

principles of wave propagation are appreciated. One of these is the *path loss* calculation, which is the determination of how much power reaches a radio receiver from a distant transmitter. In fact, the path loss calculation is usually one of the first in the design of a radio communication system.

Previously we introduced the isotropic radiator assumption. The isotropic radiator is defined to be a 0 dB gain antenna, and the power radiated from it is called the *effective isotropic radiated power (EIRP)*. As was noted earlier, it is impossible to build an isotropic radiator, but the concept is useful for calculating how much power would be available at a distance were one to have been employed. Then it is a simple matter to adjust the isotropic result so obtained to that which applies given the relative gains of the actual antennas used for transmission and reception. *The isotropic antenna uniformly "paints" the surface of a sphere having a radius equal to the distance R between transmitter and receiver.*

Figure 7.33-1 depicts a transmitter at the center of a sphere of radius R. If the transmitter power P_T is radiated by an isotropic antenna, then, since the surface area of a sphere is $4\pi R^2$, the power density at R is $P_T/4\pi R^2$. If the isotropic antenna is replaced by a transmitting antenna of gain G_T, then the power density in the direction of its maximum intensity (the direction of its main beam) is given by

$$P_D = G_T \frac{P_T}{4\pi r^2} \qquad (7.33\text{-}1)$$

If a receiving antenna at distance R is oriented to accept maximum power from
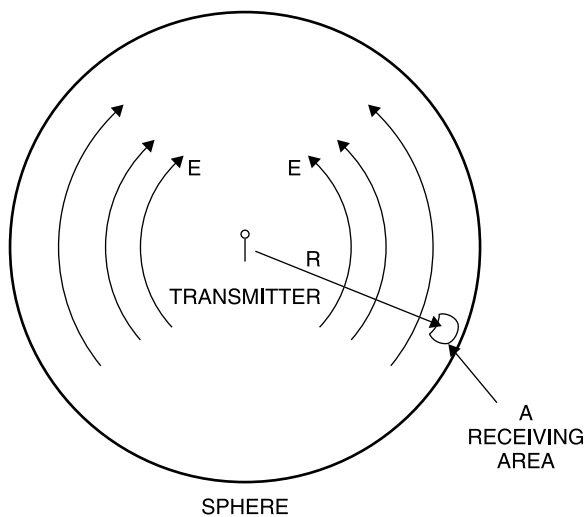
Figure 7.33-1 Effect of power reduction as a result of spreading out isotropically (painting an imaginary sphere) at a distance R from the transmitter.

this direction (it is aimed at the transmitter), the power it will receive is

$$P_R = A_E G_T \frac{P_T}{4\pi R^2} \tag{7.33-2}$$

where A_E is the effective capture area of the receiving antenna. Note that P_R is actually the available power at the terminals of the receiving antenna. Applying (7.32-33) this can be written

$$P_R = \left(\frac{\lambda_0}{4\pi R}\right)^2 G_T G_R P_T \tag{7.33-3}$$

Usually the radio system engineer is interested in determining the loss between the terminals of the transmitting antenna and those of the receiving antenna. This is called the *path loss*. For this purpose (7.33-3) can be cast in an insertion loss format as

$$\text{IL}_{\text{PATH}} = \frac{P_T}{P_R} = \left(\frac{4\pi R}{\lambda_0}\right)^2 \frac{1}{G_T G_R} = \frac{16\pi^2}{G_T G_R} \left(\frac{R}{\lambda_0}\right)^2 \tag{7.33-4}$$

Note that numeric values (not dB) must be used for G_T and G_R and that the same length units must be used for A_E, R, and λ_0 in (7.33-1) to (7.33-4).

As an example, consider the case of a 1-W, 900-MHz cellular telephone transmitter radiating its signal with a circularly polarized antenna having a gain factor of 1.64 above isotropic. The receiving monopole antenna on the exterior of an automobile is vertically polarized and has a gain factor of 3.28. The automobile is 3 miles from the transmitter, and there is a line-of-sight transmission path. What is the path loss between transmitter and receiver?

First note that since the receiving antenna is linearly polarized it can capture only half of the circularly polarized power incident upon it. To take this into account, the gain of the receiving antenna is reduced by one half to 1.64.

Substituting the system values into (7.33-4) yields

$$\text{IL}_{\text{PATH}} = \frac{16\pi^2}{G_T G_R} \left(\frac{R}{\lambda_0}\right)^2 = \frac{16\pi^2}{(1.64)(1.64)} \left(\frac{3 \times 1610}{0.33}\right)^2 = 1.23 \times 10^{10} = 101 \text{ dB}$$

since 1 mile = 1610 m and $\lambda_0 = 0.33$ m at 900 MHz. Note that since $P_T = 1$ W = 1000 mW = 30 dBm, $P_R = 30$ dBm − 101 dB = −71 dBm.

The path loss calculation is only a first step in the design of a radio communication link. Implicit in the path loss calculation is the assumption that the transmitter and receiver are on a line-of-sight with each other, that is, that there are no intervening obstacles. For real radio systems (Fig. 7.33-2), this is rarely the case. Buildings, foliage, hills, and other obstacles occupy the propagation path.
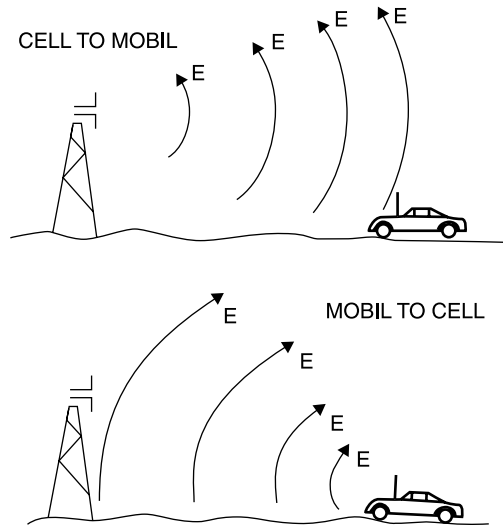
Figure 7.33-2   Cellular radio system depicting cell-to-mobile and mobile-to-cell transmissions. For illustration, a dipole antenna is depicted at the base station, but customarily a circularly polarized antenna is employed.

For a cellular telephone communication system there is almost never an unobstructed path. The transmitted signal may undergo several reflections on its roundabout path to the receiver, and the strongest signal reaching the receiver may actually be coming from a direction opposite that of the transmitting antenna. The resultant signal may arrive at the receiver as the phasor addition of several separate signals with random phases, resulting from multiple paths, a process called *multipathing*.

When this occurs, the signals combine with random phases at the receiver. Often, multiple signals result in a combined signal lower than any of its constituents, a process called *Rayleigh fading*. Rayleigh was a physicist who, among other accomplishments, formulated the distribution associated with random events. To overcome deep fades, a cellular base station may switch between different antennas (*diversity switching*) to create a more favorable addition of the multipath signals at the receiver. This can occur hundreds or thousands of times during a cellular conversation conducted between the cell site and a moving automobile containing the cellular system client.

In specifying a communication system, once the *minimum detectable signal level* is determined (more about this in Chapter 10), the minimum required transmitter power can be determined. However, a transmitted power well beyond that required by the path loss calculation is generally specified in order to create a *system margin* and a more robust communication link. *The system margin can overcome the obstacles to delivering a detectable signal much of the time, but never always*. The greater the system margin, the less the probability

that the communication link will prove to be unacceptable, and the greater the percentage of time that the link will operate satisfactorily. However, no matter how great the allowance, there will always be some probability, however small, that a sufficiently deep fade will be encountered to render the link unusable for some period of time or at all times in some locations. Cellular telephone users operating from moving vehicles experience *drop-outs* regularly and can attest to this fact. *Wireless communication links are never completely reliable*.

7.34   ELECTROMAGNETIC (EM) SIMULATION

When AC analysis is used to evaluate a component, we assume that it behaves according to its equivalent circuit model. However, this is an idealization. Real circuits have physical sizes that often are appreciable compared to the operating wavelength, causing a change of voltage and current over their dimensions. This is not included in the equivalent circuit used by a circuit simulator or in the manual application of AC analysis. The resulting circuit performance estimates are based on this idealization; consequently, they will contain some errors as a result. Sometimes the error is quite significant, as the following example demonstrates.

Suppose that we wish to employ a 25 Ω characteristic impedance, open-circuited shunt stub on a 50-Ω transmission line as a filter. The filter is to block the second harmonic of a 1920-MHz cellular system transmitter, so that the spurious requirements of its radio license will not be violated. Therefore, it is desired that the open-circuited stub presents a short circuit (or very low impedance) to the transmission line at 3840 MHz. The stub's length is to be chosen equal to 90° at 3840 MHz, resulting in a short circuit across the 50-Ω line and blocking the transmission of spurious power at 3840 MHz.

Using an ideal equivalent circuit, the open-circuited stub is modeled as shown in Figure 7.34-1, in which the 57.5° line lengths on either side of the stub correspond to 50-Ω line sections used later in the EM simulation modeling. These lengths have no effect on the stub's insertion loss performance. The calculated performance using a circuit simulator is shown in Figure 7.34-2, from
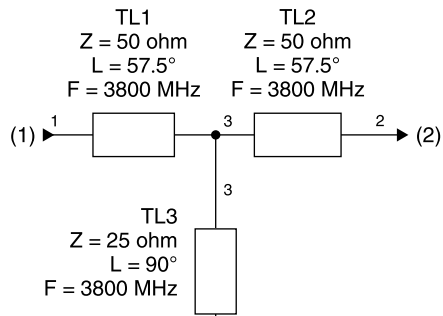
Figure 7.34-1   Ideal electrical schematic of an open-circuited, 25 Ω, shunt stub on a 50-Ω transmission line.
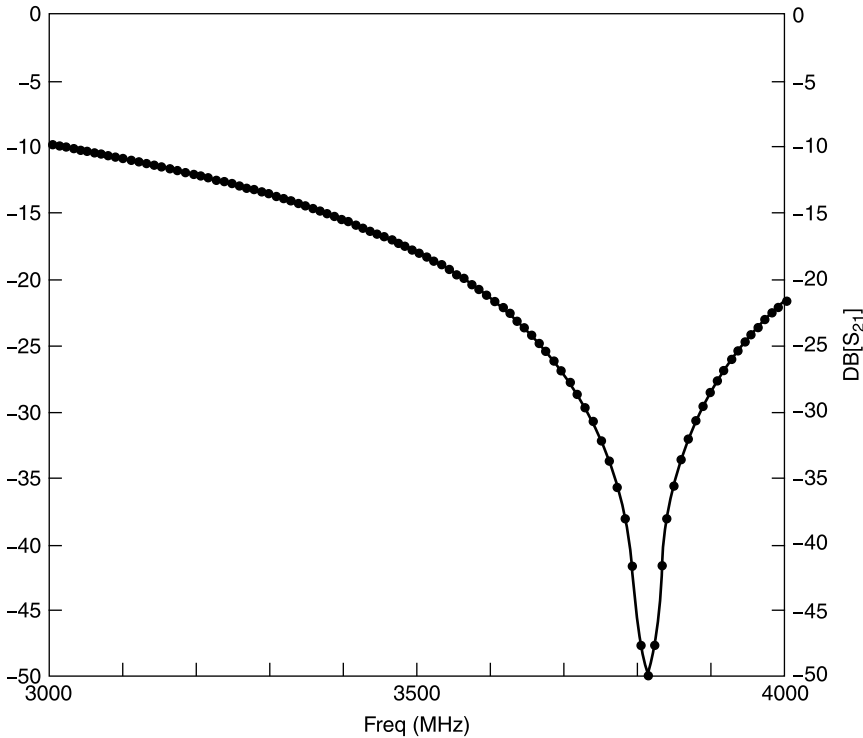
Figure 7.34-2 Calculated response of an open-circuited 25-Ω stub obtained using ideal electrical schematic.

which it is seen that 40 dB or more of isolation can be obtained over a 50-MHz bandwidth.

To build this circuit it is necessary to create a circuit pattern. The determination of the geometry is accomplished by determining the width of a 50-Ω transmission line and a 25-Ω stub. We also require determination of the wavelength on the stub at the operating frequency in the medium in which the stubs will be realized. Using the Genesys auxiliary program TLINE and selecting a design medium consisting of a microstrip line having a 0.031-thick substrate with dielectric constant 2.22, the wavelength at 3840 MHz (at which we want the stub to block transmission) is 2.165 in. (5.50 cm) and the width of a 25-Ω stub is 240 mils (0.240 in. or 61 mm) (Fig. 7.34-3).

The width of a 50-Ω line in this medium is 93 mils (0.093 in. or 2.36 mm). The wavelength at 3840 MHz is 2.5506 inches on the 50-Ω line, considerably different that the wavelength at the same frequency on the 25-Ω stub. This is a consequence of the nonhomogeneous nature of microstrip line. Were stripline used, the wavelength would be the same for both impedances.

When we attempt to draw the center conductor pattern layout for this circuit, some questions arise, such as: Where does the stub begin? Does it begin at
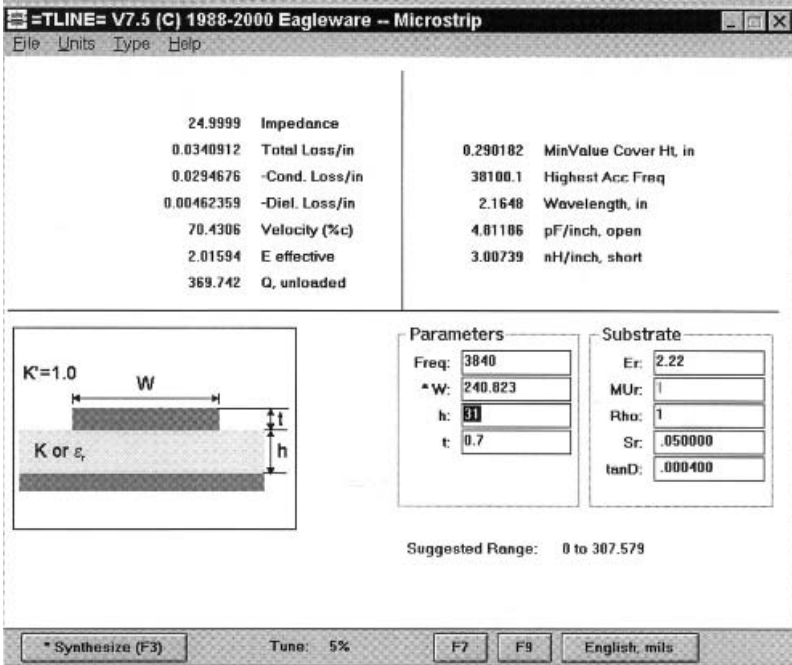
Figure 7.34-3 Calculated dimensions and wavelength for a 25-Ω microstrip line using 0.031-in.-thick substrate with relative dielectric of 2.22.

the edge of the 50-Ω main transmission line, or should one measure from the center line of the 50-Ω line? Should an allowance be made for the fringing electric field at the end of the stub, the effect of which would make the electrical length of the stub appear longer?

To answer these questions experimentally, as was the practice prior to the availability of EM simulation, it was customary to build and test the stub, usually with various geometries, and then to deduce design rules that could be used to approximate the behavior of future designs. The availability of *electro-magnetic simulators (EM simulators)* eliminates the need for time-consuming model building by permitting the experimental "construction and test of the circuit" on the computer.

The detail in Figure 7.34-4 represents dimensioning using the edge of the transmission line as the beginning of the stub and no length allowance for any fringing electric field at the open-circuited end of the stub. The 25-Ω stub impedance requires a 0.240-in.-wide line. At 3840 MHz the wavelength of 2.165 in. results in 0.541 in. for a quarter wavelength stub. Notice that the width of the stub is nearly half its length! The horizontal conductor at the top is a 50-Ω line and has a 0.093-in. width.

The EM *simulation grid* used in Figure 7.34-4 is 20 mils square. It is over this granularity that the $\vec{E}$ and $\vec{H}$ fields of adjacent cells, measuring $0.020 \times$
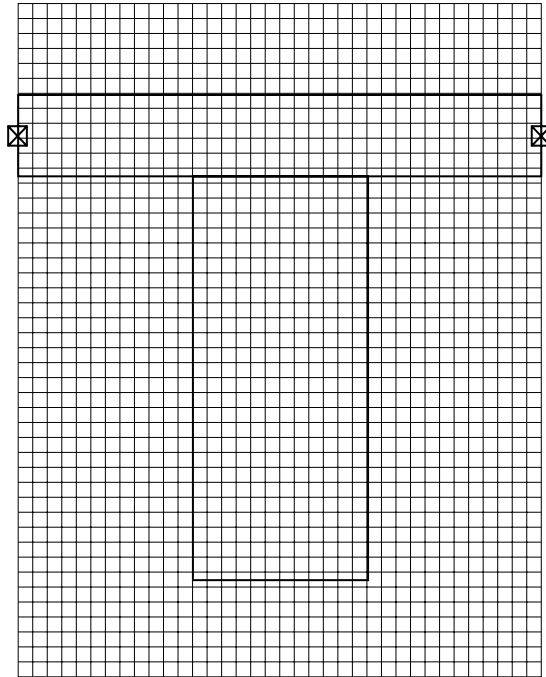
Figure 7.34-4 Layout of the 25-Ω, open-circuited shunt stub used for EM simulation. A theoretical perfectly conducting "box" containing the circuit measures 0.720 in. wide by 0.900 in. high is assumed in the simulation. The analysis also assumes a perfectly conducting box cover separated by 0.150 in. of air above the circuit. The *simulation grid* shown is 20-by-20 mils.

0.020 × 0.020 in. are forced to be continuous by the simulator. Basically, the simulation software begins by assuming starting values for the $\vec{E}$ and $\vec{H}$ fields, matching the boundary conditions that require that the tangential $\vec{E}$ field and the normal $\vec{H}$ field on conductors must be zero. Then the program determines the $\vec{E}$ and $\vec{H}$ fields within each 20 mil^3 cell and checks the continuity of the fields on the cell boundaries from cell to cell.

Usually the initially assumed values for the fields are in error, and the computer must guess a modified set of field values in all of the cells and retry the solution. This process is repeated until the combined errors are below a tolerance set by the software, and satisfactory agreement is obtained for the field continuity at all cell-shared boundaries. Using a cell size of $(0.020 \text{ in.})^3$ in a box size of $0.720 \times 0.900 \times 0.180 \text{ in.}^3$ results in $36 \times 45 \times 9$ grids or 14,580 cells for each evaluation frequency. For each calculation vector field calculations must be performed and fields on the boundaries of all adjacent cells matched.

This is a formidable task, even for a computer capable of millions of mathematical operations per second. A typical execution analysis for this model can

take several minutes to converge to a satisfactory field distribution throughout the box containing the circuit, and this process must be redone at each evaluation frequency, in the case of this example for 101 frequencies between 3000 and 4000 MHz.

The final result, once the fields sufficiently match on the cell boundaries, permits an evaluation of the fields at the input and output ports of the network. These fields are used to calculate the voltages and currents at the ports when matched source and loads are connected to them. From these voltage and current values, the scattering parameters of the network are calculated for each frequency of evaluation. Given the scattering matrix, the performance of the stub when embedded in a transmission line network is easily calculated using the network simulator portion of the software.

The EM simulation results are shown in Figure 7.34-5 and compared with the initial circuit simulation response. As can be seen, there is a significant difference in the two simulations. Since this stub was to form part of a filter in a 1920-MHz cellular system to remove the second harmonic at 3820 MHz, the discrepancy shown in Figure 7.34-5 suggests that a design based upon only a
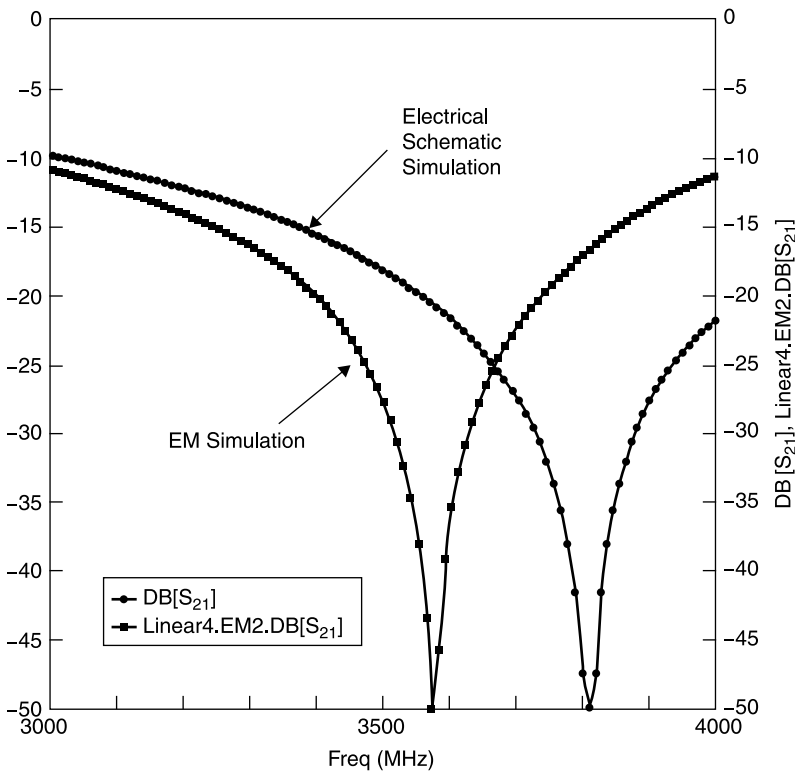
Figure 7.34-5   Comparison of EM and electrical schematic performance simulations.

network simulation would likely prove inadequate, necessitating additional experimental trials. However, with the EM simulation a design likely to be usable after the first design iteration would much more likely be obtained.

In this case, the error in using only an electrical schematic to model the filter would result in a stub resonance error of about 240 MHz at 3800 MHz, a 6% error. While this is a small percentage, it results in a difference in the null depth of 30 dB at the stub's stopband center, demonstrating the usefulness of electromagnetic simulation.

REFERENCES

1. Peter A. Rizzi, *Microwave Engineering, Passive Circuits*, Prentice-Hall, Englewood Cliffs, NJ, 1988. *Excellent microwave engineering textbook covering theory and design of transmission lines, couplers, filters, and numerous other passive devices.*

2. Simon Ramo, John R. Whinnery, and Theodore Van Duzer, *Fields and Waves in Communication Electronics*, 3rd ed., Wiley, New York, 1994. *This is a revision of [6].*

3. Paul J. Nahian, *Oliver Heaviside: Sage in Solitude*, IEEE Press, New York, 1988.

4. Hugh Hildreth Skilling, *Fundamentals of Electric Waves*, Wiley, New York, 1942. *This is a superbly written text on field theory, with many well-illustrated drawings of fields and vector functions.*

5. Sears and Zemansky, *University Physics, Vols. 1 and 2*, Addison-Wesley, Reading, MA, 1956.

6. Simon Ramo and John R. Whinnery, *Fields and Waves in Modern Radio*, Wiley, New York, 1944, 1953 (later revised with a third author, Van Duzer, in 1965, 2nd ed., 1984, and 3rd ed., 1994). *This is a classic introductory text describing fields and Maxwell's equations.*

7. George B. Collins, *Microwave Magnetrons, Radiation Laboratory Series*, McGraw-Hill, New York, 1948.

8. Jerome L. Altman, *Microwave Circuits*, D. van Nostrand Company, Inc., Princeton, NJ, 1964.

9. Robert E. Collin, *Field Theory of Guided Waves*, McGraw-Hill, New York, 1960. *Collin is a master of field theory as is demonstrated by this summary work.*

10. Robert E. Collin, *Foundations of Microwave Engineering*, McGraw-Hill, New York, 1966.

11. S. A. Schelkunoff, "The electromagnetic theory of coaxial transmission lines and cylindrical shields," *Bell System Technical Journal*, Vol. 13, October, 1934, pp. 532–579.

12. Joseph F. White, "Simplified theory for post coupling Gunn diodes to waveguide," *IEEE MTT Trans.*, Vol. MTT-20, No. 6, June, 1972, pp. 372–378.

13. N. Marcuvitz, *Waveguide Handbook*, Vol. 10 of the Radiation Laboratory Series, McGraw-Hill, New York, 1951.

14. Edward C. Jordan and Keith G. Balmain, *Electromagnetic Waves and Radiating Systems*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1968, Section 11.12.

15. Constantine A. Balanis, *Antenna Theory—Analysis and Design*, 2nd ed., Wiley, New York, 1997.

EXERCISES

E7.3-1   **a.** Find the electric field due to an isolated charge of 1 electron at a distance of 1 m in free space.

   b. How many electrons would be required to provide an electric field of 1 V/m at this same distance?

E7.3-2   Assume that there are exactly 10^{23} atoms/cm^3 of solid copper metal. (a) What is the size of the side of a copper cube large enough to provide the charge in Exercise 7.3-1b if every copper atom is ionized by gaining one electron? (b) What is the electric field if each copper atom gives up one electron?

E7.3-3   An electric dipole consists of two point charges, $+q$ and $-q$, spaced a distance 2d apart along the x axis. The midpoint between them is at x = 0. Show that for x ≫ d, the E field along the x axis is inversely proportional to x^3.

E7.4-1   A television set with a screen height of 0.5 m has a 525-line picture. It is turned on and has a beam accelerating voltage of V = 20,000 V. Once the electron beam passes the accelerating electric field, its velocity, v (assume v is in the $-z$ direction), is constant and given by $v = \sqrt{2eV/m}$ where m and e are, respectively, the mass and charge of the electron. When the beam is undeflected, the spot appears at the center of the screen. The distance from the electron gun to the screen is 0.7 m. A long wire carrying 10 A (assume this is in the $+z$ direction) of peak AC current is located under the TV. It is 0.5 m below the beam and parallel to it.

   a. In which direction will the beam be deflected when the current is traveling in the $+z$ direction?

   b. What will be the peak-to-peak deflection of the beam due to the AC current?

   c. The viewable height of the screen is 0.5 m. The picture consists of 525 lines. Will the deflection caused by the 10-A AC current be noticeable?

   d. If the deflection is noticeable, comment on how frequently this occurrence might be expected.

E7.5-1   A large farm field is in the shape of a parallelogram. Two adjacent sides are defined by the vectors:

$$\vec{A} = (3000 \text{ ft})\vec{E} + (4000 \text{ ft})\vec{N} \quad \text{and} \quad \vec{B} = (10{,}000 \text{ ft})\vec{E}$$

where the directions are $\vec{E} = $ east and $\vec{N} = $ north. Use the vector cross product to determine how many acres of land are contained in the field. Note that 1 acre $= 44{,}000 \text{ ft}^2$.

E7.7-1   **a.** A coaxial transmission line consists of an inner conductor of radius a and an outer conductor of radius b. Between these conductors is a dielectric having a relative dielectric constant of ε_R. Derive an expression for the capacitance per unit length of this line.

   b. A certain model of this transmission line has an inner conductor diameter of 1 cm and an outer conductor inside diameter of 3.35 cm, with a relative dielectric constant between the conductors of 2.1. What is the capacitance per unit length of this transmission line?

E7.7-2   The energy stored in an electric field is given by

$$U_E = \frac{\varepsilon_0 \varepsilon_R}{2} \int_V |E|^2 \, dV$$

and for a capacitor it is also given by

$$U_E = \tfrac{1}{2} C V^2$$

Show that both of these expressions give the same result for U_E for a unit length of coaxial line L, where C is the capacitance/unit length examined in E7.7-1.

E7.8-1   A static electric potential function exists in a region described by a (x, y, z) coordinate system. The function is

$$\Phi = 3x^3 + 5(y^2 - y) + 17z + 6 \text{ volts}$$

Calculate the amount of energy required to move an electron from the origin to the point $(4, 5, 2)$.

E7.8-2   What is the gradient of the potential function

$$\Phi = 3x^3 + 5(y^2 - y) + 17z + 6 \text{ volts}$$

at the point $(4, 5, 2)$?

E7.10-1 Two straight parallel conductors are a distance D apart and carry currents +I and −I, respectively, "into and out of the paper." Find an expression for and graph the magnitude of the total H field in the plane of the paper as a function of position between and outside of the linear interval containing the conductors.

E7.11-1 Find the curl of the vector field

$$\vec{A} = (y^2 z)\vec{x} + (x^3 y)\vec{y} + (z^2 xy)\vec{z}$$

E7.12-1 The circular path shown encloses a uniform magnetic field into the page. The radius of the path is $r = 0.2$ m and the magnetic flux density is given by $\vec{B} = 20(1 - e^{-t})$ teslas, where t is in seconds. (a) Calculate V and I at t = 1 s. (b) Graph B and V for 0 < t < 4 s.
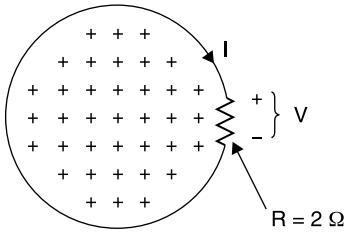
E7.14-1 From Ampere's law, the H field surrounding a wire carrying a DC current I is given by (7.10-2). Find the curl of H as a function of r for $r > a$, the radial distance from the center of the conductor. Assume the conductor is a straight vertically oriented cylinder of radius a.

E7.14-2 Solve for the curl of H inside the conductor of E7.14-1 as a function of r for $0 < r \leq a$.

E7.15-1 Find the Laplacian for a potential function that has cylindrical symmetry and is proportional to ln r, by first evaluating the gradient and then taking its divergence. Interpret the result and describe what geometry could produce this potential.

E7.16-1 It is said that Maxwell's four equations can be reduced to two, since the first two can be derived from the last two. Show that this is so. *Hint:* Start by taking the divergence of the last two equations.

E7.16-2 Using rectangular coordinates show that the defining equation for the vector Laplacian is, in fact, an identity. This identity is key to numerous field derivations, including the proof that plane waves can propagate in free space:

$$\nabla^2 \vec{A} = \nabla(\nabla \cdot \vec{A}) - \nabla \times \nabla \times \vec{A} \quad \text{where } \vec{A} = A_x\vec{x} + A_y\vec{y} + A_z\vec{z}$$

E7.17-1 The graphite form of carbon has about the lowest conductivity of the materials used as conductors, with $\sigma \approx 7 \times 10^4$ ℧/m. Applying the criterion of (7.17-9), up to what frequency would graphite be considered a "good conductor."

E7.18-1 Create a table showing skin depth in decade frequency increments from 1 to 100,000 MHz for copper, aluminum, brass, and stainless steel (0.1 C, 18 Cr, 8 Ni, balance Fe). Use as respective resistivity values 1.72, 2.62, 3.9, and 90×10^{-6} Ω-cm. Then graph your results on log-log paper. Note that all of these metals are nonmagnetic.

E7.19-1 The manufacturer of a measuring system requires a precise time delay for 0.5-μs-long pulses and uses a temperature controlled, 150-m long air dielectric copper coaxial line of 50 Ω impedance as the delay path for the pulses. Small dielectric beads support the center conductor at various intervals, but the effective relative dielectric constant is approximately unity. The inner conductor diameter is 0.43 cm and the outer conductor has an inside diameter of 1 cm (same as the example in Section 7.19). The pulse generator and load are matched to the coaxial line at both respective ends.

   The manufacturer has hired you as a consultant to advise on the electrical effects, if any, of replacing the copper coaxial cable with one made of aluminum because it is less expensive. In this replacement it is proposed that the dielectric and cross-sectional dimensions as well as the length of the line would be unchanged.

   a. Estimate the ohmic insertion loss of the copper and aluminum cables at 1 GHz.

   b. Estimate the electrical length of each cable relative to a path having the same dielectric constant and physical length at 1 GHz.

   c. What assumptions and approximations did you make in arriving at your recommendations?

E7.23-1 Suppose that we radiate a sinusoidal signal into free space having fairly constant $\vec{E}$ and $\vec{H}$ fields over a 1 m^2 area with a power density of 1 W/m^2 at some distance from the antenna. What is the $\vec{E}$ field strength of this wave? *Hint:* Assume we can use an analogy with circuit theory by which $P = V_A^2 / Z_0$.

E7.23-2 A sinusoidally varying E_θ *and* H_φ set of fields (using spherical coordinates) is radiated uniformly in all directions (isotropically). Show what the expressions for E_θ *and* H_φ must be as functions of time t and distance r from the transmitter. Assume that the transmitter is an ideal point source of radiation and that the ratio of E_θ *and* H_φ fields is the same as that for plane wave propagation.

E7.24-1 Show that a vertically polarized field can be represented as the sum of judiciously chosen RHCP and LHCP fields.

E7.24-2   Find the equations of a LHCP field and a RHCP field that combine to produce a linearly polarized field oriented at an angle α from the x axis (the vertical direction).

E7.24-3   For the left-hand polarized wave described in Figure 7.24-5:
a. Verify that E is in the negative y direction at $z = 3\lambda/4$ when $t = 0$.
b. What is the direction of E at $z = 3\lambda/4$ when t = T/4?
c. Sketch the wave in the xy plane for $z = 3\lambda/4$ and t increasing from 0 to T/4.

E7.25-1   Solve for the characteristic impedance of coaxial line, deriving the result in (7.25-15) using (7.25-7) for the distributed C of the line and deriving the distributed L of the line based on the stored magnetic energy of the line related to L by $U_M = \frac{1}{2}LI_0^2 = \frac{1}{2}\mu \int_V H^2\, dV$ where I_0 is the peak line current and H is the magnetic field.

E7.25-2   Use Poynting's vector to calculate the average power flow along a coaxial transmission line having the form shown in Figure 7.25-2. Assume sinusoidal excitation and use the phasor forms of $\vec{E}$ and $\vec{H}$.

E7.25-3   Verify that the ratio of E_r to H_φ for the TEM mode in coaxial line is $\eta = \sqrt{\mu/\varepsilon}$. Use the phasor form of Maxwell's third equation (Faraday's law). Note that for the TEM mode, $k = \omega\sqrt{\mu\varepsilon}$.

E7.25-4   Prove that the propagation constant for a sinusoidal wave on a coaxial transmission line is $k = \omega\sqrt{\mu\varepsilon}$. *Hint:* Begin with the wave equation (7.21-1), assume that k has the value sought and verify that the equation produces an identity.

E7.26-1   Find the approximate characteristic impedance of the parallel-plate transmission line shown in Figure 7.26-1a, assuming that the line is lossless and that w ≫ d (hence, all fields are uniform and confined to the space between the plates).

E7.26-2   Use Poynting's theorem to derive an expression for the average power flow in an air-filled rectangular waveguide operating in the TE_{10} mode.

E7.26-3   An air-filled rectangular waveguide is operated in the TE_{10} mode. Its dimensions are $a = 2.25$ cm and $b = 1$ cm.
a. Calculate the cutoff frequency of the TE_{10} mode.
b. Which mode has the next lowest cutoff frequency?
c. Repeat part (a) for a dielectric filled rectangular waveguide filled with a nonmagnetic dielectric having $\varepsilon_r = 2.0$.

E7.26-4   In general, the propagation constant for waveguide is $\gamma = \alpha + j\beta$. If γ is real, then $\gamma = \alpha = $ an attenuation constant (in nepers/unit length). The dimensions of an air-filled rectangular waveguide are

$a = 4.0$ cm and $b = 2.0$ cm. Calculate the attenuation constant in decibels/centimeter at 2.5 GHz, assuming TE$_{10}$ mode operation.

E7.26-5   For most transmission line circuits, the usable bandwidth is related to the frequency sensitivity of the line's electrical length, $\theta = 2\pi l/\lambda$ for TEM lines, and $\theta = 2\pi l/\lambda_g$ for waveguides. The average time required for a group of frequencies to transverse a network is called the *group delay* and is equal to $d\theta/d\omega$. To avoid distortion of signals comprising a band of frequencies, all frequencies should traverse the network in the same time (i.e., the same delay for all frequencies, or $d\theta/d\omega = a\ constant$, independent of f).

   a. Derive expressions for $d\theta/d\omega$ in both the TEM and TE$_{10}$ cases.

   b. Calculate $d\theta/d\omega$ at 7.5 GHz and at 12.0 GHz for air-filled coaxial line.

   c. Repeat (b) for air-filled rectangular waveguide with $a = 2.25$ cm and $b = 1.0$ cm, assuming operation in the TE$_{10}$ mode.

   d. Compare the results of parts (b) and (c) at the two frequencies.

E7.27-1   A slender cylindrical 20-Ω resistor is mounted across a very narrow height, air-filled, rectangular waveguide on its centerline. The waveguide width is $a = 5$ cm and its height is $b = 0.5$ cm.

   a. Calculate the insertion loss caused by the resistor at 4 GHz when the generator and load are matched to the impedance Z_g of the waveguide.

   b. Calculate the insertion loss when the centerline of the resistor is 0.5 cm from the narrow wall of the waveguide.

E7.27-2   Use transmission line theory (Chapter 4) to verify (7.27-20).

E7.28-1   A manufacturer has designed an intrusion alarm circuit to operate at 10.525 GHz. The network is realized as an integrated circuit chip that measures only 0.25×0.25 in. The manufacturer wishes to install it in a standard metal "can" measuring $0.625 \times 0.625 \times 0.125$ in. (internal dimensions). The circuit contains transistors having gain up to 18 GHz. Is higher order moding in the circuit enclosure possible?

E7.29-1   A steady DC current exists in a short, straight, thin conductor located at the origin of the right-hand coordinate system (x, y, z). The conductor is in the z direction. Ignore the fact that this short DC current with no return path cannot exist in practice.

   a. Sketch the geometry, identifying the rectangular and spherical coordinates.

   b. Find the vector potential $\vec{A}$ resulting from this current and express it in spherical coordinates (r, θ, φ).

   c. Find the $\vec{H}$ field components and express them in spherical coordinates.

E7.32-1   In the development of (7.32-16) only the terms varying as $(1/r)$ in E_θ and H_φ were used to form Poynting's vector. However, there were other terms in both E_θ and H_φ. Show that including them would not change the result.

E7.32-2   Three short wire antennas of height h are spaced a quarter wavelength apart on the z axis and oriented in the $+z$ direction. The center antenna's midpoint is at $z = 0$. The sinusoidal excitation current in the center antenna is twice as large as the current in each of the other two antennas (currents are 1:2:1).

   a. Sketch the antenna array. Then derive an expression for $|E_\theta|$ in the far field for the three-element array. Assume the radii from each antenna element to the far-field point are approximately parallel and hence that the angles (θ) of these radii measured from the z axis are approximately equal to each other. However, do take into account the phase differences due to the differences in path length from the elements to the far-field point (this is necessary to describe the antenna pattern).

   b. With power density P proportional to $|E_\theta|^2$, calculate P/P_M at $\theta = 0, \pi/6, \pi/4, \pi/3$, and $\pi/2$ where θ is the cylindrical coordinate angle and P_M is the relative power at $\theta = \pi/2$ (on the horizon). Express your answers both numerically and in decibels.

   c. Estimate from your calculations the 3-dB beam width of this array antenna.

E7.33-1   A 10-GHz point-to-point digital communication link is to be established between downtown Boston and the suburb of Lexington. The Boston antenna will be on the top of a 60-story building and the Lexington antenna will be about 100 ft high, resulting in a "line-of-sight path" of 25 miles. Dish antennas 2 ft in diameter will be employed at both sites. A system power margin of 40 dB is judged necessary to accommodate added path loss due to rain and snow and to provide a low bit error rate (BER). How much power must be radiated if the minimum detectable signal strength for the modulation to be used is –90 dBm? Assume the dish antennas are 50% efficient.

E7.34-1   Use a network simulator to plot the insertion loss/isolation of the single open-circuited stub used as the example in Section 7.34.

   a. What is the insertion loss of this filter at 1920 MHz?

   b. Use a pair of identical stubs spaced appropriately to reduce the insertion loss to less than 1.0 dB from 1900 to 1940 MHz.

   c. Use an EM simulator to design the actual layout of this two-stub filter and plot the results over frequency from 1000 to 5000 MHz.

Directional Couplers

8.1 WAVELENGTH COMPARABLE DIMENSIONS

The fact that circuit dimensions are comparable to the operating wavelength is used to advantage in the design of circuits that couple energy according to its direction of propagation. These are called *directional couplers*.

An example is the waveguide multihole coupler shown in Figure 8.1-1. Its operation can be appreciated by noting that a wave incident at port 1 exits at port 4 with a portion of the incident power exiting at port 3. However, no power exits at port 2 because the separate waves that pass through the coupling holes have a differential path length of $\lambda_g/2$ and arrive with equal amplitudes but 180° out of phase at port 2. This is an example of *forward coupling* because the coupled wave travels in the same direction as the incident wave at port 1.

The directivity just described varies with frequency because the physical separation of the coupling holes must be related to the wavelength. This coupler type can be made broadband by using not two but many coupling holes, varying their coupling (hole size) according to an optimal distribution, say a binomial or Chebyshev distribution. More about mathematical distributions in Chapter 9 on filters. Using similar reasoning, it can be seen that a wave entering at port 4 will exit at port 1 with a coupled portion exiting at port 2.

This coupler example shows how a simple measurement system can be arranged. A detector placed at port 3 samples the *incident wave* while a detector placed at port 2 samples the *wave reflected* of a load connected to port 4. The ratio of these two waves is the reflection coefficient of the load at port 4. This is the basis of the *network analyzer*. More about network analyzers later in this chapter.

8.2 THE BACKWARD WAVE COUPLER

Placing two transmission lines in parallel and close together so that energy propagating on one is coupled to the other can form a transmission line cou-
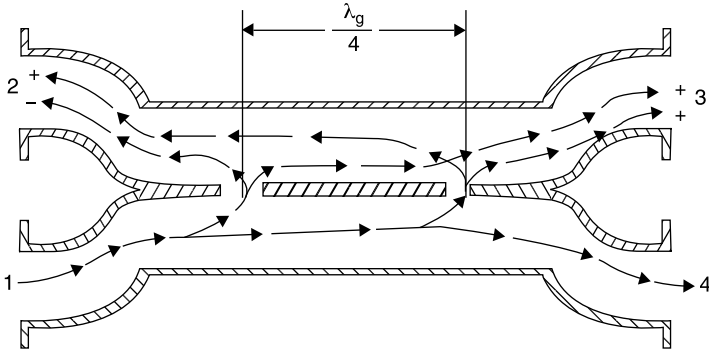
Figure 8.1-1   Cross-sectional view of a two-hole waveguide directional coupler.

pler. Intuitively, one would think that the coupled energy should travel in the same direction as does that on the coupled line, but this is not so. The *coupled power travels backward*, and therefore this coupler is called a *backward wave coupler* (Fig. 8.2-1).

This TEM mode backward wave coupler provides maximum coupling when the electrical length θ of the coupled lines is $90°$. When the cross-sectional dimensions of the coupler are properly chosen, power entering port 1 exits at ports 2 and 4, with no energy exiting at port 3. Furthermore, the device presents a matched load at input port 1 and a $90°$ phase difference between the signals emerging from ports 2 and 4 *at all frequencies*! This coupler is inherently broadband yielding up to an octave bandwidth (the highest frequency twice that of the lowest), as will be seen. Further increases in bandwidth can be
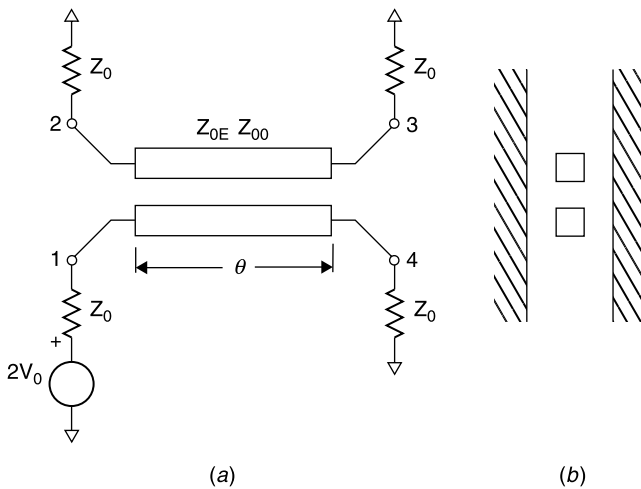
Figure 8.2-1   The TEM backward wave coupler: (*a*) top view and (*b*) cross section.

achieved by using cascaded 90° coupling sections and choosing their coupling characteristics appropriately [1, Sec. 8.3 and 2].

8.3   EVEN- AND ODD-MODE ANALYSIS

Remarkably, all of the progress that we have made so far in analyzing transmission lines, including analysis aided by the Smith chart, has not provided any

(a)

(b)
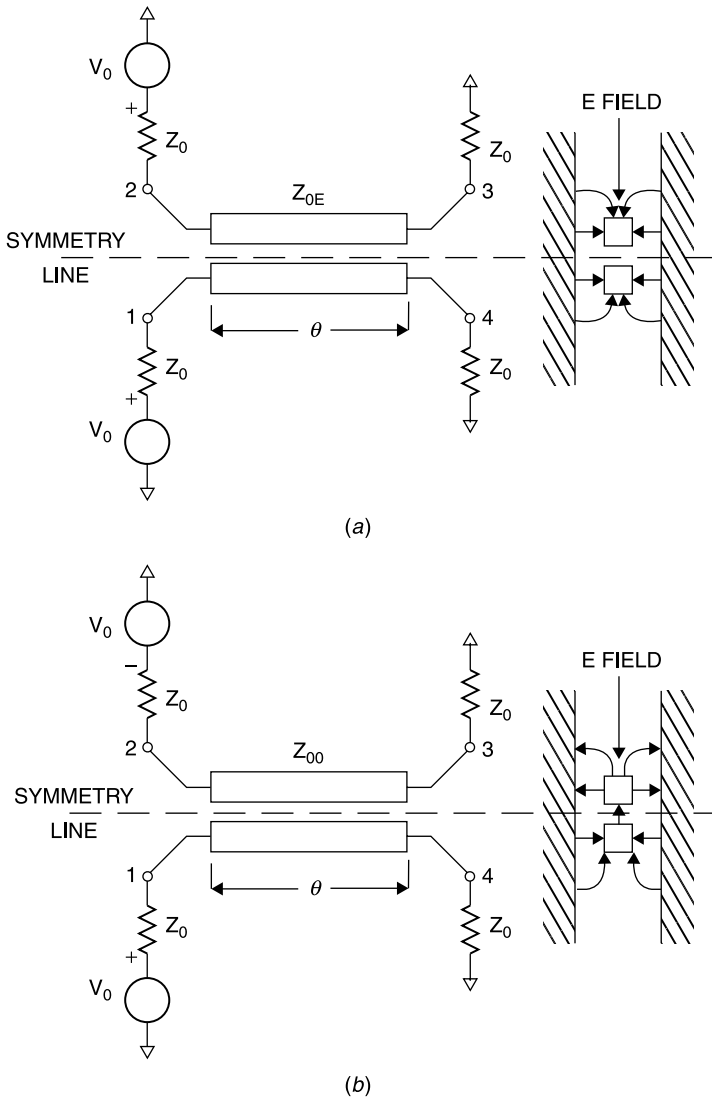
Figure 8.3-1   Even- (a) and odd-mode (b) excitations of the TEM coupled lines.

means to analyze the four-port network consisting of the two coupled transmissions lines shown in Figure 8.2-1. Our methods only permit the determination of the input impedance to an isolated transmission line given any load. To analyze coupled transmission lines requires a new method. The *even- and odd-mode analysis is* such an approach, a clever insight to accomplish coupled line analysis using single transmission line analysis methods. It consists of representing any voltage excitation of the coupled lines as the sum of an *even mode* and an *odd mode* (Fig. 8.3-1).

Studying Figure 8.3-1, it can be seen that the superposition of the four generator voltages, $\pm V_0$, applied at ports 1 and 2 is equivalent to the application of the single voltage $2V_0$ at port 1 in Figure 8.2-1. The clever part of the even- and odd-mode analysis is that now each transmission line in Figure 8.3-1 can be split away from its companion coupled line and analyzed as a single line with load. In doing so, however, it is necessary to use the *even-mode characteristic impedance* for the circuit in Figure 8.3-1a and the *odd-mode characteristic impedance* for Figure 8.3-1b.

The even-mode impedance Z_{OE} is defined when the adjacent line has the same voltage excitation. This corresponds to placing a "magnetic wall" (on which the tangential H field is zero) midway between the conductors. The odd-mode impedance Z_{OO} is defined when the adjacent line has the opposite voltage and corresponds to the condition of an electric wall (on which the tangential E field is zero) midway between the conductors. The geometries used for calculating these impedances are shown in Figure 8.3-2.
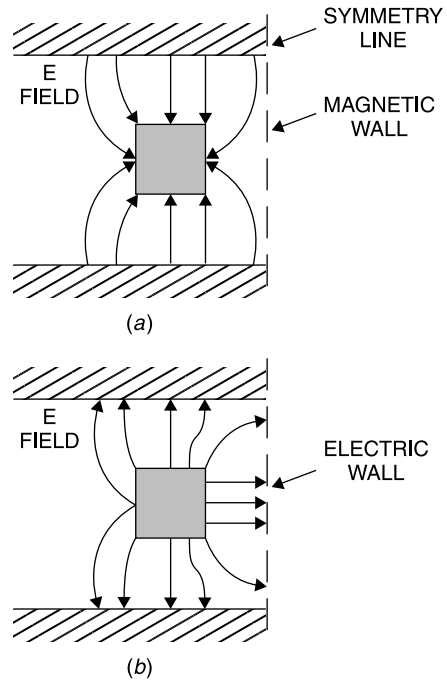
Figure 8.3-2 Geometries for determining Z_{OE} (a) and Z_{OO} (b).
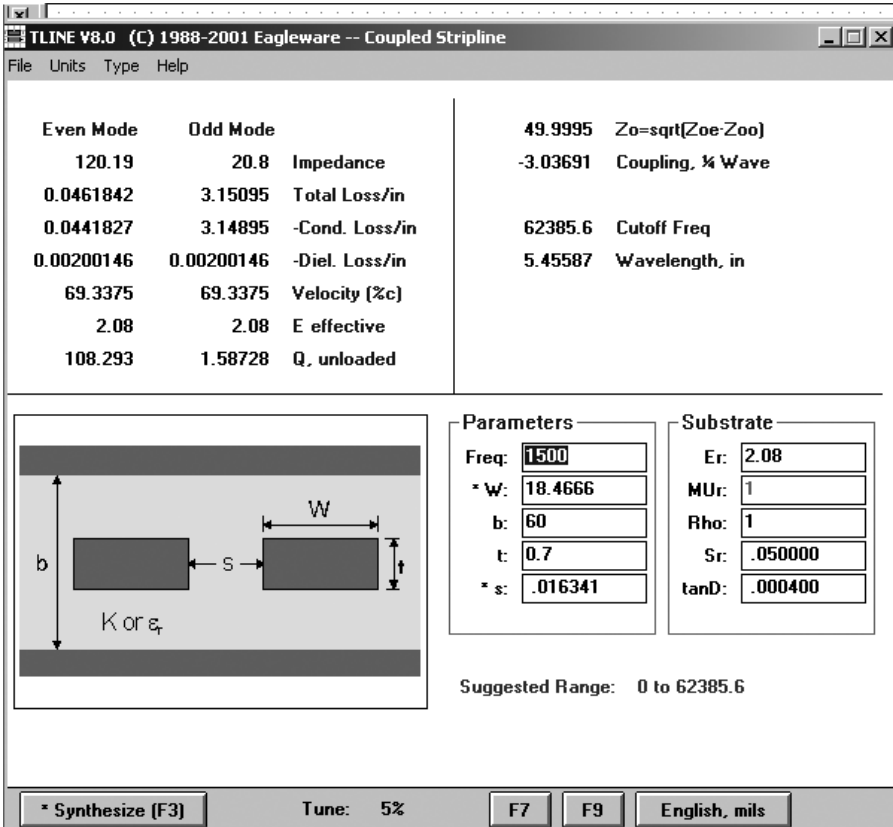
Figure 8.3-3 Software outputs showing the Z_{OE} and Z_{OO} and other parameters required for a 3-dB backward wave coupler. Obtained using Eagleware *TLine*.

These impedances are found in the same manner as was used to determine the characteristic impedance of coaxial line in Section 7.25, albeit with somewhat more mathematical complexity because the geometry is not as simple as that of coaxial line. The even- and odd-mode method dates to the 1950s when the Z_{OE} and Z_{OO} impedances were tabulated for various coupler cross-sectional geometries, but the current method for determining them is to use software. A sample software is shown in Figure 8.3-3 for the geometry to be used for a 3-dB coupler for which $Z_{OE} = 120$ Ω and $Z_{OO} = 20.8$ Ω.

Why this choice of even- and odd-mode impedances yields a 3-dB coupler will be evident once the analysis of the coupler is completed. The analysis method [3, Chapter 6] will be demonstrated by solving for the voltages at the coupler's ports 1, 2, 3, and 4 given a generator voltage of $+2V_0$ with Z_0 loads as shown in Figure 8.2-1. We begin by determining the reflection coefficients at ports 1 and 2 for the even and odd mode excitations of Figure 8.3-1. The equivalent circuit used for the individual excitation analysis is shown in Figure 8.3-4.
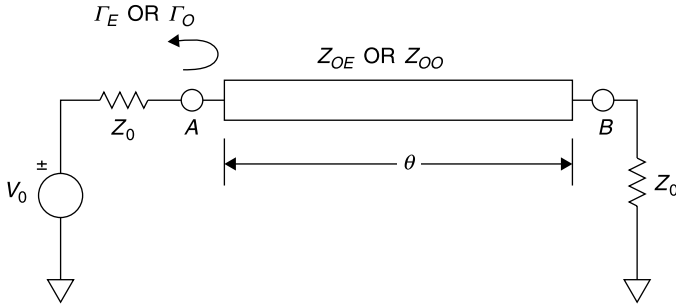
Figure 8.3-4 Circuit used to compute the input reflection coefficients to even- and odd-mode excitations.

The single-mode equivalent circuit in Figure 8.3-4 has input port A and output port B to distinguish these terminals from those of the coupler in Figure 8.2-1. The characteristic impedance of the line is Z_{OE} for the even-mode calculation and $-V_0$ and Z_{OO} for the odd mode. Notice that Z_0 is not a match termination for either the Z_{OE} or the Z_{OO} transmission line.

Given these procedures, the input impedance seen at A for the even mode is found from the input impedance formula (4.14-2):

$$Z_{AE} = Z_{OE} \frac{Z_0 + jZ_{OE} \tan \theta}{Z_{OE} + jZ_0 \tan \theta} \tag{8.3-1}$$

and the reflection coefficient at A is

$$\Gamma_E = \frac{Z_{AE} - Z_0}{Z_{AE} + Z_0} \tag{8.3-2}$$

Substituting (8.3-1) into (8.3-2) and simplifying gives

$$\Gamma_E = \frac{j[Z_{OE}^2 - Z_0^2] \tan \theta}{2Z_0 Z_{OE} + j[Z_{OE}^2 + Z_0^2] \tan \theta} \tag{8.3-3}$$

Similarly, for the odd excitation,

$$\Gamma_O = \frac{j[Z_{OO}^2 - Z_0^2] \tan \theta}{2Z_0 Z_{OO} + j[Z_{OO}^2 + Z_0^2] \tan \theta} \tag{8.3-4}$$

The total reflected voltage at port 1 for the circuit of Figure 8.2-1 is the sum of the even and odd reflection voltages given by

$$V_{1R} = \frac{V_0}{2} \Gamma_E + \frac{V_0}{2} \Gamma_O \tag{8.3-5}$$

At port 2, the even- and odd-mode voltages are of opposite polarity. Therefore,

$$V_2 = \frac{V_0}{2}\Gamma_E - \frac{V_0}{2}\Gamma_O \tag{8.3-6}$$

Substituting (8.3-3) and (8.3-4) into (8.3-5) and (8.3-6) gives a result for V_{1R} (using the plus sign option in (8.3-7) and V_2 by using the minus sign option:

$$V_{1R,2} = \frac{V_0}{2}\Gamma_E \pm \frac{V_0}{2}\Gamma_O$$

$$V_{1R,2} = \frac{N_A^{\pm}}{D_A}$$

$$= \frac{\dfrac{V_0}{2}\left\{\begin{array}{l}(j\tan\theta)(Z_{OE}^2 - Z_0^2)[2Z_0Z_{OO} + (j\tan\theta)(Z_{OO}^2 + Z_0^2)] \\ \pm(j\tan\theta)(Z_{OO}^2 - Z_0^2)[2Z_0Z_{OE} + (j\tan\theta)(Z_{OE}^2 + Z_0^2)]\end{array}\right\}}{[2Z_0Z_{OE} + (j\tan\theta)(Z_{OE}^2 + Z_0^2)][2Z_0Z_{OO} + (j\tan\theta)(Z_{OO}^2 + Z_0^2)]} \tag{8.3-7}$$

The denominator, D_A, is nonzero for all θ and for all finite Z_0, Z_{OE}, and Z_{OO}. Furthermore, when Z_{OE} and Z_{OO} are selected to satisfy

$$Z_0 = \sqrt{Z_{OE}Z_{OO}} \tag{8.3-8}$$

the even-mode numerator, N_A^+, is identically equal to zero for all θ. This means that the coupler will be *matched to Z_0 at input port 1 at all frequencies* when (8.3-8) is satisfied. This is a remarkable result, doubly so in view of the algebraic complexity with which the early researchers [4] of this coupler had to cope.

After algebraic manipulation, the value of $V_2 = N_A^-/D_A$ can be written

$$V_2 = V_0 \frac{(j\sin\theta)\left(\dfrac{Z_{OE} - Z_{OO}}{Z_{OE} + Z_{OO}}\right)}{\dfrac{2Z_0\cos\theta}{Z_{OE} + Z_{OO}} + j\sin\theta} \tag{8.3-9}$$

and when the coupling length is 90°, (8.3-9) reduces to

$$V_2 = V_0 \frac{Z_{OE} - Z_{OO}}{Z_{OE} + Z_{OO}} \tag{8.3-10}$$

This suggests the definition of a *voltage coupling coefficient k* defined as

$$k = \frac{Z_{OE} - Z_{OO}}{Z_{OE} + Z_{OO}} \tag{8.3-11}$$

and the coupling in decibels is

$$\text{Coupling (dB)} = -20 \log k \qquad (8.3\text{-}12)$$

where the negative sign is used so that the coupling is positive when expressed in decibels.

When (8.3-8) and (8.3-11) are substituted into (8.3-9), the expression for V_2 becomes

$$V_2 = V_0 \frac{jk \sin \theta}{\sqrt{1 - k^2} \cos \theta + j \sin \theta} \qquad (8.3\text{-}13)$$

If a 50-Ω, 3-dB coupler (at $\theta = 90°$) is required, $k = 1/\sqrt{2}$, since $V_2^2/Z_0 = k^2 V_0^2/Z_0 = \frac{1}{2} V_0^2/Z_0$, where V_0^2/Z_0 is the available power from port 1. Substituting this value of k into (8.3-11) and applying the match criterion of (8.3-8) gives the values $Z_{OE} = 120.19$ Ω and $Z_{OO} = 20.8$ Ω for a 3-dB, 50-Ω coupler that were stated without proof earlier.

Now it remains to calculate the V_3 and V_4 for the coupler in Figure 8.3-1 to complete the analysis. We use the $ABCD$ matrix method and include both the source and load impedances (Fig. 8.3-5).

For the even mode the $ABCD$ matrix cascade is

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}_E = \begin{pmatrix} 1 & Z_0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos \theta & jZ_{OE} \sin \theta \\ jY_{OE} \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} 1 & 0 \\ Y_0 & 1 \end{pmatrix}$$

Performing the indicated matrix multiplications and simplifying, this becomes

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}_E = \begin{pmatrix} 2 \cos \theta + j \sin \theta (Y_0 Z_{OE} + Z_0 Y_{OE}) & jZ_{OE} \sin \theta + Z_0 \cos \theta \\ jY_{OE} \sin \theta + \cos \theta & \cos \theta \end{pmatrix}$$
$$(8.3\text{-}14)$$

If this network is excited by a voltage V_0 (Fig. 8.3-2a) and the output current is
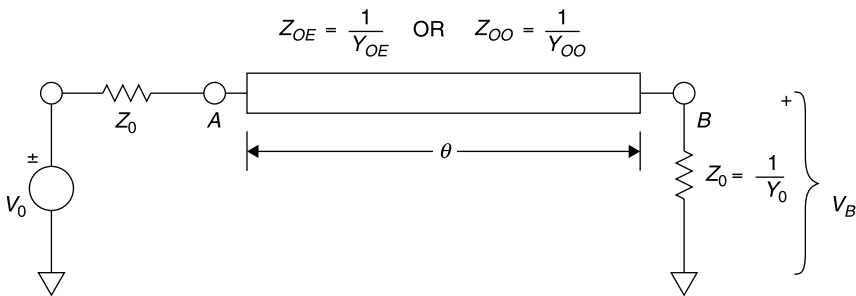
Figure 8.3-5 Circuit used to calculate the output voltages V_3 and V_4 at ports 3 and 4 of the coupler in Figure 8.2-1.

zero (since we included the load in the *ABCD* network) the output voltage is

$$V_{BE} = \frac{V_0}{A_E} = \frac{V_0}{2\cos\theta + (j\sin\theta)(Y_0 Z_{OE} + Z_0 Y_{OE})} \qquad (8.3\text{-}15)$$

Similarly, since the odd excitation corresponds to the application of V_0, at port 1 of the coupler and the use of odd-mode impedances,

$$V_{BO} = \frac{V_0}{A_O} = \frac{V_0}{2\cos\theta + (j\sin\theta)(Y_0 Z_{OO} + Z_0 Y_{OO})} \qquad (8.3\text{-}16)$$

The output voltages at ports 3 and 4 that we seek are given by

$$V_3 = V_{BE} - V_{BO} \qquad (8.3\text{-}17a)$$

$$V_4 = V_{BE} + V_{BO} \qquad (8.3\text{-}17b)$$

Combining (8.3-15) and (8.3-16), the optional plus sign gives V_3 and the minus sign gives V_4. Thus,

$$V_{BE} \pm V_{BO} = \frac{N_B^{\pm}}{D_B}$$

$$= V_0 \frac{2\cos\theta + (j\sin\theta)(Y_0 Z_{OO} + Z_0 Y_{OO}) \pm 2\cos\theta + (j\sin\theta)(Y_0 Z_{OE} + Z_0 Y_{OE})}{[2\cos\theta + (j\sin\theta)(Y_0 Z_{OE} + Z_0 Y_{OE})][2\cos\theta + (j\sin\theta)(Y_0 Z_{OO} + Z_0 Y_{OO})]}$$
$$(8.3\text{-}18)$$

where (8.3-17a) is the minus option and (8.3-17b) is the plus option. For finite Z_0, Z_{OE}, and Z_{OO}, the denominator D_B is nonzero for all θ; and, when the matching condition (8.3-8) is substituted into the N_B^- expression of (8.3-18), the result is identically zero for all θ. Thus the voltage V_3 *is identically zero for all frequencies!* For this backward wave coupler, port 2 is the intended coupled port and port 3 is the *isolated port*. No power exits port 3 when power is applied to port 1 for any frequency and for any coupling coefficient k!

   To complete the analysis of the backward wave coupler, we determine the direct voltage V_4 output. This is obtained by using the plus sign option in (8.3-18). When this is taken and a common factor taken from numerator and denominator, the result is

$$V_4 = \frac{2V_0}{2\cos\theta + (j\sin\theta)[\sqrt{Z_{OE}/Z_{OO}} + \sqrt{Z_{OO}/Z_{OE}}]} \qquad (8.3\text{-}19)$$

Then imposition of the matching condition (8.3-8) and the coupling coefficient definition (8.3-11) allows this expression to be written as

$$V_4 = V_0 \frac{\sqrt{1-k^2}}{\sqrt{1-k^2}\cos\theta + j\sin\theta} \tag{8.3-20}$$

Notice that when the coupling length is 90°, from (8.3-13) and (8.3-20), $V_2 = V_4 = V_0/\sqrt{2}$ and exactly one half of the incident power, V_0^2/Z_0 exits the two ports 1 and 4. This makes sense because the coupling is 3 dB and no power is reflected and none exits the decoupled port 3. Of course, practical couplers have mismatches at the ports that degrade this ideal performance. Nevertheless, the backward wave TEM coupler of this analysis is a frequently used component for its broadband performance. The coupling varies with frequency, having maximum coupling (Fig. 8.3-6) when the coupled line length is an odd multiple of 90°. Thus, the performance is repeated in the frequency domain at
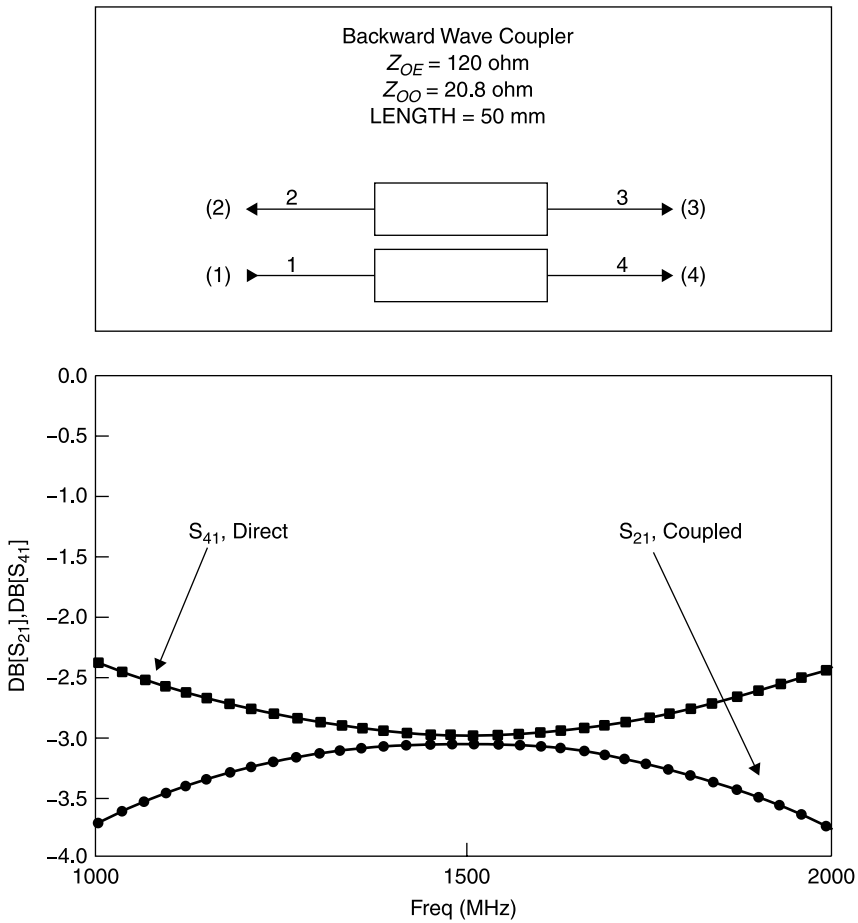
Figure 8.3-6   Direct power and coupled power of the backward wave coupler calculated using *Genesys* simulator.

the third, fifth, seventh, and so forth harmonics. The simulated performance of a 3-dB coupler with air dielectric ($\lambda/4 = 50$ mm) is shown in Figure 8.3-6.

In designing a coupler for a given amount of coupling, it is convenient to express the even- and odd-mode impedances in terms of the desired coupling coefficient. This is done by solving (8.3-8) and (8.3-9) for Z_{OE} and Z_{OO}:

$$Z_{OE} = Z_0 \sqrt{\frac{1+k}{1-k}} \qquad (8.3\text{-}21)$$

$$Z_{OO} = Z_0 \sqrt{\frac{1-k}{1+k}} \qquad (8.3\text{-}22)$$

For a given coupling tolerance, a broader bandwidth can be obtained by *overcoupling* at the center frequency. For example, if a center frequency coupling of 2.7 dB is chosen, $k = 0.724$ and (8.3-21) and (8.3-22) yield $Z_{OE} = 125\ \Omega$ and $Z_{OO} = 20\ \Omega$. For this design the variation of coupling is within 0.3 dB of 3.0 dB from 1000 to 2000 MHz, an octave bandwidth, as shown in Figure 8.3-7.
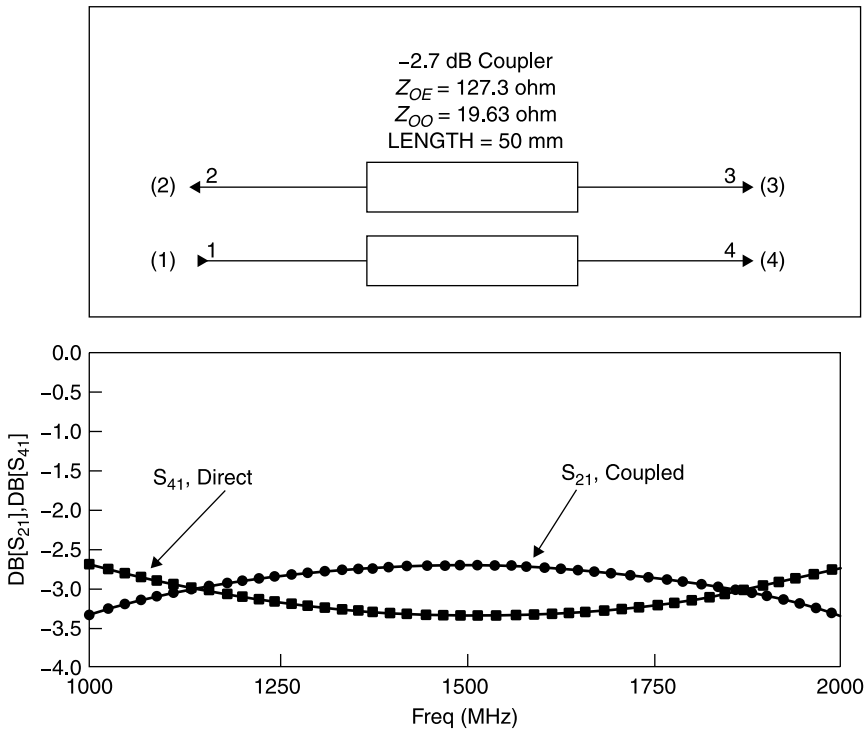
Figure 8.3-7 A -2.7 dB, over coupled, backward wave coupler.
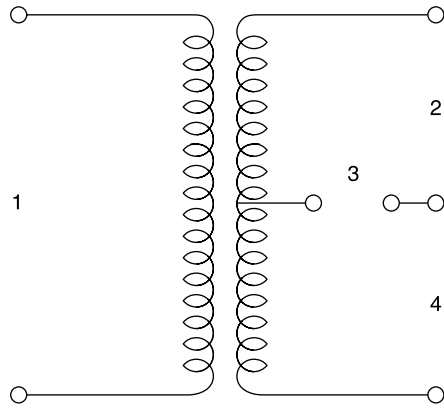
Figure 8.3-8 Hybrid coil with ports defined for comparison with the backward wave coupler.

The remarkable performance of the homogeneous (same dielectric constant for even and odd modes) TEM backward wave coupler can be summarized as follows:

1. The coupler is matched ($V_1 = V_0$) *at all frequencies!*
2. The directivity is infinite ($V_3 = 0$) *at all frequencies!*
3. The direct (V_4) and coupled (V_2) voltages are exactly 90° out of phase *at all frequencies!*
4. The magnitude of the coupled output voltage relative to the input voltage is $k = (Z_{OE} - Z_{OO})/(Z_{OE} + Z_{OO})$.

When designed such that the coupled and direct waves are equal, or nearly so, it is also called a *hybrid coupler* [5, p. 280]. This nomenclature may have been derived from the *hybrid coil* (Fig. 8.3-8) used widely in lower frequency telephone systems [6, p. 282]. When properly terminated, the hybrid coil has the property that a signal at 1 divides evenly between ports 2 and 4, and there is no output at 3. A signal at 2 also divides between 1 and 3 with no output at 4, and so forth. The 3 dB backward coupler has performance similar to the hybrid coil, but it also has a 90° phase shift between the two equal outputs whereas the hybrid coil does not.

The single quarter-wave coupler provides up to an octave (2/1) bandwidth for about 3 dB $\pm$ 0.4 dB of coupling. Considerably broader bandwidth can be obtained by cascading odd numbers of coupling sections. A 5/1 bandwidth with 3 dB $\pm$ 0.4 dB can be obtained by cascading three sections having 12.64, 1.28, and 12.64 dB coupling, respectively. This three-section coupler is described in a classic paper by Seymour Cohn [2]. The simulated performance of an air dielectric version of the model that he built is shown in Figure 8.3-9.

To achieve the very tight coupling in the center coupler Cohn invented a transmission line format called a *reentrant section*. It consists of a pair of con-

12.64 dB Coupler 1.28 dB Coupler 12.64 dB Coupler2
$Z_{OE} = 63.5$ ohm $Z_{OE} = 183$ ohm $Z_{OE} = 63.5$ ohm
$Z_{OO} = 39.4$ ohm $Z_{OO} = 13.6$ ohm $Z_{OO} = 39.4$ ohm
LENGTH = 25 mm LENGTH = 25 mm LENGTH = 25 mm
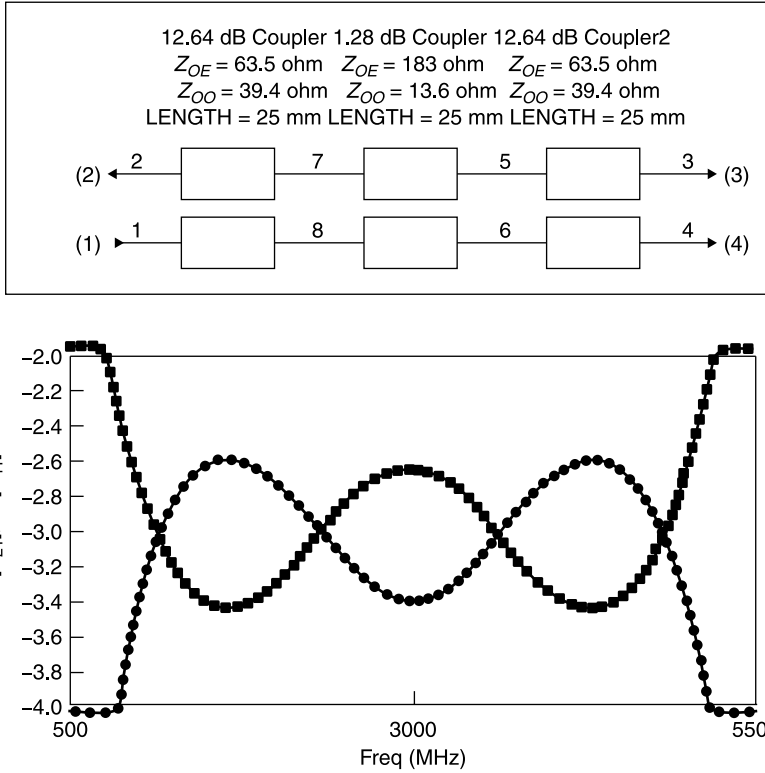
Figure 8.3-9 Performance of a three-section coupler cascade [2].

ductors, A and B, embedded as separate coaxial lines within an ungrounded (electrically "floating") intermediate conductor C that is suspended within an outer conductor D, as shown in Figure 8.3-10. The intermediate conductor C might seem to shield A and B from each other. However, C is not grounded, and its presence increases the distributed capacitance between A and B and, with it, their coupling to each other.

For the reentrant section

$$Z_{OE} = Z_{02} + 2Z_{01} \qquad (8.3\text{-}23)$$

$$Z_{OO} = Z_{02} \qquad (8.3\text{-}24)$$

where Z_{01} is the characteristic impedance of a solid conductor having the same outer dimensions as the intermediate conductor C, within the outer conductor D; and Z_{02} is the characteristic impedance of A (or B) within the coaxial region inside C.
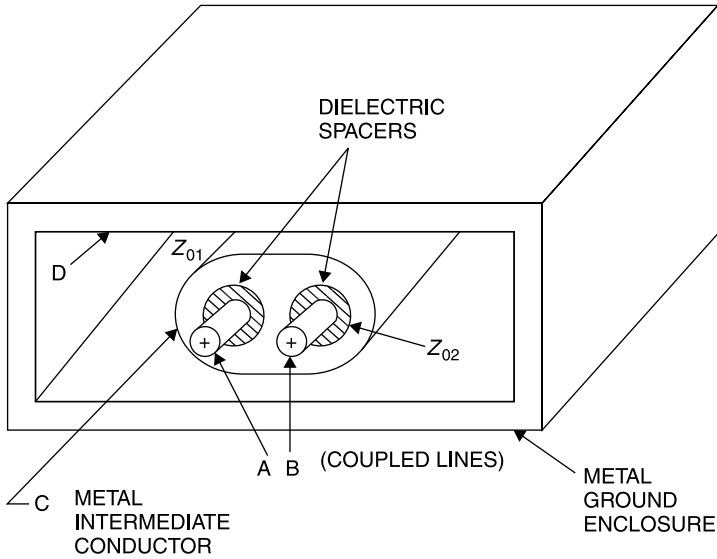
Figure 8.3-10 Cohn's reentrant coupling section [2].

8.4 REFLECTIVELY TERMINATED 3-dB COUPLER

As was derived in Section 8.3, when terminated in matched Z_0 loads defined by (8.3-8) and fed by a Z_0 generator at port 1, the backward wave coupler has no reflection at port 1 and no power output at port 3. The output voltages at ports 2 and 4 are given by (8.3-13) and (8.3-20), respectively. With this information the scattering matrix for this four-port network can be written directly by alternately assuming an input at each port with the remaining ports terminated in matched loads. The result is

$$
S = \begin{pmatrix}
0 & f_1 & 0 & f_2 \\
f_1 & 0 & f_2 & 0 \\
0 & f_2 & 0 & f_1 \\
f_2 & 0 & f_1 & 0
\end{pmatrix}
\tag{8.4-1a}
$$

where

$$
f_1 = \frac{jk \sin \theta}{\sqrt{1-k^2}\cos\theta + j\sin\theta}
\quad \text{and} \quad
f_2 = \frac{\sqrt{1-k^2}}{\sqrt{1-k^2}\cos\theta + j\sin\theta}
\tag{8.4-1b,c}
$$

The scattering matrix is useful for determining the performance of the coupler when it has multiple inputs as, for example, when reflective circuits terminate ports 2 and 4.
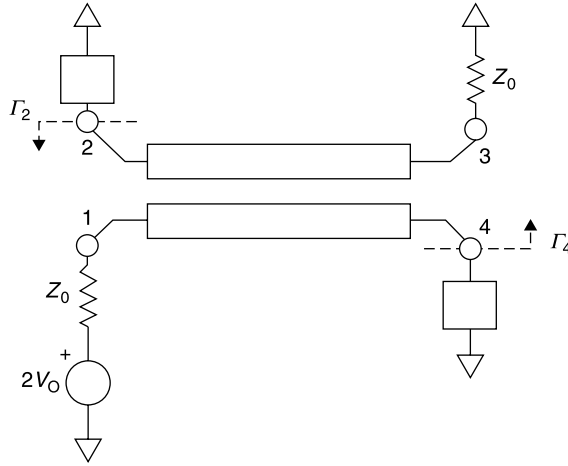
Figure 8.4-1 The 3-dB backward wave coupler with symmetric reflecting terminations on ports 2 and 4.

A frequent application consists of employing a 3-dB coupler to convert a symmetric pair of reflective circuits into a transmission network. This need may arise, for example, when it is necessary to design a diode phase shifter. The symmetric reflective terminations are connected to the direct arm (port 4) and coupled arm (port 2) of the coupler. Because of their equal amplitude and 90° phase difference for each pass through the coupler, when the energy is reflected back into the coupler (Fig. 8.4-1), it emerges from the normally decoupled port 3.

This process requires that the complex reflection coefficients at ports 2 and 4 are identical, or

$$\Gamma_2 = \Gamma_4 = \Gamma \tag{8.4-3}$$

With an input at port 1, symmetric reflections at ports 2 and 4, and a matched Z_0 load at port 3, the scattering matrix equations for the network are

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = \begin{pmatrix} 0 & f_1 & 0 & f_2 \\ f_1 & 0 & f_2 & 0 \\ 0 & f_2 & 0 & f_1 \\ f_2 & 0 & f_1 & 0 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} 0 & f_1 & 0 & f_2 \\ f_1 & 0 & f_2 & 0 \\ 0 & f_2 & 0 & f_1 \\ f_2 & 0 & f_1 & 0 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ \Gamma f_1 a_1 \\ 0 \\ \Gamma f_2 a_1 \end{pmatrix}$$

$$\tag{8.4-2a,b,c,d}$$

The reflection coefficient at input port 1 is found from (8.4-2a), specifically

$$b_1 = f_1 \Gamma f_1 a_1 + f_2 \Gamma f_2 a_1 \tag{8.4-5}$$

and the input reflection coefficient is

$$\Gamma_1 = \frac{b_1}{a_1} = \Gamma(f_1^2 + f_2^2) = \Gamma \frac{1 - k^2(1 + \sin^2 \theta)}{[\sqrt{1 - k^2} \cos \theta + j \sin \theta]^2} \qquad (8.4\text{-}6)$$

From (8.4-4) it can be seen that when $\theta = 90°$ and $k = 1/\sqrt{2}$ (for 3-dB coupling), then $\Gamma_1 = 0$, and the input port is matched. From (8.4-2c)

$$b_3 = f_2 \Gamma f_1 a_1 + f_1 \Gamma f_2 a_1 \qquad (8.4\text{-}7)$$

This can be written as

$$b_3 = a_1 \Gamma \frac{j(2k \sin \theta)\sqrt{1 - k^2}}{\sqrt{1 - k^2} \cos \theta + j \sin \theta} \qquad (8.4\text{-}8)$$

We note from (8.4-8) that when $\theta = 90°$ and $k = 1/\sqrt{2}$, and $|\Gamma| = 1$, the output wave $b_3 = \Gamma a_1$. That is, when symmetric, totally reflecting circuits are connected to ports 2 and 4 of a 3-dB coupler, the output wave equals the input in amplitude but is changed by the angle of the reflection coefficient of the reflective terminations at ports 2 and 4. This is what is desired in a phase shifter circuit.

As frequency is varied, the value of θ varies, and the power split of the coupler is no longer exactly 3 dB, as is required for a perfect match at the input with symmetric reflections at ports 2 and 4. As a result, some power is reflected at port 1 and a corresponding insertion loss is experienced in the transmission from port 1 to port 3. Despite this fact, the reflection coefficient at port 1 remains fairly low due to the fact that the waves reflected back into the coupler at ports 2 and 4 remain 90° out of phase with one another at all frequencies and hence combine 180° out of phase at port 1.

The input VSWR at port 1 and the transmission loss from port 1 to port 3 is shown in Figure 8.4-2 for a coupler with center frequency at 1500 MHz and symmetric, totally reflecting circuits at ports 2 and 4. The VSWR is 1.1 or less from 1100 to 1900 MHz, a bandwidth of 800 MHz. Notice that *for TEM devices whose bandwidth depends upon line lengths, the center frequency f_0 is the arithmetic mean of the band edge frequencies.*

The matched performance at port 1 with symmetric reflections at the 3-dB ports can be obtained with other types of 3-dB, 90° couplers such as the *branch line coupler* and the *hybrid ring coupler* (or *rat race coupler*) fitted with an extra 90° line length. The principles of these couplers will be covered shortly. However, these devices do not maintain the 90° phase difference at all frequencies for the signals emerging from their coupled and direct ports and do not achieve as broad a bandwidth in the reflection mode as that of the backward wave hybrid.

−2.7 dB Coupler
$Z_{OE} = 125$ ohm
$Z_{OO} = 20$ ohm
LENGTH = 50 mm
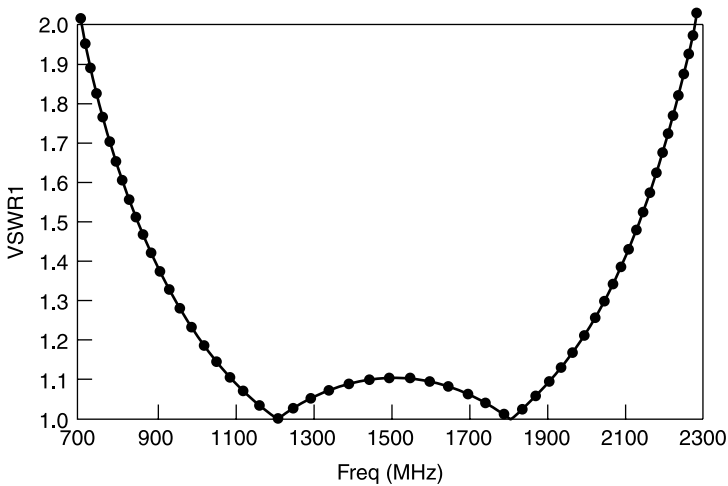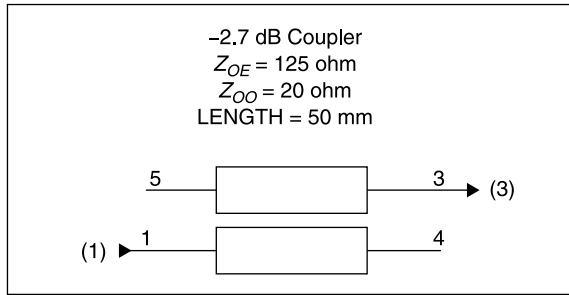
Figure 8.4-2 Input VSWR for a reflectively terminated −2.7-dB backward wave, hybrid coupler.

8.5 COUPLER SPECIFICATIONS

In the practical construction of a hybrid coupler, the region of the coupled lines can be constructed in a nearly ideal realization; however, it is necessary to make transitions from the ends of the coupled lines in order to attach connectors at the ports. It is at these transitions that significant reflections occur in a practical coupler, causing the operation of the coupler to depart from the ideal performance previously described.

Consider the coupler in Figure 8.5-1. Power P_I enters the coupler at port 1. Ideally, the desired coupled power P_C exits at port 2, and the remaining power P_L exits at port 4. However, due to coupler imperfections some power, P_D, exits at port 3 (the "decoupled" port). Furthermore, some power, P_R, is reflected at the input port.

The parameters of couplers are defined as numbers greater than unity so that the values are positive when expressed in decibels [1, p. 369].
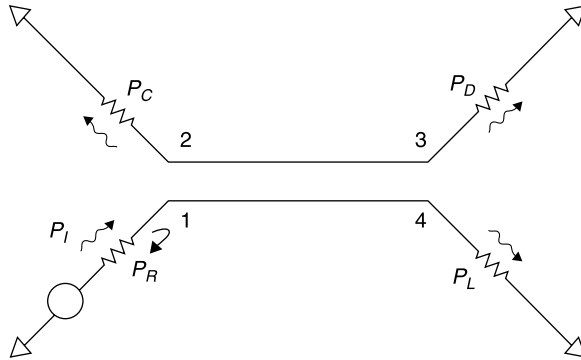
Figure 8.5-1 Practical directional coupler has return loss and finite directivity.

The *coupling* (in decibels) is defined as

$$\text{Coupling} = 10 \log\left(\frac{P_I}{P_C}\right)$$

The *directivity* (in decibels) is defined as

$$\text{Directivity} = 10 \log\left(\frac{P_C}{P_D}\right)$$

The *isolation* (in decibels) is defined as

$$\text{Isolation} = 10 \log\left(\frac{P_I}{P_D}\right)$$

Note that when expressed in decibels the isolation is equal to the coupling plus the directivity.

The *return loss* (in decibels) is defined as

$$\text{Return loss} = 10 \log\left(\frac{P_I}{P_R}\right)$$

The *insertion loss** (in decibels), which includes the diversion of coupled power, is

$$\text{Insertion loss} = 10 \log\left(\frac{P_I}{P_L}\right)$$

---

*Some manufacturers may subtract the coupled power from insertion loss ratings, defining the insertion loss as the amount of power loss to reflection and absorption in the coupler.

Note that in a symmetrically constructed coupler the same imperfections that give rise to reflections at the input also produce the undesired leakage at the directive port.

In a symmetric coupler the return loss equals the directivity. For example, a symmetric coupler having 20-dB directivity has a return loss of 20 dB as well. This corresponds to a VSWR of 1.22.

8.6 MEASUREMENTS USING DIRECTIONAL COUPLERS

Suppose we wish to use a directional coupler to determine the VSWR of a device under test (DUT) (Fig. 8.6-1). Consider the signals emanating from the coupler's measurement port. First, there is a voltage wave proportional to the desired measurement ρ_L. But then there is also an error voltage wave D due to the directivity. *It is to be emphasized that D is a voltage wave (not power); we can only use superposition for linear quantities (voltage, current, a and b waves, etc.)* Accordingly, there is a wave resulting from the interaction of the source and load reflections. This wave is the result of an infinite number of re-reflections. But this infinite series converges, as we've seen previously:

$$\rho_S\rho_L + (\rho_S\rho_L)^2 + (\rho_S\rho_L)^3 + \cdots = \frac{\rho_S\rho_L}{1 - \rho_S\rho_L} \tag{8.6-1}$$

We ignore the phases of the measured and error waves since we want to know the worst-case limits when the error waves either add or subtract from the measured ρ_L wave.

The total (combined) error wave can have magnitudes between

$$E_T = \text{total error wave magnitude} = \pm D\left[1 + \frac{\rho_S\rho_L}{1 - \rho_S\rho_L}\right] \tag{8.6-2}$$

Say that the coupler has 20-dB directivity, that the VSWR of the source is 1.3, that the VSWR of the coupler input is 1.22, and that the load we wish to measure actually has a VSWR of 1.25. Then $D = 0.1$, $\rho_S = 0.13$, and $\rho_L = 0.11$.
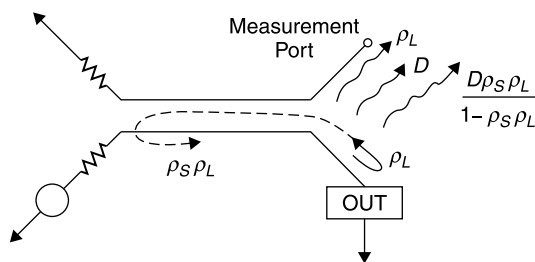
Figure 8.6-1 Measurement of the reflection coefficient of a device under test (DUT) using a backward wave coupler. Two error waves are also shown.

Then the error wave can be

$$E_T = \pm(0.1)\left[1 + \frac{(0.13)(0.11)}{1 - (0.13)(0.11)}\right] = \pm(0.1)[1.015] = \pm0.10 \quad (8.6\text{-}3)$$

This wave can combine in any phase with that of Γ_L, whose magnitude is 0.11. The resulting measured magnitude of Γ_L will be within

$$0.01 \leq \text{measured } \rho_L \leq 0.21 \qquad (8.6\text{-}4)$$

The corresponding values for measured VSWR are

$$1.02 \leq \text{VSWR}_{\text{MEAS}} \leq 1.53 \qquad (8.6\text{-}5)$$

This is insufficient accuracy for practical measurements. To circumvent this inaccuracy, the same coupler can be used in a *network analyzer* system, whose computer determines the error waves (phase and magnitude) during the calibration part of the measurement (when precision short circuit, open circuit and matched loads are connected) and then subtracts these error phasors from the actual measurements at each test frequency. This is described in the next section.

8.7   NETWORK ANALYZER IMPEDANCE MEASUREMENTS

Network analyzers (Fig. 8.7-1) measure both impedance (by measuring reflection coefficient at any port) and scattering parameters (insertion loss and inser-
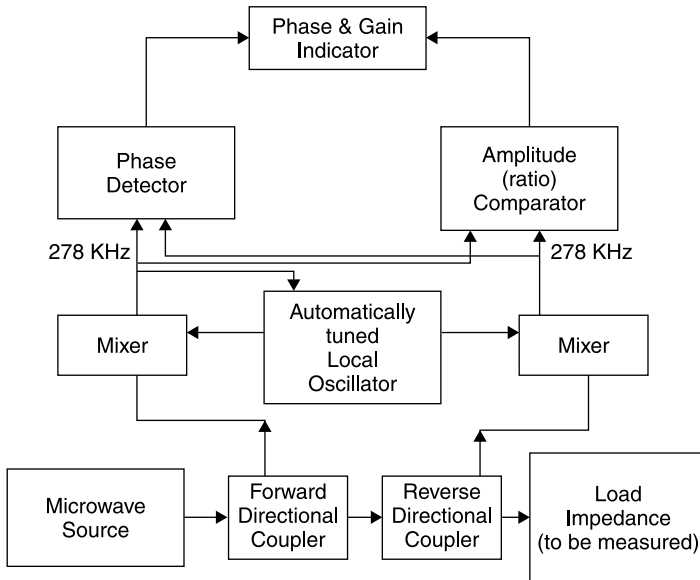
Figure 8.7-1   Network analyzer in the impedance measurement mode (after Gupta [7], reprinted with permission).

tion phase to any port). Impedance measurement is based on the direct measurement of reflection coefficient, the complex ratio of the reflected and incident waves of a terminating impedance. The incident and reflected waves are separated using directional couplers. The outputs of the couplers are fed to a harmonic frequency converter (mixer), which translates the test frequency to a fixed frequency near 300 kHz [7, pp. 114–115]. Usually an autotuning local oscillator and two identical mixers are used for this purpose. The mixing process preserves amplitude and phase information that are readily measurable at low frequency. The reflected wave vector, normalized to the incident wave (and corrected for coupler errors), can be displayed on a CRT (cathode ray tube) screen with a Smith chart overlay for direct determination of normalized impedance.

8.8 TWO-PORT SCATTERING MEASUREMENTS

The network analyzer (Fig. 8.8-1) contains microwave switches of high precision (good reproducibility and low VSWR). Using the switches the procedure to measure S_{nm} is [7, p. 115]:

1. Connect a source to port m and a matched load to port n. These two ports are connected to a network analyzer via two directional couplers as shown in Figure 8.8-1.
2. Terminate all other ports in matched loads so that all other a's (incident wave amplitudes) are zero.
3. The network analyzer provides the complex value of S_{nm} in terms of amplitude and phase (for $n \neq m$, i.e., a two-port measurement).
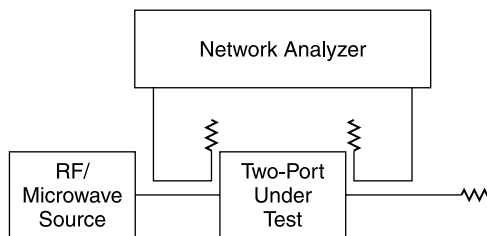
Figure 8.8-1   Network analyzer in the two-port measurement mode used to determine S parameters.

8.9 BRANCH LINE COUPLER

Over a narrow bandwidth, the performance of the 3-dB backward wave hybrid coupler can be obtained with the *branch line coupler*. This consists of four 90° transmission lines connected together in a square. It has the form of a 50-Ω line
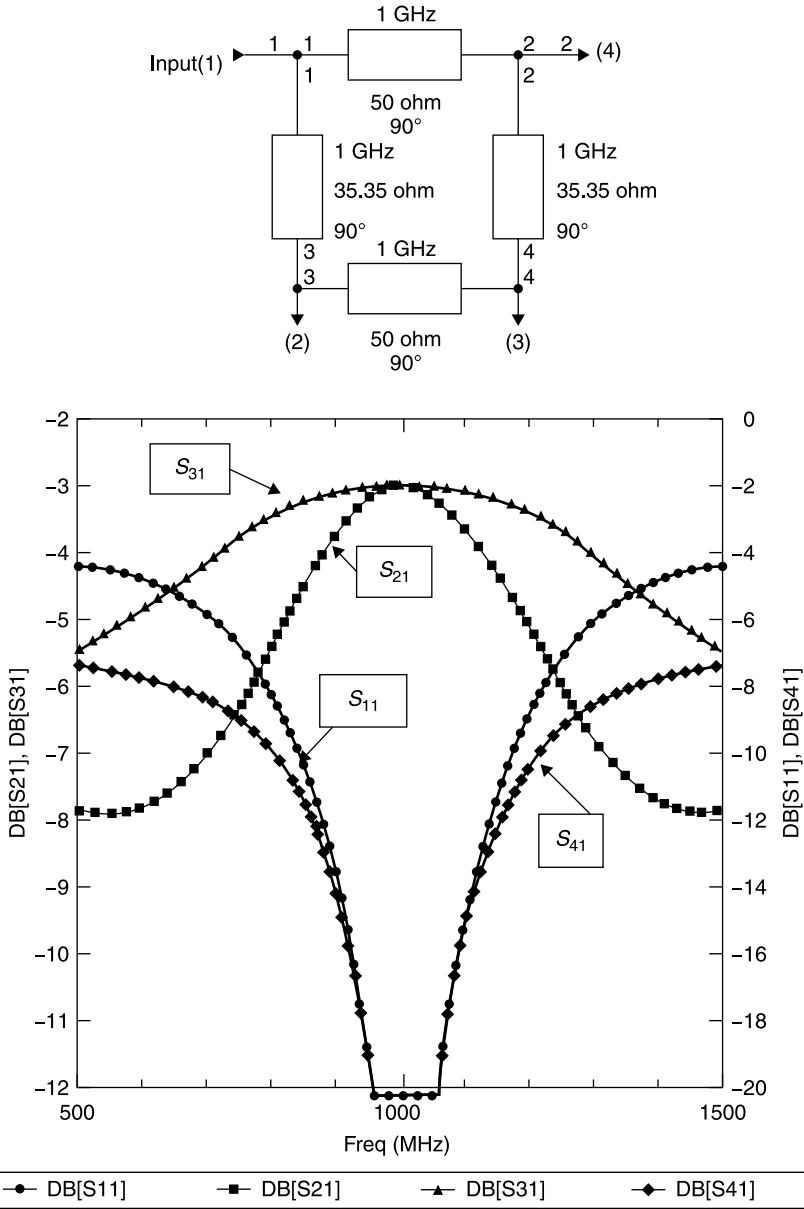
Figure 8.9-1 Equivalent circuit and performance of the 3-dB branch line coupler.

loaded at 90° intervals by 35.35-Ω (more precisely, $50/\sqrt{2}$) stubs that are interconnected at their ends with another 50-Ω line. The "stub ends" form the 3-dB ports. Like the backward wave hybrid, a 90° phase difference occurs between these two equal outputs, but for the branch line coupler this occurs only at the center frequency.
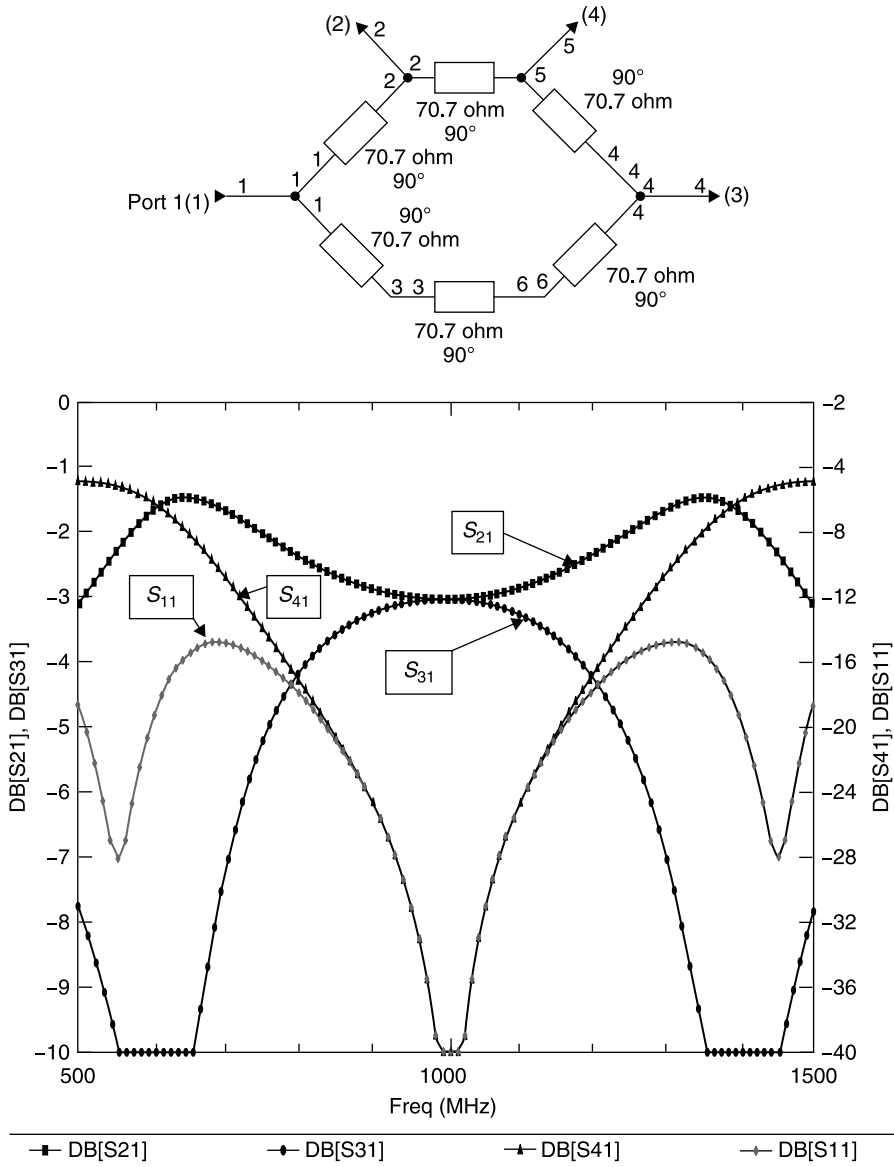
Figure 8.10-1   Circuit and performance of 3-dB hybrid ring coupler when realized in a homogeneous TEM transmission line format.

When power enters at port 1 (Fig. 8.9-1), it divides evenly between ports 2 and 3. At the center frequency, port 4 is isolated when match loads are placed at ports 2 and 3.

This coupler has the constructional advantage that it lies entirely in one center conductor plane, hence it is easy to print or incorporate within an integrated circuit. However, this coupler is fairly narrowband, usually having only about a 10% bandwidth. Furthermore, the 35.35-Ω lines are wide relative to their 90° length at microwave frequencies, making it hard to distinguish where those lines end and the 50-Ω lines begin, the problem encountered with the stub filter in Section 7.34. Accordingly, the branch line coupler is a candidate for EM simulation.

8.10   HYBRID RING COUPLER

The disadvantage of the branch line coupler is the need to make 35-Ω lines. Their low impedance line sections make the junctions so large it is difficult to estimate where one 90° line section ends and the next begins, necessitating considerable trial and error or EM simulation to realize a satisfactorily performing design.

The hybrid ring coupler, also called the *rat race coupler*, actually requires 1.5 wavelengths of transmission line (one section is 270° long), but it is easier to realize because the lines are 70.7 Ω (more precisely $Z_0\sqrt{2}$), quite narrow, and consequently their electrical lengths are more easily predicted. While this coupler is called the *hybrid ring*, the name is misleading, since its 3-dB ports are 180° out of phase rather than 90° (Fig. 8.10-1). The hybrid ring coupler can be realized in a TEM transmission line format or in a waveguide. In a waveguide, its circular chamber is likely the basis for the ''rat race'' name.

Remarkably, for a given VSWR specification the rat race usually is broader band than the branch line coupler. Although its 3-dB outputs are 180° out of phase rather than 90°, a 90° phase difference in its −3-dB outputs can be obtained simply by adding a 90° length of 50-Ω line to one of the 3-dB outputs. The 90° phase difference between 3-dB outputs is necessary if the coupler is to be used in a reflection transmission circuit, as for a phase shifter.

8.11   WILKINSON POWER DIVIDER

An equal-phase output, 3-dB divider can be obtained by transforming two 50-Ω loads to 100 Ω and connecting them in parallel to an input 50-Ω line (Fig. 8.11-1). When the output ports are bridged with a 100-Ω resistor, the result is a power divider whose output ports are matched and perfectly isolated at the design frequency [1, Sec. 8.2]. The resulting circuit is called a *Wilkinson power divider* (Fig. 8.11-2). To the extent that the output loads at 2 and 3 are
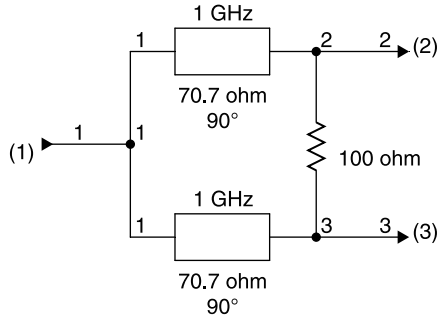
Figure 8.11-1   Wilkinson power divider provides two in-phase, 3-dB outputs and some isolation of the input from asymmetrical.
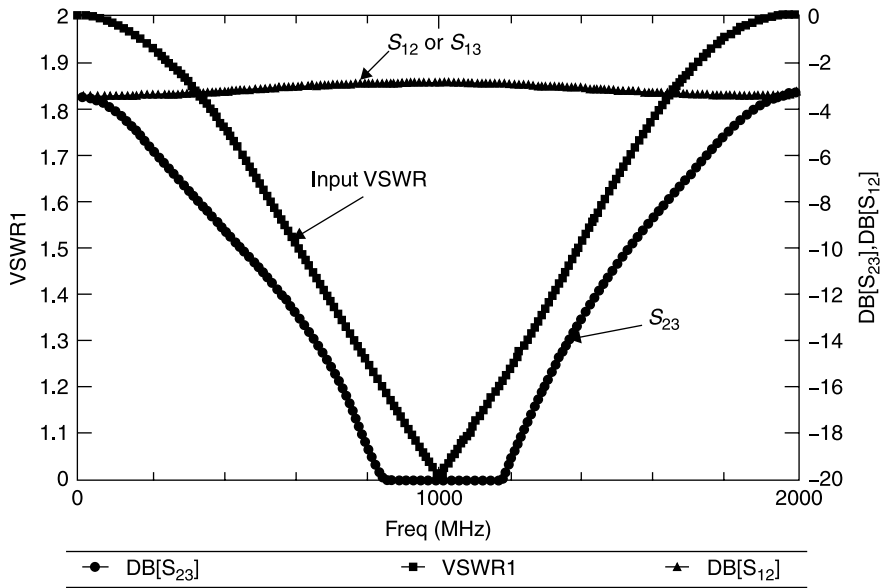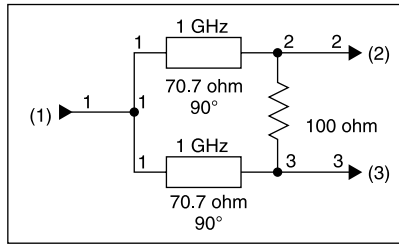
Figure 8.11-2   Performance of the Wilkinson power divider.

asymmetric, their reflections are absorbed by the 100-Ω resistor. When the loads are symmetric, no power will be absorbed in this resistor.

Another advantage of the Wilkinson divider is that very broad bandwidth can be obtained by using multiple section transformers (quarter-wave line sections) to transform the 50-Ω loads to 100-Ω inputs that, when paralleled, give an input impedance of 50 Ω. [4]

REFERENCES

1. Peter A. Rizzi, *Microwave Engineering, Passive Circuits*, Prentice-Hall, Englewood Cliffs, NJ, 1988. *Fundamental microwave engineering textbook covering theory and design of transmission lines, couplers, filters, and numerous other passive devices.*

2. Seymour Cohn, "The Re-Entrant Cross Section and Wide-Band 3-dB Hybrid Couplers," *IEEE Transactions on Microwave Theory and Techniques, MTT-11*, July, 1963, pp. 254–258.

3. Joseph F. White, *Microwave Semiconductor Engineering*, Noble Publishing, Norcross GA. *Extensive treatment of PIN diodes and their switching and phase shifting applications. Also includes fundamentals of microwave circuits.*

4. Matthei, Young, and Jones, *Microwave Filters, Impedance Matching Networks, and Coupling Structures*, McGraw-Hill, New York, 1964 (now also available from Artech House, Norwood, MA). *This is considered the Bible of the filter world.*

5. Fred E. Gardiol, *Introduction to Microwaves*, Artech House, Norwood, MA, 1984. *This is a good text on microwaves, containing many practical experimental setups.*

6. Robert E. Collin, *Foundations of Microwave Engineering*, McGraw-Hill, New York, 1966. *Collin is a superb theoretician.*

7. K. C. Gupta, *Microwaves*, Halsted Press, Wiley, New York, 1978. *Easily readable introduction to microwave engineering. Originally published in India.*

EXERCISES

E8.3-1   Derive the expressions (8.3-23) and (8.3-24) for Z_{OE} and Z_{OO} of the backward wave coupler in terms of the characteristic impedance Z_0 and the voltage coupling coefficient k, beginning with (8.3-8) and (8.3-11).

E8.3-2   **a.** What even and odd impedances are needed to form a 50-Ω, 15-dB coupler? *Hint:* The result from E8.3-1 is helpful.

   b. What is the physical length of the coupling region if the relative dielectric constant throughout the insulating region surrounding the coupled lines is 2.1?

E8.3-3   Verify that the numerator N_A^+ in (8.3-7) is zero given the matching condition of (8.3-8).

E8.3-4   Derive the expression for the straight-through voltage V_4 of the backward coupler (8.3-20) beginning with (8.3-18) and using the conditions of (8.3-8) and (8.3-11).

E8.4-1   **a.** Determine the coupling coefficient k for a 50-Ω, backward wave coupler that when used with perfectly reflecting short circuits on its coupled and direct arms will have a perfect input match at $f_0 - 0.3f_0$ and at $f_0 + 0.3f_0$ where $f_0 = 1000$ MHz is its center frequency (at which the coupling region is 90° long).

      b. What is the center frequency coupling value (in decibels) of the coupler?

      c. What are Z_{OE} and Z_{OO} for this coupler?

      d. What is the VSWR at the input at 1000 MHz?

      e. Model the circuit on a network simulator and verify your predicted values.

E8.4-2   **a.** Use a network simulator to model a 3-dB backward wave coupler with open circuits on its direct and coupled arms (ports 2 and 4) and center frequency of 1 GHz (at which the coupled line length is 90°). Plot the input VSWR (at port 1) from 500 to 1500 MHz.

      b. Repeat (a) with short-circuit terminations on ports 2 and 4 instead of open circuits.

      c. Using a network simulator, plot the transmission phase $(\text{ang}[S_{31}])$ over the 500- to 1500-MHz band of the circuits in (a) and (b) and show that the difference (phase shift obtained by switching from an open to a short circuit on ports 2 and 4) is equal to 180° over a frequency band of 10% or more. This is an example of a transmission phase shifter obtained by converting variable reflection angle terminations into a two-port network using a 3-dB coupler.

      d. Plot the insertion loss of the circuits in (a) and (d) over the 500- to 1500-MHz bandwidth. With what minimum efficiency, in percent of incident power, do the circuits in (a) and (b) convert reflecting terminations into a transmission circuit over the 700- to 1300-MHz bandwidth?

      e. Show the VSWR of the cascade connection of the circuits in (a) and (b) over the 500- to 1500-MHz bandwidth.

E8.5-1   A symmetrically designed backward wave coupler has 30 dB of directivity at its center frequency. What is its input VSWR at this frequency.

E8.6-1   An engineering team at your microwave components company proposes to design a dedicated test setup to measure products having a VSWR of 1.25 at a single frequency. They propose to build a coupler and tune it for low VSWR and high directivity at a single test frequency. The source available has a VSWR of 1.3. It is necessary to

measure an actual VSWR value of 1.25 as being between 1.23 and 1.27.

    a. What coupler directivity is required?

    b. What is the corresponding input VSWR of the coupler?

    c. Is this measurement setup practically achievable?

E8.9-1   **a.** Use a circuit simulator to model the branch line 3-dB coupler with open circuits at its 3 dB ports. What is its bandwidth for a maximum input VSWR = 1.2 (20 dB return loss)?

    b. Repeat (a) but place short circuits on the two 3-dB ports. What is the bandwidth for VSWR = 1.2 or less?

    c. Next, cascade two of the circuits used in (a) and interconnect them with a 50-Ω line that is 600 long at 1 GHz. What is the 1.2 maximum VSWR bandwidth of the cascade? How does this compare with the similar experiment of exercise E8.4-2e?

E8.10-1 Repeat Exercise 8.9-1 this time using the TEM mode hybrid ring (rat race) coupler.

    a. How does the VSWR bandwidth compare to that of the branch line?

    b. When you cascade two circuits, how does the performance compare with the cascaded branch line couplers?

E8.11-1 The Wilkinson power divider shown in Figure 8.11-1 uses single quarter-wave inverters to transform the 50-Ω loads to 100 Ω at its tee junction.

    a. Design new inverters with a center frequency of 1000 MHz having three cascaded 90° line sections to obtain broader bandwidth. What are the characteristic impedance values for the three line sections? *Hint:* Use the constant ratio step method used in the multisection Q matching.

    b. Use a network simulator to plot the VSWR of your divider and compare it to that of the single inverter mode over the 0- to 2000-MHz frequency interval. Model your design using a network simulator. How broad a −20 dB return loss bandwidth can you get?

    c. Based on the single inverter design shown in Figure 8.11-1, select the values of the three bridging resistors to be used in your design and use the network simulator to plot the isolation obtained over the 0- to 2000-MHz band with your choices. *Hint:* Guess starting values for the three resistors and then use the optimizer of the network simulator with a goal of 30-dB minimum isolation between the output ports from 500 to 1500 MHz.

Filter Design

9.1 VOLTAGE TRANSFER FUNCTION

In the specification and design of filters we are interested in the ratio of the sinusoidal voltage at the load, V_L, relative to that of the generator, V_G, as a function of frequency. Classical filter theory begins with LLFPB (lumped, linear, finite, passive, and bilateral) elements. These designs can be extended to high frequencies at which distributed elements are employed having resonances, such as reactively terminated transmission lines with zero and infinite input impedance values.

In Section 2.9 we proved that maximum power is transferred between a generator and load when the impedance of the generator, Z_G, is related to the impedance of the load, Z_L, by $Z_G = Z_L^*$. In our treatment of filters we consider only cases in which the source and load impedances are both real. Then $Z_G = R_G$ and $Z_L = R_L$ as shown in Figure 9.1-1. Under these conditions [1, pp. 11–13], *the available voltage V_{avail} is defined as the voltage across the load when all of the available power from the generator is transferred to the load.*

When source and load have different resistances

$$V_{\text{avail}} = \sqrt{\frac{R_L}{R_G}}\,\frac{V_G}{2} \tag{9.1-1}$$

Dividing both sides by V_L,

$$\frac{V_{\text{avail}}}{V_L} = \frac{1}{2}\sqrt{\frac{R_L}{R_G}}\,\frac{V_G}{V_L} \tag{9.1-2}$$

The *voltage transmission coefficient t* is defined as

$$t = \frac{V_L}{V_{\text{avail}}} = 2\sqrt{\frac{R_G}{R_L}}\,\frac{V_G}{V_L} \tag{9.1-3}$$
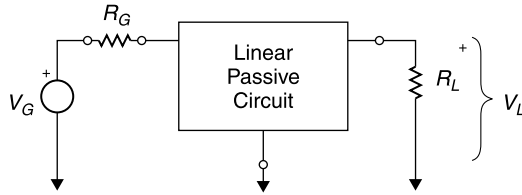
Figure 9.1-1 Two-port network driven by a voltage source and terminated at both ports.

For the filters that we shall consider, t will be unity or less. Since we wish to describe the insertion loss ratio as a number equal to unity or greater, we define the *transfer function of a network* as H where

$$H = \frac{1}{t} \qquad (9.1\text{-}4)$$

Then the *attenuation of the network*, also called the *transducer loss of the network* (previously defined in Section 2.10), in decibels, is

$$TL = 10 \log \frac{P_{\text{avail}}}{P_L} = 20 \log|H| \qquad (9.1\text{-}5)$$

where $P_L = V_L^2/R_L$. These definitions are the basis by which the attenuation of filters can be described. For the filter examples to be presented in this chapter $R_G = R_L = Z_0$ in which case *transducer loss* and *insertion loss* are equal, and the attenuation of the network can be written as

$$IL = -20 \log|S_{21}| \qquad (9.1\text{-}6)$$

where S_{21} is one of the S parameters of the network defined under the condition that source and load are equal to the real value Z_0. The format of (9.1-6) is used because this is an available output of the network simulator used to evaluate the filter design examples.

9.2 LOW-PASS PROTOTYPE

Many microwave filters are designed based upon a *low-pass filter prototype*, consisting of alternating shunt admittances and series impedances, or vice versa, series impedances and shunt admittances (Fig. 9.2-1). For nearly all designs lossless elements are employed, thus the alternating elements are actually reactances and susceptances. As lumped circuits these are L and C elements.

When the input impedance Z_{IN} of a three-element ladder network with a 1-Ω load and maximally flat low-pass response with 3-dB cutoff frequency of 1
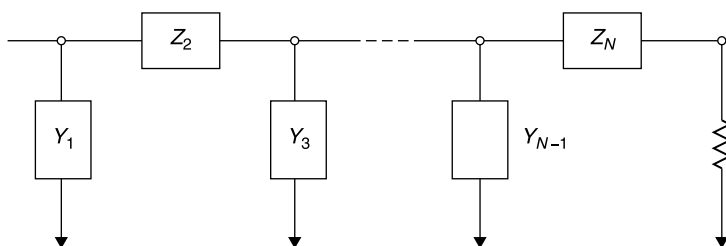
Figure 9.2-1 Filter prototype ladder network, synthesized by continued fraction expansion. Shunt elements are susceptances and series elements are reactances. A resistive termination loads the network.

rad/s is written as a fractional expansion, it takes the form [1, p. 26]

$$Z_{\text{IN}} = \cfrac{1}{s + \cfrac{1}{2s + \cfrac{1}{s + 1}}} \qquad (9.2\text{-}1)$$

where $s = \sigma + j\omega$ is used to express the transfer function of the filter. The corresponding values for the elements of the three-element ladder network are

$$Y_3 = 1s \qquad C_3 = 1 \text{ F} \qquad\qquad (9.2\text{-}1a)$$

$$Z_2 = 2s \qquad L_2 = 2 \text{ H} \qquad\qquad (9.2\text{-}1b)$$

$$Y_1 = 1s \qquad C_1 = 1 \text{ F} \qquad\qquad (9.2\text{-}1c)$$

This example defines an $N = 3$, low-pass filter.

In a similar fashion the element values for filters having 2 to N elements can be determined. The element values are called g *values* because the filter can be initiated with either a series L or a shunt C to which alternating elements are added. Thus the same set of g values for an N element filter can be used for either of two designs, depending upon whether one begins with a series or a shunt element. Thus, starting with a series inductor, the g_1 value is its inductance in henrys, while starting with a shunt capacitor requires interpreting the g_1 value as the capacitance in farads. Thereafter the elements of the filter alternate both in kind (inductance or capacitance) *and* in topology (series or shunt).

9.3 BUTTERWORTH OR MAXIMALLY FLAT FILTER

The maximally flat low-pass prototype filter is called a *Butterworth* response. It has the attributes that *the transducer loss and all of its derivatives are zero at zero frequency*; hence it provides maximum flatness.

TABLE 9.3-1 g Values for Butterworth Low-Pass Prototype Filter

N	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}
2	1.4142	1.4142	1								
3	1.0000	2.0000	1.0000	1							
4	0.7654	1.8478	1.8478	0.7654	1						
5	0.6180	1.6180	2.0000	1.6180	0.6180	1					
6	0.5176	1.4142	1.9318	1.9318	1.4142	0.5176	1				
7	0.4450	1.2470	1.8019	2.0000	1.8019	1.2470	0.4450	1			
8	0.3902	1.1111	1.6629	1.9616	1.9616	1.6629	1.1111	0.3902	1		
9	0.3473	1.0000	1.5321	1.8794	2.0000	1.8794	1.5321	1.0000	0.3473	1	
10	0.3129	0.9080	1.4142	1.7820	1.9754	1.9754	1.7820	1.4142	0.9080	0.3129	1

Source: After Rhea [1, p. 47]. Used with permission.

This is because it is a mathematical fact that *if two polynomials of order N have N points in common they are identical.* Consequently, if by alternate means we find a filter function having a zero value and zero derivatives at zero frequency, we can be confident that it is the one and only such solution, the Butterworth.

The order N of a filter is equal to the number of poles and zeros in its transmission response. In the stopband of a filter N determines how rapidly the isolation increases after the cutoff frequency (it increases at $6N$ dB/octave). In the passband N equals the number of derivatives of the response characteristic. This affects the flatness of the passband response or, in the case of the Chebyshev response to be described, the number of ripples in the passband.

For a lumped-element filter, N is equal to the number of nontrivial elements in the filter network. Nontrivial elements are those which alternate in type between L and C elements. In a filter, nontrivial L and C elements each produce a pole at zero frequency and a zero at infinite frequency or vice versa.

Trivial elements are those that can be combined without changing the response. For example, two capacitors in series or in parallel count as only one *nontrivial element* because they could be replaced by a single capacitor having an appropriately defined value. Similarly, two inductors in series or parallel also would count as only one nontrivial element.

In a filter network a resonator consisting of one L in series or in parallel with one C element counts as two nontrivial elements. Such a resonator produces a pole or zero in the filter response at a finite frequency.

The low-pass prototype Butterworth filter response is obtained by having alternating L and C elements with the values listed in Table 9.3-1, normalized to 1 Ω and having a cutoff frequency (3 dB loss) of 1 rad/s. Notice that there is an additional g value listed for each filter type. For example, for $N = 2$ there is a g_3 value listed. This is the value of the load with which the filter delivers its response. The generator impedance in these tables is always assumed to be a unity resistance. Note that *the g_i values in the prototype filter tables correspond to the value of series reactance or shunt susceptance of the respective lumped elements at the cutoff frequency f_C.* For the prototype filter circuit f_C is 1 rad/s.

All Butterworth filters have a unity load, but we shall see that for certain filters a nonunity load would be required to obtain the specified response (Fig. 9.3-1). Since there is no practical way to build an ideal transformer to transform a Z_0 load to another real value at all frequencies simultaneously, we will avoid filter design examples requiring unequal source and load resistances, except in the special case of diplexer filters.

9.4   DENORMALIZING THE PROTOTYPE RESPONSE

Of course, we do not want a 1-Ω filter with cutoff at 1 rad/s (0.16 Hz). Therefore, we scale both the impedance and frequency of the filter by adjusting the L
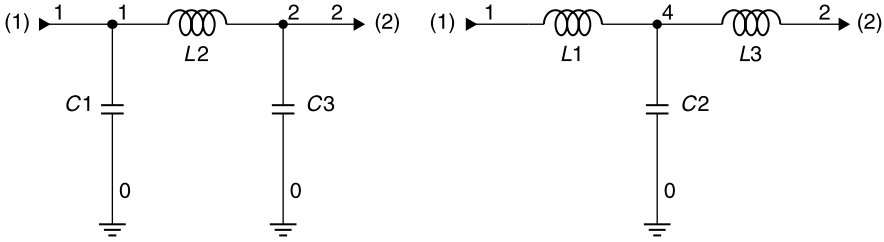
Figure 9.3-1 Either of these networks yields the same Butterworth low-pass response into a unity load.

and C values. Since the reactance of an inductor is

$$X_i = \omega L_i \tag{9.4-1}$$

it follows that to obtain a series impedance at our desired characteristic impedance Z_0 and cutoff frequency f_C, we must multiply the prototype g_i value so that

$$L_i = \frac{g_i Z_0}{2\pi f_C} \tag{9.4-2}$$

Similarly, since the reactance of a capacitor is

$$X_i = \frac{1}{\omega C_i} \tag{9.4-3}$$

then the corresponding value of C_i for our Z_0, f_C and g_i is

$$C_i = \frac{g_i}{Z_0 2\pi f_C} \tag{9.4-4}$$

In these filter design examples Z_0 is the impedance of the system. Later, if a lumped-element filter is used in a high frequency network, the Z_0 impedance of source and load will be presented to the filter when the filter is connected to them using transmission lines of Z_0 characteristic impedance.

As a filter design example, suppose that we wish to design a low-pass Butterworth filter with 3-dB cutoff at 1 GHz and $Z_0 = 50 \ \Omega$. The g values are

$$g_1 = 1.000 \tag{9.4-5a}$$

$$g_2 = 2.000 \tag{9.4-5b}$$

$$g_3 = 1.000 \tag{9.4-5c}$$
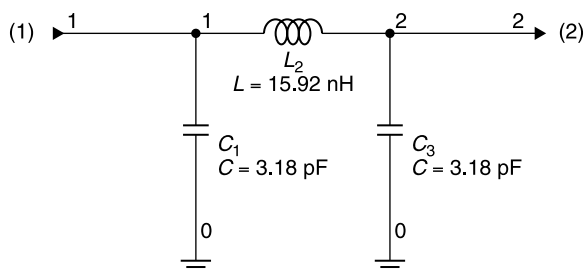
Figure 9.4-1 Three-element Butterworth low-pass filter with 3-dB cutoff at 1 GHz and $Z_0 = 50 \, \Omega$.

Furthermore, let's say that we want the filter to be in the π configuration, having a C as the first (shunt) element. Then

$$C_1 = C_3 = \frac{1.000}{(50)(2\pi)(1 \times 10^9)} = 3.18 \text{ pF} \qquad (9.4\text{-}6a,b)$$

$$L_2 = \frac{(2.000)(50)}{(2\pi)(1 \times 10^9)} = 15.92 \text{ nH} \qquad (9.4\text{-}6c)$$

The filter, scaled to our frequency and impedance, is shown in Figure 9.4-1.

In the network simulation of the circuit in Figure 9.4-1 the implicit source and load impedances are 50 Ω in the specifications of input port (1) and output port (2). The response of the filter in Figure 9.4-1 is shown in Figure 9.4-2.

In Figure 9.4-2 the lower curve uses the left y axis coordinates (expanded insertion loss), and the upper curve uses the right y axis (to show isolation), where the terms *insertion loss* (commonly used to describe transducer loss) and *isolation* are selected to describe the intended *passband* and *stopband* of the filter, respectively. Note that the insertion loss is 3 dB at 1 GHz as predicted.

Next, note that the isolation at 2 GHz is 18 dB and that the isolation at 4 GHz is 36 dB, an increase of 18 dB/octave. This also is expected for a three-element filter for which the isolation increases by $6N$ dB/octave, or 18 dB/octave. Had we used a five-element filter, the isolation would have increased by 30 dB/octave.

Filters made with lossless L and C elements achieve their frequency selective attenuation by reflecting power. For the low-pass Butterworth example the insertion loss at 1 GHz is 3 dB. This means that the reflected power is 50% of that which is incident. The corresponding reflection coefficient magnitude is 0.707 and the VSWR is 5.82. Figure 9.4-3 shows the VSWR of this low-pass filter example.

It should be emphasized that because reactive filters have high reflection coefficients their performance can be changed if they are interconnected with
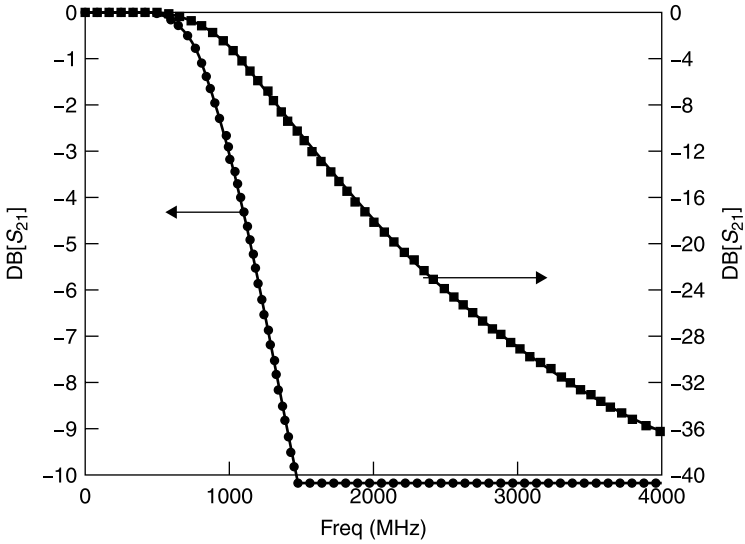
Figure 9.4-2   Insertion loss/isolation of the three-element low-pass filter with 1 GHz, 3-dB cutoff, and 50 Ω impedance.

other networks that also have appreciable reflections. The reflection interaction or *mismatch error*, as described in Section 4.7, can dramatically change the insertion loss performance from that expected using filter design theory. For example, suppose two of the three-element Butterworth low-pass filters as described in the preceding example are cascaded with a 60° long interconnecting
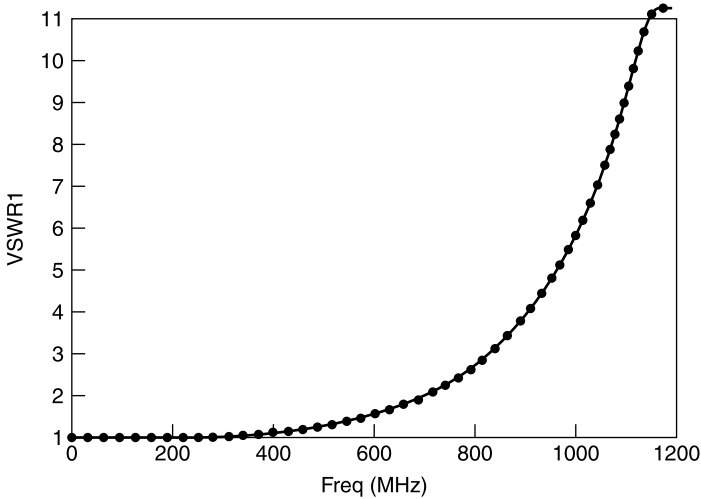
Figure 9.4-3   VSWR of the three-element Butterworth low-pass filter.
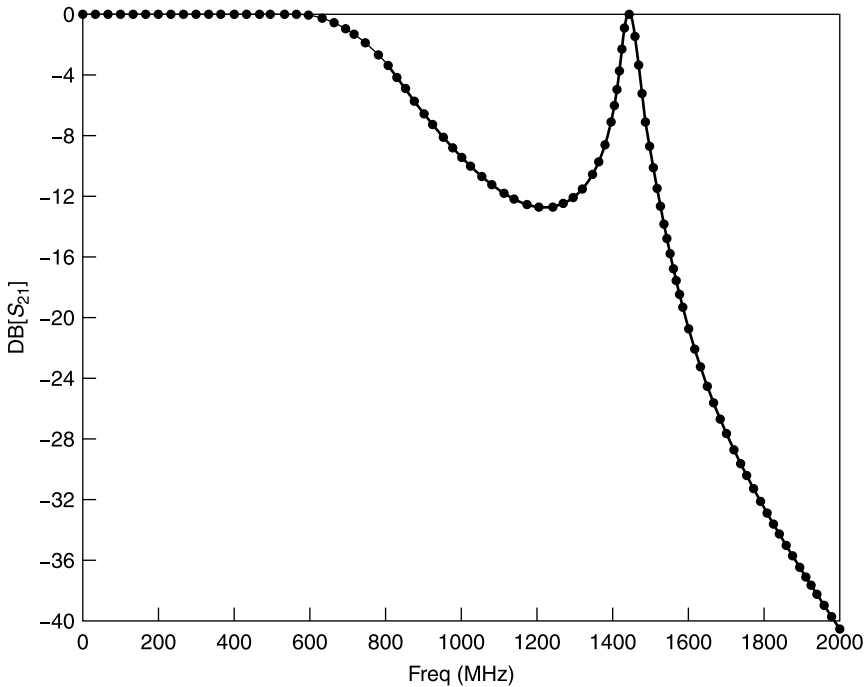
Figure 9.4-4 Two of the three-element Butterworth low-pass filters cascaded with a 60° long 50-Ω transmission line between them.

length of 50-Ω transmission line between them. The resulting insertion loss with frequency is shown in Figure 9.4-4.

From Figure 9.4-2 we might have expected about 8 dB of isolation from each filter at 1.4 GHz. However, with this spacing the reflective interactions of the two filters cause a canceling of their separate reflective attenuations, resulting in no net attenuation for the cascade at 1.4 GHz. For this reason, *it is not sufficient to design a reactive filter without also considering the environment in which it will be used.*

9.5 HIGH-PASS FILTERS

To design a highpass filter:

1. Begin with the low-pass prototype, replacing all *C*'s with *L*'s and all *L*'s with *C*'s, respectively.
2. Invert all g_i values.
3. Denormalize the resulting high-pass prototype *C*'s and *L*'s using
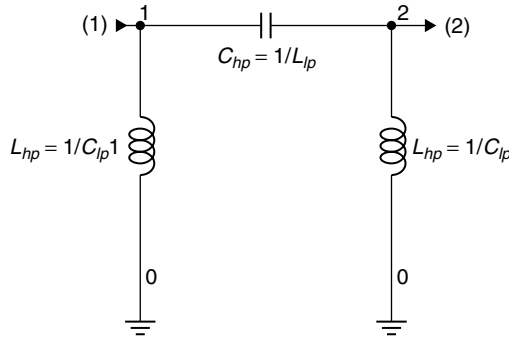
Figure 9.5-1 Conversion of the low-pass prototype filter to the high-pass prototype filter.

$$L_i = \frac{g_i Z_0}{2\pi f_C} \tag{9.5-1}$$

$$C_i = \frac{g_i}{Z_0 2\pi f_C} \tag{9.5-2}$$

For example, if we wish to convert our three-element low-pass filter to a high-pass filter, the prototype circuit is as shown in Figure 9.5-1 and the highpass prototype element values are

$$C_{hp} = \frac{1}{L_{lp}} = \frac{1}{2.000} = 0.500 \tag{9.5-3}$$

$$L_{hp} = \frac{1}{C_{lp}} = \frac{1}{1.000} = 1.000 \tag{9.5-4}$$

Denormalizing these to our 50 Ω impedance and 1 GHz cutoff frequency gives

$$L_i = \frac{g_i Z_0}{2\pi f_C} \tag{9.5-5}$$

$$C_i = \frac{g_i}{Z_0 2\pi f_C} \tag{9.5-6}$$

$$L_1 = L_3 = \frac{(1.000)(50)}{(2\pi)(1 \times 10^9)} = 7.96 \text{ nH} \tag{9.5-7}$$

$$C_2 = \frac{0.500}{(50)(2\pi)(1 \times 10^9)} = 1.59 \text{ pF} \tag{9.5-8}$$

The resulting circuit and performance is shown in Figure 9.5-2. Notice that the isolation at 250 MHz is about 18 dB greater than at 500 MHz, consistent with the 18-dB/octave isolation slope for a third-order filter.
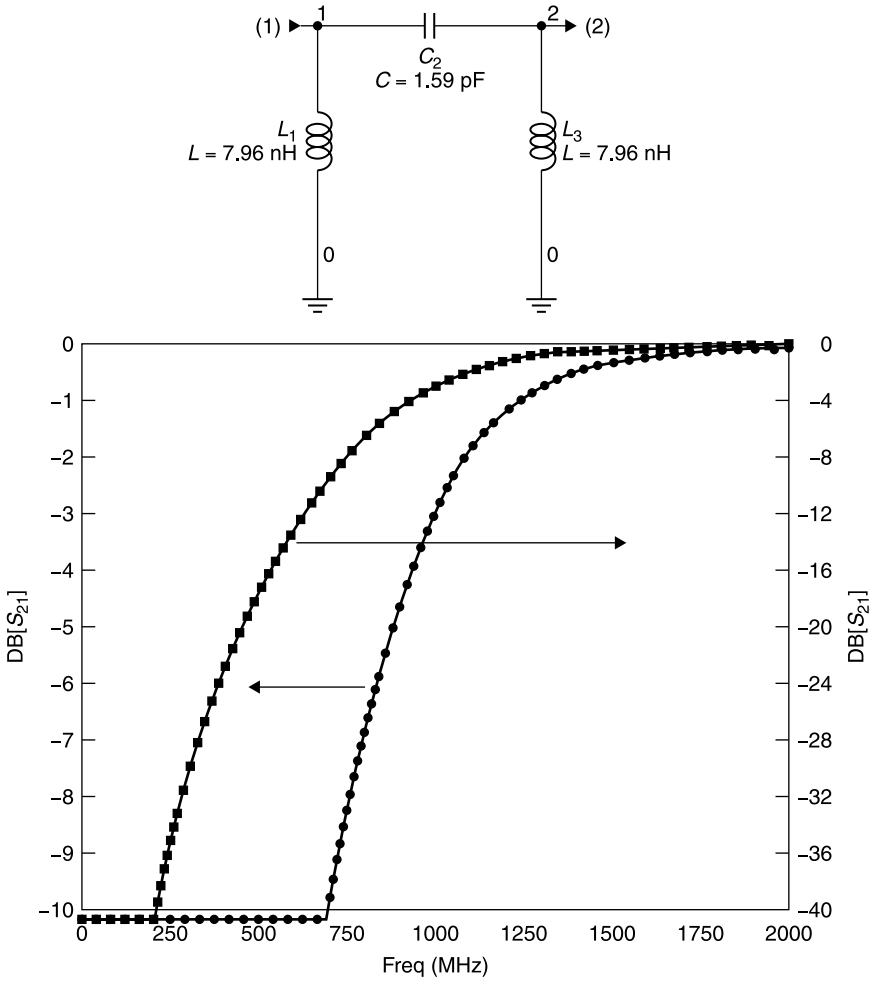
Figure 9.5-2   Loss/isolation of the high-pass filter.

9.6   BANDPASS FILTERS

Lumped-element bandpass and bandstop filters can be designed by starting with the low-pass and high-pass prototypes presented in the preceding sections. It is also possible to design transmission equivalents to these filters using Richard's transformation and Kuroda's identities [2, Sec. 9.5]. We will cover these topics shortly. For these reasons the low-pass and high-pass prototypes are important beyond their immediate importance as lumped-element filters. This and the following sections will concentrate on the lumped-element filters, which are common in high frequency applications. Their use extends well into
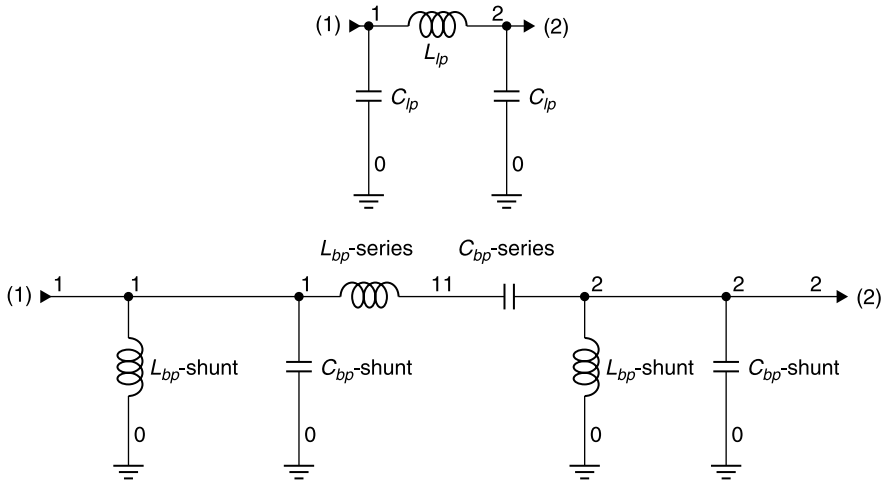
Figure 9.6-1 Transformation relations to convert the low-pass prototype filter to a bandpass prototype filter.

the microwave frequency bands as integrated circuit techniques allow realization of the small values of L and C that they require.

To design a bandpass filter starting with a low-pass prototype [1, p. 146] (Fig. 9.6-1):

1. Define the following:

$$f_l = \text{lower 3-dB cutoff frequency} \qquad (9.6\text{-}1)$$

$$f_u = \text{upper 3-dB cutoff frequency} \qquad (9.6\text{-}2)$$

$$f_0 = \text{geometric center frequency}$$

$$f_0 = \sqrt{f_l f_u} \qquad (9.6\text{-}3)$$

$$\text{BW} = f_u - f_l = \text{absolute bandwidth} \qquad (9.6\text{-}4)$$

$$\text{bw} = \frac{\text{BW}}{f_0} = \text{fractional bandwidth} \qquad (9.6\text{-}5)$$

$$\text{Percentage bandwidth} = 100\%(\text{bw}) \qquad (9.6\text{-}6)$$

2. Replace each series inductor with a series capacitor and series inductor resonator.
3. Replace each shunt capacitor with a shunt capacitor and parallel shunt inductor resonator.
4. Create the prototype (normalized) bandpass filter, assigning its elements the values shown below.

Transformed shunt elements:

$$C_{\text{bp-shunt}} = \frac{C_{\text{lp}}}{\text{bw}} \qquad (9.6\text{-}7)$$

$$L_{\text{bp-shunt}} = \frac{1}{C_{\text{bp-shunt}}} \qquad (9.6\text{-}8)$$

Transformed series elements:

$$L_{\text{bp-series}} = \frac{L_{\text{lp}}}{\text{bw}} \qquad (9.6\text{-}9)$$

$$C_{\text{bp-series}} = \frac{1}{L_{\text{bp-series}}} \qquad (9.6\text{-}10)$$

5. Denormalize all elements to obtain the final filter design. Note: Use f_0 in place of f_c. Thus,

$$L_i = \frac{g_i Z_0}{2\pi f_0} \qquad (9.6\text{-}11)$$

$$C_i = \frac{g_i}{Z_0 2\pi f_0} \qquad (9.6\text{-}12)$$

We now design a bandpass example based on our previous three-element low-pass prototype (Fig. 9.6-2). Suppose we wish a 3-dB passband from 800 to 1200 MHz. Then, BW = 400 MHz and

$$f_0 = \sqrt{(800)(1200)} = 979.8 \text{ MHz}$$

$$\text{bw} = 400/979.8 = 0.4082$$
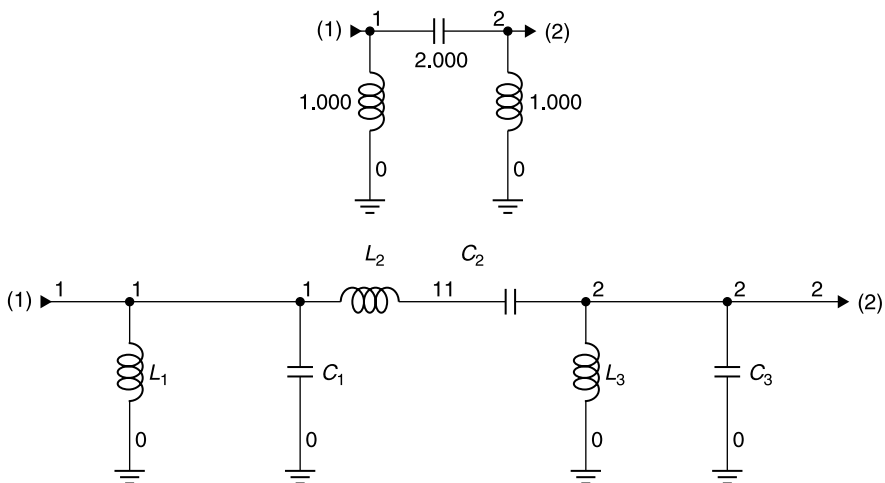
Figure 9.6-2   Schematic for the transition from low-pass to band-pass filter prototype.
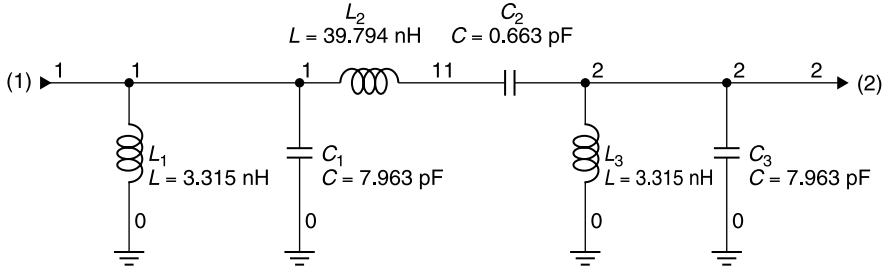
Figure 9.6-3   Denormalized bandpass filter schematic.

First, performing the prototype (normalized circuit) conversion:

$$c_1 = c_3 = \frac{c_{\text{lp}}}{\text{bw}} = \frac{1.000}{0.4082} = 2.4498$$

$$l_1 = l_3 = \frac{1}{c_1} = \frac{1}{2.4498} = 0.4082$$

$$l_2 = \frac{l_{\text{lp}}}{\text{bw}} = \frac{2.000}{0.4082} = 4.8996$$

$$c_2 = \frac{1}{l_2} = \frac{1}{4.8996} = 0.2041$$

where the lowercase letters designate the impedance and frequency normalized prototype filter. Next, denormalize the circuit (Fig. 9.6-3).

$$C_1 = C_3 = \frac{c_1}{Z_0 2\pi f_0} = \frac{2.4498}{(50)(2\pi)(979.8 \times 10^6)} = 7.9628 \text{ pF}$$

$$L_1 = L_3 = \frac{l_1 Z_0}{2\pi f_0} = \frac{(0.4082)(50)}{(2\pi)(979.8 \times 10^6)} = 3.315 \text{ nH}$$

$$L_2 = \frac{l_2 Z_0}{2\pi f_0} = \frac{(4.8996)(50)}{(2\pi)(979.8 \times 10^6)} = 39.7936 \text{ nH}$$

$$C_2 = \frac{c_2}{Z_0 2\pi f_0} = \frac{0.2041}{(50)(2\pi)(979.8 \times 10^6)} = 0.66306 \text{ pF}$$

The performance of this design can be seen in Figure 9.6-4, which shows the expanded insertion loss on the left y axis and the isolation on the right y axis. The 3-dB bandwidth extends from 800 to 1200 MHz, as was intended by design. Also note the steep isolation ''skirts'' of the performance plot. At 500 and 2000 MHz, for frequencies that are an octave removed from the center frequency of about 1000 MHz, the isolation is nearly 36 dB. This is appropriate for a $N = 6$ filter, the number of nontrivial elements in the circuit of Figure
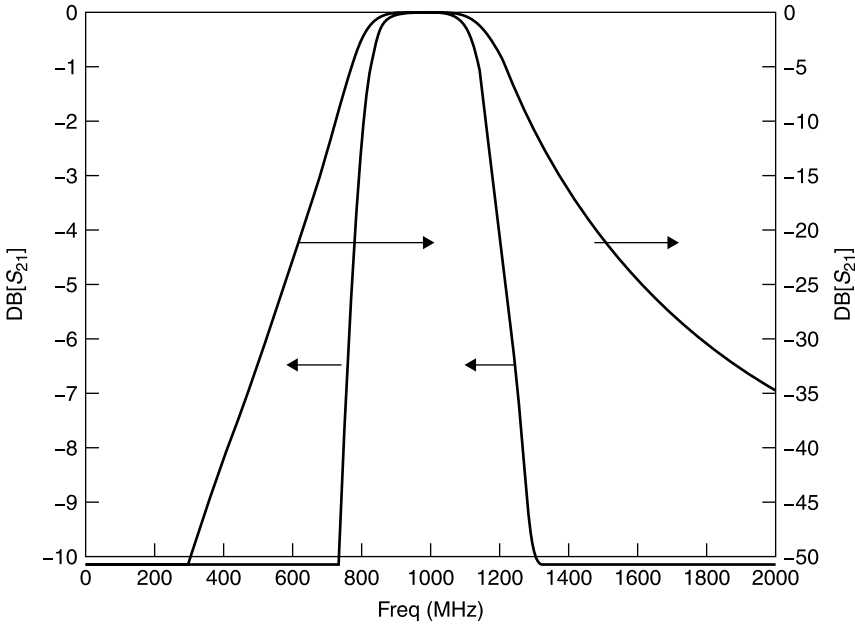
Figure 9.6-4   Calculated loss/isolation of the denormalized bandpass filter.

9.6-3. A consequence of scaling a bandpass filter from a low-pass prototype is that the order, N, of the filter is doubled.

9.7 BANDSTOP FILTERS

The design procedure for bandstop filters is similar to that of bandpass filters. However, for the bandstop filters the series resonant circuits are in shunt with the transmission line and the parallel resonant circuits are in series (Fig. 9.7-1).

The shunt resonator elements are found from [1, p. 151]

$$L_{\text{bs-shunt}} = \frac{1}{C_{\text{lp}}\text{bw}} \tag{9.7-1}$$

$$C_{\text{bs-shunt}} = \frac{1}{L_{\text{bs-shunt}}} \tag{9.7-2}$$

and the series resonator elements are found as

$$C_{\text{bs-series}} = \frac{1}{L_{\text{lp}}\text{bw}} \tag{9.7-3}$$

$$L_{\text{bs-series}} = \frac{1}{C_{\text{bs-series}}} \tag{9.7-4}$$
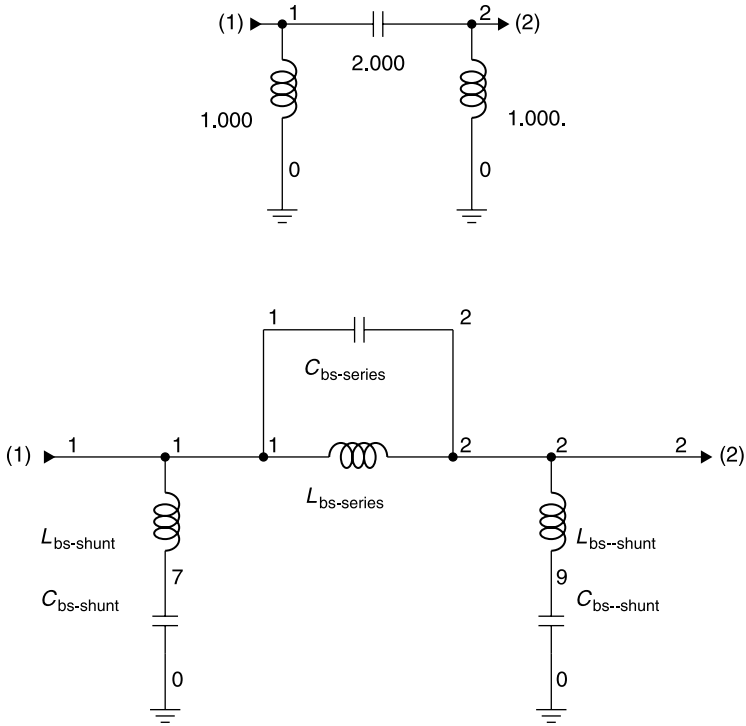
Figure 9.7-1   Developing a bandstop filter from the low-pass prototype.

Following our three-element filter examples, suppose we wish to design a bandstop filter with stopband from 800 to 1200 MHz. The steps are as follows. Using (9.7-1) to (9.7-4),

$$l_1 = l_3 = \frac{1}{c_{lp}bw} = \frac{1}{(1.000)(0.4082)} = 2.4498$$

$$c_1 = c_3 = \frac{1}{l_1} = \frac{1}{2.4498} = 0.4082$$

$$c_2 = \frac{1}{l_{lp}bw} = \frac{1}{(2.000)(0.4082)} = 1.2249$$

$$l_2 = \frac{1}{c_2} = \frac{1}{1.2249} = 0.8164$$

After computing the values of the normalized bandstop prototype from (9.7-1) to (9.7-4), the elements are denormalized, in the manner used for the bandpass filter. Denormalizing,
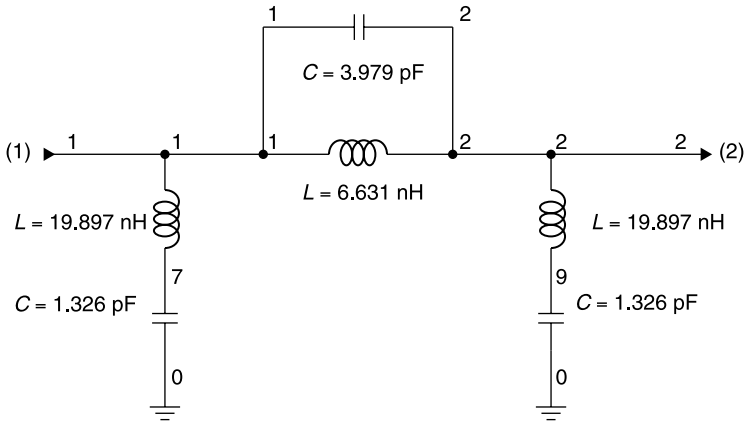
Figure 9.7-2   The 800- to 1200-MHz bandstop filter.

$$L_1 = L_3 = \frac{l_1 Z_0}{2\pi f_0} = \frac{(2.4498)(50)}{(2\pi)(979.8 \times 10^6)} = 19.8968 \text{ nH}$$

$$C_1 = C_3 = \frac{c_1}{Z_0 2\pi f_0} = \frac{0.4082}{(50)(2\pi)(979.8 \times 10^6)} = 1.3261 \text{ pF}$$

$$L_2 = \frac{l_2 Z_0}{2\pi f_0} = \frac{(0.8164)(50)}{(2\pi)(979.8 \times 10^6)} = 6.6306 \text{ nH}$$

$$C_2 = \frac{c_2}{Z_0 2\pi f_0} = \frac{1.2249}{(50)(2\pi)(979.8 \times 10^6)} = 3.9794 \text{ pF}$$

The resulting circuit is shown in Figure 9.7-2 and its performance in Figure 9.7-3. The 3-dB loss values border the stopband at 800 and 1200 MHz as planned. The isolation for this lossless filter is infinite at the band center near 1000 MHz.

9.8 CHEBYSHEV FILTERS

Chebyshev (his name is anglicized and sometimes written as Tchebyscheff, also anglicized) was a Russian mathematician credited with the Chebyshev polynomials. The polynomials have the interesting property that they have equal ripples (values at zero slope) the number of which is equal to half the order N of the polynomial. Since the order of a filter is equal to its number of nontrivial elements, an N element filter can have N changes of slope, or $N/2$ ripples, in its passband. Matching the filter's response to a Chebyshev polynomial provides an *equal ripple* response in the passband.

The reasoning behind matching a filter's transmission characteristic to a Chebyshev polynomial is that, if one can accept a certain amount of insertion
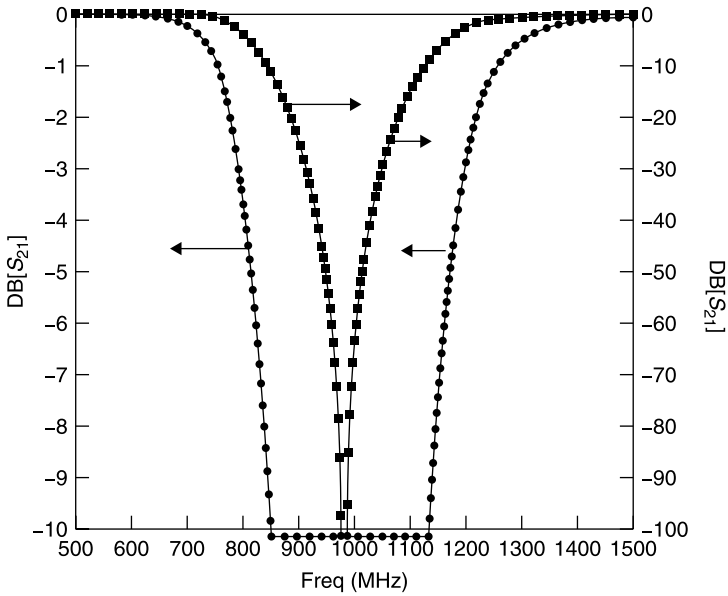
Figure 9.7-3   Calculated loss/isolation of the 800- to 1200-MHz bandstop filter.

loss at the band edges of a filter, perhaps that same loss allowance could be tolerated anywhere within the passband. If this is acceptable, then, for a given width passband, the Chebyshev filter response yields a lower peak value of passband loss or a wider bandwidth for the same peak passband loss.

However, to design Chebyshev filters, we require a *separate table of proto-type values for each value of maximum ripple*. Another consequence of the loss ripple is that *even-order Chebyshev filters cannot have equal source and load resistances*. This is because the even-order filters must have a mismatch at zero frequency, and this can only occur if source and load resistances are correspondingly different. Designing filters that are only odd-order circumvents this problem; source and load can have equal impedances.

For the filter tables, it is assumed that the source impedance, $g(0)$, is unity, and the last value in the table for a given order filter is the magnitude of the required load resistance. We confine our examples to those having unity load resistance.

Given these distinctions, the design of Chebyshev filters proceeds exactly as presented for the Butterworth designs. Tables 9.8-1 to 9.8-5 give the normalized low-pass prototype values. Transformations to high-pass, bandpass, and band-stop filters proceed as previously described. There are other sources for Chebyshev filter tables; however, these tables by Rhea are especially convenient to use because the g values yield filters having a cutoff frequency at the maximum ripple value rather than at the 3-dB loss frequency.

TABLE 9.8-1 Chebyshev 0.01-dB Equal Ripple (RL = 26.4 dB)

N	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}
2	0.4489	0.4078	0.9085								
3	0.6292	0.9703	0.6292	1							
4	0.7129	1.2004	1.3213	0.6476	0.9085						
5	0.7563	1.3049	1.5773	1.3049	0.7563	1					
6	0.7814	1.3600	1.6897	1.5350	1.4970	0.7098	0.9085				
7	0.7970	1.3924	1.7481	1.6331	1.7481	1.3924	0.7970	1			
8	0.8073	1.4131	1.7824	1.6833	1.8529	1.6193	1.5555	0.7334	0.9085		
9	0.8145	1.4271	1.8044	1.7125	1.9058	1.7125	1.8044	1.4271	0.8145	1	
10	0.8197	1.4370	1.8193	1.7311	1.9362	1.7590	1.9055	1.6528	1.5817	0.7446	1

Source: After Rhea [1, pp. 48–49]. Used with permission.

TABLE 9.8-2 Chebyshev 0.0432-dB Equal Ripple (RL = 20 dB)

N	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}
2	0.6648	0.5445	0.8190								
3	0.8516	1.1032	0.8516	1							
4	0.9314	1.2920	1.5775	0.7628	0.8190						
5	0.9714	1.3721	1.8014	1.3721	0.9714	1					
6	0.9940	1.4131	1.8933	1.5506	1.7253	0.8141	0.8190				
7	1.0080	1.4368	1.9398	1.6220	1.9398	1.4368	1.0080	1			
8	1.0171	1.4518	1.9667	1.6574	2.0237	1.6107	1.7726	0.8330	0.8190		
9	1.0235	1.4619	1.9837	1.6778	2.0649	1.6778	1.9837	1.4619	1.0235	1	
10	1.0281	1.4690	1.9952	1.6906	2.0882	1.7102	2.0642	1.6341	1.7936	0.8420	0.8190

Source: After Rhea [1, pp. 48–49]. Used with permission.

TABLE 9.8-3 Chebyshev 0.10-dB Equal Ripple (RL = 16.4 dB)

N	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}
2	0.8431	0.6220	0.7378								
3	1.0316	1.1474	1.0316	1							
4	1.1088	1.3062	1.7704	0.8181	0.7378						
5	1.1468	1.3712	1.9750	1.3712	1.1468	1					
6	1.1681	1.4040	2.0562	1.5171	1.9029	0.8618	0.7378				
7	1.1812	1.4228	2.0967	1.5734	2.0967	1.4228	1.1812	1			
8	1.1898	1.4346	2.1199	1.6010	2.1700	1.5641	1.9445	0.8778	0.7378		
9	1.1957	1.4426	2.1346	1.6187	2.2054	1.6167	2.1346	1.4426	1.1957	1	
10	1.2000	1.4482	2.1445	1.6266	2.2254	1.6419	2.2046	1.5822	1.9629	0.8853	0.7378

Source: After Rhea [1, pp. 48–49]. Used with permission.

TABLE 9.8-4 Chebyshev 0.20-dB Equal Ripple (RL = 13.5 dB)

N	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}
2	1.0379	0.6746	0.6499								
3	1.2276	1.1525	1.2276	1							
4	1.3029	1.2844	1.9762	0.8468	0.6499						
5	1.3395	1.3370	2.1661	1.3370	1.3395	1					
6	1.3598	1.3632	2.2395	1.4556	2.0974	0.8838	0.6499				
7	1.3723	1.3782	2.2757	1.5002	2.2757	1.3782	1.3723	1			
8	1.3804	1.3876	2.2964	1.5218	2.3414	1.4925	2.1349	0.8972	0.6499		
9	1.3861	1.3939	2.3094	1.5340	2.3728	1.5340	2.3094	1.3939	1.3861	1	
10	1.3901	1.3983	2.3181	1.5417	2.3905	1.5537	2.3720	1.5066	2.1514	0.9035	0.6499

Source: After Rhea [1, pp. 48–49]. Used with permission.

TABLE 9.8-5 Chebyshev 0.50-dB Equal Ripple (RL = 9.6 dB)

N	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}
2	1.4029	0.7071	0.5040								
3	1.5963	1.0967	1.5963	1							
4	1.6704	1.1926	2.3662	0.8419	0.5040						
5	1.7058	1.2296	2.5409	1.2296	1.7058	1					
6	1.7254	1.2478	2.6064	1.3136	2.4759	0.8696	0.5040				
7	1.7373	1.2582	2.6383	1.3443	2.6383	1.2582	1.7373	1			
8	1.7451	1.2647	2.6565	1.3590	2.6965	1.3389	2.5093	0.8795	0.5040		
9	1.7505	1.2690	2.6678	1.3673	2.7240	1.3673	2.6678	1.2690	1.7505	1	
10	1.7543	1.2722	2.6755	1.3725	2.7393	1.3806	2.7232	1.3484	2.5239	0.8842	0.5040

Source: After Rhea [1, pp. 48–49]. Used with permission.

(a)

(b)
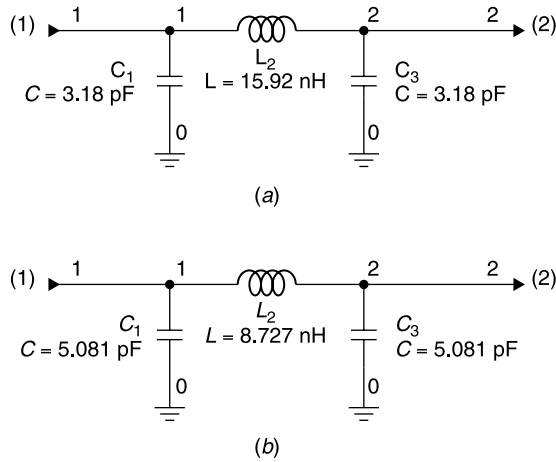
Figure 9.8-1 Comparison of element values for Butterworth and Chebyshev filters. (a) Three-section 50-Ω Butterworth low-pass filter with $f_C = 1$ GHz. (b) Three-section, 50-Ω, 0.5-dB ripple, Chebyshev low-pass filter with $f_C = 1$ GHz.

As an example, a 50-Ω, three-section, 0.2-dB ripple Chebyshev low-pass filter is designed using the procedure described for the low-pass Butterworth filter. The circuit is shown in Figure 9.8-1 and its performance compared to the three-element Butterworth low-pass filter in Figure 9.8-2.

Notice that the Chebyshev filter has a maximum loss up to 1 GHz of only 0.5 dB, while the Butterworth filter has 3.0-dB loss at 1 GHz. Also, the Chebyshev filter has a steeper slope outside the passband, yielding slightly greater isolation at frequencies above the passband.

9.9 PHASE AND GROUP DELAY

Previously, we described the transfer function H of the network. In the filter performance described up to now, we have been concerned only with the amplitude of the transfer function versus frequency. Usually, this is the predominant characteristic that comes to mind when considering filters.

However, in communication systems the variation of phase shift with frequency through a network becomes of importance when the signal bandwidth is large, as when rapid rise time pulses or video information is transmitted. Then the phase variation with frequency through the network also must be controlled.

The *phase delay* φ (or *phase shift*) through a network is just the argument (angle) of the complex transfer function H. Any network having reactive elements produces a delay in the signal propagation with time [1, pp. 34–36].

If a network passes all frequencies with equal amplitude attenuation and a transmission phase φ that increases linearly with frequency, the signal will be
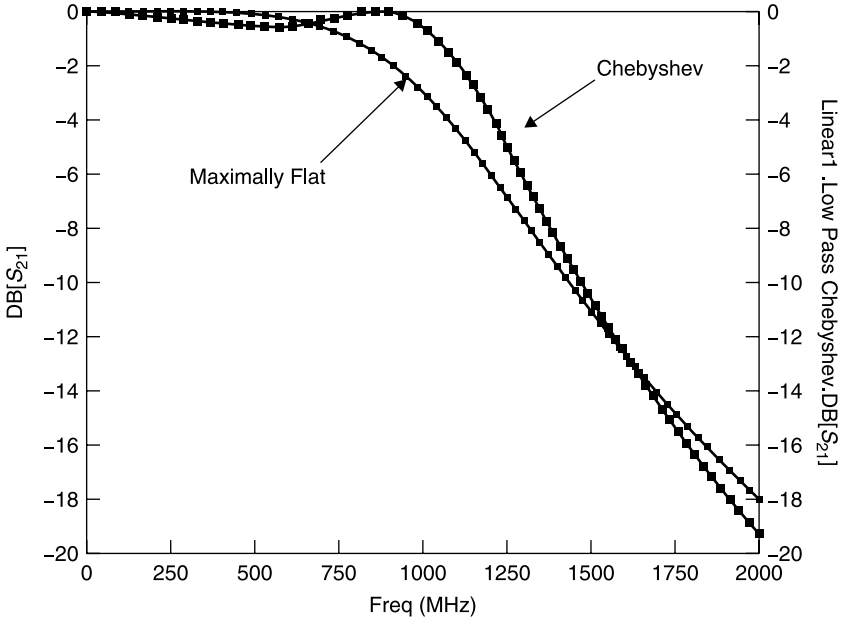
Figure 9.8-2   Calculated loss comparison of Butterworth and 0.5-dB-ripple Chebyshev low-pass filters.

delayed and level shifted but not distorted. The *phase delay* of a network is

$$t_p = \frac{\varphi}{\omega} \tag{9.9-1}$$

and the *group delay* is

$$t_d = -\frac{\partial \varphi}{\partial \omega} \tag{9.9-2}$$

The group delay (also called *envelope delay*) is related to the time required for the envelope of a signal to transverse the network. These characteristics were introduced in Section 4.12 through phase and group velocities.

Unfortunately, good frequency selectivity and constant group delay are mutually exclusive in the passive ladder networks we have been considering. Signal distortion occurs when different frequency components of the composite signal are unequally delayed. This difference is called *differential delay*. Differential group delay is the absolute difference in the group delay at two specified frequencies. It will be seen that the more frequency selective the filter network, the greater the differential delay within the passband. A class of filters that minimizes this distortion is the *Bessel filter*. Table 9.9-1 lists the Bessel *g* values for this response.

TABLE 9.9-1 Bessel Prototype Filter g Values[a]

N	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}
2	0.5755	2.1478	1								
3	0.3374	0.9705	2.2034	1							
4	0.2334	0.6725	1.0815	2.2404	1						
5	0.1743	0.5072	0.8040	1.1110	2.2582	1					
6	0.1365	0.4002	0.6392	0.8538	1.1126	2.2645	1				
7	0.1106	0.3259	0.5249	0.7020	0.8690	1.1052	2.2659	1			
8	0.0919	0.2719	0.4409	0.5936	0.7303	0.8695	1.0956	2.2656	1		
9	0.0780	0.2313	0.3770	0.5108	0.6306	0.7407	0.8639	1.0863	2.2649	1	
10	0.0672	0.1998	0.3270	0.4454	0.5528	0.6493	0.7420	0.8561	1.0781	2.2641	1

Source: After Rhea [1, p. 49]. Used with permission.
[a]Cutoff frequency loss = 3 dB and source resistance = 1.0.

TABLE 9.9-2 Bessel Equal-Ripple Phase Error 0.05° Filter g Values[a]

N	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}
2	0.6480	2.1085	1								
3	0.4328	0.0427	2.2542	1							
4	0.3363	0.7963	1.1428	2.2459	1						
5	0.2751	0.6541	0.8892	1.1034	2.2873	1					
6	0.2374	0.5662	0.7578	0.8760	1.1163	2.2448	1				
7	0.2085	0.4999	0.6653	0.7521	0.8749	1.0671	2.2845	1			
8	0.1891	0.4543	0.6031	0.6750	0.7590	0.8427	1.0901	2.2415	1		
9	0.1718	0.4146	0.5498	0.6132	0.6774	0.7252	0.8450	1.0447	2.2834	1	
10	0.1601	0.3867	0.5125	0.5702	0.6243	0.6557	0.7319	0.8178	1.0767	2.2387	1

Source: After Rhea [1, p. 50]. Used with permission.
[a]Cutoff frequency loss = 3 dB, source resistance = 1.0.

TABLE 9.9-3 Bessel Equal-Ripple Phase Error 0.5° Filter g Values[a]

N	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}
2	0.8245	1.9800	1								
3	0.5534	1.0218	2.4250	1							
4	0.4526	0.7967	1.2669	2.0504	1						
5	0.3658	0.6768	0.9513	1.0113	2.4446	1					
6	0.3313	0.5984	0.8390	0.7964	1.2734	2.0111	1				
7	0.2876	0.5332	0.7142	0.6988	0.9219	0.9600	2.4404	1			
8	0.2718	0.4999	0.6800	0.6312	0.8498	0.7447	1.3174	1.9626	1		
9	0.2347	0.4493	0.5914	0.5747	0.7027	0.6552	0.8944	0.9255	2.4332	1	
10	0.2359	0.4369	0.5887	0.5428	0.7034	0.5827	0.8720	0.6869	1.4317	1.8431	1

Source: Rhea [1, p. 50]. Used with permission.

[a] $A_a = 3$ dB, $G(0) = 1.0$.
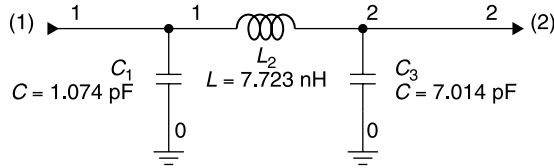
Figure 9.9-1   Bessel 50-Ω, low-pass filter with 3-dB cutoff at 1 GHz.

In addition to the pure Bessel prototype, additional tables provide an "equal phase ripple" in the transmission phase similar to the Chebyshev designs, which provide equal loss ripple in the passband attenuation. Allowing increasing phase ripple in the passband improves the rejection band performance of the Bessel filter. Tables 9.9-2 and 9.9-3 for 0.05° and 0.5° equal-ripple phase error give the g values for these filters.

Constructing our three-element filter using the Bessel response yields the circuit shown in Figure 9.9-1 and the performance shown in Figure 9.9-2. Note that the filter is not symmetric, as were the Butterworth and Chebyshev filters.

Viewing only amplitude response, one might consider the Bessel filter to be so inferior, even to the Butterworth filter, as to warrant no further consider-
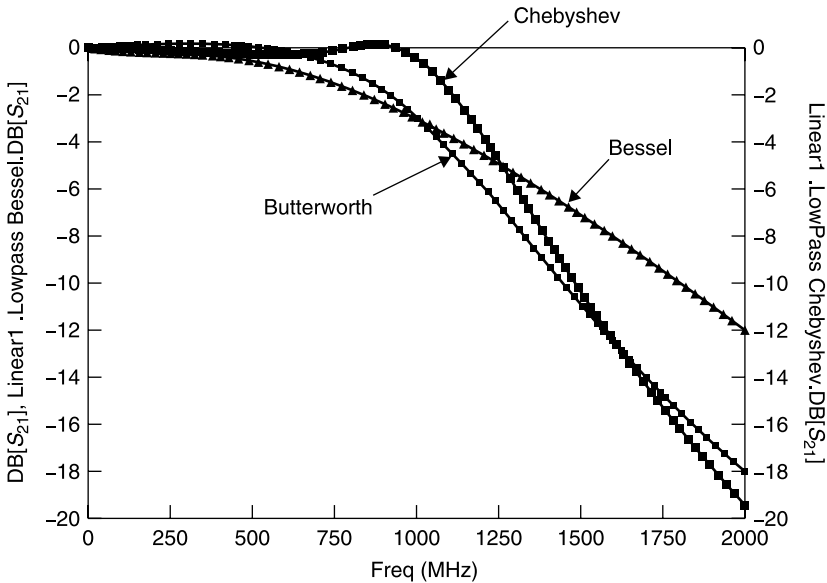
Figure 9.9-2   Bessel low-pass filter compared in amplitude response to Butterworth and Chebyshev filters of the same order.
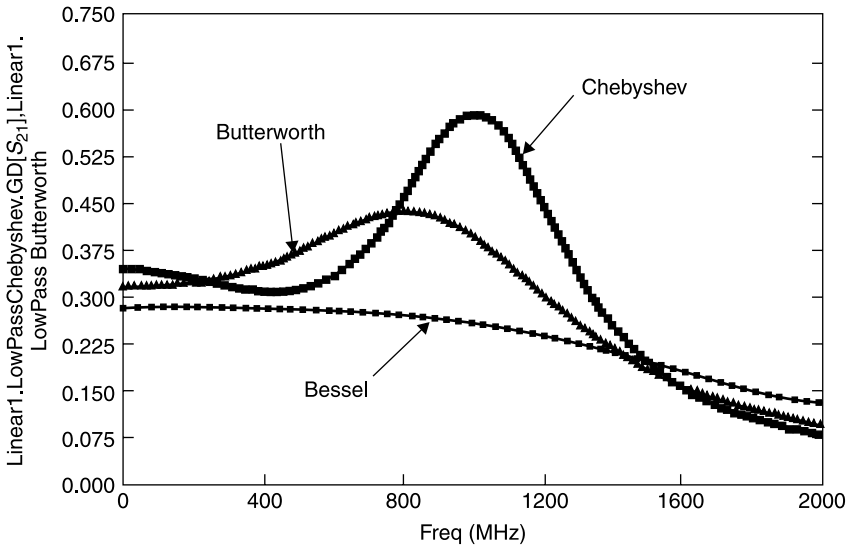
Figure 9.9-3   Group delay (in nanoseconds) comparison of Bessel, Chebyshev, and Butterworth filters, all having three sections and 1 GHz cutoff frequency.

ation. However, considering the group delay performance reveals a considerable advantage of the Bessel design.

The Bessel low-pass filter has nearly constant group delay up to 1000 MHz, its 3-dB cutoff frequency. The Chebyshev has a nearly 2 to 1 variation in group delay to its 0.5-dB cutoff, and the Butterworth has an intermediate value of group delay over the 0 to 1000 MHz band (Fig. 9.9-3). These data confirm that *the greater the selectivity, the greater the delay variation and the greater the potential signal distortion.*

9.10   FILTER *Q*

To illustrate filter types, we have reviewed only lossless representations. Real filters, of course, must be made with components having finite *Q*. The effect of finite *Q* components on filter insertion loss in the passband is greatest in those having the most frequency selectivity. The effect of finite *Q* can be seen in the Chebyshev low-pass filter shown in Figure 9.10-1. Both the L and C components are assumed to have the *Q* value listed.

In the case of the three-section Chebyshev low-pass filter, the lossless component assumption (actually *Q* = 1,000,000 is the Genesys default if *Q* is not specified) yields the expected 0.5-dB insertion loss over the passband of 0 to 1000 MHz. However, setting all circuit elements to have *Q* = 50 increases the
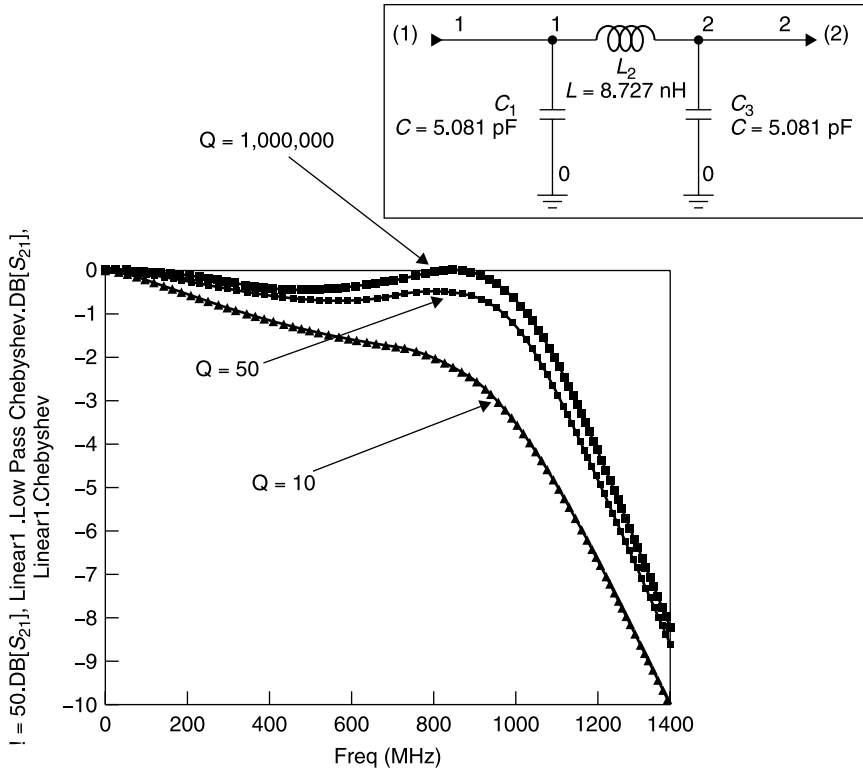
Figure 9.10-1   Effect of finite Q circuit elements on the performance of the Chebyshev 0.5-dB ripple, 1 GHz cutoff frequency filter.

loss at the cutoff frequency of 1000 MHz to about 1 dB. Further reducing element Q to 10 results in about 3.5 dB of insertion loss at 1000 MHz, as seen in Figure 9.10-1.

However, this example should not be considered a reference scale of the passband loss increase when circuit Q is 50 or 10. The change in passband loss depends upon the selectivity of the filter as well.

Figure 9.10-2 shows a five-section Chebyshev filter having 0.1 dB passband ripple for the bandwidth 175 to 225 MHz. Its insertion loss response is simulated using $Q = 1,000,000$ for all element values. This is shown in Figure 9.10-3.

Next, the filter shown in Figure 9.10-2 has its element Q changed to 50. The resulting performance is shown in Figure 9.10-4. A change of scale was made because the center frequency loss increased by 5 dB. Notice that not only has the center frequency loss increased by 5 dB, but the loss increase near the band
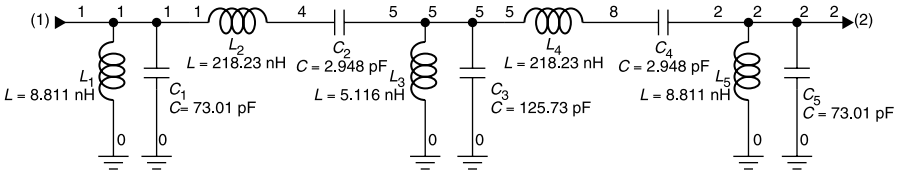
Figure 9.10-2   Five-section 0.1-dB-ripple Chebyshev bandpass filter for 175 to 225 MHz.

edges is much more, about 8 dB, resulting in a severe change in the passband shape. This filter design is not intended as representing a filter whose execution is practical, but rather to show that the effect of *Q* on passband loss depends on the selectivity of the filter, itself.

This design example has very sharp passband skirts, theoretically. But realizing such a design would require very low loss elements. Also note that the elements used in this simulation did not have parasitic inductance and capacitances, as discussed in Section 2.11. Inclusion of parasitics would make this design even less practical to realize. The point of this example is that it is important to check proposed filter designs using realistic models for their elements.
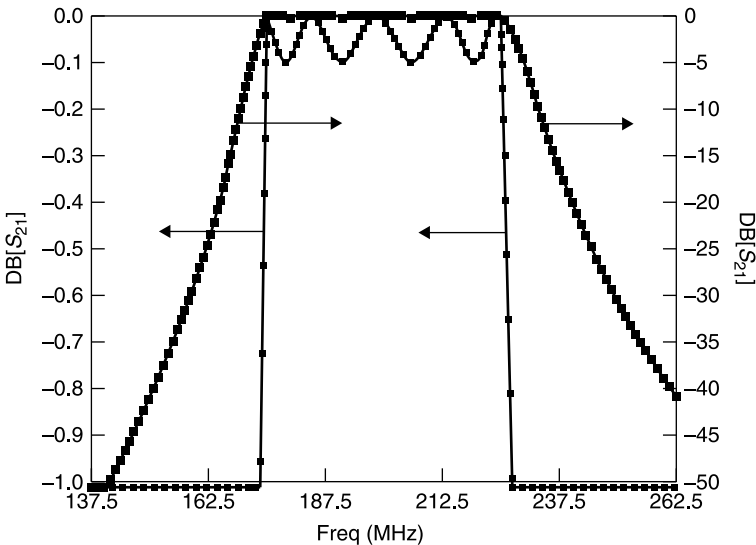
Figure 9.10-3   Calculated loss/isolation of five-section Chebyshev filter when *L* and *C* elements have $Q_U = 1,000,000$.
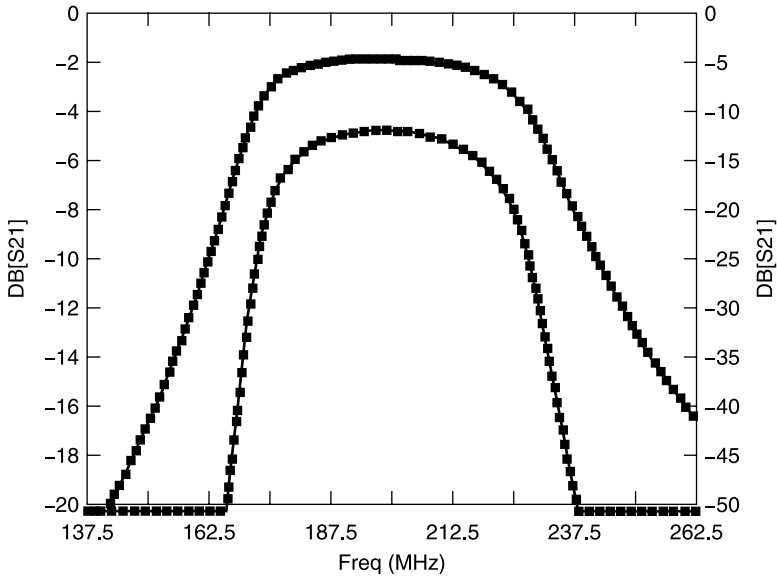
Figure 9.10-4 Calculated loss/isolation of the Chebyshev filter when elements have $Q = 50$.

9.11 DIPLEXER FILTERS

A *frequency diplexer* is a network that connects two networks that operate at different frequencies to a common terminal. For example, two radios operating at different frequency bands could be connected to a common antenna using a frequency diplexer. One commonly used realization of a frequency diplexer consists of two complementary filters connected to a tee junction. One filter is a high pass and the other a low pass. For the diplexer function, both filters have the same 3-dB cutoff frequency. Frequencies above the cutoff are directed along one side of the tee junction; frequencies below the cutoff are directed along the other side of the tee junction.

The two filters are designed using the *singly terminated* method by which the tee end of the filter is designed for an infinite load impedance for the high-pass arm and a zero load impedance for the low-pass arm. Although it is not intuitively apparent, when two such filters are connected together it is found that *a match is seen at all frequencies at the input to the tee*. The input impedance of the two filters, R_{in}, is made equal to the desired system impedance, in this case 50 Ω.

The tables given in Sections 9.3, 9.8 and 9.9 for Butterworth, Chebyshev, and Bessel filter designs are for doubly terminated filters, i.e., those having
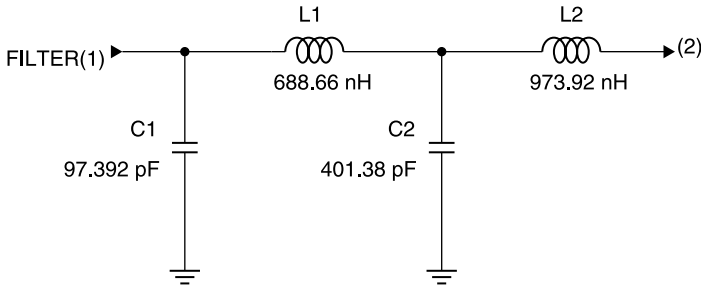
Figure 9.11-1 Singly terminated low-pass Butterworth filter with 3-dB cutoff frequency of 12.5 MHz.

the same characteristic impedance at both ports, therefore cannot be used for this design. Singly terminated designs are available with software, such as the FILTER program within the Genesys software suite.

Suppose we wish to build a diplexer having a 3-dB cutoff (transition frequency) of 12.5 MHz. This might be used to connect two radios to a common antenna, with one radio operating below 10 MHz and the other operating above 16 MHz. The two singly terminated designs realized using the FILTER software are shown in Figures 9.11-1 and 9.11-2. Notice that input port 1 is well matched. The return loss is between 50 and 60 dB and the signal input at port 1 is routed smoothly between ports 2 and 3 according to the frequency. Notice also that this is a means of providing filtering one port of which port 1 is matched at all frequencies.

When the two filters are connected together at their respective zero and infinite impedance ports the resulting diplexer shown in Fig. 9.11-3 is obtained.
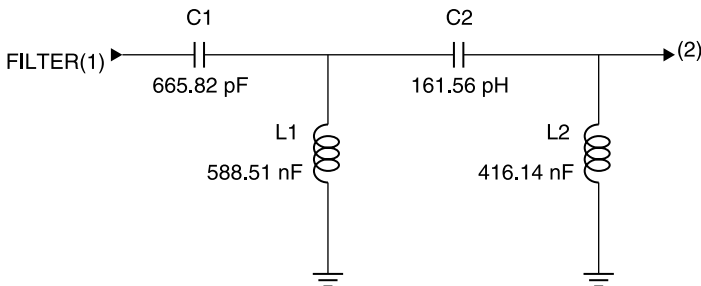
Figure 9.11-2 Singly terminated high-pass Butterworth filter with 3-dB cutoff frequency of 12.5 MHz.
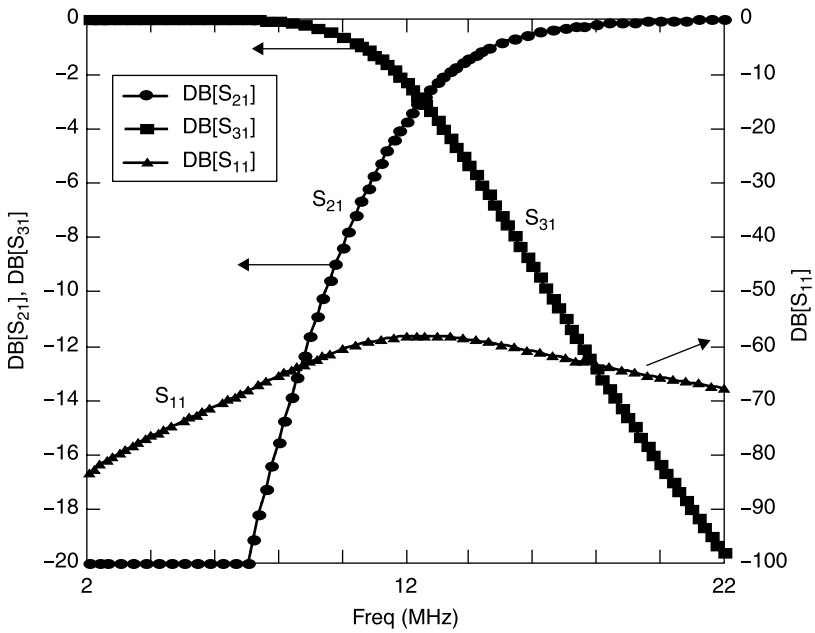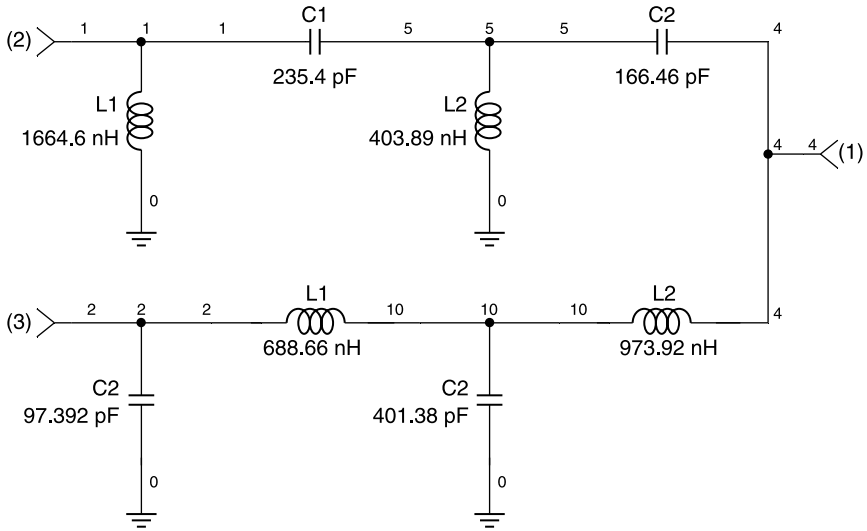
Figure 9.11-3 Frequency diplexer having matched transmission at all frequencies from its common port.
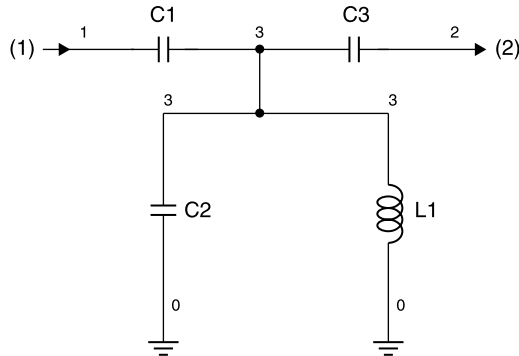
Figure 9.12-1 Top-coupled filter.

9.12 TOP-COUPLED BANDPASS FILTERS

In Section 3.4 we examined top-C, lightly coupled resonators. When used as a filter, this circuit is called a *top-coupled filter*. This simple circuit shown in Figure 9.12-1 often can provide sufficient filtering for many applications. Since
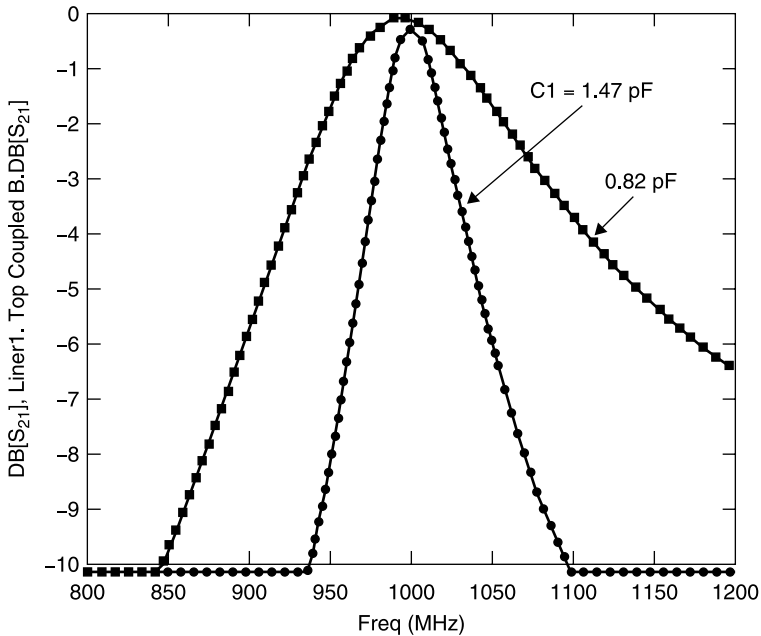
Figure 9.12-2 Varying the coupling of the top-coupled resonator changes the loaded Q.
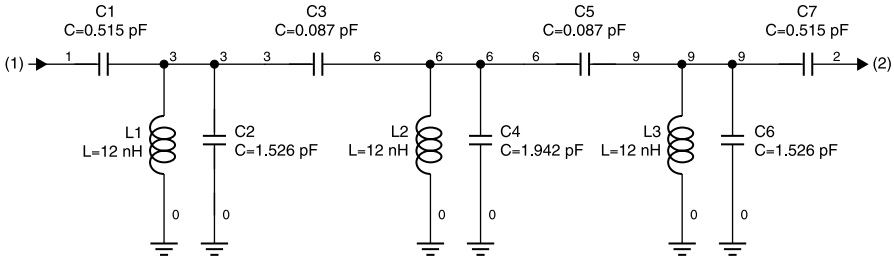
Figure 9.12-3 Three-section, Chebyshev, 0.25-dB ripple, 975 to 1025 MHz, top-coupled filter.

there are so few degrees of engineering freedom for the circuit, the design can be accomplished using the tune feature or the optimizer of a network simulator. The $L_1 C_2$ product determines the center frequency $(f_0 = 1/\sqrt{L_1 C_2})$ and $C_1 = C_3$ (for symmetry) controls the loaded Q. However, there is some interaction between resonance and coupling because C_1 affects the total capacitance of the resonator somewhat, therefore retuning of L_1 and/or C_2 is necessary to return to f_0 as C_1 and C_2 are tuned (synchronously) to adjust the coupling. The higher the reactance of C_1, the higher the loaded Q.

As an example, the resonance was set equal to 1 GHz by selecting values of $L_1 = 3.18$ nH and C_2 to resonate at 1 GHz (about 6 pF, as needed according to
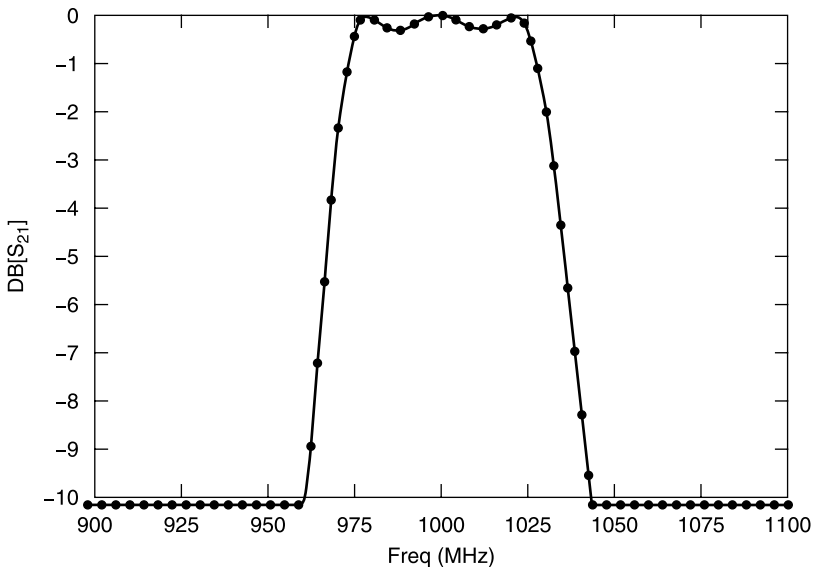
Figure 9.12-4 Calculated loss/isolation of the filter in Figure 9.12-3.

the coupling). If placed directly in parallel with a 50-Ω source and 50-Ω load, the resonator would have a loaded Q of nearly unity and a very broad 3-dB bandwidth. Using a coupling capacitance of $C_1 = C_3 = 0.82$ pF, the 3-dB bandwidth is 50 MHz with f_0 loss of 0.3 dB. This is equivalent to $Q_L = 20$ as shown in Figure 9.12-2. To accomplish the same Q_L with direct coupling would require that the inductance be reduced by a factor of 20, or $L_2 = 0.2$ nH, a very impractically small value. Increasing the coupling capacitors to 1.47 pF gives a 3-dB bandwidth of 150 MHz with f_0 loss of 0.1 dB.

The coupling that results in 0-dB insertion loss at f_0 is called critical coupling. For this simple circuit, critical coupling and the desired 3-dB bandwidth may not be accomplished simultaneously. More selectivity can be obtained using multiple resonators. A three-section, Chebyshev filter with 0.25 dB ripple over the 975 to 1025 MHz bandwidth is shown in Figure 9.12-3 and its loss performance shown in Figure 9.12-4. Notice that the resonators have 12-nH inductors, more practical to realize than those which would be necessary for a direct coupled resonator.

9.13 ELLIPTIC FILTERS

Just as it is desirable to place a ripple response in the passband, there can be advantages to having equal isolation ripple in the stopband. This can make possible a more rapid increase of isolation with frequency at the passband edge. Filters having a ripple in the stopband can be realized using *elliptic filter* designs.

While the sharp increase in isolation just outside the passband can be very advantageous for some applications, *it comes at the price of reduced isolation at frequencies considerably higher than the filter's cutoff frequency*.

We will compare the performance of a Chebyshev 0.5-dB- (passband) ripple, five-section, 1-GHz low-pass filter with a comparably designed elliptic filter designed for the same cutoff. The straight Chebyshev filter is shown in Figure 9.13-1.

The elliptic filter, designed using the Genesys SFILTER software, to similar specifications, is shown in Figure 9.13-2. Notice that the stopband attenuation drops more rapidly just above the 1 GHz cutoff frequency, due to the parallel resonant series elements, but that the maximum isolation is only 48 dB, even at frequencies far from cutoff. This limit on the isolation is the performance trade-off for the more rapid isolation increase near the cutoff frequency.
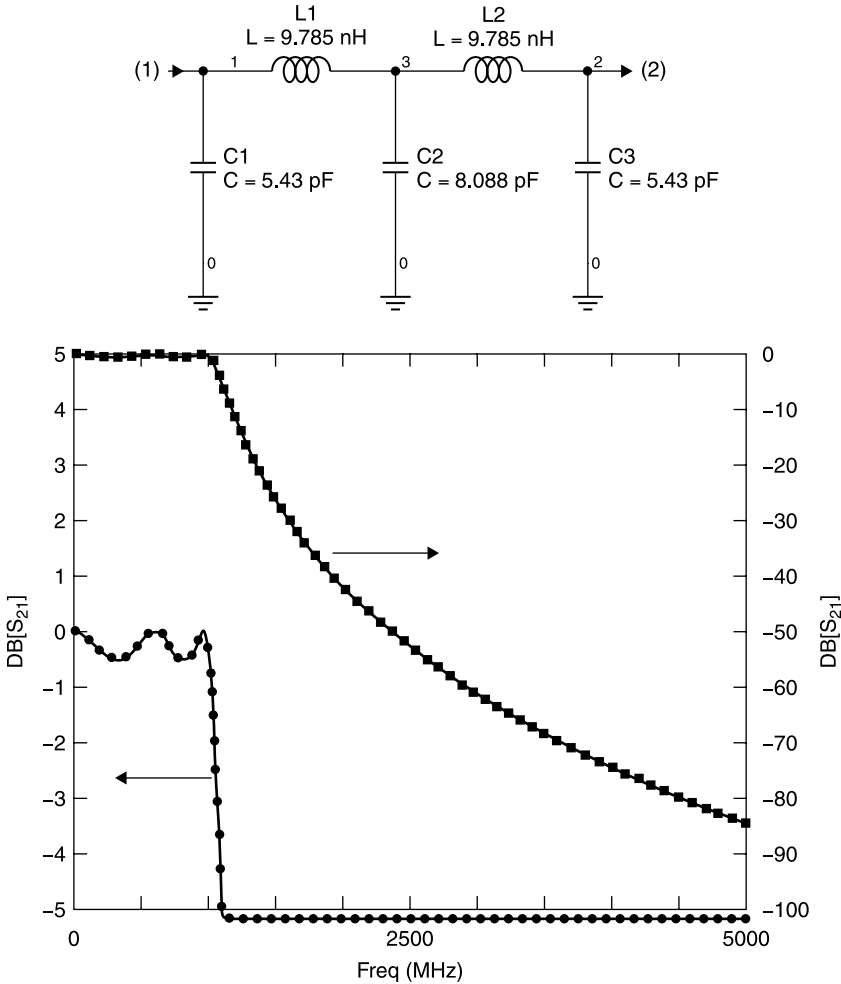
Figure 9.13-1   Five-section, low-pass, 0.5-dB-ripple Chebyshev filter.

9.14   DISTRIBUTED FILTERS

Most lumped-element filter designs are limited to realizations up to only a few hundred megahertz before element values become impractically small. To design filters for use above 500 MHz, it is more practical to design *distributed filters*, that is, *filters using distributed structures, such as lengths of transmission lines, as resonant elements*. The top-coupled filter in Figure 9.12-1 is a resonator capacitively coupled near its high-voltage node. A similar resonant structure can be obtained using a half-wave transmission line capacitively coupled to
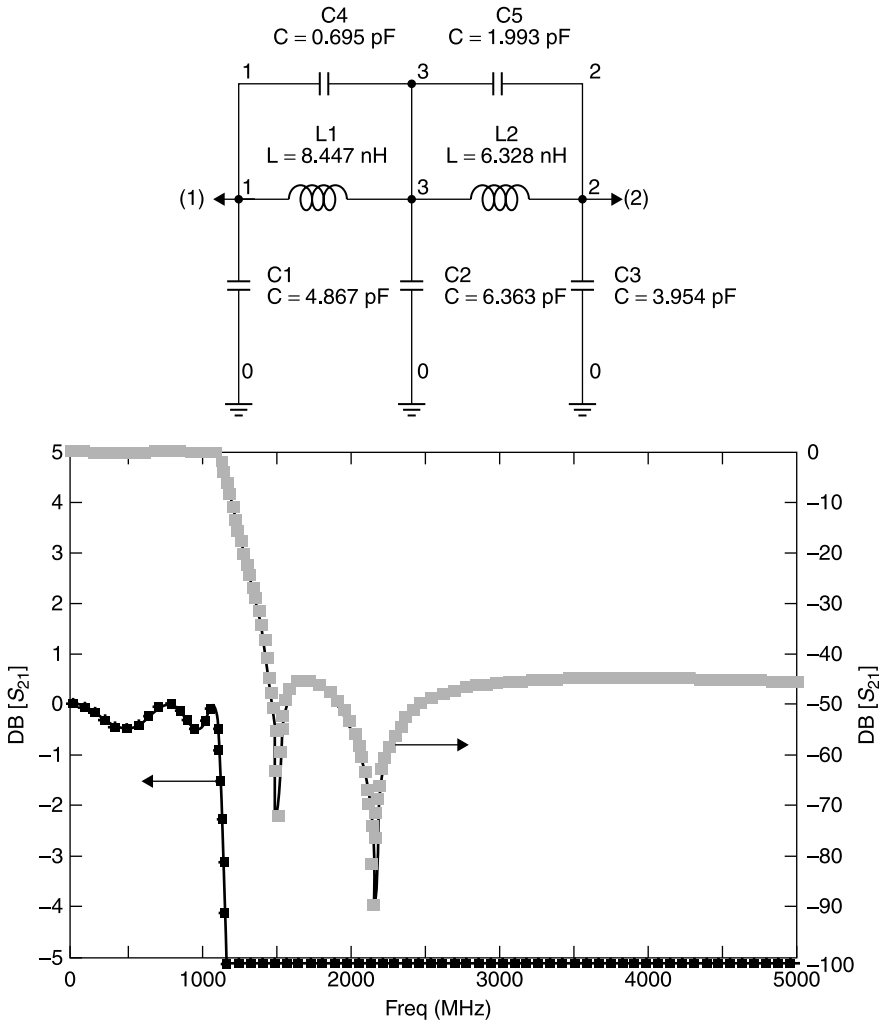
Figure 9.13-2   Five-section, low-pass elliptic filter.

generator and load, as shown in Figure 9.14-1 and its loss performance is shown in Figure 9.14-2.

The half-wave resonator stores energy in the form of standing waves with voltage maxima at its two ends. Thus, the capacitive coupling shown is electrically similar to that of the top-coupled filter previously described. The 3-dB bandwidth of this example filter is only 8 MHz with $f_0 = 1000$ MHz, yielding an effective $Q_L = 125$. Yet the resonator consists of only a length of 50-Ω transmission line, an easily realized structure. The coupling capacitors might be realized as gaps in the transmission line and designed either using an EM sim-

Transmission Line
Z = 50 ohm
L = 171.3° [Theta]
F = 1000 MHz

(1) 1   4   3   2 (2)

C1
C = 0.242 pF [C]

C2
C = 0.242 pF [C]
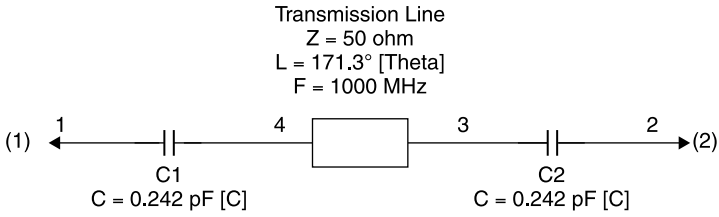
Figure 9.14-1   End-coupled, half-wave resonator filter.
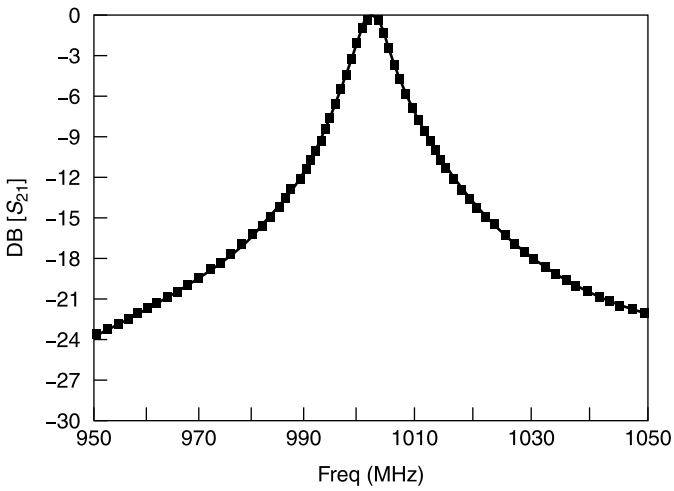
Figure 9.14-2   Calculated loss/isolation of the half-wave filter shown in Figure 9.14-1.
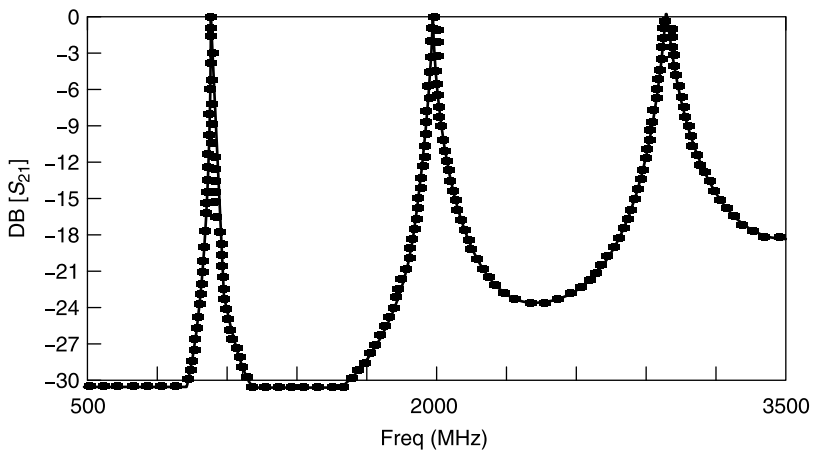
Figure 9.14-3   Recurring passbands of the half-wave resonator filter.

(1) ◄─── 1 ──┤├── 8 ──┤ TL1 ├── 3 ──┤├── 4 ──┤ TL2 ├── 5 ──┤├── 6 ──┤ TL3 ├── 7 ──┤├── 2 ──► (2)
C1   TL1   C2   TL2   C3   TL3   C4

C = 1.11 pF [C1] Z = 50 ohm [ZI] C = 0.271 pF [C2] Z = 50 ohm [ZI] C = 0.271 pF [C3] Z = 50 ohm [ZI] C = 1.11 pF [C4]
L = 157.777° [L1]   L = 170.365° [L2]   L = 157.777° [L3]
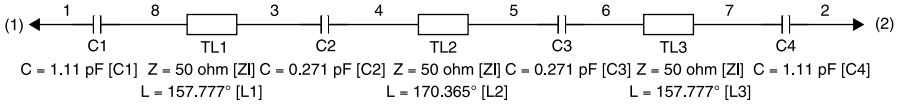
Figure 9.14-4 Three-section, half-wave resonator filter with $f_0 = 1$ GHz designed using the optimizer of the network simulator.

ulator or empirically, but once designed, could be reproduced by means of the printing process used to form the transmission line circuitry.

The circuit was modeled using a network simulator, and the resonator's length was adjusted in conjunction with the adjustment of the coupling capacitors, C_1 and C_2, to return the resonance to 1 GHz. As with the lumped-element, top-coupled filter, the coupling capacitors lower the resonant frequency; so readjustment of the line length was needed to return f_0 to 1 GHz. Notice that the isolation is greater than 20 dB at 960 and 1040 MHz. On the other hand, transmission line filters of which this is an example have recurring passbands, in this case roughly at integer multiples of f_0, as shown in Figure 9.14-3.

As with lumped filters, multiple half-wave resonators can be used. Figure 9.14-4 shows a three-resonator circuit designed using the optimizer of the network simulator. Initial values of 180° at 1 GHz and 50 Ω were selected for the
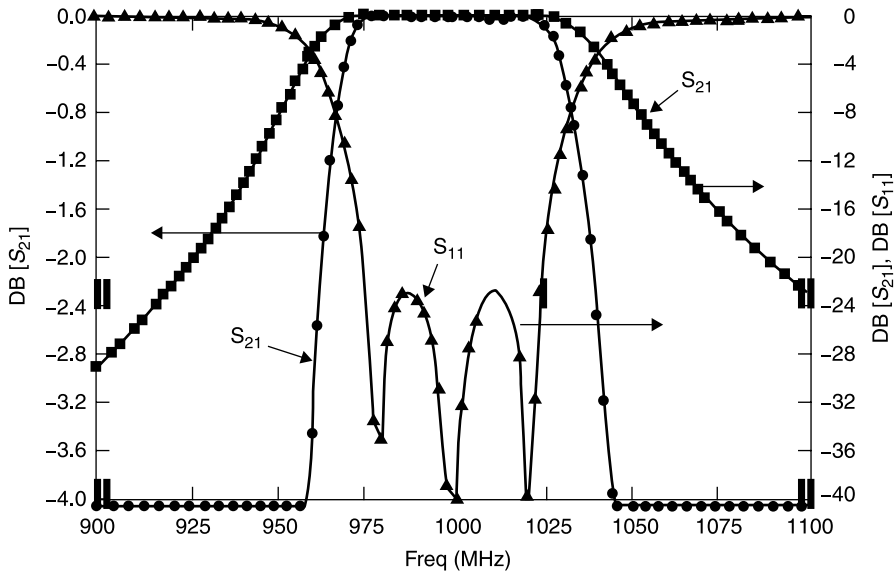
Figure 9.14-5 Calculated loss/isolation of the filter described in Figure 9.14-4.

resonators and 1 pF for the coupling capacitors. The optimization goals were $|S_{11}| < -20$ dB for $975 \le f \le 1025$ MHz and $|S_{21}| < -20$ dB at 900 and 1100 MHz. After some cut-and-try techniques with the optimizer, the goals were changed from -20 dB to -24 dB. The circuit shown in Figure 9.14-4 resulted and its performance is shown in Figure 9.14-5.

Unlike the lossless lumped element simulations previously performed, this simulation has an allowance for component losses, and so the insertion loss performance is a reasonable approximation to what might actually be obtained. A value of 0.05 dB/wavelength of loss is assumed in the calculation for all three half-wave resonator lines, and the coupling capacitors are assumed to have $Q = 100$.

There are numerous topologies for distributed filters, some of which use distributed coupling as was demonstrated for the backward wave coupler in Section 8.2. They can be designed using proprietary software such as the FILTER program, which provides electrical schematics and mechanical layout drawings. Many can also be designed using cut-and-try techniques aided by the optimizer of the network simulator. Further descriptions and discussion of distributed filters is provided by Rhea [1].

9.15 THE RICHARDS TRANSFORMATION

The previous sections of this chapter have shown that there is a large body of filter synthesis theory applicable to lumped elements. It would be desirable if this can be applied to distributed filters. An insight into how to do this can be gained by comparing the impedance variation with frequency of a lumped inductor and a short-circuited length of transmission line [2, 3]. In particular a short-circuited length of transmission line can be substituted for a lumped inductor by defining an equivalence between the two elements described by (9.15-1) and (9.15-2), respectively:

$$Z_L = j\omega L \qquad (9.15\text{-}1)$$

$$Z_{\text{shorted line}} = jZ_0 \tan \theta = jZ_0 \tan\left(\frac{\pi}{2}\frac{f}{f_R}\right) \qquad (9.15\text{-}2)$$

where f_R is the frequency at which the shorted stub length is 90° long. Note that the shorted line has the same mathematical variation as the inductor if we equate

$$L = Z_0 \qquad (9.15\text{-}3)$$
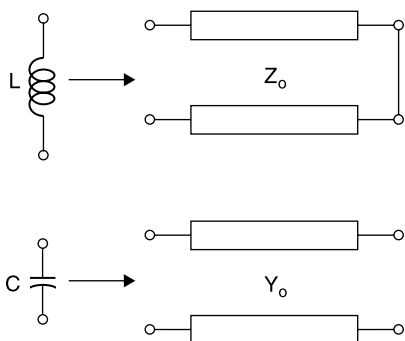
Figure 9.15-1 With the Richards transformation, inductors are replaced by short-circuited transmission lines and capacitors by open-circuited transmission lines.

where L is in henrys and Z_0 in ohms, and define a new frequency variable, Ω, according to

$$j\Omega = j \tan\left(\frac{\pi}{2} \frac{f}{f_R}\right) \tag{9.15-4}$$

The expression in (9.15-4) is the *Richards frequency transformation* (Fig. 9.15-1).

Under this same transformation the behavior of an open-circuited transmission line is mathematically equivalent to that of a capacitor:

$$Y_C = j\omega C \tag{9.15-5}$$

$$Y_{\text{open-circuited line}} = jY_0 \tan\theta = jY_0 \tan\left(\frac{\pi}{2} \frac{f}{f_R}\right) \tag{9.15-6}$$

The mathematical equivalence that we propose does not mean that an inductor's reactance has the same frequency variation as a shorted line, or that a capacitor's reactance variation is the same as that of an open-circuited line, except in the new frequency domain defined by (9.15-4).

As an example of this equivalence, we will design a three-element, low-pass, Chebyshev filter with 0.5-dB ripple, $f_C = 1$ GHz and stub resonant frequency $f_R = 1.4$ GHz. Notice that f_R is an additional degree of freedom in the distributed circuit design. In the low-pass prototype $f_R = \infty$.

Since this is a Chebyshev 0.5-dB ripple design, the loss at 1 GHz will be 0.5 dB and will not exceed this value between 0 and 1 GHz. For this example, we use the tee circuit. The g_i values, from Table 9.8-5 are

$$g_i = 1.5963 \quad 1.0967 \quad 1.5963 \quad 1.000 \text{ (load)}$$
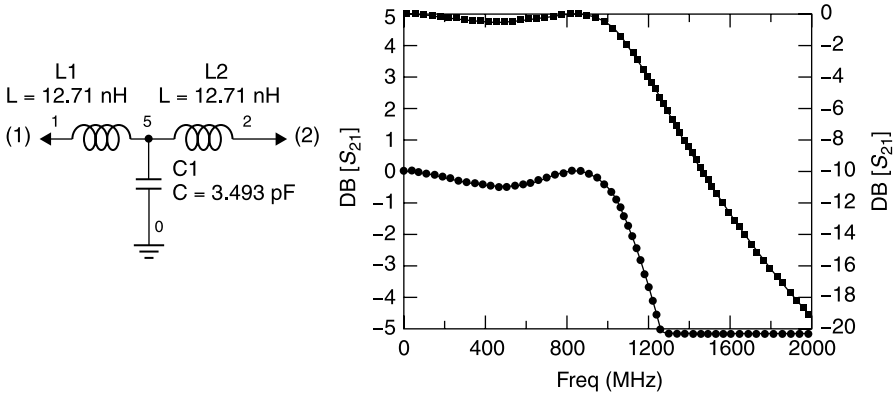
Figure 9.15-2   The 0.5-dB ripple, 50-Ω, Chebyshev low-pass filter.

The corresponding lumped-element tee circuit is obtained by frequency and impedance scaling the g_i values using (9.4-2) and (9.4-4) as shown in Figure 9.15-2 along with its frequency response.

Recall from Section 9.4 that the g_i values are the respective series reactances and shunt susceptances of the filter elements at f_C. This means that the shorted transmission line Z_{01} that replaces a series inductor must have a reactance at f_C of

$$g_1 Z_0 = Z_{01}\Omega_C$$

where

$$\Omega_C = \tan\left(\frac{\pi}{2}\frac{1.0 \text{ GHz}}{1.4 \text{ GHz}}\right) = 2.08$$

Then

$$Z_{01} = Z_{03} = \frac{g_1 Z_0}{\Omega_C} = \frac{1.5963(50 \text{ }\Omega)}{2.08} = 38.4 \text{ }\Omega$$

Similarly, the open-circuited stub Y_{02} that replaces a shunt capacitor in the low-pass prototype filter must have a susceptance at f_C given by

$$g_2 Y_0 = \frac{g_2}{Z_0} = Y_{02}\Omega_C$$

from which

$$Y_{02} = \frac{g_2}{Z_0 \Omega_C} = \frac{1.0967}{(50 \text{ }\Omega)2.08} = 0.01050 \text{ } \mho$$
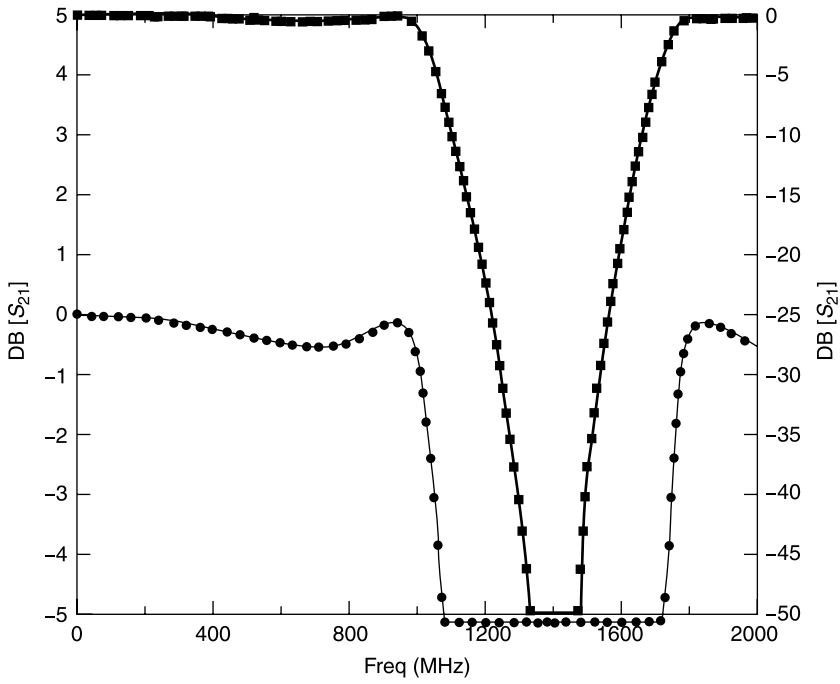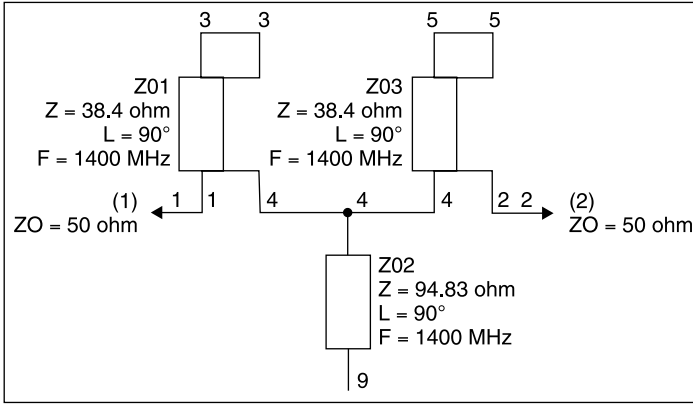
Figure 9.15-3 Circuit and performance of the three-element, 0.5-dB ripple, Chebyshev distributed element filter.
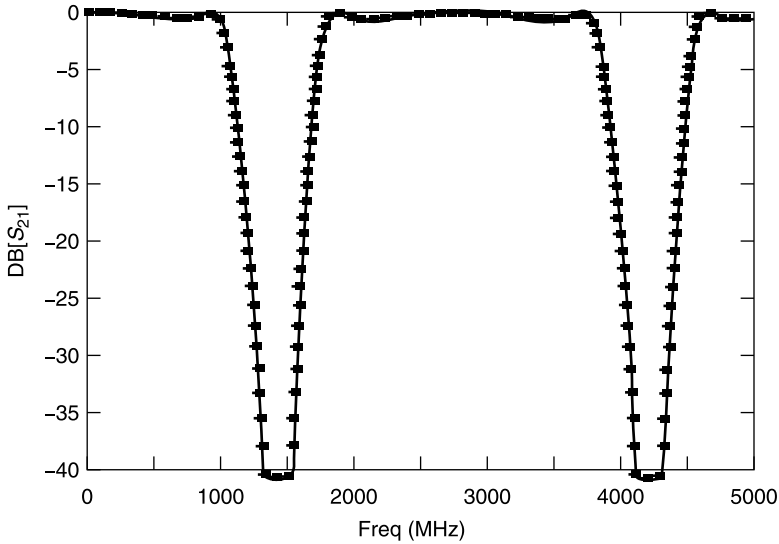
Figure 9.15-4   Extended frequency plot of the distributed filter of Figure 9.15-3.

and

$$Z_{02} = \frac{1}{Y_{02}} = 94.83 \ \Omega$$

This circuit and its performance is shown in Figure 9.15-3.

Notice that the frequency translation has shrunk the frequency domain of the low-pass prototype circuit of Figure 9.15-1 from $0 \leq f \leq \infty$ to the new Ω domain for which $0 \leq f \leq f_R$. That is, while the entire frequency response of the low-pass lumped-element circuit extended from zero to infinity in frequency, that loss performance region has been condensed into the range from zero to f_R, the resonant frequency of the stubs in the distributed filter. At the same time the transformation has introduced a new realm of circuit behavior beyond f_R. In this new domain the loss performance is repeated, mirrorlike, about f_R thereby creating a bandstop filter response. In a similar fashion, a bandpass filter could be created based on the design of a high-pass lumped-element filter.

The stopband of the new distributed filter extends equally on either side of f_R out to f_C. Thus the stopband is 1.0 to 1.8 GHz. Since the filter's performance is defined by transmission line lengths, *the center frequency of the stopband is the arithmetic mean, 1.4 GHz, of its f_C band edge frequencies, 1.0 and 1.8 GHz*. Also, since this is a distributed element filter using quarter-wave resonators, its performance recurs at odd integer multiples of f_R ($f_R, 3f_R, 5f_R, \ldots$), as shown in Figure 9.15-4.

This distributed filter can be realized in coaxial transmission line by embedding the two short-circuit terminated stubs within the center conductor of the main transmission line, as shown in Figure 9.15-5.
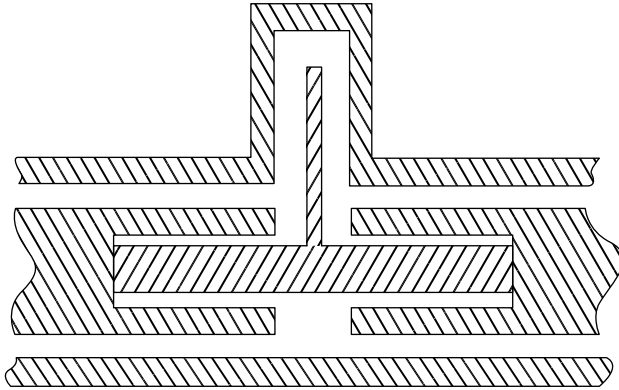
Figure 9.15-5 Cross-sectional sketch of a coaxial line realization of the distributed bandpass filter with series stubs embedded within the center conductor of the main line.

9.16 KURODA'S IDENTITIES

The circuit of Figure 9.15-5 can be realized in coaxial transmission line but not in microstrip or stripline, as would often be desirable, due to the necessity for series short-circuited stubs. *Kuroda's identities of the first kind* (Fig. 9.16-1) al-
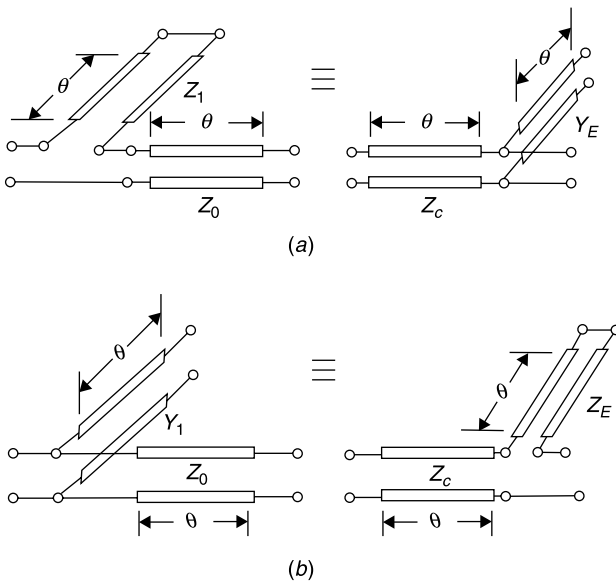
Figure 9.16-1 Kuroda's identities of the first kind applied to transmission lines: (a) $Z_C = nZ_0$ and $Y_E = (n - 1)/nZ_0$ where $n = 1 + Z_1/Z_0$. (b) $Z_C = Z_0/n$ and $Z_E = ((n - 1)/n)Z_0$ where $n = 1 + Y_1Z_0$.

low the replacement of a series-shorted stub and a cascade Z_0 line with a shunt open-circuited stub and a different cascade line, and vice versa [2, 4]. Note that all lines have the same electrical length, θ. Given this requirement and the formulas in Figure 9.16-1, *the Kuroda equivalences apply at all frequencies.*

Notice that the series-shorted stub must be combined with a section of Z_0 line before the equivalence with an open-circuited shunt stub can be made. The addition to the input and output of sections of Z_0 transmission line has no effect on the VSWR or insertion loss of the filter, but it does cause the transformed filter to have added electrical length, which may be significant in some applications. The circuits otherwise are equivalent.

We transform the distributed filter of Figure 9.15-3 to one using all shunt stubs using Kuroda's identity in Figure 9.16-1*a* to transform the series short-circuited stubs to shunt open-circuited stubs. In this filter $Z_{01} = Z_{03} = 38.4$ Ω.
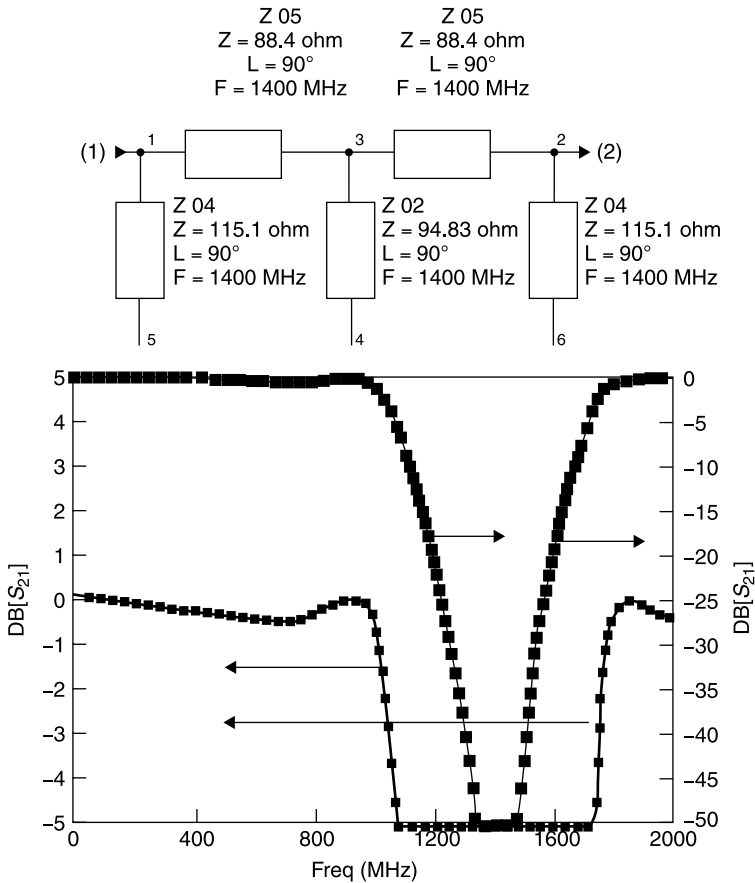
Figure 9.16-2   Circuit and performance of the Chebyshev filter transformed to an all shunt stub format.

(a)

(b)
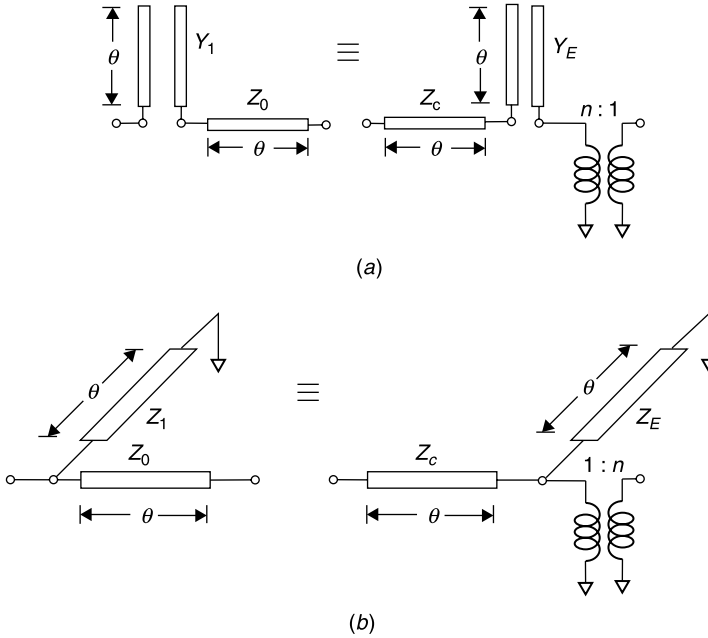
Figure 9.16-3 Kuroda's identities of the second kind: (a) $Z_C = nZ_0$ and $Y_E = 1/[n(n-1)Z_0]$ where $n = 1 + 1/Y_1Z_0$. (b) $Z_C = Z_0/n$ and $Z_E = Z_0/[n(n-1)]$ where $n = 1 + Z_0/Z_1$.

Then, since $Z_0 = 50\ \Omega$, $n = 1.768$, $Z_C = 88.4\ \Omega$, and $Z_E = 1/Y_E = 115.1\ \Omega$. The transformed filter and its response is shown in Figure 9.16-2, from which it can be seen that the insertion loss is identical to the filter of Figure 9.15-3.

Another pair of equivalences is shown in Figure 9.16-3. These are called *Kuroda's identities of the second kind*. The presence of an ideal transformer in the equivalent circuit can be eliminated by the judicious use of combinations of transformations whose total impedance transformation cancels, eliminating the need for the ideal transformers. With these transformations, for example, a three-stub filter with impedance Z_0 and interconnecting Z_0 line could be transformed into a single stub filter with two cascade lines of impedance different from Z_0.

9.17 MUMFORD'S MAXIMALLY FLAT STUB FILTERS

A simple form of distributed bandpass filter consists of shorted TEM mode stubs that are a quarter-wave long at the center frequency, f_0. They are spaced

TABLE 9.17-1 Normalized Stub Admittances for Mumford's Quarter-Wave Shorted Stub Filters [5]

(a) Three-Stub Filters

$10 \log K_3$	k_1	k_2
−12.728	0.100	0.200
−0.944	0.300	0.600
+5.460	0.500	1.000
+10.138	0.700	1.400
+15.560	1.000	2.000
+21.156	1.400	2.800
+27.604	2.000	4.000
+31.904	2.500	5.000
+35.563	3.000	6.000

(b) Four-Stub Filters

$10 \log K_4$	k_1	k_2
−5.170	0.100	0.292
+3.253	0.200	0.571
+13.329	0.400	1.109
+25.668	0.800	2.141
+35.909	1.300	3.395
+44.873	1.900	4.877
+56.734	3.000	7.568

(c) Five-Stub Filters

$10 \log K_5$	k_1	k_2	k_3
+3.452	0.100	0.366	0.532
+13.577	0.200	0.694	0.989
+20.523	0.300	1.005	1.410
+26.002	0.400	1.304	1.808
+30.601	0.500	1.596	2.193
+38.160	0.700	2.166	2.933
+44.324	0.900	2.724	3.648
+54.172	1.300	3.819	5.038
+66.970	2.000	5.702	7.403
+77.874	2.800	7.829	10.058

Source: From [5].

at quarter-wave intervals along the TEM mode main transmission line path. Mumford [5] analyzed this filter type for the maximally flat (Butterworth) transmission response and created tables of values for the characteristic admittances of the stubs relative to that of the main line. The stub admittances nor-
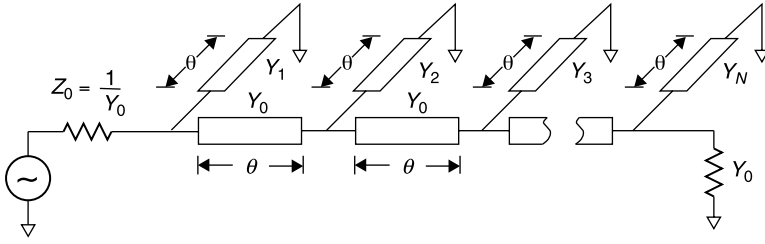
Figure 9.17-1   Schematic of Mumford's quarter-wave stub filters.

malized to Y_0, the characteristic admitttance of the main line and the admittances of the matched generator and load, are listed in Table 9.17-1 for 3, 4 and 5 stub designs. The filters are symmetrical; therefore, only half of the values are listed. The 2-stub design is not listed because, due to symmetry, both stubs are identical. For additional designs using 6 to 10 stubs, refer to Mumford's paper [5] (Fig. 9.17-1). The insertion loss (or isolation) ratio of a filter having N stubs of length l can be computed for any frequency using

$$\text{IL} = 1 + K_N \frac{\cos^{2N} \theta}{\sin^2 \theta} \tag{9.17-1}$$

$$\theta = \frac{2\pi l}{\lambda} \tag{9.17-2}$$

$$K_N = \tfrac{1}{4}\{k_1(k_2 + 2) \cdots (k_i + 2) \cdots (k_N + 2)\}^2 \tag{9.17-3}$$

$$k_i = \frac{Y_i}{Y_0} = y_i \tag{9.17-4}$$

where y_i is the normalized characteristic admittance of the ith stub, λ is the wavelength in the TEM mode medium, and K_N is a measure of the steepness of the isolation slope of the filter. For example, the insertion loss of a two-stub filter whose stubs have the same characteristic impedance as the main line is
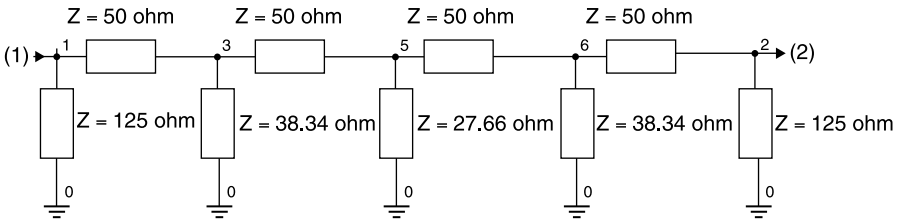
Figure 9.17-2   Five-stub Mumford filter with $K_5 = 26$ dB. For this example, all stubs and interconnecting line lengths are 90° at 1 GHz.
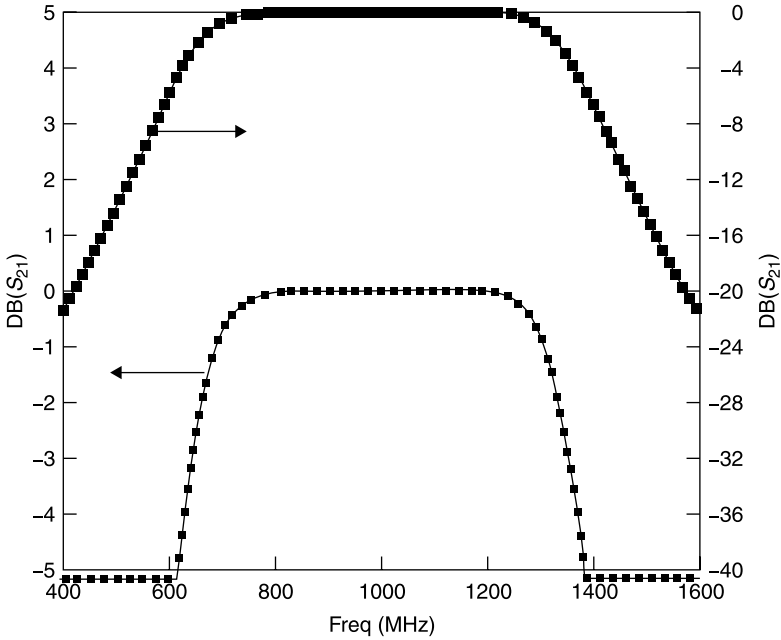
Figure 9.17-3   Calculated insertion loss of the filter in Figure 9.17-2.

$$IL = 1 + \frac{9}{4}\frac{\cos^4 \theta}{\sin^2 \theta} \qquad (9.17\text{-}5)$$

As a filter design example, consider the five-stub model for which $K_5 = 26$ dB in Table 9.17-1(c). When this filter is designed for $Z_0 = 50\ \Omega$ the stub impedances are as shown in Figure 9.17-2, and the performance when the stubs and their spacings are $90°$ at 1 GHz is shown in Figure 9.17-3. The stub impedance values are practical for realization in microstrip or stripline.

9.18 FILTER DESIGN WITH THE OPTIMIZER

Prior to the ready availability of computers, designers were constrained to employ established techniques to design filters and other circuits. Random changing of variables or cut-and-try tuning in the lab could quickly become impractically laborious. Presently, however, even the most basic personal computer can perform thousands of insertion loss evaluations (or other calculations) per second for circuits having 10 or more elements and can be directed through optimizer software to vary the element values to better satisfy a set of performance goals. In principle, any of the filters discussed thus far in this chapter
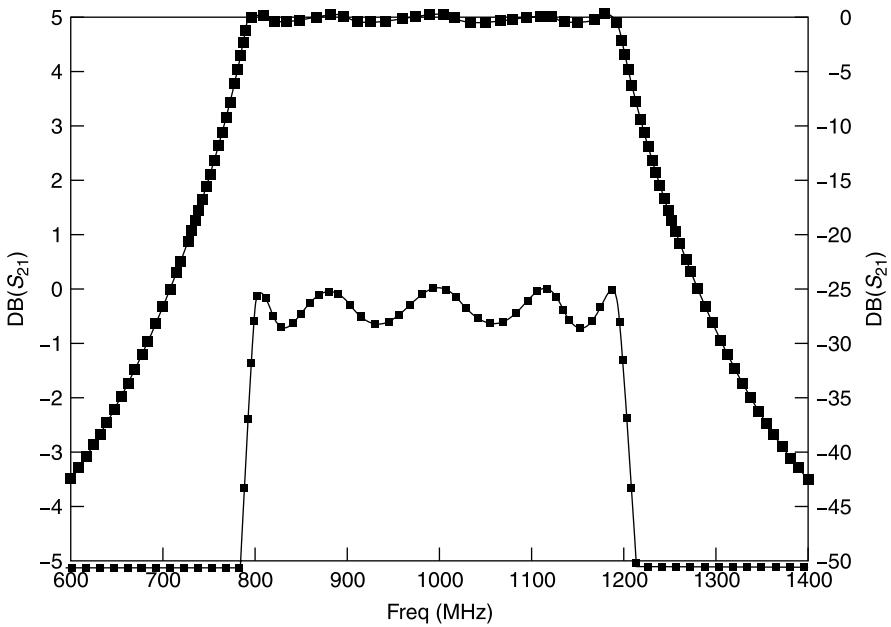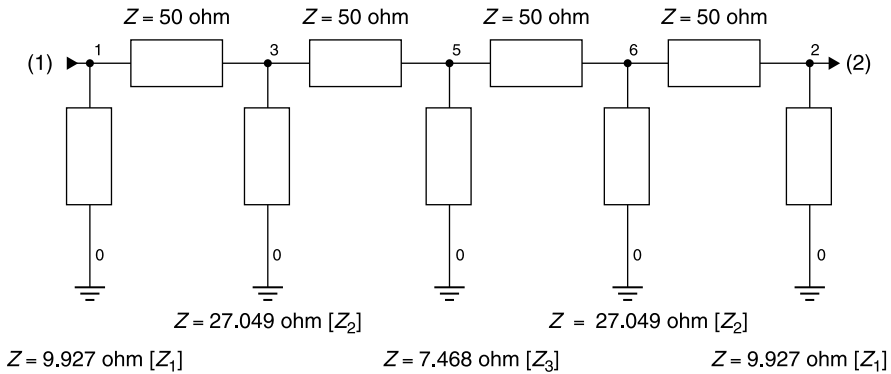
Figure 9.18-1   An 800- to 1200-MHz equal-ripple bandpass filter designed using the network simulator optimizer beginning with the Butterworth filter of Figure 9.18-2.

could have been determined using the optimizer of a network simulator, given an initial circuit of the correct topology.

However, with a completely random search and no good starting point, the optimizer may not converge on the desired solution, and some knowledge of filter performance is required merely to specify an initial circuit topology for the optimization. Furthermore, without prior knowledge of what is achievable,

it is difficult to know what goals are realistic. This is why a study of basic filter designs remains useful. Each different filter type can serve as a starting point from which a design tailored to the application at hand can be optimized.

As an example, consider the five-stub filter described in Figure 9.18-2. Suppose that a passband from 800 to 1200 MHz is required, within which a 0.5-dB maximum insertion loss is acceptable, and that outside of this bandwidth it is desirable to have as much insertion loss (isolation) as possible. Further, suppose that the filter is to be symmetric with respect to input and output, that the main line impedance is to remain 50 Ω, and that the stub lengths and spacings will be 90° at the center frequency of 1 GHz. This leaves only three design variables, the characteristic impedances of the first three stubs.

An optimization was performed with this circuit using as initial goals that $S_{21} > -0.5$ dB from 800 to 1200 MHz and $S_{21} < -20$ dB at 600 and 1400 MHz. This goal was reached by the optimizer almost immediately and therefore was judged to be too easily achieved. After a few additional optimizer executions, it was found that the isolation value could be reduced to < -40 dB at 625 and 1375 MHz. Note that goals must be entered as negative decibels. Optimizers do not recognize that "loss" implies a negative sign.

The resulting circuit and performance are shown in Figure 9.18-1. The isolation is greater than 25 dB only 100 MHz outside of this broad passband of 800 to 1200 MHz. The resulting loss characteristic is a reasonably close approximation to a 0.5-dB equal-ripple Chebyshev filter, even though the circuit optimization began with a Butterworth filter design. The 7- and 10-Ω characteristic impedance stubs required by this design would not be realized in microstrip, but the design might be practical in a coaxial transmission line format.

9.19 STATISTICAL DESIGN AND YIELD ANALYSIS

Using Standard Part Values

In the initial design of a component or system, element values are chosen to provide exactly the performance that is desired. However, in practice, circuits are generally fabricated using standard component values. For example, suppose that a Butterworth low-pass filter with a 3-dB cutoff frequency of 100 MHz is required. The theoretical solution for this filter is shown in Figure 9.19-1.

In checking for the availability of standard parts (Table 9.19-1) that can be provided with a $\pm 10\%$ element value tolerance, it is found that standard values include 47, 56, 150, and 180. Substituting the nearest values to those shown in Figure 9.19-1 leads to the filter shown in Figure 9.19-2.

The adoption of the nearest standard part values to the design has increased the loss at the 100 MHz cutoff frequency from 3.0 to 3.6 dB. The markers on the graph of Figure 9.19-2 indicate that the filter has less than 1 dB of insertion loss up to 86 MHz and greater than 20 dB of isolation above 156 MHz.
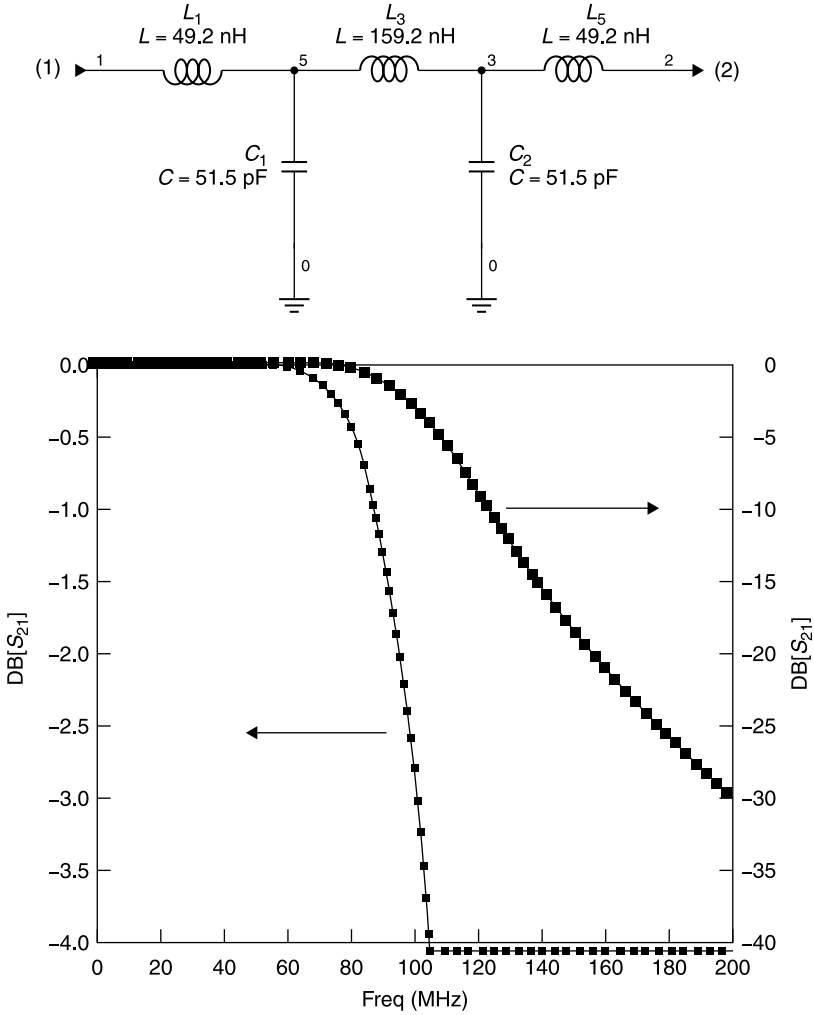
Figure 9.19-1   Insertion loss/isolation of an ideal Butterworth filter with 100 MHz, 3-dB cutoff frequency.

The Normal Distribution

Next, we recognize that even standard value components are not exactly realized but have a distribution of values. When a large quantity of parts are made according to a repetitive process, their distribution usually can be described by a "bell-shaped" curve having the form described by (9.19-1) and sketched in Figure 9.19-3:

$$N(x) = A \, e^{[\frac{-1}{2}(\frac{x-x_0}{\sigma})^2]} \qquad (9.19\text{-}1)$$

where A and σ are constants.

TABLE 9.19-1 U.S. Standard C83.2[a]

Part Tolerance	$\pm 5\%$	$\pm 10\%$	$\pm 20\%$
Step Multiplier	$10^{1/24} = 1.10$	$10^{1/12} = 1.21$	$10^{1/6} = 1.47$
	10	10	10
	11		
	12	12	
	13		
	15	15	15
	16		
	18	18	
	20		
	22	22	22
	24		
	27	27	
	30		
	33	33	33
	36		
	39	39	
	43		
	47	47	47
	51		
	56	56	
	62		
	68	68	68
	75		
	82	82	
	91		
	100	100	100

[a]Smaller or larger values use decimal multiplication of the values for the separate $\pm 5\%$, $\pm 10\%$, and $\pm 20\%$ series values [8].

The distribution in Figure 9.19-3 is so frequently encountered that it is called the *normal distribution*. It describes the likely density of parts, $N(x)$, distributed about a center value x_0. It is also called a *Gaussian distribution* about a nominal value. The expression in (9.19-1) is a continuous quantity; therefore, it can only be applied with good approximation when the quantity of parts is large. The constant σ defines the width of the curve. When $|x - x_0| = \sigma$ the density $N(x)$ is about 60% of its peak value. The constant A is evaluated such that the density of parts integrated over all possible values is equal to N, the total number of parts. That is,

$$N \equiv \int_0^\infty N(x)\,dx = A \int_0^\infty e^{\left[\frac{-1}{2}\left(\frac{x-x_0}{\sigma}\right)^2\right]}\,dx \qquad (9.19\text{-}2)$$

In order to separate this characteristic function from the actual quantity of
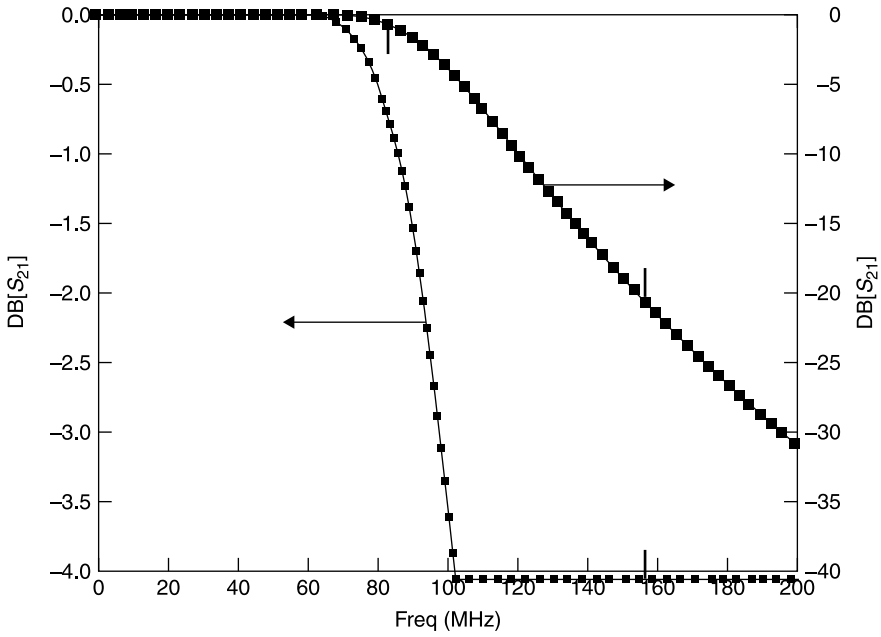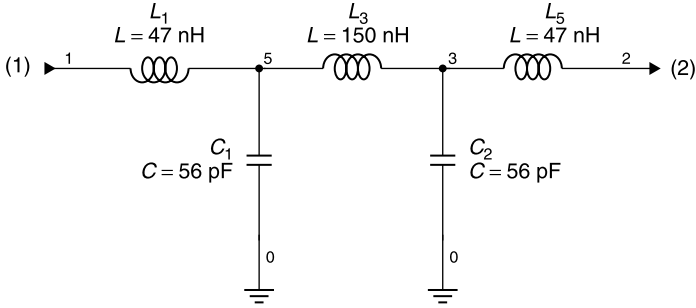
Figure 9.19-2 Low-pass filter realized using standard part values.

parts, the practice is to express $N(x)$ as a product of the total number of parts and the *probability factor*, $\varphi(x)$. With this definition

$$N(x) = N\varphi(x) \tag{9.19-3}$$

where

$$\varphi(x) = C\, e^{\left[\frac{-1}{2}\left(\frac{x-x_0}{\sigma}\right)^2\right]} \tag{9.19-4}$$
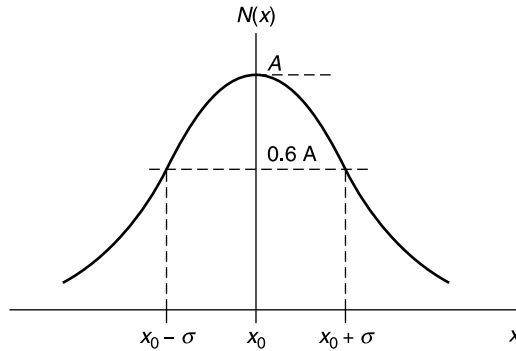
Figure 9.19-3   Normal distribution of parts, also called a Gaussian distribution.

Since $\varphi(x)$ is a probability factor, its value is bounded by

$$0 \le \varphi(x) \le 1 \qquad (9.19\text{-}5)$$

That is, the probability of an event lies somewhere between zero (when the event has no chance of occurring) and unity (when the event is a certainty). Suppose that these statistics are to represent a manufacturing lot of capacitors being made with an intended standard value of 47 pF. Then, theoretically, the range of part values is all positive values for x, and the constant, C, must be defined to satisfy

$$\int_0^\infty \varphi(x)\,dx \equiv 1 = C \int_0^\infty e^{\left[\frac{-1}{2}\left(\frac{x-x_0}{\sigma}\right)^2\right]}\,dx \qquad (9.19\text{-}6)$$

Imposition of the condition in (9.19-6) will also satisfy that of (9.19-5). The integration in (9.19-6) required to evaluate C is tricky [7, Chapter 8, Sec. 10]. To perform the integration, express the integral with either of two dummy variables, x and y:

$$I = \int_{-\infty}^{\infty} e^{-x^2/2}\,dx = \int_{-\infty}^{\infty} e^{-y^2/2}\,dy \qquad (9.19\text{-}7)$$

The integration is carried out from $-\infty$ to $+\infty$ because in general the probability function might have values for both negative and positive values of x. Squaring I and using both the x and y variable forms of (9.19-7) gives

$$I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dx\,dy \qquad (9.19\text{-}8)$$

The integration in (9.19-8) corresponds to integrating the function $e^{-(x^2+y^2)/2}$ over the entire xy plane. Expressing (9.19-8) in polar coordinates, the same function and area of integration can be rewritten as

$$I^2 = \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r\, dr\, d\theta = 2\pi \int_0^{\infty} e^{-r^2/2}\, d\left(\frac{r^2}{2}\right) = 2\pi \qquad (9.19\text{-}9)$$

Hence,

$$I = \int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2\pi} \qquad (9.19\text{-}10)$$

and the constant C in (9.19-6) must be equal to $(1/\sigma)\sqrt{1/2\pi}$ so that the normal, or Gaussian, probability function satisfies (9.19-6). The probability function is then defined by (9.19-11) and plotted in Figure 9.19-4:

$$\varphi(x) = \frac{1}{\sigma}\sqrt{\frac{1}{2\pi}}\, e^{\left[\frac{-1}{2}\left(\frac{x-x_0}{\sigma}\right)^2\right]} \qquad (9.19\text{-}11)$$

and the density of parts $N(x)$ in terms of the total number of parts N is given by

$$N(x) = N\varphi(x) = \frac{N}{\sigma}\sqrt{\frac{1}{2\pi}}\, e^{\left[\frac{-1}{2}\left(\frac{x-x_0}{\sigma}\right)^2\right]} \qquad (9.19\text{-}12)$$

Given a sufficiently large N for applying statistics (say, a number of capacitors), the number of capacitors N_{SPEC} having a specified value (such as the capacitance value being within $x_1 \le x \le x_2$) is given by

$$N_{\text{SPEC}} = N\int_{x_1}^{x_2} \varphi(x)\, dx = \frac{N}{\sigma}\sqrt{\frac{1}{2\pi}}\int_{x_1}^{x_2} e^{\left[\frac{-1}{2}\left(\frac{x-x_0}{\sigma}\right)^2\right]}\, dx \qquad (9.19\text{-}13)$$

The indefinite integral in (9.19-12) does not have a known analytic solution. The number of in-specification parts can be estimated by performing a graphical integration of the curve in Figure 9.19-4, below which the number of squares contained within $-3\sigma \le x \le +3\sigma$ is approximately 240.

For example, suppose that a manufacturer makes 1000 capacitors intended to have a standard value of *47* pF. However, due to an error in the process, the distribution is centered at *52* pF ($x_0 = 52$ pF), with $\sigma = 0.096 x_0 = 5$ pF. How many parts will meet a specification of 47 pF $\pm$ 5 pF (*a $\pm$10.6% tolerance*)? The distribution is shown in Figure 9.19-5. The shaded area represents the usable parts that meet the capacitance specification, a little less than one half of those produced.

Other Distributions

While the normal distribution is the most common for part manufacture, many factors can influence the distributional outcome. One should not assume that parts obtained with a $\pm$ tolerance can be accurately represented by a normal
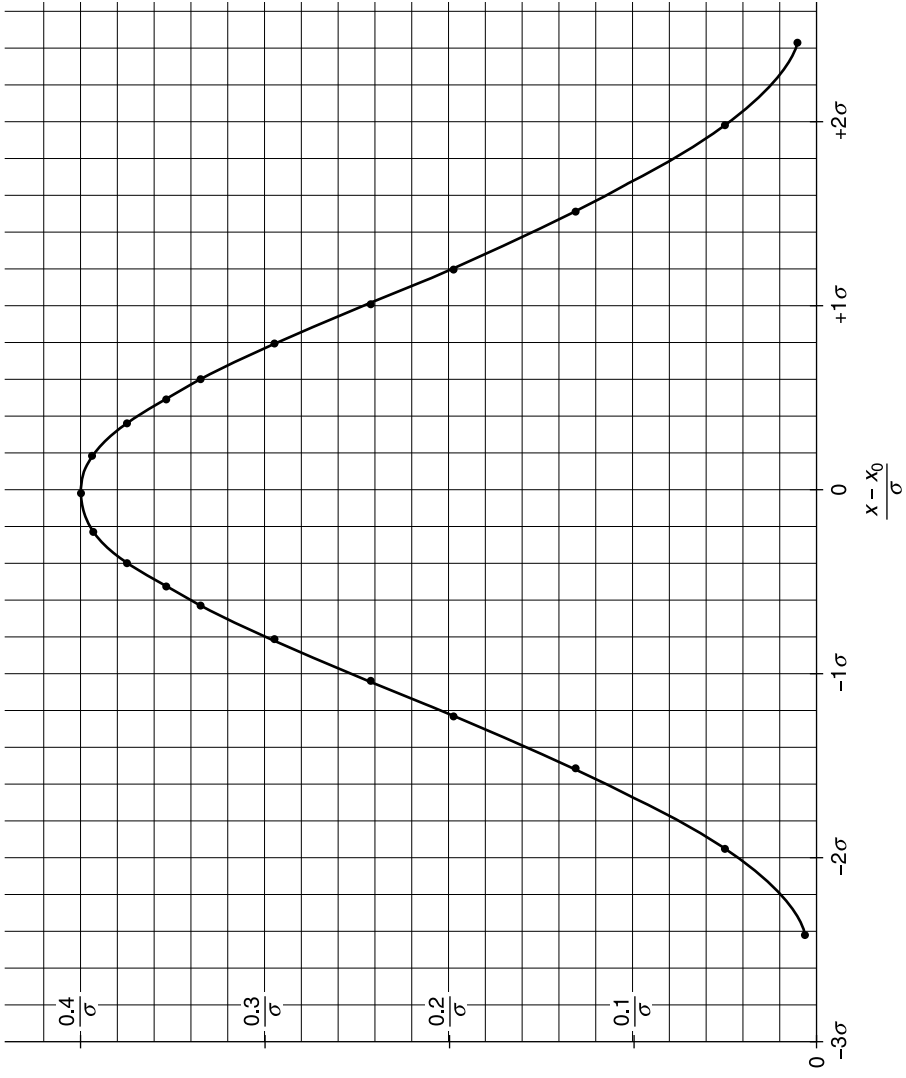
Figure 9.19-4 Normal, or Gaussian, probability function. The area under the curve is approximately 240 squares.
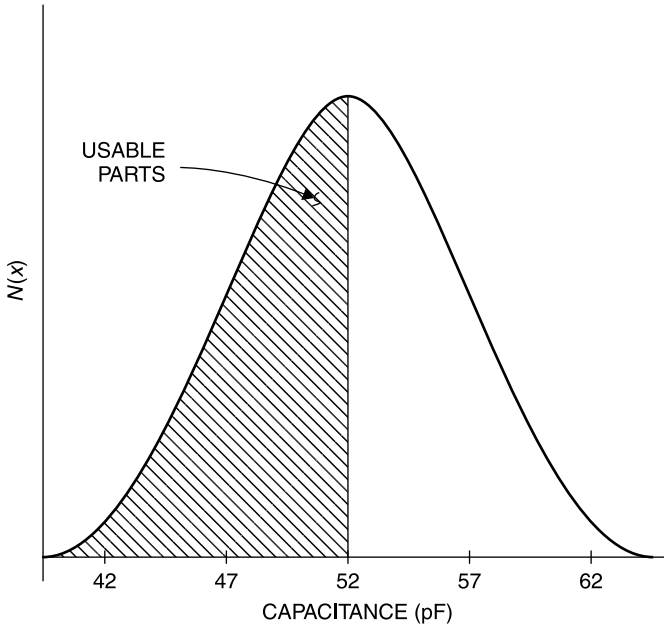
Figure 9.19-5   Normal distribution of capacitors centered at 52 pF with one sigma $=$ 5 pF and the quantity that meet a 47-pF $\pm$ 5 pF specification (shaded area).

distribution centered on their nominal value. As was seen from the previous example, even when the normal distribution does apply, its center value, x_0, may not be the nominal value for that part because the manufacturer's process may not be sufficiently controlled to center the distribution on the part value desired.

Even when the distribution is centered properly on the nominal value, parts delivered to a $\pm 20\%$ tolerance will not necessarily be represented by a normal
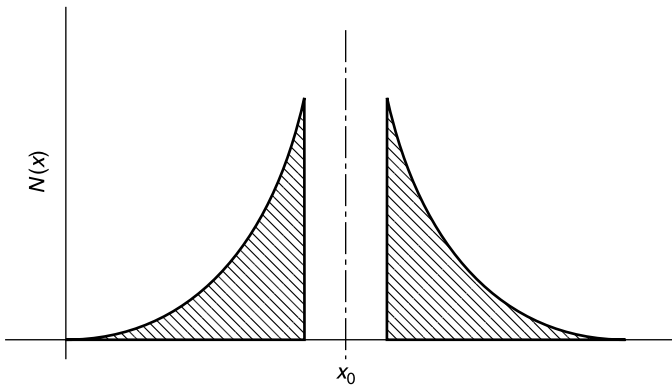
Figure 9.19-6   Gaussian (normal) distribution showing how the center values might be removed to leave a *depleted normal distribution*.

distribution because the supplier may have sold the parts that were within 10% of the nominal value to another customer at a premium price, leaving a *depleted normal distribution* for the parts that will be delivered to a $\pm20\%$ tolerance (Fig. 9.19-6).

There are numerous other possible distributions for component values. Another common type is the uniform distribution in which all values between the plus and minus component tolerances are equally probable. This might be the approximate distribution if a manufacturer combines parts made from several production runs, each of which has a different x_0. From the uniform distribution, a depleted uniform distribution can result for the same reason as that described for the depleted normal distribution.

To control manufacturing yield, the actual distribution of parts must be controlled, either by first measuring a set of available parts or else procuring parts to a distributional specification (and then verifying on their receipt that they meet that specification). *In manufacture, controlling yield is no less important than designing the circuit.*

Returning to the Butterworth low-pass filter example, the insertion loss does not exceed 1 dB for frequencies up to 86 MHz, and the isolation is greater than 20 dB for frequencies above 156 MHz. Suppose we set a specification for our
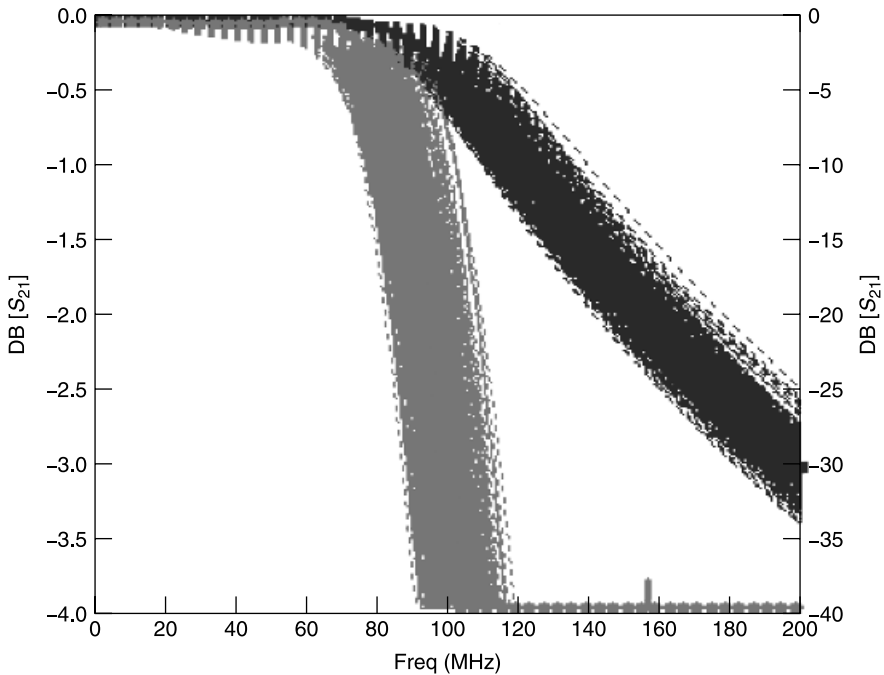
Figure 9.19-7   Calculated performance distribution of 500 filters made using one-sigma values that are 15% of the nominal component value. Only 50 filters passed the 1-dB loss and 20-dB isolation specification for a total yield of 10%.
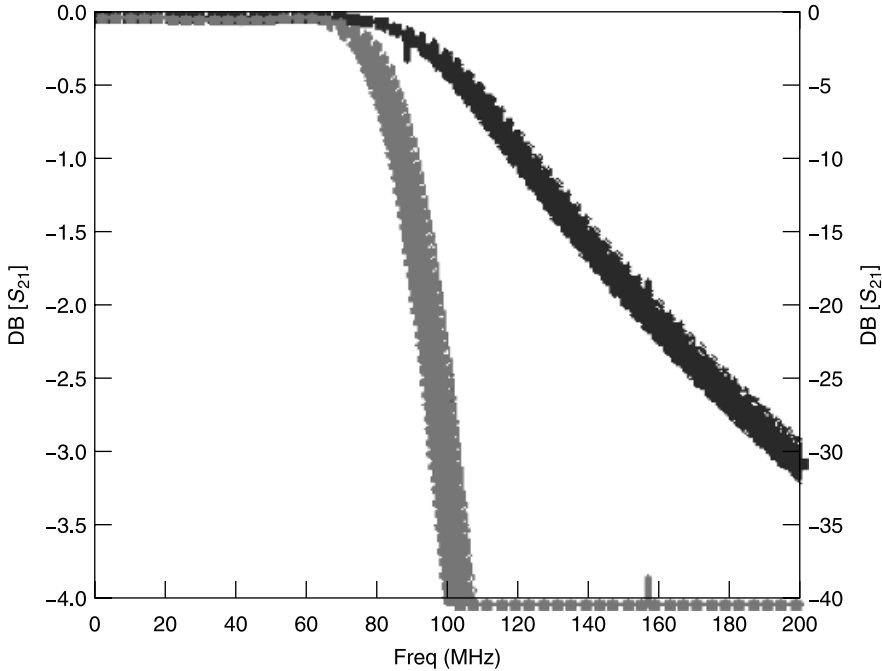
Figure 9.19-8   Distribution of Butterworth filter performance when the components have normal distributions with a one-sigma value that is 5% of the nominal component value. A total of 60 units passed the 1-dB-loss and 20-dB-isolation specification for a yield of 12%.

filter that the loss must be below 1.0 dB up to 86 MHz and the isolation must be greater than 20 dB above 156 MHz.

Further, assume for this example that all components have a normal distribution of values with a one sigma equal to 15% of the nominal value. The resulting performance of 500 filters "built" using these parts is shown in Figure 9.19-7.

Of course, the specifications could be relaxed to allow more filters to pass. Alternatively, components could be procured with a tighter distribution. If the one-sigma value is reduced to 5% of the nominal value, the resulting distribution is that shown in Figure 9.19-8.

While the spread is much tighter than for the 15% one sigma, the yield increased to only 59 units, a 12% yield. If the specification limits are relieved to require less than 1 dB of loss up to 80 MHz and greater than 20 dB of isolation at 170 MHz and above, and if the one-sigma tolerance is made 10%, then 473 units meet the specifications, for a yield of nearly 94.6%. Clearly, both the specifications and the control of parts are crucial to the achievement of high yields.

REFERENCES

1. Randall W. Rhea, *HF Filter Design and Computer Simulation*, Noble Publishing, Atlanta, GA, 1994. *This book is must reading for the filter designer. It contains an excellent introduction to filter theory with numerous examples of both lumped element and distributed filters. It includes practical effects such as coil winding and the element Q's that are practically realizable.*

2. Peter A. Rizzi, *Microwave Engineering, Passive Circuits*, Prentice-Hall, Englewood Cliffs, NJ, 1988. *Excellent microwave engineering textbook covering theory and design of transmission lines, couplers, filters, and numerous other passive devices.*

3. P. I. Richards, "Resistor-Transmission Line Circuits," *Proceedings of the Institute of Radio Engineers (IRE)*, Vol. 36, February, 1948, pp. 217–220.

4. Ralph Levy, "A general equivalent circuit transformation for distributed networks," *IEEE Transactions on Circuit Theory*, CT-12, September 1965, pp. 457–458.

5. W. W. Mumford, "Tables of stub admittances for maximally flat filters using shorted quarter wave stubs," *IEEE Transactions on Microwave Theory and Techniques*, Vol. MTT-13 No. 5, September 1965, pp. 695–696.

6. Matthei, Young, and Jones, *Microwave Filters, Impedance Matching Networks, and Coupling Structures*, McGraw-Hill, New York, 1964 (now also available from Artech House, Norwood, Massachusetts). *Often called the Bible of the filter world.*

7. I. S. Sokolnikoff and R. M. Redheffer, *Mathematics of Physics and Modern Engineering*, McGraw-Hill, New York, 1958.

8. Howard W. Sams & Co., *Reference Data for Radio Engineers*, 5th ed., Howard W. Sams & Co., 1974.

EXERCISES

E9.4-1   **a.** Design a 50-Ω, five-element, lumped-element, Butterworth low-pass filter with a 3-dB cutoff at 1000 MHz to take out the harmonics in a cellular phone transmitter. Use a series *L* as the first element.

  b. How much attenuation would you expect for the second harmonic of an 850-MHz signal?

  c. Suppose we need at least 40-dB attenuation of the second harmonic of 850 MHz. What can you do to achieve this and also minimize passband insertion loss?

  d. Prove your results using a circuit simulator.

  e. If the system bandwidth is 825 to 875 MHz and you wish to eliminate the second harmonic, which could be between 1650 and 1750 MHz, what changes would you make to the filter to minimize its loss in the passband and maximize its isolation to the second-harmonic band?

E9.5-1   **a.** Convert the filter you designed in E9.4-1 into a high-pass filter of the same type and cutoff frequency.

b. Examine your result using a circuit simulator (create a new schematic to analyze this circuit).

c. Are the elements as easy to realize?

E9.8-1 A 50-Ω, 0.0432-dB ripple (20 dB return loss), Chebyshev three-section bandpass filter with 0.0432-dB band edge frequencies of 700 and 1000 MHz is required.

a. What is f_0?

b. What is the fractional bandwidth?

c. What are the low-pass g_i values?

d. What are the actual L and C values of the filter?

e. Plot the loss versus frequency performance using a network simulator.

f. What is the order of the filter?

g. At what rate does the isolation increase in decibels/octave?

h. What is the effect if inductors have a Q of 50 and capacitors have a Q of 100?

i. What would you say is the highest practical frequency at which a filter of this type can be built?

E9.8-2 **a.** Design a Chebyshev 0.0432-dB ripple (20-dB return loss) bandpass filter in the same manner used for E9.8-1 but for a passband of only 800 to 900 MHz.

b. Verify the design using a network simulator.

c. What can you say about the relative practicality of the required components for this filter versus the 700- to 1000-MHz filter designed in E9.8-1? What do you believe causes the added difficulty?

d. Use the circuit simulator to determine the passband insertion loss when the filter is realized using capacitor Q's of 100 and inductor Q's of 50.

E9.14-1 **a.** Repeat the design of the 800- to 900-MHz, 0.0432-dB ripple (20-dB return loss), Chebyshev bandpass filter, but this time realize it as a microstrip end-coupled (half-wavelength resonator) filter with a 0.031-in.-thick substrate of dielectric constant 2.22. *Hint:* Use a filter program if you have one. If not, use a network simulator and begin with three, 50-Ω, half-wavelength resonators end-coupled by capacitors. Use the optimizer to hold loss to 0.0432-dB in the passband. Make the line lengths and capacitor values adjustable in the optimization.

b. Can you get the same loss performance as the lumped-element filter?

c. Can you get the same isolation performance?

d. Compare the practicality of building this filter with the lumped-element filter.

E9.15-1 The Richards transformation is to be used to design a three-stub bandstop filter with $f_R = 1.30$ GHz and a 3-dB bandwidth of 600 MHz (i.e., $f_C = 1$ GHz).

    a. Base the design on a three-element tee configuration Butterworth low-pass filter. Determine the characteristic impedances of the series-connected shorted stubs, Z_{SS}, and the shunt-connected open-circuited stub, Z_{OC}. Also calculate the stub lengths l. Assume a 50-Ω system impedance ($Z_0 = 50\ \Omega$) and air dielectric. Make sketches of the equivalent circuit of the lumped-element and stub filters showing element values.

    b. Plot the loss/isolation versus frequency response from 1 to 2 GHz. Verify that the 3-dB bandwidth is 600 MHz. What is the 20-dB bandwidth?

E9.16-1 Use the appropriate Kuroda identity to convert the distributed tee filter design of E9.15-1 to a three shunt-connected open-circuit stub filter with the same performance.

    a. Calculate the characteristic impedance of the end stubs Z_{ES} and the interconnection lines Z_{01}.

    b. Plot the loss versus frequency response of the shunt stub filter and compare it to that of the tee configuration filter of E9.15-1.

E9.16-2 Prove the equivalence of Kuroda's identity in Figure 9.16-1*a*. *Hint:* If two circuits have identical behavior, they must have the same *ABCD* matrix.

E9.19-1 **a.** Make a scaled sketch of the distribution curve for a manufacturing run of 1000 resistors having a center of distribution value, $x_0 = 68\ \Omega$, and a standard deviation, $\sigma = 5\ \Omega$.

    b. Use your drawing to estimate the number of resistors that will meet a $\pm 5\%$ tolerance (say, $\pm 3.5\ \Omega$) about 68 Ω.

Transistor Amplifier Design

10.1 UNILATERAL DESIGN

Evaluating *S* Parameters

The design of a transistor amplifier is the creation of a circuit environment in which a selected transistor performs as closely as possible to a set of design objectives. The circuit designer may have no influence over the transistor's parameters. These are established by its geometry, quality of manufacture, and semiconductor physics. Usually the circuit designer selects a transistor from a catalog, guided by and limited to the properties described by its *S* parameters.

To the circuit designer the transistor is a two-port network described by a table of *S* parameters taken over the entire frequency domain over which it has gain. After the bias and heat sinking needs of the transistor have been satisfied, the RF design proceeds using the *S* parameters. At this point it is of no consequence whether the device is a bipolar or field-effect transistor (FET) or any other device whose *S* parameters suggest the prospect of gain.

As was shown in Chapter 6, when the source and load impedances are the same as those used to determine the *S* parameters, the magnitude of S_{21} is the ratio of the outgoing wave b_2 to the incoming wave a_1. Hence it is equivalent to the *voltage or current gain* of the amplifier. Similarly, the magnitude of the square of S_{21} is equal to the power gain.

As an example, suppose it is desired to design an amplifier to operate at 1 GHz with 50-Ω source and load impedances, and that it is desirable to obtain as much gain as practical at 1 GHz. On reviewing tables of *S* parameters provided in a catalog or in an electronic file within a circuit simulator, suppose that the Motorola 2N6679A bipolar transistor is selected for the amplifier design. Its *S* parameters are shown in the standard format in Table 10.1-1.

We see that the magnitude of S_{21} is *6.6 at 1 GHz*. Then the basic gain, without tuning, in a *50*-Ω system is

$$G = 20 \log|S_{21}| = 20 \log 6.6 = 16.4 \text{ dB} \qquad (10.1\text{-}1)$$

TABLE 10.1-1   The S Parameter File for the Motorola 2N667A Bipolar Transistor[a]

! 2N6679A.S2P
! 2N6679
! VCE = 15 V; IC = 25 mA

GHZ S MA R 50

! S-PARAMETER DATA

0.1	0.60	−76	38.6	141	0.01	55	0.83	−20
0.5	0.67	−158	12.7	95	0.02	40	0.50	−27
1	0.68	−178	6.6	77	0.03	53	0.46	−32
1.5	0.68	170	4.4	64	0.04	54	0.47	−41
2	0.69	162	3.4	54	0.05	54	0.47	−50
2.5	0.69	154	2.7	42	0.06	55	0.49	−59
3	0.69	146	1.3	31	0.07	55	0.53	−70
3.5	0.69	138	1.9	21	0.08	54	0.55	−79
4	0.69	131	1.7	11	0.09	51	0.57	−89
4.5	0.69	123	1.5	1	0.10	49	0.59	−97
5	0.69	114	1.4	−9	0.12	44	0.62	−106
5.5	0.69	106	1.2	−19	0.14	39	0.64	−113
6	0.69	98	1.1	−28	0.15	33	0.68	−122
6.5	0.69	90	1.0	−37	0.17	31	0.69	−130

Source: From the *Genesys S* parameter directory.

[a] By convention the columns contain left to right: freq (in GHz), $S_{11}, S_{21}, S_{12}, S_{22}$, where each S parameter is contained in two columns, the first is the magnitude (numeric, not decibel) and the second is the phase angle in degrees.

Transistor Biasing

At this point suppose that we consider this to be desirable performance. It is necessary to design a bias circuit [1, pp. 273–283]. For the bias a 20-V source is selected, since the S parameter file indicates the performance was obtained with a 15-V bias between collector and emitter. From the transistor data, we see that the performance was obtained with $V_{CE} = 15$ V and $I_C = 25$ mA. The DC current gain $\beta = I_C/I_B$ is not listed in the S parameter file, but on contacting the manufacturer, suppose that at room temperature $\beta = 40$. Then a suitable biasing network is as shown in Figure 10.1-1.

The resistor values are calculated as follows. A base current of 0.63 mA is required (25 mA/40). But we want a shunt path to carry about 5 times this value, say 3 mA, so that when the 15-V level falls, the base current will fall nearly proportionately. Since the transistor is a silicon NPN, the base emitter junction is a silicon diode. Hence it will have a voltage drop at turn-on of about 0.75 V. The emitter-base shunting resistor is calculated as:

$$R_3 = \frac{0.75 \text{ V}}{3 \text{ mA}} = 250 \ \Omega \qquad (10.1\text{-}2)$$
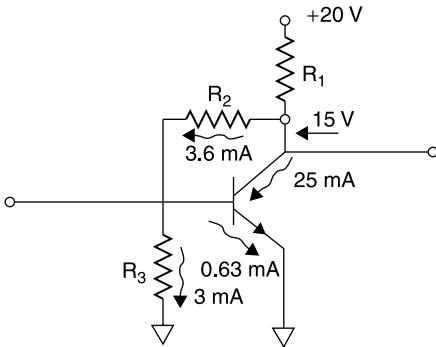
Figure 10.1-1 Bias circuit for the 2N6679A transistor.

Next,

$$R_2 = \frac{15 \text{ V} - 0.75 \text{ V}}{3.63 \text{ mA}} = 3925 \ \Omega \tag{10.1-3}$$

Finally,

$$R_1 = \frac{20 \text{ V} - 15 \text{ V}}{28.6 \text{ mA}} = 175 \ \Omega \tag{10.1-4}$$

In practice, the closest values of standard resistors may be used. This bias circuit tends to self-compensate. As temperature increases, the base current increases for a given V_{BE}. However, this causes an increase in I_C, causing a larger voltage drop in R_1, reducing V_{BE}.

By themselves, the resistors R_2 and R_3 are large compared to the 50-Ω RF impedance level at the input. However, their parasitics may cause them to place an excessive load on the RF circuit. Furthermore, it is undesirable to have RF signals present in the bias circuit since this may cause undesired RF radiation. Therefore, circuitry to isolate the bias from the RF signals is employed, as, for example, that shown in Figure 10.1-2.

Basically the bias isolation circuitry presents a high RF impedance from the transistor terminals to those of the bias circuitry. This can be accomplished using discrete components (Fig. 10.1-2*a*). For the present 1 GHz, 50-Ω amplifier example isolating elements of $L = 100$ nH and $C = 100$ pF might be employed. When space permits, the isolating elements can be "printed" as distributed elements along with the RF circuit. The transmission lines shown are made a quarter wavelength long at the RF center frequency of the amplifier. For the *50* Ω amplifier example, the high impedance lines made $Z_H = 150$ Ω, and the low impedance lines made $Z_L = 25$ Ω.

The bias isolating networks are crucially important in amplifier design and should not be given short shrift in the design effort. As we shall later discuss, transistors may have gain well above the design bandwidth. In the present

(a)

(b)
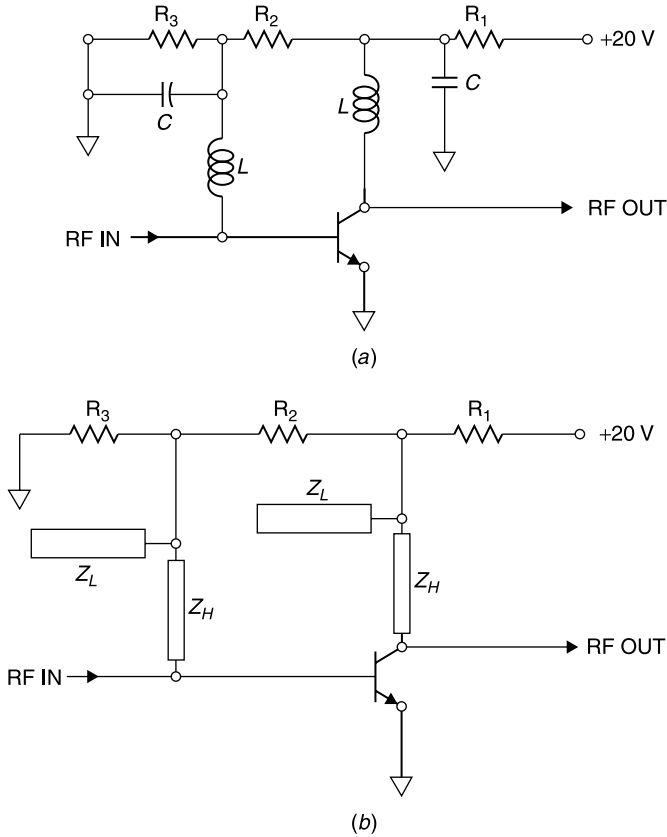
Figure 10.1-2   Transistor with bias isolation components: (a) discrete bias isolation and (b) distributed bias isolation.

1-GHz amplifier example, it can be seen from Table 10.1-1 that the transistor has gain above 5 GHz. Therefore, it could break into oscillation given certain out-of-band impedance loading.

The *single-stage* bias circuits shown in Figure 10.1-2 have only single high impedance and low impedance elements for isolation. They are shown for illustration of the bias isolating methods and are not presented as adequate designs for all applications. Where practical, *multistage* bias isolation is preferable as shown in Figure 10.1-3.

An informal measure of the effectiveness of the distributed bias structure is to bring a screwdriver tip near the open-circuit ends of the *bias flags* (the open-circuited ends of the low-impedance line sections) while measuring the output power of the amplifier. A significant power change experienced when so probing the end of the second bias flag, Z_{L2}, indicates the bias circuit may be resonant at the wrong frequency or that more isolating stages are required.

(a)

(b)
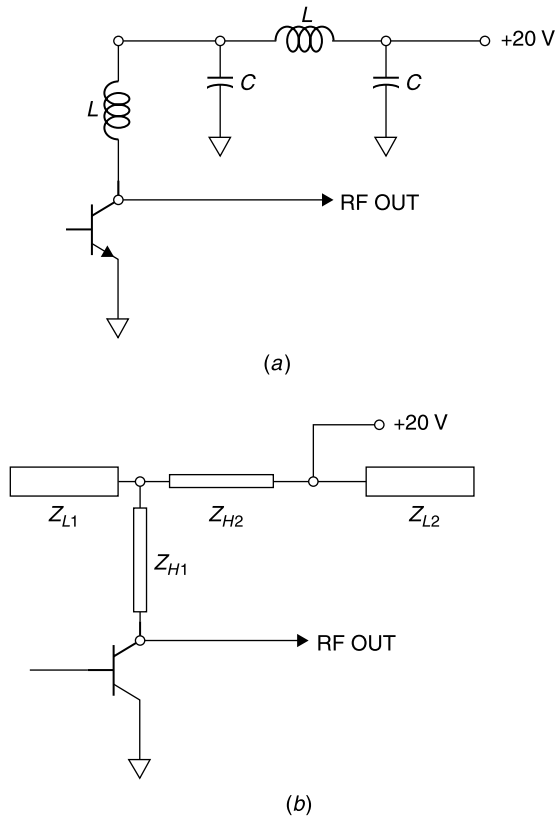
Figure 10.1-3 Two-stage bias isolating circuits: (a) discrete and (b) distributed.

Evaluating RF Performance

Most circuit simulators can accept the S parameter table to provide an analysis over frequency, interpolating the data to calculate the circuit behavior between frequencies for which S parameter data are available. The circuit evaluation could be performed with bias elements, but we will assume in this text that sufficient bias isolation is achieved, and only the RF circuit will be shown for purposes of illustration. However, when the amplifier design is critical, as for an integrated circuit, it is wise to include the bias circuit and any peripheral circuit elements on the broadband simulation of the amplifier. The gain of the present example, using the 2N6679A transistor, is shown in Figure 10.1-4. As will be common in this text, the impedance of the input and output ports is 50 Ω unless otherwise noted.

At this point, *if this performance is satisfactory*, the electrical portion of the amplifier design is complete. On the other hand, on examining the input and output reflection coefficients, S_{11} and S_{22}, respectively, in Table 10.1-1, it can
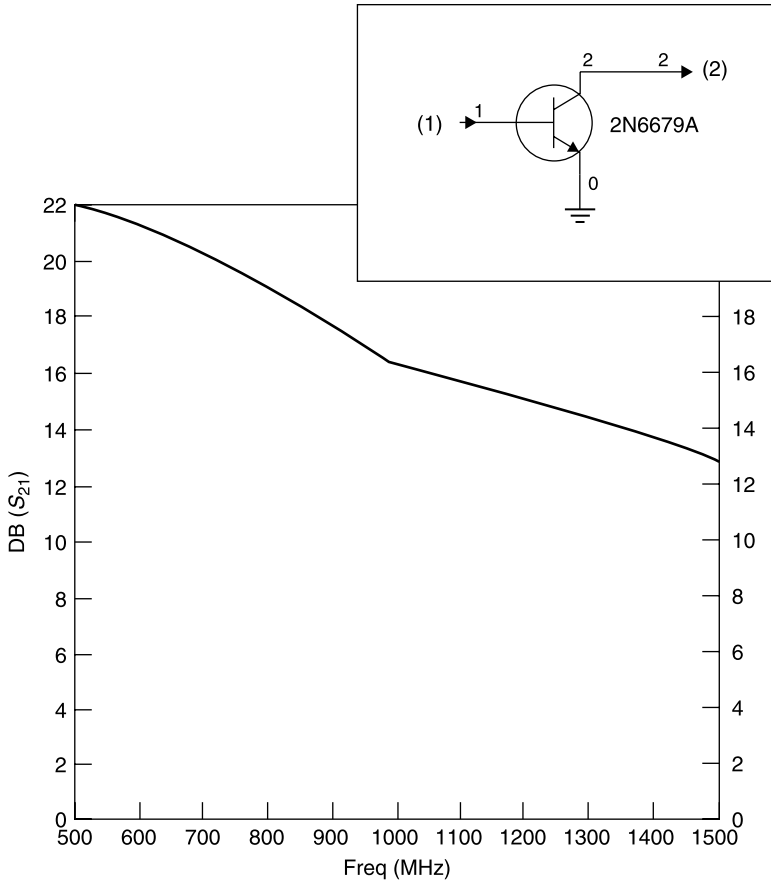
Figure 10.1-4 Amplifier consisting of a biased Motorola 2N6679A silicon bipolar transistor with 50-Ω source and load terminations.

be seen that a good amount of power is lost to reflection at the input and mismatch at the output. At 1 GHz

$$|S_{11}| = 0.68$$

$$\text{Input mismatch loss} = 1 - |S_{11}|^2 = 1 - 0.68^2 = 0.54 \qquad (10.1\text{-}5)$$

$$= 10 \log(1 - |S_{11}|^2) = -2.7 \text{ dB}$$

$$|S_{22}| = 0.46$$

$$\text{Output mismatch loss} = 1 - |S_{22}|^2 = 1 - 0.46^2 = 0.79 \qquad (10.1\text{-}6)$$

$$= 10 \log(1 - |S_{22}|^2) = -1.0 \text{ dB}$$

These mismatch losses correspond to 46% power loss at the input and 21% at the output. Therefore, by matching input and output, we could have up to 3.7 dB of additional gain and have a matched circuit at input and output as well.

10.2 AMPLIFIER STABILITY

Thus far, even if we are satisfied with the performance obtained by installing a transistor in the same impedances in which it was measured, there is a point that we have not considered. We have treated the transistor as if it were a *unilateral device*. Put in other words, we have assumed that signals pass from the input to the output, but not in the reverse direction.

Making the *unilateral assumption is equivalent to assuming that* S_{12} *is zero*. The S_{12} parameter provides the feedback term by which power from the output circuit (which is relatively high due to the transistor's amplification) can feed back to the input. When it does so, it may combine with reflections already present at the input to produce an effective S_{11} whose magnitude exceeds unity.

This corresponds to *reflection gain*, and a transistor amplifier that can experience this gain, is termed *conditionally unstable*, the condition being certain combination(s) of load impedance, S_{12} and S_{11}, that could produce self-oscillation (instability).

In the current 2N6679A example, the magnitudes of S_{11} and S_{22} of the transistor are less than one for all frequencies. This means that the transistor is *stable when embedded between 50-Ω source and load* (Fig. 10.2-1). That is, it will not oscillate in the 50-Ω input/output environment. However, this would not be generally considered sufficient within the industry. Rather, the expectation is that *a properly designed (stabilized) amplifier will not oscillate no matter*
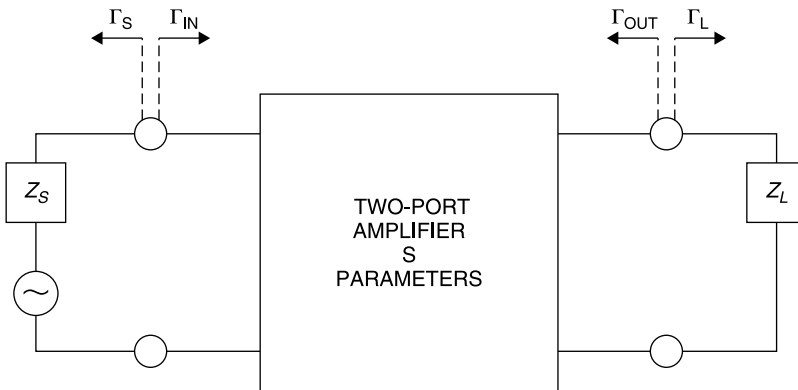
Figure 10.2-1   An *S* parameter represented network embedded between source and load of general complex reflection coefficients.

what passive source and load impedances are presented to it, including short or open circuits of any phase.

Because the S_{12} of practical transistors is not zero, a signal path exists from the output (where power levels are higher due to the device's gain) to the input. It is possible that for certain values of load Γ_L, Γ_{IN} can exceed unity, turning the circuit into a reflection amplifier at the input. Similarly, certain values of Γ_S at the input might cause Γ_{OUT} to exceed unity magnitude. When either or both of these conditions can occur, at any frequency, the circuit is said to be only *conditionally stable* or equivalently, *potentially unstable.*

The boundary line between stable and unstable operation at the input is the $|\Gamma_{IN}| = 1$ circle centered at the origin of the Smith chart. The contour of values for Γ_L (also a circle) that produces a unity magnitude reflection coefficient at the input is called the *input stability circle*. This is shown in Figure 10.2-2 for the 2N6679A transistor. Keep in mind that *the input stability circle represents a set of load reflection coefficients* that cause the input reflection coefficient to have a unity magnitude. Also note that the *input stability circle* represents a *boundary*, on one side of which $|\Gamma_{IN}| < 1$ resulting in load impedances that

1000 MHz: 1.972, 39.022°
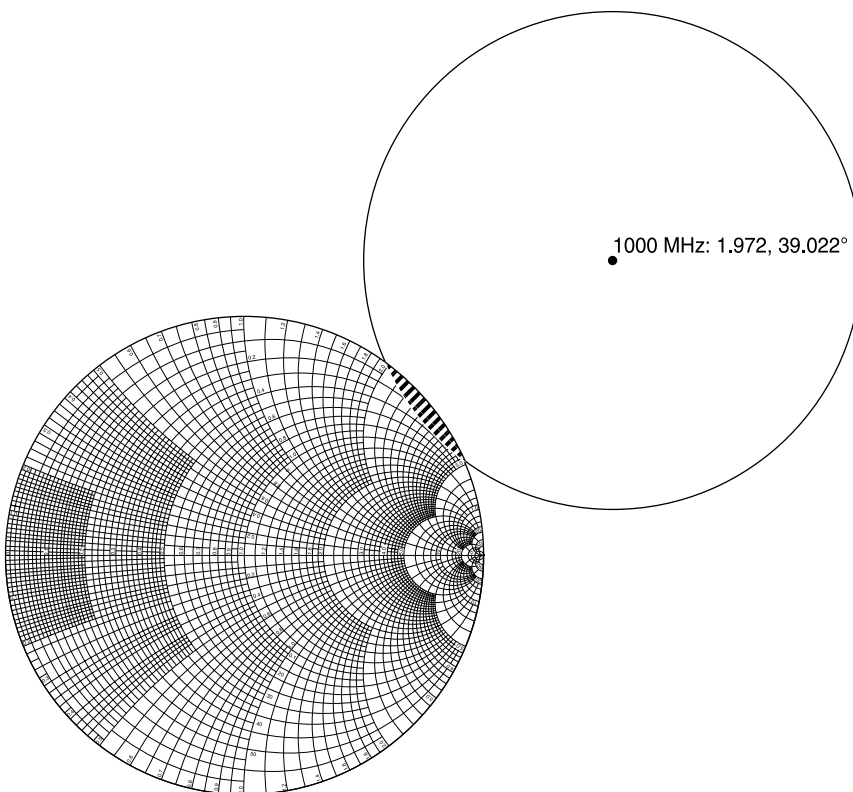
Figure 10.2-2 Load instability circle for the Motorola 2N6679A transistor at 1 GHz.

result in stable operation and on the other side of which $|\Gamma_{IN}| > 1$, corresponding to load impedances that produce potential instability.

The Γ_L values are related to the Γ_{IN} values according to (10.2-1), called a mapping function because it maps a Γ_{IN} contour into a Γ_L contour [2, pp. 140–145]. This function maps the $|\Gamma_{IN}| = 1$ into a Γ_L contour, in this case a circle, to a circle:

$$\Gamma_L = \frac{\Gamma_{IN} - S_{11}}{S_{22}\Gamma_{IN} - \Delta} \quad \text{where } \Delta = S_{11}S_{22} - S_{12}S_{21} \qquad (10.2\text{-}1)$$

Inserting the condition $|\Gamma_{IN}| = 1$ into the above equation creates a mapping of the $|\Gamma_{IN}| = 1$ circle into the input stability circle in the Γ_L plane with center at

$$C_L = \frac{S_{22}^* - \Delta^* S_{11}}{|S_{22}|^2 - |\Delta|^2} \qquad (10.2\text{-}2)$$

and a radius of

$$R_L = \left| \frac{S_{12}S_{21}}{|S_{22}|^2 - |\Delta|^2} \right| \qquad (10.2\text{-}3)$$

The input stability circle for the 2N6679A untuned amplifier at 1 GHz, is shown in Figure 10.2-2.

The *input stability circle* is also called the *load instability circle* because it describes the Γ_L values that border on causing instability, that is, which result in $|\Gamma_{IN}| = 1$. *Henceforth, we will refer to the input stability circle as the load instability circle* because this places the focus on the load impedances that cause instability.

The circle intersects and intrudes into the unity radius area of the Smith chart. This means that the transistor is *potentially unstable* for certain passive load impedances bounded by the circle. Note, again, that the circle is the *boundary* of the unstable impedances. Since it is only the boundary, when we first plot the circle we do not know whether the unstable load impedances are those inside the circle or outside of the circle. Of course, in the example in Figure 10.2-2 it would seem highly likely that the unstable points are inside the circle since most of the circle is in the negative resistance domain of the Smith chart. However, there are instances in which the entire circle may lie within the unity radius, passive impedance portion of the Smith chart. Therefore, some means for determining which side of the boundary represents the unstable impedances is needed. The determination is facilitated by the availability of the S parameters, which apply with a known load, usually 50 Ω. *To determine whether the unstable load impedances are inside or outside the circle, consider the "easy point"* S_{11}.

We know from the S parameter file for the 2N5579A transistor that when the transistor is loaded with 50 Ω, the center of the Smith chart, $|S_{11}| < 1$.

Therefore, the center of the Smith chart is a stable load point. Since the center of the Smith chart is outside the load instability circle, it follows that in this case all points outside the circle are stable, and that the points inside the circle in Figure 10.2-2 are the unstable loads.

From Figure 10.2-2 it can be seen that high inductive load impedances at the output can cause instability (at the input). Also, notice that points outside the $\Gamma = 1$ circle correspond to loads with negative resistance. These do not concern us because any amplifier can be made unstable by subjecting it to an appropriate source or load impedance with a negative real part. *Amplifier stability only requires that the circuit be stable with passive input and output reflection coefficients, that is, passive source and load impedances.*

Next consider the source reflection coefficients that cause $|\Gamma_{OUT}| = 1$. The relation of Γ_S to Γ_{OUT} is given by [2, p. 142]

$$\Gamma_S = \frac{S_{22} - \Delta\Gamma_{OUT}}{1 - S_{11}\Gamma_{OUT}} \tag{10.2-4}$$

In a similar fashion, this can be used to map the $|\Gamma_{OUT}| = 1$ circle onto the Γ_S plane. This maps as a circle with center at

$$C_S = \frac{S_{11}^* - \Delta^* S_{22}}{|S_{11}|^2 - |\Delta|^2} \tag{10.2-5}$$

and radius

$$R_S = \left| \frac{S_{12}S_{21}}{|S_{11}|^2 - |\Delta|^2} \right| \tag{10.2-6}$$

Plotting this for our transistor at 1 GHz gives the left circle outside the Smith chart in Figure 10.2-3.

The *output stability circle* is also called the *source instability circle* because it represents Γ_S values that cause instability in the input circuit. *Henceforth we will call it the source instability circle*. Once again, we know that the points within the output stability circle are the unstable ones because, from the S parameters of the transistor, $|S_{22}| < 1$ with a 50-Ω termination on the input. Notice that the output circuit may become unstable ($|\Gamma_{OUT}| > 1$) when the input sees low source impedances.

What do these circles mean? If we attempt to get more gain from the transistor by adding matching circuits to the input and output, they must not present those impedances which would cause instability. In fact, even if we accept the transistor as it is and specify that only 50-Ω loads should be applied, if the source or load is removed and there is a transmission line whose length transforms an open circuit to one of the unstable impedances, the circuit may oscillate. In other words, the circuit will be *potentially unstable*.
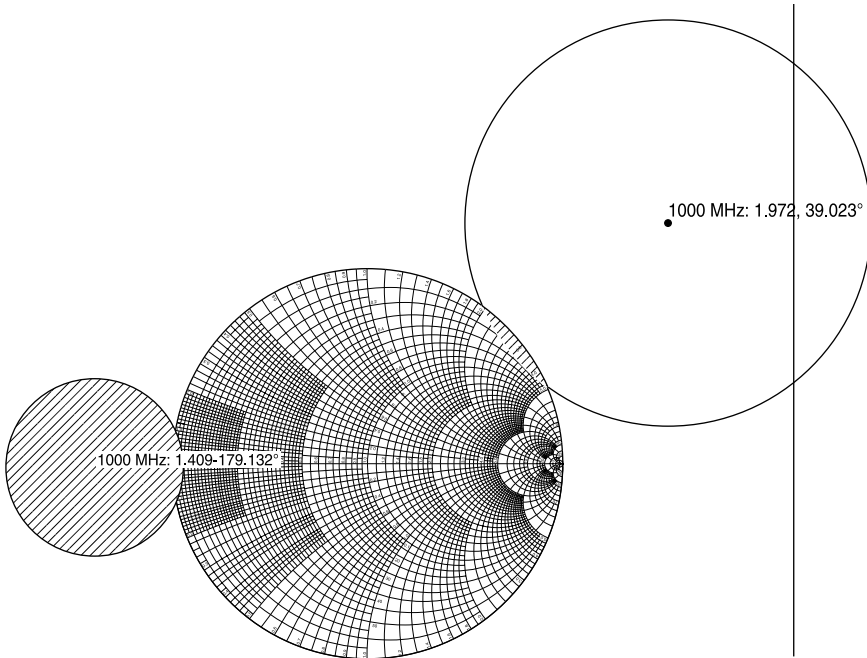
Figure 10.2-3 Source instability circle (at left) and load instability circle (at right) for the 2N6679A transistor at 1 GHz.

Thus far, in our examples of stability circles, we have only examined the transistor's behavior at 1 GHz. But, for an amplifier to be unconditionally stable, it must be so at all frequencies. To assure this requires that it be found to be unconditionally stable at all frequencies for which the transistor has more than unity gain. The resulting *source and load instability circles are two separate families of circles* (Fig. 10.2-4).

10.3   *K* FACTOR

Since the stability circles can be calculated directly from the *S* parameters, it would seem possible that stability could be determined from the *S* parameters themselves, without need for the elaborate graphical construction of the stability circles, and this is so. We define the *stability factor K* as [2, p. 145]

$$K = \frac{1 - |S_{11}|^2 - |S_{22}|^2 + |\Delta|^2}{2|S_{12}S_{21}|} \tag{10.3-1}$$

where
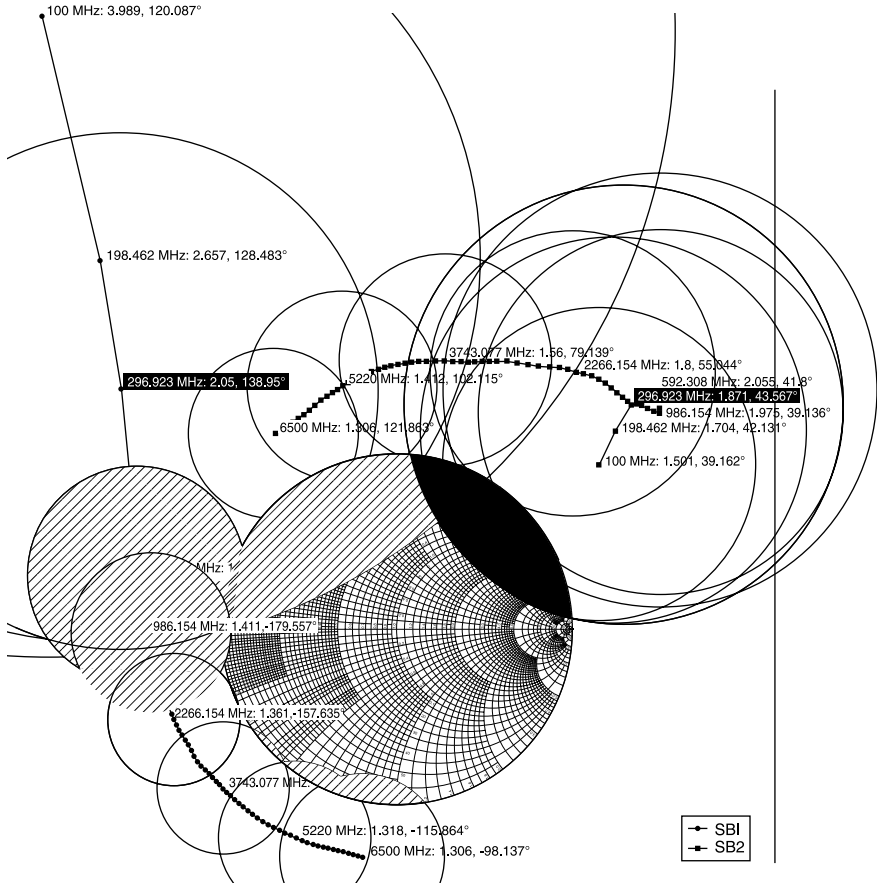
Figure 10.2-4   Source and load instability circles for a band of frequencies from 100 to 6500 MHz.

$$|\Delta|^2 = |S_{11}S_{22} - S_{12}S_{21}|^2 \qquad (10.3\text{-}2)$$

The amplifier is unconditionally stable provided that

$$K > 1 \quad \text{and} \quad |\Delta|^2 < 1 \qquad (10.3\text{-}3a,b)$$

Equivalently [3, p. 324], the amplifier is unconditionally stable if

$$K > 1 \quad \text{and} \quad B_1 > 0 \qquad (10.3\text{-}4a,b)$$

where

$$B_1 = 1 + |S_{11}|^2 - |S_{22}|^2 - |\Delta|^2 > 0 \qquad (10.3\text{-}5)$$

TABLE 10.3-1   Stability Criteria for 2N6679A Transistor

Freq (MHz)	K	B_1
0	0.195	0.203
500	0.684	1.153
1000	0.875	1.23
1500	0.964	1.219
2000	0.962	1.231
2500	0.972	1.205
3000	1.755	1.119
3500	0.913	1.117
4000	0.863	1.086
4500	0.842	1.051
5000	0.663	1.009
5500	0.611	0.975
6000	0.526	0.902
6500	0.498	0.879

Applying this test to the 2N6679A transistor and using the Genesys program to calculate K and B_1 gives the results in Table 10.3-1.

Thus we see that the transistor is potentially unstable at 1 GHz, as had already been deduced from the input and output stability circles. It is also potentially unstable at other frequencies. However, the source and load instability circles provide insight into what can be done to make the transistor unconditionally stable. Notice that the source instability circles indicate that low impedances at the input cause instability. This suggests that a resistance placed in series with the base lead can reduce the intersection of the source instability circles with the $|\Gamma| \leq 1$ circle. This will reduce gain. However, since device instability usually peaks at low frequencies (at which the transistor has the highest gain), the resistor can be shunted with a capacitor to minimize its effect near 1 GHz. One such stabilizing circuit is shown in Figure 10.3-1.

This measure almost removed all instabilities. It moved both the source and load instability circles away from the $|\Gamma| = 1$ (passive Smith chart) portion of the reflection coefficient plane (Fig. 10.3-2). Rather than add more resistance to the base lead, it is often found to be more effective to add some damping to the output. Corrections to the input stability tend to help the output stability, and vice versa.

Examining the load instability circles reveals that high output load impedances cause instability. To counter this, add shunt resistance at the output, but place it at the end of a 90° line at 1 GHz to minimize its effect at 1 GHz. The addition of shunt resistance limits how large the load impedance can be, minimizing the likelihood that a high impedance that can cause instability will be encountered at the output.
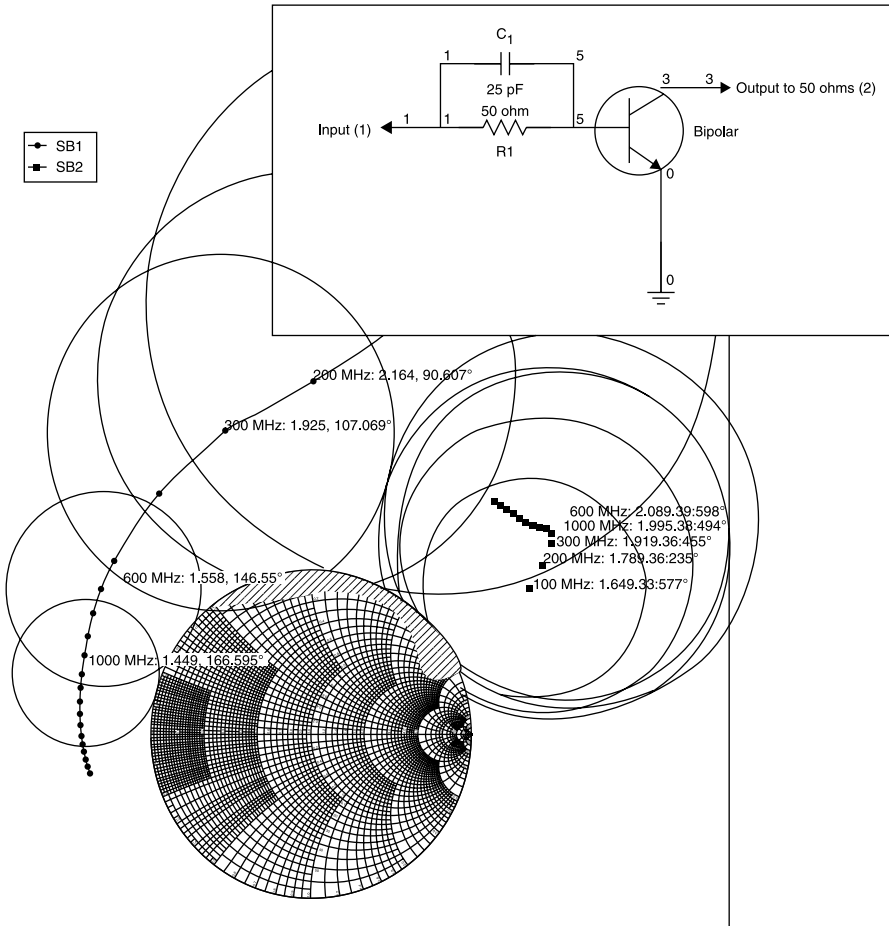
Figure 10.3-1   Revised instability circles when series resistance is placed in the base lead of the 2N6679A transistor.

This addition to the output circuit has unconditionally stabilized the device over the 100- to 2000-MHz range, in which it previously was potentially unstable (Table 10.3-2). From the K factor calculations it can be seen that the low frequency end is especially well stabilized. This is important because S parameter data below 100 MHz were not included in the S parameter file.

The resulting gain of the 2N6679A with the stabilizing elements is shown in Figure 10.3-3. Notice that, in comparing this gain with that of the unstabilized transistor in Figure 10.1-4, we have sacrificed only about 0.6 dB of gain at 1 GHz, dropping from 16.4 dB to about 15.8 dB at this frequency.
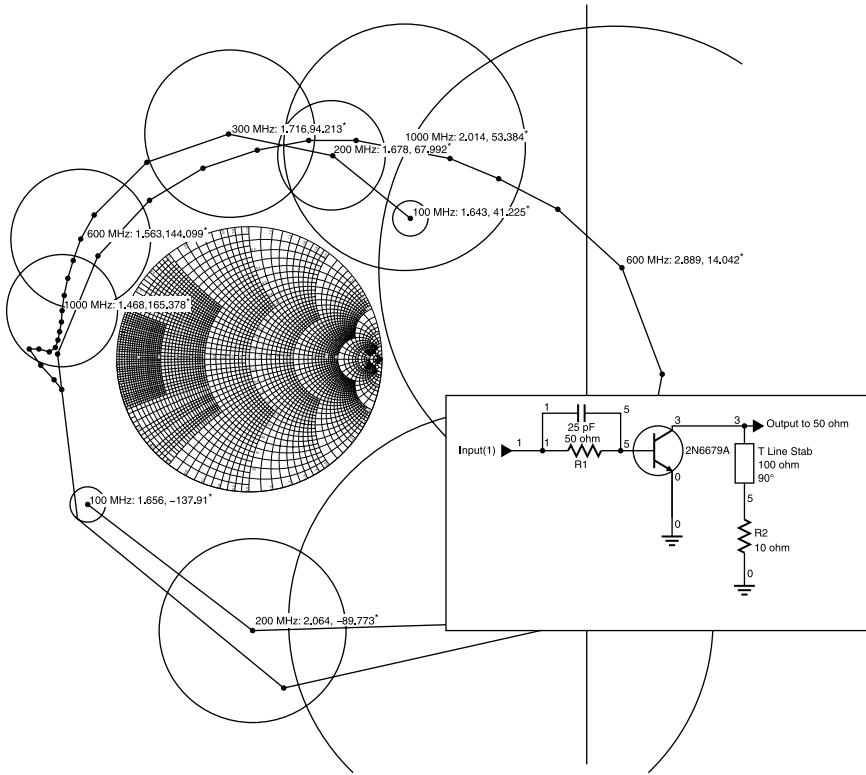
Figure 10.3-2 Source and load instability circles for the 2N6679A with stabilizing circuits in the input and outputs.

We have reduced the low frequency gain even more, but this is of no concern. In fact, it is a benefit. *Gain outside the desired bandwidth is a liability because it can cause instabilities*. We are now in a position to match the input and output of the transistor to realize additional gain.

10.4 TRANSDUCER GAIN

We have seen that the S parameters are a valuable aid both for collecting data for a transistor and then using the data to predict performance and design an amplifier circuit. Unfortunately, unlike Z, Y, or $ABCD$ parameters, *the values of S parameters depend not only upon the properties of the transistor but also upon the source and load circuits used to measure them*. This is because they measure transmitted and reflected waves, and these depend upon both the transistor and the source and load used to test it.

TABLE 10.3-2   Calculated *K* Factor for Stabilized 2N6679A Transistor

	Freq (MHz)	K	B_1
1	100	6.093	0.81
2	200	2.008	0.875
3	300	1.237	0.965
4	400	1.074	1.076
5	500	1.221	1.186
6	600	1.081	1.216
7	700	1.02	1.23
8	800	1.015	1.232
9	900	1.059	1.224
10	1000	1.165	1.204
11	1100	1.173	1.162
12	1200	1.222	1.105
13	1300	1.34	1.031
14	1400	1.608	0.934
15	1500	2.339	0.817
16	1600	5.172	0.709
17	1700	18.87	0.712
18	1800	5.059	0.925
19	1900	2.247	1.184
20	2000	1.608	1.328

Note that *the concepts of a reflection coefficient and traveling waves can be used even if there are no actual transmission lines at the device ports* (Fig. 10.4-1). One might expect that the input reflection coefficient, Γ_{IN}, would simply be equal to S_{11}, and that $\Gamma_{OUT} = S_{22}$. However, because of feedback these must be corrected [1, p. 214]:

$$\Gamma_{IN} = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L} \quad \text{and} \quad \Gamma_{OUT} = S_{22} + \frac{S_{12}S_{21}\Gamma_S}{1 - S_{11}\Gamma_S} \quad \text{(10.4-1a,b)}$$

The general *transducer gain* of a two-port network having S_{21} and S_{12} values, whether it is a transistor or not, is

$$G_T = \frac{\text{power delivered to load}}{\text{power available from the source}} \quad \text{(10.4-2)}$$

$$G_T = \frac{P_{LOAD}}{P_{AVAIL}} = \frac{1 - |\Gamma_S|^2}{|1 - \Gamma_{IN}\Gamma_S|^2} |S_{21}|^2 \frac{1 - |\Gamma_L|^2}{|1 - S_{22}\Gamma_L|^2}$$

$$= \frac{1 - |\Gamma_S|^2}{|1 - S_{11}\Gamma_S|^2} |S_{21}|^2 \frac{1 - |\Gamma_L|^2}{|1 - \Gamma_{OUT}\Gamma_L|^2} \quad \text{(10.4-3a,b)}$$
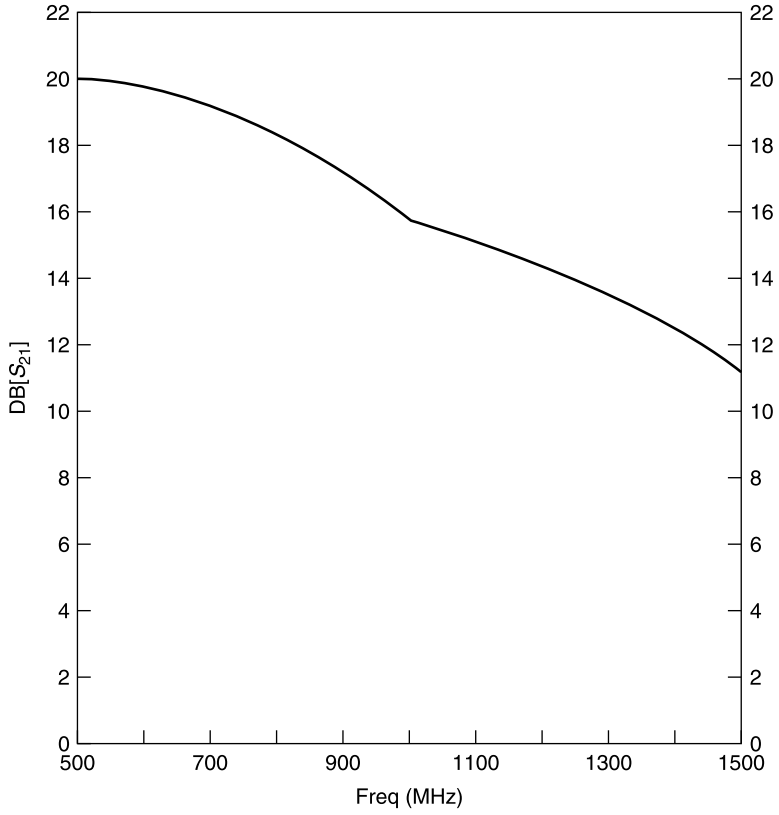
Figure 10.3-3   Gain of stabilized 2N6679A transistor.
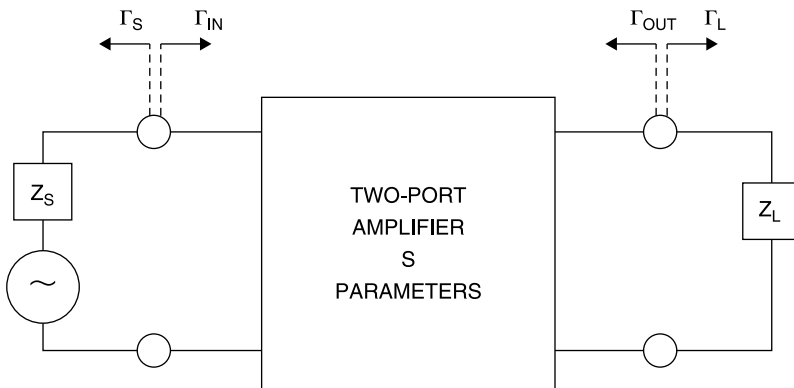
Figure 10.4-1   Diagram of input and output reflection coefficients.
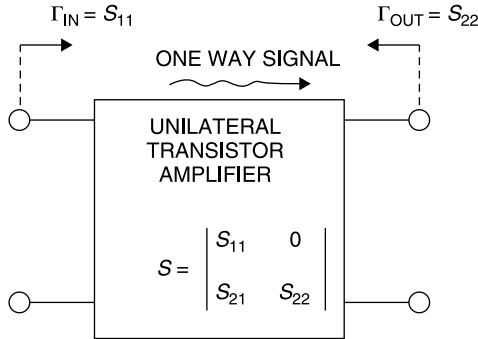
Figure 10.5-1 Input and output reflection coefficients for the unilateral gain assumption.

10.5 UNILATERAL GAIN DESIGN

The *transducer gain* expressions of (10.4-3) are too complex for manual design. To effect an approximate gain solution, let us ignore the feedback, that is, assume that $S_{12} = 0$. If an amplifier has no feedback, signals pass one way through it. Accordingly, this is called the *unilateral gain* assumption (Fig. 10.5-1).

If $S_{12} = 0$, then $\Gamma_{IN} = S_{11}$ and $\Gamma_{OUT} = S_{22}$. Furthermore, if $S_{12} = 0$ the transducer gain becomes [1, p. 228]

$$G_{TU} = \frac{1 - |\Gamma_S|^2}{|1 - S_{11}\Gamma_S|^2} |S_{21}|^2 \frac{1 - |\Gamma_L|^2}{|1 - S_{22}\Gamma_L|^2} \qquad (10.5\text{-}1)$$

This can be considered as three separate gain factors (Fig. 10.5-2):

$$G_{TU} = G_S G_0 G_L \qquad (10.5\text{-}2)$$

in which

$$G_S = \frac{1 - |\Gamma_S|^2}{|1 - S_{11}\Gamma_S|^2} \qquad G_0 = |S_{21}|^2 \qquad G_L = \frac{1 - |\Gamma_L|^2}{|1 - S_{22}\Gamma_L|^2} \qquad (10.5\text{-}3\text{a,b,c})$$
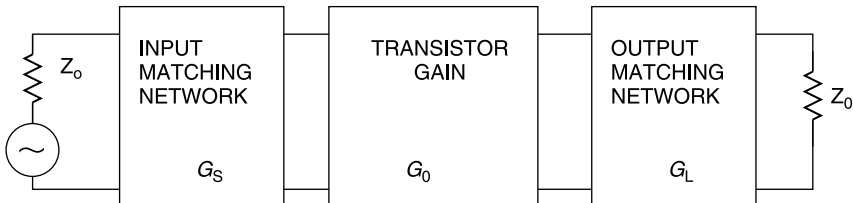
Figure 10.5-2 Three-stage, unilateral gain design model for which $S_{12} = 0$.

The unilateral gain expressions of (10.5-3) apply for any Γ_S and Γ_L. To maximize G_S, we select

$$\Gamma_S = S_{11}^* \tag{10.5-4}$$

Then

$$G_{S\text{-MAX}} = \frac{1 - |\Gamma_S|^2}{|1 - S_{11}\Gamma_S|^2} = \frac{1 - |S_{11}|^2}{(1 - |S_{11}|^2)^2} = \frac{1}{1 - |S_{11}|^2} \tag{10.5-5}$$

Similarly, to maximize G_L, we select

$$\Gamma_L = S_{22}^* \tag{10.5-6}$$

and then

$$G_{L\text{-MAX}} = \frac{1}{1 - |S_{22}|^2} \tag{10.5-7}$$

Therefore, under the unilateral assumption, $S_{12} = 0$, the maximum gain to be obtained from a transistor is

$$G_{\text{TU-MAX}} = \frac{1}{1 - |S_{11}|^2} |S_{21}|^2 \frac{1}{1 - |S_{22}|^2} \tag{10.5-8}$$

The overall gain (in decibels) to be obtained is then

$$G_{\text{TU}}(\text{dB}) = G_S(\text{dB}) + G_0(\text{dB}) + G_L(\text{dB}) \tag{10.5-9}$$

in which we recognize that G_S and G_L are the gains (or losses) to be obtained by matching (or deliberately further mismatching) the input and output circuits, respectively. Of course, there is an error in the gain calculations of (10.5-8) and (10.5-9) if, in the actual transistor, $S_{12} \neq 0$. In that case the true gain G_T is related to the calculated unilateral gain G_{TU} by [1, p. 239]

$$\frac{1}{(1 + U)^2} < \frac{G_T}{G_{\text{TU}}} < \frac{1}{(1 - U)^2} \tag{10.5-10}$$

where

$$U = \frac{|S_{11}||S_{21}||S_{12}||S_{22}|}{(1 - |S_{11}|^2)(1 - |S_{22}|^2)} \tag{10.5-11}$$

The value of U varies with frequency because of its dependence on the S parameters, and it is called the *unilateral figure of merit*. For the 2N6679A,

applying the S parameter values at 1 GHz from Table 10.1-1, U is calculated as

$$U = \frac{(0.68)(6.6)(0.03)(0.46)}{(1 - 0.68^2)(1 - 0.46^2)} = 0.146 \tag{10.5-12}$$

$$\frac{1}{(1 - U)^2} = \frac{1}{(1 - 0.146)^2} = 1.37 = +1.37 \text{ dB} \tag{10.5-13}$$

$$\frac{1}{(1 + U)^2} = \frac{1}{(1 + 0.146)^2} = 0.76 = -1.18 \text{ dB} \tag{10.5-14}$$

From this it can be seen that the unilateral gain approximation can be used for the 2N6679A at 1 GHz with an error somewhat greater than ± 1 dB. To obtain the "maximum gain" using the unilateral gain design, we transform the 50-Ω source to $Z_S = Z_{IN}^*$, and the 50-Ω load is transformed to $Z_L = Z_{OUT}^*$. From the S parameters,

$$z_{IN} = \frac{1 + S_{11}}{1 - S_{11}} \quad \text{and} \quad z_{OUT} = \frac{1 + S_{22}}{1 - S_{22}} \tag{10.5-15}$$

To unnormalize,

$$Z_{IN} = z_{IN}Z_0 \quad \text{and} \quad Z_{OUT} = z_{OUT}Z_0 \tag{10.5-16}$$

Since we have modified the S parameters by adding the stabilizing components, we must use revised S parameters for the stabilized 2N6679A transistor. This would be a laborious task but is easy and straightforward using the network simulator software. The revised S parameters are shown in Table 10.5-1. The corresponding input and output impedances are shown in Table 10.5-2.

There are numerous methods by which the amplifier can be matched. As an example, the Q matching method can be employed (Fig. 10.5-3). Since the input impedance at 1 GHz is

$$Z_{IN} = (10.43 - j7.238) \ \Omega \tag{10.5-17}$$

it is necessary to transform the 50-Ω source to Z_{IN}^*. Thus,

TABLE 10.5-1   Revised S Parameters at 1 GHz for Stabilized 2N6679A Transistor Amplifier

	Magnitude	Angle (deg)
S_{11}	0.661	-162.8
S_{21}	6.25	83.23
S_{12}	0.028	59.23
S_{22}	0.414	-31.86

TABLE 10.5-2   Input and Output Impedances for Stabilized 2N6679A Amplifier

Z_{IN}	Z_{OUT}
$(10.43 - j7.238)\ \Omega$	$(88.493 - j46.646)\ \Omega$

$$Z_S = (10.43 + j7.238)\ \Omega \qquad (10.5\text{-}18)$$

We begin by transforming the 50-Ω source to 10.43 Ω:

$$\frac{50}{10.43} = 4.794 = 1 + Q^2$$
$$Q = 1.948 \qquad (10.5\text{-}19)$$

Since the transformation from 50 Ω is to a lower resistance, we begin with a shunt reactance in parallel with the 50 Ω.

We have selected a shunt inductor rather than a capacitor because the final transformed value for Z_S has an inductive part. For a Q of 1.948, the reactance of L_1 is +25.667 Ω. This can be provided using an L_1 given by

$$L_1 = \frac{25.667\ \Omega}{6.28\ \Omega/\text{nH}} = 4.09\ \text{nH} \qquad (10.5\text{-}20)$$

The series equivalent circuit on the right has the required 10.43-Ω real part and, because the circuit Q is unchanged, a $+j20.32$-Ω reactive part. However, we require a $+7.238$-Ω reactance, hence we must tune out part of this reactance using a series capacitor C_2. The reactance magnitude of C_2 is $20.32 - 7.24 = 13.08\ \Omega$. This is provided by a capacitor C_2 with the value

$$C_2 = \frac{159}{13.08} = 12.16\ \text{pF} \qquad (10.5\text{-}21)$$

The resulting circuit and performance are shown in Figure 10.5-4.

Notice that with the input tuning the gain, S_{21}, is about 18.4 dB at 1 GHz, compared with about 15.9 dB for the stabilized transistor alone, a gain im-
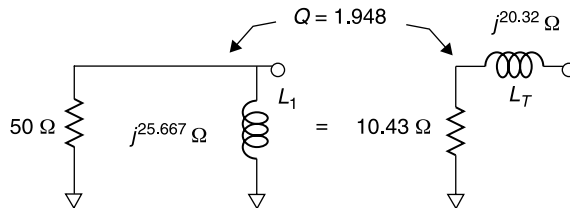
Figure 10.5-3 Transforming the 50-Ω source to 10.43 Ω using the Q matching method.
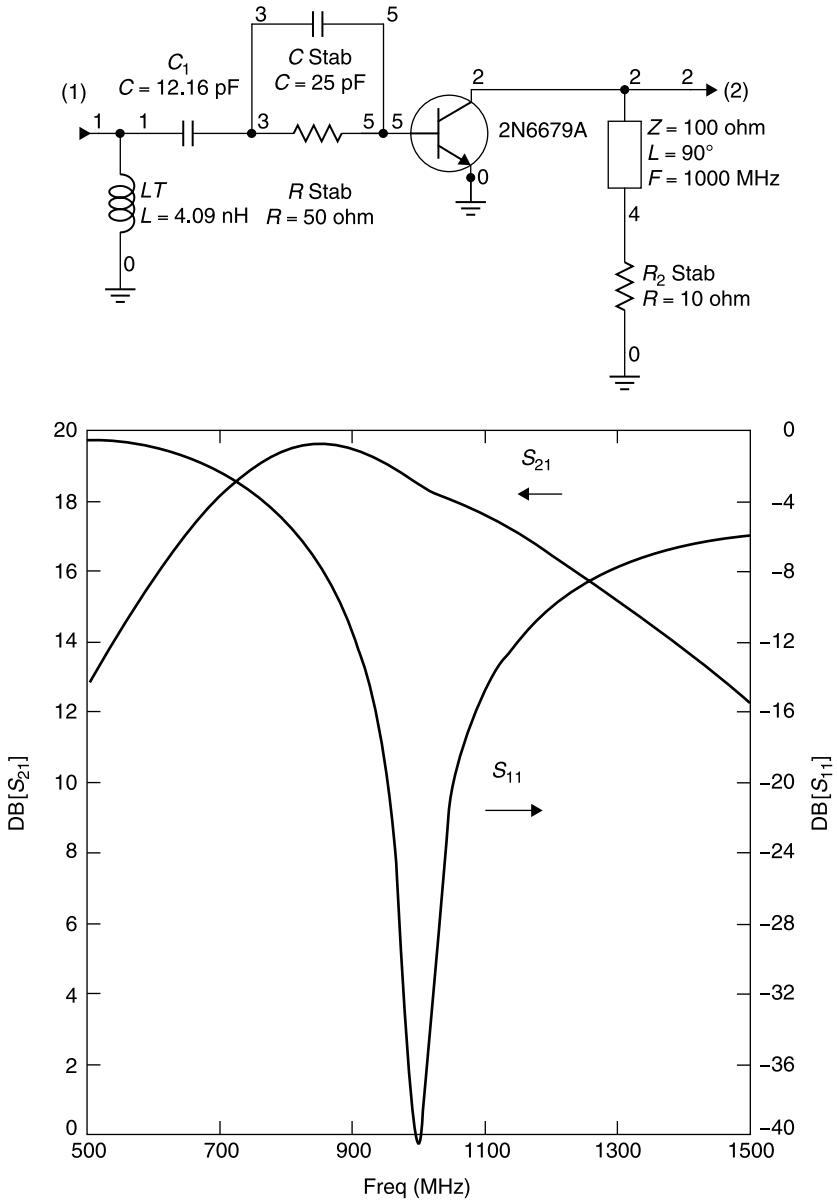
Figure 10.5-4 S_{21} and S_{11} of the stabilized 2N6679A with unilaterally matched input.

provement of 2.5 dB. This is consistent with the result to be expected in tuning the 2.5-dB mismatch loss of the stabilized transistor, having an S_{11} magnitude of 0.661 (Table 10.5-1). Keep in mind that this S_{21} is the gain of the overall two-port network in Figure 10.5-4.

Also from Table 10.5-1, $|S_{22}| = 0.414$. This means that 17% of the power is

being lost due to the output mismatch. If this were recovered, the gain could increase by another 0.8 dB. To tune the output port, the output impedance at 1 GHz is seen from Table 10.5-2 to be $(88.493 - j46.646)$ Ω.

The 50-Ω load is to be transformed to the complex conjugate of this value or

$$Z_L = (88.493 + j46.646) \ \Omega \tag{10.5-22}$$

Using the Q matching method, and starting at the load, we first transform 50 to 88.493 Ω for which, arbitrarily, a series inductance L_2 is chosen (Fig. 10.5-5). The resistive transformation ratio is 88.493/50 for which the required Q is 0.877 and L_2 is 6.99 nH. The parallel equivalent circuit consists of the desired 88.493 Ω resistance shunted by a parallel inductive reactance of $88.493/0.877 =$
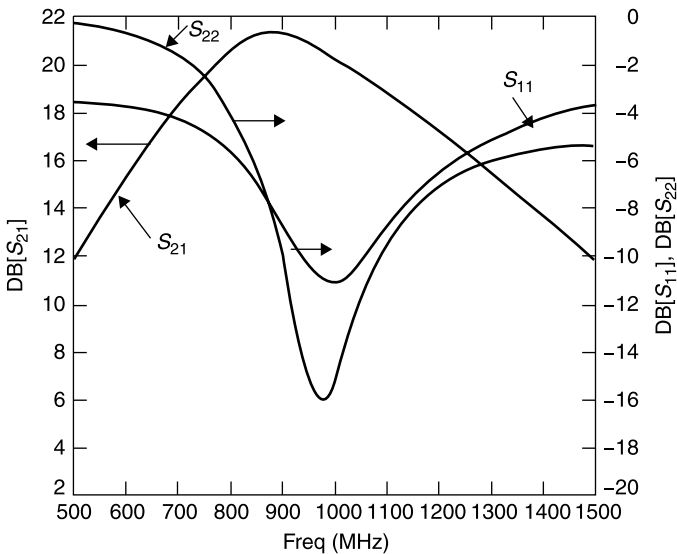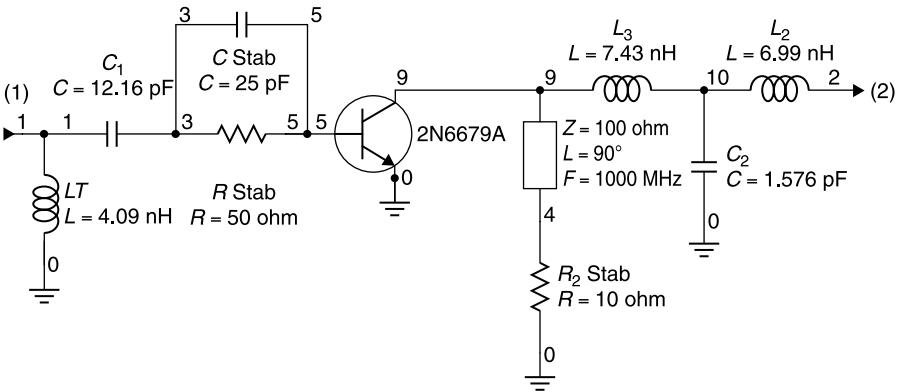
Figure 10.5-5   The 2N6679A transistor stabilized and unilaterally input and output tuned.

100.86 Ω. This is parallel resonated at 1 GHz by a capacitance C_2 of 1.576 pF. An additional reactive impedance of $+j46.646$ Ω is required to transform the 50-Ω load to the Z_L given in (10.5-22), and this is achieved with a series inductance L_3 of 7.43 nH. The stabilized 2N6977A with unilaterally tuned input and output is shown in Figure 10.5-5 along with its performance.

The gain when both input and output are matched is 20 dB at 1 GHz. This is within 1 dB of the 19.2-dB maximum gain expected. Recall that the unilateral figure of merit analysis indicated that the error in gain estimate could be between -1.18 and $+1.37$ dB. The 20-dB gain is an increase over the input matched case (Fig. 10.5-4) of 1.6 dB. This is more than the 0.8 dB expected improvement. Also, both the input and output are imperfectly matched as can be seen from the plots of S_{11} and S_{22} in Figure 10.5-5. These inaccuracies are to be expected in the unilateral gain method, when the transistor's internal feedback S_{12} is ignored.

10.6   UNILATERAL GAIN CIRCLES

Input Gain Circles

Sometimes we do not wish to obtain all of the gain available. Instead, we may wish a lower gain level or we may wish to know how much gain can be obtained with a given impedance source or load. For these applications, constant *unilateral gain circles* are useful. We have already seen that the unilateral gain can be represented as three gain factors as expressed in (10.5-3a,b,c).

Previously, we were concerned with matching the input and/or output totally to realize fully G_S and/or G_L. But what if we want to obtain only a partial amount of G_S or G_L or even to detune one of them to obtain a lower value of gain? This is quite practical. We simply do not quite match either the input or the output, or both, as perfectly as we might.

This process is described by impedance circles on the Smith chart for input and output loads that produce the same gain at various decibel levels below the maximum obtainable gain. When the input gain circles are plotted on the Smith chart [1, pp. 230–231], they have centers given by

$$C_{gs} = \frac{G_S S_{11}^*}{1 - |S_{11}|^2 (1 - G_s)} \tag{10.6-1}$$

with corresponding radii

$$r_{gs} = \frac{\sqrt{1 - G_s}(1 - |S_{11}|^2)}{1 - |S_{11}|^2 (1 - G_s)} \tag{10.6-2}$$

in which the gain values (G_S) are numeric power ratios (not decibels).
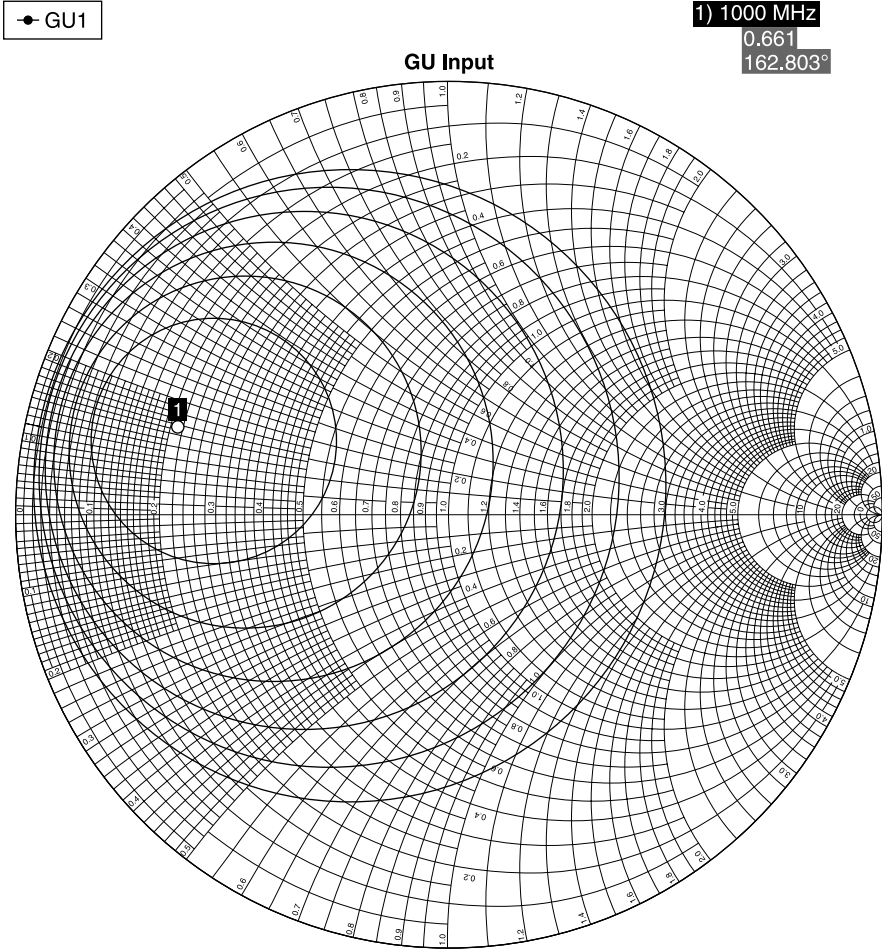
Figure 10.6-1 Circles of z_s impedance values (normalized to 50 Ω) that provide constant gain. Center point ($z_s = 0.21 + j0.14$) corresponds to tuning input and adds 2.5-dB unilateral gain to the untuned but stabilized 2N6679 transistor. Remaining circles provide 1, 2, 3, 4, 5, and 6 dB less gain, respectively.

Consider the input gain circles described by (10.6-1) and (10.6-2). For each value of gain G_S, including negative decibel values (corresponding to less gain than is achieved with 50-Ω source and load), a circle can be plotted on the Smith chart (Fig. 10.6-1) with center at C_{gs} and radius r_{gs}. The circle having zero radius corresponds to matching the input and, as can be seen, this point corresponds to $z_s = 0.21 + j0.14$, the value of Z_S given in Table 10.5-2 when normalized to 50 Ω. Given that the magnitude of S_{11} for the stabilized 2N6679A is 0.661 (Table 10.5-1), the maximum gain addition, G_S, is 2.5 dB.

The circles drawn about this point in Figure 10.6-1 are for lesser gains by 1, 2, 3, 4, 5, and 6 dB, respectively.

Notice that applying a source impedance of 50 Ω results in a reduction from the maximum G_S gain of -2.5 dB, as would be expected. The circles corresponding to 3, 4, 5, and 6 dB less gain correspond to the application of z_S impedances that "detune" the transistor from the gain it would have were a 50-Ω source impedance used. The family of gain circles in Figure 10.6-1 are also useful as a means of estimating the change of gain to be expected should a source impedance other than that intended be applied.

Output Gain Circles

Similarly, at the output, the constant gain circles have centers at

$$C_{gL} = \frac{G_L S_{22}^*}{1 - |S_{22}|^2 (1 - G_L)} \qquad (10.6\text{-}3)$$

with radii

$$r_{gL} = \frac{\sqrt{1 - G_L}(1 - |S_{22}|^2)}{1 - |S_{22}|^2 (1 - G_L)} \qquad (10.6\text{-}4)$$

in which the gain values (G_L) are numeric power ratios (not decibels). When these circles are plotted about the normalized load impedance z_L that provides best tuning of the output of the stabilized 2N6679A, the results are as shown in Figure 10.6-2. Notice that the -1-dB gain circle includes the origin of the Smith chart. In other words, that there is less than 1 dB gain difference between placing a match at the output versus using a 50-Ω load. This is expected. The magnitude of S_{22} at 1 GHz is 0.414, which corresponds to a mismatch loss of only 0.8 dB.

Figures (10.6-1) and (10.6-2) permit the ready determination that the unilateral gain improvement to be expected from tuning both the input and output is 2.5 dB + 0.8 dB = 3.2 dB. Of course, the caution applies that the total unilateral gain estimates for this transistor, from (10.5-13) and (10.5-14) at 1 GHz have a tolerance of -1.18 and $+1.37$ dB, respectively.

An advantage of designing to less than the fullest gain available is that we have a range of input and output matching impedances from which to choose. For example, suppose that we can accept 1 dB less gain at the input. Then designing Z_S (z_S, normalized to be precise) to be anywhere on the -1-dB circle in Figure 10.6-1 will suffice. For our convenience, choose the point on the circle at which $z_S = 0.50$ (purely resistive), corresponding to a Z_S of $(0.50)(50\ \Omega) = 25\ \Omega$. This can be accommodated by a quarter-wave inverter of characteristic impedance Z_T given by
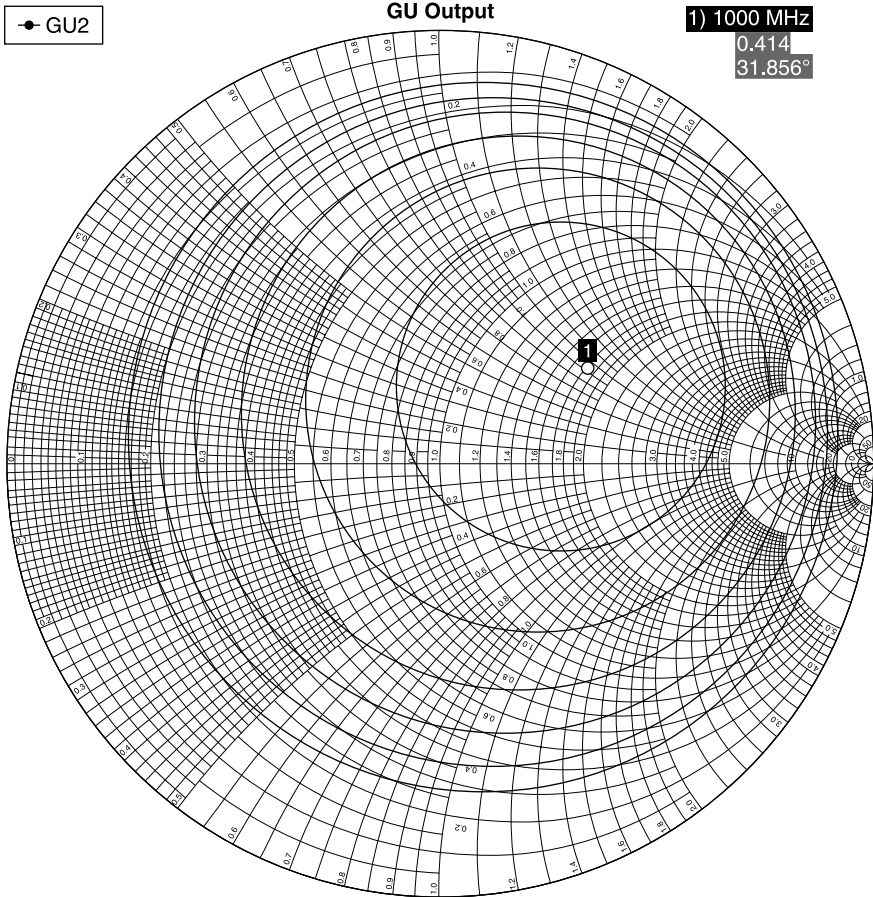
Figure 10.6-2 Constant unilateral gain circles for the output of the stabilized 2N6679A transistor. Circles about optimum value are $1, 2, 3, 4, 5$, and 6 dB lower in gain, respectively.

$$Z_T = \sqrt{Z_S Z_0} = \sqrt{(50\ \Omega)(25\ \Omega)} = 35.36\ \Omega$$

Similarly, choose the $z_L = 0.90$ point on the -1-dB circle for the output in Figure 10.6-2. This load can be realized using a $90°$ line (inverter) of $Z_T = 47.43\ \Omega$. The resulting circuit and performance is shown in Figure 10.6-3. It can be seen that neither the input nor the output is well matched, S_{11} and S_{22} having values of -5.5 and -7 dB, respectively, at 1 GHz. The gain at 1 GHz is 17.2 dB. This is 2 dB below the maximum unilateral gain of 19.2 dB, remarkably close given the accuracy of the unilateral estimate.

Making gain adjustments to the design is especially easy when the impedance inverter matching has been employed. We simply choose another real
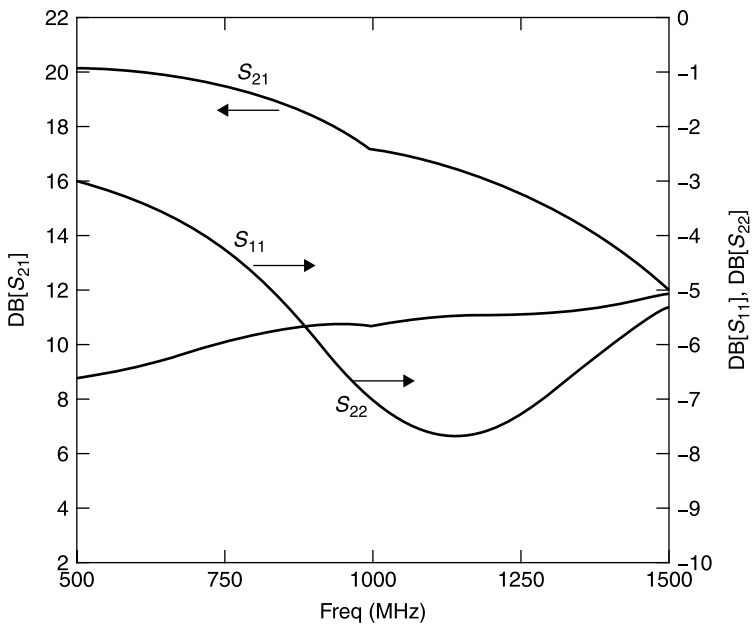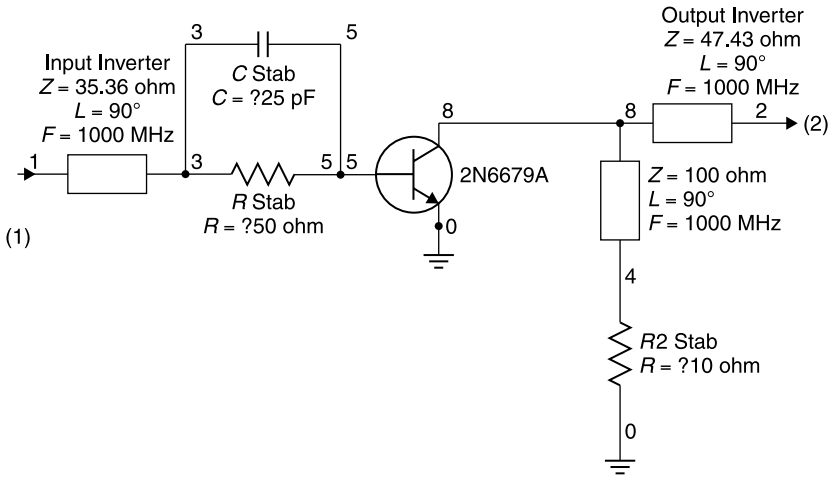
Figure 10.6-3   Unilateral design of the stabilized 2N6679A at 1 GHz for 1 dB less gain at input and 1 dB less at the output relative to maximum gain.

impedance on the gain circle desired and change the quarter-wave inverter's characteristic impedance.

Suppose that we wish to achieve a gain that is 0.5 dB below that which was obtained using the -2-dB design in Figure 10.6-2. We will effect this 0.5-dB reduction by adjusting the input source impedance. From the input plot of
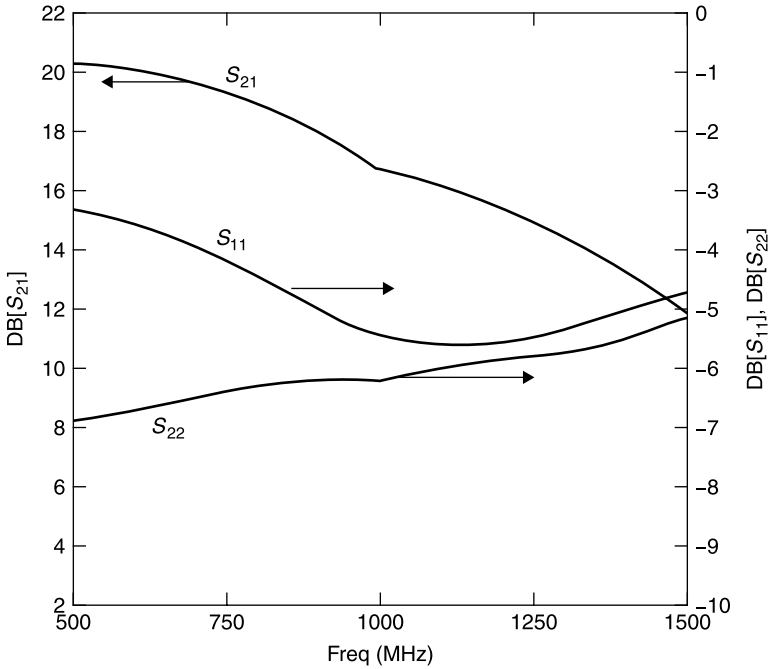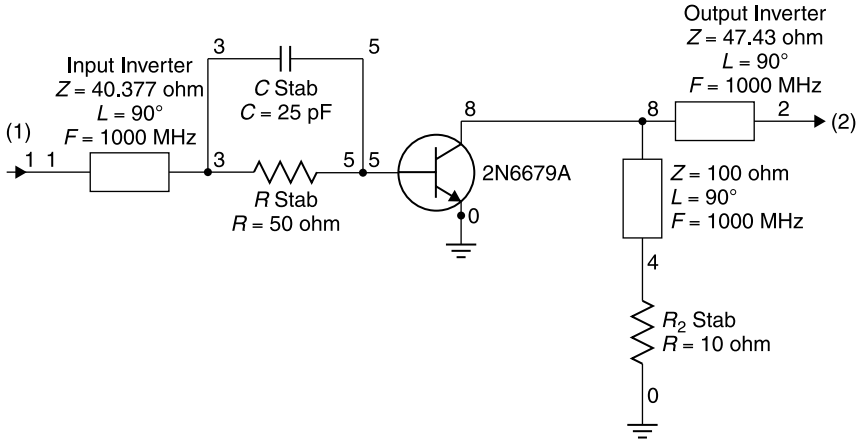
Figure 10.6-4 Gain reduction of about 0.5 dB made by changing the characteristic impedance of the input impedance inverter.

unilateral gain circles, we see that, between the -1- and -2-dB circles in Figure 10.6-1, a reduction of 1 dB occurs between $z_S = 0.5$ *and* $z_S = 0.83$. We select a midpoint $z_S = 0.67$. Then $Z_S = 33.5$ Ω, for which we need an inverter with characteristic impedance of 40.77 Ω. The result is shown in Figure 10.6-4. The gain at 1 GHz is 16.7 dB, 0.5 dB less than that of the previous value of 17.2 dB.

10.7   SIMULTANEOUS CONJUGATE MATCH DESIGN

The unilateral gain design was intended to simplify amplifier design by providing an approximate solution, ignoring feedback in the transistor and with it the interaction of source and load impedances. However, the need to design for stability required the addition of input and/or output circuitry with the consequent need to perform arduous complex calculations of stability circles. Furthermore, addition of the stabilizing circuitry also required recalculation of the S parameters of the stabilized transistor. The result is that the design of a stable amplifier, even using the unilateral design method is too complex for hand calculation. A circuit simulator or other software aid is needed to perform the considerable design labor of RF and microwave amplifier design.

The unilateral design method is also useful for providing initial insight into the various roles played by the input and output loads placed on the transistor. In fact, selection of these impedances constitutes the only RF circuit design options after the choice of a candidate transistor. Given that suitable computer aid is necessary for comprehensive amplifier design, and having observed the effects of load interactions with the unilateral design, there is no further reason to ignore the feedback term S_{12}. Rather it is appropriate to include it from the start in any amplifier design.

The inaccuracies encountered by applying the unilateral design demonstrate that the feedback term, S_{12}, causes the value of the input impedance required for a perfect match to be affected by the load impedance and vice versa. It might seem that finding a simultaneous set of source and load impedances to match input and output perfectly would require an endless series of cut-and-try designs to arrive at the optimum set of Z_S and Z_L. But this is not the case.

For an unconditionally stable transistor (or an unstable one that has been stabilized), it is possible to find a *simultaneous conjugate match* solution yielding an amplifier design for which the input and output ports are perfectly and simultaneously matched to the load and source. This approach accurately takes the feedback due to S_{12} into account. This can be accomplished at any frequency for which S parameters of a stable or stabilized transistor are available and provides the *maximum stable gain* (*MSG*) of which the transistor is capable.

The solution [2, pp. 146–147] for the reflection coefficient Γ_{SM} to be presented by the source to the stable (or stabilized) transistor is

$$\Gamma_{\text{SM}} = C_1^* \left[\frac{B_1 \pm \sqrt{B_1^2 - 4|C_1|^2}}{2|C_1|^2} \right] \qquad (10.7\text{-}1)$$

where

$$C_1 = S_{11} - \Delta S_{22}^* \qquad B_1 = 1 + |S_{11}|^2 - |S_{22}|^2 - |\Delta|^2 \qquad \Delta = S_{11}S_{22} - S_{12}S_{21}$$
$$(10.7\text{-}2a,b,c)$$

At the output port, the simultaneous match load Γ_{LM} is given by

$$\Gamma_{LM} = C_2^* \left[\frac{B_2 \pm \sqrt{B_2^2 - 4|C_2|^2}}{2|C_2|^2} \right]$$

where

$$C_2 = S_{22} - \Delta S_{11}^* \qquad B_2 = 1 - |S_{11}|^2 + |S_{22}|^2 - |\Delta|^2 \qquad \Delta = S_{11}S_{22} - S_{12}S_{21}$$
$$(10.7\text{-}3\text{a,b,c})$$

For an unconditionally stable transistor, the minus signs (where the option is $\pm$) in the above expressions produce the useful results. When provided with Γ_{SM} and Γ_{LM} terminations, the transistor has its maximum gain [2, pp. 157–158], $G_{T,\max}$:

$$G_{T,\max} = \frac{(1 - |\Gamma_{SM}|^2)|S_{21}|^2(1 - |\Gamma_{LM}|^2)}{|(1 - S_{11}\Gamma_{SM})(1 - S_{22}\Gamma_{LM}) - S_{12}S_{21}\Gamma_{SM}\Gamma_{LM}|^2} \quad (10.7\text{-}4)$$

Interestingly, this gain expression, after some complex algebra, also can be written as

$$G_{T,\max} = \frac{|S_{21}|}{|S_{12}|}[K - \sqrt{K^2 - 1}] \qquad (10.7\text{-}5)$$

where K previously has been defined in (10.3-1). When $K < 1$, the two-port network is potentially unstable. At a given frequency, the *maximum stable gain* is a figure of merit for a transistor. However, it should be noted that when providing this gain it borders on being conditionally unstable. The maximum stable gain occurs when $K = 1$. Then,

$$\text{MSG} = \frac{|S_{21}|}{|S_{12}|} \qquad (10.7\text{-}6)$$

The MSG is easily calculated from the S parameters, and transistor suppliers are fond of citing the MSG for their transistors because it gives the highest applicable gain for the device. However, if operated with this gain, the device may be on the threshold of oscillation. Practical amplifier designers must back away from this gain by a safe margin to ensure stability.

The computations of Γ_{SM} and Γ_{LM} or the corresponding source impedance Z_{SM} and load impedance Z_{LM} are complex. However, these calculations can be performed using network simulation software. For the stabilized 2N6679A transistor, the results in impedance form are given in Table 10.7-1.

The simultaneous conjugate match impedances Z_{SM} and Z_{LM} are those that must be presented to the transistor at source and load, respectively. One does

TABLE 10.7-1   Simultaneous Match Input and Output Impedances for Stabilized 2N6679A Transistor

Freq. (MHz)	Z_{SM} (Ω)	Z_{LM} (Ω)
900	$2.879 + j8.446$	$59.433 + j154.129$
1000	$4.144 + j6.335$	$54.355 + j117.564$
1100	$8.872 + j4.838$	$35.046 + j100.924$
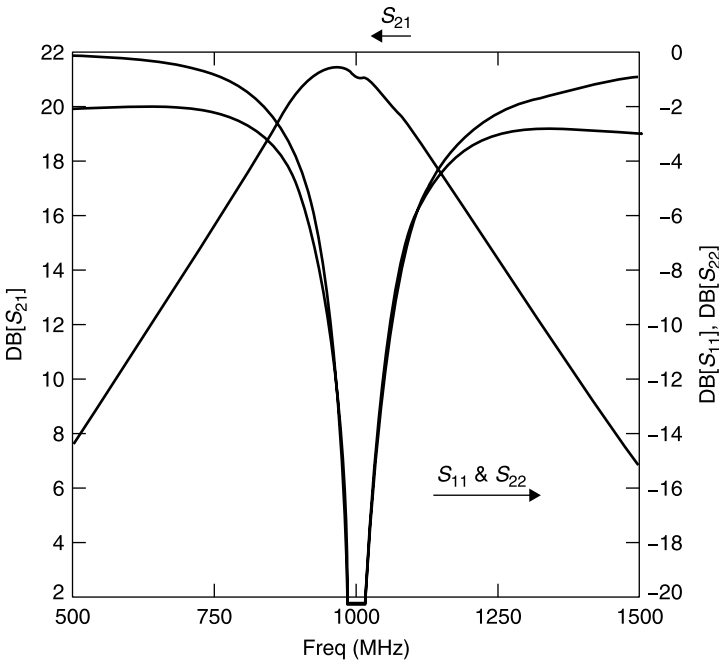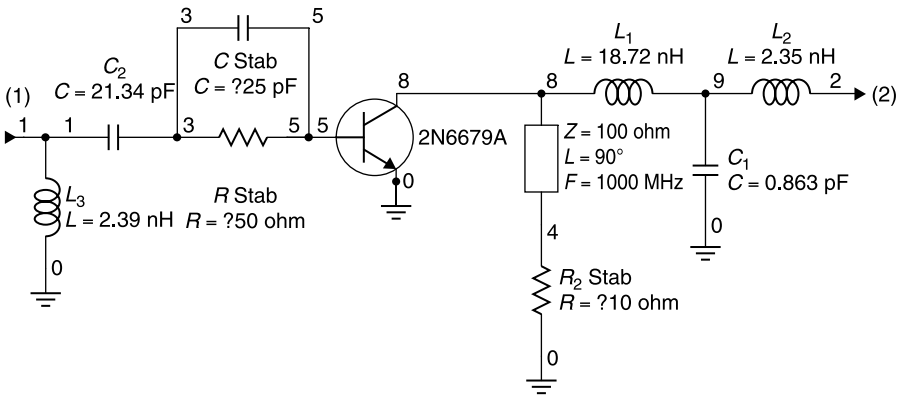
Figure 10.7-1   Simultaneous conjugate match design for the stabilized 2N6679A.

not form the complex conjugates of these impedances. Notice that they are similar but certainly not identical to the Z_S and Z_L values used for the unilateral gain design.

Unilateral design at 1 GHz:

$$Z_S = 10.494 + j9.796 \ \Omega \quad \text{and} \quad Z_L = 88.493 + j46.646$$

Simultaneous match design at 1 GHz:

$$Z_{\text{SM}} = 4.144 + j6.335 \ \Omega \quad \text{and} \quad Z_{\text{LM}} = 54.355 + j117.564$$

As an example of the application of the simultaneous match design to the stabilized 2N6679A transistor, the Q matching method of transforming the 50-Ω source and load to the required transistor load impedances will be used. The resulting circuit and performance is shown in Figure 10.7-1. It can be seen that the input and output matches, S_{11} and S_{22}, show return loss of 40 dB at 1 GHz, essentially perfect input and output matches. The gain is 21.1 dB at 1 GHz.

The expected gain can be found from the magnitudes of S_{21} and S_{12} along with the K factor. At this frequency, the stabilized 2N6679 has $|S_{21}| = 6.25$, $|S_{12}| = 0.028$, and $K = 1.151$. Then applying (10.7-5),

$$G_{T,\max} = \frac{|S_{21}|}{|S_{12}|}[K - \sqrt{K^2 - 1}] = \frac{6.25}{0.028}[1.151 - \sqrt{1.151^2 - 1}] = 129.7 = 21.1 \text{ dB}$$

This agrees with the value obtained from the circuit simulation of Figure 10.7-1.

10.8 VARIOUS GAIN DEFINITIONS

For the 50-Ω loaded amplifier, for the unilateral amplifier designs, and for the simultaneous conjugate match design, we used the *transducer gain (G_T)* definition. For a network described by S parameters, there are two other commonly used gain definitions [1, p. 213], the *operating gain (G_P)* and the *available gain (G_A)*, special cases derived from the transducer gain (Fig. 10.8-1):

$$G_T = \frac{P_L}{P_{\text{AVS}}} = \frac{\text{power delivered to the load}}{\text{power available from the source}} \tag{10.8-1}$$

$$G_P = \frac{P_L}{P_{\text{IN}}} = \frac{\text{power delivered to the load}}{\text{power delivered to the network}} \tag{10.8-2}$$

$$G_A = \frac{P_{\text{AVN}}}{P_{\text{AVS}}} = \frac{\text{power available from the network}}{\text{power available from the source}} \tag{10.8-3}$$
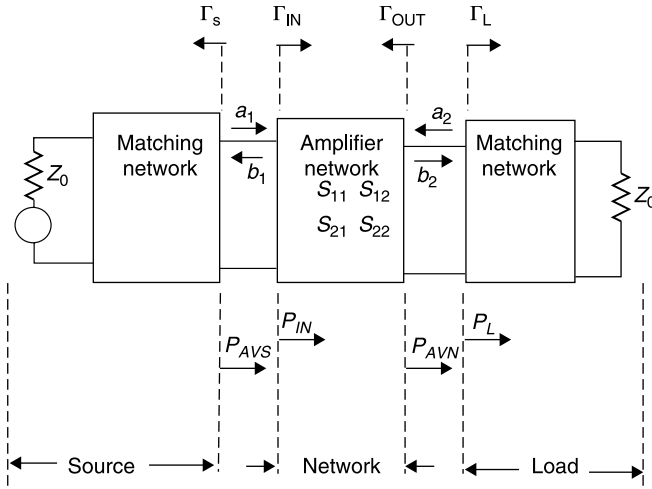
Figure 10.8-1   Gain terms.

where

$$P_{IN} = \tfrac{1}{2}|a_1|^2 - \tfrac{1}{2}|b_1|^2 \quad \text{and} \quad P_{AVS} = P_{IN} \quad (\text{when } \Gamma_{IN} = \Gamma_S^*) \qquad (10.8\text{-}4a,b)$$

$$P_L = \tfrac{1}{2}|b_2|^2 - \tfrac{1}{2}|a_2|^2 \quad \text{and} \quad P_{AVN} = P_L \quad (\text{when } \Gamma_L = \Gamma_{OUT}^*) \qquad (10.8\text{-}5a,b)$$

The three gain definitions can be expressed in terms of the S parameters and the reflection coefficients of the network, source and load [1, 2].

Transducer gain:

$$G_T = \frac{P_L}{P_{AVS}} = \frac{1 - |\Gamma_S|^2}{|1 - \Gamma_{IN}\Gamma_S|^2} |S_{21}|^2 \frac{1 - |\Gamma_L|^2}{|1 - S_{22}\Gamma_L|^2} \qquad (10.8\text{-}6)$$

or

$$G_T = \frac{P_L}{P_{AVS}} = \frac{1 - |\Gamma_S|^2}{|1 - S_{11}\Gamma_S|^2} |S_{21}|^2 \frac{1 - |\Gamma_L|^2}{|1 - \Gamma_{OUT}\Gamma_L|^2} \qquad (10.8\text{-}7)$$

Operating gain:

$$G_P = \frac{P_L}{P_{IN}} = \frac{1}{1 - |\Gamma_{IN}|^2} |S_{21}|^2 \frac{1 - |\Gamma_L|^2}{|1 - S_{22}\Gamma_L|^2} \qquad (10.8\text{-}8)$$

Available gain:

$$G_A = \frac{P_{AVN}}{P_{AVS}} = \frac{1 - |\Gamma_S|^2}{|1 - S_{11}\Gamma_S|^2} |S_{21}|^2 \frac{1 - |\Gamma_L|^2}{|1 - S_{22}\Gamma_L|^2} \qquad (10.8\text{-}9)$$

where

$$\Gamma_{IN} = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L} \quad \text{and} \quad \Gamma_{OUT} = S_{22} + \frac{S_{12}S_{21}\Gamma_S}{1 - S_{11}\Gamma_S} \quad (10.8\text{-}10a,b)$$

10.9   OPERATING GAIN DESIGN

The operating gain design is used with a matched source and optional load. The design is exact, and S_{12} is not neglected. The operating gain procedure consists of selecting a Γ_L from constant gain circles (to be presented shortly) and then finding the corresponding input match Γ_S.

The operating gain method is particularly useful in the case of power amplifiers, for which a specific load impedance for the transistor is often required for maximum power output. The load Z_L is typically specified by the manufacturer that was found empirically to yield the highest power output for the device at a specified frequency. Other applications for this method arise in which the Z_L is predetermined, perhaps by the input impedance of a subsequent filter.

Unlike the unilateral gain and simultaneous conjugate match designs, the *operating gain method can be used with potentially unstable networks.* However, if the amplifier is not unconditionally stable, care must be taken that the chosen Z_S and Z_L values do not lie within or too near the impedances that cause potential instability. Furthermore, if the design is applied with a potentially unstable network, there will be no safeguard against the network oscillating when the designed Z_S and/or Z_L may be absent, as might occur when the intended source or load is disconnected. Therefore, while the operating gain method can be used with potentially unstable networks, good engineering practice suggests that one first make the network unconditionally stable, if this can be done within the amplifier's performance requirements.

The operating gain can be rewritten as [1, pp. 247–248]

$$G_P = |S_{21}|^2 g_P \qquad (10.9\text{-}1)$$

where

$$g_P = \frac{G_P}{|S_{21}|^2} = \frac{1 - |\Gamma_L|^2}{1 - |S_{11}|^2 + |\Gamma_L|^2(|S_{22}|^2 - |\Delta|^2) - 2\,\text{Re}(\Gamma_L C_2)} \qquad (10.9\text{-}2)$$

and

$$\Delta = S_{11}S_{22} - S_{12}S_{21} \qquad (10.9\text{-}3)$$

$$C_2 = S_{22} - \Delta S_{11}^* \qquad (10.9\text{-}4)$$

These relations yield circles in the Γ_L plane (the Smith chart) having constant gain. The circles are described by

$$|\Gamma_L - C_P| = r_P \tag{10.9-5}$$

where the center of the circle C_P is located at

$$C_P = \frac{g_P C_2^*}{1 + g_P(|S_{22}|^2 - |\Delta|^2)} \tag{10.9-6}$$

and the radius of the circle is

$$r_P = \frac{[1 - 2K|S_{12}S_{21}|g_P + |S_{12}S_{21}|^2 g_P^2]^{1/2}}{|1 + g_P(|S_{22}|^2 - |\Delta|^2)|} \tag{10.9-7}$$

The procedure to design to a given transducer power gain, $G_T = G_P$, is:

1. For a given G_P, plot the center and radius of the gain circle on the Smith chart.
2. Select a desired Γ_L (or equivalently, z_L).
3. For the particular Γ_L, maximum power is obtained by matching the input according to

$$\Gamma_S = \Gamma_{IN}^* \tag{10.9-8}$$

where

$$\Gamma_{IN} = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L} \tag{10.9-9}$$

These steps are laborious if performed by hand. Alternatively, they can be accomplished using the network simulator. Constant gain circles in increments of $-1, -2, -3, -4, -5,$ and -6 dB below the optimum load Γ_M (which gives the simultaneous conjugate match) are shown in Figure 10.9-1. These have been calculated for the 2N6679A transistor alone—without its stabilizing elements.

Notice that the load instability circle (the innermost circle) lies partially within the $|\Gamma| \leq 1$ circle of the Smith chart (shaded region). The network simulator does this to indicate that these are unstable load impedances to be avoided in the selection of Γ_L. Suppose a point on the -1-dB circle is selected corresponding to the normalized impedance $z_L = 1 + j1.6$.

This means that we must load the transistor with an unnormalized impedance of $(50 + j80)$ Ω. This choice is convenient since the 50-Ω resistive part is already included in the matched termination. The reactance is provided by an

GP

Operating Gain Circles
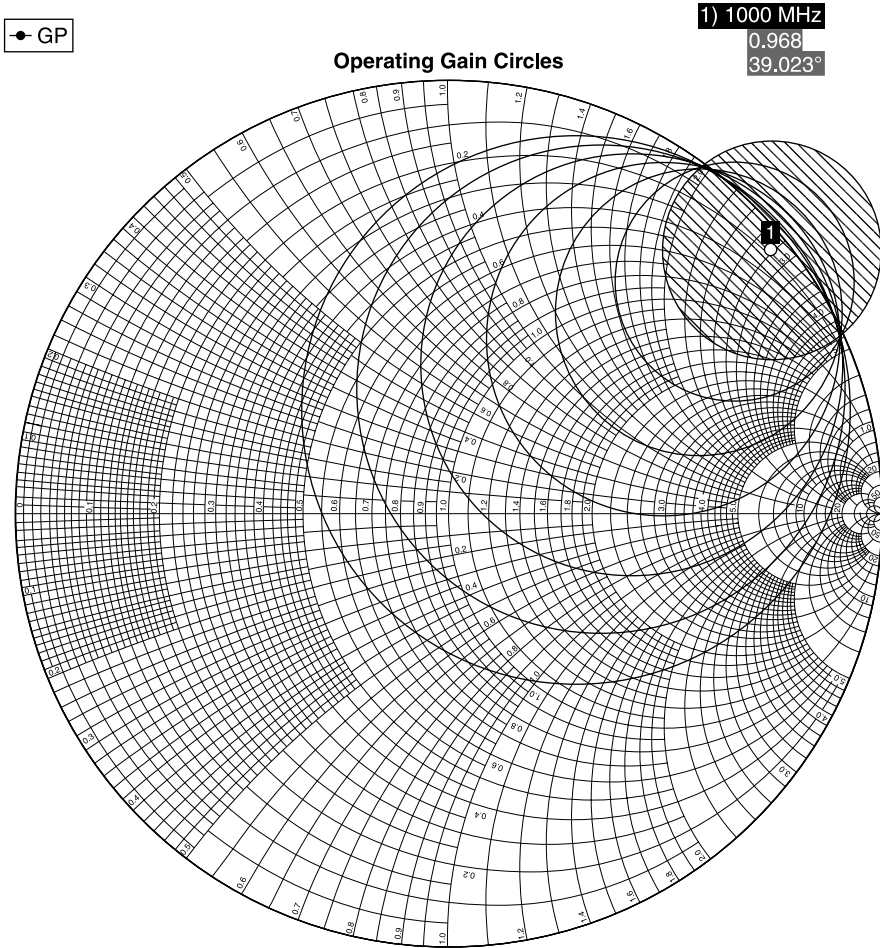
Figure 10.9-1 Operating gain circles for the unstabilized 2N6679A transistor at $-1, -2,$ $-3, -4, -5,$ and -6 dB below simultaneous conjugate match gain. Also shown is the load instability circle (shaded).

inductor

$$L_1 = \frac{80\ \Omega}{6.28\ \Omega/\text{nH}} = 12.74\ \text{nH}$$

The resulting circuit is shown in Figure 10.9-2.

Next, we determine the input impedance Z_{IN} required for the -1-dB gain, using the network simulator for the calculation (Table 10.9-1). The required source impedance is the complex conjugate of Z_{IN}, thus $Z_S = Z_{\text{IN}}^* = (4.133 +$
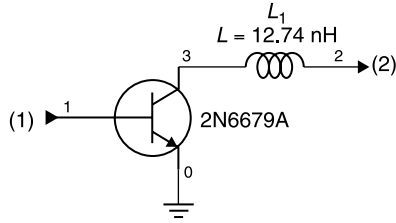
Figure 10.9-2   The 2N6679A with a load from the -1-dB operating gain circle.

$j1.448$) Ω. We can use a quarter-wave impedance inverter to obtain the real part. Its characteristic impedance Z_T is

$$Z_T = \sqrt{(4.133 \ \Omega)(50 \ \Omega)} = 14.38 \ \Omega$$

The required reactive part of Z_S can be obtained with a series inductance L_2:

$$L_2 = \frac{1.448 \ \Omega}{6.28 \ \Omega/\text{nH}} = 0.23 \ \text{nH}$$

The complete circuit and performance is shown in Figure 10.9-3. The gain at 1 GHz is 22.4 dB, higher than that obtained when input and output were matched with the *stabilized* 2N6679A circuit. As expected the input circuit is matched, $|S_{11}| = -40$ dB, but the output is not matched, and this mismatch results in a 23% power loss or about 1.1 dB. However, to obtain this additional gain would require selection of a load quite near those that cause instability.

This operating gain result might appear to be preferable to the simultaneous conjugate match and unilateral designs performed earlier. It has more gain at 1 GHz, and also has far more gain below 1 GHz, should that prove desirable. However, this design did not provide unconditional stability, as was provided by the earlier designs. Rechecking the K and B_1 values we find (Table 10.9-2) that the amplifier we have designed is potentially unstable from 100 to 1500 MHz. Checking the instability circles (Fig. 10.9-4) confirms this fact.

However, the amplifier will be stable if 50-Ω source and loads are used since none of the input and output instability circles includes the origin of the Smith chart.

TABLE 10.9-1   Input Impedance of Circuit in Figure 10.9-2

Frequency (MHz)	Re[Z_{IN}] (Ω)	Im[Z_{IN}] (Ω)
1000	4.133	-1.448
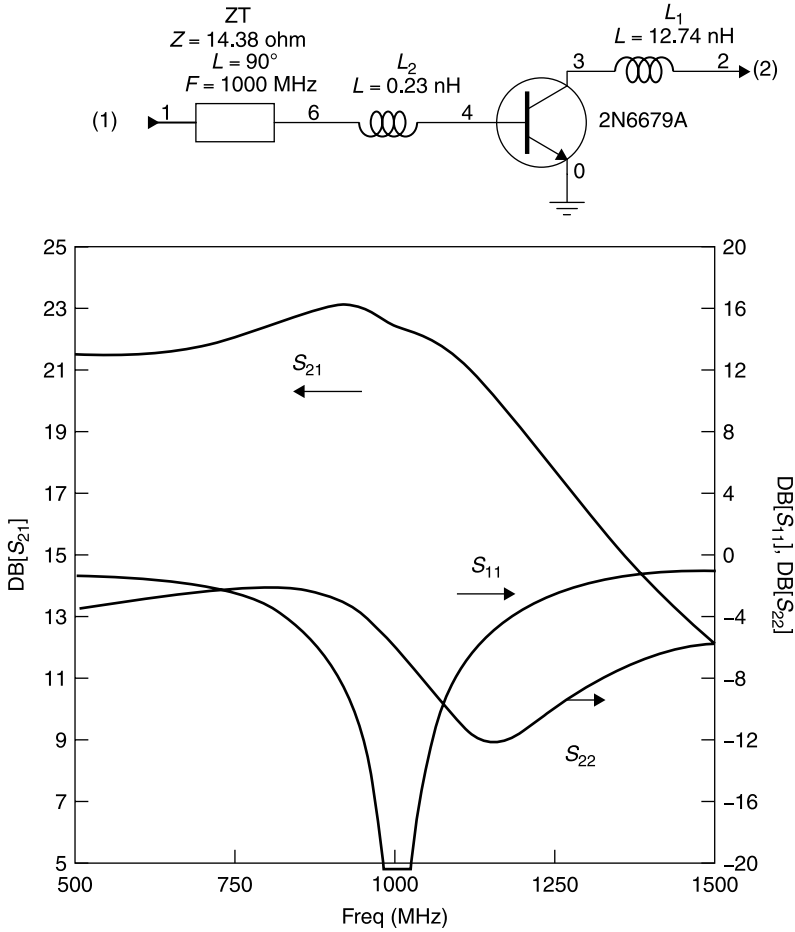
Figure 10.9-3 Schematic and performance of the completed amplifier designed using the *operating gain* method.

10.10 AVAILABLE GAIN DESIGN

The available gain design uses an optional source impedance and a matched load. This method is the complement of the operating gain. *The source impedance Z_S is selected, and then the output is matched to the corresponding Z_{OUT}, making $Z_L = Z_{OUT}^*$.* Otherwise, the procedure is the same.

Equivalently, in terms of reflection coefficients, *with the available gain approach, the source reflection coefficient Γ_S is selected and then the output is matched, making $\Gamma_L = \Gamma_{OUT}^*$.*

TABLE 10.9-2 K **and** B_1 **Factors for Operating Gain Design**

Frequency (MHz)	K	B_1
100	0.265	0.570
200	0.317	0.724
300	0.377	0.785
400	0.474	0.812
500	0.685	0.825
600	0.683	0.640
700	0.696	0.390
800	0.728	0.072
900	0.784	−0.180
1000	0.877	0.134
1100	0.874	0.591
1200	0.880	1.005
1300	0.897	1.163
1400	0.924	1.183
1500	0.966	1.151

Using the 2N6679A transistor for an example, the process is begun by selecting Z_S on an available gain circle (Fig. 10.10-1).

In this case the circles are symmetric with respect to the real axis facilitating selection of a purely real source impedance. On the −1-dB circle the real axis intercept is $z_S = 0.16$. Then $Z_S = (50\ \Omega)(0.16) = 8\ \Omega$ (Fig. 10.10-2). A quarter-wave impedance inverter can transform the 50-Ω source to the required Z_S when it has the characteristic impedance, Z_T, given by

$$Z_T = \sqrt{(50\ \Omega)(8\ \Omega)} = 20\ \Omega$$

The circuit then becomes as shown in Figure 10.10-2.

Next we determine Z_{OUT} and conjugately match to it (Table 10.10-1). Matching to the complex conjugate,

$$Z_L = (59.919 + j118.348)\ \Omega$$

To obtain this Z_L, we use a quarter-wave impedance inverter to obtain the real part and then a series inductor for the imaginary part:

$$Z_{T2} = \sqrt{(59.919\ \Omega)(50\ \Omega)} = 54.74\ \Omega$$

and the reactive part is obtained by an inductance L_2, having the value
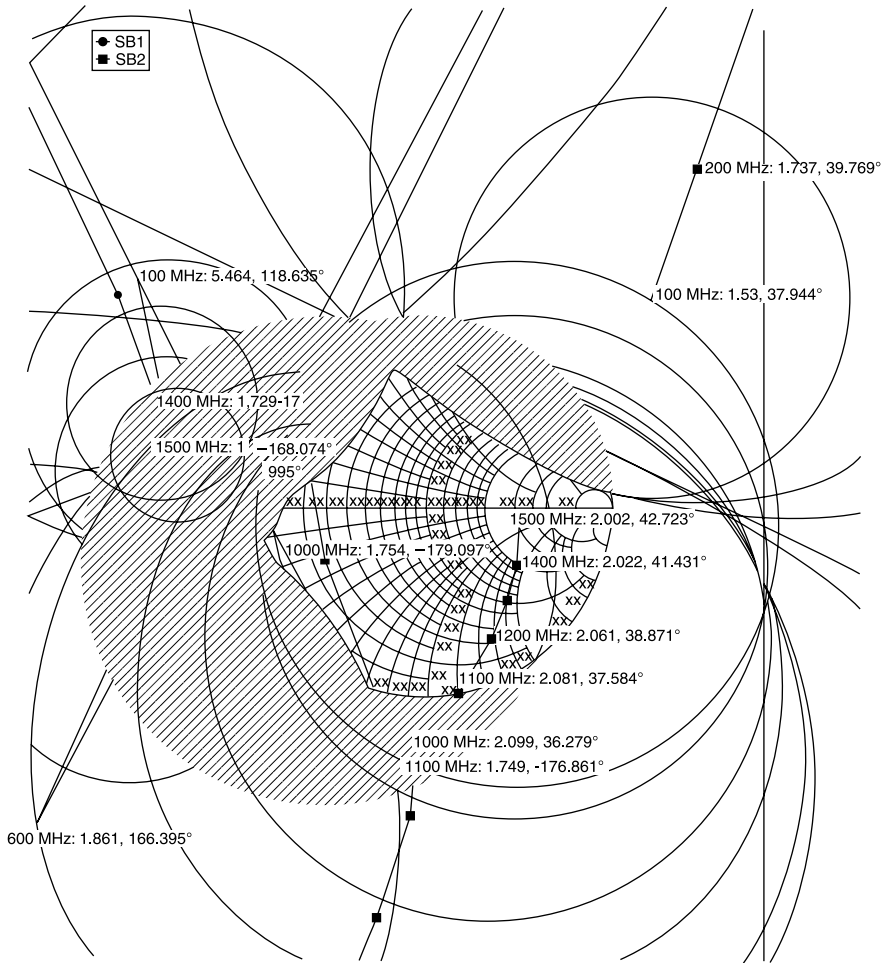
Figure 10.9-4   Input and output instability circles (100 to 1500 MHz) for the operating gain designed amplifier. The network is seen to be stable only when loaded with source and load impedances near the center of the Smith chart.

$$L_2 = \frac{118.348\ \Omega}{6.28\ \Omega/\text{nH}} = 18.85\ \text{nH}$$

The -1-dB available gain circuit design is shown in Figure 10.10-3.

The gain at 1 GHz is the same for the available gain amplifier as for the operating gain design performed previously, 22.4 dB. We would expect this since both were designed to be 1 dB below the simultaneous conjugate matched gain. Also, the nonmatched input port has a return loss of -6.2 dB, also similar to the operating gain result. Also, like the operating gain amplifier, the available gain design is potentially unstable. The K and B_1 values are shown in Table 10.10-2.
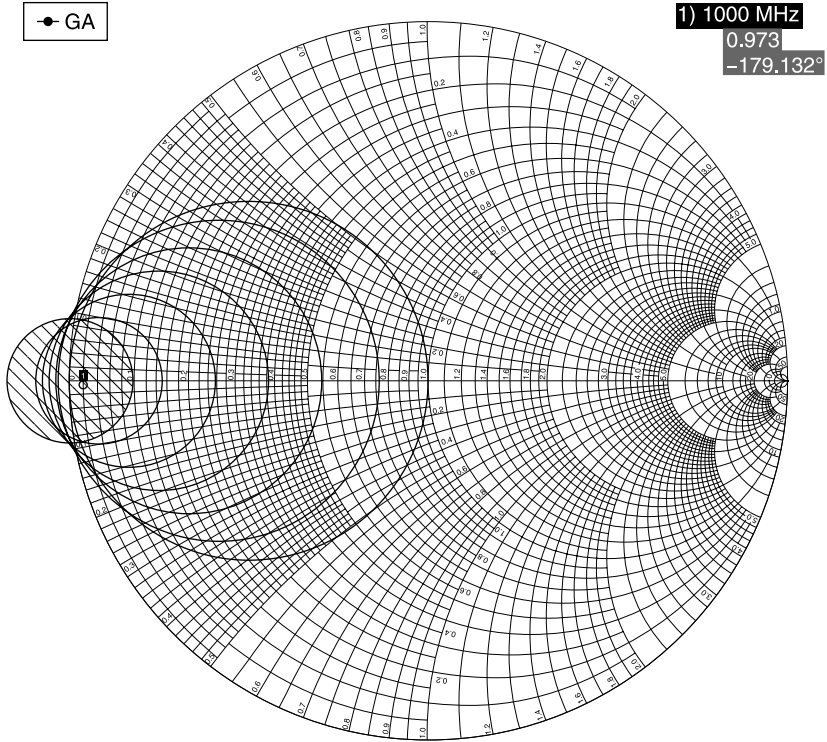
Figure 10.10-1 Available gain circles at 1 GHz for the 2N6679A transistor for $-1, -2,$ $-3, -4, -5,$ and -6 dB. Also shown is the source instability circle (shaded).
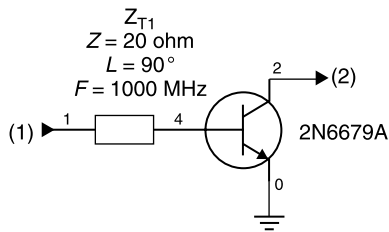
Figure 10.10-2 Transistor provided with an 8-Ω source impedance to provide an available gain 1 dB less than maximum.

TABLE 10.10-1 Calculated Output Impedance of Circuit in Figure 10.10-2

Frequency (MHz)	Re[Z_{OUT}]	Im[Z_{OUT}]
1000	59.919	-118.348
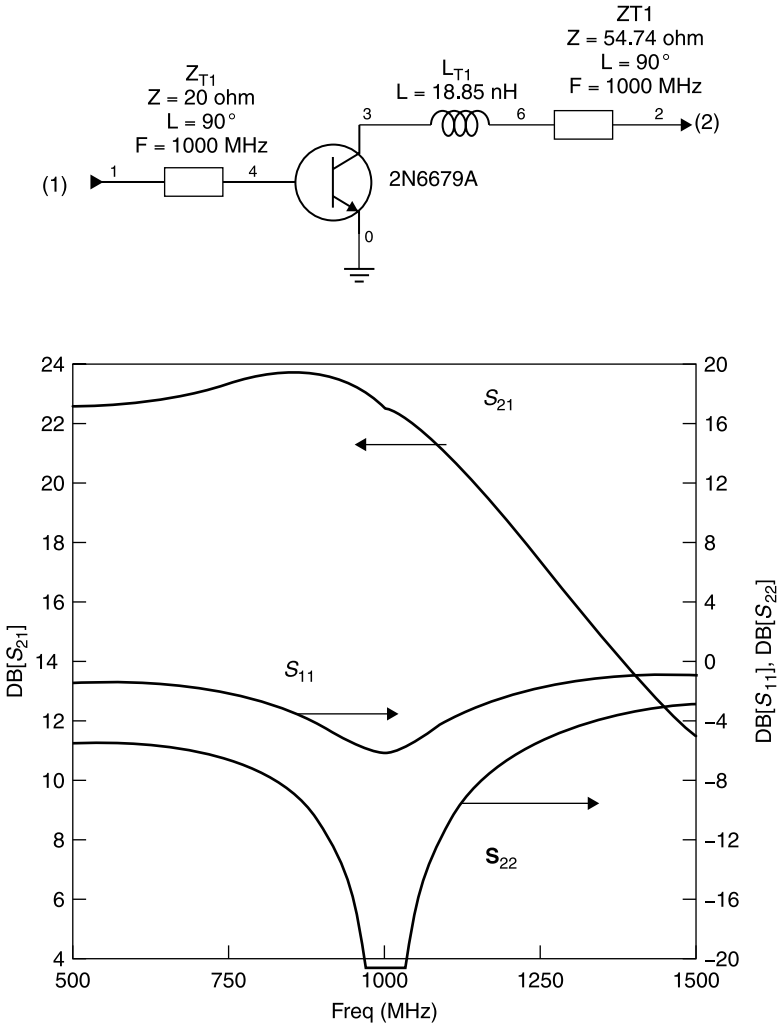
Figure 10.10-3 Design and performance of the available gain amplifier.

**TABLE 10.10-2   *K* and *B*$_1$ Factors for *Available Gain*
Design Shown in Figure 10.10-2**

Frequency (MHz)	K	B_1
0	0.195	0.203
1000	0.875	0.61
2000	0.962	0.447
3000	1.755	0.101
4000	0.863	0.044
5000	0.663	0.017
6000	0.526	8.75×10^{-3}
7000	0.491	6.009×10^{-3}
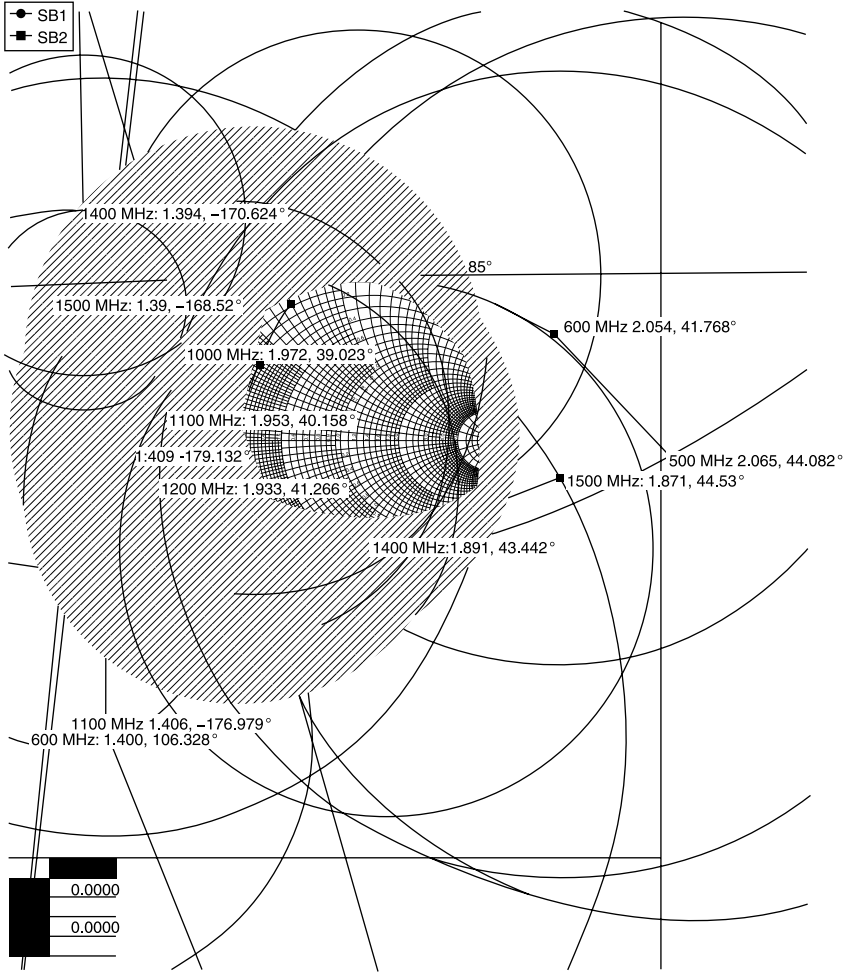
Figure 10.10-4 Input and output instability circles for the *available gain* amplifier design of Figure 10.10-3.

Finally, as with the operating gain design, the available gain design requires loads near the center of the Smith chart (Fig. 10.10-4) to be stable.

10.11   NOISE IN SYSTEMS

Thermal Noise Limit

One might wonder why it is a concern what the path loss of a radio signal is when amplification can be inexpensively added to an arbitrary degree, raising

any signal, no matter how weak, to a comfortable listening level. The answer is in the *thermal noise limit*.

An important and unavoidable source of noise in electronic systems is thermal noise, whose power level is proportional to absolute temperature. This *thermal noise* is also known as *Johnson noise*. It is *white noise* because *its density is frequency independent over much of the electromagnetic spectrum*. The open-circuit noise voltage across a resistance R has a zero time average value and a mean-squared value of [2, Chapter 8]

$$\langle |e|^2 \rangle = 4kTRB \tag{10.11-1}$$

where e = mean-squared value of the noise voltage
$\langle \ \rangle$ = mean value of the quantity inside over time
k = Boltzmann's constant = 1.38044×10^{-23} J/K
$\qquad \qquad = 8.63 \times 10^{-5}$ eV/K
T = absolute temperature (K)
R = resistance of the noise source, and
B = bandwidth (in Hz) over which the noise is measured

From the equation for available power from a voltage source, the available noise power to a network is $e^2/4R$. It follows that the spectral power density of thermal noise is

$$P_N = kTB \text{ (in watts)} \tag{10.11-2}$$

Notice that *thermal noise power* (Fig. 10.11-1) *is independent of resistance (or source impedance)!* At room temperature (290 K, 20°C, or 68°F) the noise power in a 1 Hz bandwidth is

$$
\begin{aligned}
P_N &= (1.38044 \times 10^{-23} \text{ J/K})(290 \text{ K})(1 \text{ Hz}) \\
&= (1.4 \text{ dB} - 230 \text{ dB} + 24.6 \text{ dB}) \text{ W/Hz} \\
&= -204 \text{ dBW/Hz} \\
&= -174 \text{ dBm/Hz} \tag{10.11-3}
\end{aligned}
$$

For example, in an AMPS (Advanced Mobile Phone System) signal bandwidth of 30 kHz, the thermal noise at 290 K is

$$30 \text{ kHz} = 30{,}000 \text{ Hz} = +44.8 \text{ Hz (dB)}$$

$$P_N = -174 \text{ dBm/Hz} + 44.8 \text{ Hz (dB)} = -129.2 \text{ dBm}$$

For a standard television channel, having a 6 MHz bandwidth, the noise is
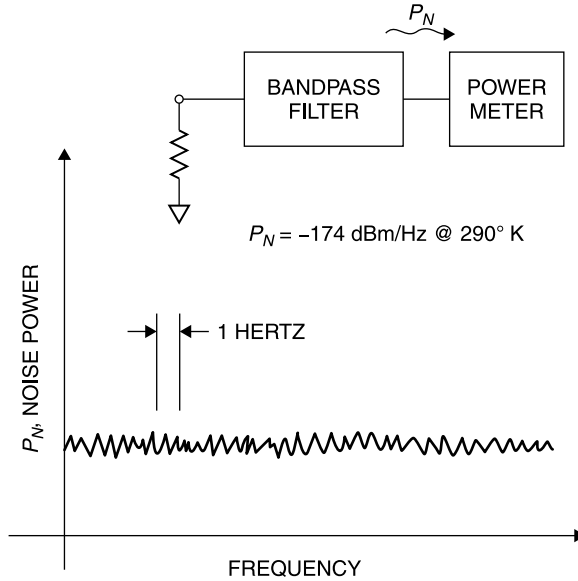
Figure 10.11-1   Thermal noise power available from a resistor.

$$6 \text{ MHz} = 6{,}000{,}000 \text{ Hz} = +67.8 \text{ dB}$$

$$P_N = -174 \text{ dBm/Hz} + 67.8 \text{ Hz (dB)} = -106.2 \text{ dBm}$$

Therefore, other things being equal, a stronger signal must be received for television reception than for voice communication since the broader bandwidth required for television contains more thermal noise. Spread spectrum signals, which artificially employ more bandwidth than the signal information requires, are a special case. Spread spectrum signals can be received below the noise floor [4].

Other Noise Sources

There are other sources of noise, but the thermal noise is unavoidable and serves as a baseline for the noise to be expected. Some other sources of noise include [5, Sec. 27]:

Flicker noise (also known as $1/f$ noise), in amplifiers at low frequencies (below 50 kHz) [1, p. 79]

Man-made noise (auto ignitions, sparking motors, etc.)

Atmospheric noise (lightning)

Precipitation static (caused by rain, hail, snow)

Galactic noise (exoatmospheric noises, e.g., the sun)

Carrier noise, caused by the discrete nature of charge carriers in diodes, transistors, and tubes

Despite the number of additional noise sources beyond thermal noise, *most radio system performances estimates are based on limitations imposed only by thermal noise*. The apparent reasoning for this is that other noise sources are not always prevalent, and when they are, may be compensated by the system margin. The *noise margin* is the difference between the thermal noise floor and the anticipated signal strength at the receiver under ideal (usually line of sight) transmission conditions. A portion of the noise margin is required to compensate for the signal losses attendant with multiple bounces around buildings and other obstacles.

Noise Figure of a Two-Port Network

The noise factor F of a linear two-port network is the ratio of the available noise power at its output, P_{NO}, divided by the product of the available noise power at its input, P_{NI}, times the network's numeric gain G. Thus,

$$F = \frac{P_{NO}}{P_{NI}G} = \frac{\bar{P}_{NO}}{P_{NI}} \tag{10.11-4}$$

where $\bar{P}_{NO}$ is the output noise power referred to the input terminals, that is, the output noise power divided by the gain G. An equivalent statement is: *Noise factor F is the ratio of the signal/noise power at the input to the signal/ noise power at the output.*

When expressed in decibels the noise power ratio is called the *noise figure, NF:*

$$Noise\ figure = \text{NF} = 10 \log F \tag{10.11-5}$$

Since F is a function of P_{NI}, it does not uniquely characterize the noise performance of the two-port network. For example, *if a very high noise temperature is used at the input in the determination of F, it will tend to swamp out the noise introduced by the two-port network, resulting in a lower measured noise factor or noise figure.*

To overcome the uncertainty in the specification of the noise factor or noise figure of a two-port network, *the Institute of Electrical and Electronics Engineers (IEEE) has standardized the definition by specifying P_{NI} as the noise available from a resistor at 290 K. Thus*

$$P_{NI} \equiv kT_0B \quad \text{where} \quad T_0 = 290 \text{ K } (17°\text{C or } 62.6°\text{F}) \tag{10.11-6}$$

Accordingly, for noise factor measurements, one must use

$$kT_0 = 4 \times 10^{-21} \text{ W/Hz} \tag{10.11-7}$$

and the standard noise factor becomes

$$F = \frac{\bar{P}_{NO}}{kT_0B} \tag{10.11-8}$$
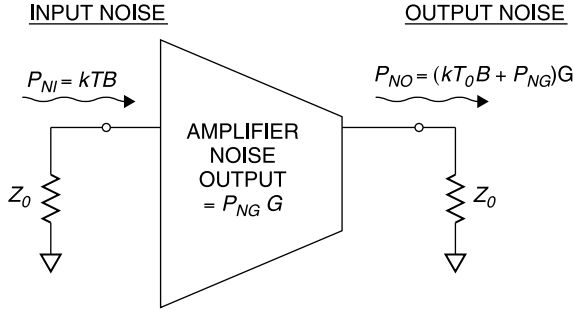
Figure 10.11-2   Calculation of amplifier noise factor F by referring all noise sources to the input of the amplifier.

where $\bar{P}_{NO} = P_{NO}/G$ is the total output noise power referred to the input port when the input signal channel is terminated by a resistor at temperature T_0.

As an example, consider an amplifier with a gain of 20 dB and a bandwidth of 5 MHz. Assume that the output noise power generated by the amplifier in its 5 MHz bandwidth is $GP_{NG} = -83$ dBm, where P_{NG} is the noise generated by the amplifier, referred to the input port. The noise figure (Fig. 10.11-2) of the amplifier is calculated as

$$kT_0B = (4 \times 10^{-21} \text{ W/Hz})(5 \times 10^6 \text{ Hz}) = 2 \times 10^{-14} \text{ W}$$

$$P_{NG}G = 5 \times 10^{-12} \text{ W}$$

Since the two noise sources are uncorrelated, their power values can be added, thus

$$P_{NO} = 100(2 \times 10^{-14} \text{ W}) + 5 \times 10^{-12} \text{ W} = 7 \times 10^{-12} \text{ W}$$

$$\bar{P}_{NO} = 7 \times 10^{-14} \text{ W}$$

$$F = \frac{\bar{P}_{NO}}{kT_0B} = \frac{7 \times 10^{-14} \text{ W}}{2 \times 10^{-14} \text{ W}} = 3.5 \quad \text{(or 5.4 dB)}$$

Note that both noise sources can be referred to the input port. Thus

$$\bar{P}_{NO} = kT_0B + P_{NG}$$

$$F = \frac{\text{total output noise (referred to the input)}}{\text{thermal noise at input}} \qquad (10.11\text{-}9)$$

$$= \frac{kT_0B + P_{NG}}{kT_0B} = 1 + \frac{P_{NG}}{kT_0B}$$
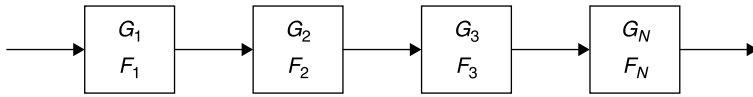
Figure 10.11-3   Noise factor for a cascade of two ports, each with its own noise factor F_i and gain G_i.

Noise Factor of a Cascade

When separate networks are cascaded each having its own gain G_i and noise factor F_i, all networks add noise to the signal that travels through them, but the contribution to the overall noise factor from succeeding networks is reduced when previous stages have amplified the signal (Fig. 10.11-3). The overall noise factor is calculated using

$$F_{\text{Cascade}} = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} + \cdots + \frac{F_i - 1}{G_1 G_2 \cdots G_{i-1}} + \cdots \quad (10.11\text{-}10)$$

Note that the calculation in (10.11-10) must be performed using the numeric value for gain as well as noise factor F, not NF (in decibels). The implications of this simple cascade noise figure formula can be considerable in real system designs.

The noise factor of a matched lossy two-port network is

$$F = L \qquad (10.11\text{-}11)$$

where L is the insertion loss factor of the two-port network. This relation assumes that the physical temperature of the loss element is approximately equal to $T_0 = 290$ K. It has a gain equal to the reciprocal of its loss ratio, or

$$G = 1/L \qquad (10.11\text{-}12)$$

For example, a matched two-port network at 290 K having 3 dB of dissipative loss has a loss ratio $L = 2$, and this is also its noise factor, $F = 2$. The gain is $\frac{1}{2} = 0.5$.

As a system noise figure example, consider the block diagram of a satellite receiver system in Figure 10.11-4. Suppose that the system has an outdoor dish
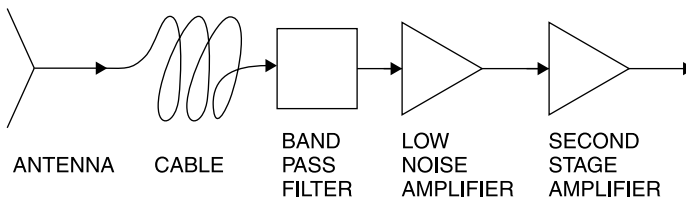
Figure 10.11-4   Block diagram of a satellite receiver front end.

antenna pointed at a satellite, and that, to place the receiver electronics in-doors, a cable (with 3-dB loss) from the antenna into the building is used. The system has a bandpass filter (with 2-dB loss) to eliminate out-of-band noise, and a low-noise amplifier with a gain of 15 dB and noise figure of only 1.0 dB. Finally, the low-noise amplifier is followed by a high gain amplifier, having 20 dB of gain and noise figure of 2.0 dB. The noise figure of the system is found as follows.

The noise factor of the cascade in Figure 10.11-4 is found using (10.11-10):

$$F_{\text{Cascade}} = 2.00 + \frac{0.58}{0.5} + \frac{0.26}{(0.5)(0.63)} + \frac{0.58}{(0.5)(0.63)(31.6)}$$

$$= 4.04 \quad (\text{or } 6.06 \text{ dB})$$

This is a very high value, given the low noise figure of the first amplifier. Even though we used a low-noise amplifier with a noise figure of only 1.0 dB, our front-end noise figure is 6 dB. After some consideration, the low-noise and second-stage amplifiers are located in the antenna feed and their output cabled into the building. The new system block diagram is shown in Figure 10.11-5.
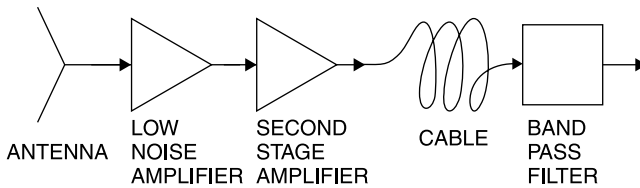
Figure 10.11-5   Revised satellite receiver front end.

For the revised circuit the noise factor is calculated as

$$F_{\text{Cascade}} = 1.26 + \frac{0.58}{31.6} + \frac{1.00}{(31.6)(100)} + \frac{0.58}{(31.6)(100)(0.5)}$$

$$= 1.28 \quad (\text{or } 1.07 \text{ dB})$$

This revised design has a noise figure that is nearly 5 dB less. Accordingly, a satellite dish antenna with one-third the area, or a little over half the diameter, can be used, yielding the same signal-to-noise ratio.

Noise Temperature

As an alternative to specifying the noise figure of a linear two-port network, one can equivalently specify the effective noise temperature T_e (Fig. 10.11-6). This method directly characterizes the noise generated by the two port net-work, P_{NG}, in terms of an equivalent amount of thermal noise produced by a
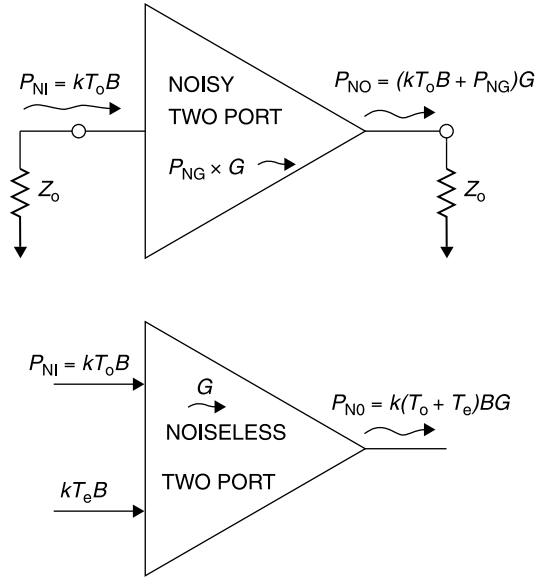
Figure 10.11-6   Use of noise temperature T_e as an alternative to noise factor.

resistor at the temperature T_e. The equivalence is established using the previously derived equation for noise factor F:

$$F = 1 + \frac{P_{NG}}{kT_0 B} = 1 + \frac{kT_e B}{kT_0 B} = 1 + \frac{T_e}{T_0} \qquad (10.11\text{-}13)$$

or conversely

$$T_e = (F - 1)T_0 \qquad (10.11\text{-}14)$$

It is important to note that T_e is not the physical temperature of the two-port network, but merely a way of characterizing its noisiness.

For example, if NF = 6 dB, $F = 4$ and $T_e = 3 \times 290 = 870$ K. The noise temperature of a cascade of networks can be found in a manner similar to that used for noise factor. If each two-port network has an equivalent noise temperature T_{ei}, the cascade has an overall noise temperature $T_{e\text{-Cascade}}$ given by

$$T_{e\text{-Cascade}} = T_{e1} + \frac{T_{e2}}{G_1} + \frac{T_{e3}}{G_1 G_2} + \cdots + \frac{T_{ei}}{G_1 G_2 \cdots G_{i-1}} \qquad (10.11\text{-}15)$$

For the previous satellite system the total noise temperature can be calculated using the respective effective noise temperatures of the components. These values are calculated using the noise factor values defined previously and the conversion formula:

$$F = 1 + \frac{T_e}{T_0} \quad \text{or} \quad T_e = (F - 1)T_0$$

Using this relation, the effective noise temperatures of the component two ports are: The LNA (75.4 K), the second-stage amp (168.2 K), the cable (290 K), and the bandpass filter (168.2 K). Applying (10.11-13) gives

$$T_{e\text{-Cascade}} = 75.4 + \frac{168.2}{31.6} + \frac{290}{(31.6)(100)} + \frac{168.2}{(31.6)(100)(0.5)}$$

$$= 75.4 + 5.32 + 0.09 + 0.11$$

$$= 80.9 \text{ K}$$

Converting this 80.9 K effective noise temperature to noise factor F,

$$F = 1 + \frac{T_e}{T_0} = 1 + \frac{80.9}{290} = 1.28$$

Expressed in decibels, the noise figure (NF) is then 1.07 dB, as we calculated previously.

10.12   LOW-NOISE AMPLIFIERS

Transistors require an optimum source reflection coefficient, Γ_{OPT}, or equivalently an optimum source impedance, Z_{OPT}, at their input in order to deliver lowest noise factor, F_{MIN}. Since source reflection coefficient (or equivalently, source impedance) is specified, *the available gain design method is used for low-noise amplifiers*. If the source is not equal to Γ_{OPT}, then the actual noise factor F of the amplifier is given by [2, Sec. 9.2]

$$F = F_{\text{MIN}} + \frac{R_n}{Z_0} \frac{|\Gamma_S - \Gamma_{\text{OPT}}|}{(1 - |\Gamma_S|^2)|1 + \Gamma_{\text{OPT}}|^2} \tag{10.12-1}$$

where R_n is the *correlation resistance* and Z_0 is the characteristic *impedance of the system in which* Γ_{OPT} *is measured*. The values of $F_{\text{MIN}}, \Gamma_{\text{OPT}}$, and R_n are specified by the manufacturer for each test frequency. The values of Γ_S that provide a constant noise factor value F form circles on the Smith chart, just as constant-gain loci are circles on the Smith chart. Given $F_{\text{MIN}}, \Gamma_{\text{OPT}}$, and R_n, these circles can be plotted for various values of F in excess of F_{MIN} for a given transistor and frequency. The circles [2] have centers at

$$c = \frac{\Gamma_{\text{OPT}}}{N + 1} \tag{10.12-2}$$

and radii

$$r = \frac{\sqrt{N[N + (1 - |\Gamma_{OPT}|^2)]}}{N + 1}$$

(10.12-3)

where

$$N = \frac{Z_0(F - F_{MIN})|1 + \Gamma_{OPT}|^2}{4R_n}$$

(10.12-4)

As an example, consider the design of an amplifier to operate at 2000 MHz. To obtain a low noise figure, a gallium arsenide field-effect transistor (FET), model NE67300, is selected, S and noise parameters for which are shown in Table 10.12-1 [6]. The data for the unpackaged device (transistor chip) are used for this example.

TABLE 10.12-1   Fifty-Ohm System S and Noise Parameters for NE67300 Low-Noise Transistor[a]

! FILENAME:		NE67300.S2P					
! NEC PART NUMBER:		NE67300					
! BIAS CONDITIONS		VDS = 3 V, IDS = 10 mA					
# GHz S MA R 50							
! S-Parameter DATA							
2	0.95	−26	3.79	161	0.04	79	0.59	−13
4	0.89	−50	3.26	141	0.06	66	0.58	−24
6	0.82	−70	2.83	126	0.08	56	0.54	−33
8	0.78	−88	2.55	114	0.09	51	0.50	−42
10	0.73	−102	2.21	104	0.10	48	0.47	−48
12	0.71	−114	2.16	93	0.10	43	0.45	−55
14	0.71	−122	2.11	90	0.11	44	0.47	−62
16	0.67	−128	1.92	76	0.11	43	0.49	−64
18	0.66	−140	1.81	63	0.11	40	0.52	−70
! NOISE PARAMETERS							
1	0.30	0.90	17	0.65			
2	0.35	0.84	40	0.57			
4	0.40	0.72	79	0.48			
6	0.55	0.62	112	0.39			
8	0.80	0.56	143	0.33			
10	1.1	0.50	168	0.28			
12	1.4	0.46	−165	0.24			
14	1.7	0.43	−140	0.20			
16	2.0	0.40	−112	0.18			
18	2.5	0.40	−84	0.16			

[a] Noise parameters are (left to right): freq (GHz), F_{MIN} (dB), Mag[Γ_{OPT}], angle[Γ_{OPT}], and (R_n/Z_0).
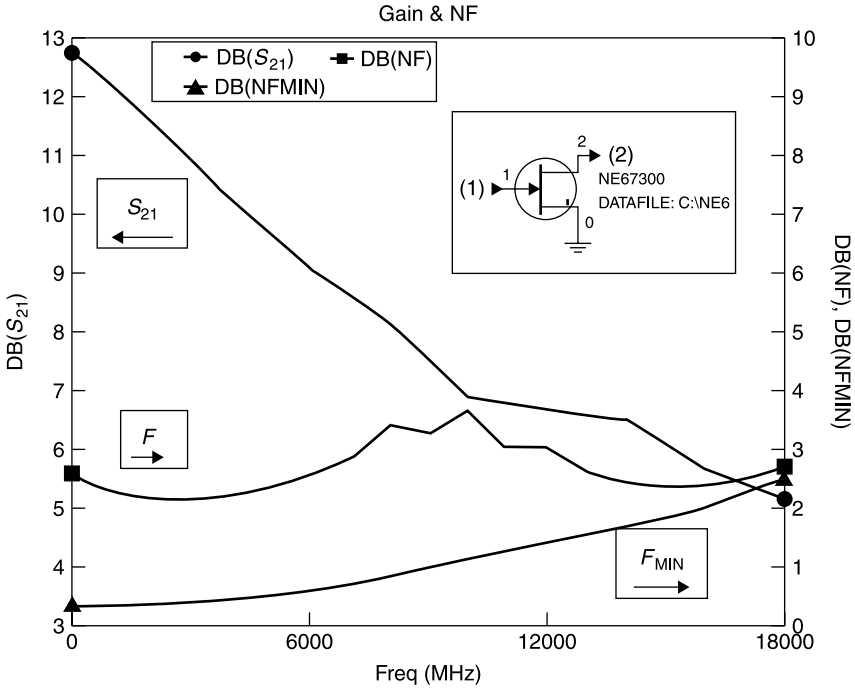
Figure 10.12-1   Performance of the NE67300 with frequency.

The gain (S_{21}), minimum noise figure, NF_{MIN} and actual noise figure, NF when no matching is provided are shown in Figure 10.12-1. It can be seen that without the necessary circuitry to present Γ_{OPT} to the input, the transistor has a much higher noise figure than the minimum of which it is capable. At 2 GHz, $NF = 2.0$ dB while $NF_{MIN} = 0.4$ dB.

From the calculated data in Figure 10.12-1 it is evident that an amplifier built using the NE67300 could yield a noise figure as low as 0.4 dB at 2 GHz. However, the device is potentially unstable over the 2 to 18 GHz bandwidth, as seen from the values of K and B_1 in Figure 10.12-2.

Therefore, before embarking on the amplifier design, we will add some circuit elements to improve the stability conditions. Often it is found that use of an inductor in series with the common lead, in this case the drain, produces negative feedback that improves stability, tends to make Γ_{OPT} and Γ_{1M} (the source reflection coefficient for maximum gain) move closer to each other, and does not materially increase NF_{MIN}. Figure 10.12-3 shows the result of using a 1-nH inductor for this purpose.

The drain lead inductor has improved the stability in that the K factor is much closer to unity over a broad portion of the gainful bandwidth of the transistor. This has been obtained at the expense of gain; however, some of the gain loss can be recovered by tuning the output. Since the circuit is not yet
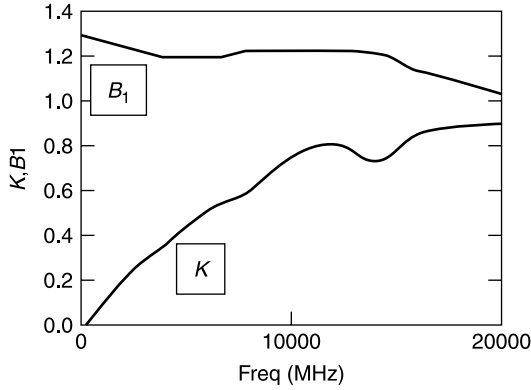
Figure 10.12-2 Stability values K and B_1 for the NE67300.

fully stabilized, particularly at the operating frequency of 2 GHz, we add some resistive damping to the output of the transistor. Placing resistive elements in the input circuit would more seriously reduce NF_{MIN}. Figure 10.12-4 shows the results of adding a series RC ($R = 300\ \Omega$ *and* $C = 10$ pF*)* circuit across the output to provide increasing damping with frequency.

It can be seen that the transistor is almost stable over the entire bandwidth for which it has gain. It is not always necessary to obtain unconditional stabil-
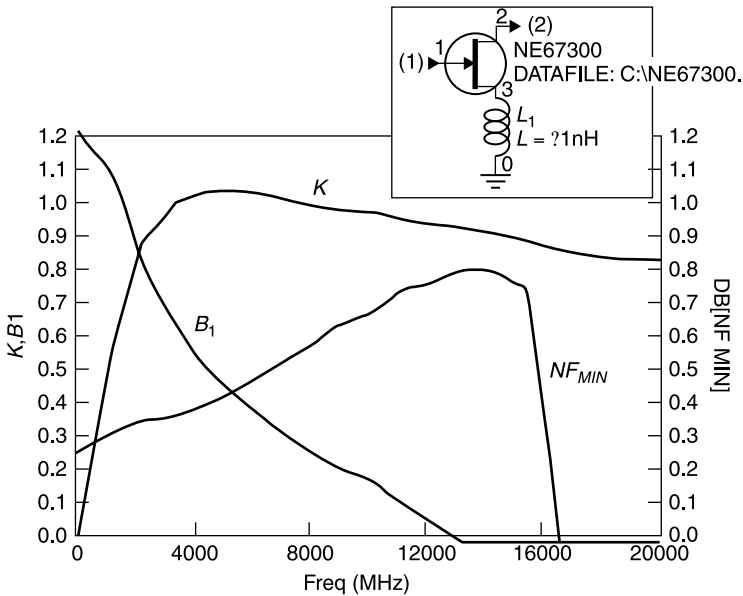
Figure 10.12-3 Stability factors with a 1-nH drain inductor.
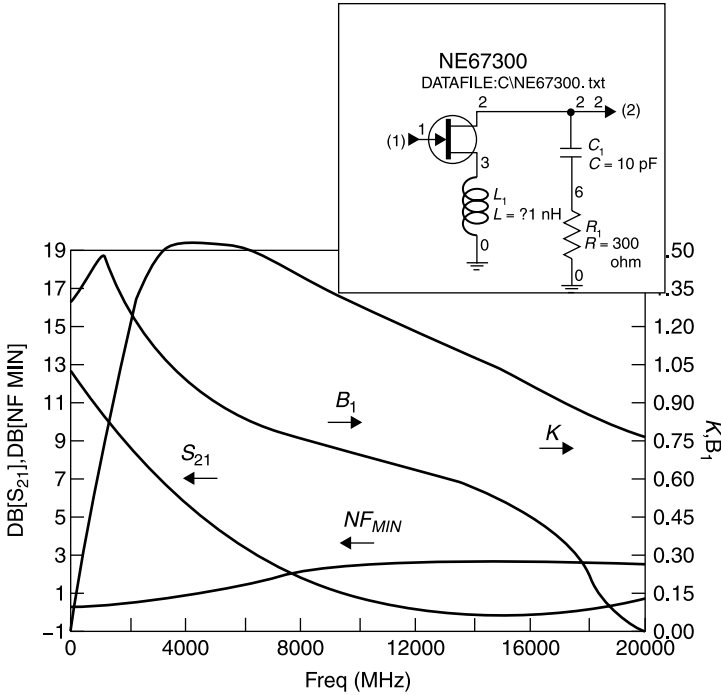
Figure 10.12-4   Amplifier performance with drain inductor and collector RC network for stability.

ity for a low-noise amplifier because its source and load impedances usually are well controlled and not subject to variation. However, this example shows how stability can be improved without sacrificing much of NF_{MIN}. For the circuit of Figure 10.12-4, NF_{MIN} is only 0.5 dB. The new value of Γ_{OPT} for this network is $0.785\angle 41.4°$. The corresponding impedance, $Z_{OPT} = (43.9 + j118.4)\ \Omega$, and the constant NF circles are shown in Figure 10.12-5.

It can be seen from Figure 10.12-5 that the Γ_{OPT} and Γ_{1M} source reflections are fairly close to each other. Nevertheless, it would require a sacrifice of 0.5 dB in noise figure to match to Γ_{1M}; therefore, we will transform the 50-Ω source to Z_{OPT}. This is done by first transforming 50 Ω to the required 43.9 Ω using an impedance inverter of 46.85 Ω, and then adding a series inductor of 9.43 nH. With this change in the input circuit, the input impedance looking into the output is $Z_{IN2} = (162.1 - j75.54)\ \Omega$. To transform the 50-$\Omega$ load to the complex conjugate of Z_{IN2}, we first transform 50 to 162.1 Ω using an impedance inverter of 90° Ω and add an inductor of 5.86 nH. The final low-noise amplifier circuit and performance are shown in Figure 10.12-6. The completed low-noise amplifier circuit has 13.1 dB of gain and a noise figure of 0.5 dB at 2 GHz.
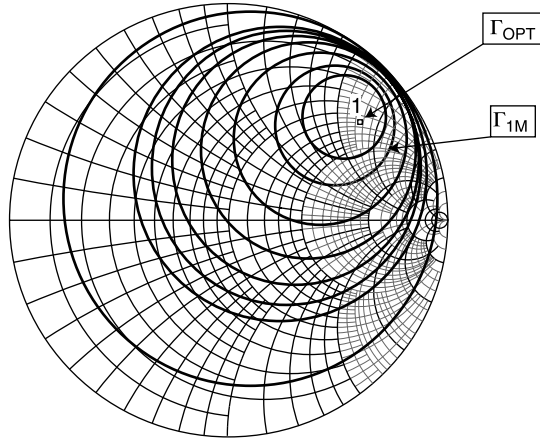
Figure 10.12-5   Constant NF circles for the amplifier in Figure 10.12-4. Circles are for $-0.25, -0.5, -1, -1.5, -2, -2.5, -3,$ *and* -6 *dB* below $NF_{MIN} = 0.5$ dB.

10.13   AMPLIFIER NONLINEARITY

Gain Saturation

There is no widely accepted definition for the linearity of a network. Indeed, the appropriate definition of nonlinear effects depends upon the network characteristics of concern. For an amplifier, one might wish to know at what power level the output ceases to increase in direct proportion to the input signal. For this the *1 dB compression* definition is useful. When a voltage $v(t)$ is applied to the input of an amplifier, an output voltage $v_0(t)$ appears at the output of the amplifier given by [2, p. 246]

$$v_0(t) = a_0 + a_1 v(t) + a_2 v^2(t) + a_3 v^3(t) + \cdots + a_n v^n(t) \qquad (10.13\text{-}1)$$

When the input voltage is small (10.13-1) can be approximated as

$$v_0(t) \approx a_1 v(t) \qquad (10.13\text{-}2)$$

and the amplifier's output is very nearly a constant times the input voltage, with voltage gain a_1. As the input voltage is increased, the remaining terms in (10.13-1) become significant and the amplifier displays a nonlinear response. When the coefficients a_i in (10.13-1) are evaluated for practical amplifiers, it is found that the sign of a_2 is negative, and this is the term responsible for saturation and the 1-dB compression.

   The 1-dB compression point, $P_{\text{1-dB } C}$, *is defined as the output power at which the output power of the network is 1 dB less than it would have been had its input to output characteristic remained linear* (Fig. 10.13-1). This definition can be

NE67300
DATAFILE: C:\NE67300.txt
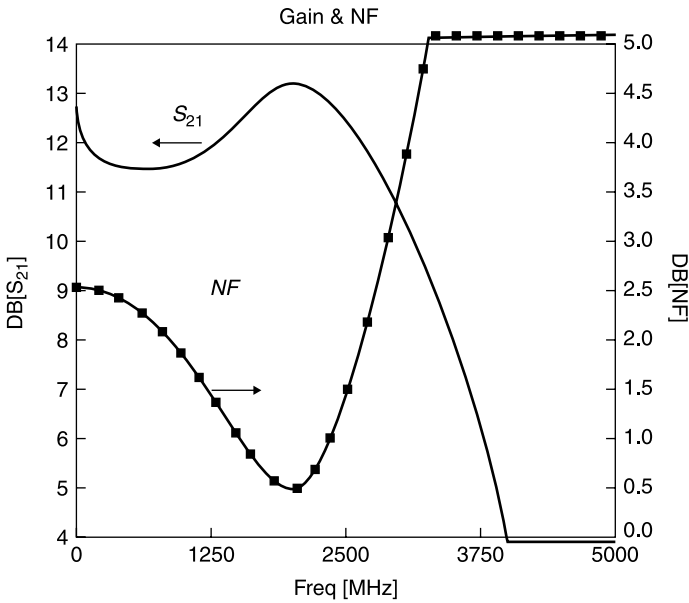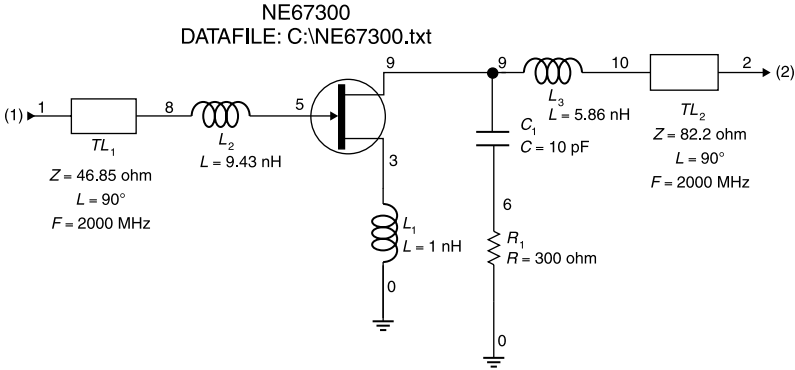
Figure 10.12-6 Final low-noise amplifier and its performance.

applied to any network. It is useful to determine up to what power levels do the performance parameters of the amplifier designs presented thus far remain reasonably linear. Beyond the 1-dB compression point, as input power to the amplifier is further increased, the output power reaches a maximum *saturated power output*, P_{SAT}.

Intermodulation Distortion

Well below the 1-dB compression point the amplifier's nonlinearities generate harmonics of the input signal as well as mixing products if two or more input
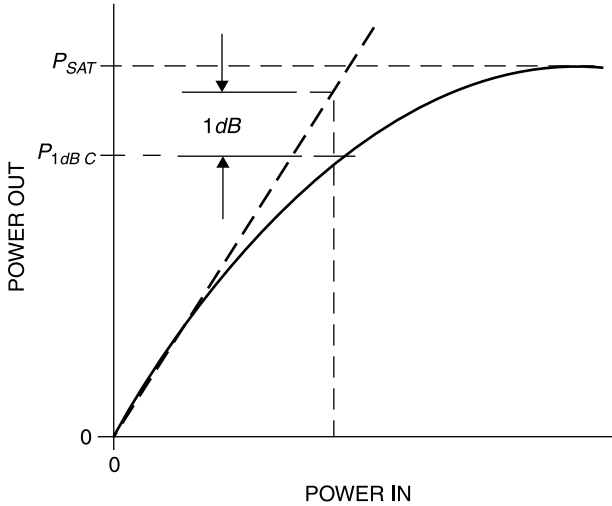
Figure 10.13-1   The 1-dB compression point, $P_{1\text{-dB }C}$.

signals are applied simultaneously. Simple harmonics of the input frequency, f_0, occur at $2f_0, 3f_0, 4f_0$, and so forth. Since these are multiples of the input frequency and since typical communication systems have bandwidths of no more that 10 or 20%, these harmonics can be filtered from the output. Hence, they are of minimal concern. However, intermodulation terms cannot be so disregarded.

The *two-tone test* consists of applying two closely spaced, equal amplitude frequencies, f_1 and f_2, to a nonlinear network. The output of the network contains frequency components at DC, $f_1, f_2, 2f_1, 2f_2, (2f_1 - f_2), (2f_2 - f_1), 3f_1$, $3f_2$, and so forth, an infinite number of these multiplying and mixing products. However, the products $(2f_1 - f_2)$ and $(2f_2 - f_1)$ are the most troublesome because they occur within the bandwidth of the system (Fig. 10.13-2) and cannot be removed by filtering. Other mixing products of higher order also appear in the passband but, because of their higher orders, have less amplitudes than the third-order products $(2f_1 - f_2)$ and $(2f_2 - f_1)$.

In the language of mixing, the fundamental applied frequencies f_1 and f_2 are termed *first-order* signals. In this nomenclature the frequencies $(2f_1 - f_2)$ and $(2f_2 - f_1)$ are *third-order mixing products*. This is because their generation requires that one of the input frequencies be doubled (a *second-order* process) and then that frequency mixed with the other input frequency (a third-order process). Accordingly, the solution to avoiding intermodulation products is to operate the network at a sufficiently low input power that third-order products have negligible amplitudes, that is, amplitudes at the level of the background noise or *noise floor*.

To identify this power range, we use the fact that, as input power is increased, at some input threshold *harmonics of the fundamental appear and then*
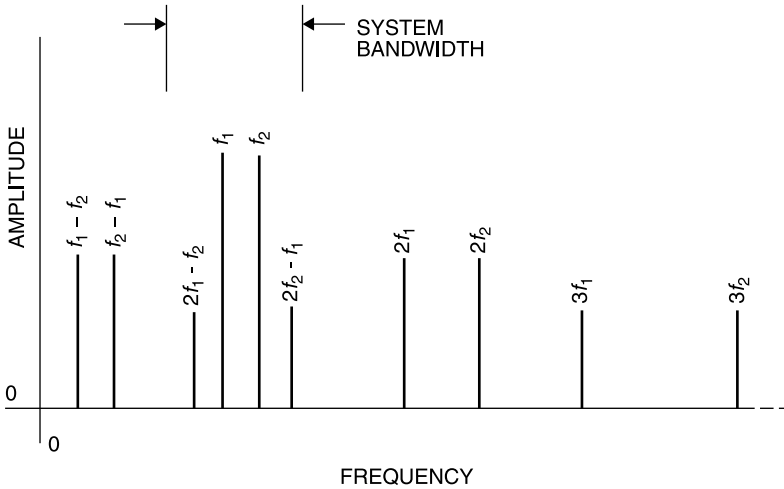
Figure 10.13-2   Frequency spectrum showing principal multiplying and mixing products when two closely spaced frequencies, f_1 and f_2, are applied to a nonlinear network (the *two-tone test*).

increase at a slope equal to their respective harmonic numbers (Fig. 10.13-3). To measure this effect a spectrum analyzer is connected to the output of the network. As input power at f_0 reaches a power level P_A the second harmonic, $2f_0$, is seen to rise above the noise level. Thereafter, for each 1-dB increase in the level of f_0, the second harmonic, $2f_0$, increases by 2 dB. With a further increase of the input signal level to P_B, it is found that the third harmonic, $3f_0$, appears above the noise and thereafter grows by 3 dB for each 1-dB increase in the level of input power at f_0. In short, the *harmonics grow with a slope equal to their harmonic number*.

The levels P_A and P_B are different for different networks, but the slopes at which the harmonics increase are analytically and experimentally found to be nearly equal to the harmonic number. A method for determining the power range over which noticeable third-order distortion occurs is based on these observations. The *third-order intercept*, P_3, is the output power level at which the extended third-order harmonic slope meets that of the fundamental. At this output power the ratio of the third-order harmonic to the fundamental theoretically would be 0 dB. Descending from this power level, for each 1-dB decrease in the power of f_0, all third-order products decrease by 3 dB.

Of course, operation at P_3 is impossible since the network's output power usually saturates below this level. In fact, it can be shown both experimentally and analytically using the first three terms in (10.13-1) that the third-order intercept is approximately 10 dB above the 1-dB compression point [1, p. 363]. However, P_3 *serves as a convenient power reference for derating the network to a level that reduces the third-order product to a negligible value*. Due to their
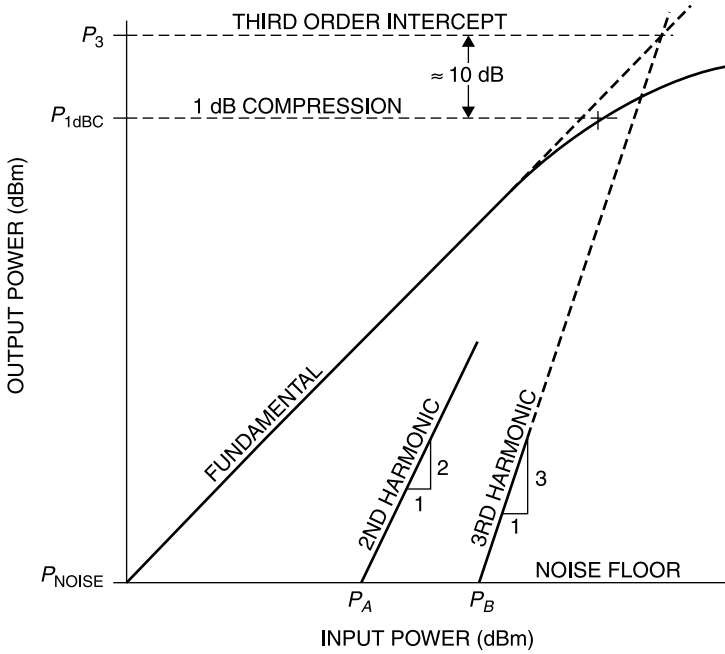
Figure 10.13-3   Growth of harmonics as a function of input power.

3-to-1 ratio, reducing the input power by one-third of the distance in decibels from the noise floor to P_3 causes the third harmonic to drop to the noise level (Fig. 10.13-4). The corresponding *output power range* for f_0 is called [1, Sec. 4.7] the *spurious-free dynamic range (SFDR)* and given by

$$\text{SFDR (dB)} = \tfrac{2}{3}[P_3 - P_{\text{NOISE}}]$$
$$= \tfrac{2}{3}[P_3 + 174 \text{ dBm} - 10 \log \text{BW (dB)}$$
$$- \text{NF (dB)} - G \text{ (dB)}] \qquad (10.13\text{-}3)$$

where

$$P_3 = \text{third-order intercept (dBm)}$$
$$\text{BW} = \text{amplifier bandwidth (Hz)}$$
$$\text{NF} = \text{amplifier noise figure (dB)}$$
$$G = \text{amplifier gain (dB)}$$

For example, an amplifier having 25 dB of gain, a noise figure of 4 dB, a bandwidth of 10 MHz, and a 1-dB compression point at 1 W has a SFDR
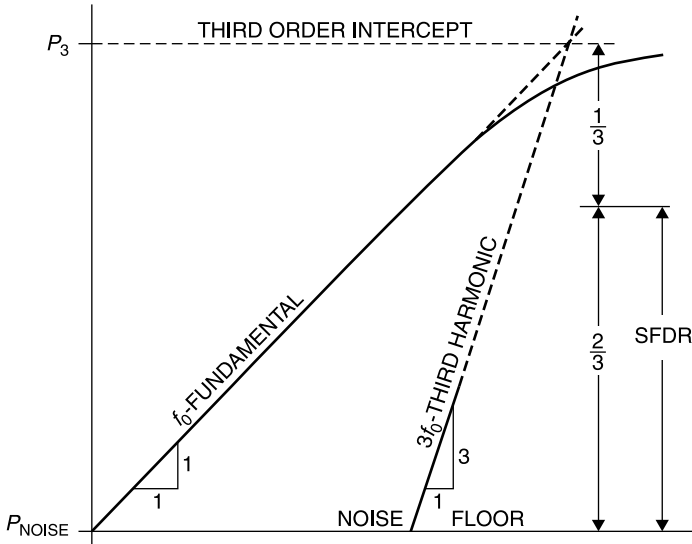
Figure 10.13-4   Determining the *spurious-free dynamic range (SFDR)*.

calculated as

$$\text{SFDR (dB)} = \tfrac{2}{3}[(30 \text{ dBm} + 10 \text{ dB}) + 174 \text{ dBm} - 10 \log 10^7 - 4 \text{ dB} - 25 \text{ dB}]$$
$$= 77 \text{ dB}$$

Sometimes it is desirable to take into account the fact that the minimum detectable signal S_{MIN} is somewhat above the noise level, and to adjust the useful dynamic range of the amplifier accordingly [2, p. 363]. To distinguish this range from that defined in (10.13-3) we define the *useful dynamic range (UDR)* as

$$\text{UDR (dB)} = \tfrac{2}{3}[P_3 + 174 \text{ dBm} - 10 \log \text{BW (dB)}$$
$$- \text{NF (dB)} - G \text{ (dB)} - S_{\text{MIN}} \text{ (dBm)}] \qquad (10.13\text{-}4)$$

where

$$S_{\text{MIN}} = \text{minimum detectable signal (dBm)}$$

It can be seen that the useful dynamic range is smaller than the SFDR by an amount equal to the minimum detectable signal.

10.14   BROADBANDING WITH FEEDBACK

In general, the S_{21} gain of a transistor falls off with frequency. If we want an amplifier which has fairly constant gain over a broad bandwidth, a *broadband*

amplifier, then we must arrange the transistor's input and output circuits to favor the high end of the band and, possibly, to increase the mismatch at the low end of the band.

There is no closed-form solution to this task. Some very clever engineering has been applied to the problem, but ultimately a certain amount of creativity and guesswork is required to obtain good results. This is a perfect application for computerized assisted guessing, customarily called *optimization*. Usually, one stabilizes the transistor first. Then, to create a design starting point, the unilateral design is used, neglecting the feedback term S_{12}. Then matching networks are designed using optimization to tailor the input and output gains to obtain a broadband result.

We will try a different tack by using feedback. This example uses the HP AT415868 bipolar transistor, having the S parameter data shown in Table 10.14-1.

It is a good idea to look at the file to be sure there are no typographical errors in it (there is one here, a missing decimal point for the 700-MHz data). Customarily, there is no data for zero frequency since the network analyzer used for measurements does not operate at DC. The network simulator will interpolate the data to zero frequency, but this is a mathematical interpolation.

TABLE 10.14-1   *S* Parameter File for HP AT 415868 Transistor

!AT-41586
!S-PARAMETERS at Vce = 8 V Ic = 25 mA. LAST UPDATED 08-03-92

ghz s ma r 50

0.100	0.64	−61	39.4	154	0.014	64	0.82	−24
0.200	0.59	−101	28.7	169	0.022	53	0.64	−35
0.300	0.56	−125	21.4	124	0.026	49	0.53	−38
0.400	0.55	−140	17.0	111	0.030	49	0.47	−39
0.500	0.54	−151	14.0	104	0.033	50	0.43	−38
0.600	0.54	−159	11.7	97	0.036	52	0.40	−38
0.700	0.54	−166	10.1	91	0.039	53	0.40	−37
0.800	0.54	−171	8.9	86	0.042	55	0.38	−37
0.900	0.54	−176	7.9	81	0.045	56	0.37	−37
1.000	0.55	177	7.2	77	0.048	57	0.36	−37
1.500	0.57	164	4.8	64	0.064	59	0.34	−42
2.000	0.57	152	3.6	55	0.080	57	0.32	−49
2.500	0.60	141	2.9	44	0.100	55	0.31	−58
3.000	0.62	132	2.4	34	0.120	52	0.31	−68
3.500	0.64	124	2.1	24	0.140	49	0.31	−80
4.000	0.67	116	1.9	18	0.180	45	0.32	−94
4.500	0.70	109	1.6	9	0.160	45	0.30	−109
5.000	0.73	102	1.5	1	0.170	42	0.30	−123
5.500	0.77	96	1.3	−7	0.190	38	0.32	−138
6.000	0.76	90	1.2	−14	0.200	33	0.35	−152

TABLE 10.14-2   Edited S Parameter File for the HP AT41586

!AT-41586
!S-PARAMETERS at Vce = 8 V Ic = 25 mA. LAST UPDATED 08-03-92

ghz s ma r 50

0.000	0.70	−40	45	180	0.010	75	0.90	−10
0.100	0.64	−61	39.4	154	0.014	64	0.82	−24
0.200	0.59	−101	28.7	169	0.022	53	0.64	−35
0.300	0.56	−125	21.4	124	0.026	49	0.53	−38
0.400	0.55	−140	17.0	111	0.030	49	0.47	−39
0.500	0.54	−151	14.0	104	0.033	50	0.43	−38
0.600	0.54	−159	11.7	97	0.036	52	0.40	−38
0.700	0.54	−166	10.1	91	0.039	53	0.40	−37
0.800	0.54	−171	8.9	86	0.042	55	0.38	−37
0.900	0.54	−176	7.9	81	0.045	56	0.37	−37
1.000	0.55	177	7.2	77	0.048	57	0.36	−37
1.500	0.57	164	4.8	64	0.064	59	0.34	−42
2.000	0.57	152	3.6	55	0.080	57	0.32	−49
2.500	0.60	141	2.9	44	0.100	55	0.31	−58
3.000	0.62	132	2.4	34	0.120	52	0.31	−68
3.500	0.64	124	2.1	24	0.140	49	0.31	−80
4.000	0.67	116	1.9	18	0.180	45	0.32	−94
4.500	0.70	109	1.6	9	0.160	45	0.30	−109
5.000	0.73	102	1.5	1	0.170	42	0.30	−123
5.500	0.77	96	1.3	−7	0.190	38	0.32	−138
6.000	0.76	90	1.2	−14	0.200	33	0.35	−152

We know that the phase angle for zero frequency is 180° (perfect negative feedback), so it is best to modify the file accordingly. The corrected and revised file is shown in Table 10.14-2. With these parameters the calculated stability data are shown in Figure 10.14-1. The transistor is potentially unstable over most of the 0 to 6000 MHz band; however, K is nearly equal to unity over much of the band.

The input and output stability circles for 1 GHz are shown in Figure 10.14-2. The device is nearly stable at 1 GHz except for very low impedances presented to the input circuit. Therefore, we will add a low value of resistance in series with the base lead.

Examining the S_{21} parameter in Table 10.14-2 reveals that the gain with 50-Ω source and load varies from 31.9 dB at 100 MHz to 17.1 dB at 1000 MHz. Suppose that we wish the amplifier to have a flat gain characteristic from 50 to 1050 MHz. This bandwidth encompasses the entire VHF and UHF television transmission bands. To design the amplifier we will use negative feedback, more feedback at low frequencies than at high frequencies.

Feedback from the collector to base is negative. To favor high frequencies we will interconnect these leads with a series RL circuit. An impedance in the
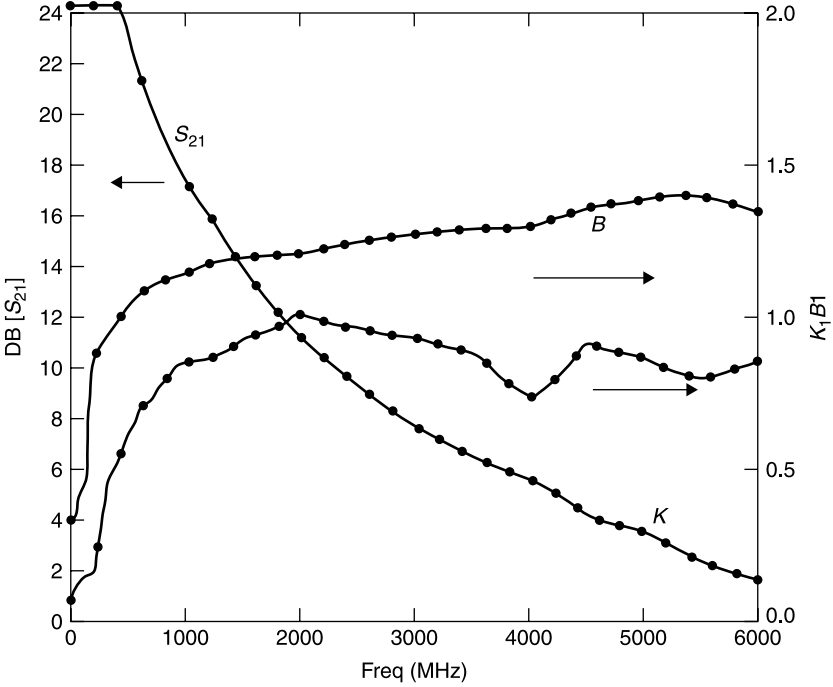
Figure 10.14-1   Stability and gain of the HP AT415868 bipolar transistor.

emitter to ground path also provides negative feedback. To favor high fre-
quencies we will use a parallel *RC* network there. Using this circuit with nom-
inal values, the result is that shown in Figure 10.14-3. Prior to optimization the
performance does not look promising with respect to the design goals.

   Next the circuit of Figure 10.14-3 was optimized using the network simula-
tor, with performance goals of $\text{Mag}[S_{21}] = 15$ dB, $K > 1.1$ and $B_1 > 0.1$. The
question marks ahead of the element values indicate that the optimizer is
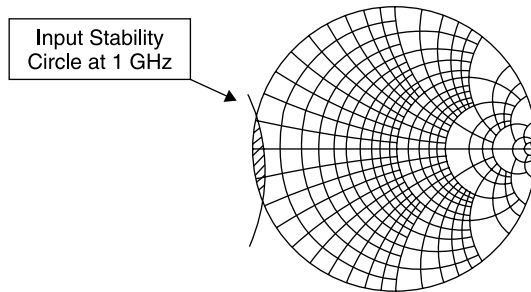
Figure 10.14-2   Input stability circle for HP AT415 at 1 GHz (output stability circle
does not intersect unity Smith chart).
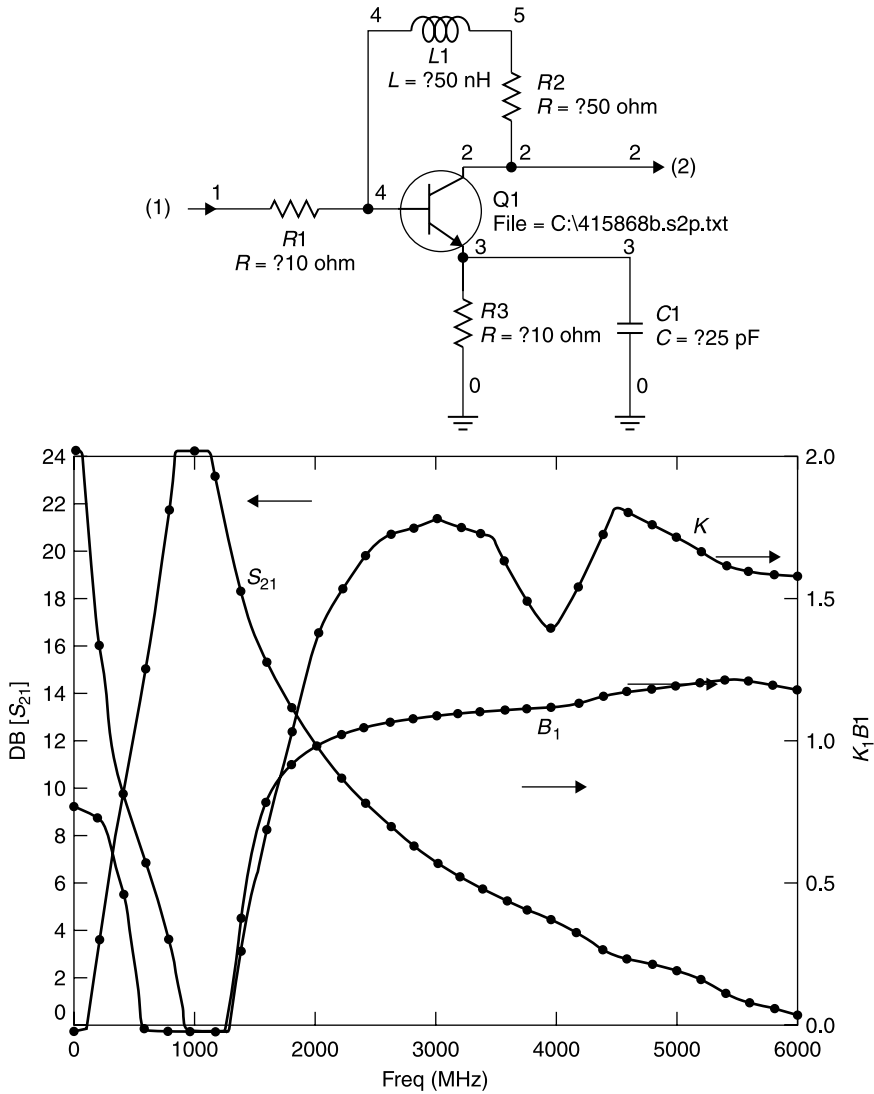
Figure 10.14-3   Amplifier circuit and performance with feedback element values prior to optimization.

allowed to vary that value. Relative weightings (W) of the errors used in the simulator optimization were unity for all variables except K, for which $W = 10$. This causes an error of *0.1* in K is to be considered equally with a gain error of *1* dB by the optimizer. Note that the *15* dB goal for the gain was set below the minimum gain of *17.1* dB established from the S parameter data.

Starting with nominal initial component values shown in Figure 10.14-3, the
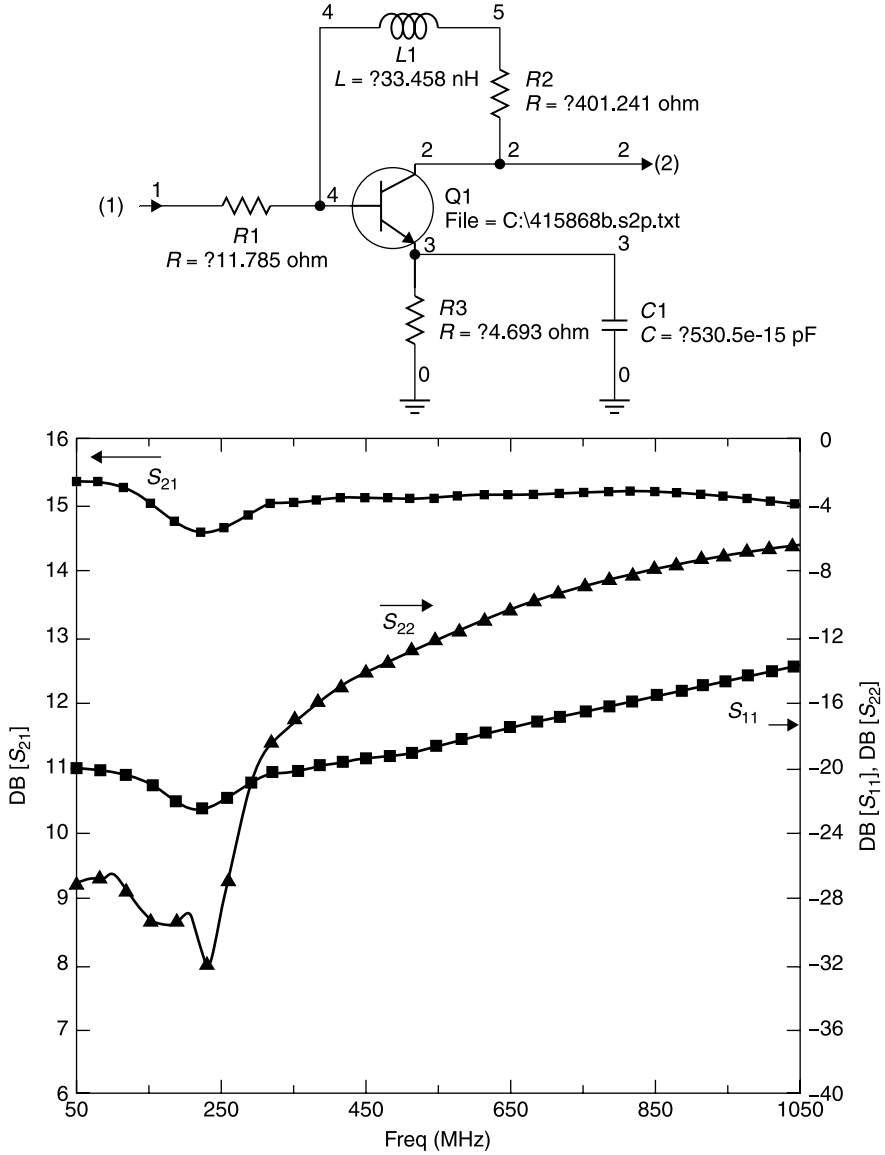
Figure 10.14-4 Final circuit and performance of the 50- to 1000-MHz amplifier using feedback and the 415858b transistor.

network simulator optimized the circuit, and the results are shown in Figure 10.14-4. The capacitance C_1 can be eliminated because the optimized value is near zero. The gain is within 0.4 dB of 15 dB over the 50 to 1050 MHz band. The stability factor K was 1.1 or greater and B_1 was greater than 0.1 over the whole gainful range of the transistor, from 0 to 6000 MHz. Probably less gain

variation and/or a wider bandwidth could be obtained, but this result is remarkable for the fact that only four components (not counting C_1) were used to achieve the result and their function had to include unconditionally stabilizing the transistor over the 0 to 6000 MHz frequency range.

In an initial optimization trial a gain goal of 16 dB was set. This gain was obtained with even smaller variation with frequency, but the resulting amplifier was potentially unstable $(K < 1)$ over most of the 0 to 6000 MHz bandwidth. Reducing the gain goal to 15 dB did not yield a stable amplifier until the weighting of K was increased from 1 to 10, emphasizing the importance of this goal. From this it can be seen that fairly narrow margins exist around the achievable goals with a particular transistor and its circuit. Setting goals within the margins is a crucial part of the optimization.

10.15 CASCADING AMPLIFIER STAGES

When two amplifiers having the circuit of Figure 10.14-4 are cascaded, their gains add without very much mismatch error, and the resulting gain is within 1 dB of 30 dB from 50 to 1100 MHz, as shown in Figure 10.15-1. This is unusual and occurs because the S_{11} and S_{22} magnitudes (Fig. 10.15-2) of each amplifier section are fairly small, especially S_{11}. When either the input or output of a single-stage amplifier is well matched, the cascade combination of two of them will have low mismatch interaction. The feedback method of broadbanding provides a better match than does broadbanding by reflecting unwanted gain.

For example, suppose that we had applied the same set of goals to optimize the circuit shown in Figure 10.15-2. Its performance after optimization is also shown in Figure 10.15-2.

This circuit offers nearly the same gain flatness over the *50 to 1050 MHz* bandwidth and also is unconditionally stable from *0 to 6000 MHz*. However, the input and output mismatches, S_{11} and S_{22}, of this circuit are greater than those of the feedback optimized circuit of Figure 10.15-1; consequently, when two of these stages are cascaded together, as shown in Figure 10.15-3, a larger variation of gain, *almost ± 3 dB over the 50 to 1050 MHz* band, occurs due to the *mismatch error* (Section 4.8).

Designing broadband amplifiers and cascaded amplifier stages does require guesswork, or optimizing. However, certain strategies are useful in this pursuit. First, since transistor gain diminishes as frequency increases, circuitry that favors the passage of higher frequencies (series capacitors and shunt inductors) is more likely to give uniform gain with frequency.

Circuits which flatten gain using feedback may provide lower S_{11} and S_{22} values, minimizing the mismatch error when the circuits are cascaded or used with other reflective components, such as reactive filters. Finally, S_{11} and S_{22} can be made part of the optimization goals. If either S_{11} or S_{22} has a small value, then two such circuits cascaded will have a relatively low mismatch error.
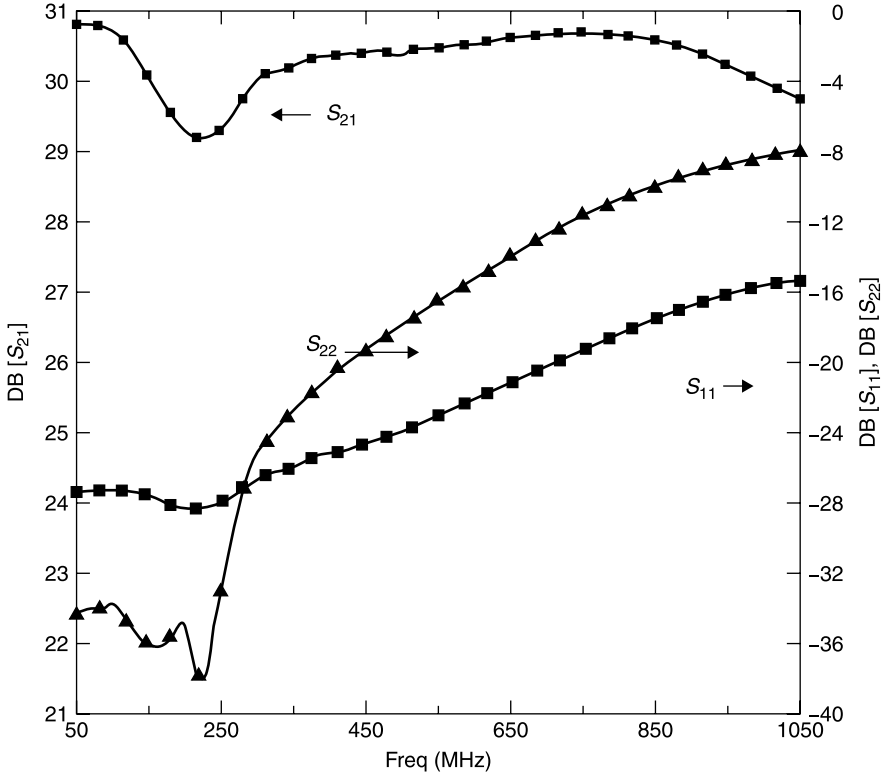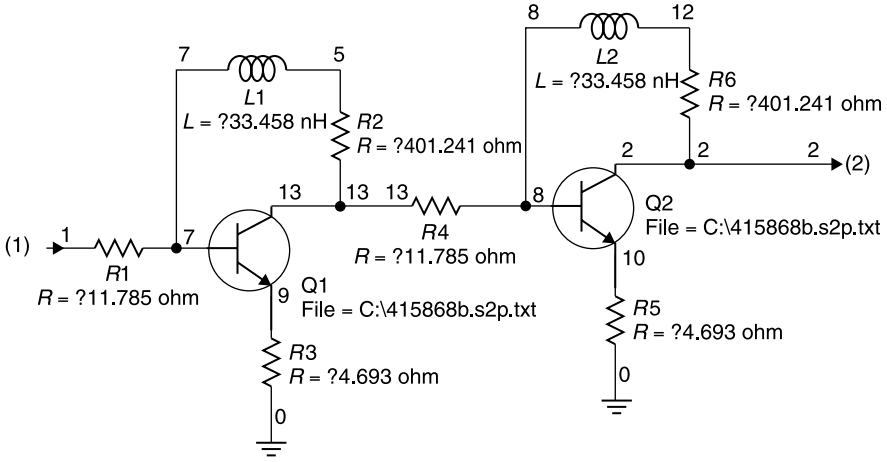
Figure 10.15-1   Two amplifiers cascaded.
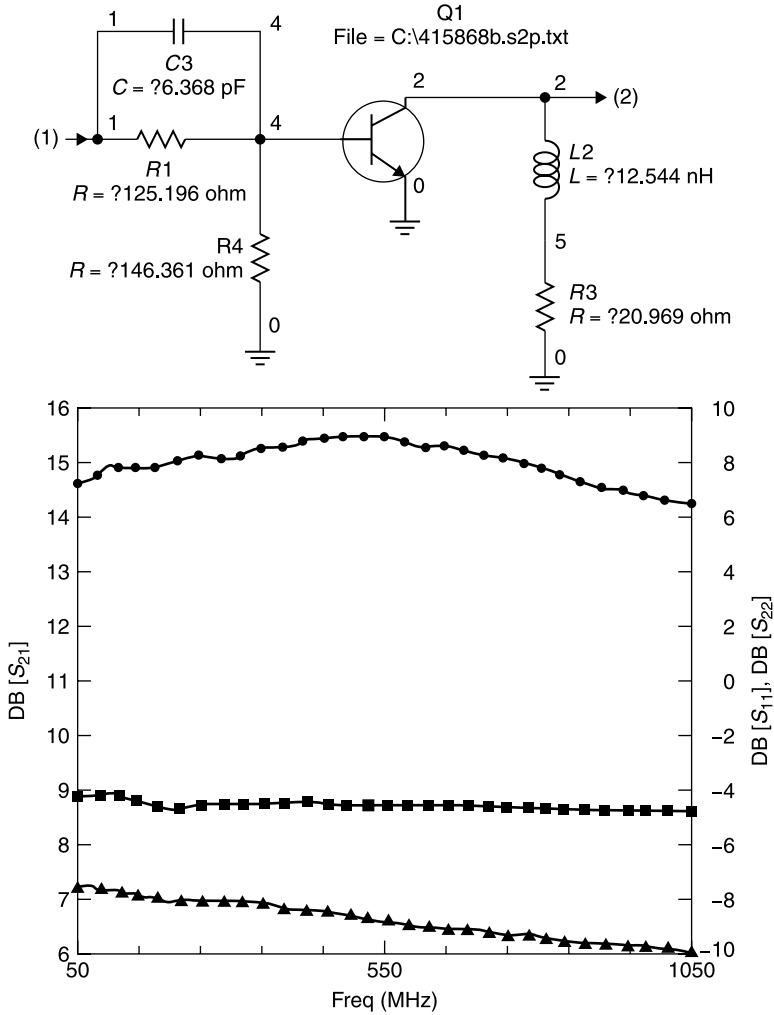
Figure 10.15-2   Alternate 50- to 1050-MHz amplifier design using the 415858b transistor.

10.16   AMPLIFIER DESIGN SUMMARY

For most amplifier designs the choices amount to what input and output impedance environments will be presented to the transistor over all of the frequency range for which it has gain. One must balance the need for gain against the requirements for unconditional stability since an amplifier that breaks into oscillation is not only useless, it is a liability. Inevitably, the design choices lead to amplifier networks that do not present a match to either the input or output
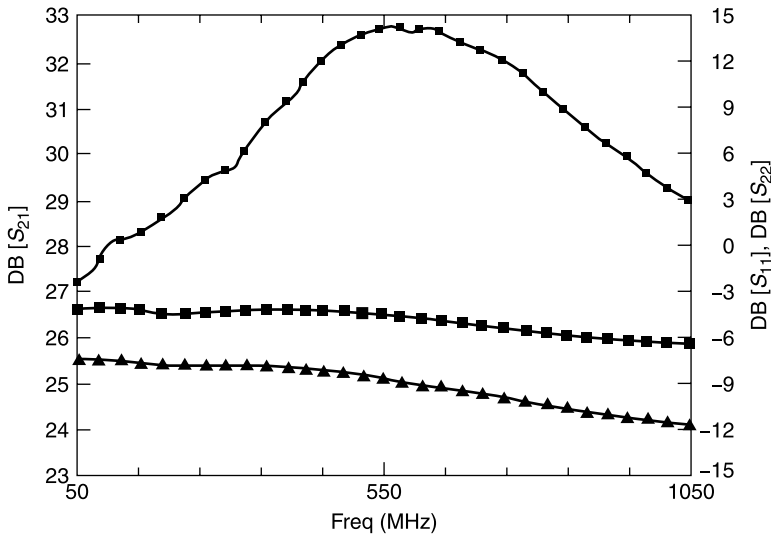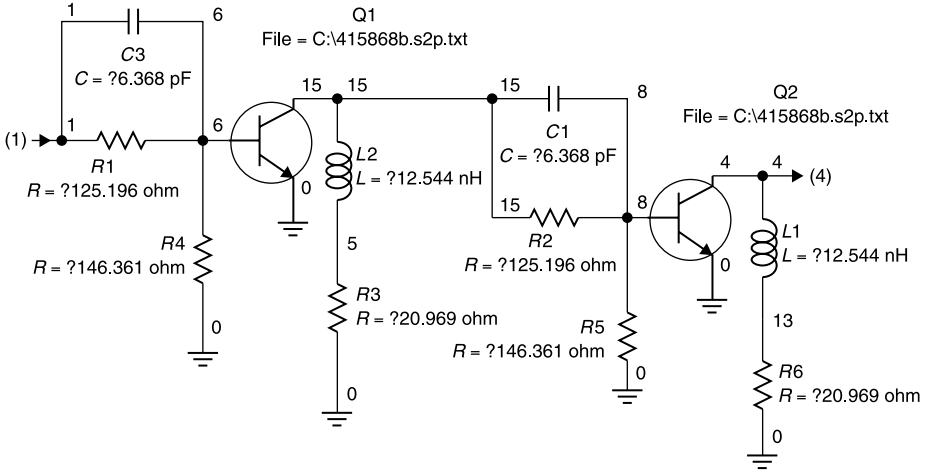
Figure 10.15-3   Gain of the cascade combination of two of the circuit shown in Figure 10.15-2.

transmission lines. If the remainder of the system also has high mismatches, such as would be true if reactive filters are connected to the amplifier, the performance will not be predictable unless those networks are modeled along with the amplifier. Many points have been covered in this chapter, but in the end some artistry is required of the amplifier designer to balance all of the desired amplifier attributes with the relatively few degrees of design freedom available.

REFERENCES

1. Guillermo Gonzalez, *Microwave Transistor Amplifiers, Analysis and Design*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, 1984. *An excellent engineering reference for transistor amplifier design.*

2. Theodore Grosch, *Small Signal Microwave Amplifier Design*, Noble Publishing, Norcross, GA, 1999. *The author presents a thorough derivation of all of the important formulas for transistor amplifier design and evaluation, based on the S parameters of the transistor and the circuit, which surrounds it.*

3. *Genesys 7 Reference Manual*, Eagleware, Norcross, GA, 1986–2000. *This is the reference manual for the CAD program used as a network simulator for the examples of this text. Note that early versions of this manual have an error in the equation for B1 on page 324. Use the corrected expression (10.3-5) given in this text.*

4. Joseph F. White, "What is CDMA?" *Applied Microwave & Wireless*, Fall, 1993, pp. 5–8. *A simplified presentation of the CDMA format.*

5. *Reference Data for Radio Engineers*, 5th ed., Howard W. Sams & Co., New York, 1974. *Newer editions are available. This is a classic reference for useful radio data.*

6. California Eastern Laboratories (CEL), *RF and Microwave Semiconductors (*the CD*), Selection Guide and Design Parameter Library*, CEL, Santa Clara, CA, 1999–2000. *This is the CD of the CEL catalog and contains data for thousands of transistors, both bipolar and FET. In addition, there are very well written application notes describing transistor amplifier design. All of this is provided free of charge.*

EXERCISES

E10.3-1 **a.** Using the S parameter table below, calculate the K and B_1 factors for the 2N6603A transistor at 1 GHz.

b. What can you conclude from these values?

Freq (GHz)	S_{11}		S_{21}		S_{12}		S_{22}	
0.1	0.69	−30	12.16	160	0.03	72	0.95	16
0.2	0.65	−61	11.03	143	0.05	59	0.84	31
0.5	0.63	−122	7.05	111	0.07	36	0.56	54
1	0.64	−158	4.13	88	0.09	28	0.39	68
2	0.65	170	2.14	61	0.11	29	0.33	91

E10.5-1 **a.** Calculate the unilateral gain figure of merit for the 2N6603A transistor.

b. What does this figure of merit indicate about the accuracy of estimating the gain of the transistor as $20 \log|S_{21}|$?

E10.5-2 **a.** Estimate the $G_S, G_0,$ and G_L gains at 1000 MHz of the 2N6603A transistor based on its S parameters.

b. What is the total estimated unilateral gain if source and load are transformed to S_{11}^* and S_{21}^*, respectively?

E10.5-3   **a.** Verify the G_0 gain for the 2N6603A using a network simulator and plot gain, K and B_1 over the frequency of the S parameters.

b. Match the input at *1 GHz* to the untuned transistor's Z_{IN}. By how much did the gain increase at 1 GHz?

c. Next add matching to the output, matching to the untuned transistor's Z_{OUT}.

d. How well matched are the resulting S_{11} and S_{22}?

e. How close is the total gain at 1 GHz to the sum $G_S + G_0 + G_L$?

f. Is the total gain "error" within the bounds predicted by the unilateral figure of merit calculated in E10.5-1?

E10.7-1   **a.** Use the network simulator to determine the input and output impedances Z_{SM} *and* Z_{LM} to be presented to the 2N6603A for a simultaneous conjugate match at 1 GHz.

b. Would the simulator provide the required Z_{SM} *and* Z_{LM} impedances? Interpret this result.

E10.7-2   **a.** Examine the input stability of the 2N6603A by plotting the input stability circle on the Smith chart (use a network simulator to do this if available) and determine a series resistance value at the input, which will unconditionally stabilize the transistor at 1000 MHz.

b. Using this circuit find the Z_{SM} *and* Z_{LM} impedances.

c. Add the matching circuits for Z_{SM} *and* Z_{LM} to the stabilized schematic.

d. Plot the gain and S_{11} and S_{22} for the conjugately matched circuit.

e. Is this what you expected?

f. What performance compromise have you made?

g. Is your design unconditionally stable at all frequencies?

E10.7-3   **a.** Stabilize the 2N6603A transistor at all frequencies for which the transistor has gain, including DC. Modify the S parameter file to correct the DC S parameters so that the argument of S_{21} is 180° and its magnitude (β) is appropriate to DC. Extrapolate its β from the RF S_{21} parameters.

b. Recalculate Z_{SM} and Z_{LM} at 1 GHz and redesign the amplifier to use these values.

c. What is your final performance?

d. What performance compromise have you made for unconditional stability at all frequencies?

e. Was your stabilizing technique the most efficient possible?

E10.9   **a.** Design an amplifier using the operating gain method and the Motorola 2N66l79A bipolar transistor to have 20 dB of gain at 1 GHz. The S parameters are listed in Table 10.1-1. For this exercise, the circuit need only be stable at 1 GHz using the source and load impedances you employ to obtain the 20 dB of gain.

      b. Plot S_{21}, S_{11}, and S_{22} of the resulting circuit over the bandwidth for which S parameters are given.

E10.12-1   **a.** Design a low-noise amplifier for use at 850 MHz using the HP AT30511 (S and noise parameters in table below). Obtain as much gain as you can, consistent with a noise figure of no more than 0.5 dB.

      b. Plot gain and noise figure from 600 to 1100 MHz.

!AT-30511								
!S and NOISE PARAMETERS at Vce $= 2.7$ V Ic $= 1$ mA.								
# ghz s ma r 50								
0.1	0.97	−5	3.50	175	0.01	86	0.999	−2
0.5	0.95	−23	3.38	156	0.05	74	0.96	−13
0.9	0.86	−39	3.20	139	0.08	63	0.93	−23
1	0.84	−43	3.10	135	0.08	60	0.92	−25
1.5	0.72	−63	2.80	115	0.11	49	0.85	−34
1.8	0.65	−73	2.60	105	0.12	43	0.82	−38
2	0.61	−80	2.53	99	0.13	40	0.79	−41
!Noise Parameters								
0.5	0.3	0.96	10	1.49				
0.9	0.4	0.92	19	1.33				
1.8	0.9	0.83	43	0.98				

E10.13-1   **a.** You are testing a newly designed amplifier that has a 1-dB noise figure and 20 dB of small signal gain using a spectrum analyzer at its output. The output bandwidth is limited by an integral filter to 800 to 1000 MHz. You observe that with a 900-MHz input signal of varying amplitude the third harmonic (at 2700 MHz) is just discernible near the noise level when the output power of the fundamental is 0.1 mW (−10 dBm). What is the third-order intercept for this amplifier–filter network?

      b. What is the spurious-free dynamic range (SFDR) of this network?

E10.14-1   **a.** Design a broadband amplifier to operate from 50 to 1000 MHz (the entire television band) using the 2N6603A transistor. Start using a circuit unconditionally stabilized over the 0 to 2000 MHz band. *Hint:* Use the stabilized transistor circuit developed in

E10.7-3 as a starting point. Add *L-C* networks to the input and output for matching and use an optimizer. Try for a flat 10 dB of gain from 50 to 1000 MHz and add the requirement that $K > 1.1$ and $B_1 > 0.1$ for *0 to 2000 MHz* (the whole range for which we have *S* parameters).

b. Plot the gain, K and B_1 over the 0 to 2000 MHz bandwidth.

E10.15-1 **a.** Cascade two copies of the broadband amplifier designed in E10.14-1 and plot the combined gain from 0 to 2000 MHz, combining the paralleled capacitors at the interconnection of the two stages.

b. Use the optimizer to design a 20-dB flat gain amplifier from 50 to 1000 MHz. Also require unconditional stability from 0 to 2000 MHz (the whole range for which we have *S* parameters).

c. How well matched is the cascade amplifier?

d. Could you have accomplished the same result but with better input and output VSWR? How would you alter the circuit to accomplish this?

e. Alter the circuit and see if you can obtain a maximum return loss at the input and output of -15 dB while maintaining stability and fairly flat 20-dB gain over 50 to 1000 MHz.

Symbols and Units

This appendix lists the principal symbols used in this book in two groups, roman and Greek characters. The listing may not include symbols used in brief descriptions and derivations or standard symbols such as e, j, π, and so forth. Where applicable the *standard international (SI) unit* is listed along with its abbreviation. The general notational conventions used in the text (to connote general or sinusoidal time variation, vector, and normalized quantities) as well as coordinate systems are described in Section 1.3.

Symbol	Quantity	Standard International Unit and Abbreviation
	Roman Characters	
$\vec{A}$	Three-dimensional vector $= A_x\vec{x} + A_y\vec{y} + A_z\vec{z}$	Dimensions of vector quantity (e.g., volts/meter for $\vec{E}$)
$\vec{A}$	Vector potential	
A, B, C, D	*ABCD* matrix elements	A and D are dimensionless, B is in ohms (Ω), C in mhos ($\mho$)
A_C	Antenna capture area	Meters^2 (m^2)
A_{EFF}	Antenna effective area	Meters^2 (m^2)
a_k	Incident wave at port k, used with S matrix	Watts$^{1/2}$ (W$^{1/2}$)
B	Magnetic flux density	Webers/meter^2 (W/m^2) or teslas (T)
B	AC susceptance	Mhos ($\mho$)
$\bar{b}$	Normalized susceptance	Dimensionless
b_k	Scattered wave from port k, used with S matrix	Watts$^{1/2}$ (W$^{1/2}$)
C	Capacitance	Farads (F)
c	Velocity of light in a vacuum	Meters/second (m/s)
D	Electric flux density	Coulombs/meter^2 (C/m^2)
d	Distance from load end of transmission line	Meters (m)

Symbol	Quantity	Standard International Unit and Abbreviation
E	Electric field intensity	Volts/meter (V/m)
e	Charge on the electron	Coulombs (C)
F	Force	Newtons (N)
f	AC frequency	Hertz (Hz)
f_C	Cutoff frequency	Hertz (Hz)
f_0	Design center frequency	Hertz (Hz)
f_R	Self-resonant frequency due to parasitics of a lumped R, L, or C	Hertz (Hz)
G	Conductance	Mhos (℧)
$\bar{g}$	Normalized conductance	Dimensionless
G	Gain	Dimensionless[a]
g_k	Normalized elements values of low-pass filter	Dimensionless
g_k'	Normalized elements of high-pass filter	Dimensionless
H	Magnetic field intensity	Amperes/meter (A/m)
I, I	Electric current	Amperes (A)
IL	Insertion loss or isolation	Dimensionless[a]
J, J	Current density	Amperes/meter^2 (A/m^2)
k	Wave number $= \omega\sqrt{\mu\varepsilon} = 2\pi/\lambda$	Meter^{-1} (m^{-1})
k_C	Coupling coefficient of a directional coupler	Dimensionless
k_T	Transmission coefficient of a directional coupler	Dimensionless
L	Inductance	Henry (or henries) (H)
l	Length, usually of a transmission line	Meters (m)
ln	Natural logarithm (base e)	Dimensionless
log	Common logarithm (base 10)	Dimensionless
M	Magnetization	Amperes/meter (A/m)
n	Transformer turns ratio	Dimensionless
P	Power	Watts (W)
P_A	Available generator power	Watts (W)
Q	Circuit quality	Dimensionless
Q_U, Q_L, Q_E	Unloaded, loaded, and external Qs of a resonator	Dimensionless
q	Electric charge	Coulombs (C)
R	Resistance	Ohms (Ω)
$\bar{r}$	Normalized resistance	Dimensionless
$\vec{r}, \vec{\theta}$, and $\vec{\varphi}$	Unit vectors in the r, θ, and φ directions of spherical coordinate system	Dimensionless
$\vec{r}, \vec{\varphi}$, and $\vec{z}$	Unit vectors in the r, φ, and z directions of cylindrical coordinate system	Dimensionless

Symbol	Quantity	Standard International Unit and Abbreviation
S	Surface area	Meters^2 (m^2)
SWR *or* VSWR	Standing-wave ratio	Dimensionless
S_{nn}, S_{mn}	Scattering coefficient at port n or between ports mn	Dimensionless
T	Period of an AC signal	Seconds (s)
T	Transmission coefficient of a network	Dimensionless
T_n	Chebyshev polynomial of order n	Dimensionless
$T_{11}, T_{12}, T_{21}, T_{22}$	Elements of the transmission matrix	Dimensionless
t	Time	Seconds (s)
$\tan \delta$	Dielectric loss tangent	Dimensionless
TL	Transducer loss (or gain)	Dimensionless[a]
U	Energy	Joules (J)
U_E	Stored electric energy	Joules (J)
U_M	Stored magnetic energy	Joules (J)
V	Electric potential or voltage	Volts (V)
v	Velocity	Meters/second (m/s)
v	EM wave velocity in an unbounded medium $(= c/\sqrt{\mu_R \varepsilon_R})$	Meters/second (m/s)
v_g	Group velocity	Meters/second (m/s)
v_p	Phase velocity	Meters/second (m/s)
X	AC reactance	Ohms (Ω)
$\vec{x}, \vec{y}, \vec{z}$	Unit vectors in the x, y, and z directions of rectangular coordinate system	Dimensionless
$\bar{x}$	Normalized reactance	Dimensionless
Y	AC admittance	Mhos ($\mho$)
Y_0	Transmission line characteristic admittance	Mhos ($\mho$)
y_{ij}	Y matrix element	Mhos ($\mho$)
Z	AC impedance	Ohms (Ω)
Z_{OE}	Even mode impedance of coupled line pair	Ohms (Ω)
Z_{OO}	Odd mode impedance of coupled line pair	Ohms (Ω)
Z_{TE}	Wave impedance of TE mode	Ohms (Ω)
Z_{TM}	Wave impedance of TM mode	Ohms (Ω)
z_{ij}	Z matrix element	Ohms (Ω)

Greek Symbols

α	Attenuation constant	Nepers/meter (Np/m) or decibels/meter (dB/m)

Symbol	Quantity	Standard International Unit and Abbreviation
α_C	Attenuation constant associated with conductors	Nepers/meter (Np/m) or decibels/meter (dB/m)
α_D	Attenuation constant associated with dielectrics	Nepers/meter (Np/m) or decibels/meter (dB/m)
β	Phase constant	Radian/meter (rad/m) or degrees/meter (°/m)
γ	Propagation constant	Meters^{-1} (m^{-1})
Γ	Reflection coefficient	Dimensionless
Γ_G	Generator reflection coefficient	Dimensionless
Γ_L	Load reflection coefficient	Dimensionless
δ_S	Skin depth	Meters (m)
ε_0	Permittivity of free space	Farads/meter (F/m)
ε_R	Relative permittivity factor or *dielectric constant*	Dimensionless
ε	Permittivity ($= \varepsilon_0 \varepsilon_R$)	Farads/meter (F/m)
η	Intrinsic impedance of an unbounded medium	Ohms (Ω)
θ	Electrical length ($= \beta l$) of a transmission line	Radians (rad) or degrees (°)
θ	Power factor angle (between voltage and current)	Radians (rad) or degrees (°)
λ_0	Wavelength in free space	Meters (m)
λ	Wavelength in an unbounded medium ($= \lambda_0/\sqrt{\mu_R \varepsilon_R}$)	Meters (m)
λ_C	Cutoff wavelength	Meters (m)
λ_G	Guide wavelength	Meters (m)
μ_0	Permeability of free space	Henries/meter (H/m)
μ_R	Relative permeability	Dimensionless
μ	Permeability ($= \mu_R \mu_0$)	Henries/meter (H/m)
ρ	Resistivity	Ohm-meters (Ω-m)
ρ	Charge density	Coulombs/meter^3 (C/m^3)
σ	Conductivity	Mhos/meter ($\mho$/m)
ω	Radian frequency	Radians/second (rad/s)

[a] The quantity is dimensionless, however, the numeric value may be expressed in decibels (dB).

COMPLEX MATHEMATICS

THE IMAGINARY NUMBER

Complex numbers, having a real and imaginary part, probably arose out of a purely mathematical exploration of the description of what would happen if there were a square root of -1. Since this seemed unreal, mathematicians called this the *imaginary number*, denoted as i, thus

$$i = \sqrt{-1} \tag{B.1}$$

Mathematicians evolved a complete set of operations about how *complex numbers* [1], having *real and imaginary parts*, would behave, probably without regard to any practical application such a number system might have.

This peculiar inquiry was to have a profound effect on electrical engineering. Many circuit applications involved finding the voltage and currents resulting when a sinusoidally varying excitation was applied to a network. When a sinusoid is applied to a linear network consisting of *RLC* elements, *all resulting voltages and currents are sinusoidal at the same frequency*. This meant that, at most, the analysis of a network involved finding the magnitude and phase of the resulting sinusoidal response. The rules established by mathematicians for complex numbers proved to be perfect for analyzing alternating current (AC) circuits.

Electrical engineers customarily use the letter i to represent electric current. Therefore, for electrical engineering purposes, the imaginary number was assigned the symbol j, thus

$$j = \sqrt{-1} \tag{B.2}$$

The mathematics of complex numbers is extensive [1], but for AC analysis, we need only consider the basic complex number functions: *addition*, *subtraction*, *multiplication*, and *division*, the *rectangular* and *polar* formats, and forming the *complex conjugate*.

First, a complex number can be expressed in *rectangular form* as

$$Z = a + jb \tag{B.3}$$

COMPLEX ADDITION

To add two complex numbers, use either rectangular form and add their real and imaginary parts together, respectively:

$$(a + jb) + (c + jd) = (a + c) + j(b + d) \tag{B.4}$$

For example:

$$(3 + j9) + (4 + j2) = 7 + j11$$
$$(3 + j9) + (4 - j2) = 7 + j7$$
$$(3 + j9) + (-4 + j13) = -1 + j22$$
$$(-3 - j9) + (4 - j5) = 1 - j14$$

COMPLEX SUBTRACTION

To subtract one complex number from another, use their rectangular form and subtract their real and imaginary parts, respectively:

$$(a + jb) - (c + jd) = (a - c) + j(b - d) \tag{B.5}$$

For example,

$$(3 + j9) - (4 + j2) = -1 + j7$$
$$(3 + j9) - (4 - j2) = -1 + j11$$
$$(3 + j9) - (-4 + j13) = 7 - j4$$
$$(-3 - j9) - (4 - j5) = -7 - j4$$

For multiplication or division, the complex number is best converted into *polar form*.

RECTANGULAR TO POLAR CONVERSION

The complex number has real and imaginary parts orthogonal to each other. It can be visualized as a vector. However, in electrical engineering electric and
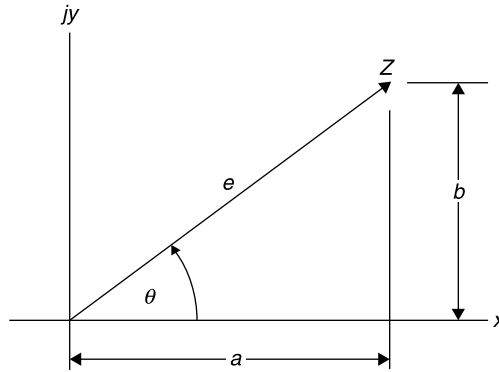
Figure B.1   Vector representation of a complex number Z in the complex plane.

magnetic fields are vector quantities. In order to differentiate complex quantities representing voltage, current, or field magnitudes from spatial field vectors, the *Institute of Electrical and Electronics Engineers (IEEE)* suggests that the complex number representing a sinusoidal quantity be termed a *phasor*. This will be the special term used to describe the two-dimensional complex number. Think of the complex number as having a *real part x along the horizontal axis* and an *imaginary part of magnitude y along the j axis* (Fig. B.1).

Let

$$Z = x + jy = a + jb \tag{B.6}$$

Since, in this case, $x = a$ and $y = b$, and applying the Pythagorean formula, the *polar form* of Z is $Z = \rho \angle \theta$, this is formed as

$$Z = a + jb = \sqrt{(a^2 + b^2)} \angle \tan^{-1} \frac{b}{a} = \rho \angle \theta \tag{B.7}$$

where ρ *is the magnitude* and θ *is the angle* of Z. For example,

$$Z = 4 + j3 = \sqrt{(4^2 + 3^2)} \angle \tan^{-1} \tfrac{3}{4} = 5 \angle 37° \tag{B.8}$$

POLAR TO RECTANGULAR CONVERSION

To convert from polar to rectangular form (Fig. B.2):

$$Z = \rho \angle \theta = (\rho \cos \theta) + j(\rho \sin \theta) \tag{B.9}$$

For example,

$$Z = 5 \angle 37° = (5 \cos 37°) + j(5 \sin 37°) = 4 + j3 \tag{B.10}$$
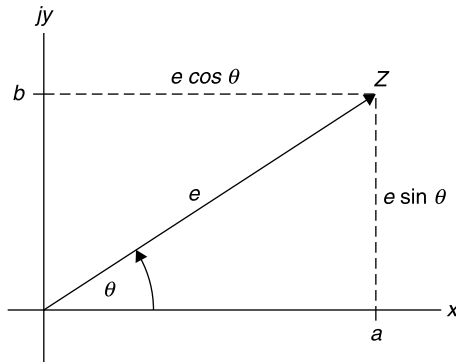
Figure B.2   Conversion between polar and rectangular forms of a complex number, Z.

COMPLEX MULTIPLICATION

To multiply two complex numbers convert them to polar format; then multiply their magnitudes and add their angles, respectively:

$$Z_1 Z_2 = (\rho_1 \angle \theta_1)(\rho_2 \angle \theta_2) = (\rho_1 \rho_2) \angle (\theta_1 + \theta_2) \tag{B.11}$$

For example,

$$(3 + j4)(5 - j8) = (5 \angle 53°)(9.43 \angle -56°)$$
$$= 47.15 \angle -3° \tag{B.12}$$

COMPLEX DIVISION

To divide two complex numbers, use polar format; then divide the magnitude of the numerator by the magnitude of the denominator and subtract the angle of the
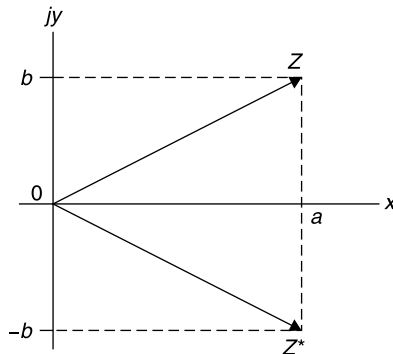
Figure B.3   Complex conjugate of Z.

denominator from the angle of the numerator:

$$\frac{Z_1}{Z_2} = \frac{\rho_1 \angle \theta_1}{\rho_2 \angle \theta_2} = \frac{\rho_1}{\rho_2} \angle (\theta_1 - \theta_2) \tag{B.13}$$

For example,

$$(3 + j4)/(5 - j8) = (5\angle 53°)/(9.43\angle -56°)$$
$$= 0.53\angle 109° \tag{B.14}$$

FORMING THE COMPLEX CONJUGATE

The *complex conjugate* Z^* of a complex number Z (Fig. B.3) is formed by changing the sign of the imaginary part (*in the rectangular form*) or changing the sign of the angle (*in the polar form*):

$$Z = a + jb = \rho \angle \theta$$
$$Z^* = a - jb = \rho \angle -\theta \tag{B.15}$$

For example,

$$Z = 4 + j3 = 5\angle 37° \quad \text{and} \quad Z^* = 4 - j3 = 5\angle -37° \tag{B.16}$$

REFERENCE

1. Kenneth S. Miller, *Advanced Complex Calculus*, Harper & Brothers, New York, 1960. *An old reference, much too advanced for our purposes, but its development of the bilinear transformation is fundamental to the Smith chart.*

Diameter and Resistance of Annealed Copper Wire by Gauge Size

AWG Wire B & S Gauge	Diameter (mils)	Diameter (mm)	DC Resistance at 20°C, 68°F per 100 ft (Ω)	Weight (lb/100 ft)	Weight (g/100 ft)
0000	460.0	116.8	0.00490	64.05	29,053
000	409.6	104.0	0.00618	50.78	23,035
00	364.8	92.66	0.00779	40.28	18,272
0	324.9	82.52	0.00982	31.95	14,493
1	289.3	73.48	0.01239	25.33	11,491
2	257.6	65.43	0.01563	20.09	9,111
3	229.4	58.27	0.01971	15.93	7,225
4	204.3	51.89	0.0248	12.63	5,731
5	181.9	46.20	0.0313	10.02	4,543
6	162.0	41.15	0.0395	7.944	3,603
7	144.3	36.65	0.0498	6.303	2,859
8	128.5	32.64	0.0628	4.998	2,267
9	114.4	29.06	0.0792	3.961	1,797
10	101.9	25.88	0.0999	3.143	1,426
11	90.74	23.05	0.126	2.492	1,130
12	80.81	20.53	0.159	1.977	896.6
13	71.96	18.28	0.200	1.567	711.0
14	64.08	16.28	0.253	1.243	563.8
15	57.07	14.50	0.318	0.9859	447.2
16	50.82	12.91	0.402	0.7818	354.6
17	45.26	11.50	0.506	0.6201	281.3
18	40.30	10.24	0.639	0.4916	223.0
19	35.89	9.116	0.805	0.3899	176.9
20	31.96	8.118	1.02	0.3092	140.2
21	28.46	7.229	1.28	0.2452	111.2
22	25.35	6.439	1.61	0.1945	88.23
23	22.57	5.733	2.04	0.1542	69.94

AWG Wire B & S Gauge	Diameter (mils)	Diameter (mm)	DC Resistance at 20°C, 68°F per 100 ft (Ω)	Weight (lb/100 ft)	Weight (g/100 ft)
24	20.10	5.105	2.57	0.1223	55.47
25	17.90	4.547	3.24	0.09699	43.99
26	15.94	4.049	4.08	0.07691	34.89
27	14.20	3.607	5.14	0.06104	27.69
28	12.64	3.211	6.49	0.04836	21.94
29	11.26	2.860	8.18	0.03838	17.41
30	10.03	2.548	10.3	0.03045	13.81
31	8.928	2.268	13.0	0.02413	10.94
32	7.950	2.019	16.4	0.01913	8.678
33	7.080	1.798	20.7	0.01517	6.882
34	6.305	1.601	26.1	0.01203	5.458
35	5.615	1.426	32.9	0.00954	4.329
36	5.000	1.270	41.5	0.00757	3.432
37	4.453	1.131	52.3	0.00600	2.723
38	3.965	1.007	66.0	0.00476	2.159
39	3.531	0.897	83.2	0.00377	1.712
40	3.145	0.799	105	0.00299	1.358

Properties of Some Materials

Material	Density (g/cm^3)	Specific Heat (cal./g·°C)	Thermal Resistivity (°C·cm/W)	Linear Expan. Coefficient (×10^{-6}/°C)	Relative Dielectric Constant	Resistivity (×10^{-6} Ω·cm)
Aluminum	2.70	0.226	0.46	22.9		2.65
Alumina (Al$_2$O$_3$)			5.41		9.8	
Boron-nitride	2.20		3.15			
Brass (66 Cu 34 Zn)			0.9	0.85		39
Carbon	2.2	0.165	0.15			1375
Copper	8.96	0.092	0.25	16.5		1.67
Glass					4–6	
Gold	19.3	0.03	0.34	14.2		2.4
Indium	7.31		4.2	33		8.37
Iron (99.99%)	7.87	0.108	1.3	11.7		9.7
Lead	11.34	0.03	2.9	28.7		20.6
Mercury	113.55	0.33	12			98
Mylar			1040		3	
Nylon			482		3	
Oil (transil)					2.2	
Paper (royal gray)					3	
Platinum	21.45	0.177	1.45	8.9		11
Polyethylene			304		2.3	
Polystyrene			965	70	2.55	
Quartz (SiO$_2$)			7–15		3.78	
Rexolite (1422)					2.53	
Silicon	2.4	0.176	1.2	4.2	11.8	
Silver	10.49	0.056	0.25	18.9		1.59
Sodium chloride					5.9	
Teflon pure			482	100, variable	2.03	
Tin	7.3	0.054	1.6	23		11
Titanium	4.5	0.142		8.5		42
Tungsten	19.3	0.034	0.5	4.3		5.6
Water 20°C (distilled)	1.00	1.00	160		80	
Zinc	7.14	0.09	0.9	17–39		6

Standard Rectangular Waveguides

EIA Designation WR-	Recom. Frequency Range (GHz)	TE$_{10}$ Cutoff Frequency (GHz)	Attenuation dB/100 ft Lo to Hi Frequency	Material	a Inside Width (in.)	b Inside Height (in.)	Wall Thickness or Outside Diameter (in.)
975	0.75–1.12	0.605	0.14–0.10	Aluminum	9.750	4.875	0.125
770	0.96–1.45	0.766	0.20–0.14	Aluminum	7.700	3.850	0.125
650	1.12–1.70	0.908	0.27–0.18	Aluminum	6.500	3.250	0.080
510	1.45–2.20	1.157			5.100	2.550	0.080
430	1.70–2.60	1.372	0.50–0.33	Aluminum	4.300	2.150	0.080
340	2.20–3.30	1.736	0.75–0.49	Aluminum	3.400	1.700	0.080
284	2.60–3.95	2.078	0.94–0.64	Aluminum	2.840	1.340	0.080
229	3.30–4.90	2.577		Aluminum	2.290	1.145	0.064
187	3.95–5.85	3.152	1.8–1.1	Aluminum	1.872	0.872	0.064
159	4.90–7.05	3.711		Aluminum	1.590	0.795	0.064
137	5.85–8.20	4.301	2.5–1.9	Aluminum	1.372	0.622	0.064
112	7.05–10.00	5.259	3.5–2.7	Aluminum	1.122	0.497	0.064
90	8.20–12.40	6.557	5.5–3.8	Aluminum	0.900	0.400	0.050
75	10.00–15.00	7.868		Aluminum	0.750	0.375	0.050
62	12.4–18.00	9.486	9.5–8.3	Brass	0.622	0.311	0.040
51	15.00–22.00	11.574			0.510	0.255	0.040
42	18.00–26.50	14.047	21–15	Brass	0.420	0.170	0.040
34	22.00–33.00	17.328			0.340	0.170	0.040
28	26.50–40.00	21.081	22–15	Silver	0.280	0.140	0.040
22	33.00–50.00	26.342	31–21	Silver	0.224	0.112	0.040
19	40.00–60.00	31.357			0.188	0.094	0.040
15	50.00–75.00	39.863	53–39	Silver	0.148	0.074	0.040
12	60.00–90.00	48.350	93–52	Silver	0.122	0.061	0.040
10	75.00–110.00	59.010			0.100	0.050	0.040
8	90.00–140.00	73.840	152–99	Silver	0.080	0.040	0.156 Dia.
7	110–170	90.840	163–137	Silver	0.065	0.0325	0.156 Dia.
5	140–220	115.750	308–193	Silver	0.051	0.0255	0.156 Dia.
4	170–260	137.520	384–254	Silver	0.043	0.0215	0.156 Dia.
3	220–325	173.280	512–348	Silver	0.034	0.0170	0.156 Dia.

Reprinted with permission of Microwave Development Labs, 135 Crescent Rd., Needham Hts. MA 02494.
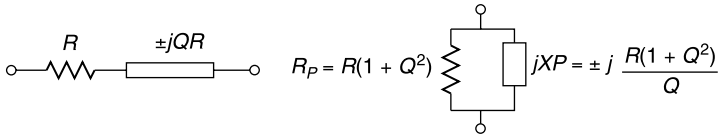
FREQUENTLY USED RELATIONS

Reactances

$$X_L = \omega L = 2\pi f L: \quad \text{At 1 GHz, 1 nH has } +6.28 \ \Omega$$

$$X_C = \frac{-1}{\omega C} = \frac{-1}{2\pi f C}: \quad \text{At 1 GHz, 1 pF has } -159 \ \Omega$$

Wavelength

$$\text{At 1 GHz } \lambda_0 = 30 \text{ cm} \approx 11.8 \text{ in.}$$

Series–Parallel Equivalent Circuits

Transmission Lines
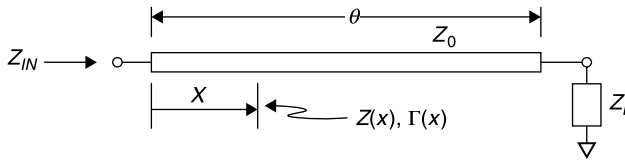
$$Z_{\text{IN}} = Z_0 \frac{Z_L + jZ_0 \tan \theta}{Z_0 + jZ_L \tan \theta} \qquad z_{\text{IN}} = \frac{z_L + j \tan \theta}{1 + jz_L \tan \theta} \qquad z_{\text{IN}} = \frac{Z_{\text{IN}}}{Z_0} \qquad z_L = \frac{Z_L}{Z_0}$$

$$\Gamma(x) = \rho \angle \varphi = \frac{V_R(x)}{V_I(x)} = \frac{Z(x) - Z_0}{Z(x) + Z_0} = \frac{z - 1}{z + 1} = \frac{Y_0 - Y(x)}{Y_0 + Y(x)} = \frac{1 - y}{1 + y}$$

$$Z(x) = Z_0 \frac{1 + \Gamma(x)}{1 - \Gamma(x)}$$

$$\rho = \frac{\text{VSWR} - 1}{\text{VSWR} + 1} \qquad \text{VSWR} = \frac{1 + \rho}{1 - \rho}$$

$$\text{Return loss} = \rho^2 = 20 \log \rho \text{ (dB)};$$

$$\text{Mismatch loss} = 1 - \rho^2 = 10 \log(1 - \rho^2) \text{ (dB)}$$

$$\text{Mismatch error (max \& min values)} = 20 \log \left[\frac{1}{(1 \pm \rho_1 \rho_2)} \right]$$

2 × 2 Matrix Multiplication

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \cdot \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}$$

Noise Factor F and Noise Figure NF

$$F = \frac{\text{SNR}_{\text{IN}}}{\text{SNR}_{\text{OUT}}} = \frac{P_{\text{Noise OUT}}}{GP_{\text{Noise IN}}} = 1 + \frac{T_E}{T_0} \quad \text{where } P_{\text{Noise IN}} = kT_0B \text{ and } T_0 = 290 \text{ K}$$

$$F_{\text{CASCADE}} = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} + \cdots + \frac{F_N - 1}{G_1 G_2 G_3 \dots G_{N-1}} \qquad \text{NF} = 10 \log F$$

Spurious-Free Dynamic Range

$$\text{SFDR (dB)} = \tfrac{2}{3}[P_3(\text{dBm}) + 174 \text{ dBm} - 10 \log B - G(\text{dB}) - \text{NF(dB)}]$$

Maxwell's Equations

1. $\nabla \cdot \vec{\mathbf{D}} = \rho \qquad \oint_S \vec{\mathbf{D}} \cdot \vec{\mathbf{dS}} = \int_v \rho \, dv$

2. $\nabla \cdot \vec{\mathbf{B}} = 0 \qquad \oint_S \vec{\mathbf{B}} \cdot \vec{\mathbf{dS}} = 0$

3. $\nabla \times \vec{\mathbf{E}} = -\dfrac{\partial \vec{\mathbf{B}}}{\partial t} \qquad \oint \vec{\mathbf{E}} \cdot \vec{\mathbf{dl}} = -\int_S \dfrac{\partial \vec{\mathbf{B}}}{\partial t} \cdot \vec{\mathbf{dS}}$

4. $\nabla \times \vec{\mathbf{H}} = \vec{\mathbf{J}}_C + \dfrac{\partial \vec{\mathbf{D}}}{\partial t} \qquad \oint \vec{\mathbf{H}} \cdot \vec{\mathbf{dl}} = \oint_S \vec{\mathbf{J}} \cdot \vec{\mathbf{dS}} + \oint_S \dfrac{\partial \vec{\mathbf{D}}}{\partial t} \cdot \vec{\mathbf{dS}}$

$$\vec{\mathbf{D}} = \varepsilon\vec{\mathbf{E}} \qquad \vec{\mathbf{B}} = \mu\vec{\mathbf{H}} \qquad \vec{\mathbf{J}} = \sigma\vec{\mathbf{E}} \qquad \vec{\mathbf{F}} = q\vec{\mathbf{v}} \times \vec{\mathbf{B}}$$

Vector Operations (Rectangular Coordinates)

Dot product:

$$\vec{\mathbf{A}} \cdot \vec{\mathbf{B}} = A_x B_x + A_y B_y + A_z B_z$$

Cross product:

$$\vec{\mathbf{A}} \times \vec{\mathbf{B}} = \begin{vmatrix} \vec{\mathbf{x}} & \vec{\mathbf{y}} & \vec{\mathbf{z}} \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix}$$

$$= \vec{\mathbf{x}}(A_y B_z - B_y A_z) + \vec{\mathbf{y}}(A_z B_x - B_z A_x) + \vec{\mathbf{z}}(A_x B_y - B_x A_y)$$

Divergence:

$$\nabla \cdot \vec{\mathbf{A}} = \frac{\partial}{\partial x} A_x + \frac{\partial}{\partial y} A_y + \frac{\partial}{\partial z} A_z$$

Curl:

$$\nabla \times \vec{\mathbf{A}} = \begin{vmatrix} \vec{\mathbf{x}} & \vec{\mathbf{y}} & \vec{\mathbf{z}} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ A_x & A_y & A_z \end{vmatrix} = \vec{\mathbf{x}}\left[\dfrac{\partial A_z}{\partial y} - \dfrac{\partial A_y}{\partial z}\right] + \vec{\mathbf{y}}\left[\dfrac{\partial A_x}{\partial z} - \dfrac{\partial A_z}{\partial x}\right] + \vec{\mathbf{z}}\left[\dfrac{\partial A_y}{\partial x} - \dfrac{\partial A_x}{\partial y}\right]$$

Gradient:

$$\nabla \Phi = \vec{\mathbf{x}}\dfrac{\partial \Phi}{\partial x} + \vec{\mathbf{y}}\dfrac{\partial \Phi}{\partial y} + \vec{\mathbf{z}}\dfrac{\partial \Phi}{\partial z}$$

Scalar Laplacian

$$\nabla^2 \Phi = \nabla \cdot \nabla \Phi = \dfrac{\partial^2 \Phi}{\partial x^2} + \dfrac{\partial^2 \Phi}{\partial y^2} + \dfrac{\partial^2 \Phi}{\partial z^2}$$

Vector Laplacian

$$\nabla^2 \vec{\mathbf{A}} = \vec{\mathbf{x}}\left(\dfrac{\partial^2 A_x}{\partial x^2} + \dfrac{\partial^2 A_x}{\partial y^2} + \dfrac{\partial^2 A_x}{\partial z^2}\right) + \vec{\mathbf{y}}\left(\dfrac{\partial^2 A_y}{\partial x^2} + \dfrac{\partial^2 A_y}{\partial y^2} + \dfrac{\partial^2 A_y}{\partial z^2}\right)$$

$$+ \vec{\mathbf{z}}\left(\dfrac{\partial^2 A_z}{\partial x^2} + \dfrac{\partial^2 A_z}{\partial y^2} + \dfrac{\partial^2 A_z}{\partial z^2}\right)$$

Physical Constants and Parameter Values

Description	Symbol and Value
Vacuum permittivity	$\varepsilon_0 = 8.854 \times 10^{-12}$ F/m
	$\approx 10^{-9}/36\pi$ F/m
Vacuum permeability	$\mu_0 = 4\pi \times 10^{-7}$ H/m
Vacuum speed of light	$c = 1/\sqrt{\mu_0 \varepsilon_0}$
	$= 2.998 \times 10^8$ m/s
	$\approx 3.00 \times 10^8$ m/s
	≈ 11.8 in./ns
Boltzmann's constant	$k = 1.380 \times 10^{-23}$ J/K
Electronic charge	$e = 1.602 \times 10^{-19}$ C
Electron volt	$e = 1.602 \times 10^{-19}$ J
Electron rest mass	$m_0 = 9.11 \times 10^{-31}$ kg
Planck's constant	$h = 6.62517 \times 10^{-34}$ J·s
Copper conductivity	$\sigma = 5.80 \times 10^7$ ℧/m
Value of Pi	$\pi = 3.14159$
Radian	$= 360°/2\pi = 57.29578° \approx 57.3°$
Loss (decibels)	$= 8.686$ loss (nepers)
Loss (nepers)	$= 0.115$ loss (decibels)

$$y = ax \qquad dy/dx = a$$
$$y = e^{ax} \qquad dy/dx = ae^{ax}$$
$$y = \sin ax \qquad dy/dx = a \cos ax$$
$$y = \sinh u \qquad dy/dx = \cosh u \, du/dx$$
$$y = uv \qquad dy/dx = v \, du/dx + u \, dv/dx$$
$$y = x^n \qquad dy/dx = nx^{n-1}$$
$$y = \ln x \qquad dy/dx = 1/x$$
$$y = \cos ax \qquad dy/dx = -a \sin ax$$
$$y = \cosh u \qquad dy/dx = \sinh u \, du/dx$$
$$y = u/v \qquad dy/dx = \frac{v \, du/dx - u \, dv/dx}{v^2}$$

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots \qquad |x| < 1$$
$$\sinh x = \tfrac{1}{2}(e^x - e^{-x}) \qquad \cosh x = \tfrac{1}{2}(e^x + e^{-x})$$
$$\tanh x = \frac{\sinh x}{\cosh x} \qquad \coth x = \frac{\cosh x}{\sinh x}$$
$$\sin^2 x + \cos^2 = 1$$
$$\sin(A + B) = \sin A \cos B + \cos A \sin B$$
$$\cos(A + B) = \cos A \cos B - \sin A \sin B$$
$$e^{jx} = \cos x + j \sin x$$
$$\nabla(\Phi + \Psi) = \nabla\Phi + \nabla\Psi$$
$$\nabla \times (\vec{A} + \vec{B}) = \nabla \times \vec{A} + \nabla \times \vec{B}$$
$$\nabla \cdot (\Psi\vec{A}) = \vec{A} \cdot \nabla\Psi + \Psi\nabla \cdot \vec{A}$$
$$\nabla \times (\Phi\vec{A}) = \nabla\Phi \times \vec{A} + \Phi\nabla \times \vec{A}$$
$$\nabla \cdot \nabla \times \vec{A} = 0$$
$$\nabla \times \nabla \times \vec{A} = \nabla(\nabla \cdot \vec{A}) - \nabla^2\vec{A}$$
$$\cosh^2 x - \sinh^2 x = 1$$
$$\sin(A - B) = \sin A \cos B - \cos A \sin B$$
$$\cos(A - B) = \cos A \cos B + \sin A \sin B$$
$$e^{-jx} = \cos x - j \sin x$$
$$\nabla \cdot (\vec{A} + \vec{B}) = \nabla \cdot \vec{A} + \nabla \cdot \vec{B}$$
$$\nabla(\Phi\Psi) = \Phi\nabla\Psi + \Psi\nabla\Phi$$
$$\nabla \cdot (\vec{A} \times \vec{B}) = \vec{B} \cdot \nabla \times \vec{A} - \vec{A} \cdot \nabla \times \vec{B}$$
$$\nabla \cdot \nabla\Phi = \nabla^2\Phi$$
$$\nabla \times \nabla\Phi = 0$$
$$\vec{A} \times (\vec{B} \times \vec{C}) = \vec{B}(\vec{A} \cdot \vec{C}) - \vec{C}(\vec{A} \cdot \vec{B})$$

A practical guide for today's wireless engineer

High Frequency Techniques: An Introduction to RF and Microwave Engineering is a clearly written classical circuit and field theory text illustrated with modern computer simulation software. The book's ten chapters cover:

- The origins and current uses of wireless transmission
- A review of AC analysis, Kirchhoff's laws, RLC elements, skin effect, and introduction to the use of computer simulation software
- Resonators, Q definitions, and Q-based impedance matching
- Transmission lines, waves, VSWR, reflection phenomena, Fano's reflection bandwidth limits, telegrapher, and impedance transformation equations
- Development and in-depth use of the Smith Chart
- Matrix algebra with Z, Y, *ABCD*, S, and *T* matrix applications
- An unusually thorough introduction to electromagnetic field theory, step-by-step development of vector calculus, Maxwell's equations, waveguides, propagation, and antennas
- Backward wave, branch line, rat race and Wilkinson couplers, impedance measurements, and detailed even and odd mode analysis
- Filter designs for Butterworth, Chebyshev, Bessel and elliptic responses, Kuroda's identities, Richards's transformation, and computer optimized designs
- Transistor amplifier design using Unilateral Gain, Simultaneous Match, Available Gain and Operating Gain approaches, insuring stability, cascading stages, broadbanding, noise theory, and intermodulation effects

Using informal language, *High Frequency Techniques* takes the reader step-by-step through RF and microwave theory and design, providing a lasting practical reference for the practicing wireless engineer.

JOSEPH F. WHITE, PhD, is an instructor and consultant at JFW Industries, Inc. He has twenty-five years of design experience, was technical director at M/A-COM, Inc., and received the IEEE Microwave Theory and Techniques Society's Application Award for "Contributions to Phased Array Antennas." He edited the *Microwave Journal* and *Applied Microwave and Wireless* magazines, wrote *Microwave Semiconductor Engineering*, and is a Fellow of the IEEE. He can be reached at jfwhite@ieee.org.
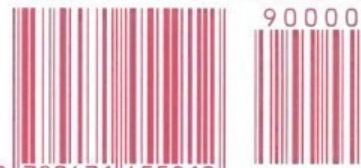
◆ **IEEE**
IEEE PRESS

WILEY-INTERSCIENCE
wiley.com