Handbook of

ENGINEERING ELECTROMAGNETICS

Handbook of
ENGINEERING ELECTROMAGNETICS

Edited by
Rajeev Bansal

University of Connecticut
Storrs, Connecticut, U.S.A.

MARCEL DEKKER

MARCEL DEKKER NEW YORK

Although great care has been taken to provide accurate and current information, neither the author(s) nor the publisher, nor anyone else associated with this publication, shall be liable for any loss, damage, or liability directly or indirectly caused or alleged to be caused by this book. The material contained herein is not intended to provide specific advice or recommendations for any specific situation.

Trademark notice: Product or corporate names may be trademarks or registered trademarks and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress.

ISBN: 0-8247-5628-2

This book is printed on acid-free paper.

Headquarters
Marcel Dekker, Inc., 270 Madison Avenue, New York, NY 10016, U.S.A.
tel: 212-696-9000; fax: 212-685-4540

Distribution and Customer Service
Marcel Dekker, Inc., Cimarron Road, Monticello, New York 12701, U.S.A.
tel: 800-228-1160; fax: 845-796-1772

Eastern Hemisphere Distribution
Marcel Dekker AG, Hutgasse 4, Postfach 812, CH-4001 Basel, Switzerland
tel: 41-61-260-6300; fax: 41-61-260-6333

World Wide Web
http://www.dekker.com

The publisher offers discounts on this book when ordered in bulk quantities. For more information, write to Special Sales/Professional Marketing at the headquarters address above.

Current printing (last digit):

10 9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

To the memory of my parents

Preface

This handbook is intended as a desk reference for the broad area of engineering electromagnetics. Since electromagnetics provides the underpinnings for many technological fields such as wireless communications, fiber optics, microwave engineering, radar, electromagnetic compatibility, material science, and biomedicine, there is a great deal of interest in and need for training in the engineering applications of electromagnetics. Practicing engineers in these diverse fields need to understand how engineering electromagnetic principles can be applied to the formulation and solution of actual engineering problems. As technologies wax and wane and engineers move around, they find themselves learning new applications on the run.

The *Handbook of Engineering Electromagnetics* should serve as a bridge between standard textbooks in electromagnetic theory and specialized references such as a handbook on wireless antenna design. While textbooks are comprehensive in terms of the theoretical development of the subject matter, they are usually deficient in the practical application of that theory. Specialized handbooks, on the other hand, often provide detailed lists of formulas, tables, and graphs, but do not provide the insight needed to appreciate the underlying physical concepts. This handbook will permit a practicing engineer/scientist to

Review the necessary electromagnetic *theory* in the context of the application
Gain an appreciation for the key electromagnetic terms and parameters
Learn how to apply the theory to formulate engineering problems
Obtain guidance to the specialized literature for additional details.

Since the *Handbook of Engineering Electromagnetics* is intended to be useful to engineers engaged in electromagnetic applications in a variety of professional settings, the coverage of topics is correspondingly broad in scope (as can be inferred from the table of contents). In terms of *fundamental concepts*, the book includes coverage of Maxwell equations, static fields, electromagnetic induction, waves, transmission lines, waveguides, antennas, and electromagnetic compatibility (Chapters 1–10). In terms of *electromagnetic technologies*, radar, wireless communication, satellite communication, and optical communication are covered (Chapters 11–14). Chapter 15 provides an introduction to various numerical techniques being used for computer-aided solutions to complex electromagnetic problems. Given the ubiquitous nature of electromagnetic fields, it is important to consider their biological effects and safety standards (Chapter 16). Chapter 17 presents a concise survey of current and evolving biomedical applications, while Chapter 18 is a review of the techniques used for measuring the electromagnetic properties of biological materials. In terms of *frequency range*, this book spans the spectrum from static fields to light waves, with special emphasis on the radio frequency/ microwave range. Pertinent data in the form of tables and graphs are provided within the context of the subject matter. In addition, Appendixes A and B are brief compilations of important electromagnetic constants and units, respectively. Finally, Appendix C is a convenient tutorial on vector analysis and coordinate systems.

To keep the size of this handbook manageable, certain topics (e.g., electrical machines and semiconductor devices) had to be excluded. The primary guiding principle has been to exclude applications where an analysis based on electromagnetic theory does not play a significant role in actual engineering practice or areas where a meaningful coverage could not be provided within the framework of the handbook.

First and foremost, I thank all the contributors, whose hard work is reflected in these pages. I would like to express my appreciation to Dr. Amir Faghri, Dean of the School of Engineering, and Dr. Robert Magnusson, Head of the Electrical and Computer Engineering Department, University of Connecticut, for supporting my request for a sabbatical leave (spring 2003), which facilitated the completion of this project. My editors at Marcel Dekker, Inc., especially Taisuke Soda, provided valuable help and advice throughout the project. I thank Anthony Palladino for his help in preparing the manuscript of Appendix C. Finally, I would like to express my gratitude to my family for their unfailing support and encouragement.

Rajeev Bansal

Contents

Contributors

Nathan Blaunstein *Ben-Gurion University of the Negev, Beer Sheva, Israel*

Christo Christopoulos *University of Nottingham, Nottingham, England*

Afshin Daryoush *Drexel University, Philadelphia, Pennsylvania, U.S.A.*

Kenneth R. Demarest *The University of Kansas, Lawrence, Kansas, U.S.A.*

Riadh Habash *University of Ottawa, Ottawa, Ontario, Canada*

Randy L. Haupt *Utah State University, Logan, Utah, U.S.A.*

Mark N. Horenstein *Boston University, Boston, Massachusetts, U.S.A.*

David R. Jackson *University of Houston, Houston, Texas, U.S.A.*

Mohammad Kolbehdari *Intel Corporation, Hillsboro, Oregon, U.S.A.*

James C. Lin *University of Illinois at Chicago, Chicago, Illinois, U.S.A.*

Joseph C. Palais *Arizona State University, Tempe, Arizona, U.S.A.*

Branko D. Popović[†] *University of Belgrade, Belgrade, Yugoslavia*

Milica Popović *McGill University, Montreal, Quebec, Canada*

Zoya Popović *University of Colorado, Boulder, Colorado, U.S.A.*

N. Narayana Rao *University of Illinois at Urbana-Champaign, Urbana, Illinois, U.S.A.*

[†] *Deceased.*

Matthew N. O. Sadiku *Prairie View A&M University, Prairie View, Texas, U.S.A.*

Levent Sevgi *DOGUS University, Istanbul, Turkey*

David Thiel *Griffith University, Nathan, Queensland, Australia*

Mohammad-Reza Tofighi *Drexel University, Philadelphia, Pennsylvania, U.S.A.*

Andreas Weisshaar *Oregon State University, Corvallis, Oregon, U.S.A.*

Jeffrey T. Williams *University of Houston, Houston, Texas, U.S.A.*

Donald R. Wilton *University of Houston, Houston, Texas, U.S.A.*

1
Fundamentals of Engineering Electromagnetics Revisited

N. Narayana Rao
University of Illinois at Urbana-Champaign
Urbana, Illinois, U.S.A.

In this chapter, we present in a nutshell the fundamental aspects of engineering electromagnetics from the view of looking back in a reflective fashion at what has already been learned in undergraduate electromagnetics courses as a novice. The first question that comes to mind in this context is on what constitutes the fundamentals of engineering electromagnetics. If the question is posed to several individuals, it is certain that they will come up with sets of topics, not necessarily the same or in the same order, but all containing the topic of Maxwell's equations at some point in the list, ranging from the beginning to the end of the list. In most cases, the response is bound to depend on the manner in which the individual was first exposed to the subject. Judging from the contents of the vast collection of undergraduate textbooks on electromagnetics, there is definitely a heavy tilt toward the traditional, or historical, approach of beginning with statics and culminating in Maxwell's equations, with perhaps an introduction to waves. Primarily to provide a more rewarding understanding and appreciation of the subject matter, and secondarily owing to my own fascination resulting from my own experience as a student, a teacher, and an author [1–7] over a few decades, I have employed in this chapter the approach of beginning with Maxwell's equations and treating the different categories of fields as solutions to Maxwell's equations. In doing so, instead of presenting the topics in an unconnected manner, I have used the thread of statics–quasistatics–waves to cover the fundamentals and bring out the frequency behavior of physical structures at the same time.

1.1. FIELD CONCEPTS AND CONSTITUTIVE RELATIONS

1.1.1. Lorentz Force Equation

A region is said to be characterized by an electric field if a particle of charge q moving with a velocity $\mathbf{v}$ experiences a force $\mathbf{F}_e$, independent of $\mathbf{v}$. The force, $\mathbf{F}_e$, is given by

$$\mathbf{F}_e = q\mathbf{E} \tag{1.1}$$

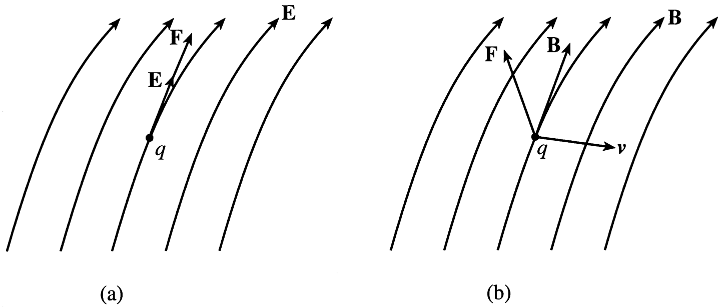


Figure 1.1 Illustrates that (a) the electric force is parallel to **E** but (b) the magnetic force is perpendicular to **B**.

where **E** is the electric field intensity, as shown in Fig. 1.1a. We note that the units of **E** are newtons per coulomb (N/C). Alternate and more commonly used units are volts per meter (V/m), where a volt is a newton-meter per coulomb. The line integral of **E** between two points A and B in an electric field region, $\int_A^B \mathbf{E} \cdot d\mathbf{l}$, has the meaning of voltage between A and B. It is the work per unit charge done by the field in the movement of the charge from A to B. The line integral of **E** around a closed path C is also known as the *electromotive force* (emf) around C.

If the charged particle experiences a force which depends on **v**, then the region is said to be characterized by a magnetic field. The force, $\mathbf{F}_m$, is given by

$$\mathbf{F}_m = q\mathbf{v} \times \mathbf{B} \tag{1.2}$$

where **B** is the magnetic flux density. We note that the units of **B** are newtons/(coulomb-meter per second), or (newton-meter per coulomb) × (seconds per square meter), or volt-seconds per square meter. Alternate and more commonly used units are webers per square meter (Wb/m^2) or tesla (T), where a weber is a volt-second. The surface integral of **B** over a surface S, $\int_S \mathbf{B} \cdot d\mathbf{S}$, is the magnetic flux (Wb) crossing the surface.

Equation (1.2) tells us that the magnetic force is proportional to the magnitude of **v** and orthogonal to both **v** and **B** in the right-hand sense, as shown in Fig. 1.1b. The magnitude of the force is $qvB \sin \alpha$, where α is the angle between **v** and **B**. Since the force is normal to **v**, there is no acceleration along the direction of motion. Thus the magnetic field changes only the direction of motion of the charge and does not alter the kinetic energy associated with it.

Since current flow in a wire results from motion of charges in the wire, a wire of current placed in a magnetic field experiences a magnetic force. For a differential length $d\mathbf{l}$ of a wire of current I placed in a magnetic field **B**, this force is given by

$$d\mathbf{F}_m = I\, d\mathbf{l} \times \mathbf{B} \tag{1.3}$$

as shown in Fig. 1.2.

Combining Eqs. (1.1) and (1.2), we obtain the expression for the total force $\mathbf{F} = \mathbf{F}_e + \mathbf{F}_m$, experienced by a particle of charge q moving with a velocity **v** in a region of

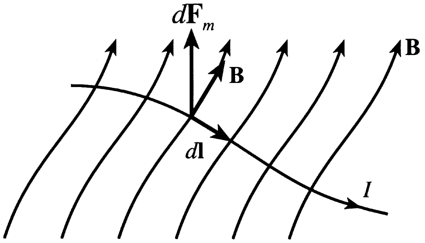


Figure 1.2 Force experienced by a current element in a magnetic field.

electric and magnetic fields, **E** and **B**, respectively, as

$$\begin{aligned}\mathbf{F} &= q\mathbf{E} + q\mathbf{v} \times \mathbf{B} \\ &= q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \qquad (1.4)\end{aligned}$$

Equation (1.4) is known as the *Lorentz force equation.*

1.1.2. Material Parameters and Constitutive Relations

The vectors **E** and **B** are the fundamental field vectors that define the force acting on a charge moving in an electromagnetic field, as given by the Lorentz force Eq. (1.4). Two associated field vectors **D** and **H**, known as the *electric flux density* (or the *displacement flux density*) and the *magnetic field intensity*, respectively, take into account the dielectric and magnetic properties, respectively, of material media. Materials contain charged particles that under the application of external fields respond giving rise to three basic phenomena known as *conduction*, *polarization*, and *magnetization*. Although a material may exhibit all three properties, it is classified as a *conductor*, a *dielectric*, or a *magnetic* material depending upon whether conduction, polarization, or magnetization is the predominant phenomenon. While these phenomena occur on the atomic or ‘‘microscopic’’ scale, it is sufficient for our purpose to characterize the material based on ‘‘macroscopic’’ scale observations, that is, observations averaged over volumes large compared with atomic dimensions.

In the case of conductors, the effect of conduction is to produce a current in the material known as the *conduction current*. Conduction is the phenomenon whereby the free electrons inside the material move under the influence of the externally applied electric field with an average velocity proportional in magnitude to the applied electric field, instead of accelerating, due to the frictional mechanism provided by collisions with the atomic lattice. For linear isotropic conductors, the conduction current density, having the units of amperes per square meter (A/m^2), is related to the electric field intensity in the manner

$$\mathbf{J}_c = \sigma\mathbf{E} \qquad (1.5)$$

where σ is the conductivity of the material, having the units siemens per meter (S/m). In semiconductors, the conductivity is governed by not only electrons but also holes.

While the effect of conduction is taken into account explicitly in the electromagnetic field equations through Eq. (1.5), the effect of polarization is taken into account implicitly

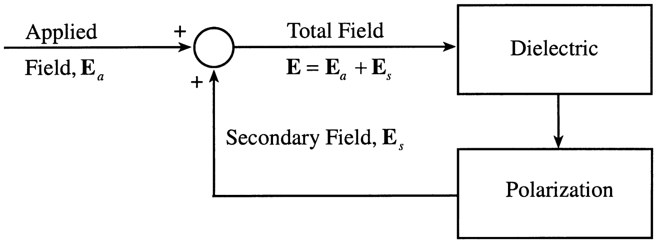

Applied
Field, $\mathbf{E}_a$
+
+
Total Field
$\mathbf{E} = \mathbf{E}_a + \mathbf{E}_s$
Dielectric
Secondary Field, $\mathbf{E}_s$
Polarization

Figure 1.3 Illustrates the effect of polarization in a dielectric material.

through the relationship between **D** and **E**, which is given by

$$\mathbf{D} = \varepsilon \mathbf{E} \tag{1.6}$$

for linear isotropic dielectrics, where ε is the *permittivity* of the material having the units coulomb squared per newton-squared meter, commonly known as *farads per meter* (F/m), where a *farad* is a coulomb square per newton-meter.

Polarization is the phenomenon of creation and net alignment of electric dipoles, formed by the displacements of the centroids of the electron clouds of the nuclei of the atoms within the material, along the direction of an applied electric field. The effect of polarization is to produce a secondary field that acts in superposition with the applied field to cause the polarization. Thus the situation is as depicted in Fig. 1.3. To implicitly take this into account, leading to Eq. (1.6), we begin with

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P} \tag{1.7}$$

where ε_0 is the permittivity of free space, having the numerical value 8.854×10^{-12}, or approximately $10^{-9}/36\pi$, and **P** is the polarization vector, or the dipole moment per unit volume, having the units (coulomb-meters) per cubic meter or coulombs per square meter. Note that this gives the units of coulombs per square meter for **D**. The term $\varepsilon_0 \mathbf{E}$ accounts for the relationship between **D** and **E** if the medium were free space, and the quantity **P** represents the effect of polarization. For linear isotropic dielectrics, **P** is proportional to **E** in the manner

$$\mathbf{P} = \varepsilon_0 \chi_e \mathbf{E} \tag{1.8}$$

where χ_e, a dimensionless quantity, is the electric susceptibility, a parameter that signifies the ability of the material to get polarized. Combining Eqs. (1.7) and (1.8), we have

$$\begin{aligned} \mathbf{D} &= \varepsilon_0 (1 + \chi_e) \mathbf{E} \\ &= \varepsilon_0 \varepsilon_r \mathbf{E} \\ &= \varepsilon \mathbf{E} \end{aligned} \tag{1.9}$$

where ε_r $(= 1 + \chi_e)$ is the relative permittivity of the material.

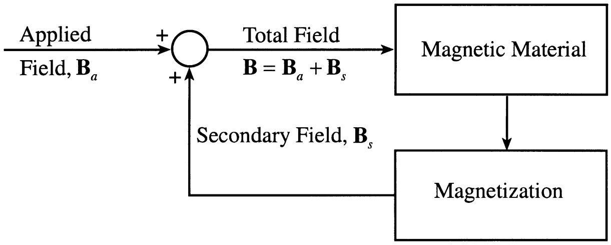

Applied
Field, $\mathbf{B}_a$
+
+
Total Field
$\mathbf{B} = \mathbf{B}_a + \mathbf{B}_s$
Magnetic Material
Secondary Field, $\mathbf{B}_s$
Magnetization

Figure 1.4 Illustrates the effect of magnetization in a magnetic material.

In a similar manner, the effect of magnetization is taken into account implicitly through the relationship between **H** and **B**, which is given by

$$\mathbf{H} = \frac{\mathbf{B}}{\mu} \tag{1.10}$$

for linear isotropic magnetic materials, where μ is the permeability of the material, having the units newtons per ampere squared, commonly known as *henrys per meter* (H/m), where a *henry* is a newton-meter per ampere squared.

Magnetization is the phenomenon of net alignment of the axes of the magnetic dipoles, formed by the electron orbital and spin motion around the nuclei of the atoms in the material, along the direction of the applied magnetic field. The effect of magnetization is to produce a secondary field that acts in superposition with the applied field to cause the magnetization. Thus the situation is as depicted in Fig. 1.4. To implicitly take this into account, we begin with

$$\mathbf{B} = \mu_0\mathbf{H} + \mu_0\mathbf{M} \tag{1.11}$$

where μ_0 is the permeability of free space, having the numerical value $4\pi \times 10^{-7}$, and **M** is the magnetization vector or the magnetic dipole moment per unit volume, having the units (ampere-square meters) per cubic meter or amperes per meter. Note that this gives the units of amperes per square meter for **H**. The term $\mu_0\mathbf{H}$ accounts for the relationship between **H** and **B** if the medium were free space, and the quantity $\mu_0\mathbf{M}$ represents the effect of magnetization. For linear isotropic magnetic materials, **M** is proportional to **H** in the manner

$$\mathbf{M} = \chi_m\mathbf{H} \tag{1.12}$$

where χ_m, a dimensionless quantity, is the magnetic susceptibility, a parameter that signifies the ability of the material to get magnetized. Combining Eqs. (1.11) and 1.12), we have

$$\mathbf{H} = \frac{\mathbf{B}}{\mu_0(1+\chi_m)} \\ = \frac{\mathbf{B}}{\mu_0\mu_r} \\ = \frac{\mathbf{B}}{\mu} \tag{1.13}$$

where μ_r $(=1+\chi_m)$ is the relative permeability of the material.

Equations (1.5), (1.6), and (1.10) are familiarly known as the *constitutive relations*, where σ, ε, and μ are the material parameters. The parameter σ takes into account explicitly the phenomenon of conduction, whereas the parameters ε and μ take into account implicitly the phenomena of polarization and magnetization, respectively.

The constitutive relations, Eqs. (1.5), (1.6), and (1.10), tell us that $\mathbf{J}_c$ is parallel to **E**, **D** is parallel to **E**, and **H** is parallel to **B**, independent of the directions of the field vectors. For anisotropic materials, the behavior depends upon the directions of the field vectors. The constitutive relations have then to be written in matrix form. For example, in an anisotropic dielectric, each component of **P** and hence of **D** is in general dependent upon each component of **E**. Thus, in terms of components in the Cartesian coordinate system, the constitutive relation is given by

$$\begin{bmatrix} D_x \\ D_y \\ D_z \end{bmatrix} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} \\ \varepsilon_{31} & \varepsilon_{32} & \varepsilon_{33} \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} \tag{1.14}$$

or, simply by

$$[\mathbf{D}] = [\varepsilon][\mathbf{E}] \tag{1.15}$$

where [**D**] and [**E**] are the column matrices consisting of the components of **D** and **E**, respectively, and $[\varepsilon]$ is the permittivity matrix (tensor of rank 2) containing the elements ε_{ij}, $i=1$, 2, 3 and $j=1$, 2, 3. Similar relationships hold for anisotropic conductors and anisotropic magnetic materials.

Since the permittivity matrix is symmetric, that is, $\varepsilon_{ij}=\varepsilon_{ji}$, from considerations of energy conservation, an appropriate choice of the coordinate system can be made such that some or all of the nondiagonal elements are zero. For a particular choice, all of the nondiagonal elements can be made zero so that

$$[\varepsilon] = \begin{bmatrix} \varepsilon_1 & 0 & 0 \\ 0 & \varepsilon_2 & 0 \\ 0 & 0 & \varepsilon_3 \end{bmatrix} \tag{1.16}$$

Then

$$D_{x'} = \varepsilon_1 E_{x'} \tag{1.17a}$$

$$D_{y'} = \varepsilon_2 E_{y'} \tag{1.17b}$$

$$D_{z'} = \varepsilon_3 E_{z'} \tag{1.17c}$$

so that **D** and **E** are parallel when they are directed along the coordinate axes, although with different values of *effective permittivity*, that is, ratio of **D** to **E**, for each such direction. The axes of the coordinate system are then said to be the *principal axes* of the medium. Thus when the field is directed along a principal axis, the anisotropic medium can be treated as an isotropic medium of permittivity equal to the corresponding effective permittivity.

1.2. MAXWELL'S EQUATIONS, BOUNDARY CONDITIONS, POTENTIALS, AND POWER AND ENERGY

1.2.1. Maxwell's Equations in Integral Form and the Law of Conservation of Charge

In Sec. 1.1, we introduced the different field vectors and associated constitutive relations for material media. The electric and magnetic fields are governed by a set of four laws, known as *Maxwell's equations*, resulting from several experimental findings and a purely mathematical contribution. Together with the constitutive relations, Maxwell's equations form the basis for the entire electromagnetic field theory. In this section, we shall consider the time variations of the fields to be arbitrary and introduce these equations and an auxiliary equation in the time domain form. In view of their experimental origin, the fundamental form of Maxwell's equations is the integral form. In the following, we shall first present all four Maxwell's equations in integral form and the auxiliary equation, the law of conservation of charge, and then discuss several points of interest pertinent to them. It is understood that all field quantities are real functions of position and time; that is, $\mathbf{E} = \mathbf{E}(\mathbf{r}, t) = \mathbf{E}(x, y, z, t)$, etc.

Faraday's Law

Faraday's law is a consequence of the experimental finding by Michael Faraday in 1831 that a time-varying magnetic field gives rise to an electric field. Specifically, the electromotive force around a closed path C is equal to the negative of the time rate of increase of the magnetic flux enclosed by that path, that is,

$$\oint_C \mathbf{E}\cdot d\mathbf{l} = -\frac{d}{dt}\int_S \mathbf{B}\cdot d\mathbf{S} \tag{1.18}$$

where S is any surface bounded by C, as shown, for example, in Fig. 1.5.

Ampere's Circuital Law

Ampere's circuital law is a combination of an experimental finding of Oersted that electric currents generate magnetic fields and a mathematical contribution of Maxwell that time-varying electric fields give rise to magnetic fields. Specifically, the magnetomotive force (mmf) around a closed path C is equal to the sum of the current enclosed by that path due

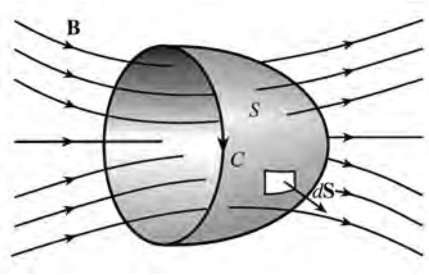


Figure 1.5 Illustrates Faraday's law.

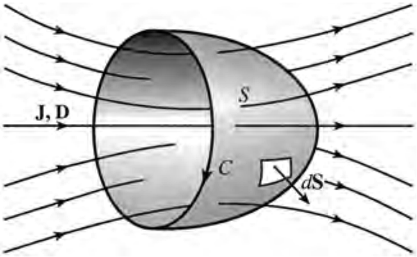


Figure 1.6 Illustrates Ampere's circuital law.

to actual flow of charges and the displacement current due to the time rate of increase of the electric flux (or displacement flux) enclosed by that path; that is,

$$\oint_C \mathbf{H}\cdot d\mathbf{l} = \int_S \mathbf{J}\cdot d\mathbf{S} + \frac{d}{dt}\int_S \mathbf{D}\cdot d\mathbf{S} \tag{1.19}$$

where S is any surface bounded by C, as shown, for example, in Fig. 1.6.

Gauss' Law for the Electric Field

Gauss' law for the electric field states that electric charges give rise to electric field. Specifically, the electric flux emanating from a closed surface S is equal to the charge enclosed by that surface, that is,

$$\oint_S \mathbf{D}\cdot d\mathbf{S} = \int_V \rho\, dv \tag{1.20}$$

where V is the volume bounded by S, as shown, for example, in Fig. 1.7. In Eq. (1.20), the quantity ρ is the volume charge density having the units coulombs per cubic meter (C/m^3).

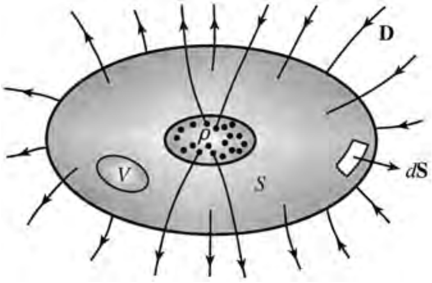


Figure 1.7 Illustrates Gauss' law for the electric field.

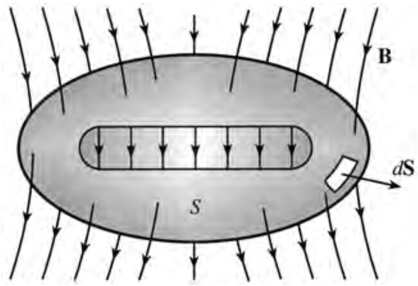


Figure 1.8 Illustrates Gauss' law for the magnetic field.

Gauss' Law for the Magnetic Field

Gauss' law for the magnetic field states that the magnetic flux emanating from a closed surface S is equal to zero, that is,

$$\oint_S \mathbf{B} \cdot d\mathbf{S} = 0 \tag{1.21}$$

Thus, whatever magnetic flux enters (or leaves) a certain part of the closed surface must leave (or enter) through the remainder of the closed surface, as shown, for example, in Fig. 1.8.

Law of Conservation of Charge

An auxiliary equation known as the *law of conservation of charge* states that the current due to flow of charges emanating from a closed surface S is equal to the time rate of decrease of the charge inside the volume V bounded by that surface, that is,

$$\oint_S \mathbf{J} \cdot d\mathbf{S} = -\frac{d}{dt}\int_V \rho \, dv$$

or

$$\oint_S \mathbf{J} \cdot d\mathbf{S} + \frac{d}{dt}\int_V \rho \, dv = 0 \tag{1.22}$$

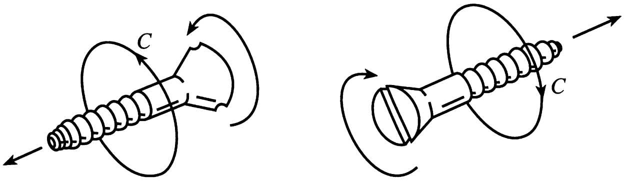


Figure 1.9 Right-hand-screw-rule convention.

There are certain procedures and observations of interest pertinent to Eqs. (1.18)–(1.22), as follows.

1. The direction of the infinitesimal surface vector $d\mathbf{S}$ in Figs. 1.5 and 1.6 denotes that the magnetic flux and the displacement flux, respectively, are to be evaluated in accordance with the right-hand screw rule (RHS rule), that is, in the sense of advance of a right-hand screw as it is turned around C in the sense of C, as shown in Fig. 1.9. The RHS rule is a convention that is applied consistently for all electromagnetic field laws involving integration over surfaces bounded by closed paths.
2. In evaluating the surface integrals in Eqs. (1.18) and (1.19), any surface S bounded by C can be employed. In addition in Eq. (1.19), the same surface S must be employed for both surface integrals. This implies that the time derivative of the magnetic flux through all possible surfaces bounded by C is the same in order for the emf around C to be unique. Likewise, the sum of the current due to flow of charges and the displacement current through all possible surfaces bounded C is the same in order for the mmf around C to be unique.
3. The minus sign on the right side of Eq. (1.18) tells us that when the magnetic flux enclosed by C is increasing with time, the induced voltage is in the sense opposite to that of C. If the path C is imagined to be occupied by a wire, then a current would flow in the wire that produces a magnetic field so as to oppose the increasing flux. Similar considerations apply for the case of the magnetic flux enclosed by C decreasing with time. These are in accordance with Lenz' law, which states that the sense of the induced emf is such that any current it produces tends to oppose the change in the magnetic flux producing it.
4. If loop C contains more than one turn, such as in an N-turn coil, then the surface S bounded by C takes the shape of a spiral ramp, as shown in Fig. 1.10. For a tightly wound coil, this is equivalent to the situation in which N separate, identical, single-turn loops are stacked so that the emf induced in the N-turn coil is N times the emf induced in one turn. Thus, for an N-turn coil,

$$\text{emf} = -N\frac{d\psi}{dt} \tag{1.23}$$

   where ψ is the magnetic flux computed as though the coil is a one-turn coil.
5. Since magnetic force acts perpendicular to the motion of a charge, the magnetomotive (mmf) force, that is, $\oint_C \mathbf{H}\cdot d\mathbf{l}$, does not have a physical meaning similar to that of the electromotive force. The terminology arises purely from analogy with electromotive force for $\oint_C \mathbf{E}\cdot d\mathbf{l}$.

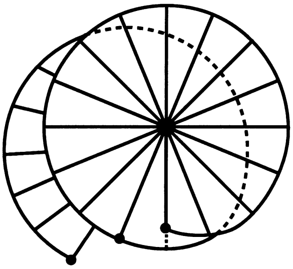


Figure 1.10 Two-turn loop.

6. The charge density ρ in Eq. (1.20) and the current density $\mathbf{J}$ in Eq. (1.19) pertain to true charges and currents, respectively, due to motion of true charges. They do not pertain to charges and currents resulting from the polarization and magnetization phenomena, since these are implicitly taken into account by the formulation of these two equations in terms of **D** and **H**, instead of in terms of **E** and **B**.
7. The displacement current, $d(\int_S \mathbf{D}\cdot d\mathbf{S})/dt$ is not a true current, that is, it is not a current due to actual flow of charges, such as in the case of the conduction current in wires or a convection current due to motion of a charged cloud in space. Mathematically, it has the units of $d[(\mathrm{C/m^2}) \times \mathrm{m^2}]/dt$ or amperes, the same as the units for a true current, as it should be. Physically, it leads to the same phenomenon as a true current does, even in free space for which **P** is zero, and **D** is simply equal to $\varepsilon_0\mathbf{E}$. Without it, the uniqueness of the mmf around a given closed path C is not ensured. In fact, Ampere's circuital law in its original form did not contain the displacement current term, thereby making it valid only for the static field case. It was the mathematical contribution of Maxwell that led to the modification of the original Ampere's circuital law by the inclusion of the displacement current term. Together with Faraday's law, this modification in turn led to the theoretical prediction by Maxwell of the phenomenon of electromagnetic wave propagation in 1864 even before it was confirmed experimentally 23 years later in 1887 by Hertz.
8. The observation concerning the time derivative of the magnetic flux crossing all possible surfaces bounded by a given closed path C in item 2 implies that the time derivative of the magnetic flux emanating from a closed surface S is zero, that is,

$$\frac{d}{dt}\oint_S \mathbf{B}\cdot d\mathbf{S} = 0 \qquad (1.24)$$

   One can argue then that the magnetic flux emanating from a closed surface is zero, since at an instant of time when no sources are present the magnetic field vanishes. Thus, Gauss' law for the magnetic field is not independent of Faraday's law.
9. Similarly, combining the observation concerning the sum of the current due to flow of charges and the displacement current through all possible surfaces

bounded by a given closed path C in item 2 with the law of conservation of charge, we obtain for any closed surface S,

$$\frac{d}{dt}\left(\oint_S \mathbf{D}\cdot d\mathbf{S} - \int_V \rho\, dv\right) = 0 \tag{1.25}$$

where V is the volume bounded by S. Once again, one can then argue that the quantity inside the parentheses is zero, since at an instant of time when no sources are present, it vanishes. Thus, Gauss' law for the electric field is not independent of Ampere's circuital law in view of the law of conservation of charge.

10. The cut view in Fig. 1.8 indicates that magnetic field lines are continuous, having no beginnings or endings, whereas the cut view in Fig. 1.7 indicates that electric field lines are discontinuous wherever there are charges, diverging from positive charges and converging on negative charges.

1.2.2 Maxwell's Equations in Differential Form and the Continuity Equation

From the integral forms of Maxwell's equations, one can obtain the corresponding differential forms through the use of Stoke's and divergence theorems in vector calculus, given, respectively, by

$$\oint_C \mathbf{A}\cdot d\mathbf{l} = \int_S (\nabla \times \mathbf{A})\cdot d\mathbf{S} \tag{1.26a}$$

$$\oint_S \mathbf{A}\cdot d\mathbf{S} = \int_V (\nabla\cdot\mathbf{A})\, dv \tag{1.26b}$$

where in Eq. (1.26a), S is any surface bounded by C and in Eq. (1.26b), V is the volume bounded by S. Thus, Maxwell's equations in differential form are given by

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{1.27}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \tag{1.28}$$

$$\nabla\cdot\mathbf{D} = \rho \tag{1.29}$$

$$\nabla\cdot\mathbf{B} = 0 \tag{1.30}$$

corresponding to the integral forms Eqs. (1.18)–(1.21), respectively. These differential equations state that at any point in a given medium, the curl of the electric field intensity is equal to the time rate of decrease of the magnetic flux density, and the curl of the magnetic field intensity is equal to the sum of the current density due to flow of charges and the displacement current density (time derivative of the displacement flux density); whereas the divergence of the displacement flux density is equal to the volume charge density, and the divergence of the magnetic flux density is equal to zero.

Auxiliary to the Maxwell's equations in differential form is the differential equation following from the law of conservation of charge Eq. (1.22) through the use of Eq. (1.26b). Familiarly known as the *continuity equation*, this is given by

$$\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} = 0 \tag{1.31}$$

It states that at any point in a given medium, the divergence of the current density due to flow of charges plus the time rate of increase of the volume charge density is equal to zero.

From the interdependence of the integral laws discussed in the previous section, it follows that Eq. (1.30) is not independent of Eq. (1.27), and Eq. (1.29) is not independent of Eq. (1.28) in view of Eq. (1.31).

Maxwell's equations in differential form lend themselves well for a qualitative discussion of the interdependence of time-varying electric and magnetic fields giving rise to the phenomenon of electromagnetic wave propagation. Recognizing that the operations of curl and divergence involve partial derivatives with respect to space coordinates, we observe that time-varying electric and magnetic fields coexist in space, with the spatial variation of the electric field governed by the temporal variation of the magnetic field in accordance with Eq. (1.27), and the spatial variation of the magnetic field governed by the temporal variation of the electric field in addition to the current density in accordance with Eq. (1.28). Thus, if in Eq. (1.28) we begin with a time-varying current source represented by **J**, or a time-varying electric field represented by $\partial\mathbf{D}/dt$, or a combination of the two, then one can visualize that a magnetic field is generated in accordance with Eq. (1.28), which in turn generates an electric field in accordance with Eq. (1.27), which in turn contributes to the generation of the magnetic field in accordance with Eq. (1.28), and so on, as depicted in Fig. 1.11. Note that **J** and ρ are coupled, since they must satisfy Eq. (1.31). Also, the magnetic field automatically satisfies Eq. (1.30), since Eq. (1.30) is not independent of Eq. (1.27).

The process depicted in Fig. 1.11 is exactly the phenomenon of electromagnetic waves propagating with a velocity (and other characteristics) determined by the parameters of the medium. In free space, the waves propagate unattenuated with the velocity $1/\sqrt{\mu_0\varepsilon_0}$, familiarly represented by the symbol c. If either the term $\partial\mathbf{B}/\partial t$ in Eq. (1.27) or the term $\partial\mathbf{D}/\partial t$ in Eq. (1.28) is not present, then wave propagation would not occur. As already stated in the previous section, it was through the addition of the term

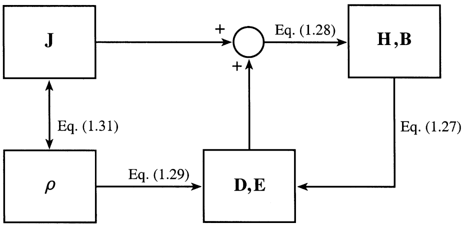


Figure 1.11 Generation of interdependent electric and magnetic fields, beginning with sources **J** and ρ.

$\partial\mathbf{D}/\partial t$ in Eq. (1.28) that Maxwell predicted electromagnetic wave propagation before it was confirmed experimentally.

Of particular importance is the case of time variations of the fields in the sinusoidal steady state, that is, the frequency domain case. In this connection, the frequency domain forms of Maxwell's equations are of interest. Using the phasor notation based on

$$A\ \cos(\omega t+\phi)=\mathrm{Re}\left[Ae^{j\phi}e^{j\omega t}\right]=\mathrm{Re}\left[\bar{A}e^{j\omega t}\right] \tag{1.32}$$

where $\bar{A}=Ae^{j\phi}$ is the phasor corresponding to the time function, we obtain these equations by replacing all field quantities in the time domain form of the equations by the corresponding phasor quantities and $\partial/\partial t$ by $j\omega$. Thus with the understanding that all phasor field quantities are functions of space coordinates, that is, $\bar{\mathbf{E}}=\bar{\mathbf{E}}(\mathbf{r})$, etc., we write the Maxwell's equations in frequency domain as

$$\nabla\times\bar{\mathbf{E}}=-j\omega\bar{\mathbf{B}} \tag{1.33}$$

$$\nabla\times\bar{\mathbf{H}}=\bar{\mathbf{J}}+j\omega\bar{\mathbf{D}} \tag{1.34}$$

$$\nabla\cdot\bar{\mathbf{D}}=\bar{\rho} \tag{1.35}$$

$$\nabla\cdot\bar{\mathbf{B}}=0 \tag{1.36}$$

Also, the continuity equation, Eq. (1.31), transforms to the frequency domain form

$$\nabla\cdot\bar{\mathbf{J}}+j\omega\bar{\rho}=0 \tag{1.37}$$

Note that since $\nabla\cdot\nabla\times\bar{\mathbf{E}}=0$, Eq. (1.36) follows from Eq. (1.33), and since $\nabla\cdot\nabla\times\bar{\mathbf{H}}=0$, Eq. (1.35) follows from Eq. (1.34) with the aid of Eq. (1.37).

Now the constitutive relations in phasor form are

$$\bar{\mathbf{D}}=\varepsilon\bar{\mathbf{E}} \tag{1.38a}$$

$$\bar{\mathbf{H}}=\frac{\bar{\mathbf{B}}}{\mu} \tag{1.38b}$$

$$\bar{\mathbf{J}}_c=\sigma\bar{\mathbf{E}} \tag{1.38c}$$

Substituting these into Eqs. (1.33)–(1.36), we obtain for a material medium characterized by the parameters ε, μ, and σ,

$$\nabla\times\bar{\mathbf{E}}=-j\omega\mu\bar{\mathbf{H}} \tag{1.39}$$

$$\nabla\times\bar{\mathbf{H}}=(\sigma+j\omega\varepsilon)\bar{\mathbf{E}} \tag{1.40}$$

$$\nabla\cdot\bar{\mathbf{H}}=0 \tag{1.41}$$

$$\nabla\cdot\bar{\mathbf{E}}=\frac{\bar{\rho}}{\varepsilon} \tag{1.42}$$

Note however that if the medium is homogeneous, that is, if the material parameters are independent of the space coordinates, Eq. (1.40) gives

$$\nabla \cdot \bar{\mathbf{E}} = \frac{1}{\sigma + j\omega\varepsilon} \nabla \cdot \nabla \times \bar{\mathbf{H}} = 0 \tag{1.43}$$

so that $\bar{\rho} = 0$ in such a medium.

A point of importance in connection with the frequency domain form of Maxwell's equations is that in these equations, the parameters ε, μ, and σ can be allowed to be functions of ω. In fact, for many dielectrics, the conductivity increases with frequency in such a manner that the quantity $\sigma/\omega\varepsilon$ is more constant than is the conductivity. This quantity is the ratio of the magnitudes of the two terms on the right side of Eq. (1.40), that is, the conduction current density term $\sigma\bar{\mathbf{E}}$ and the displacement current density term $j\omega\varepsilon\bar{\mathbf{E}}$.

1.2.3. Boundary Conditions

Maxwell's equations in differential form govern the interrelationships between the field vectors and the associated source densities at points in a given medium. For a problem involving two or more different media, the differential equations pertaining to each medium provide solutions for the fields that satisfy the characteristics of that medium. These solutions need to be matched at the boundaries between the media by employing "boundary conditions," which relate the field components at points adjacent to and on one side of a boundary to the field components at points adjacent to and on the other side of that boundary. The boundary conditions arise from the fact that the integral equations involve closed paths and surfaces and they must be satisfied for all possible closed paths and surfaces whether they lie entirely in one medium or encompass a portion of the boundary.

The boundary conditions are obtained by considering one integral equation at a time and applying it to a closed path or a closed surface encompassing the boundary, as shown in Fig. 1.12 for a plane boundary, and in the limit that the area enclosed by the closed path, or the volume bounded by the closed surface, goes to zero. Let the quantities pertinent to medium 1 be denoted by subscript 1 and the quantities pertinent to medium 2 be denoted by subscript 2, and $\mathbf{a}_n$ be the unit normal vector to the surface and directed into medium 1. Let all normal components at the boundary in both media be directed along $\mathbf{a}_n$ and denoted by an additional subscript n and all tangential components at the boundary in both media be denoted by an additional subscript t. Let the surface charge density (C/m^2) and the surface current density (A/m) on the boundary be ρ_S and $\mathbf{J}_S$, respectively. Then, the boundary conditions corresponding to the Maxwell's equations in integral form can be summarized as

$$\mathbf{a}_n \times (\mathbf{E}_1 - \mathbf{E}_2) = \mathbf{0} \tag{1.44a}$$

$$\mathbf{a}_n \times (\mathbf{H}_1 - \mathbf{H}_2) = \mathbf{J}_S \tag{1.44b}$$

$$\mathbf{a}_n \cdot (\mathbf{D}_1 - \mathbf{D}_2) = \rho_S \tag{1.44c}$$

$$\mathbf{a}_n \cdot (\mathbf{B}_1 - \mathbf{B}_2) = 0 \tag{1.44d}$$

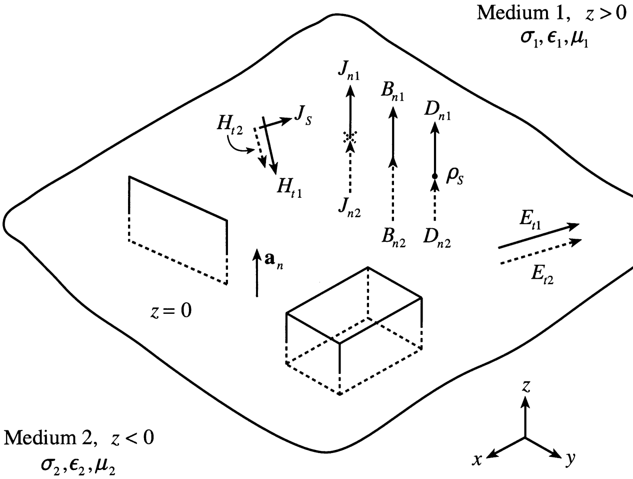


Figure 1.12 For deriving the boundary conditions at the interface between two arbitrary media.

or in scalar form,

$$E_{t1} - E_{t2} = 0 \tag{1.45a}$$

$$H_{t1} - H_{t2} = J_S \tag{1.45b}$$

$$D_{n1} - D_{n2} = \rho_S \tag{1.45c}$$

$$B_{n1} - B_{n2} = 0 \tag{1.45d}$$

In words, the boundary conditions state that at a point on the boundary, the tangential components of **E** and the normal components of **B** are continuous, whereas the tangential components of **H** are discontinuous by the amount equal to J_S at that point, and the normal components of **D** are discontinuous by the amount equal to ρ_S at that point, as illustrated in Fig. 1.12. It should be noted that the information concerning the direction of $\mathbf{J}_S$ relative to that of $(\mathbf{H}_1 - \mathbf{H}_2)$, which is contained in Eq. (1.44b), is not present in Eq. (1.45b). Hence, in general, Eq. (1.45b) is not sufficient and it is necessary to use Eq. (1.44b).

While Eqs. (1.44a)–(1.44d) or Eqs. (1.45a)–(1.45d) are the most commonly used boundary conditions, another useful boundary condition resulting from the law of conservation of charge is given by

$$\mathbf{a}_n \cdot (\mathbf{J}_1 - \mathbf{J}_2) = -\nabla_S \cdot \mathbf{J}_S - \frac{\partial \rho_S}{\partial t} \tag{1.46}$$

In words, Eq. (1.46) states that, at any point on the boundary, the components of $\mathbf{J}_1$ and $\mathbf{J}_2$ normal to the boundary are discontinuous by the amount equal to the negative of the sum of the two-dimensional divergence of the surface current density and the time derivative of the surface charge density at that point.

1.2.4. Electromagnetic Potentials and Potential Function Equations

Maxwell's equations in differential form, together with the constitutive relations and boundary conditions, allow for the unique determination of the fields **E**, **B**, **D**, and **H** for a given set of source distributions with densities **J** and ρ. An alternate approach involving the electric scalar potential Φ and the magnetic vector potential **A**, known together as the *electromagnetic potentials* from which the fields can be derived, simplifies the solution in some cases. This approach leads to solving two separate differential equations, one for Φ involving ρ alone, and the second for **A** involving **J** alone.

To obtain these equations, we first note that in view of Eq. (1.30), **B** can be expressed as the curl of another vector. Thus

$$\mathbf{B} = \nabla \times \mathbf{A} \tag{1.47}$$

Note that the units of **A** are the units of **B** times meter, that is, Wb/m. Now, substituting Eq. (1.47) into Eq. (1.27), interchanging the operations of $\partial/\partial t$ and curl, and rearranging, we obtain

$$\begin{aligned} &\nabla \times \left[\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t}\right] = 0 \\ &\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} = -\nabla\Phi \\ &\mathbf{E} = -\nabla\Phi - \frac{\partial \mathbf{A}}{\partial t} \end{aligned} \tag{1.48}$$

where the negative sign associated with $\nabla\Phi$ is chosen for a reason to be evident later in Sec. 1.3.2. Note that the units of Φ are the units of **E** times meter, that is, V. Note also that the knowledge of Φ and **A** enables the determination of **E** and **B**, from which **D** and **H** can be found by using the constitutive relations.

Now, using Eqs. (1.6) and (1.10) to obtain **D** and **H** in terms of Φ and **A** and substituting into Eqs. (1.29) and (1.28), we obtain

$$\nabla^2\Phi + \nabla\cdot\left[\frac{\partial \mathbf{A}}{\partial t}\right] = -\frac{\rho}{\varepsilon} \tag{1.49a}$$

$$\nabla \times \nabla \times \mathbf{A} + \mu\varepsilon\,\frac{\partial}{\partial t}\left[\nabla\Phi + \frac{\partial \mathbf{A}}{\partial t}\right] = \mu\mathbf{J} \tag{1.49b}$$

where we have assumed the medium to be homogeneous and isotropic, in addition to being linear. Using the vector identity

$$\nabla \times \nabla \times \mathbf{A} = \nabla(\nabla\cdot\mathbf{A}) - \nabla^2\mathbf{A} \tag{1.50}$$

and interchanging the operations of $\partial/\partial t$ and divergence or gradient depending on the term, and rearranging, we get

$$\nabla^2\Phi + \frac{\partial}{\partial t}(\nabla\cdot\mathbf{A}) = -\frac{\rho}{\varepsilon} \tag{1.51a}$$

$$\nabla^2\mathbf{A} - \nabla\left[\nabla\cdot\mathbf{A} + \mu\varepsilon\frac{\partial\Phi}{\partial t}\right] - \mu\varepsilon\,\frac{\partial^2\mathbf{A}}{\partial t^2} = -\mu\mathbf{J} \tag{1.51b}$$

These equations are coupled. To uncouple them, we make use of Helmholtz's theorem, which states that a vector field is completely specified by its curl and divergence. Therefore, since the curl of **A** is given by Eq. (1.47), we are at liberty to specify the divergence of **A**. We do this by setting

$$\nabla\cdot\mathbf{A} = -\mu\varepsilon\,\frac{\partial\Phi}{\partial t} \tag{1.52}$$

which is known as the Lorenz condition, resulting in the uncoupled equations

$$\nabla^2\Phi - \mu\varepsilon\,\frac{\partial^2\Phi}{\partial t^2} = -\frac{\rho}{\varepsilon} \tag{1.53}$$

$$\nabla^2\mathbf{A} - \mu\varepsilon\,\frac{\partial^2\mathbf{A}}{\partial t^2} = -\mu\mathbf{J} \tag{1.54}$$

which are called the *potential function equations*. While the Lorenz condition may appear to be arbitrary, it actually implies the continuity equation, which can be shown by taking the Laplacian on both sides of Eq. (1.52) and using Eqs. (1.53) and (1.54).

It can be seen that Eqs. (1.53) and (1.54) are not only uncoupled but they are also similar, particularly in Cartesian coordinates since Eq. (1.54) decomposes into three equations involving the three Cartesian components of **J**, each of which is similar to (1.53). By solving Eqs. (1.53) and (1.54), one can obtain the solutions for Φ and **A**, respectively, from which **E** and **B** can be found by using Eqs. (1.48) and (1.47), respectively. In practice, however, since ρ is related to **J** through the continuity equation, it is sufficient to find **B** from **A** obtained from the solution of Eq. (1.54) and then find **E** by using the Maxwell's equation for the curl of **H**, given by Eq. (1.28).

1.2.5. Power Flow and Energy Storage

A unique property of the electromagnetic field is its ability to transfer power between two points even in the absence of an intervening material medium. Without such ability, the effect of the field generated at one point will not be felt at another point, and hence the power generated at the first point cannot be put to use at the second point.

To discuss power flow associated with an electromagnetic field, we begin with the vector identity

$$\nabla\cdot(\mathbf{E}\times\mathbf{H}) = \mathbf{H}\cdot(\nabla\times\mathbf{E}) - \mathbf{E}\cdot(\nabla\times\mathbf{H}) \tag{1.55}$$

and make use of Maxwell's curl equations, Eqs. (1.27) and (1.28), to write

$$\nabla\cdot(\mathbf{E}\times\mathbf{H}) = -\mathbf{E}\cdot\mathbf{J} - \mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t} - \mathbf{H}\cdot\frac{\partial\mathbf{B}}{\partial t} \tag{1.56}$$

Allowing for conductivity of a material medium by denoting $\mathbf{J} = \mathbf{J}_0 + \mathbf{J}_c$, where $\mathbf{J}_0$ is that part of **J** that can be attributed to a source, and using the constitutive relations (1.5), (1.6), and (1.10), we obtain for a medium characterized by σ, ε, and μ,

$$-\mathbf{E}\cdot\mathbf{J}_0 = \sigma\varepsilon^2 + \frac{\partial}{\partial t}\left[\frac{1}{2}\varepsilon E^2\right] + \frac{\partial}{\partial t}\left[\frac{1}{2}\mu H^2\right] + \boldsymbol{\nabla}\cdot(\mathbf{E}\times\mathbf{H}) \tag{1.57}$$

Defining a vector $\mathbf{P}$ given by

$$\mathbf{P} = \mathbf{E}\times\mathbf{H} \tag{1.58}$$

and taking the volume integral of both sides of Eq. (1.58), we obtain

$$-\int_V (\mathbf{E}\cdot\mathbf{J}_0)\,dv = \int_V \sigma E^2\,dv + \frac{\partial}{\partial t}\int_V \left(\frac{1}{2}\varepsilon E^2\right)dv + \frac{\partial}{\partial t}\int_V \left(\frac{1}{2}\mu H^2\right)dv + \oint_S \mathbf{P}\cdot d\mathbf{S} \tag{1.59}$$

where we have also interchanged the differentiation operation with time and integration operation over volume in the second and third terms on the right side and used the divergence theorem for the last term.

In Eq. (1.59), the left side is the power supplied to the field by the current source $\mathbf{J}_0$ inside V. The quantities σE^2, $(1/2)\varepsilon E^2$, and $(1/2)\mu H^2$ are the power dissipation density (W/m^3), the electric stored energy density (J/m^3), and the magnetic stored energy density (J/m^3), respectively, due to the conductive, dielectric, and magnetic properties, respectively, of the medium. Hence, Eq. (1.59) says that the power delivered to the volume V by the current source $\mathbf{J}_0$ is accounted for by the power dissipated in the volume due to the conduction current in the medium, plus the time rates of increase of the energies stored in the electric and magnetic fields, plus another term, which we must interpret as the power carried by the electromagnetic field out of the volume V, for conservation of energy to be satisfied. It then follows that the vector $\mathbf{P}$ has the meaning of power flow density vector associated with the electromagnetic field. The statement represented by Eq. (1.59) is known as the *Poynting's theorem*, and the vector $\mathbf{P}$ is known as the *Poynting vector*. We note that the units of $\mathbf{E}\times\mathbf{H}$ are volts per meter times amperes per meter, or watts per square meter (W/m^2) and do indeed represent power density. In particular, since $\mathbf{E}$ and $\mathbf{H}$ are instantaneous field vectors, $\mathbf{E}\times\mathbf{H}$ represents the instantaneous Poynting vector. Note that the Poynting's theorem tells us only that the power flow out of a volume V is given by the surface integral of the Poynting vector over the surface S bounding that volume. Hence we can add to $\mathbf{P}$ any vector for which the surface integral over S vanishes, without affecting the value of the surface integral. However, generally, we are interested in the total power leaving a closed surface and the interpretation of $\mathbf{P}$ alone as representing the power flow density vector is sufficient.

For sinusoidally time-varying fields, that is, for the frequency domain case, the quantity of importance is the time-average Poynting vector instead of the instantaneous Poynting vector. We simply present the important relations here, without carrying out the derivations. The time-average Poynting vector, denoted by $\langle\mathbf{P}\rangle$, is given by

$$\langle\mathbf{P}\rangle = \mathrm{Re}\big[\bar{\mathbf{P}}\big] \tag{1.60}$$

where $\bar{\mathbf{P}}$ is the complex Poynting vector given by

$$\bar{\mathbf{P}} = \frac{1}{2}\bar{\mathbf{E}} \times \bar{\mathbf{H}}^* \tag{1.61}$$

where the star denotes complex conjugate. The Poynting theorem for the frequency domain case, known as the *complex Poynting's theorem*, is given by

$$-\int_V \left(\frac{1}{2}\bar{\mathbf{E}}\cdot\bar{\mathbf{J}}_0^*\right) dv = \int_V \langle p_d\rangle dv + j2\omega \int_V (\langle w_m\rangle - \langle w_e\rangle) dv + \oint_S \bar{\mathbf{P}}\cdot d\mathbf{S} \tag{1.62}$$

where

$$\langle p_d\rangle = \frac{1}{2}\sigma\bar{\mathbf{E}}\cdot\bar{\mathbf{E}}^* \tag{1.63a}$$

$$\langle w_e\rangle = \frac{1}{4}\varepsilon\bar{\mathbf{E}}\cdot\bar{\mathbf{E}}^* \tag{1.63b}$$

$$\langle w_m\rangle = \frac{1}{4}\mu\bar{\mathbf{H}}\cdot\bar{\mathbf{H}}^* \tag{1.63c}$$

are the time-average power dissipation density, the time-average electric stored energy density, and the time-average magnetic stored energy density, respectively. Equation (1.62) states that the time-average, or real, power delivered to the volume V by the current source is accounted for by the time-average power dissipated in the volume plus the time-average power carried by the electromagnetic field out of the volume through the surface S bounding the volume and that the reactive power delivered to the volume V by the current source is equal to the reactive power carried by the electromagnetic field out of the volume V through the surface S plus a quantity that is 2ω times the difference between the time-average magnetic and electric stored energies in the volume.

1.3. STATIC FIELDS, QUASISTATIC FIELDS, AND WAVES

1.3.1. Classification of Fields

While every macroscopic field obeys Maxwell's equations in their entirety, depending on their most dominant properties, it is sufficient to consider a subset of, or certain terms only, in the equations. The primary classification of fields is based on their time dependence. Fields which do not change with time are called *static*. Fields which change with time are called *dynamic*. Static fields are the simplest kind of fields, because for them $\partial/\partial t = 0$ and all terms involving differentiation with respect to time go to zero. Dynamic fields are the most complex, since for them Maxwell's equations in their entirety must be satisfied, resulting in wave type solutions, as provided by the qualitative explanation in Sec. 1.2.2. However, if certain features of the dynamic field can be analyzed as though the field were static, then the field is called *quasistatic*.

If the important features of the field are not amenable to static type field analysis, they are generally referred to as *time-varying*, although in fact, quasistatic fields are also time-varying. Since in the most general case, time-varying fields give rise to wave phenomena, involving velocity of propagation and time delay, it can be said that quasistatic fields are those time-varying fields for which wave propagation effects can be neglected.

1.3.2. Static Fields and Circuit Elements

For static fields, $\partial/\partial t = 0$. Maxwell's equations in integral form and the law of conservation of charge become

$$\oint_C \mathbf{E}\cdot d\mathbf{l} = 0 \tag{1.64a}$$

$$\oint_C \mathbf{H}\cdot d\mathbf{l} = \int_S \mathbf{J}\cdot d\mathbf{S} \tag{1.64b}$$

$$\oint_S \mathbf{D}\cdot d\mathbf{S} = \int_V \rho\, dv \tag{1.64c}$$

$$\oint_S \mathbf{B}\cdot d\mathbf{S} = 0 \tag{1.64d}$$

$$\oint_S \mathbf{J}\cdot d\mathbf{S} = 0 \tag{1.64e}$$

whereas Maxwell's equations in differential form and the continuity equation reduce to

$$\nabla \times \mathbf{E} = 0 \tag{1.65a}$$

$$\nabla \times \mathbf{H} = \mathbf{J} \tag{1.65b}$$

$$\nabla \cdot \mathbf{D} = \rho \tag{1.65c}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{1.65d}$$

$$\nabla \cdot \mathbf{J} = 0 \tag{1.65e}$$

Immediately, one can see that, unless **J** includes a component due to conduction current, the equations involving the electric field are completely independent of those involving the magnetic field. Thus the fields can be subdivided into *static electric fields*, or *electrostatic fields*, governed by Eqs. (1.64a) and (1.64c), or Eqs. (1.65a) and (1.65c), and *static magnetic fields*, or *magnetostatic fields*, governed by Eqs. (1.64b) and (1.64d), or Eqs. (1.65b) and (1.65d). The source of a static electric field is ρ, whereas the source of a static magnetic field is **J**. One can also see from Eq. (1.64e) or (1.65e) that no relationship exists between **J** and ρ. If **J** includes a component due to conduction current, then since $\mathbf{J}_c = \sigma\mathbf{E}$, a coupling between the electric and magnetic fields exists for that part of the total field associated with $\mathbf{J}_c$. However, the coupling is only one way, since the right side of Eq. (1.64a) or (1.65a) is still zero. The field is then referred to as *electromagnetostatic field.* It can also be seen then that for consistency, the right sides of Eqs. (1.64c) and (1.65c) must be zero, since the right sides of Eqs. (1.64e) and (1.65e) are zero. We shall now consider each of the three types of static fields separately and discuss some fundamental aspects.

Electrostatic Fields and Capacitance

The equations of interest are Eqs. (1.64a) and (1.64c), or Eqs. (1.65a) and (1.65c). The first of each pair of these equations simply tells us that the electrostatic field is a conservative field, and the second of each pair of these equations enables us, in principle, to determine the electrostatic field for a given charge distribution. Alternatively, the potential function equation, Eq. (1.53), which reduces to

$$\nabla^2\Phi = -\frac{\rho}{\varepsilon} \tag{1.66}$$

can be used to find the electric scalar potential, Φ, from which the electrostatic field can be determined by using Eq. (1.48), which reduces to

$$\mathbf{E} = -\nabla\Phi \tag{1.67}$$

Equation (1.66) is known as the *Poisson's equation*, which automatically includes the condition that the field be conservative. It is worth noting that the potential difference between two points A and B in the static electric field, which is independent of the path followed from A to B because of the conservative nature of the field is

$$\begin{aligned}\int_A^B \mathbf{E}\cdot d\mathbf{l} &= \int_A^B [-\nabla\Phi]\cdot d\mathbf{l}\\ &= \Phi_A - \Phi_B\end{aligned} \tag{1.68}$$

the difference between the value of Φ at A and the value of Φ at B. The choice of minus sign associated with $\nabla\Phi$ in Eq. (1.48) is now evident.

The solution to Poisson's equation, Eq. (1.66), for a given charge density distribution $\rho(\mathbf{r})$ is given by

$$\Phi(\mathbf{r}) = \frac{1}{4\pi\varepsilon}\int_{V'} \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,dv' \tag{1.69}$$

where the prime denotes source point and no prime denotes field point. Although cast in terms of volume charge density, Eq. (1.69) can be formulated in terms of a surface charge distribution, a line charge distribution, or a collection of point charges. In particular, for a point charge $Q(\mathbf{r}')$, the solution is given by

$$\Phi(\mathbf{r}) = \frac{Q(\mathbf{r}')}{4\pi\varepsilon|\mathbf{r}-\mathbf{r}'|} \tag{1.70}$$

It follows from Eq. (1.67) that the electric field intensity due to the point charge is given by

$$\mathbf{E}(\mathbf{r}) = \frac{Q(\mathbf{r}')\,(\mathbf{r}-\mathbf{r}')}{4\pi\varepsilon\,|\mathbf{r}-\mathbf{r}'|^3} \tag{1.71}$$

which is exactly the expression that results from Coulomb's law for the electric force between two point charges.

Equation (1.69) or its alternate forms can be used to solve two types of problems:

1. finding the electrostatic potential for a specified charge distribution by evaluating the integral on the right side, which is a straightforward process with the help of a computer but can be considerably difficult analytically except for a few examples, and
2. finding the surface charge distribution on the surfaces of an arrangement of conductors raised to specified potentials, by inversion of the equation, which is the basis for numerical solution by the well-known *method of moments*.

In the case of type 1, the electric field can then be found by using Eq. (1.67).

In a charge-free region, $\rho = 0$, and Poisson's equation, Eq. (1.66), reduces to

$$\nabla^2 \Phi = 0 \tag{1.72}$$

which is known as the *Laplace equation*. The field is then due to charges outside the region, such as surface charge on conductors bounding the region. The situation is then one of solving a boundary value problem. In general, for arbitrarily shaped boundaries, a numerical technique, such as the *method of finite differences*, is employed for solving the problem. Here, we consider analytical solution involving one-dimensional variation of Φ.

A simple example is that of the parallel-plate arrangement shown in Fig. 1.13a, in which two parallel, perfectly conducting plates ($\sigma = \infty$, $\mathbf{E} = 0$) of dimensions w along the y direction and l along the z direction lie in the $x = 0$ and $x = d$ planes. The region between the plates is a perfect dielectric ($\sigma = 0$) of material parameters ε and μ. The thickness of the plates is shown exaggerated for convenience in illustration. A potential difference of V_0 is maintained between the plates by connecting a direct voltage source at the end $z = -l$. If fringing of the field due to the finite dimensions of the structure normal to the x direction is neglected, or if it is assumed that the structure is part of one which is infinite in extent

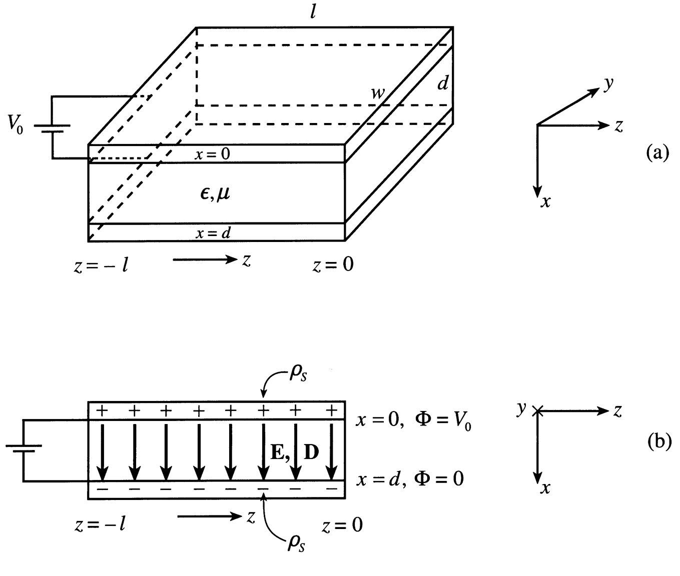


Figure 1.13 Electrostatic field in a parallel-plate arrangement.

normal to the x direction, then the problem can be treated as one-dimensional with x as the variable, and Eq. (1.72) reduces to

$$\frac{d^2\Phi}{dx^2} = 0 \tag{1.73}$$

The solution for the potential in the charge-free region between the plates is given by

$$\Phi(x) = \frac{V_0}{d}(d - x) \tag{1.74}$$

which satisfies Eq. (1.73), as well as the boundary conditions of $\Phi = 0$ at $x = d$ and $\Phi = V_0$ at $x = 0$. The electric field intensity between the plates is then given by

$$\mathbf{E} = -\nabla\Phi = \frac{V_0}{d}\mathbf{a}_x \tag{1.75}$$

as depicted in the cross-sectional view in Fig. 1.13b, and resulting in displacement flux density

$$\mathbf{D} = \frac{\varepsilon V_0}{d}\mathbf{a}_x \tag{1.76}$$

Then, using the boundary condition for the normal component of **D** given by Eq. (1.44c) and noting that there is no field inside the conductor, we obtain the magnitude of the charge on either plate to be

$$Q = \left(\frac{\varepsilon V_0}{d}\right)(wl) = \frac{\varepsilon wl}{d} V_0 \tag{1.77}$$

We can now find the familiar circuit parameter, the capacitance, C, of the parallel-plate arrangement, which is defined as the ratio of the magnitude of the charge on either plate to the potential difference V_0. Thus

$$C = \frac{Q}{V_0} = \frac{\varepsilon wl}{d} \tag{1.78}$$

Note that the units of C are the units of ε times meter, that is, farads. The phenomenon associated with the arrangement is that energy is stored in the capacitor in the form of electric field energy between the plates, as given by

$$\begin{aligned} W_e &= \left(\frac{1}{2}\varepsilon E_x^2\right)(wld) \\ &= \frac{1}{2}\left(\frac{\varepsilon wl}{d}\right)V_0^2 \\ &= \frac{1}{2}CV_0^2 \end{aligned} \tag{1.79}$$

the familiar expression for energy stored in a capacitor.

Magnetostatic Fields and Inductance

The equations of interest are Eqs. (1.64b) and (1.64d) or Eqs. (1.65b) and (1.65d). The second of each pair of these equations simply tells us that the magnetostatic field is solenoidal, which as we know holds for any magnetic field, and the first of each pair of these equations enables us, in principle, to determine the magnetostatic field for a given current distribution. Alternatively, the potential function equation, Eq. (1.54), which reduces to

$$\nabla^2\mathbf{A} = -\mu\mathbf{J} \tag{1.80}$$

can be used to find the magnetic vector potential, **A**, from which the magnetostatic field can be determined by using Eq. (1.47). Equation (1.80) is the Poisson's equation for the magnetic vector potential, which automatically includes the condition that the field be solenoidal.

The solution to Eq. (1.80) for a given current density distribution $\mathbf{J}(\mathbf{r})$ is, purely from analogy with the solution Eq. (1.69) to Eq. (1.66), given by

$$\mathbf{A}(\mathbf{r}) = \frac{\mu}{4\pi}\int_{V'} \frac{\mathbf{J}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, dv' \tag{1.81}$$

Although cast in terms of volume current density, Eq. (1.81) can be formulated in terms of a surface current density, a line current, or a collection of infinitesimal current elements. In particular, for an infinitesimal current element $I\,d\mathbf{l}(\mathbf{r}')$, the solution is given by

$$\mathbf{A}(\mathbf{r}) = \frac{\mu I\, d\mathbf{l}(\mathbf{r}')}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{1.82}$$

It follows from Eq. (1.47) that the magnetic flux density due to the infinitesimal current element is given by

$$\mathbf{B}(\mathbf{r}) = \frac{\mu I\, d\mathbf{l}(\mathbf{r}') \times (\mathbf{r}-\mathbf{r}')}{4\pi|\mathbf{r}-\mathbf{r}'|^3} \tag{1.83}$$

which is exactly the law of Biot-Savart that results from Ampere's force law for the magnetic force between two current elements. Similar to that in the case of Eq. (1.69), Eq. (1.81) or its alternate forms can be used to find the magnetic vector potential and then the magnetic field by using Eq. (1.47) for a specified current distribution.

In a current-free region, $\mathbf{J} = 0$, and Eq. (1.80) reduces to

$$\nabla^2\mathbf{A} = 0 \tag{1.84}$$

The field is then due to currents outside the region, such as surface currents on conductors bounding the region. The situation is then one of solving a boundary value problem as in the case of Eq. (1.72). However, since the boundary condition Eq. (1.44b) relates the magnetic field directly to the surface current density, it is straightforward and more convenient to determine the magnetic field directly by using Eqs. (1.65b) and (1.65d).

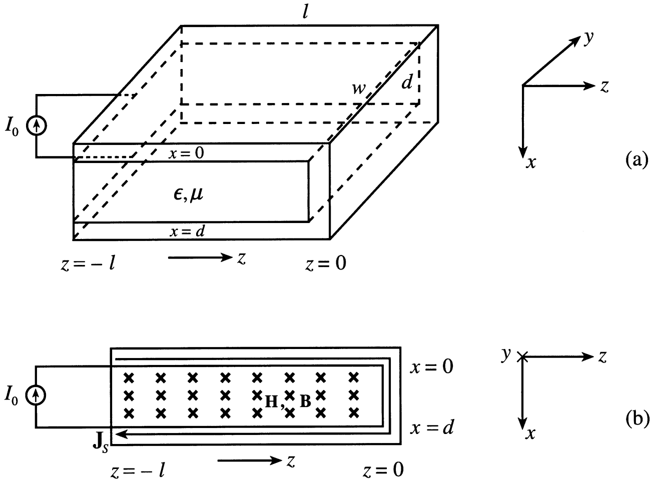


Figure 1.14 Magnetostatic field in a parallel-plate arrangement.

A simple example is that of the parallel-plate arrangement of Fig. 1.13a with the plates connected by another conductor at the end $z=0$ and driven by a source of direct current I_0 at the end $z=-l$, as shown in Fig. 1.14a. If fringing of the field due to the finite dimensions of the structure normal to the x direction is neglected, or if it is assumed that the structure is part of one which is infinite in extent normal to the x direction, then the problem can be treated as one-dimensional with x as the variable and we can write the current density on the plates to be

$$\mathbf{J}_S = \begin{cases} \left(\dfrac{I_0}{w}\right)\mathbf{a}_z & \text{on the plate } x = 0 \\ \left(\dfrac{I_0}{w}\right)\mathbf{a}_x & \text{on the plate } z = 0 \\ -\left(\dfrac{I_0}{w}\right)\mathbf{a}_z & \text{on the plate } x = d \end{cases} \tag{1.85}$$

In the current-free region between the plates, Eq. (1.65b) reduces to

$$\begin{vmatrix} \mathbf{a}_x & \mathbf{a}_y & \mathbf{a}_z \\ \dfrac{\partial}{\partial x} & 0 & 0 \\ H_x & H_y & H_z \end{vmatrix} = 0 \tag{1.86}$$

and Eq. (1.65d) reduces to

$$\frac{\partial B_x}{\partial x} = 0 \tag{1.87}$$

so that each component of the field, if it exists, has to be uniform. This automatically forces H_x and H_z to be zero since nonzero value of these components do not satisfy the boundary conditions Eqs. (1.44b) and (1.44d) on the plates, keeping in mind that the field is entirely in the region between the conductors. Thus, as depicted in the cross-sectional view in Fig. 1.14b,

$$\mathbf{H} = \frac{I_0}{w}\mathbf{a}_y \tag{1.88}$$

which satisfies the boundary condition Eq. (1.44b) on all three plates, and results in magnetic flux density

$$\mathbf{B} = \frac{\mu I_0}{w}\mathbf{a}_y \tag{1.89}$$

The magnetic flux, ψ, linking the current I_0, is then given by

$$\psi = \left(\frac{\mu I_0}{w}\right)(dl) = \frac{\mu dl}{w}I_0 \tag{1.90}$$

We can now find the familiar circuit parameter, the inductance, L, of the parallel-plate arrangement, which is defined as the ratio of the magnetic flux linking the current to the current. Thus

$$L = \frac{\psi}{I_0} = \frac{\mu dl}{w} \tag{1.91}$$

Note that the units of L are the units of μ times meter, that is, henrys. The phenomenon associated with the arrangement is that energy is stored in the inductor in the form of magnetic field energy between the plates, as given by

$$\begin{aligned} W_m &= \left(\frac{1}{2}\mu H^2\right)wld \\ &= \frac{1}{2}\left(\frac{\mu dl}{w}\right)I_0^2 \\ &= \frac{1}{2}LI_0^2 \end{aligned} \tag{1.92}$$

the familiar expression for energy stored in an inductor.

Electromagnetostatic Fields and Conductance

The equations of interest are

$$\oint_C \mathbf{E}\cdot d\mathbf{l} = 0 \tag{1.93a}$$

$$\oint_C \mathbf{H}\cdot d\mathbf{l} = \int_S \mathbf{J}_c\cdot d\mathbf{S} = \sigma\int_S \mathbf{E}\cdot d\mathbf{S} \tag{1.93b}$$

$$\oint_S \mathbf{D}\cdot d\mathbf{S} = 0 \tag{1.93c}$$

$$\oint_S \mathbf{B}\cdot d\mathbf{S} = 0 \tag{1.93d}$$

or in differential form,

$$\nabla \times \mathbf{E} = 0 \tag{1.94a}$$

$$\nabla \times \mathbf{H} = \mathbf{J}_\mathrm{c} = \sigma\mathbf{E} \tag{1.94b}$$

$$\nabla \cdot \mathbf{D} = 0 \tag{1.94c}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{1.94d}$$

From Eqs. (1.94a) and (1.94c), we note that Laplace's equation, Eq. (1.72), for the electrostatic potential is satisfied, so that, for a given problem, the electric field can be found in the same manner as in the case of the example of Fig. 1.13. The magnetic field is then found by using Eq. (1.94b) and making sure that Eq. (1.94d) is also satisfied.

A simple example is that of the parallel-plate arrangement of Fig. 1.13a but with an imperfect dielectric material of parameters σ, ε, and μ, between the plates, as shown in Fig. 1.15a. Then, the electric field between the plates is the same as that given by Eq. (1.75), that is,

$$\mathbf{E} = \frac{V_0}{d}\mathbf{a}_x \tag{1.95}$$

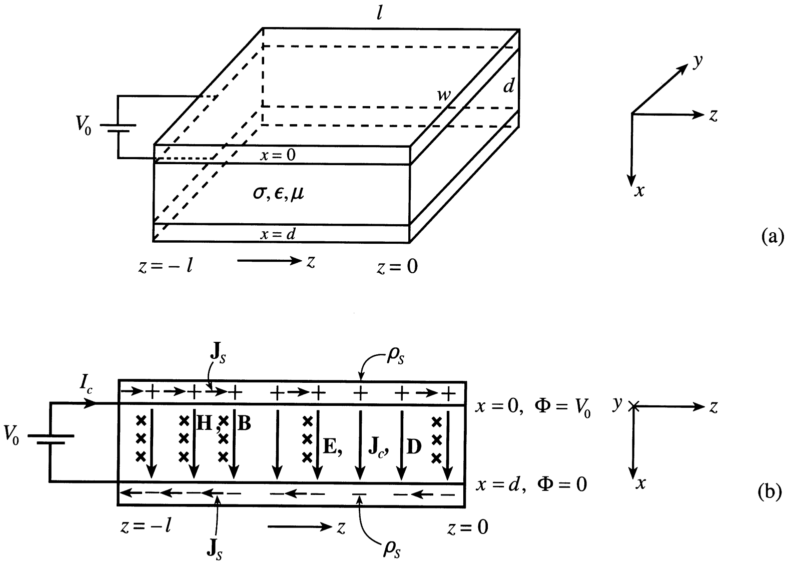


Figure 1.15 Electromagnetostatic field in a parallel-plate arrangement.

resulting in a conduction current of density

$$\mathbf{J}_c = \frac{\sigma V_0}{d}\mathbf{a}_x \tag{1.96}$$

from the top plate to the bottom plate, as depicted in the cross-sectional view of Fig. 1.15b. Since $\partial\rho/\partial t = 0$ at the boundaries between the plates and the slab, continuity of current is satisfied by the flow of surface current on the plates. At the input $z = -l$, this surface current, which is the current drawn from the source, must be equal to the total current flowing from the top to the bottom plate. It is given by

$$I_c = \left(\frac{\sigma V_0}{d}\right)(wl) = \frac{\sigma wl}{d} V_0 \tag{1.97}$$

We can now find the familiar circuit parameter, the conductance, G, of the parallel-plate arrangement, which is defined as the ratio of the current drawn from the source to the source voltage V_0. Thus

$$G = \frac{I_c}{V_0} = \frac{\sigma wl}{d} \tag{1.98}$$

Note that the units of G are the units of σ times meter, that is, siemens (S). The reciprocal quantity, R, the resistance of the parallel-plate arrangement, is given by

$$R = \frac{V_0}{I_c} = \frac{d}{\sigma wl} \tag{1.99}$$

The unit of R is ohms. The phenomenon associated with the arrangement is that power is dissipated in the material between the plates, as given by

$$\begin{aligned} P_d &= (\sigma E^2)(wld) \\ &= \left(\frac{\sigma wl}{d}\right) V_0^2 \\ &= G\,V_0^2 \\ &= \frac{V_0^2}{R} \end{aligned} \tag{1.100}$$

the familiar expression for power dissipated in a resistor.

Proceeding further, we find the magnetic field between the plates by using Eq. (1.94b), and noting that the geometry of the situation requires a y component of $\mathbf{H}$, dependent on z, to satisfy the equation. Thus

$$\mathbf{H} = H_y(z)\,\mathbf{a}_y \tag{1.101a}$$

$$\frac{\partial H_y}{\partial z} = -\frac{\sigma V_0}{d} \tag{1.101b}$$

$$\mathbf{H} = -\frac{\sigma V_0}{d} z\, \mathbf{a}_y \tag{1.101c}$$

where the constant of integration is set to zero, since the boundary condition at $z=0$ requires H_y to be zero for z equal to zero. Note that the magnetic field is directed in the positive y direction (since z is negative) and increases linearly from $z=0$ to $z=-l$, as depicted in Fig. 1.15b. It also satisfies the boundary condition at $z=-l$ by being consistent with the current drawn from the source to be $w\left[H_y\right]_{z=-1} = (\sigma V_0/d)(wl) = I_c$.

Because of the existence of the magnetic field, the arrangement is characterized by an inductance, which can be found either by using the flux linkage concept or by the energy method. To use the flux linkage concept, we recognize that a differential amount of magnetic flux $d\psi' = \mu H_y d(dz')$ between z equal to $(z' - dz')$ and z equal to z', where $-l < z' < 0$, links only that part of the current that flows from the top plate to the bottom plate between $z = z'$ and $z = 0$, thereby giving a value of $(-z'/l)$ for the fraction, N, of the total current linked. Thus, the inductance, familiarly known as the *internal inductance*, denoted L_i, since it is due to magnetic field internal to the current distribution, as compared to that in Eq. (1.91) for which the magnetic field is external to the current distribution, is given by

$$\begin{aligned} L_i &= \frac{1}{I_c} \int_{z'=-l}^{0} N\, d\psi' \\ &= \frac{1}{3} \frac{\mu dl}{w} \end{aligned} \tag{1.102}$$

or 1/3 times the inductance of the structure if $\sigma = 0$ and the plates are joined at $z = 0$, as in Fig. 1.14b.

Alternatively, if the energy method is used by computing the energy stored in the magnetic field and setting it equal to (1/2) $L_i I_c^2$, then we have

$$\begin{aligned} L_i &= \frac{1}{I_c^2}(dw) \int_{z=-l}^{0} \mu H_y^2\, dz \\ &= \frac{1}{3} \frac{\mu dl}{w} \end{aligned} \tag{1.103}$$

same as in Eq. (1.102).

Finally, recognizing that there is energy storage associated with the electric field between the plates, we note that the arrangement has also associated with it a capacitance C, equal to $\varepsilon wl/d$. Thus, all three properties of conductance, capacitance, and inductance are associated with the structure. Since for $\sigma = 0$ the situation reduces to that of Fig. 1.13, we can represent the arrangement of Fig. 1.15 to be equivalent to the circuit shown in Fig. 1.16. Note that the capacitor is charged to the voltage V_0 and the current through it is zero (open circuit condition). The voltage across the inductor is zero (short circuit condition), and the current through it is V_0/R. Thus, the current drawn from the voltage

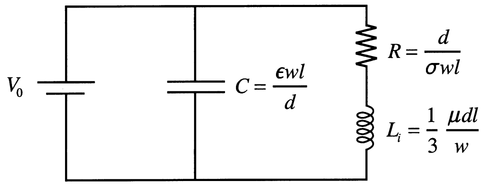

V_0
$C = \frac{\epsilon wl}{d}$
$R = \frac{d}{\sigma wl}$
$L_i = \frac{1}{3}\frac{\mu dl}{w}$

Figure 1.16 Circuit equivalent for the arrangement of Fig. 1.15.

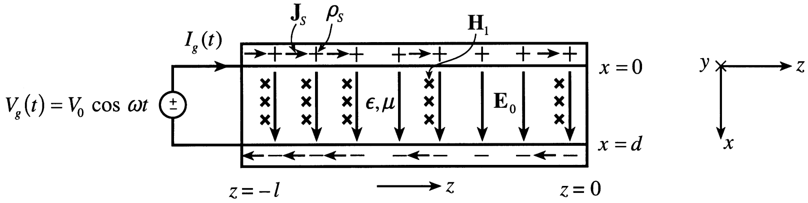


Figure 1.17 Electroquasistatic field analysis for the parallel-plate structure of Fig. 1.13.

source is V_0/R and the voltage source views a single resistor R, as far as the current drawn from it is concerned.

1.3.3. Quasistatic Fields and Low-frequency Behavior

As mentioned in Sec. 1.3.1, quasistatic fields are a class of dynamic fields for which certain features can be analyzed as though the fields were static. In terms of behavior in the frequency domain, they are low-frequency extensions of static fields present in a physical structure, when the frequency of the source driving the structure is zero, or low-frequency approximations of time-varying fields in the structure that are complete solutions to Maxwell's equations. Here, we use the approach of low-frequency extensions of static fields. Thus, for a given structure, we begin with a time-varying field having the same spatial characteristics as that of the static field solution for the structure and obtain field solutions containing terms up to and including the first power (which is the lowest power) in ω for their amplitudes. Depending on whether the predominant static field is electric or magnetic, quasistatic fields are called *electroquasistatic fields* or *magnetoquasistatic fields.* We shall now consider these separately.

Electroquasistatic Fields

For electroquasistatic fields, we begin with the electric field having the spatial dependence of the static field solution for the given arrangement. An example is provided by the arrangement in Fig. 1.13a excited by a sinusoidally time-varying voltage source $V_g(t) = V_0 \cos \omega t$, instead of a direct voltage source, as shown by the cross-sectional view in Fig. 1.17. Then,

$$\mathbf{E}_0 = \frac{V_0}{d} \cos \omega t \; \mathbf{a}_x \tag{1.104}$$

where the subscript 0 denotes that the amplitude of the field is of the zeroth power in ω. This results in a magnetic field in accordance with Maxwell's equation for the curl of **H**, given by Eq. (1.28). Thus, noting that $\mathbf{J}=0$ in view of the perfect dielectric medium, we have for the geometry of the arrangement,

$$\frac{\partial H_{y1}}{\partial z} = -\frac{\partial D_{x0}}{\partial t} = \frac{\omega \varepsilon V_0}{d} \sin \omega t \tag{1.105}$$

$$\mathbf{H}_1 = \frac{\omega \varepsilon V_0 z}{d} \sin \omega t \, \mathbf{a}_y \tag{1.106}$$

where we have also satisfied the boundary condition at $z=0$ by choosing the constant of integration such that $\left[H_{y1}\right]_{z=0}$ is zero, and the subscript 1 denotes that the amplitude of the field is of the first power in ω. Note that the amplitude of H_{y1} varies linearly with z, from zero at $z=0$ to a maximum at $z=-l$.

We stop the solution here, because continuing the process by substituting Eq. (1.106) into Maxwell's curl equation for **E**, Eq. (1.27) to obtain the resulting electric field will yield a term having amplitude proportional to the second power in ω. This simply means that the fields given as a pair by Eqs. (1.104) and (1.106) do not satisfy Eq. (1.27) and hence are not complete solutions to Maxwell's equations. The complete solutions are obtained by solving Maxwell's equations simultaneously and subject to the boundary conditions for the given problem.

Proceeding further, we obtain the current drawn from the voltage source to be

$$\begin{aligned} I_g(t) &= w\left[H_{y1}\right]_{z=-l} \\ &= -\omega\left(\frac{\varepsilon w l}{d}\right) V_0 \sin \omega t \\ &= C\frac{dV_g(t)}{dt} \end{aligned} \tag{1.107}$$

or,

$$\bar{I}_g = j\omega C \bar{V}_g \tag{1.108}$$

where $C=\varepsilon wl/d$ is the capacitance of the arrangement obtained from static field considerations. Thus, the input admittance of the structure is $j\omega C$ so that its low-frequency input behavior is essentially that of a single capacitor of value same as that found from static field analysis of the structure. Indeed, from considerations of power flow, using Poynting's theorem, we obtain the power flowing into the structure to be

$$\begin{aligned} P_{\text{in}} &= wd\left[E_{x0}H_{y1}\right]_{z=0} \\ &= -\left(\frac{\varepsilon w l}{d}\right)\omega V_0^2 \sin \omega t \cos \omega t \\ &= \frac{d}{dt}\left(\frac{1}{2}CV_g^2\right) \end{aligned} \tag{1.109}$$

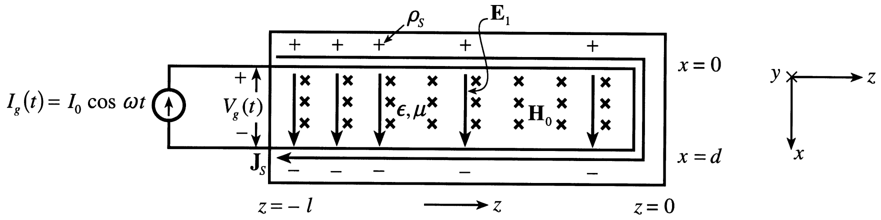


Figure 1.18 Magnetoquasistatic field analysis for the parallel-plate structure of Fig. 1.14.

which is consistent with the electric energy stored in the structure for the static case, as given by Eq. (1.79).

Magnetoquasistatic Fields

For magnetoquasistatic fields, we begin with the magnetic field having the spatial dependence of the static field solution for the given arrangement. An example is provided by the arrangement in Fig. 1.14a excited by a sinusoidally time-varying current source $I_g(t) = I_0 \cos \omega t$, instead of a direct current source, as shown by the cross-sectional view in Fig. 1.18. Then,

$$\mathbf{H}_0 = \frac{I_0}{w} \cos \omega t \, \mathbf{a}_y \tag{1.110}$$

where the subscript 0 again denotes that the amplitude of the field is of the zeroth power in ω. This results in an electric field in accordance with Maxwell's curl equation for **E**, given by Eq. (1.27). Thus, we have for the geometry of the arrangement,

$$\frac{\partial E_{x1}}{\partial z} = -\frac{\partial B_{y0}}{\partial t} = \frac{\omega \mu I_0}{w} \sin \omega t \tag{1.111}$$

$$\mathbf{E}_1 = \frac{\omega \mu I_0 z}{w} \sin \omega t \, \mathbf{a}_x \tag{1.112}$$

where we have also satisfied the boundary condition at $z = 0$ by choosing the constant of integration such that $[E_{x1}]_{z=0} = 0$ is zero, and again the subscript 1 denotes that the amplitude of the field is of the first power in ω. Note that the amplitude of E_{x1} varies linearly with z, from zero at $z = 0$ to a maximum at $z = -l$.

As in the case of electroquasistatic fields, we stop the process here, because continuing it by substituting Eq. (1.112) into Maxwell's curl equation for **H**, Eq. (1.28), to obtain the resulting magnetic field will yield a term having amplitude proportional to the second power in ω. This simply means that the fields given as a pair by Eqs. (1.110) and (1.112) do not satisfy Eq. (1.28), and hence are not complete solutions to Maxwell's equations. The complete solutions are obtained by solving Maxwell's equations simultaneously and subject to the boundary conditions for the given problem.

Proceeding further, we obtain the voltage across the current source to be

$$\begin{aligned} V_g(t) &= d[E_{x1}]_{z=-l} \\ &= -\omega\left(\frac{\mu\, dl}{w}\right) I_0 \sin \omega t \\ &= L\frac{dI_g(t)}{dt} \end{aligned} \tag{1.113}$$

or

$$\bar{V}_g = j\omega L\, \bar{I}_g \tag{1.114}$$

where $L = \mu dl/w$ is the inductance of the arrangement obtained from static field considerations. Thus, the input impedance of the structure is $j\omega L$, such that its low-frequency input behavior is essentially that of a single inductor of value the same as that found from static field analysis of the structure. Indeed, from considerations of power flow, using Poynting's theorem, we obtain the power flowing into the structure to be

$$\begin{aligned} P_{\text{in}} &= wd\left[E_{x1} H_{y0}\right]_{z=-l} \\ &= -\left(\frac{\mu dl}{w}\right)\omega I_0^2 \sin \omega t \cos \omega t \\ &= \frac{d}{dt}\left(\frac{1}{2} L I_g^2\right) \end{aligned} \tag{1.115}$$

which is consistent with the magnetic energy stored in the structure for the static case, as given by Eq. (1.92).

Quasistatic Fields in a Conductor

If the dielectric slab in the arrangement of Fig. 1.17 is conductive, as shown in Fig. 1.19a, then both electric and magnetic fields exist in the static case because of the conduction current, as discussed under electromagnetostatic fields in Sec. 1.3.2. Furthermore, the electric field of amplitude proportional to the first power in ω contributes to the creation of magnetic field of amplitude proportional to the first power in ω, in addition to that from electric field of amplitude proportional to the zeroth power in ω.

Thus, using the results from the static field analysis for the arrangement of Fig. 1.15, we have for the arrangement of Fig. 1.19a

$$\mathbf{E}_0 = \frac{V_0}{d} \cos \omega t\, \mathbf{a}_x \tag{1.116}$$

$$\mathbf{J}_{c0} = \sigma \mathbf{E}_0 = \frac{\sigma V_0}{d} \cos \omega t\, \mathbf{a}_x \tag{1.117}$$

$$\mathbf{H}_0 = -\frac{\sigma V_0 z}{d} \cos \omega t\, \mathbf{a}_y \tag{1.118}$$

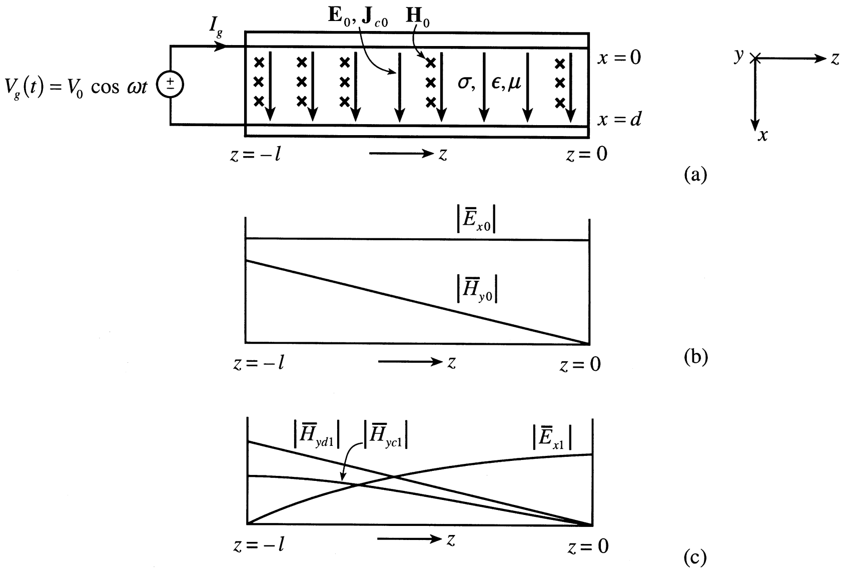


Figure 1.19 (a) Zero-order fields for the parallel-plate structure of Fig. 1.15. (b) Variations of amplitudes of the zero-order fields along the structure. (c) Variations of amplitudes of the first-order fields along the structure.

as depicted in the figure. Also, the variations with z of the amplitudes of E_{x0} and H_{y0} are shown in Fig. 1.19b.

The magnetic field given by Eq. (1.118) gives rise to an electric field having amplitude proportional to the first power in ω, in accordance with Maxwell's curl equation for **E**, Eq. (1.27). Thus

$$\frac{\partial E_{x1}}{\partial z} = -\frac{\partial B_{y0}}{\partial t} = -\frac{\omega\mu\sigma V_0 z}{d} \sin \omega t \tag{1.119}$$

$$E_{x1} = -\frac{\omega\mu\sigma V_0}{2d}\left(z^2 - l^2\right) \sin \omega t \tag{1.120}$$

where we have also made sure that the boundary condition at $z = -l$ is satisfied. This boundary condition requires that E_x be equal to V_g/d at $z = -l$. Since this is satisfied by E_{x0} alone, it follows that E_{x1} must be zero at $z = -l$.

The electric field given by Eq. (1.116) and that given by Eq. (1.120) together give rise to a magnetic field having terms with amplitudes proportional to the first power in ω, in accordance with Maxwell's curl equation for **H**, Eq. (1.28). Thus

$$\begin{aligned}\frac{\partial H_{y1}}{\partial z} &= -\sigma E_{x1} - \varepsilon \frac{\partial E_{x0}}{\partial t} \\ &= \frac{\omega\mu\sigma^2 V_0}{2d}\left(z^2 - l^2\right) \sin \omega t + \frac{\omega\varepsilon V_0}{d} \sin \omega t \end{aligned} \tag{1.121}$$

$$H_{y1} = \frac{\omega\mu\sigma^2 V_0\left(z^3 - 3zl^2\right)}{6d} \sin \omega t + \frac{\omega\varepsilon V_0 z}{d} \sin \omega t \tag{1.122}$$

where we have also made sure that the boundary condition at $z=0$ is satisfied. This boundary condition requires that H_y be equal to zero at $z=0$, which means that all of its terms must be zero at $z=0$. Note that the first term on the right side of Eq. (1.122) is the contribution from the conduction current in the material resulting from E_{x1} and the second term is the contribution from the displacement current resulting from E_{x0}. Denoting these to be H_{yc1} *and* H_{yd1}, respectively, we show the variations with z of the amplitudes of all the field components having amplitudes proportional to the first power in ω, in Fig. 1.19c.

Now, adding up the contributions to each field, we obtain the total electric and magnetic fields up to and including the terms with amplitudes proportional to the first power in ω to be

$$E_x = \frac{V_0}{d}\cos\omega t - \frac{\omega\mu\sigma V_0}{2d}\left(z^2 - l^2\right)\sin\omega t \tag{1.123a}$$

$$H_y = -\frac{\sigma V_0 z}{d}\cos\omega t + \frac{\omega\varepsilon V_0 z}{d}\sin\omega t + \frac{\omega\mu\sigma^2 V_0\left(z^3 - 3zl^2\right)}{6d}\sin\omega t \tag{1.123b}$$

or

$$\bar{E}_x = \frac{\bar{V}_g}{d} + j\omega\frac{\mu\sigma}{2d}\left(z^2 - l^2\right)\bar{V}_g \tag{1.124a}$$

$$\bar{H}_y = -\frac{\sigma z}{d}\bar{V}_g - j\omega\frac{\varepsilon z}{d}\bar{V}_g - j\omega\frac{\mu\sigma^2\left(z^3 - 3zl^2\right)}{6d}\bar{V}_g \tag{1.124b}$$

Finally, the current drawn from the voltage source is given by

$$\begin{aligned}\bar{I}_g &= w\left[\bar{H}_y\right]_{z=-l} \\ &= \left(\frac{\sigma wl}{d} + j\omega\frac{\varepsilon wl}{d} - j\omega\frac{\mu\sigma^2 wl^3}{3d}\right)\bar{V}_g\end{aligned} \tag{1.125}$$

The input admittance of the structure is given by

$$\begin{aligned}\bar{Y}_{\text{in}} &= \frac{\bar{I}_g}{\bar{V}_g} = j\omega\frac{\varepsilon wl}{d} + \frac{\sigma wl}{d}\left(1 - j\omega\frac{\mu\sigma l^2}{3}\right) \\ &\approx j\omega\frac{\varepsilon wl}{d} + \frac{1}{(d/\sigma wl)[1 + j\omega(\mu\sigma l^2/3)]}\end{aligned} \tag{1.126}$$

where we have used the approximation $[1 + j\omega(\mu\sigma l^2/3)]^{-1} \approx [1 - j\omega(\mu\sigma l^2/3)]$. Proceeding further, we have

$$\begin{aligned}\bar{Y}_{\text{in}} &= j\omega\frac{\varepsilon wl}{d} + \frac{1}{(d/\sigma wl) + j\omega(\mu dl/3w)} \\ &= j\omega C + \frac{1}{R + j\omega L_i}\end{aligned} \tag{1.127}$$

where $C = \varepsilon wl/d$ is the capacitance of the structure if the material is a perfect dielectric, $R = d/\sigma wl$ is the resistance of the structure, and $L_i = \mu dl/3w$ is the internal inductance of the structure, all computed from static field analysis of the structure.

The equivalent circuit corresponding to Eq. (1.127) consists of capacitance C in parallel with the series combination of resistance R and internal inductance L_i, same as in Fig. 1.16. Thus, the low-frequency input behavior of the structure is essentially the same as that of the equivalent circuit of Fig. 1.16, with the understanding that its input admittance must also be approximated to first-order terms. Note that for $\sigma = 0$, the input admittance of the structure is purely capacitive. For nonzero σ, a critical value of σ equal to $\sqrt{3\varepsilon/\mu l^2}$ exists for which the input admittance is purely conductive. For values of σ smaller than the critical value, the input admittance is complex and capacitive, and for values of σ larger than the critical value, the input admittance is complex and inductive.

1.3.4. Waves and the Distributed Circuit Concept

In Sec. 1.3.3, we have seen that quasistatic field analysis of a physical structure provides information concerning the low-frequency input behavior of the structure. As the frequency is increased beyond that for which the quasistatic approximation is valid, terms in the infinite series solutions for the fields beyond the first-order terms need to be included. While one can obtain equivalent circuits for frequencies beyond the range of validity of the quasistatic approximation by evaluating the higher order terms, no further insight is gained through that process, and it is more straightforward to obtain the exact solution by resorting to simultaneous solution of Maxwell's equations when a closed form solution is possible.

Wave Equation and Solutions

Let us, for simplicity, consider the structures of Figs. 1.17 and 1.18, for which the material between the plates is a perfect dielectric ($\sigma = 0$). Then, regardless of the termination at $z = 0$, the equations to be solved are

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} = -\mu \frac{\partial \mathbf{H}}{\partial t} \tag{1.128a}$$

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} = \varepsilon \frac{\partial \mathbf{E}}{\partial t} \tag{1.128b}$$

For the geometry of the arrangements, $\mathbf{E} = E_x(z, t)\mathbf{a}_x$ and $\mathbf{H} = H_y(z, t)\mathbf{a}_y$, so that Eqs. (1.128a) and (1.128b) simplify to

$$\frac{\partial E_x}{\partial z} = -\mu \frac{\partial H_y}{\partial t} \tag{1.129a}$$

$$\frac{\partial H_y}{\partial z} = -\varepsilon \frac{\partial E_x}{\partial t} \tag{1.129b}$$

Combining the two equations by eliminating H_y, we obtain

$$\frac{\partial^2 E_x}{\partial z^2} = \mu\varepsilon \frac{\partial^2 E_x}{\partial t^2} \tag{1.130}$$

which is the *wave equation.* It has solutions of the form

$$E_x(z, t) = A \cos\omega\left(t - \sqrt{\mu\varepsilon}\, z + \phi^+\right) + B \cos\omega\left(t + \sqrt{\mu\varepsilon}\, z + \phi^-\right) \tag{1.131}$$

The terms on the right side correspond to traveling waves propagating in the $+z$ and $-z$ directions, which we shall call the (+) and (−) waves, respectively, with the velocity $1/\sqrt{\mu\varepsilon}$, or $c/\sqrt{\mu_r\varepsilon_r}$, where $c = 1/\sqrt{\mu_0\varepsilon_0}$ is the velocity of light in free space. This can be seen by setting the derivative of the argument of the cosine function in each term equal to zero or by plotting each term versus z for a few values of t, as shown in Fig. 1.20a and b for the (+) and (−) waves, respectively. The corresponding solution for H_y is given by

$$H_y(z, t) = \frac{1}{\sqrt{\mu/\varepsilon}}\left[A \cos\omega\left(t - \sqrt{\mu\varepsilon}\, z + \phi^+\right) - B \cos\omega\left(t + \sqrt{\mu\varepsilon}\, z + \phi^-\right)\right] \tag{1.132}$$

For sinusoidal waves, which is the case at present, the velocity of propagation is known as the *phase velocity*, denoted by v_p, since it is the velocity with which a constant phase surface moves in the direction of propagation. The quantity $\omega\sqrt{\mu\varepsilon}$ is the magnitude

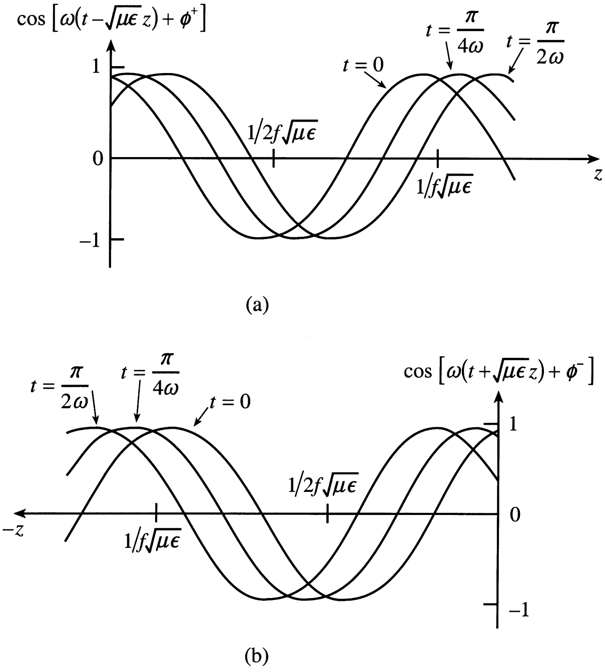


Figure 1.20 Plots of (a) $\cos\left[\omega\left(t - \sqrt{\mu\varepsilon}\, z\right) + \phi^+\right]$and (b) $\cos\left[\omega\left(t + \sqrt{\mu\varepsilon} z\right) + \phi^-\right]$, versus z, for a few values of t.

of the rate of change of phase at a fixed time t, for either wave. It is known as the *phase constant* and is denoted by the symbol β. The quantity $\sqrt{\mu/\varepsilon}$, which is the ratio of the electric field intensity to the magnetic field intensity for the (+) wave, and the negative of such ratio for the (−) wave, is known as the *intrinsic impedance* of the medium. It is denoted by the symbol η. Thus, the phasor electric and magnetic fields can be written as

$$\bar{E}_x = \bar{A}e^{-j\beta z} + \bar{B}e^{j\beta z} \tag{1.133}$$

$$\bar{H}_y = \frac{1}{\eta}\left(\bar{A}e^{-j\beta z} - \bar{B}e^{j\beta z}\right) \tag{1.134}$$

We may now use the boundary conditions for a given problem and obtain the specific solution for that problem. For the arrangement of Fig. 1.17, the boundary conditions are $\bar{H}_y = 0$ at $z = 0$ and $\bar{E}_x = \bar{V}_g/d$ at $z = -l$. We thus obtain the particular solution for that arrangement to be

$$\bar{E}_x = \frac{\bar{V}_g}{d\ \cos\beta l}\cos\beta z \tag{1.135}$$

$$\bar{H}_y = \frac{-j\bar{V}_g}{\eta d\ \cos\beta l}\sin\beta z \tag{1.136}$$

which correspond to complete standing waves, resulting from the superposition of (+) and (−) waves of equal amplitude. Complete standing waves are characterized by pure half-sinusoidal variations for the amplitudes of the fields, as shown in Fig. 1.21. For values of z at which the electric field amplitude is a maximum, the magnetic field amplitude is zero, and for values of z at which the electric field amplitude is zero, the magnetic field amplitude is a maximum. The fields are also out of phase in time, such that at any value of z, the magnetic field and the electric field differ in phase by $t = \pi/2\omega$.

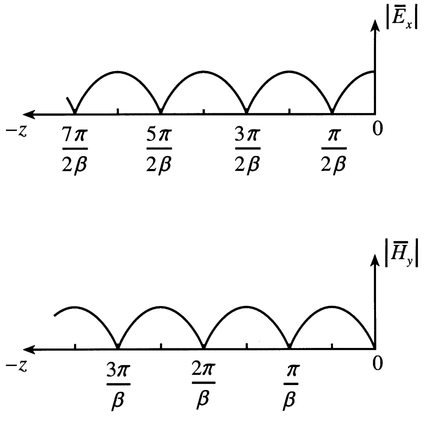


Figure 1.21 Standing wave patterns for the fields for the structure of Fig. 1.17.

Now, the current drawn from the voltage source is given by

$$\begin{aligned}\bar{I}_g &= w\left[\bar{H}_y\right]_{z=-l}\\ &= \frac{jw\bar{V}_g}{\eta d}\tan\beta l\end{aligned} \tag{1.137}$$

so that the input impedance of the structure is

$$\bar{Y}_{\text{in}} = \frac{\bar{I}_g}{\bar{V}_g} = j\frac{w}{\eta d}\tan\beta l \tag{1.138}$$

which can be expressed as a power series in βl. In particular, for $\beta l < \pi/2$,

$$\bar{Y}_{\text{in}} = j\frac{w}{\eta d}\left[\beta l + \frac{(\beta l)^3}{3} + \frac{2(\beta l)^5}{15} + \cdots\right] \tag{1.139}$$

The first term on the right side can be identified as belonging to the quasistatic approximation. Indeed for $\beta l \ll 1$, the higher order terms can be neglected, and

$$\begin{aligned}\bar{Y}_{\text{in}} &\approx \frac{jw}{\eta d}(\beta l)\\ &= j\omega\left(\frac{\varepsilon wl}{d}\right)\end{aligned} \tag{1.140}$$

same as that given by Eq. (1.111).

It can now be seen that the condition $\beta l \ll 1$ dictates the range of validity for the quasistatic approximation for the input behavior of the structure. In terms of the frequency f of the source, this condition means that $f \ll v_p/2\pi l$, or in terms of the period $T = 1/f$, it means that $T \gg 2\pi(l/v_p)$. Thus, as already mentioned, quasistatic fields are low-frequency approximations of time-varying fields that are complete solutions to Maxwell's equations, which represent wave propagation phenomena and can be approximated to the quasistatic character only when the times of interest are much greater than the propagation time, l/v_p, corresponding to the length of the structure. In terms of space variations of the fields at a fixed time, the wavelength $\lambda\,(=2\pi/\beta)$, which is the distance between two consecutive points along the direction of propagation between which the phase difference is 2π, must be such that $l \ll \lambda/2\pi$; thus, the physical length of the structure must be a small fraction of the wavelength. In terms of amplitudes of the fields, what this means is that over the length of the structure, the field amplitudes are fractional portions of the first one-quarter sinusoidal variations at the $z = 0$ end in Fig. 1.21, with the boundary conditions at the two ends of the structure always satisfied. Thus, because of the cos βz dependence of $\bar{E}_x$ on z, the electric field amplitude is essentially a constant, whereas because of the sin βz dependence of $\bar{H}_y$ on z, the magnetic field amplitude varies linearly with z. These are exactly the nature of the variations of the zero-order electric field and the first-order magnetic field, as discussed under electroquasistatic fields in Sec. 1.3.3.

For frequencies slightly beyond the range of validity of the quasistatic approximation, we can include the second term in the infinite series on the right side of Eq. (1.139) and deduce the equivalent circuit in the following manner.

$$\begin{aligned}\bar{Y}_{\text{in}} &\approx j\frac{w}{\eta d}\left[\beta l + \frac{(\beta l)^3}{3}\right]\\ &= j\omega\left(\frac{\varepsilon wl}{d}\right)\left[1+\left(\omega\frac{\varepsilon wl}{d}\right)\left(\omega\frac{\mu dl}{3w}\right)\right]\end{aligned} \tag{1.141}$$

or

$$\begin{aligned}\bar{Z}_{\text{in}} &= \frac{1}{j\omega(\varepsilon wl/d)[1+\omega(\varepsilon wl/d)(\omega\mu dl/3w)]}\\ &\approx \frac{1}{j\omega(\varepsilon wl/d)} + j\omega(\mu dl/3w)\end{aligned} \tag{1.142}$$

Thus the input behavior is equivalent to that of a capacitor of value same as that for the quasistatic approximation in series with an inductor of value 1/3 times the inductance found under the quasistatic approximation for the same arrangement but shorted at $z = 0$, by joining the two parallel plates. This series inductance is familiarly known as the *stray inductance*. But, all that has occurred is that the fractional portion of the sinusoidal variations of the field amplitudes over the length of the structure has increased, because the wavelength has decreased. As the frequency of the source is further increased, more and more terms in the infinite series need to be included, and the equivalent circuit becomes more and more involved. But throughout all this range of frequencies, the overall input behavior is still capacitive, until a frequency is reached when βl crosses the value $\pi/2$ and $\tan \beta l$ becomes negative, and the input behavior changes to inductive! In fact, a plot of $\tan\ \beta l$ versus f, shown in Fig. 1.22, indicates that as the frequency is varied, the input behavior alternates between capacitive and inductive, an observation unpredictable without the complete solutions to Maxwell's equations. At the frequencies at which the input behavior changes from capacitive to inductive, the input admittance becomes infinity (short-circuit condition). The field amplitude variations along the length of the structure are then exactly odd integer multiples of one-quarter sinusoids. At the frequencies at

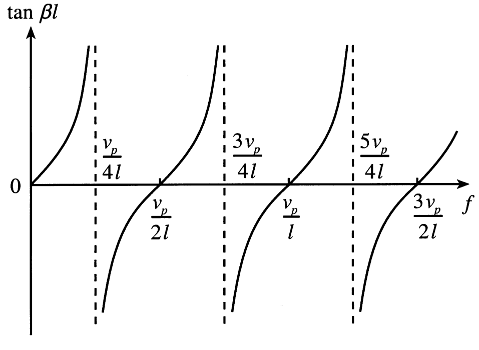


Figure 1.22 Frequency dependence of tan βl.

which the input behavior changes from inductive to capacitive, the input admittance becomes zero (open-circuit condition). The field amplitude variations along the length of the structure are then exactly even integer multiple of one-quarter sinusoids, or integer multiples of one-half sinusoids.

Turning now to the arrangement of Fig. 1.18, the boundary conditions are $\bar{E}_x = 0$ at $z = 0$ and $\bar{H}_y = \bar{I}_g/w$ at $z = -l$. We thus obtain the particular solution for that arrangement to be

$$\bar{E}_x = -\frac{j\eta\bar{I}_g}{w\ \cos\beta l}\ \sin\beta z \tag{1.143}$$

$$\bar{H}_y = \frac{\bar{I}_g}{w\ \cos\beta l}\ \cos\beta z \tag{1.144}$$

which, once again, correspond to complete standing waves, resulting from the superposition of (+) and (−) waves of equal amplitude, and characterized by pure half-sinusoidal variations for the amplitudes of the fields, as shown in Fig. 1.23, which are of the same nature as in Fig. 1.21, except that the electric and magnetic fields are interchanged.

Now, the voltage across the current source is given by

$$\begin{aligned}\bar{V}_g &= d\left[\bar{E}_x\right]_{z=-l}\\ &= \frac{j\eta d\bar{I}_g}{w}\ \tan\beta l\end{aligned} \tag{1.145}$$

so that the input impedance of the structure is

$$\bar{Z}_{\text{in}} = \frac{\bar{V}_g}{\bar{I}_g} = j\frac{\eta d}{w}\ \tan\beta l \tag{1.146}$$

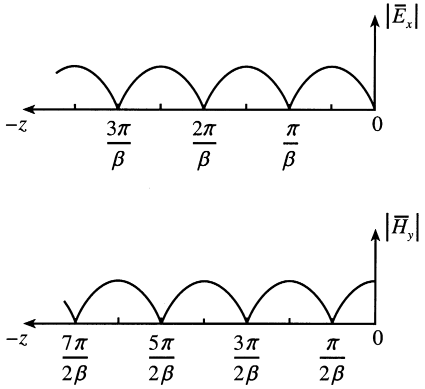


Figure 1.23 Standing wave patterns for the fields for the structure of Fig. 1.18.

which can be expressed as a power series in βl. In particular, for $\beta l < \pi/2$,

$$\bar{Z}_{in} = j\frac{\eta d}{w}\left[\beta l + \frac{(\beta l)^3}{3} + \frac{2(\beta l)^5}{15} + \cdots\right] \tag{1.147}$$

Once again, the first term on the right side can be identified as belonging to the quasistatic approximation. Indeed for $\beta l \ll 1$,

$$\begin{aligned}\bar{Z}_{in} &\approx \frac{\eta d}{w}(\beta l)\\ &= j\omega\left(\frac{\mu dl}{w}\right)\end{aligned} \tag{1.148}$$

same as that given by Eq. (1.118), and all the discussion pertinent to the condition for the validity of the quasistatic approximation for the structure of Fig. 1.17 applies also to the structure of Fig. 1.18, with the roles of the electric and magnetic fields interchanged. For $l \ll \lambda/2\pi$, the field amplitudes over the length of the structure are fractional portions of the first one-quarter sinusoidal variations at the $z = 0$ end in Fig. 1.23, with the boundary conditions at the two ends always satisfied. Thus, because of the $\cos\beta z$ dependence of $\bar{H}_y$ on z, the magnetic field amplitude is essentially a constant, whereas because of the $\sin \beta z$ dependence of $\bar{E}_x$ on z, the electric field amplitude varies linearly with z. These are exactly the nature of the variations of the zero-order magnetic field and the first-order electric field, as discussed under magnetoquasistatic fields in Sec. 1.3.3.

For frequencies slightly beyond the range of validity of the quasistatic approximation, we can include the second term in the infinite series on the right side of Eq. (1.147) and deduce the equivalent circuit in the following manner.

$$\begin{aligned}\bar{Z}_{\text{in}} &\approx \frac{j\eta d}{w}\left[\beta l + \frac{(\beta l)^3}{3}\right]\\ &= j\omega\left(\frac{\mu dl}{w}\right)\left[1 + \left(\omega\frac{\mu dl}{w}\right)\left(\omega\frac{\varepsilon wl}{3d}\right)\right]\end{aligned} \tag{1.149}$$

or

$$\begin{aligned}\bar{Y}_{\text{in}} &= \frac{1}{j\omega(\mu dl/w)[1 + (\omega\mu dl/w)(\omega\varepsilon wl/3d)]}\\ &\approx \frac{1}{j\omega(\mu dl/w)} + j\omega\left(\frac{\varepsilon wl}{3d}\right)\end{aligned} \tag{1.150}$$

Thus the input behavior is equivalent to that of an inductor of value same as that for the quasistatic approximation in parallel with a capacitor of value 1/3 times the capacitance found under the quasistatic approximation for the same arrangement but open at $z = 0$, without the two plates joined. This parallel capacitance is familiarly known as the *stray capacitance*. But again, all that has occurred is that the fractional portion of the sinusoidal variations of the field amplitudes over the length of the structure has increased, because the wavelength has decreased. As the frequency of the source is further increased, more and more terms in the infinite series need to be included and the equivalent circuit becomes more and more involved. But throughout all this range of frequencies, the overall input behavior is still inductive, until a frequency is reached when βl crosses the value $\pi/2$ and $\tan \beta l$ becomes negative and the input behavior changes to capacitive. In fact, the plot of $\tan \beta l$ versus f, shown in Fig. 1.22, indicates that as the frequency is varied, the input behavior alternates between inductive and capacitive, an observation unpredictable without the complete solutions to Maxwell's equations. At the frequencies at which the input behavior changes from inductive to capacitive, the input impedance becomes infinity (open-circuit condition). The field amplitude variations along the length of the structure are then exactly odd integer multiples of one-quarter sinusoids. At the frequencies at which the input behavior changes from capacitive to inductive, the input impedance becomes zero (short-circuit condition). The field amplitude variations along the length of the structure are then exactly even integer multiples of one-quarter sinusoids, or integer multiples of one-half sinusoids.

Distributed Circuit Concept

We have seen that, from the circuit point of view, the structure of Fig. 1.13 behaves like a capacitor for the static case and the capacitive character is essentially retained for its input behavior for sinusoidally time-varying excitation at frequencies low enough to be within the range of validity of the quasistatic approximation. Likewise, we have seen that from a circuit point of view, the structure of Fig. 1.14 behaves like an inductor for the static case and the inductive character is essentially retained for its input behavior for sinusoidally time-varying excitation at frequencies low enough to be within the range of validity of the quasistatic approximation. For both structures, at an arbitrarily high enough frequency, the input behavior can be obtained only by obtaining complete (wave) solutions to Maxwell's equations, subject to the appropriate boundary conditions. The question to ask then is whether there is a circuit equivalent for the structure itself, independent of the termination, that is representative of the phenomenon taking place along the structure and valid at any arbitrary frequency, to the extent that the material parameters themselves are independent of frequency? The answer is, yes, under certain conditions, giving rise to the concept of the *distributed circuit*.

To develop and discuss the concept of the distributed circuit using a more general case than that allowed by the arrangements of Figs. 1.13 and 1.14, let us consider the case of the structure of Fig. 1.15 driven by a sinusoidally time-varying source, as in Fig. 1.19a. Then the equations to be solved are

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} = -\mu \frac{\partial \mathbf{H}}{\partial t} \tag{1.151a}$$

$$\nabla \times \mathbf{H} = \mathbf{J}_c + \frac{\partial \mathbf{D}}{\partial t} = \sigma \mathbf{E} + \varepsilon \frac{\partial \mathbf{E}}{\partial t} \tag{1.151b}$$

For the geometry of the arrangement, $\mathbf{E} = E_x(z, t)\mathbf{a}_x$ and $\mathbf{H} = H_y(z, t)\mathbf{a}_y$, so that Eqs. (1.151a) and (1.151b) simplify to

$$\frac{\partial E_x}{\partial z} = -\mu \frac{\partial H_y}{\partial t} \tag{1.152a}$$

$$\frac{\partial H_y}{\partial z} = -\sigma E_x - \varepsilon \frac{\partial E_x}{\partial t} \tag{1.152b}$$

Now, since E_z and H_z are zero, we can, in a given $z =$ constant plane, uniquely define a voltage between the plates in terms of the electric field intensity in that plane and a current crossing that plane in one direction on the top plate and in the opposite direction on the bottom plate in terms of the magnetic field intensity in that plane. These are given by

$$V(z, t) = dE_x(z, t) \tag{1.153a}$$

$$I(z, t) = wH_y(z, t) \tag{1.153b}$$

Substituting Eqs. (1.153a) and (1.153b) in Eqs. (1.152a) and (1.152b), and rearranging, we obtain

$$\frac{\partial V(z, t)}{\partial z} = -\left[\frac{\mu d}{w}\right]\frac{\partial I(z, t)}{\partial t} \tag{1.154a}$$

$$\frac{\partial I(z, t)}{\partial z} = -\left[\frac{\sigma w}{d}\right]V(z, t) - \left[\frac{\varepsilon w}{d}\right]\frac{\partial V(z, t)}{\partial t} \tag{1.154b}$$

Writing the derivates with respect to z on the left sides of the equations in terms of limits as $\Delta z \to 0$, and multiplying by Δz on both sides of the equations provides the equivalent circuit for a section of length Δz of the structure, as shown in Fig. 1.24, in which the quantities $\mathcal{L}$, $\mathcal{C}$, and $\mathcal{G}$, given by

$$\mathcal{L} = \frac{\mu d}{w} \tag{1.155a}$$

$$\mathcal{C} = \frac{\varepsilon w}{d} \tag{1.155b}$$

$$\mathcal{G} = \frac{\sigma w}{d} \tag{1.155c}$$

are the inductance per unit length, capacitance per unit length, and conductance per unit length, respectively, of the structure, all computed from static field analysis, except that now they are expressed in terms of "per unit length" and not for the entire structure in a "lump." It then follows that the circuit representation of the entire structure consists of an infinite number of such sections in cascade, as shown in Fig. 1.25. Such a circuit is known as a *distributed circuit*. The distributed circuit notion arises from the fact that the

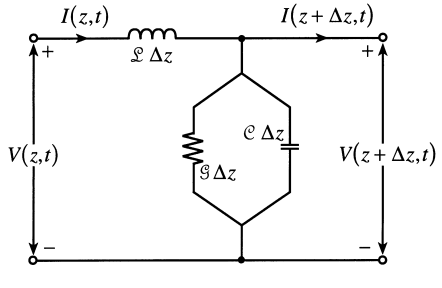


Figure 1.24 Circuit equivalent for Eqs. (1.159a and b), in the limit $\Delta z \to 0$.

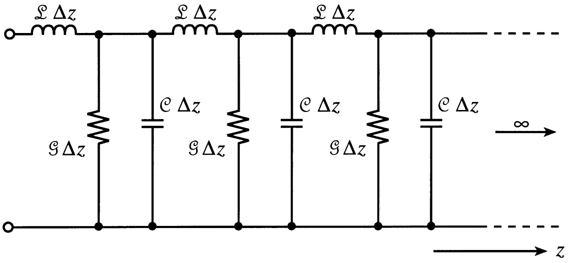


Figure 1.25 Distributed circuit representation of the structure of Fig. 1.19a.

inductance, capacitance, and conductance are distributed uniformly and overlappingly along the structure. A physical interpretation of the distributed-circuit concept follows from energy considerations, based on the properties that inductance, capacitance, and conductance are elements associated with energy storage in the magnetic field, energy storage in the electric field, and power dissipation due to conduction current flow, in the material. Since these phenomena occur continuously and overlappingly along the structure, the inductance, capacitance, and conductance must be distributed uniformly and overlappingly along the structure.

A physical structure for which the distributed circuit concept is applicable is familiarly known as a *transmission line*. The parallel-plate arrangement of Figs. 1.13–1.15 is a special case of a transmission line, known as the *parallel-plate* line, in which the waves are called *uniform plane waves*, since the fields are uniform in the $z=$ constant planes. In general, a transmission line consists of two parallel conductors having arbitrary cross sections and the waves are transverse electromagnetic, or TEM, waves, for which the fields are nonuniform in the $z=$ constant planes but satisfying the property of both electric and magnetic fields having no components along the direction of propagation, that is, parallel to the conductors. For waves for which the electric field has a component along the direction of propagation but the magnetic field does not, as is the case for transverse magnetic or TM waves, the current on the conductors crossing a given transverse plane cannot be expressed uniquely in terms of the magnetic field components in that plane. Likewise, for waves for which the magnetic field has a component along the direction of propagation but the electric field does not, as is the case for transverse electric or TE waves, the voltage between the conductors in a given transverse plane cannot be expressed uniquely in terms of the electric field components in that plane. Structures which support TM and TE waves are generally known as *waveguides*, although transmission lines are also waveguides in the sense that TEM waves are guided parallel to the conductors of the line.

All transmission lines having perfect conductors are governed by the equations

$$\frac{\partial V(z,t)}{\partial z} = -\mathcal{L}\,\frac{\partial I(z,t)}{\partial t} \tag{1.156a}$$

$$\frac{\partial I(z,t)}{\partial z} = -\mathcal{G}V(z,t) - \mathcal{C}\frac{\partial V(z,t)}{\partial t} \tag{1.156b}$$

which are known as the *transmission-line equations*. The values of $\mathcal{L}$, $\mathcal{C}$, and $\mathcal{G}$ differ from one line to another, and depend on the cross-sectional geometry of the conductors. For the parallel-plate line, $\mathcal{L}$, $\mathcal{C}$, and $\mathcal{G}$ are given by Eqs. (1.155a), (1.155b), and (1.155c), respectively. Note that

$$\mathcal{L}\mathcal{C} = \mu\varepsilon \tag{1.157a}$$

$$\frac{\mathcal{G}}{\mathcal{C}} = \frac{\sigma}{\varepsilon} \tag{1.157b}$$

a set of relations, which is applicable to any line governed by Eqs. (1.156a) and (1.156b). Thus for a given set of material parameters, only one of the three parameters, $\mathcal{L}$, $\mathcal{C}$, and $\mathcal{G}$, is independent.

In practice, the conductors are imperfect, adding a resistance per unit length and additional inductance per unit length in the series branches of the distributed circuit. Although the waves are then no longer exactly TEM waves, the distributed circuit is commonly used for transmission lines with imperfect conductors. Another consideration that arises in practice is that the material parameters and hence the line parameters can be functions of frequency.

1.3.5. Hertzian Dipole Fields via the Thread of Statics–Quasistatics–Waves

In the preceding three sections, we have seen the development of solutions to Maxwell's equations, beginning with static fields and spanning the frequency domain from quasistatic approximations at low frequencies to waves for beyond quasistatics. In this section, we shall develop the solution for the electromagnetic field due to a Hertzian dipole by making use of the thread of statics–quasistatics–waves, as compared to the commonly used approach based on the magnetic vector potential, for a culminating experience of revisiting the fundamentals of engineering electromagnetics.

The Hertzian dipole is an elemental antenna consisting of an infinitesimally long piece of wire carrying an alternating current $I(t)$, as shown in Fig. 1.26. To maintain the current flow in the wire, we postulate two point charges $Q_1(t)$ and $Q_2(t)$ terminating the wire at its two ends, so that the law of conservation of charge is satisfied. Thus, if

$$I(t) = I_0 \cos \omega t \tag{1.158}$$

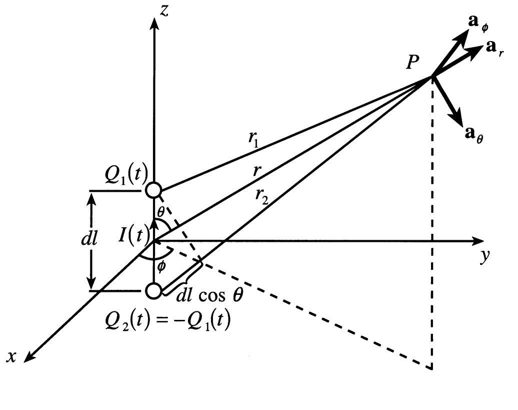


Figure 1.26 For the determination of the electromagnetic field due to the Hertzian dipole.

then

$$Q_1(t) = \frac{I_0}{\omega} \sin \omega t \tag{1.159a}$$

$$Q_2(t) = -\frac{I_0}{\omega} \sin \omega t = -Q_1(t) \tag{1.159b}$$

For $d/dt = 0$, the charges are static and the current is zero. The field is simply the electrostatic field due to the electric dipole made up of $Q_1 = -Q_2 = Q_0$. Applying Eq. (1.70) to the geometry in Fig. 1.26, we write the electrostatic potential at the point P due to the dipole located at the origin to be

$$\Phi = \frac{Q_0}{4\pi\varepsilon}\left(\frac{1}{r_1} - \frac{1}{r_2}\right) \tag{1.160}$$

In the limit $dl \to 0$, keeping the dipole moment $Q_0(dl)$ fixed, we get

$$\Phi = \frac{Q_0(dl)\ \cos\theta}{4\pi\varepsilon r^2} \tag{1.161}$$

so that the electrostatic field at the point P due to the dipole is given by

$$\mathbf{E} = -\nabla\Phi = \frac{Q_0(dl)}{4\pi\varepsilon r^3}(2\ \cos\theta\ \mathbf{a}_r + \sin\theta\ \mathbf{a}_\theta) \tag{1.162}$$

With time variations in the manner $Q_1(t) = -Q_2(t) = Q_0 \sin \omega t$, so that $I_0 = \omega Q_0$, and at low frequencies, the situation changes to electroquasistatic with the electric field of amplitude proportional to the zeroth power in ω given by

$$\mathbf{E}_0 = \frac{Q_0(dl)\ \sin\omega t}{4\pi\varepsilon r^3}(2\ \cos\theta\ \mathbf{a}_r + \sin\theta\ \mathbf{a}_\theta) \tag{1.163}$$

The corresponding magnetic field of amplitude proportional to the first power in ω is given by the solution of

$$\nabla \times \mathbf{H}_1 = \frac{\partial \mathbf{D}_0}{\partial t} = \varepsilon \frac{\partial \mathbf{E}_0}{\partial t} \tag{1.164}$$

For the geometry associated with the arrangement, this reduces to

$$\begin{vmatrix} \dfrac{\mathbf{a}_r}{r^2\ \sin\theta} & \dfrac{\mathbf{a}_\theta}{r\ \sin\theta} & \dfrac{\mathbf{a}_\phi}{r} \\ \dfrac{\partial}{\partial r} & \dfrac{\partial}{\partial\theta} & 0 \\ 0 & 0 & r\ \sin\theta\ H_{\phi 1} \end{vmatrix} = \varepsilon \frac{\partial \mathbf{E}_0}{\partial t} \tag{1.165}$$

so that

$$\mathbf{H}_1 = \frac{\omega Q_0(dl)\ \cos\omega t}{4\pi r^2}\ \sin\theta\,\mathbf{a}_\phi \tag{1.166}$$

To extend the solutions for the fields for frequencies beyond the range of validity of the quasistatic approximation, we recognize that the situation then corresponds to wave propagation. With the dipole at the origin, the waves propagate radially away from it so that the time functions sin ωt and cos ωt in Eqs. (1.163) and (1.166) need to be replaced by sin $(\omega t - \beta r)$ and cos $(\omega t - \beta r)$, respectively, where $\beta = \omega\sqrt{\mu\varepsilon}$ is the phase constant. Therefore, let us on this basis alone and without any other considerations, write the field expressions as

$$\mathbf{E} = \frac{I_0(dl)\ \sin(\omega t - \beta r)}{4\pi\varepsilon\omega r^3}(2\ \cos\theta\,\mathbf{a}_r + \sin\theta\,\mathbf{a}_\theta) \tag{1.167}$$

$$\mathbf{H} = \frac{I_0(dl)\ \cos(\omega t - \beta r)}{4\pi r^2}\ \sin\theta\,\mathbf{a}_\phi \tag{1.168}$$

where we have also replaced Q_0 by I_0/ω, and pose the question as to whether or not these expressions represent the solution for the electromagnetic field due to the Hertzian dipole. The answer is "no," since they do not satisfy Maxwell's curl equations

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} = -\mu\frac{\partial \mathbf{H}}{\partial t} \tag{1.169a}$$

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} = \varepsilon\frac{\partial \mathbf{E}}{\partial t} \tag{1.169b}$$

which can be verified by substituting them into the equations.

There is more than one way of resolving this discrepancy, but we shall here do it from physical considerations. Even a cursory look at the solutions for the fields given by Eqs. (1.167) and (1.168) points to the problem, since the Poynting vector $\mathbf{E} \times \mathbf{H}$ corresponding to them is proportional to $1/r^5$, and there is no real power flow associated with them because they are out of phase in ωt by $\pi/2$. But, we should expect that the fields contain terms proportional to $1/r$, which are in phase, from considerations of real power flow in the radial direction and from the behavior of the waves viewed locally over plane areas normal to the radial lines emanating from the Hertzian dipole, and electrically far from it ($\beta r \gg 1$), to be approximately that of uniform plane waves with the planes as their constant phase surfaces, as shown in Fig. 1.27.

To elaborate upon this, let us consider two spherical surfaces of radii r_a and r_b and centered at the dipole and insert a cone through these two surfaces such that its vertex is at the antenna, as shown in the Fig. 1.27. Then the power crossing any portion of the spherical surface of radius r_b must be the same as the power crossing the spherical surface of radius r_a inside the cone. Since these surface areas are proportional to the square of the radius and since the surface integral of the Poynting vector gives the power, the Poynting vector must have an r component proportional to $1/r^2$, and it follows that the solutions for E_θ and H_ϕ must contain terms proportional to $1/r$ and in phase.

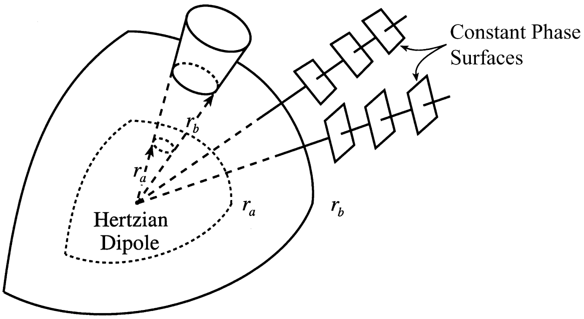


Figure 1.27 Radiation of electromagnetic waves far from the Hertzian dipole.

Thus let us modify the expression for **H** given by Eq. (1.168) by adding a second term containing $1/r$ in the manner

$$\mathbf{H} = \frac{I_0(dl)\,\sin\theta}{4\pi}\left[\frac{\cos(\omega t - \beta r)}{r^2} + \frac{A\,\cos(\omega t - \beta r + \delta)}{r}\right]\mathbf{a}_\phi \tag{1.170}$$

where A and δ are constants to be determined. Then, from Maxwell's curl equation for **H**, given by Eq. (1.169b), we obtain

$$\begin{aligned}\mathbf{E} = {} & \frac{2I_0(dl)\,\cos\theta}{4\pi\varepsilon\omega}\left[\frac{\sin(\omega t - \beta r)}{r^3} + \frac{A\,\sin(\omega t - \beta r + \delta)}{r^2}\right]\mathbf{a}_r \\ & + \frac{I_0(dl)\,\sin\,\theta}{4\pi\varepsilon\omega}\left[\frac{\sin(\omega t - \beta r)}{r^3} + \frac{\beta\,\sin(\omega t - \beta r)}{r^2}\right. \\ & \left. + \frac{A\beta\,\cos(\omega t - \beta r + \delta)}{r}\right]\mathbf{a}_\theta \end{aligned} \tag{1.171}$$

Now, substituting this in Maxwell's curl equation for **E** given by Eq. (1.169a), we get

$$\begin{aligned}\mathbf{H} = {} & \frac{I_0(dl)\,\sin\theta}{4\pi}\left[\frac{2\,\sin(\omega t - \beta r)}{\beta r^3} + \frac{2A\,\cos(\omega t - \beta r + \delta)}{\beta^2 r^3}\right. \\ & \left. + \frac{\cos(\omega t - \beta r)}{r^2} + \frac{A\,\cos(\omega t - \beta r + \delta)}{r}\right]\mathbf{a}_\phi \end{aligned} \tag{1.172}$$

But Eq. (1.172) must be the same as Eq. (1.170). Therefore, we set

$$\frac{2\,\sin(\omega t - \beta r)}{\beta r^3} + \frac{2A\,\cos(\omega t - \beta r + \delta)}{\beta^2 r^3} = 0 \tag{1.173}$$

which gives us

$$\delta = \frac{\pi}{2} \tag{1.174}$$

$$A = \beta \tag{1.175}$$

Substituting Eqs. (1.174) and (1.175) in Eqs. (1.171) and (1.172), we then have the complete electromagnetic field due to the Hertzian dipole given by

$$\begin{aligned}\mathbf{E} = {} & \frac{2I_0(dl)\,\cos\theta}{4\pi\varepsilon\omega}\left[\frac{\sin(\omega t - \beta r)}{r^3} + \frac{\beta\,\cos(\omega t - \beta r)}{r^2}\right]\mathbf{a}_r \\ & + \frac{I_0(dl)\,\sin\theta}{4\pi\varepsilon\omega}\left[\frac{\sin(\omega t - \beta r)}{r^3} + \frac{\beta\,\cos(\omega t - \beta r)}{r^2}\right. \\ & \left. - \frac{\beta^2\,\sin(\omega t - \beta r)}{r}\right]\mathbf{a}_\theta\end{aligned} \tag{1.176}$$

$$\mathbf{H} = \frac{I_0(dl)\,\sin\theta}{4\pi}\left[\frac{\cos(\omega t - \beta r)}{r^2} - \frac{\beta\,\sin(\omega t - \beta r)}{r}\right]\mathbf{a}_\phi \tag{1.177}$$

Expressed in phasor form and with some rearrangement, the field components are given by

$$\bar{E}_r = \frac{2\beta^2\eta I_0(dl)\,\cos\theta}{4\pi}\left[-j\frac{1}{(\beta r)^3} + \frac{1}{(\beta r)^2}\right]e^{-j\beta r} \tag{1.178}$$

$$\bar{E}_\theta = \frac{\beta^2\eta I_0(dl)\,\sin\theta}{4\pi}\left[-j\frac{1}{(\beta r)^3} + \frac{1}{(\beta r)^2} + j\frac{1}{\beta r}\right]e^{-j\beta r} \tag{1.179}$$

$$\bar{H}_\phi = \frac{\beta^2 I_0(dl)\,\sin\theta}{4\pi}\left[\frac{1}{(\beta r)^2} + j\frac{1}{\beta r}\right]e^{-j\beta r} \tag{1.180}$$

The following observations are pertinent to these field expressions:

1. They satisfy all Maxwell's equations exactly.
2. For any value of r, the time-average value of the θ component of the Poynting vector is zero, and the time-average value of the r component of the Poynting vector is completely from the $1/r$ terms, thereby resulting in the time-average power crossing all possible spherical surfaces centered at the dipole to be the same.
3. At low frequencies such that $\beta r \ll 1$, the $1/(\beta r)^3$ terms dominate the $1/(\beta r)^2$ terms, which in turn dominate the $1/(\beta r)$ terms, and $e^{-j\beta r} \approx (1 - j\beta r)$, thereby reducing the field expressions to the phasor forms of the quasistatic approximations given by Eqs. (1.163) and (1.166).

Finally, they are the familiar expressions obtained by using the magnetic vector potential approach.

REFERENCES

There is a multitude of textbooks on engineering electromagnetics, let alone electromagnetics, and it is difficult to prepare a list without inadvertently omitting some of them. Therefore, I have not attempted to include a bibliography of these books; instead, I refer the reader to his or her favorite book(s), while a student or later during the individual's career, and I list below my own books, which are referenced on the first page of this chapter.

1. Rao, N. N. *Basic Electromagnetics with Applications*; Prentice Hall: Englewood Cliffs, NJ, 1972.
2. Rao, N. N. *Elements of Engineering Electromagnetics*; Prentice Hall: Englewood Cliffs, NJ, 1977.
3. Rao, N. N. *Elements of Engineering Electromagnetics*; 2nd Ed.; Prentice Hall: Englewood Cliffs, NJ, 1987.
4. Rao, N. N. *Elements of Engineering Electromagnetics*; 3rd Ed.; Prentice Hall: Englewood Cliffs, NJ, 1991.
5. Rao, N. N. *Elements of Engineering Electromagnetics*; 4th Ed.; Prentice Hall; Englewood Cliffs, NJ, 1994.
6. Rao, N. N. *Elements of Engineering Electromagnetics*; 5th Ed.; Prentice Hall; Upper Saddle River, NJ, 2000.
7. Rao, N. N. *Elements of Engineering Electromagnetics*; 6th Ed.; Pearson Prentice Hall; Upper Saddle River, NJ, 2004.

2
Applied Electrostatics

Mark N. Horenstein
Boston University
Boston, Massachusetts, U.S.A.

2.1. INTRODUCTION

The term *electrostatics* brings visions of Benjamin Franklin, the "kite and key" experiment, Leyden jars, cat fur, and glass rods. These and similar experiments heralded the discovery of electromagnetism and were among some of the first recorded in the industrial age. The forces attributable to electrostatic charge have been known since the time of the ancient Greeks, yet the discipline continues to be the focus of much research and development. Most electrostatic processes fall into one of two categories. Sometimes, electrostatic charge produces a desired outcome, such as motion, adhesion, or energy dissipation. Electrostatic forces enable such diverse processes as laser printing, electrophotography, electrostatic paint spraying, powder coating, environmentally friendly pesticide application, drug delivery, food production, and electrostatic precipitation. Electrostatics is critical to the operation of micro-electromechanical systems (MEMS), including numerous microsensors, transducers, accelerometers, and the microfluidic "lab on a chip". These microdevices have opened up new vistas of discovery and have changed the way electronic circuits interface with the mechanical world. Electrostatic forces on a molecular scale lie at the core of nanodevices, and the inner workings of a cell's nucleus are also governed by electrostatics. A myriad of self-assembling nanodevices involving coulombic attraction and repulsion comprise yet another technology in which electrostatics plays an important role.

Despite its many useful applications, electrostatic charge is often a nuisance to be avoided. For example, sparks of electrostatic origin trigger countless accidental explosions every year and lead to loss of life and property. Less dramatically, static sparks can damage manufactured products such as electronic circuits, photographic film, and thin-coated materials. The transient voltage and current of a single spark event, called an *electrostatic discharge* (ESD), can render a semiconductor chip useless. Indeed, a billion-dollar industry specializing in the prevention or neutralization of ESD-producing electrostatic charge has of necessity evolved within the semiconductor industry to help mitigate this problem.

Unwanted electrostatic charge can also affect the production of textiles or plastics. Sheets of these materials, called *webs*, are produced on rollers at high speed. Electrostatic charge can cause webs to cling to rollers and jam production lines. Similarly, the sparks that result from accumulated charge can damage the product itself, either by exposing light-sensitive surfaces or by puncturing the body of the web.

This chapter presents the fundamentals that one needs in order to understand electrostatics as both friend and foe. We first define the electrostatic regime in the broad context of Maxwell's equations and review several fundamental concepts, including Coulomb's law, force-energy relations, triboelectrification, induction charging, particle electrification, and dielectric breakdown. We then examine several applications of electrostatics in science and industry and discuss some of the methods used to moderate the effects of unwanted charge.

2.2. THE ELECTROQUASISTATIC REGIME

Like all of electromagnetics, electrostatics is governed by Maxwell's equations, the elegant mathematical statements that form the basis for all that is covered in this book. True electrostatic systems are those in which all time derivatives in Maxwell's equations are exactly zero and in which forces of magnetic origin are absent. This limiting definition excludes numerous practical electrostatic-based applications. Fortunately, it can be relaxed while still capturing the salient features of the electrostatic domain. The *electroquasistatic* regime thus refers to those cases of Maxwell's equations in which fields and charge magnitudes may vary with time but in which the forces due to the electric field always dominate over the forces due to the magnetic field. At any given moment in time, an electroquasistatic field is identical to the field that would be produced were the relevant charges fixed at their instantaneous values and locations.

In order for a system to be electroquasistatic, two conditions must be true: First, any currents that flow within the system must be so small that the magnetic fields they produce generate negligible forces compared to coulombic forces. Second, any time variations in the electric field (or the charges that produce them) must occur so slowly that the effects of any induced magnetic fields are negligible. In this limit, the curl of E approaches zero, and the cross-coupling between E and H that would otherwise give rise to propagating waves is negligible. Thus one manifestation of the electroquasistatic regime is that the sources of the electric field produce no propagating waves.

The conditions for satisfying the electroquasistatic limit also can be quantified via dimensional analysis. The curl operator $\nabla\times$ has the dimensions of a reciprocal distance ΔL, while each time derivative dt in Maxwell's equations has the dimensions of a time Δt. Thus, considering Faraday's law:

$$\nabla \times \mathbf{E} = \frac{-\partial \mu \mathbf{H}}{\partial t} \tag{2.1}$$

the condition that the left-hand side be much greater than the right-hand side becomes dimensionally equivalent to

$$\frac{E}{\Delta L} \gg \frac{\mu H}{\Delta t} \tag{2.2}$$

This same dimensional argument can be applied to Ampere's law:

$$\nabla \times \mathbf{H} = \frac{\partial \varepsilon \mathbf{E}}{\partial t} + \mathbf{J} \tag{2.3}$$

which, with $\mathbf{J} = 0$, leads to

$$\frac{H}{\Delta L} \gg \frac{\varepsilon E}{\Delta t} \tag{2.4}$$

Equation (2.4) for **H** can be substituted into Eq. (2.2), yielding

$$\frac{\mathbf{E}}{\Delta L} \gg \frac{\mu}{\Delta t} \frac{\varepsilon \mathbf{E} \Delta L}{\Delta t} \tag{2.5}$$

This last equation results in the dimensional condition that

$$\Delta L \ll \frac{\Delta t}{\sqrt{\mu \varepsilon}} \tag{2.6}$$

The quantity $1/\sqrt{\mu\varepsilon}$ is the propagation velocity of electromagnetic waves in the medium (i.e., the speed of light), hence $\Delta t/\sqrt{\mu\varepsilon}$ is the distance that a wave would travel after propagating for time Δt. If we interpret Δt as the period T of a possible propagating wave, then according to Eq. (2.6), the quasistatic limit applies if the length scale ΔL of the system is much smaller than the propagation wavelength at the frequency of excitation.

In the true electrostatic limit, the time derivatives are exactly zero, and Faraday's law Eq. (2.1) becomes

$$\nabla \times \mathbf{E} = 0 \tag{2.7}$$

This equation, together with Gauss' law

$$\nabla \cdot \varepsilon \mathbf{E} = \rho \tag{2.8}$$

form the foundations of the electrostatic regime. These two equations can also be expressed in integral form as:

$$\oint \mathbf{E} \cdot \mathbf{dl} = 0 \tag{2.9}$$

and

$$\int \varepsilon \mathbf{E} \cdot dA = \int \rho dV \tag{2.10}$$

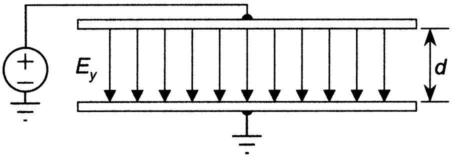


Figure 2.1 A simple system consisting of two parallel electrodes of area A separated by a distance d.

The curl-free electric field Eq. (2.7) can be expressed as the gradient of a scalar potential Φ:

$$\mathbf{E} = -\nabla\Phi \tag{2.11}$$

which can be integrated with respect to path length to yield the definition of the voltage difference between two points a and b:

$$V_{ab} = -\int_b^a \mathbf{E} \cdot \mathbf{dl} \tag{2.12}$$

Equation (2.12) applies in any geometry, but it becomes particularly simple for parallel-electrode geometry. For example, the two-electrode system of Fig. 2.1, with separation distance d, will produce a uniform electric field of magnitude

$$E_y = \frac{V}{d} \tag{2.13}$$

when energized to a voltage V. Applying Gauss' law to the inner surface of the either electrode yields a relationship between the surface charge ρ_s and E_y,

$$\varepsilon E_y = \rho_s \tag{2.14}$$

Here ρ_s has the units of coulombs per square meter, and ε is the dielectric permittivity of the medium between the electrodes. In other, more complex geometries, the solutions to Eqs. (2.9) and (2.10) take on different forms, as discussed in the next section.

2.3. DISCRETE AND DISTRIBUTED CAPACITANCE

When two conductors are connected to a voltage source, one will acquire positive charge and the other an equal magnitude of negative charge. The charge per unit voltage is called the *capacitance* of the electrode system and can be described by the relationship

$$C = \frac{Q}{V} \tag{2.15}$$

Here $\pm Q$ are the magnitudes of the positive and negative charges, and V is the voltage applied to the conductors. It is easily shown that the capacitance between two parallel plane electrodes of area A and separation d is given approximately by

$$C = \frac{\varepsilon A}{d} \tag{2.16}$$

where ε is the permittivity of the material between the electrodes, and the approximation results because field enhancements, or "fringing effects," at the edges of the electrodes have been ignored. Although Eq. (2.16) is limited to planar electrodes, it illustrates the following basic form of the formula for capacitance in any geometry:

$$\text{Capacitance} = \frac{\text{permittivity} \times \text{area parameter}}{\text{length parameter}} \tag{2.17}$$

Table 2.1 provides a summary of the field, potential, and capacitance equations for energized electrodes in several different geometries.

2.4. DIELECTRIC PERMITTIVITY

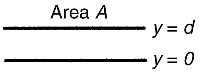
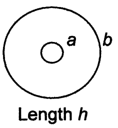
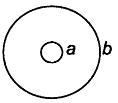
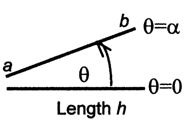
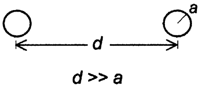
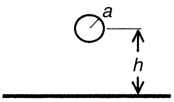
The dielectric permittivity of a material describes its tendency to become internally polarized when subjected to an electric field. Permittivity in farads per meter can also can be expressed in fundamental units of coulombs per volt-meter (C/V·m). The *dielectric constant*, or *relative permittivity*, of a substance is defined as its permittivity normalized to ε_0, where $\varepsilon_0 = 8.85 \times 10^{-12}$ F/m is the permittivity of free space. For reference purposes, relative permittivity values for several common materials are provided in Table 2.2. Note that no material has a permittivity smaller than ε_0.

2.5. THE ORIGINS OF ELECTROSTATIC CHARGE

The source of electrostatic charge lies at the atomic level, where a nucleus having a fixed number of positive protons is surrounded by a cloud of orbiting electrons. The number of protons in the nucleus gives the atom its unique identity as an element. An individual atom is fundamentally charge neutral, but not all electrons are tightly bound to the nucleus. Some electrons, particularly those in outer orbitals, are easily removed from individual atoms. In conductors such as copper, aluminum, or gold, the outer electrons are weakly bound to the atom and are free to roam about the crystalline matrix that makes up the material. These free electrons can readily contribute to the flow of electricity. In insulators such as plastics, wood, glass, and ceramics, the outer electrons remain bound to individual atoms, and virtually none are free to contribute to the flow of electricity.

Electrostatic phenomena become important when an imbalance exists between positive and negative charges in some region of interest. Sometimes such an imbalance occurs due to the phenomenon of *contact electrification* [1–8]. When dissimilar materials come into contact and are then separated, one material tends to retain more electrons and become negatively charged, while the other gives up electrons and become positively charged. This contact electrification phenomenon, called *triboelectrification*, occurs at the

Table 2.1 Field, Potential, and Capacitance Expressions for Various Electrode Geometries

Geometry	**E** field	Potential	Capacitance
Planar  	$E_y = \frac{V}{d}$	$\Phi = V\frac{y}{d}$	$C = \frac{\varepsilon A}{d}$
Cylindrical  	$E_r = \frac{V}{r\ln(b/a)}$	$\Phi = \frac{V}{\ln(b/a)}\ln\left(\frac{b}{r}\right)$	$C = \frac{2\pi\varepsilon h}{\ln(b/a)}$
Spherical  	$E_r = \frac{V}{r^2[1/a - 1/b]}$	$\Phi = V\frac{a\,(b-r)}{r\,(b-a)}$	$C = 4\pi\varepsilon\frac{ba}{b-a}$
Wedge  	$E_\theta = \frac{V}{\alpha r}$	$\Phi = V\frac{\theta}{\alpha}$	$C = \frac{\varepsilon h}{\alpha}\ln\left(\frac{b}{a}\right)$
Parallel lines (at ±V)  		$\Phi \approx \frac{2\pi\varepsilon V}{\ln(d/a)}\ln\left(\frac{r_1}{r_2}\right)$ r_1; r_2 = distances to lines	$C \approx \frac{\pi\varepsilon h}{\ln(d/a)}$
Wire to plane  			$C \approx \frac{2\pi\varepsilon}{\cosh^{-1}[(h+a)/a]}$ $h \gg a$

points of intimate material contact. The amount of charge transferred to any given contact point is related to the work function of the materials. The process is enhanced by friction which increases the net contact surface area. Charge separation occurs on both conductors and insulators, but in the former case it becomes significant only when at least one of the conductors is electrically isolated and able to retain the separated charge. This situation is commonly encountered, for example, in the handling of conducting powders. If neither conductor is isolated, an electrical pathway will exist between them, and the separated charges will flow together and neutralize one another. In the case of insulators, however, the separated charges cannot easily flow, and the surfaces of the separated objects remain charged. The widespread use of insulators such as plastics and ceramics in industry and manufacturing ensures that triboelectrification will occur in numerous situations. The pneumatic transport of insulating particles such as plastic pellets, petrochemicals, fertilizers, and grains are particularly susceptible to tribocharging.

Table 2.2 Relative Permittivities of Various Materials

Air	1	Polycarbonate	~3.0
Alumina	8.8	Polyethylene	2.3
Barium titanate ($BaTiO_3$)	1200	Polyamide	~3.4–4.5
Borosilicate glass	4	Polystyrene	2.6
Carbon tetrachloride	2.2	Polyvinyl chloride	6.1
Epoxy	~3.4–3.7	Porcelain	~5–8
Ethanol	24	Quartz	3.8
Fused quartz (SiO_2)	3.9	Rubber	~2–4
Gallium arsenide	13.1	Selenium	6
Glass	~4–9	Silicon	11.9
Kevlar	~3.5–4.5	Silicon nitride	7.2
Methanol	33	Silicone	~3.2–4.7
Mylar	3.2	Sodium chloride	5.9
Neoprene	~4–6.7	Styrofoam	1.03
Nylon	~3.5–4.5	Teflon	2.1
Paper	~1.5–3	Water	~80
Paraffin	2.1	Wood (dry)	1.4–2.9
Plexiglas	2.8		

Table 2.3 The Triboelectric Series

POSITIVE
Quartz
Silicone
Glass
Wool
Polymethyl methacrylate (Plexiglas)
Salt (NaCl)
Fur
Silk
Aluminum
Cellulose acetate
Cotton
Steel
Wood
Hard rubber
Copper
Zinc
Gold
Polyester
Polystyrene
Natural rubber
Polyurethane
Polystyrene
Polyethylene
Polypropylene
Polyvinyl chloride
Silicon
Teflon
NEGATIVE

Source: Compiled from several sources [9–13].

The relative propensity of materials to become charged following contact and separation has traditionally been summarized by the *triboelectric series* of Table 2.3. (*Tribo* is a Greek prefix meaning *frictional*.) After a contact-and-separation event, the material that is listed higher in the series will tend to become positively charged, while the one that is lower in the series will tend to become negatively charged. The vagueness of the phrase "will tend to" in the previous sentence is intentional. Despite the seemingly reliable order implied by the triboelectric series, the polarities of tribocharged materials often cannot be predicted reliably, particularly if the materials lie near each other in the series. This imprecision is evident in the various sources [9–13] cited in Table 2.3 that differ on the exact order of the series. Contact charging is an imprecise science that is driven by effects occurring on an atomic scale. The slightest trace of surface impurities or altered surface states can cause a material to deviate from the predictions implied by the triboelectric series. Two contact events that seem similar on the macroscopic level can yield entirely different results if they are dissimilar on the microscopic level. Thus contact and separation of *like* materials can sometimes lead to charging if the contacting surfaces are microscopically dissimilar. The triboelectric series of Table 2.3 should be viewed as a probabilistic prediction of polarity during multiple charge separation events. Only when two materials are located at extremes of the series can their polarities be predicted reliably following a contact-charging event.

2.6. WHEN IS "STATIC" CHARGE TRULY STATIC?

The term *static electricity* invokes an image of charge that cannot flow because it is held stationary by one or more insulators. The ability of charge to be static in fact does depend on the presence of an insulator to hold it in place. What materials can really be considered insulators, however, depends on one's point of view. Those who work with electrostatics know that the arrival of a cold, dry winter is synonymous with the onset of "static season," because electrostatic-related problems are exacerbated by a lack of humidity. When cold air enters a building and is warmed, its relative humidity declines noticeably. The tendency of hydroscopic surfaces to absorb moisture, thereby increasing their surface conductivities, is sharply curtailed, and the decay of triboelectric charges to ground over surface-conducting pathways is slowed dramatically. Regardless of humidity level, however, these conducting pathways always exist to some degree, even under the driest of conditions. Additionally, surface contaminants such as dust, oils, or residues can add to surface conduction, so that eventually all electrostatic charge finds its way back to ground. Thus, in most situations of practical relevance, no true insulator exists. In electrostatics, the definition of an insulator really depends on how long one is willing to wait. Stated succinctly, if one waits long enough, everything will look like a perfect conductor sooner or later. An important parameter associated with "static electricity" is its relaxation time constant—the time it takes for separated charges to recombine by flowing over conducting pathways. This relaxation time, be it measured in seconds, hours, or days, must always be compared to time intervals of interest in any given situation.

2.7. INDUCTION CHARGING

As discussed in the previous section, contact electrification can result in the separation of charge between two dissimilar materials. Another form of charge separation occurs when a voltage is applied between two conductors, for example the electrodes of a capacitor. Capacitive structures obey the relationship

$$Q = \pm CV \tag{2.18}$$

where the positive and negative charges appear on the surfaces of the opposing electrodes. The electrode which is at the higher potential will carry $+Q$; the electrode at the lower potential will carry $-Q$. The mode of charge separation inherent to capacitive structures is known as *inductive* charging. As Eq. (2.18) suggests, the magnitude of the inductively separated charge can be controlled by altering either C or V. This feature of induction charging lies in contrast to triboelectrification, where the degree of charge separation often depends more on chance than on mechanisms that can be controlled.

If a conductor charged by induction is subsequently disconnected from its source of voltage, the now electrically floating conductor will retain its acquired charge regardless of its position relative to other conductors. This mode of induction charging is used often in industry to charge atomized droplets of conducting liquids. The sequence of diagrams shown in Fig. 2.2 illustrates the process. The dispensed liquid becomes part the capacitive electrode as it emerges from the hollow tube and is charged by induction. As the droplet breaks off, it retains its charge, thereafter becoming a free, charged droplet. A droplet of a given size can be charged only to the maximum *Raleigh limit* [9,14,15]:

$$Q_{\max} = 8\pi\sqrt{\varepsilon_0 \gamma} R_p^{3/2} \tag{2.19}$$

Here γ is the liquid's surface tension and R_p the droplet radius. The Raleigh limit signifies the value at which self repulsion of the charge overcomes the surface tension holding the droplet together, causing the droplet to break up.

2.8. DIELECTRIC BREAKDOWN

Nature is fundamentally charge neutral, but when charges are separated by any mechanism, the maximum quantity of charge is limited by the phenomenon of *dielectric breakdown*. Dielectric breakdown occurs in solids, liquids and gases and is characterized by the maximum field magnitude that can be sustained before a field-stressed material loses its insulating properties.* When a solid is stressed by an electric field, imperfections

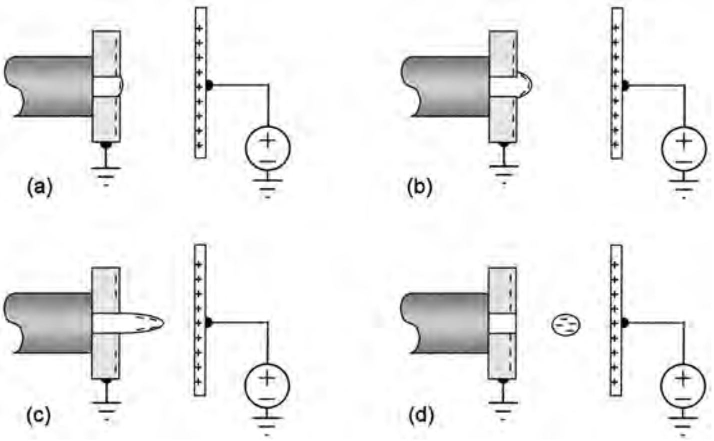


Figure 2.2 Charging a conducting liquid droplet by induction. As the droplet breaks off (d), it retains the charge induced on it by the opposing electrode.

*Breakdown in vacuum invariably occurs over the surfaces of insulating structures used to support opposing electrodes.

or stray impurities can initiate a local discharge, which degrades the composition of the material. The process eventually extends completely through the material, leading to irreversible breakdown and the formation of a conducting bridge through which current can flow, often with dramatic results. In air and other gases, ever-present stray electrons (produced randomly, for example, by ionizing cosmic rays) will accelerate in an electric field, sometimes gaining sufficient energy between collisions to ionize neutral molecules, thereby liberating more electrons. If the field is of sufficient magnitude, the sequence of ensuing collisions can grow exponentially in a self-sustaining *avalanche* process. Once enough electrons have been liberated from their molecules, the gas becomes locally conducting, resulting in a spark discharge. This phenomenon is familiar to anyone who has walked across a carpet on a dry day and then touched a doorknob or light switch. The human body, having become electrified with excess charge, induces a strong electric field on the metal object as it is approached, ultimately resulting in the transfer of charge via a rapid, energetic spark. The most dramatic manifestation of this type of discharge is the phenomenon of atmospheric lightning.

A good rule of thumb is that air at standard temperature and pressure will break down at a field magnitude of about 30 kV/cm (i.e., 3 MV/m or 3×10^6 V/m). This number increases substantially for small air gaps of 50 μm or less because the gap distance approaches the mean free path for collisions, and fewer ionizing events take place. Hence a larger field is required to cause enough ionization to initiate an avalanche breakdown. This phenomena, known as the *Paschen effect*, results in a breakdown-field versus gap-distance curve such as the one shown in Fig. 2.3 [9,10,12,18,19]. The Paschen effect is critical to the operation of micro-electromechanical systems, or MEMS, because fields in excess of 30 kV/cm are required to produce the forces needed to move structural elements made from silicon or other materials.

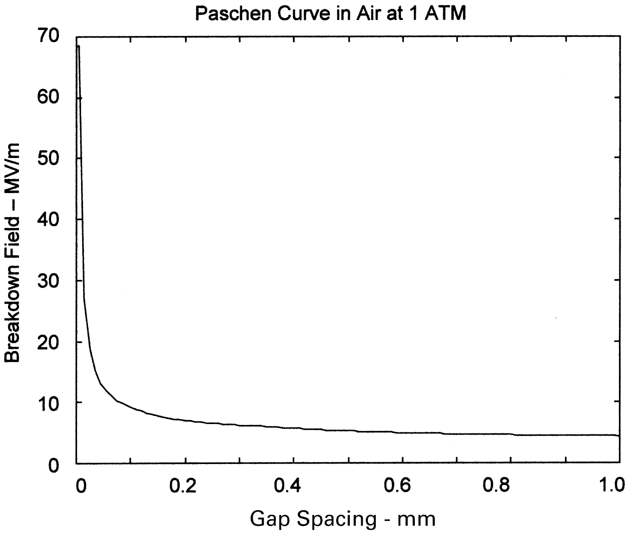


Figure 2.3 Paschen breakdown field vs. gap spacing for air at 1 atmosphere. For large gap spacings, the curve is asymptotic to 3×10^6 V/m.

2.9. CORONA DISCHARGE

One of the more common methods for intentionally producing electrostatic charge involves the phenomenon of *corona discharge*. Corona is a partial breakdown that occurs when two electrodes, one sharp and the other much less so, are energized by a voltage source. In such a configuration, the electric field around the sharp electrode is greatly enhanced. At some critical level of voltage, called the *onset voltage*, the field near the sharp electrode exceeds the dielectric breakdown strength of the gas, typically air. This localized breakdown produces free electrons and positive ions via the avalanche process. In the remainder of the electrode space, however, the field is substantially weaker, and no ionization takes place. Thus the breakdown that occurs near the stressed electrode provides a source of ions, but no spark discharge occurs. If the stressed electrode is positive, the positive ions will be repelled from it, providing an abundant source of positive ions. If the stressed electrode is negative, the free electrons will be repelled from it but will quickly attach to neutral molecules upon leaving the high field region, thereby forming negative ions. The phenomenon of corona is illustrated graphically in Fig. 2.4 for a positive source electrode.

For either ion polarity, and in most electrode configurations, the relationship between applied voltage and the resulting corona current follows an equation of the form $i_C = gV(V - V_C)$, where V_C is the critical onset voltage of the electrode system and g is a constant. The values of g and V_C will depend on many factors, including electrode geometry, spacing, radii of curvature, and surface roughness, as well as on ion mobility, air temperature, and air pressure. One must generally determine g and V_C empirically, but in coaxial geometry this relationship can be solved analytically [18]. The result is a complex formula, but for small currents, the equation for cylindrical geometry can be approximated by

$$i_L = \frac{4\pi\varepsilon_0 \kappa V(V - V_C)}{b^2 \ln(b/a)} \tag{2.20}$$

Here i_L is the current per unit axial length, b and a are the outer and inner coaxial radii, respectively, and κ is the ion mobility (about $2.2 \times 10^{-4}\,\mathrm{m^2/V{\cdot}s}$ for air at standard temperature and pressure). As the applied voltage V is increased, corona will first occur at the corona onset voltage V_C. For coaxial electrodes with an air dielectric, V_C is equal to

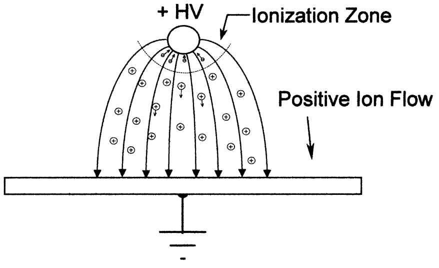


Figure 2.4 Basic mechanism of corona discharge near a highly stressed electrode. Positive corona is shown; a similar situation exists for negative corona.

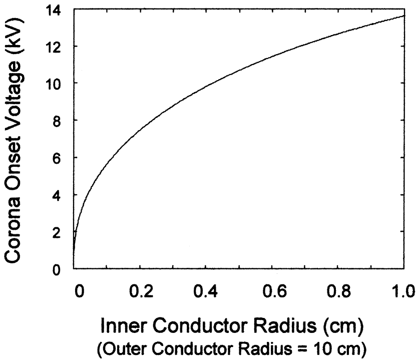


Figure 2.5 Plot of corona onset voltage V_c vs. inner conductor radius a for coaxial electrodes with 10-cm outer conductor radius.

the voltage at which the electric field on the surface of the inner electrode first reaches the value given by Peek's equation [18,20]:

$$E_{\text{peek}} = mE_{\text{bk}}\left(1 + \frac{0.0308}{\sqrt{a}}\right) \tag{2.21}$$

Here $E_{\text{bk}} = 3 \times 10^6\,\text{V/m}$ is the breakdown strength of air under uniform field conditions, a is the inner conductor radius in meters, m is an empirical surface roughness factor, and standard temperature and pressure are assumed. Note that E_{peek} will always be larger than the breakdown field E_{bk}. Peek's equation describes the field that must be established at the inner conductor surface before local breakdown (corona) can occur. The equation is also approximately valid for parallel-wire lines. For smooth conductors $m = 1$, and for rough surfaces $m = 0.8$.

In a coaxial system, the electric field magnitude at the inner radius a is given by

$$E(r) = \frac{V}{a\ln(b/a)} \tag{2.22}$$

hence the corona onset voltage becomes

$$V_C = E_{\text{peek}}\, a\ln\left(\frac{b}{a}\right) = E_{\text{bk}}\left(1 + \frac{0.0308}{\sqrt{a}}\right)a\ln\left(\frac{b}{a}\right) \tag{2.23}$$

A plot of V_C versus a for the case $b = 10\,\text{cm}$ is shown in Fig. 2.5.

2.10. CHARGES AND FORCE

The electrostatic force f_{12} between two charges q_1 and q_2 separated by a distance r is governed by Coulomb's law, a fundamental principle of physics:

$$f_{12} = \frac{q_1 q_2}{4\pi\varepsilon r^2} \tag{2.24}$$

The direction of this force is parallel to a line between the charges. All other force relationships in electrostatics derive from Coulomb's law. If a collection of charges produces a net electric field **E**, it is easily shown by integration that the collective force exerted on a solitary charge q by all the other charges becomes just $q\mathbf{E}$. This simple relationship comprises the electric field term in the Lorentz force law of electromagnetics:

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \tag{2.25}$$

In many practical situations in electrostatics, one is interested in the forces on conductors and insulators upon which charges reside. Numerous mathematical methods exist for predicting such forces, including the *force-energy* method, the *boundary element* method, and the *Maxwell stress tensor* [21–24]. Of these three methods, the force-energy method is the one most easily understood from basic principles and the most practical to use in many situations. The analysis that follows represents an abridged derivation using the force-energy method.

We first consider a constant-charge system in which two objects carrying fixed charges experience a net force F_Q (as yet unknown). One such hypothetical system is illustrated in Fig. 2.6. If one of the objects is displaced against F_Q by an incremental distance dx relative to the other object, then the mechanical work dW_m performed on the displaced object will be $F_Q\,dx$. Because the objects and their fixed charges are electrically isolated, the work transferred to the displaced body must increase the energy stored in the system. The stored electrostatic energy W_e thus will be augmented by dW_m, from which it follows that

$$F_Q = \frac{dW_m}{dx} \tag{2.26}$$

As an example of this principle, consider the parallel-electrode structure of Fig. 2.7, for which the capacitance is given by

$$C = \frac{\varepsilon A}{x} \tag{2.27}$$

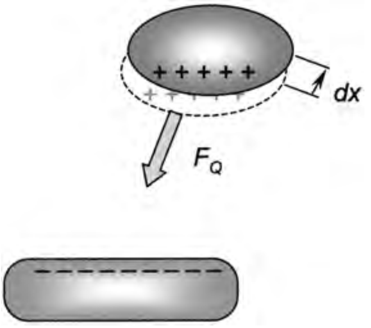


Figure 2.6 One charged object is displaced relative to another. The increment of work added to the system is equal the electrostatic force F_Q times the displacement dx.

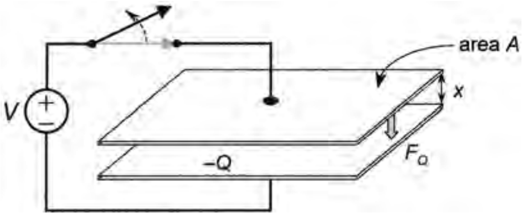


Figure 2.7 Parallel electrodes are energized by a voltage source that is subsequently disconnected. Fixed charges ±Q remain on the electrodes.

If the electrodes are precharged, then disconnected from their source of voltage, the charge will thereafter remain constant. The stored electrical energy can then be expressed as [24]

$$W_e = \frac{Q^2}{2C} \tag{2.28}$$

The force between the electrodes can be found by taking the x derivative of this equation:

$$F_Q = \frac{dW_e}{dx} = \frac{Q^2}{2}\frac{d}{dx}\left(\frac{x}{\varepsilon A}\right) = \frac{Q^2}{2\varepsilon A} \tag{2.29}$$

Equation (2.29) also describes the force between two insulating surfaces of area A that carry uniform surface charge densities $\rho_s = \pm Q/A$.

It is readily shown [21–24] that applying the energy method to two conductors left connected to the energizing voltage V yields a similar force equation:

$$F_V = \frac{dW_e}{dx} \tag{2.30}$$

Here W_e is the stored electric energy expressed as $1/2CV^2$. When this formula is applied to a system in which voltage, not charge, is constrained, the force it predicts will always be attractive.

Equation (2.30) is readily applied to the parallel-electrode structure of Fig. 2.7 with the switch closed. The force between the conductors becomes

$$\frac{dW_e}{dx} = \frac{V^2}{2}\frac{d}{dx}\left(\frac{\varepsilon A}{x}\right) = -\frac{\varepsilon A V^2}{2x^2} \tag{2.31}$$

This force is inversely proportional to the square of the separation distance x.

2.11. PARTICLE CHARGING IN AIR

Many electrostatic processes use the coulomb force to influence the transport of charged airborne particles. Examples include electrostatic paint spraying [10,16], electrostatic

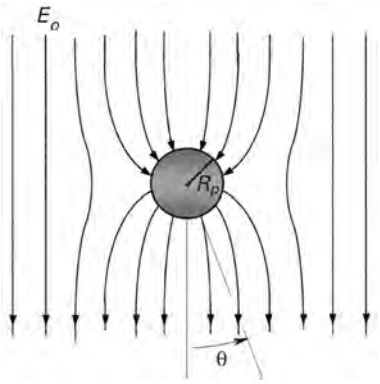


Figure 2.8 Conducting sphere distorts an otherwise uniform electric field. The field components are given by Eq. (2.32). If the source of the field produces ions, the latter will follow the field lines to the particle surface.

powder coating, electrostatic crop spraying [25,26], electrostatic drug delivery, and electrostatic precipitation. These processes are described later in this chapter. Airborne particles are sometimes charged by induction, requiring that initial contact be made with a conducting electrode. In other processes, particles are charged by ions in the presence of an electric field.

In this section, we examine the latter process in more detail. To a first approximation, many airborne particles can be treated as conducting spheres—an assumption that greatly simplifies the equations governing particle charging. The approximation requires that the particle have a shape free from prominent asymmetries and also that the intrinsic charging time of the particle, given by the ratio ε/σ of the particle's permittivity to conductivity, be much shorter than other time scales of interest. Suppose that an uncharged particle of radius R_p is situated in a uniform, downward-pointing electric field $\boldsymbol{E}_o$, as depicted in Fig. 2.8. A "uniform field" in this case is one that does not change spatially over the scale of at least several particle radii. Further suppose that a uniform, homogeneous source of unipolar ions is produced by the system and carried toward the particle by the electric field. These ions might be produced, for example, by some form of corona discharge. If we assume the ion density to be small enough such that space-charge perturbation of the field is negligible, the electric field components in the neighborhood of the particle become:

$$E_r = E_o\left(1 + \frac{2R_p^3}{r^3}\right)\cos\theta + \frac{Q}{4\pi\varepsilon_o r^2}$$

and

$$E_\theta = E_o\left(\frac{R_p^3}{r^3} - 1\right)\sin\theta \tag{2.32}$$

with $E_\varphi = 0$. Here Q represents any charge that the conducting particle may carry. If Q is positive, the second term in the equation for E_r adds a uniform radial component that points outward.

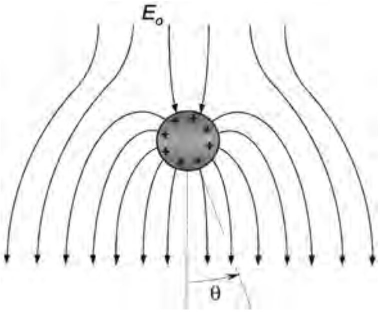


Figure 2.9 As the particle collects charge Q, the field lines are increasingly excluded from the particle surface.

Figure 2.8 shows the field pattern for the case $Q = 0$. Note that **E** is everywhere perpendicular to the particle surface, where $E_\theta = 0$. Ions will be transported to the surface of the particle by the field, thereby increasing the magnitude of Q. If the ions are positive, only field lines leading *into* the particle will contribute to its charging. Field lines that originate from the surface of the particle cannot carry ions, because no source of ions exists there. As charge accumulates on the particle and the second term for E_r in Eq. (2.32) becomes larger, the field pattern for $Q \neq 0$ takes the form shown in Fig. 2.9. The reduction in magnitude of the inward-pointing field lines restricts the flow of ions to the particle surface. When $Q/4\pi\varepsilon_0 r^2$ in Eq. (2.32) becomes equal to the factor $E_o(1 + 2R_p^3/r^3)$ at $r = R_p$, all field lines will originate from the particle itself, so that further ion charging of the particle will cease. Under this condition, E_r at $\theta = 180°$ and $r = R_p$ becomes zero. The charge limit Q_{sat} can thus be found by setting E_r in Eq. (2.32) to zero, yielding

$$\frac{Q_{\text{sat}}}{4\pi\varepsilon_0 R_p^2} = 3E_o \tag{2.33}$$

or

$$Q_{\text{sat}} = 12\pi\varepsilon_0 R_p^2 E_o \tag{2.34}$$

The value given by Eq. (2.34) is called the *saturation charge* of the particle, or sometimes the Pauthenier limit [27]. It represents the maximum charge that the particle can hold. For a 100-μm particle situated in a 100-kV/m field, for example, the saturation charge calculated from Eq. (2.34) becomes 0.33 pC.

Note that Q_{sat} increases with particle radius and the ambient field E_o, but it is not dependent on ion mobility or ion density. These latter quantities affect only the *rate* of particle charging [15,24].

For $Q < Q_{\text{sat}}$, it can be shown via surface integration of the field equation, Eq. (2.32), that the ion current to the particle is given by

$$i_Q = \frac{dQ}{dt} = 3\pi R_p^2\, E_o N q_{\text{ion}}\, \kappa \left(1 - \frac{Q}{Q_{\text{sat}}}\right)^2 \tag{2.35}$$

where N is the ambient ion density, q_{ion} the ion charge, and κ the ion mobility. Solving this differential equation results in an expression for Q as a function of time:

$$Q(t) = Q_{\text{sat}} \frac{t/\tau}{1 + t/\tau} \tag{2.36}$$

This hyperbolic charging equation is governed by the time constant $\tau = 4\varepsilon_0 / N q_{\text{ion}}\,\kappa$. For the typical values $N = 10^{15}$ ions/m^3 and $\kappa = 2 \times 10^{-4}$ m^2/V·s for singly charged ions in air, particle charging will be governed by the hyperbolic charging time constant $\tau = 1.1$ ms. Note that this latter value is independent of particle radius and electric field magnitude.

2.12. CHARGED PARTICLE MOTION

A charged, airborne particle will experience two principal forces: electrostatic and aerodynamic. The former will be given by

$$\mathbf{F}_{\text{elec}} = Q\mathbf{E} \tag{2.37}$$

where Q is the particle charge, while the latter will be given by the Stokes' drag equation [9,15]:

$$\mathbf{F}_{\text{drag}} = -6\pi\eta R_p(\mathbf{U}_{\text{p}} - \mathbf{U}_{\text{air}}) \tag{2.38}$$

Here $\mathbf{U}_{\text{p}}$ is the particle velocity, $\mathbf{U}_{\text{air}}$ the ambient air velocity (if any), and η the kinematic viscosity of air. At standard temperature and pressure, $\eta = 1.8 \times 10^{-5}$ N·s/m^2 [9]. Equation (2.38) is valid for particles in the approximate size range 0.5 to 25 μm, for which inertia can usually be ignored. For smaller particles, Brownian motion becomes the dominant mechanical force, whereas for particles larger than about 25 μm, the Reynolds number for typical values of $\mathbf{U}_{\text{p}}$ approaches unity and the Stokes' drag limit no longer applies.

The balance between $\mathbf{F}_{\text{elec}}$ and $\mathbf{F}_{\text{drag}}$ determines the net particle velocity:

$$\mathbf{U}_{\text{p}} = \mathbf{U}_{\text{air}} + \frac{Q}{6\pi\eta R_p}\text{E} \tag{2.39}$$

The quantity $Q/6\pi\eta R_p$, called the particle *mobility*, describes the added particle velocity per unit electric field. The mobility has the units of m^2/V · s.

2.13. ELECTROSTATIC COATING

Electrostatic methods are widely used in industry to produce coatings of excellent quality. Electrostatic-assisted spraying techniques can be used for water or petroleum-based paints as well as curable powder coatings, surface lacquers, and numerous chemical substrates. In electrostatic paint spraying, microscopic droplets charged by induction are driven directly to the surface of the work piece by an applied electric field. In power coating methods, dry particles of heat-cured epoxies or other polymers are first charged, then forced to the surface of the work piece by electrostatic forces. Similar spraying techniques are used to coat crops with minimal pesticides [17, 25, 26]. In one system, electrostatic methods are even used to spray tanning solution or decontamination chemicals on the human body. Electrostatic methods substantially reduce the volume of wasted coating material, because particles or droplets are forced directly to the coated surface, and only small amounts miss the target to become wasted product.

2.14. ELECTROSTATIC PAINT SPRAYING

The basic form of an electrostatic paint spray system is illustrated in Fig 2.10. Paint is atomized from a pressurized nozzle that is also held at a high electric potential relative to ground. Voltages in the range 50 kV to 100 kV are typical for this application. As paint is extruded at the nozzle outlet, it becomes part of the electrode system, and charge is induced on the surface of the liquid jet. This charge will have the same polarity as the energized nozzle. As each droplet breaks off and becomes atomized, it carries with it its induced charge and can thereafter be driven by the electric field to the work piece. Very uniform charge-to-mass ratios Q/m can be produced in this way, leading to a more uniform coating compared to nonelectrostatic atomization methods. The technique works best if the targeted object is a good conductor (e.g., the fender of an automobile), because the electric field emanating from the nozzle must terminate primarily on the work piece surface if an efficient coating process is to be realized.

Once a droplet breaks away from the nozzle, its trajectory will be determined by a balance between electrostatic forces and viscous drag, as summarized by Eq. (2.39). If the ambient air velocity $\mathbf{U}_{\text{air}}$ is zero, this equation becomes

$$\mathbf{U}_{\text{p}} = \frac{Q}{6\pi\eta R_p}\mathbf{E} \tag{2.40}$$

Of most interest is the droplet velocity when it impacts the work surface. Determining its value requires knowledge of the electric field at the work surface, but in complex

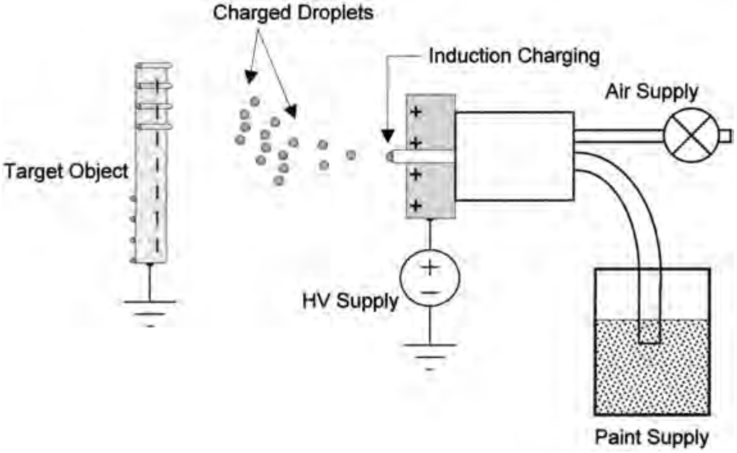


Figure 2.10 Basic electrostatic spray system. Droplets are charged by induction as they exit the atomization nozzle.

geometries, analytical solutions are seldom possible. Estimation or empirical measurement using a field mill (see Sec. 2.18) is usually required. Velocities in the range 0.1 to 100 m/s are common in electrostatic painting operations [9–16].

Note that the particle radius in Eq. (2.40) can be expressed in terms of the droplet mass, given by

$$M = \frac{4\pi R_p^3 \gamma}{3} \tag{2.41}$$

where is the mass density of the liquid. The velocity equation, Eq. (2.40), can thus be written as

$$\mathbf{U}_\mathrm{p} = \frac{2R_p^2\gamma}{9\eta}\frac{Q}{M}\mathbf{E} \tag{2.42}$$

This form of the equation illustrates the significance of the *charge-to-mass ratio* of the droplet. For a given electric field magnitude, the droplet velocity will be proportional to Q/M. Because Q has a maximum value determined by either the Raleigh limit of Eq. (2.19) or the saturation charge limit of Eq. (2.34), Eq. (2.42) will be limited as well. For a 100-μm droplet of unity density charged to its saturation limit in a 100 kV/m field, the impact velocity becomes about 0.1 m/s.

2.15. ELECTROPHOTOGRAPHY

The "simple" copy machine has become common in everyday life, but in reality, this machine is far from simple. The copier provides a good example of how electrostatics can be used to transfer particles between surfaces. The transfer process, first invented by Chester Carlson around 1939 [10], is also known as *electrophotography*, or sometimes *xerography*. Although the inner workings of a copy machine are complex [28], its basic features can be understood from the simplified diagram of Fig. 2.11. A thin photosensitive layer is deposited over a grounded surface, usually in the form of a rotating drum. The photosensitive material has the property that it remains an insulator in the dark but becomes partially conducting when exposed to light.

In the first step, the photoconductor is charged by ions from a corona source. This device, sometimes called a *corotron* [21], is scanned just over the surface of the photoconductor, allowing ions to migrate and stick to the photoconductor surface. These deposited charges are strongly attracted to their image charges in the underlying ground layer, but because the dark photoconductor is an insulator, the charges cannot move toward each other, but instead remain fixed in place.

Next, light projected from the image to be reproduced is focused on the photoconductor surface. The regions of the image corresponding to black remain insulating, while the white areas are exposed to light and become conducting. The charge deposited over these latter regions flows through the photoconductor to the ground plane, thereby discharging the photoconductor. The remaining electrostatic pattern on the drum is called a *latent image*.

The photoconductor is next exposed to toner particles that have been charged, usually by triboelectrification, to a polarity opposite that of the latent image. Some field

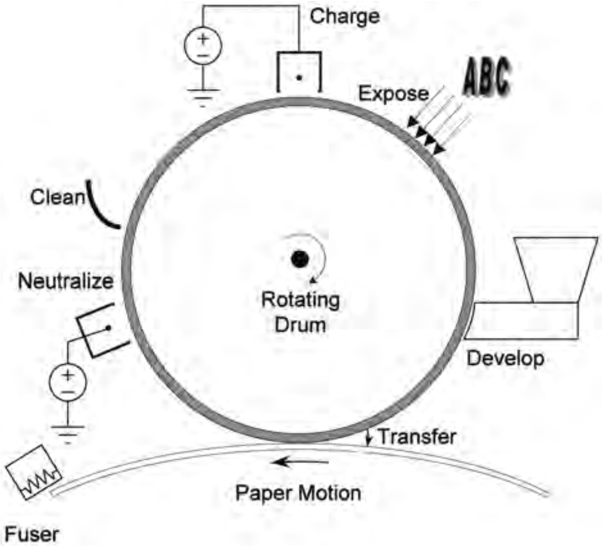


Figure 2.11 Basic elements of an electrostatic photocopier. As the light-sensitive drum rotates, it is charged, exposed to the image, dusted with toner, brought into contact with the paper, then discharged, and cleaned. The imprinted page passes through a fuser which melts the toner into the paper.

lines from the latent image extend above the surface and are of sufficient magnitude to capture and hold the charged toner particles. The latent image is thus transformed into a real image in the form of deposited toner particles.

In the next step of the process, image on the toner-coated drum is transferred to paper. The paper, backed by its own ground plane, is brought in proximity to the photoconductor surface. If the parameters are correctly chosen, the toner particles will be preferentially attracted to the paper and will jump from the photoconductor to the paper surface. The paper is then run through a high-temperature fuser which melts the toner particles into the paper.

2.16. ELECTROSTATIC PRECIPITATION

Electrostatic precipitation is used to remove airborne pollutants in the form of smoke, dust, fumes, atomized droplets, and other airborne particles from streams of moving gas [29–34]. Electrostatic precipitators provide a low cost method for removing particles of diameter 10 μm or smaller. They are often found in electric power plants, which must meet stringent air quality standards. Other applications include the cleaning of gas streams from boilers, smelting plants, blast furnaces, cement factories, and the air handling systems of large buildings. Electrostatic precipitators are also found on a smaller scale in room air cleaners, smoke abatement systems for restaurants and bars, and air cleaning systems in restaurants and hospitals (e.g., for reducing cigarette smoke or airborne bacteria). Electrostatic precipitators provide an alternative to bag house filters which operate like large vacuum-cleaner bags that filter pollutants from flowing gas.

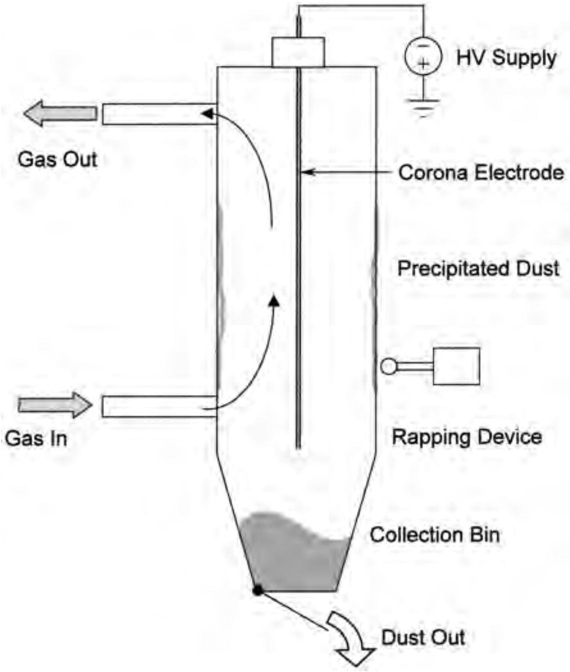


Figure 2.12 Schematic diagram of a single-stage, cylindrical electrostatic precipitator. Dust-laden air enters at the bottom of the stack; clean air exits the top. Negative corona current charges the particles, which then precipitate on the chamber walls. Mechanical "rapping" is used to help dislodge the dust to the collection bin.

The pressure drop across a bag-house filtration system can be very large, hence smaller pressure drop is one principal advantage of electrostatic precipitator systems. Another advantage of an electrostatic precipitator is its lower power consumption compared to a bag-house system, because less air handling equipment is required. The overall pressure drop in a large, industrial-scale electrostatic precipitator, for which the gas flow rate may exceed 1000 m^3/min, is typically less than 10 mm H_2O from source to exit [32].

The basic elements of a precipitator system are shown in Fig. 2.12. The particle-laden gas stream flows through a collection of corona electrodes mounted inside a rigid duct. The corona electrodes can be thin, parallel wires suspended on insulators, or a series of sharp points facing the duct walls. As discussed in Sec. 2.9, corona current will flow once the applied voltage exceeds the critical onset value expressed by Peek's formula, Eq. (2.21). In a large industrial precipitator, this onset voltage might be in the tens-of-kilovolts range, while the onset voltage in a small scale room precipitator is usually below 10 kV. It is difficult to achieve stable corona discharge below about 5 kV because the small gap sizes required to achieve Peek's field often lead to complete spark breakdown across the electrode gap.

The electrodes in an electrostatic precipitator serve two functions. The corona discharge produces a steady stream of ions which charge the airborne particles via the ion-impact charging mechanism described in Sec. 2.11. The charged particles then experience a transverse coulomb force $q\mathbf{E}$ and migrate toward the walls of the duct where they are collected and later removed by one of several cleaning methods. These methods include periodic washing of the duct walls, mechanical rapping to cause the particles to fall into a collection bin, and replacement of the duct's inner lining. This last method is usually reserved for small, bench-top systems.

Although most airborne particles will be neither spherical nor perfectly conducting, the model of Eqs. (2.32)–(2.36) often provides a reasonable estimate of particle charging dynamics. One important requirement is that the particles have enough residence time in the corona-ion flux to become charged to saturation and to precipitate on the collection walls of the duct.

Two problems of concern in the design of electrostatic precipitators include gradient force motion of dielectric or conducting particles, and a phenomenon known as *back ionization*. Gradient force, which is independent of particle charge, occurs whenever a particle is situated in an electric field whose magnitude changes with position, that is, when $\nabla|\mathbf{E}| \neq 0$. This phenomenon is illustrated schematically in Fig. 2.13 for a conducting, spherical particle. The free electrons inside the particle migrate toward the left and leave positive charge to the right, thereby forming a dipole moment. The electric field in Fig. 2.13a is stronger on the right side of the particle, hence the positive end of the dipole experiences a stronger force than does the negative end, leading to a net force to the right. In Fig. 2.13b, the field gradient and force direction, but not the orientation of the dipole, are reversed. For the simple system of Fig. 2.13, the force can be expressed by the one-dimensional spatial derivative of the field:

$$F_x = (qd)\frac{dE_x}{dx} \tag{2.43}$$

In three-dimensional vector notation, the dipole moment is usually expressed as

$$\mathbf{p} = q\mathbf{d}$$

where **d** is a vector pointing from the negative charge of the dipole to its positive charge. Hence in three dimensions, Eq. (2.43) becomes

$$\mathbf{F} = (\mathbf{p}\cdot\nabla)\mathbf{E} \tag{2.44}$$

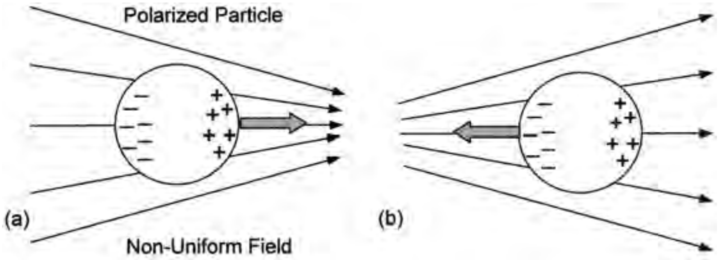


Figure 2.13 A conducting or dielectric particle in a nonuniform field. The particle is polarized, pulling the positive end in the direction of the field lines and the negative end against them. The net force on the particle will be toward the side that experiences the stronger field magnitude. (a) Positive force dominates; (b) negative force dominates.

Equation (2.44) also applies in the more general case of a dielectric particle, where the force density **f** is expressed in terms of the polarization vector $\mathbf{P} = n\mathbf{p}$:

$$\mathbf{f} = (\mathbf{P} \cdot \nabla)\mathbf{E} \tag{2.45}$$

where n is the number of polarized dipoles per unit volume.

In the corona electrode configuration of an electrostatic precipitator, the field gradient is most pronounced near the corona-producing electrode. Here the gradient force can exceed the coulomb force in magnitude and cause pollutant particles to migrate toward, and deposit on, the high voltage electrode rather than on the collector plate. This phenomenon reduces the efficiency of the precipitator but can be avoided by ensuring that the particles acquire saturation charge quickly as they flow through the duct.

The second problem in precipitation, called *back ionization*, or sometimes *back corona* [32,33], occurs when the precipitated particles have high resistivity. The corona current passing through the built-up layer on its way to the duct walls can raise the surface potential of the layer. If this surface potential exceeds the breakdown strength of air, a discharge occurs in the layer, liberating electrons and producing positive ions. These ions migrate toward the negative electrode and tend to neutralize the pollutant particles. This process can greatly reduce the collection efficiency of the precipitator.

2.17. FIELD AND CHARGE MEASUREMENT

The ability to measure electrostatic fields and charge is important in many scientific and engineering disciplines. Measuring these quantities usually requires specialized instrumentation, because a standard voltmeter is useful in only a limited set of circumstances. For example, if one attempts to measure the potential of a charged, electrically isolated conductor with a voltmeter, as in Fig. 2.14, the internal impedance of the meter will fix the conductor potential at zero and allow its charge to flow to ground, thereby obscuring the original quantity to be measured. A standard voltmeter is altogether useless for measuring the potential of a charged *insulator*, because a voltmeter requires that some current, however small, be drawn from the point of measurement. Moreover, the surface of a charged insulator need not be an equipotential; hence the concept of voltage becomes somewhat muddied.

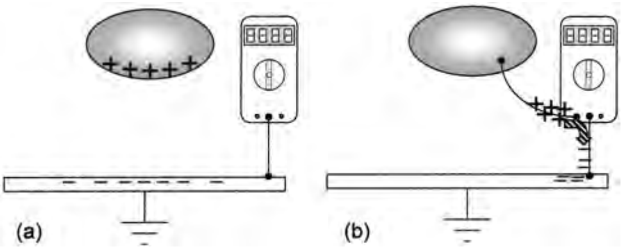


Figure 2.14 Attempting to measure the voltage of a charged, isolated conductor (a) results in a discharged object of zero potential (b).

2.18. ELECTROSTATIC FIELD MILL

Numerous devices have been developed to measure electrostatic fields and voltages, including force sensors [35] and high-impedance solid-state sensors [35–38], but the most prevalent for measuring electrostatic fields has been the variable capacitance field mill [9,10,39,40]. The term *field mill* is used here in its broadest sense to describe any electrostatic field measuring device that relies on mechanical motion to vary the capacitance between the sensor and the source of the field. The variable aperture variety is prevalent in atmospheric science, electric power measurements, and some laboratory instruments, while the vibrating capacitor version can be found in numerous laboratory instruments.

The motivation for the variable aperture field mill comes from the boundary condition for an electric field incident upon a grounded, conducting electrode:

$$\varepsilon E = \rho_s \tag{2.46}$$

or

$$E = \frac{\rho_s}{\varepsilon} \tag{2.47}$$

where ρ_s is the surface chanrge density. A variable aperture field mill modulates the exposed area of a sensing electrode, so that the current flowing to the electrode becomes

$$i = \frac{dQ}{dt} = \frac{d\rho_s A}{dt} = \varepsilon E \frac{dA}{dt} \tag{2.48}$$

For a time-varying, periodic $A(t)$, the peak current magnitude will be proportional to the electric field incident upon the field mill.

One type of variable aperture field mill is depicted in Fig. 2.15. A vibrating vane periodically blocks the underlying sense electrode from the incident field, thereby causing

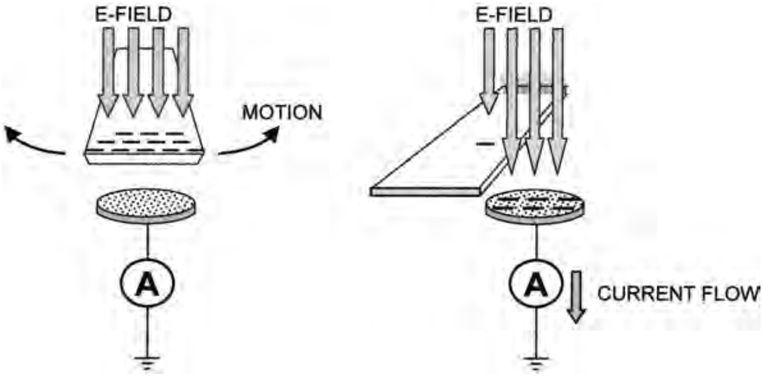


Figure 2.15 Simplified rendition of the variable-aperture field mill.

the induced charge to change periodically. If the exposed area varies sinusoidally as

$$A = A_o \frac{1 + \sin \omega t}{2} \tag{2.49}$$

then the peak current to the sensing electrode will be given by

$$i_{\text{peak}} = \omega \varepsilon E \frac{A_o}{2} \tag{2.50}$$

If the electric field strength varies spatially on a scale comparable to the span of the aperture, then the field mill will respond to the spatial average of the incident field taken over the aperture area. Fields with small-scale spatial variations are found in several industrial, biological, and micromechanical applications. Aperture diameters as small as 0.5 mm are practical and may be found inside the probes of commercially available field meters and noncontacting voltmeters.

2.19. NONCONTACTING VOLTMETER

The field mill described in the previous section is important to an instrument known as the feedback-null surface potential monitor, or *noncontacting voltmeter* [9,10,41–43]. Commercial versions of these instruments are standard equipment in most electrostatics laboratories. The most salient feature of this measurement method is that surface potentials can be measured without physical contact. The basic operating principle of the meter is illustrated in Fig. 2.16. A small field mill is mounted on the end of a hand-held probe, but its outer housing is *not* connected to ground. The output signal of the field-mill feeds a phase-sensitive detection circuit and high-voltage amplifier. The output of the latter is connected back to the probe housing, thereby forming a negative feedback loop. When the probe encounters an object at nonzero potential, the detected field signal, amplified by the high-voltage amplifier, raises the potential of the probe until the field incident on the probe approaches zero. This concept is illustrated in Fig. 2.17. The feedback loop attains equilibrium when the probe body is raised to the same potential as the surface being measured, resulting in only a small residual field at the probe aperture. The residual signal in this "null-field" condition can be made arbitrarily small by increasing the gain of the high-voltage amplifier. Under equilibrium feedback conditions,

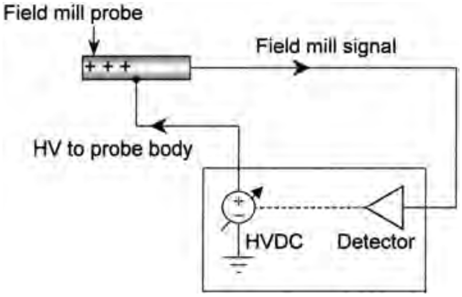


Figure 2.16 Basic structure of a noncontacting voltmeter.

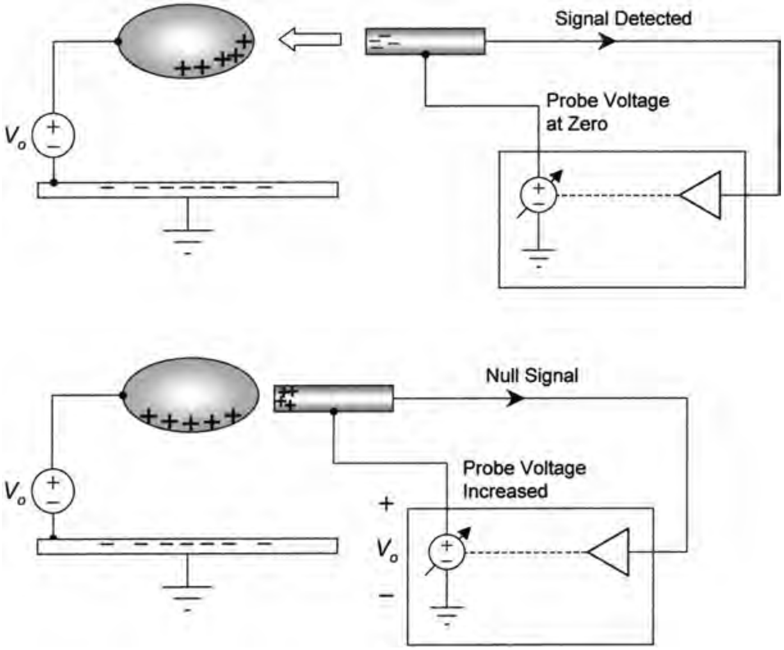


Figure 2.17 The noncontacting voltmeter in operation. Top: Probe approaches charged object to be measured. Bottom: Probe potential is raised until the field it measures is zero (null signal condition).

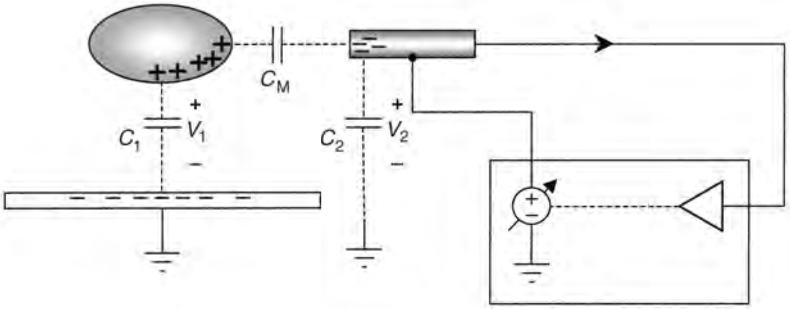


Figure 2.18 Using a noncontacting voltmeter to measure a charged, electrically isolated conductor.

the high voltage on the probe body, monitored using any suitable metering circuit, provides a measure of the surface potential. If the surface potential varies spatially, the meter output will reflect the spatial average encountered by the probe's aperture. The measuring range of the instrument is determined by the positive and negative saturation limits of the high-voltage amplifier. Values up to a few kilovolts (positive and negative) are typical for most commercial instruments.

When a noncontacting voltmeter reads the surface of a conductor connected to a fixed voltage source, the reading is unambiguous. If the probe approaches a floating conductor, the situation can be modeled by the two-body capacitance system of Fig. 2.18.

In this diagram, C_1 and C_2 denote the capacitances to ground of the conductor and probe, respectively, and C_M represents their mutual capacitance. The charge Q_1 on the conductor will be given by [44]

$$Q_1 = C_1 V_1 + C_{\mathrm{M}}(V_1 - V_2) \tag{2.51}$$

The feedback loop of the meter will raise the potential of the probe until $V_2 = V_1$, so that Eq. (2.51) becomes

$$V_1 = \frac{Q_1}{C_1} \tag{2.52}$$

This unambiguous result reflects the potential of the floating conductor with the probe absent.

One of the more common uses of noncontacting voltmeters involves the measurement of charge on insulating surfaces. If surface charge on an insulating layer is tightly coupled to an underlying ground plane, as in Fig. 2.19, the surface potential V_s of the charge layer will be well defined. Specifically, if the layer has thickness d , the surface potential becomes

$$V_s = E \cdot d = \left(\frac{\rho_s}{\varepsilon}\right) d \tag{2.53}$$

The surface charge and its ground-plane image function as a double layer that introduces a potential jump between the ground plane and the upper surface of the insulator. The potential of a noncontacting voltmeter probe placed near the surface will be raised to the same potential V_s, allowing the surface charge ρ_s to be determined from Eq. (2.52).

If the charge on the insulator is not tightly coupled to a dominant ground plane, its surface potential will be strongly influenced by the position of the probe as well as by the insulator's position relative to other conductors and dielectrics. Under these conditions, the reading of the noncontacting voltmeter becomes extremely sensitive to probe position and cannot be determined without a detailed analysis of the fields surrounding the charge [45]. Such an analysis must account for two superimposed components: the field E_Q produced by the measured charge with the probe grounded, and the field E_V created by the voltage of the probe with the surface charge absent. The voltmeter will raise the probe potential until a null-field condition with $E_Q + E_V = 0$ is reached. Determining the relationship between the resulting probe voltage and the unknown surface charge requires a detailed field solution that takes into account the probe shape, probe position, and

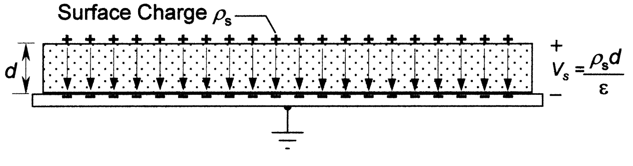


Figure 2.19 Surface charge on an insulator situated over a ground plane. The voltage on the surface of the insulator is clearly defined as $\rho_s d/\varepsilon$.

insulator geometry. Because of the difficulty in translating voltmeter readings into actual charge values, noncontacting voltmeter measurements of isolated charge distributions that are not tightly coupled to ground planes are best used for relative measurement purposes only. A noncontacting voltmeter used in this way becomes particularly useful when measuring the decay time of a charge distribution. The position of the probe relative to the surface must remain fixed during such a measurement.

2.20. MICROMACHINES

The domain of micro-electromechanical systems, or MEMS, involves tiny microscale machines made from silicon, titanium, aluminum, or other materials. MEMS devices are fabricated using the tools of integrated-circuit manufacturing, including photolithography, pattern masking, deposition, and etching. Design solutions involving MEMS are found in many areas of technology. Examples include the accelerometers that deploy safety airbags in automobiles, pressure transducers, microfluidic valves, optical processing systems, and projection display devices.

One technique for making MEMS devices is known as *bulk micromachining*. In this method, microstructures are fabricated within a silicon wafer by a series of selective etching steps. Another common fabrication technique is called *surface micromachining*. The types of steps involved in the process are depicted in Fig. 2.20. A silicon substrate is patterned with alternating layers of polysilicon and oxide thin films that are used to build up the desired structure. The oxide films serve as *sacrificial layers* that support the

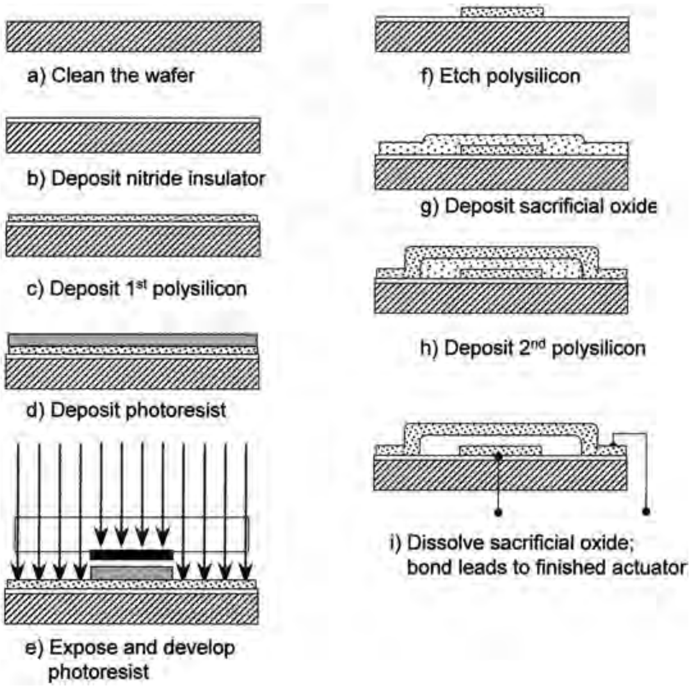

a) Clean the wafer
b) Deposit nitride insulator
c) Deposit 1st polysilicon
d) Deposit photoresist
e) Expose and develop photoresist
f) Etch polysilicon
g) Deposit sacrificial oxide
h) Deposit 2nd polysilicon
i) Dissolve sacrificial oxide; bond leads to finished actuator

Figure 2.20 Typical surface micromachining steps involved in MEMS fabrication. Oxides are used as sacrificial layers to produce structural members. A simple actuator is shown here.

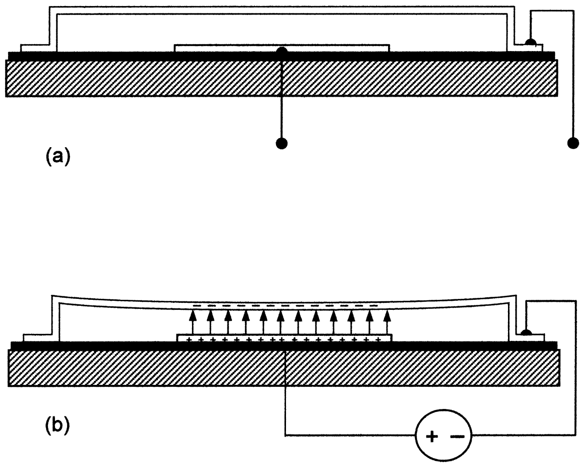


Figure 2.21 Applying a voltage to the actuator causes the membrane structure to deflect toward the substrate. The drawing is not to scale; typical width-to-gap spacing ratios are on the order of 100.

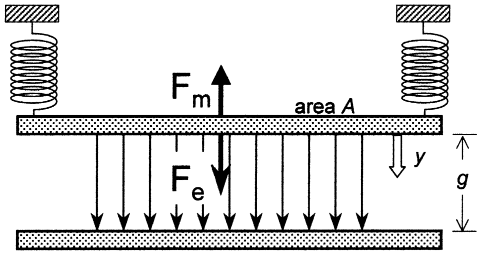


Figure 2.22 The MEMS actuator of Fig. 2.21 can be modeled by the simple mass-spring structure shown here. F_e is the electrostatic force when a voltage is applied; F_m is the mechanical restoring force.

polysilicon during sequential deposition steps but are removed in the final steps of fabrication. This construction technique is analogous to the way that stone arches were made in ancient times. Sand was used to support stone pieces and was removed when the building could support itself, leaving the finished structure.

One simple MEMS device used in numerous applications is illustrated in Fig. 2.21. This *double-cantilevered actuator* consists of a bridge supported over a fixed activation electrode. The bridge has a rectangular shape when viewed from the top and an aspect ratio (ratio of width to gap spacing) on the order of 100. When a voltage is applied between the bridge and the substrate, the electrostatic force of attraction causes the bridge to deflect downward. This vertical motion can be used to open and close valves, change the direction of reflected light, pump fluids, or mix chemicals in small micromixing chambers.

The typical bridge actuator has a gap spacing of a few microns and lateral dimensions on the order of 100 to 300 μm. This large aspect ratio allows the actuator to be modeled by the simple two-electrode capacitive structure shown in Fig. 2.22.

The electrostatic force in the y direction can be found by taking the derivative of the stored energy (see Sec. 2.10):

$$F_E = \frac{\partial}{\partial y}\frac{1}{2}CV^2 = \frac{\varepsilon_0 A V^2}{(g-y)^2} \tag{2.54}$$

Here y is the deflection of the bridge, A its surface area, and g the gap spacing at zero deflection. As Eq. (2.54) shows, the electrostatic force increases with increasing deflection and becomes infinite as the residual gap spacing $(g-y)$ approaches zero. To first order, the mechanical restoring force will be proportional to the bridge deflection and can be expressed by the simple equation

$$F_M = -ky \tag{2.55}$$

The equilibrium deflection y for a given applied voltage will occur when $F_\mathrm{M} = F_\mathrm{E}$, i.e., when

$$ky = \frac{\varepsilon_0 A V^2}{(g-y)^2} \tag{2.56}$$

Figure 2.23 shows a plot of y versus V obtained from Eq. (2.56). For voltages above the critical value V_c, the mechanical restoring force can no longer hold back the electrostatic force, and the bridge collapses all the way to the underlying electrode. This phenomenon, known as *snap-through*, occurs at a deflection of one third of the zero-voltage gap spacing. It is reversible only by setting the applied voltage to zero and sometimes cannot be undone at all due to a surface adhesion phenomenon known as *sticktion*.

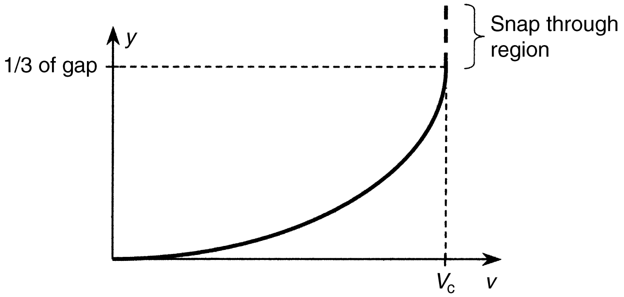


Figure 2.23 Voltage displacement curve for the actuator model of Fig. 2.22. At a deflection equal to one-third of the gap spacing, the electrostatic force overcomes the mechanical restoring force, causing the membrane to "snap through" to the substrate.

The deflection at which snap-through occurs is easily derived by noting that at $v = V_c$, the slope of the voltage–displacement curve becomes infinite, i.e., dV/dy becomes zero. Equation (2.56) can be expressed in the form

$$V = \sqrt{\frac{ky}{\varepsilon_0 A}}(g - y) \tag{2.57}$$

The y derivative of this equation becomes zero when $y = g/3$.

2.21. DIGITAL MIRROR DEVICE

One interesting application of the MEMS actuator can be found in the digital mirror device (DMD) used in computer projection display systems. The DMD is an array of electrostatically-actuated micromirrors of the type shown in Fig. 2.24. Each actuator is capable of being driven into one of two bi-stable positions. When voltage is applied to the right-hand pad, as in Fig 2.24a, the actuator is bent to the right until it reaches its mechanical limit. Alternatively, when voltage is applied to the left-hand pad, as in Fig. 2.24b, the actuator bends to the left. The two deflection limits represent the logic **0** (no light projected) and logic **1** (light projected) states of the mirror pixel.

2.22. ELECTROSTATIC DISCHARGE AND CHARGE NEUTRALIZATION

Although much of electrostatics involves harnessing the forces of charge, sometimes static electricity can be most undesirable. Unwanted electrostatic forces can interfere with materials and devices, and sparks from accumulated charge can be quite hazardous in the vicinity of flammable liquids, gases, and air dust mixtures [12, 46–51]. In this section, we examine situations in which electrostatics is a problem and where the main objective is to eliminate its effects.

Many manufacturing processes involve large moving webs of insulating materials, such as photographic films, textiles, food packaging materials, and adhesive tapes. These materials can be adversely affected by the presence of static electricity. A moving web is easily charged by contact electrification because it inevitably makes contact with rollers, guide plates, and other processing structures. These contact and separation events provide ample opportunity for charge separation to occur [52]. A charged web can be attracted to parts of the processing machinery, causing jams in the machinery or breakage of the web material. In some situations, local surface sparks may also occur that can ruin the

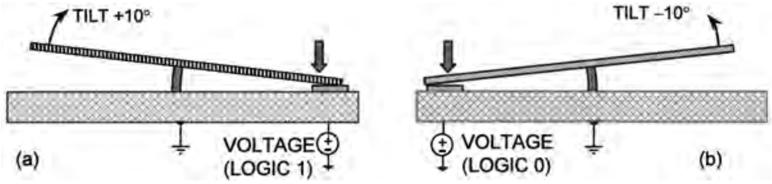


Figure 2.24 Simplified schematic of digital mirror device. Each pixel tilts ± 10° in response to applied voltages.

processed material. This issue is especially important in the manufacturing of photographic films, which can be prematurely exposed by the light from sparks or other discharges.

Electrostatic charge is very undesirable in the semiconductor industry. Sensitive semiconductor components, particularly those that rely on metal-oxide-semiconductor (MOS) technology, can be permanently damaged by the electric fields from nearby charged materials or by the discharges that occur when charged materials come into contact with grounded conductors. Discharges similar to the "carpet sparks" that plague temperate climates in winter can render semiconductor chips useless. A static-charged wafer also can attract damaging dust particles and other contaminants.

The term *electrostatic discharge* (ESD) refers to any unwanted sparking event caused by accumulated static charge. An abundance of books and other resources may be found in the literature to aid the electrostatics professional responsible for preventing ESD in a production facility [53–58].

Numerous methods exist to neutralize accumulated charge before it can lead to an ESD event. The ionizing neutralizer is one of the more important devices used to prevent the build up of unwanted static charge. An ionizer produces both positively and negatively charged ions of air that are dispersed in proximity to sensitive devices and work areas. When undesirable charge appears on an object from contact electrification or induction charging, ions of the opposite polarity produced by the ionizer are attracted to the object and quickly neutralize it. The relatively high mobility of air ions allows this neutralization to occur rapidly, usually in a matter of seconds or less.

The typical ionizer produces ions via the process of corona discharge. A coronating conductor, usually a sharp needle point, or sometimes a thin, axially mounted wire, is energized to a voltage on the order of 5 to 10 kV. An extremely high electric field develops at the electrode, causing electrons to be stripped from neutral air molecules via an avalanche multiplication process (see Sec. 2.9). In order to accommodate unwanted charge of either polarity, and to avoid inadvertent charging of surfaces, the ionizer must simultaneously produce balanced quantities of positive and negative charge. Some ionizers produce bipolar charge by applying an ac voltage to the corona electrode. The ionizer thus alternately produces positive and negative ions that migrate as a bipolar charge cloud toward the work piece. Ions having polarity opposite the charge being neutralized will be attracted to the work surface, while ions of the same polarity will be repelled. The undesired charge thus extracts from the ionizer only what it needs to be neutralized.

Other ionizers use a different technique in which adjacent pairs of electrodes are energized simultaneously, one with positive and the other with negative dc high voltage. Still other neutralizers use separate positive and negative electrodes, but energize first the positive side, then the negative side for different intervals of time. Because positive and negative electrodes typically produce ions at different rates, this latter method of electrification allows the designer to adjust the "on" times of each polarity, thereby ensuring that the neutralizer produces the proper balance of positive and negative ions.

Although the production of yet more charge may seem a paradoxical way to eliminate unwanted charge, the key to the method lies in maintaining a proper balance of positive and negative ions produced by the ionizer, so that no additional net charge is imparted to nearby objects or surfaces. Thus, one figure of merit for a good ionizer is its overall balance as measured by the lack of charge accumulation of either polarity at the work piece served by the ionizer. Another figure of merit is the speed with which an ionizer can neutralize unwanted charge. This parameter is sometimes called the ionizer's *effectiveness*. The more rapidly unwanted static charge can be neutralized, the less chance it will have to affect sensitive electronic components or interfere with a production process. Effectiveness of an ionizer is maximized by transporting the needed charge as rapidly as possible to the neutralized object [21]. Sometimes this process is assisted by air flow from a fan or blowing air stream. Increasing the density of ions beyond some minimum level does not increase effectiveness because the extra ions recombine quickly.

2.23. SUMMARY

This chapter is intended to serve as an introduction to the many applications of electrostatics in science, technology, and industry. The topics presented are not all inclusive of this fascinating and extensive discipline, and the reader is encouraged to explore some of the many reference books cited in the text. Despite its long history [59], electrostatics is an ever-evolving field that seems to emerge anew with each new vista of discovery.

REFERENCES

1. Schein, L.B.; LaHa, M.; Novotny, D. Theory of insulator charging. Phys. Lett. **1992**, *A 167*, 79–83.
2. Horn, R.G.; Smith, D.T. Contact electrification and adhesion between dissimilar materials. Science **1992**, *256*, 362–364.
3. Harper, W.R. Contact and frictional electrification. In *Monographs on the Physics and Chemistry of Materials*; Clarendon Press: Oxford, 1967.
4. Shinbrot, T. A look at charging mechanisms. J. Electrostat. **1985**, *17*, 113–123.
5. Davies, D.K. Charge generation of dielectric surfaces. J. Phys. **1969**, *D2*, 1533.
6. Schein, L.B.; Cranch, J. The static electrification of mixtures of insulating powders. J. Appl. Phys. **1975**, *46*, 5140.
7. Schein, L.B.; Castle, G.S.P.; Dean, A. Theory of monocomponent development. J. Imag. Technol **1989**, *15*, 9.
8. Schein, L.B.; LaHa, M.; Novotny, D. Theory of insulator charging. Phys. Lett. **1992**, *A 167*, 79.
9. Cross, J. *Electrostatics: Principles, Problems and Applications*; IOP Publishing: Bristol, 1987; 500.
10. Taylor, D.M.; Secker, P.E. *Industrial Electrostatics*; John Wiley and Sons: New York, 1994.
11. Montgomery, D.J. Static electrification in solids. Solid State Phys. **1959**, *9*, 139–197.
12. Glor, M. *Electrostatic Hazards in Powder Handling*; John Wiley and Sons: New York, 1988.
13. Coehn, A. Ann. Physik, **1898**, *64*, 217.
14. JW (Lord) Raleigh, On the equilibrium of liquid conducting masses charged with electricity. Phil. Mag. **1882**, *14*, 184–186.
15. Melcher, J.R. *Continuum Electromechanics*; MIT Press: Cambridge, Massachusetts, 1981, 8.44.
16. Bailey, A.G. *Electrostatic Spraying of Liquids*; John Wiley and Sons: New York, 1988.
17. Law, S.E. Electrostatic atomization and spraying. In *Handbook of Electrostatic Processes*; Chang, J.S., Kelly, A.J., Crowley, J.M., Eds.; Marcel Dekker: New York, 1995; 413–440.
18. Cobine, J.D. *Gaseous Conductors*; Dover Press: New York, 1958, 252–281.
19. Tobazéon, R. Electrical phenomena of dielectric materials. In *Handbook of Electrostatic Processes*; Chang, J.S., Kelly, A.J., Crowley, J.M., Eds.; Marcel Dekker: New York, 1995; 51–82.
20. Peek, F.W. *Dielectric Phenomena in High Voltage Engineering*; McGraw-Hill: New York, 1929, 48–108.
21. Crowley, J.M. *Fundamentals of Applied Electrostatics*; Wiley: New York, 1986, 164, 207–225.

22. Haus, H.; Melcher, J.R. *Electromagnetic Fields and Energy*; Prentice-Hall: Englewood Cliffs, NJ, 1989, 486–521.
23. Woodson, H.; Melcher, J.R. *Electromechanical Dynamics*; John Wiley and Sons: New York, 1968, Chapter 8.
24. Zahn, M., *Electromagnetic Field Theory: A Problem Solving Approach*; John Wiley and Sons: New York, 1979, 204–230.
25. Law, S.E. Electrostatic pesticide spraying: concepts and practice. IEEE Trans. **1983**, *IA-19* (2), 160–168.
26. Inculet, I.I.; Fisher, J.K. Electrostatic aerial spraying. IEEE Trans. **1989**, *25* (3).
27. Pauthenier, M.M.; Moreau-Hanot, M. La charge des particules spheriques dans un champ ionize. J. Phys. Radium (Paris) **1932**, *3*, 590.
28. Schein, L.B. *Electrophotography and Development Physics*; 2nd Ed.; Springer Verlag: New York, 1992.
29. White, H.J. *Industrial Electrostatic Precipitation*; Reading, Addison-Wesley: MA, 1962.
30. Masuda, S.; Hosokawa, H. *Electrostatic precipitation*. In *Handbook of Electrostatics*; Chang, J.S., Kelly, A.J., Crowley, J.M., Eds.; Marcel Dekker: New York, 1995; 441–480.
31. Masuda, S. Electrical precipitation of aerosols. Proc. 2nd Aerosol Int. Conf., Berlin, Germany: Pergamon Press, 1986; 694–703.
32. White, H.J. Particle charging in electrostatic precipitation. AIEE Trans. Pt. 1, *70*, 1186.
33. Masuda, S.; Nonogaki, Y. Detection of back discharge in electrostatic precipitators. Rec. IEEE/IAS Annual Conference, Cincinnati, Ohio, 1980; 912–917.
34. Masuda, S.; Obata, T.; Hirai, J. A pulse voltage source for electrostatic precipitators. Rec. IEEE/IAS Conf., Toronto, Canada, 1980; 23–30.
35. Nyberg, B.R.; Herstad, K.; Larsen, K.B.; Hansen, T. Measuring electric fields by using pressure sensitive elements. IEEE Trans. Elec. Ins, **1979**, *EI-14*, 250–255.
36. Horenstein, M. A direct gate field-effect transistor for the measurement of dc electric fields. IEEE Trans. Electr. Dev. **1985**, *ED-32* (3): 716.
37. McCaslin, J.B. Electrometer for ionization chambers using metal-oxide-semiconductor field-effect transistors. Rev. Sci. Instr. **1964**, *35* (11), 1587.
38. Blitshteyn, M. Measuring the electric field of flat surfaces with electrostatic field meters. Evaluation Engineering, **Nov. 1984**, *23* (10), 70–86.
39. Schwab, A.J. *High Voltage Measurement Techniques*; MIT Press: Cambridge, MA, 1972, 97–101.
40. Secker, P.E. Instruments for electrostatic measurements. J. Elelectrostat. **1984**, *16* (1), 1–19.
41. Vosteen, R.E.; Bartnikas, R. Electrostatic charge measurement. Engnr Dielectrics, Vol IIB, Electr Prop Sol Insul Matls, ASTM Tech Publ 926, 440–489.
42. Vosteen, W. A high speed electrostatic voltmeter technique. Proc IEEE Ind Appl Soc Annual Meeting IAS-88(2): 1988; 1617–1619.
43. Horenstein, M. Measurement of electrostatic fields, voltages, and charges. In *Handbook of Electrostatics*; Chang, J.S., Kelly, A.J., Crowley, J.M. Eds.; Marcel Dekker: New York, 1995; 225–246.
44. Popovic, Z.; Popovic, B.D. *Introductory Electromagnetics*; Prentice-Hall: Upper Saddle River, NJ, 2000; 114–115.
45. Horenstein, M. Measuring surface charge with a noncontacting voltmeter. J. Electrostat. **1995**, *35*, 2.
46. Gibson, N.; Lloyd, F.C. Incendivity of discharges from electrostatically charged plastics. Brit. J. Appl. Phys. **1965**, *16*, 619–1631.
47. Gibson, N. Electrostatic hazards. In Electrostatics ’83; Inst. Phys. Conf. Ser. No. 66, Oxford, 1983; 1–11.
48. Glor, M. Hazards due to electrostatic charging of powders. J. Electrostatics **1985**, *16*, 175–181.
49. Pratt, T.H. *Electrostatic Ignitions of Fires and Explosions*; Burgoyne: Marietta, GA, 1997, 115–152.

50. Lüttgens, G.; Wilson, N. *Electrostatic Hazards*; Butterworth-Heinemann: Oxford, 1997, 137–148.
51. Bailey, A.G. Electrostatic hazards during liquid transport and spraying. In *Handbook of Electrostatics*; Chang, J.S., Kelly, A.J., Crowley, J.M., Eds.; Marcel Dekker: New York, 1995; 703–732.
52. Hughes, J.F.; Au, A.M.K.; Blythe, A.R. Electrical charging and discharging between films and metal rollers. Electrostatics '79. Inst. Phys. Conf. Ser. No. 48, Oxford, 1979; 37–44.
53. Horvath, T.; Berta, I. *Static Elimination*; Research Studies Press: New York, 1982; 118.
54. Davies, D.K. Harmful effects and damage to electronics by electrostatic discharges. J. Electrostatics **1985**, *16*, 329–342.
55. McAteer, O.J.; Twist, R.E. Latent ESD failures, EOS/ESD Symposium Proceedings, Orlando, FL, 1982; 41–48.
56. Boxleitner, W. *Electrostatic Discharge and Electronic Equipment: A Practical Guide for Designing to Prevent ESD Problems*; IEEE Press: New York, 1989, 73–84.
57. McAteer, O.J. *Electrostatic Discharge Control*; McGraw-Hill: New York, 1990.
58. Greason, W. *Electrostatic Discharge in Electronics*; John Wiley and Sons: New York, 1993.
59. Moore, A.D. *Electrostatics and Its Applications*; John Wiley and Sons: New York, 1973.

3
Magnetostatics

Milica Popović
McGill University
Montréal, Quebec, Canada

Branko D. Popović[†]
University of Belgrade
Belgrade, Yugoslavia

Zoya Popović
University of Colorado
Boulder, Colorado, U.S.A.

> To the loving memory of our father, professor, and coauthor. We hope that he would have agreed with the changes we have made after his last edits.
>
> — *Milica and Zoya Popovic*

3.1. INTRODUCTION

The force between two static electric charges is given by Coulomb's law, obtained directly from measurements. Although small, this force is easily measurable. If two charges are *moving*, there is an *additional* force between them, the *magnetic force*. The magnetic force between *individual* moving charges is extremely small when compared with the Coulomb force. Actually, it is so small that it cannot be detected experimentally between just a pair of moving charges. However, these forces can be measured using a vast number of electrons (practically one per atom) in organized motion, i.e., electric current. Electric current exists within almost electrically neutral materials. Thus, magnetic force can be measured independent of electric forces, which are a result of charge unbalance.

Experiments indicate that, because of this vast number of interacting moving charges, the magnetic force between two current-carrying conductors can be much larger than the maximum electric force between them. For example, strong electromagnets can carry weights of several tons, while electric force cannot have even a fraction of that strength. Consequently, the magnetic force has many applications. For example, the approximate direction of the North Magnetic Pole is detected with a magnetic device—a compass. Recording and storing various data are most commonly accomplished using

[†] Deceased.

magnetic storage components, such as computer disks and tapes. Most household appliances, as well as industrial plants, use motors and generators, the operation of which is based on magnetic forces.

The goal of this chapter is to present:

Fundamental theoretical foundations for magnetostatics, most importantly Ampere's law

Some simple and commonly encountered examples, such as calculation of the magnetic field inside a coaxial cable

A few common applications, such as Hall element sensors, magnetic storage, and MRI medical imaging.

3.2. THEORETICAL BACKGROUND AND FUNDAMENTAL EQUATIONS

3.2.1. Magnetic Flux Density and Lorentz Force

The electric force on a charge is described in terms of the electric field vector, **E**. The magnetic force on a charge moving with respect to other moving charges is described in terms of the *magnetic flux density vector*, **B**. The unit for **B** is a *tesla* (T). If a point charge Q [in coulombs (C)] is moving with a velocity **v** [in meters per second (m/s)], it experiences a force [in newtons (N)] equal to

$$\mathbf{F} = Q\mathbf{v} \times \mathbf{B} \tag{3.1}$$

where "×" denotes the vector product (or cross product) of two vectors.

The region of space in which a force of the form in Eq. (3.1) acts on a moving charge is said to have a *magnetic field* present. If in addition there is an electric field in that region, the total force on the charge (the Lorentz force) is given by

$$\mathbf{F} = Q\mathbf{E} + Q\mathbf{v} \times \mathbf{B} \tag{3.2}$$

where **E** is the electric field intensity in volts per meter (V/m).

3.2.2. The Biot–Savart Law

The magnetic flux density is produced by current-carrying conductors or by permanent magnets. If the source of the magnetic field is the electric current in thin wire loops, i.e. current loops, situated in vacuum (or in air), we first adopt the orientation along the loop to be in the direction of the current in it. Next we define the product of the wire current, I, with a short vector length of the wire, $d\mathbf{l}$ (in the adopted reference direction along the wire), as the *current element*, $I\,d\mathbf{l}$ (Fig. 3.1a). With these definitions, the magnetic flux density due to the entire current loop C (which may be arbitrarily complex), is at any point given by the experimentally obtained Biot–Savart law:

$$\mathbf{B} = \frac{\mu_0}{4\pi} \oint_C \frac{I\,d\mathbf{l} \times \mathbf{a}_r}{r^2} \tag{3.3}$$

The unit vector $\mathbf{a}_r$ is directed *from the source point* (i.e., the current element) *towards the field point* (i.e., the point at which we determine **B**). The constant μ_0 is known as the

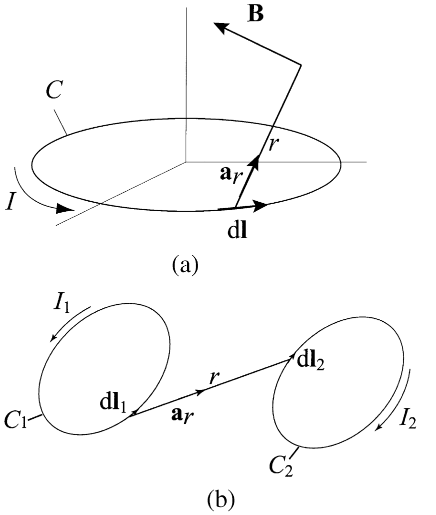


Figure 3.1 (a) A current loop with a current element. (b) Two current loops and a pair of current elements along them.

permeability of vacuum. Its value is defined to be exactly

$$\mu_0 = 4\pi \times 10^{-7}\,\mathrm{H/m}$$

Note that the magnetic flux density vector of individual current elements is perpendicular to the plane of the vectors **r** and $d\mathbf{l}$. Its orientation is determined by the right-hand rule when the vector $d\mathbf{l}$ is rotated by the shortest route towards the vector $\mathbf{a}_r$. The (vector) integral in Eq. (3.3) can be evaluated in closed form in only a few cases, but it can be always evaluated numerically.

The line current I in Eq. (3.3) is an approximation to volume current. Volume currents are described by the current density vector, **J** [amperes per meter squared (A/m^2)]. Let ΔS be the cross-sectional area of the wire. The integral in Eq. (3.3) then becomes a volume integral where $Id\mathbf{l}$ is replaced by $\mathbf{J}\cdot\Delta S\cdot d\mathbf{l}= \mathbf{J}\cdot dv$. At high frequencies (above about 1MHz), currents in metallic conductors are distributed in very thin layers on conductor surfaces (the *skin effect*). These layers are so thin that they can be regarded as geometrical surfaces. In such cases we introduce the concept of *surface current density* $\mathbf{J}_s$ (in A/m), and the integral in Eq. (3.3) becomes a surface integral, where $Id\mathbf{l}$ is replaced by $\mathrm{J}_s\,dS$.

3.2.3. Units: How Large is a Tesla?

The unit for magnetic flux density in the SI system is a tesla* (T). A feeling for the magnitude of a tesla can be obtained from the following examples. The magnetic flux density of the earth's dc magnetic field is on the order of 10^{-4} T. Around current-carrying

*The unit was named after the American scientist of Serbian origin Nikola Tesla, who won the ac–dc battle over Thomas Edison and invented three-phase systems, induction motors, and radio transmission. An excellent biography of this eccentric genius is *Tesla, Man out of Time*, by Margaret Cheney, Dorset Press, NY, 1989.

conductors in vacuum, the intensity of **B** ranges from about 10^{-6} T to about 10^{-2} T. In air gaps of electrical machines, the magnetic flux density can be on the order of 1 T. Electromagnets used in magnetic-resonance imaging (MRI) range from about 2 T to about 4 T [5,15]. Superconducting magnets can produce flux densities of several dozen T.

3.2.4. Magnetic Force

From Eq. (3.2) it follows that the magnetic force on a current element $I\,d\mathbf{l}$ in a magnetic field of flux density **B** is given by

$$d\mathbf{F} = I\,d\mathbf{l} \times \mathbf{B} \quad \text{(N)} \tag{3.4}$$

Combining this expression with the Biot–Savart law, an expression for the magnetic force between two current loops C_1 and C_2 (Fig. 3.1b) is obtained:

$$d\mathbf{F}_{C1 \text{ on } C2} = \frac{\mu_0}{4\pi} \oint_{C_1} \oint_{C_2} I_2\,d\mathbf{l}_2 \times I_1 d\mathbf{l}_1 \tag{3.5}$$

3.2.5. Magnetic Moment

For a current loop of vector area **S** (the unit vector normal to **S**, by convention, is determined by the right-hand rule with respect to the reference direction along the loop), the *magnetic moment* of the loop, **m**, is defined as

$$\mathbf{m} = I \times \mathbf{S} \tag{3.6}$$

If this loop is situated in a uniform magnetic field of magnetic flux density **B**, the *mechanical* moment, **T**, on the loop resulting from magnetic forces on its elements is

$$\mathbf{T} = \mathbf{m} \times \mathbf{B} \tag{3.7}$$

This expression is important for understanding applications such as motors and generators.

The lines of vector **B** are defined as (generally curved) imaginary lines such that vector **B** is tangential to them at all points. For example, from Eq. (3.3) it is evident that the lines of vector **B** for a single current element are circles centered along the line of the current element and in planes perpendicular to the element.

3.2.6. Magnetic Flux

The flux of vector **B** through a surface is termed the *magnetic flux*. It plays a very important role in magnetic circuits, and a fundamental role in one of the most important electromagnetic phenomena, electromagnetic induction. The magnetic flux, Φ, through a surface S is given by

$$\Phi = \int_S \mathbf{B} \cdot d\mathbf{S} \qquad \text{in webers (Wb)} \tag{3.8}$$

The magnetic flux has a very simple and important property: it is equal to zero through *any* closed surface,

$$\oint_S \mathbf{B} \cdot d\mathbf{S} = 0 \tag{3.9}$$

This relation is known as the *law of conservation of magnetic flux* and represents the fourth Maxwell's equation in integral form. In differential form, it can be written as $\nabla \cdot \mathbf{B} = 0$, using the divergence theorem. An interpretation of the law of conservation of magnetic flux is that "magnetic charges" do not exist, i.e., a south and north pole of a magnet are never found separately. The law tells us also that the lines of vector **B** do not have a beginning or an end. Sometimes, this last statement is phrased more loosely: it is said that the lines of vector **B** close on themselves.

An important conclusion follows: If we have a closed contour C in the field and imagine any number of surfaces spanned over it, *the magnetic flux through any such surface, spanned over the same contour, is the same.* There is just one condition that needs to be fulfilled in order for this to be true: the unit vector normal to all the surfaces must be the same with respect to the contour, as shown in Fig. 3.2. It is customary to orient the contour and then to define the vector unit normal on any surface on it according to the right-hand rule.

3.2.7. Ampere's Law in Vacuum

The magnetic flux density vector **B** resulting from a time-invariant current density **J** has a very simple and important property: If we compute the line integral of **B** along any closed contour C, it will be equal to μ_0 times the total current that flows through any surface spanned over the contour. This is *Ampere's law* for dc (time-invariant) currents in vacuum (Fig. 3.3):

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \int_S \mathbf{J} \cdot d\mathbf{S} \tag{3.10}$$

The reference direction of the vector surface elements of S is adopted according to the right-hand rule with respect to the reference direction of the contour. In the applications of Ampere's law, it is very useful to keep in mind that the flux of the current density vector (the current intensity) is the same through all surfaces having a common boundary

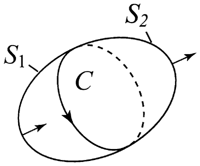


Figure 3.2 Two surfaces, S_1 and S_2, defined by a common contour C, form a closed surface to which the law of conservation of magnetic flux applies—the magnetic flux through them is the same. The direction chosen for the loop determines the normal vector directions for S_1 and S_2 according to the right-hand rule.

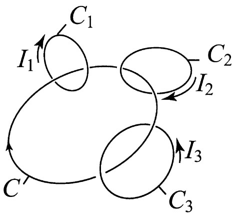


Figure 3.3 Three current loops and the illustration of Ampere's law. The line integral of **B** along C in the case shown equals I_1–I_2–I_3.

contour. Ampere's law is not a new property of the magnetic field—it follows from the Biot–Savart law, which in turn is based on experiment.

Ampere's law in Eq. (3.10) is a general law of the magnetic field of time-invariant (dc) currents in vacuum. It can be extended to cases of materials in the magnetic field, but in this form it is *not* valid for magnetic fields produced by *time-varying* (ac) currents. Since the left-hand side in Ampere's law is a vector integral, while the right-hand side is a scalar, it can be used to determine analytically vector **B** only for problems with a high level of symmetry for which the vector integral can be reduced to a scalar one. Several such practical commonly encountered cases are a cylindrical wire, a coaxial cable and parallel flat current sheets.

3.2.8. Magnetic Field in Materials

If a body is placed in a magnetic field, magnetic forces act on all *moving* charges within the atoms of the material. These moving charges make the atoms and molecules inside the material look like tiny current loops. The moment of magnetic forces on a current loop, Eq. (3.7), tends to align vectors **m** and **B**. Therefore, in the presence of the field, a substance becomes a large aggregate of oriented elementary current loops which produce their own magnetic fields. Since the rest of the body does not produce any magnetic field, a substance in the magnetic field can be visualized as a large set of oriented elementary current loops situated in vacuum. A material in which magnetic forces produce such oriented elementary current loops is referred to as a *magnetized* material. It is possible to replace a material body in a magnetic field with equivalent *macroscopic* currents *in vacuum* and analyze the magnetic field provided that we know how to find these equivalent currents. Here the word *macroscopic* refers to the fact that a small volume of a material is assumed to have a very large number of atoms or molecules.

The number of revolutions per second of an electron around the nucleus is very large—about 10^{15} revolutions/s. Therefore, it is reasonable to say that such a rapidly revolving electron is a small (elementary) current loop with an associated magnetic moment. This picture is, in fact, more complicated since in addition electrons have a magnetic moment of their own (their spin). However, each atom can macroscopically be viewed as an equivalent single current loop. Such an elementary current loop is called an *Ampere current*. It is characterized by its magnetic moment, $\mathbf{m} = I\mathbf{S}$. The macroscopic quantity called the *magnetization vector*, **M**, describes the density of the vector magnetic moments in a magnetic material at a given point and for a substance with N Ampere currents per unit volume can be written as

$$\mathbf{M} = \sum \frac{\mathbf{m}_{\text{in } dv}}{dv} = N\mathbf{m} \tag{3.11}$$

The significance of Eq. (3.11) is as follows. The magnetic field of a single current loop in vacuum can be determined from the Biot–Savart law. The vector **B** of such a loop at large distances from the loop (when compared with the loop size) is proportional to the magnetic moment, **m**, of the loop. According to Eq. (3.11) we can subdivide magnetized materials into small volumes, ΔV, each containing a very large number of Ampere currents, and represent each volume by a single larger Ampere current of moment $\mathbf{M}\,\Delta V$. Consequently, if we determine the magnetization vector at all points, we can find vector **B** by integrating the field of these larger Ampere currents over the magnetized material. This is much simpler than adding the fields of the prohibitively large number of individual Ampere currents.

3.2.9. Generalized Ampere's Law and Magnetic Field Intensity

Ampere's law in the form as in Eq. (3.10) is valid for any current distribution *in vacuum*. Since the magnetized substance is but a vast number of elementary current loops in vacuum, we can apply Ampere's law to fields in materials, provided we find how to include these elementary currents on the right-hand side of Eq. (3.10). The total current of elementary current loops "strung" along a closed contour C, i.e., the total current of all Ampere's currents through the surface S defined by contour C, is given by

$$I_{\text{Ampere through } S} = \oint_C \mathbf{M} \cdot d\mathbf{l} \tag{3.12}$$

The generalized form of Ampere's law valid for time-invariant currents therefore reads

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \mu_0 \left(\int_S \mathbf{J} \cdot d\mathbf{S} + \oint_C \mathbf{M} \cdot d\mathbf{l} \right) \tag{3.13}$$

Since the contour C is the same for the integrals on the left-hand and right-hand sides of the equation, this can be written as

$$\oint_C \left(\frac{\mathbf{B}}{\mu_0} - \mathbf{M} \right) \cdot d\mathbf{l} = \int_S \mathbf{J} \cdot d\mathbf{S} \tag{3.14}$$

The combined vector $\mathbf{B}/\mu_0 - \mathbf{M}$ has a convenient property: Its line integral along any closed contour depends only on the *actual* current through the contour. This is the only current we can control—switch it on and off, change its intensity or direction, etc. Therefore, the combined vector is defined as a new vector that describes the magnetic field in the presence of materials, known as the *magnetic field intensity*, **H**:

$$\mathbf{H} = \frac{\mathbf{B}}{\mu_0} - \mathbf{M} \qquad (\text{A/m}) \tag{3.15}$$

With this definition, the generalized Ampere's law takes the final form:

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = \int_S \mathbf{J} \cdot d\mathbf{S} \tag{3.16}$$

As its special form, valid for currents in vacuum, this form of Ampere's law is also valid *only for time-constant (dc) currents.*

The definition of the magnetic field intensity vector in Eq. (3.15) is general and valid for any material. Most materials are those for which the magnetization vector, **M**, is a linear function of the local vector **B** (the cause of material magnetization). In such cases a linear relationship exists between any two of the three vectors **H**, **B**, and **M**. Usually, vector **M** is expressed as

$$\mathbf{M} = \chi_m \mathbf{H} \qquad (\chi_m \text{ is dimensionless, } \mathbf{M} \text{ in A/m}) \tag{3.17}$$

The dimensionless factor χ_m is known as the *magnetic susceptibility* of the material. We then use Eq. (3.17) and express **B** in terms of **H**:

$$\mathbf{B} = \mu_0(1 + \chi_m)\mathbf{H} = \mu_0 \mu_r \mathbf{H} \qquad (\mu_r \text{ is dimensionless}, \mu_0 \text{ in H/m}) \tag{3.18}$$

The dimensionless factor $\mu_r = (1 + \chi_m)$ is known as the *relative permeability* of the material, and μ as the *permeability* of the material. Materials for which Eq. (3.18) holds are *linear* magnetic materials. If it does not hold, they are *nonlinear*. If at all points of the material μ is the same, the material is said to be *homogeneous*; otherwise, it is *inhomogeneous*.

Linear magnetic materials can be *diamagnetic*, for which $\chi_m < 0$ (i.e., $\mu_r < 1$), or *paramagnetic*, for which $\chi_m > 0$ (i.e., $\mu_r > 1$). For both diamagnetic and paramagnetic materials $\mu_r \cong 1$, differing from unity by less than ± 0.001. Therefore, in almost all applications diamagnetic and paramagnetic materials can be considered to have $\mu = \mu_0$.

Ampere's law in Eq. (3.16) can be transformed into a differential equation, i.e., its differential form, by applying Stokes' theorem of vector analysis:

$$\nabla \times \mathbf{H} = \mathbf{J} \tag{3.19}$$

This differential form of the generalized Ampere's law is valid only for time-invariant currents and magnetic fields.

3.2.10. Macroscopic Currents Equivalent to a Magnetized Material

The macroscopic currents *in vacuum* equivalent to a magnetized material can be both volume and surface currents. The volume density of these currents is given by

$$\mathbf{J}_m = \nabla \times \mathbf{M} \qquad (\text{A/m}^2) \tag{3.20}$$

This has a practical implication as follows. In case of a linear, homogeneous material of magnetic susceptibility χ_m, with no macroscopic currents in it,

$$\mathbf{J}_m = \nabla \times \mathbf{M} = \nabla \times (\chi_m \mathbf{H}) = \chi_m \nabla \times \mathbf{H} = \mathbf{0} \tag{3.21}$$

since $\nabla \times \mathbf{H} = \mathbf{0}$ if $\mathbf{J} = \mathbf{0}$, as assumed. Consequently, in a linear and homogeneous magnetized material with no macroscopic currents there is no volume distribution of equivalent currents. This conclusion is relevant for determining magnetic fields of magnetized materials, where the entire material can be replaced by equivalent surface currents. For example, the problem of a magnetized cylinder reduces to solving the simple case of a solenoid (coil).

3.2.11. Boundary Conditions

Quite often it is necessary to solve magnetic problems involving inhomogeneous magnetic materials that include boundaries. To be able to do this it is necessary to know the relations that must be satisfied by various magnetic quantities at two close points on the two sides of a boundary surface. Such relations are called *boundary conditions*. The two most important boundary conditions are those for the tangential components of **H** and the normal components of **B**. Assuming that there are no macroscopic surface currents on the boundary surface, from the generalized form of Ampere's law it follows that the tangential components of **H** are equal:

$$\mathbf{H}_{1\mathrm{tang}} = \mathbf{H}_{2\mathrm{tang}} \tag{3.22}$$

The condition for the normal components of **B** follows from the law of conservation of magnetic flux, Eq. (3.8), and has the form

$$\mathbf{B}_{1\mathrm{norm}} = \mathbf{B}_{2\mathrm{norm}} \tag{3.23}$$

The boundary conditions in Eqs. (3.22) and (3.23) are valid for *any* media—linear or nonlinear. If the two media are linear, characterized by permeabilities μ_1 and μ_2, the two conditions can be also written in the form

$$\frac{\mathbf{B}_{1\mathrm{tang}}}{\mu_1} = \frac{\mathbf{B}_{2\mathrm{tang}}}{\mu_2} \tag{3.24}$$

and

$$\mu_1 \mathbf{H}_{1\mathrm{norm}} = \mu_2 \mathbf{H}_{2\mathrm{norm}} \tag{3.25}$$

If two media divided by a boundary surface are linear, the lines of vector **B** and **H** refract on the surface according to a simple rule, which follows from the boundary conditions. With reference to Fig. 3.4, this rule is of the form

$$\frac{\tan\alpha_1}{\tan\alpha_2} = \frac{\mu_1}{\mu_2} \tag{3.26}$$

On a boundary between two magnetized materials, Fig. 3.5, the equivalent surface current density is given by

$$\mathbf{J}_{\mathrm{ms}} = \mathbf{n} \times (\mathbf{M}_1 - \mathbf{M}_2) \tag{3.27}$$

Note that the unit vector **n** normal to the boundary surface is directed into medium 1 (Fig. 3.5).

The most interesting practical case of refraction of magnetic field lines is on the boundary surface between air and a medium of high permeability. Let air be medium 1.

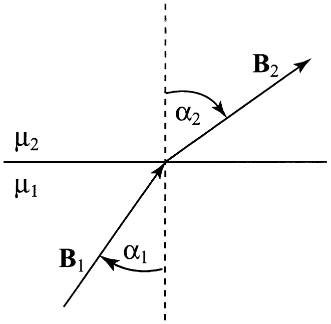


Figure 3.4 Lines of vector **B** or vector **H** refract according to Eq. (3.26).

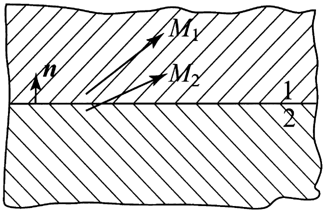


Figure 3.5 Boundary surface between two magnetized materials.

Then the right-hand side of Eq. (3.26) is very small. This means that $\tan\alpha_1$ must also be very small *for any* α_2 (except if $\alpha_2 = \pi/2$, i.e., if the magnetic field lines in the medium of high permeability are tangential to the boundary surface). Since for small angles $\tan\alpha_1 \cong \alpha_1$, the *magnetic field lines in air are practically normal to the surface of high permeability*. This conclusion is very important in the analysis of electrical machines with cores of high permeability, magnetic circuits (such as transformers), etc.

3.2.12. Basic Properties of Magnetic Materials

In the absence of an external magnetic field, atoms and molecules of many materials have no magnetic moment. Such materials are referred to as *diamagnetic materials*. When brought into a magnetic field, a current is induced in each atom and has the effect of reducing the field. (This effect is due to electromagnetic induction, and exists in *all* materials. It is very small in magnitude, and in materials that are not diamagnetic it is dominated by stronger effects.) Since their presence slightly *reduces* the magnetic field, diamagnetics evidently have a permeability slightly *smaller* than μ_0. Examples are water ($\mu_r = 0.9999912$), bismuth ($\mu_r = 0.99984$), and silver ($\mu_r = 0.999975$).

In other materials, atoms and molecules have a magnetic moment, but with no external magnetic field these moments are distributed randomly, and no macroscopic magnetic field results. In one class of such materials, known as *paramagnetics*, the atoms have their magnetic moments, but these moments are oriented statistically. When a field is applied, the Ampere currents of atoms align themselves with the field to some extent. This alignment is opposed by the thermal motion of the atoms, so it increases as the temperature decreases and as the applied magnetic field becomes stronger. The result of the alignment of the Ampere currents is a very small magnetic field added to the external field.

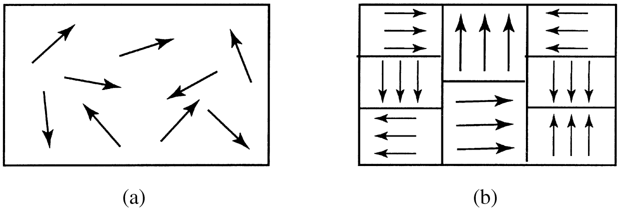


Figure 3.6 Schematic of an unmagnetized (a) paramagnetic and (b) ferromagnetic materials. The arrows show qualitatively atomic (or molecular) magnetic moments.

For paramagnetic materials, therefore, μ is slightly greater than μ_0, and μ_r is slightly greater than one. Examples are air ($\mu_r = 1.00000036$) and aluminum ($\mu_r = 1.000021$).

The most important magnetic materials in electrical engineering are known as *ferromagnetics*. They are, in fact, paramagnetic materials, but with very strong interactions between atoms (or molecules). As a result of these interactions, groups of atoms (10^{12} to 10^{15} atoms in a group) form inside the material, and in these groups the magnetic moments of all the molecules are oriented in the same direction. These groups of molecules are called *Weiss domains*. Each domain is, in fact, a small saturated magnet. A sketch of atomic (or molecular) magnetic moments in paramagnetic and ferromagnetic materials is given in Fig. 3.6.

The size of a domain varies from material to material. In iron, for example, under normal conditions, the linear dimensions of the domains are 10μm. In some cases they can get as large as a few millimeters or even a few centimeters across. If a piece of a highly polished ferromagnetic material is covered with fine ferromagnetic powder, it is possible to see the outlines of the domains under a microscope. The boundary between two domains is not abrupt, and it is called a *Bloch wall*. This is a region 10^{-8}–10^{-6}μm in width (500 to 5000 interatomic distances), in which the orientation of the atomic (or molecular) magnetic moments changes gradually.

Above a certain temperature, the *Curie temperature*, the thermal vibrations completely prevent the parallel alignment of the atomic (or molecular) magnetic moments, and ferromagnetic materials become paramagnetic. For example, the Curie temperature of iron is 770°C (for comparison, the melting temperature of iron is 1530°C).

In materials referred to as *antiferromagnetics*, the magnetic moments of adjacent molecules are antiparallel, so that the net magnetic moment is zero. (Examples are FeO, $CuCl_2$ and FeF_2, which are not widely used.) *Ferrites* are a class of antiferromagnetics very widely used at radio frequencies. They also have antiparallel moments, but, because of their asymmetrical structure, the net magnetic moment is not zero, and the Weiss domains exist. Ferrites are weaker magnets than ferromagnetics, but they have high electrical resistivities, which makes them important for high-frequency applications. Figure 3.7 shows a schematic comparison of the Weiss domains for ferromagnetic, antiferromagnetic and ferrite materials.

Ferromagnetic materials are nonlinear, i.e., $\mathbf{B} \neq \mu\mathbf{H}$. How does a ferromagnetic material behave when placed in an external magnetic field? As the external magnetic field is increased from zero, the domains that are approximately aligned with the field increase in size. Up to a certain (not large) field magnitude, this process is reversible—if the field is turned off, the domains go back to their initial states. Above a certain field strength, the domains start rotating under the influence of magnetic forces, and this process is irreversible. The domains will keep rotating until they are all aligned with the local

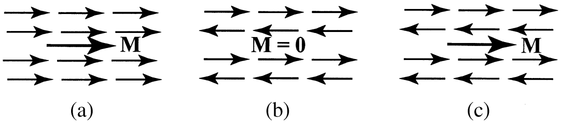


Figure 3.7 Schematic of Weiss domains for (a) ferromagnetic, (b) antiferromagnetic, and (c) ferrite materials. The arrows represent atomic (or molecular) magnetic moments.

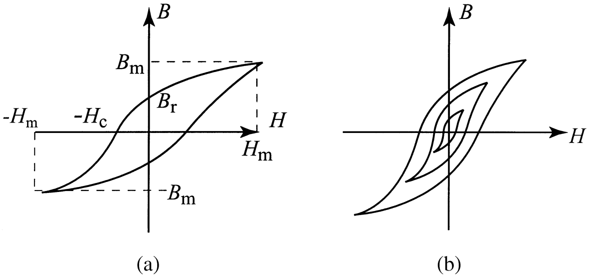


Figure 3.8 (a) Typical hysteresis loop for a ferromagnetic material. (b) The hysteresis loops for external fields of different magnitudes have different shapes. The curved line connecting the tips of these loops is known as the *normal* magnetization curve.

magnetic flux density vector. At this point, the ferromagnetic is *saturated*, and applying a stronger magnetic field does not increase the magnetization vector.

When the domains rotate, there is friction between them, and this gives rise to some essential properties of ferromagnetics. If the field is turned off, the domains cannot rotate back to their original positions, since they cannot overcome this friction. This means that some *permanent* magnetization is retained in the ferromagnetic material. The second consequence of friction between domains is loss to thermal energy (heat), and the third consequence is *hysteresis*, which is a word for a specific nonlinear behavior of the material. This is described by curves $B(H)$, usually measured on toroidal samples of the material. These curves are closed curves around the origin, and they are called *hysteresis loops*, Fig. 3.8a. The hysteresis loops for external fields of different magnitudes have different shapes, Fig. 3.8b.

In electrical engineering applications, the external magnetic field is in many cases approximately sinusoidally varying in time. It needs to pass through several periods until the $B(H)$ curve stabilizes. The shape of the hysteresis loop depends on the frequency of the field, as well as its strength. For small field strengths, it looks like an ellipse. It turns out that the ellipse approximation of the hysteresis loop is equivalent to a *complex permeability*. For sinusoidal time variation of the field, in complex notation we can write $\underline{\mathbf{B}} = \underline{\mu}\underline{\mathbf{H}} = (\mu' - j\mu'')\underline{\mathbf{H}}$, where underlined symbols stand for complex quantities. (This is analogous to writing that a complex voltage equals the product of complex impedance and complex current.) This approximation does not take saturation into account. It can be shown that the imaginary part, μ'', of the complex permeability describes ferromagnetic material hysteresis losses that are proportional to frequency (see chapter on electromagnetic induction). In ferrites, which are sometimes referred to as *ceramic ferromagnetic materials*, the dielectric losses, proportional to f^2, exist in addition (and may even be dominant).

Table 3.1 Magnetic Properties of Some Commonly Used Materials

Material	Relative permeability, μ_r	Comment
Silver	0.9999976	Diamagnetic
Copper	0.99999	Diamagnetic
Gold	0.99996	Diamagnetic
Water	0.9999901	Diamagnetic
Aluminum	1.000021	Paramagnetic
Moly permalloy	100 (few)	Ferromagnetic with air
Ferrite	1000	For example, $\mathrm{NiO} \cdot \mathrm{Fe_2O_3}$, insulator
Nickel	600	Ferromagnetic
Steel	2000	Ferromagnetic
Iron (0.2 impurity)	5000	Ferromagnetic
Purified iron (0.05 impurity)	2×10^5	Ferromagnetic
Supermalloy	As high as 10^6	Ferromagnetic

The ratio B/H (corresponding to the permeability of linear magnetic materials) for ferromagnetic materials is not a constant. It is possible to define several *permeabilities*, e.g., the one corresponding to the initial, reversible segment of the magnetization curve. This permeability is known as the *initial* permeability. The range is very large, from about 500 μ_0 for iron to several hundreds of thousands μ_0 for some alloys.

The ratio B/H along the normal magnetization curve (Fig. 3.8b) is known as the *normal* permeability. If we magnetize a material with a dc field, and then add to this field a small sinusoidal field, a resulting small hysteresis loop will have a certain ratio $\Delta B/\Delta H$. This ratio is known as the *differential* permeability. Table 3.1 shows some values of permeability for commonly used materials.

3.2.13. Magnetic Circuits

Perhaps the most frequent and important practical applications of ferromagnetic materials involve cores for transformers, motors, generators, relays, etc. The cores have different shapes, they may have air gaps, and they are magnetized by a current flowing through a coil wound around a part of the core. These problems are hard to solve strictly, but the approximate analysis is accurate enough and easy, because it resembles dc circuit analysis.

We will restrict our attention to thin linear magnetic circuits, i.e., to circuits with thickness much smaller than their length, as in Fig. 3.9, characterized approximately by a convenient permeability (e.g., initial permeability), assumed to be independent of the magnetic field intensity. The magnetic flux in the circuit is determined from the equations.

Ampere's law applied to a contour that follows the center of the magnetic core in Fig. 3.9 can be written as

$$\oint_{\mathbf{C}} \mathbf{H} \cdot d\mathbf{l} = H_1 L_1 + H_2 L_2 = NI \tag{3.28}$$

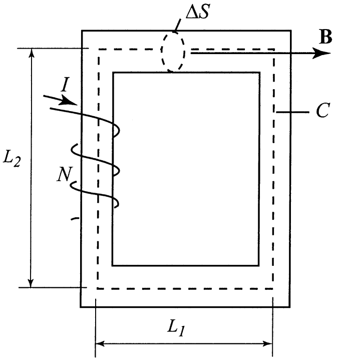


Figure 3.9 A thin magnetic circuit. L_1 and L_2 are lengths of the core sides along contour C through the center of the magnetic core of cross-section ΔS.

where

$$H_i = \frac{B_i}{\mu_i} = \frac{1}{\mu_i}\frac{\Phi_i}{S_i} \tag{3.29}$$

is the magnetic field intensity in each section of the core, assuming a linear magnetic material or a small-signal (dynamic) permeability. An additional equation is obtained for the magnetic fluxes Φ_i at the "nodes" of the magnetic circuit, recalling that

$$\oint_{S_0} \mathbf{B} \cdot d\mathbf{S} = \sum\nolimits_i \Phi_i = 0 \tag{3.30}$$

for any closed surface S_0. Equations (3.28)–(3.30) can be combined to have the same form as the analogous Kirchoff's laws for electrical circuits:

$$\sum_i \Phi_i = 0 \qquad \text{for any node}$$

which is analogous to

$$\sum_i I_i = 0 \tag{3.31}$$

$$\sum_i R_{mi}\Phi_i - \sum_i N_i I_i = 0 \qquad \text{for any closed loop}$$

analogous to

$$\sum_i R_i I_i - \sum_i V_i = 0 \tag{3.32}$$

$$R_{mi} = \frac{1}{\mu_i}\frac{L_i}{S_i} \qquad \text{for any branch}$$

analogous to

$$R_i = \frac{1}{\sigma_i} \frac{L_i}{S_i} \tag{3.33}$$

where R_m is referred to as *magnetic resistance*, and σ is the electrical conductance. The last equation is Ohm's law for uniform linear resistors.

If the magnetic circuit contains a short air gap, L_0 long, the magnetic resistance of the air gap is calculated as in Eq. (3.33), with $\mu_i = \mu_0$.

3.3. APPLICATIONS OF MAGNETOSTATICS

The sections that follow describe briefly some common applications of magnetostatic fields and forces, with the following outline:

1. Forces on charged particles (cathode ray tubes, Hall effect devices)
2. Magnetic fields and forces of currents in wires (straight wire segment, Helmholtz coils)
3. Magnetic fields in structures with some degree of symmetry (toroidal coil, solenoid, coaxial cable, two-wire line, strip-line cable)
4. Properties of magnetic materials (magnetic shielding, magnetic circuits)
5. System-level applications (magnetic storage, Magnetic Resonance Imaging—MRI).

3.3.1. Basic Properties of Magnetic Force on a Charged Particle (the Lorentz Force)

By inspecting the Lorentz force in Eq. (3.2), we come to the following conclusion: The speed of a charged particle (magnitude of its velocity) can be changed by the electric force $Q\mathbf{E}$. It *cannot* be changed by the magnetic force $Q\mathbf{v} \times \mathbf{B}$, because magnetic force is always normal to the direction of velocity. Therefore, charged particles can be accelerated only by electric forces.

The ratio of the maximal magnetic and maximal electric force on a charged particle moving with a velocity $\mathbf{v}$ equals vB/E. In a relatively large domain in vacuum, it is practically impossible to produce a magnetic flux density of magnitude exceeding 1 T, but charged particles, e.g., electrons, can easily be accelerated to velocities on the order of 1000 km/s. To match the magnetic force on such a particle, the electric field strength must be on the order of 10^6V/m, which is possible, but not easy or safe to achieve. Therefore, for example, if we need to substantially deflect an electron beam in a small space, we use magnetic forces, as in television or computer-monitor cathode-ray tubes.

The horizontal component of the earth's magnetic field is oriented along the north–south direction, and the vertical component is oriented downwards on the northern hemisphere and upwards on the southern hemisphere. Therefore, cathode-ray tubes that use magnetic field deflection have to be tuned to take this external field into account. It is likely that your computer monitor (if it is a cathode-ray tube) will not work exactly the same way if you turn it sideways (it might slightly change colors or shift the beam by a couple of millimeters) or if you use it on the other side of the globe.

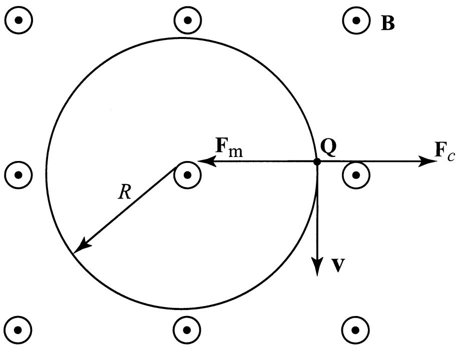


Figure 3.10 Charged particle in a uniform magnetic field.

Charged Particle Moving in a Uniform Magnetic Field

Consider a charged particle $Q > 0$ moving in a magnetic field of flux density **B** with a velocity **v** normal to the lines of vector **B**, Fig. 3.10. Since the magnetic force on the charge is always perpendicular to its velocity, it can only change the direction of the charged particle motion. To find the trajectory of the particle, note that the magnetic force on the particle is directed as indicated, tending to curve the particle trajectory. Since **v** is normal to **B**, the force magnitude is simply QvB. It is opposed by the centrifugal force, mv^2/R, where R is the radius of curvature of the trajectory. Therefore,

$$QvB = \frac{mv^2}{R} \tag{3.34}$$

so that the radius of curvature is constant, $R = mv/QB$. Thus, the particle moves in a circle. It makes a full circle in

$$t = T = \frac{2\pi R}{v} = \frac{2\pi m}{QB} \tag{3.35}$$

seconds, which means that the frequency of rotation of the particle is equal to $f = 1/T = QB/2\pi m$. Note that f does not depend on v. Consequently, all particles that have the same charge and mass make the same number of revolutions per second. This frequency is called the *cyclotron* frequency. Cyclotrons are devices that were used in the past in scientific research for accelerating charged particles. A simplified sketch of a cyclotron is shown in Fig. 3.11, where the main part of the device is a flat metal cylinder, cut along its middle. The two halves of the cylinder are connected to the terminals of an oscillator (source of very fast changing voltage). The whole system is in a uniform magnetic field normal to the bases of the cylinder, and inside the cylinder is highly rarefied air.

A charged particle from source O finds itself in an electric field that exists between the halves of the cylinder, and it accelerates toward the other half of the cylinder. While outside of the space between the two cylinder halves, the charge finds itself only in a magnetic field, and it circles around with a radius of curvature $R = mv/QB$. The time it takes to go around a semicircle does not depend on its velocity. That means that it will

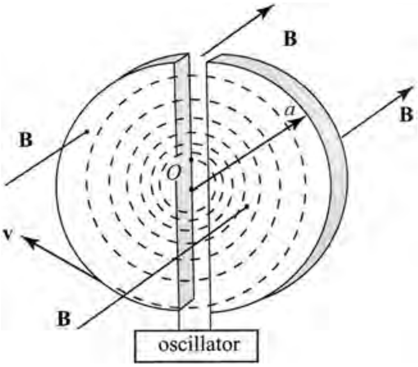


Figure 3.11 Sketch of a cyclotron used in scientific research to accelerate charged particles by means of electric and magnetic fields.

always take the charge the same amount of time to reach the gap between the two cylinders. If the electric field variation in this region is adjusted in such a way that the charge is always accelerated, the charge will circle around in larger and larger circles, with increasingly larger velocity, until it finally shoots out of the cyclotron. The velocity of the charge when it gets out of the cyclotron is $v = QBa/m$. This equation is valid only for velocities not close to the speed of light. For velocities close to the speed of light, the mass is not constant (it is increased due to relativistic effects). As a numerical example, for $B = 1\,\text{T}$, $Q = e$, $a = 0.5\,\text{m}$, $m = 1.672 \times 10^{-27}\,\text{kg}$ (a proton), the particles will be accelerated to velocities $v = 47.9 \times 10^6\,\text{m/s}$. Cyclotrons are not used any more for particle physics research, but they were instrumental in the initial phases of this science. At the University of Chicago, for example, a cyclotron was used for research that led to the development of the atomic bomb.

The Hall Effect

In 1879, Edwin Hall discovered an effect that can be used for measuring the magnetic field and for determining the sign of free charges in conductors. Let a conducting ribbon have a width d and is in a uniform magnetic field of flux density **B** perpendicular to the ribbon, Fig. 3.12. A current of density **J** flows through the ribbon. The free charges can in principle be positive (Fig. 3.12a) or negative (Fig. 3.12b). The charges that form the current are moving in a magnetic field, and therefore there is a magnetic force $\mathbf{F} = Q\mathbf{v} \times \mathbf{B}$ acting on them. Due to this force, positive charges accumulate on one side of the ribbon, and negative ones on the other side. These accumulated charges produce an electric field E_H. This electric field, in turn, acts on the free charges with a force that is in the opposite direction to the magnetic force. The charges will stop accumulating when the electric force is equal in magnitude to the magnetic force acting on each of the charges. Therefore, in steady state

$$QvB = QE_H \qquad \text{or} \qquad E_H = vB \tag{3.36}$$

Between the left and right edge of the ribbon, one can measure a voltage equal to

$$|V_{12}| = E_H d = vBd \tag{3.37}$$

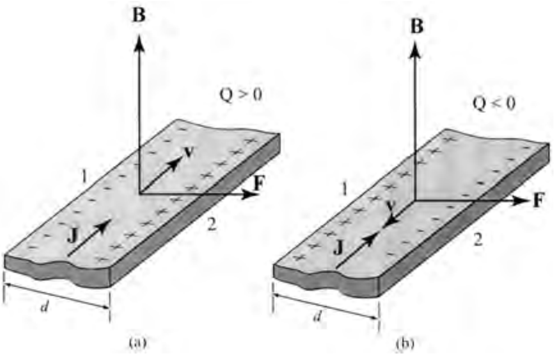


Figure 3.12 The Hall effect in case of (a) positive free-charge carriers, and (b) negative free-charge carriers.

In the case shown in Fig. 3.12a, this voltage is negative, and in Fig. 3.12b it is positive. Thus, the sign of the voltage corresponds to the sign of free charge carriers and can be determined by a voltmeter.

Since $J = NQv$, where N is the number of free charges per unit volume, the Hall voltage becomes

$$|V_{12}| = \frac{Jd}{NQ}B \tag{3.38}$$

Thus, if the coefficient Jd/NQ is determined for a ribbon such as the one sketched in Fig. 3.12, by measuring V_{12}, the magnetic flux density B can be determined. Usually, Jd/NQ is determined experimentally. This ribbon has four terminals: two for the connection to a source producing current in the ribbon, and two for the measurement of voltage across it. Such a ribbon is called a *Hall element.*

For single valence metals, e.g., copper, if we assume that there is one free electron per atom, the charge concentration is given by

$$N = \frac{N_A \rho_m}{M} \tag{3.39}$$

where N_A is Avogadro's number (6.02×10^{23}atoms/mol), ρ_m is the mass density of the metal, and M is the atomic mass.

As a result of the above properties, Hall elements are key components in devices used for a wide range of measurements:

The Hall effect is most pronounced in semiconductors. Hall-effect devices are commonly used to determine the type and concentration of free carriers of semiconductor samples, as can be deduced from Eqs. (3.38) and (3.39).

Gaussmeters (often called *teslameters*) use a Hall element to measure magnetic flux density, by generating output voltage proportional to the magnetic field. Special attention is given to the design of the accompanying Hall-effect probes. The accuracy and calibration of Hall-effect Gaussmeters is verified by standardized reference magnets.

In integrated circuits technology, the Hall effect is used for sensors and switches. In sensors, the magnetic flux density through the Hall element determines the output voltage; in switches, it determines the switching state. Hall-effect sensor operation is robust with respect to environmental conditions.

Linear Hall sensors, which generate voltage proportional to the magnetic flux perpendicular to the Hall plate, are characterized by output quiescent voltage (the output voltage in absence of the magnetic field) and sensitivity. Their industrial applications include measurement of angle, current, position and distance, and pressure, force, and torque sensors. In automotive industry, they are used for active suspension control, headlight range adjustment, liquid level sensors, power steering, and so on. With very low energy consumption (a fraction of a mW), linear Hall sensors are more efficient and cost effective than most inductive and optoelectronic sensors.

A Hall switch contains an integrated comparator with predefined switching levels and an open-drain transistor at its digital output, which can be adapted to different logic systems. The output characteristic of a Hall switch resembles a hysteresis-like (B, V_{out}) curve. The magnetic flux density B of the hysteresis ranges from B_{off} to B_{on}; if $B > B_{on}$, the output transistor is switched on, and if $B < B_{off}$, the transistor is switched off. These switches are also available in a differential form, where the output transistor is switched according to the difference of the magnetic flux between two Hall-element plates separated typically by several millimeters. Finally, in the case of two-wire Hall switches, the output signal of the switch is a current of an internal source, which is switched on or off by the magnetic field applied to the Hall plate. In all Hall switches, simplified switching ensures a clean, fast, and bounceless switch avoiding the problems present in mechanical contact switches. Hall-effect switches are more cost effective than most electromechanical switches. Among other applications, they are widely used for commutation of brushless DC motors, wheel speed sensors, measurement of rotations per minute, pressure switches, position-dependent switches, etc. The automotive industry uses Hall switches, e.g., in ignition and wiper systems, door locks, window raising controls, and retraction-roof controls and for break light switches. In the computer industry, this type of switch is used in keyboards.

3.3.2. Magnetic Fields of Currents in Wires

Biot–Savart's law in Eq. (3.3) can be used to calculate vector **B** produced by currents in wire loops of arbitrary shapes (i.e., a variety of electrical circuits). Such loops are often made of (or can be approximated by) a sequence of interconnected straight wire segments. Evaluation of **B** in such cases can greatly be simplified if we determine vector **B** produced by the current in a single straight wire segment. With reference to Fig. 3.13, using Biot–Savart's law, the following expression is obtained

$$B = \frac{\mu_0 I}{4\pi a}(\sin\theta_2 - \sin\theta_1) \tag{3.40}$$

Helmholtz Coils

To obtain in a simple manner highly uniform magnetic field in a relatively large domain of space in air, *Helmholtz coils* can be used. They consist of two thin, parallel, coaxial circular loops of radius *a* that are a distance *a* apart, Fig. 3.14a. Each loop carries a current I,

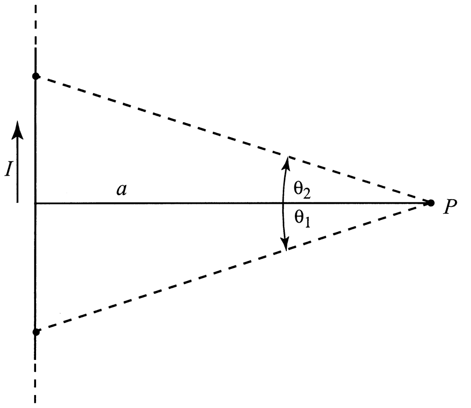


Figure 3.13 Calculating the magnetic field at point *P* due to a straight wire segment with current *I*.

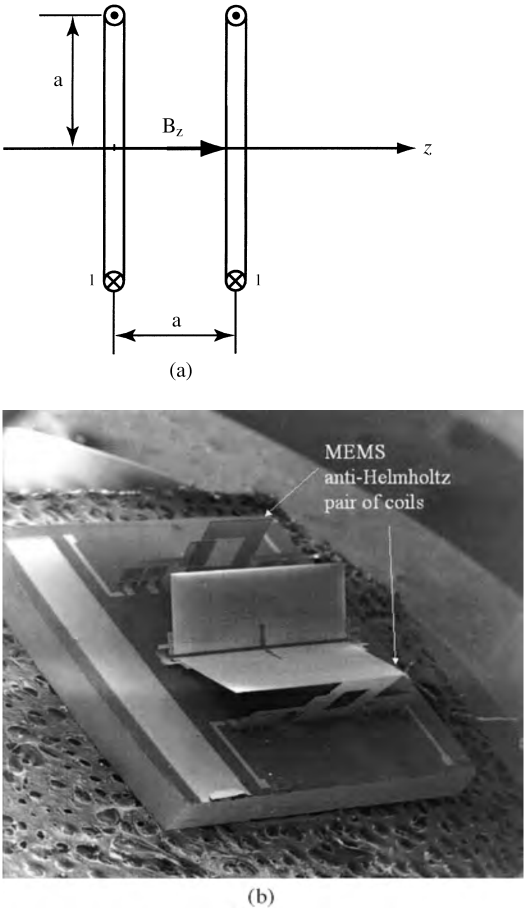


Figure 3.14 (a) Sketch of a Helmholtz pair of coils. The magnetic field in the center is highly uniform. (b) Photograph of a micro-electromachined anti-Helmholtz coils (courtesy Profs. Victor Bright and Dana Anderson, University of Colorado at Boulder). The inductors are released to spring up into position.

and the two currents are in the same direction. Starting from the Biot–Savart law, we find that the magnetic field at a distance z from the center of one loop, on their common axis, is axial, of magnitude

$$B_z(z) = \frac{\mu_0 I a^2}{2} \frac{1}{\left[a^2 + (a - z)^2\right]^{3/2}} \tag{3.41}$$

It can be proven that, at $z = a/2$, the first, second, and even the third derivatives of $B_z(z)$ are zero, which means that the magnetic flux density around that point is highly uniform.

If the currents in Helmholtz coils are flowing in opposite directions, the magnetic field has a null in the center, accompanied by a very large gradient. An interesting application of this anti-Helmholtz pair of coils is in the emerging field of atomic optics, where large gradients of the magnetic field are used to guide atoms and even Bose-Einstein condensates. A photograph of a micro-electromachined (MEM) anti-Helmholtz pair is shown in Fig. 3.14b.

Magnetic Force Between Two Long Parallel Wires: a Definition of the Ampere

Two parallel current-carrying wires can either attract or repel each other, depending on the direction of the currents in them. If the wires are in vacuum (air) and are very long (theoretically infinitely long), if currents in them are of equal magnitude I and the distance between them is d, the force per unit length of the wires is

$$F_{\text{per unit length}} = \mu_0 \frac{I^2}{2\pi d} \tag{3.42}$$

To get a feeling for the magnetic forces between current-carrying conductors, from this equation we find that, for $d = 1$ m and $I = 1$ A, the force on each of the wires is 2×10^{-7} N/m. This used to be one of the definitions of the unit for electrical current, the *ampere*.

Magnetic Force on the Short Circuit of a Two-Wire Line

As another example, the magnetic force on the segment A-A' of the two-wire-line short circuit shown in Fig. 3.15a is given by

$$F = \frac{\mu_0 I^2}{2\pi} \ln \frac{d - a}{a} \tag{3.43}$$

If a large current surge occurs in the line, the force shock on the short circuit can be quite large. For example, if there is a sudden increase of current intensity to $I = 5000$ A, $a = 0.5$ cm, and $d = 50$ cm, the force shock is about 23 N, which may be sufficient to open the short circuit if it is not firmly connected.

Magnetic Force in a Loudspeaker

Shown in Fig. 3.15b is a sketch of a permanent magnet used in loudspeakers. The lines of the magnetic flux density vector are radial, and at the position of the coil the magnitude

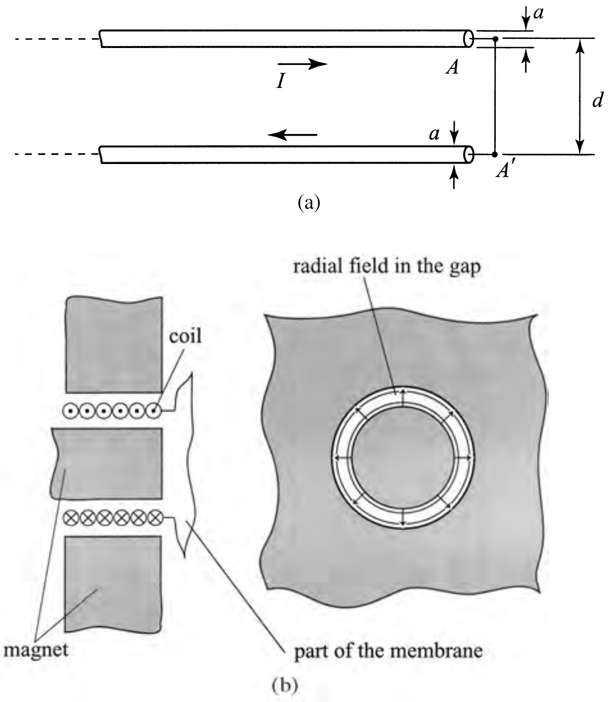


Figure 3.15 (a) A Short-circuited two-wire line. (b) Permanent magnet and coil used in a loudspeaker.

is $B = 1\,\mathrm{T}$. Let the coil have N turns, its radius be a, and the current in the coil be I. The magnetic force on the coil (which is glued to the loudspeaker membrane) is $F = 2\pi aNIB$. If, in particular, $I = 0.15\,\mathrm{A}$, $N = 10$, and $a = 0.5\,\mathrm{cm}$, we find that $F = 0.047\,\mathrm{N}$.

3.3.3. Applications of Ampere's Law

Ampere's law can be used to determine the magnetic field produced by currents in structures with a high level of symmetry. Common and practical examples are discussed below.

Magnetic Field of a Straight Wire

Consider a *straight*, very long (theoretically infinite) wire of circular cross section of radius a, Fig. 3.16a. (A wire may be considered infinitely long if it is much longer than the shortest distance from it to the observation point.) There is a current of intensity I in the wire distributed uniformly over its cross section. Note that, due to symmetry, both outside and inside the wire the lines of vectors **B** and **H** are circles centered along the wire axis and in planes normal to it. Therefore, the only unknown is the magnitude of these vectors as a function of the distance r from the wire axis. Using Ampere's law we find that

$$B(r) = \frac{\mu_0 I}{2\pi r} \qquad \text{for } r \geq a \tag{3.44}$$

As long as the point is outside the wire, the radius of the wire a is irrelevant. This expression for B outside a round wire is valid for a wire of any radius, including an infinitely thin one. Inside the wire, the magnetic flux density is given by

$$B(r) = \frac{\mu_0 I r}{2\pi a^2} \qquad \text{for } r \leq a \tag{3.45}$$

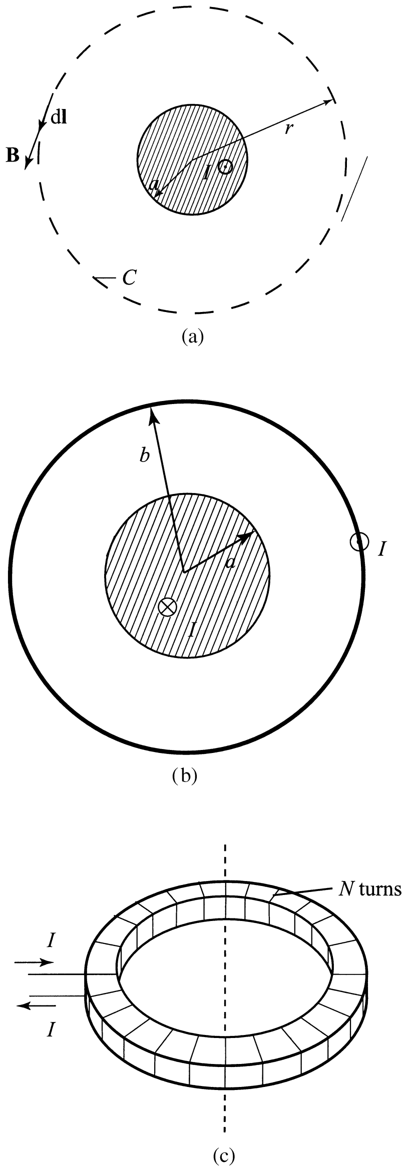
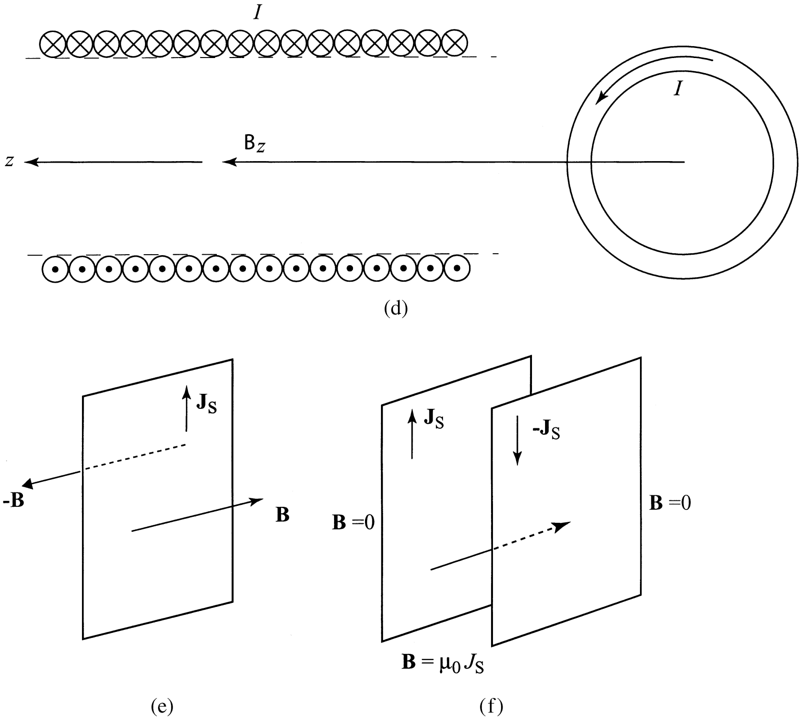


Figure 3.16 (a) Cross section of straight wire of circular cross section with a current of intensity I. (b) The cross section of a coaxial cable with very thin outer conductor. (c) A toroidal coil with N windings. (d) Longitudinal and transverse cross sections of a solenoid. (e) A current sheet. (f) Two parallel current sheets.



(d)

(e)

(f)

Figure 3.16 Continued.

Magnetic Field in a Coaxial Cable

Using the same procedure we find the magnetic flux density due to currents I and -I in conductors of a coaxial cable, Fig. 3.16b. If the outer conductor is assumed to be very thin (as it usually is), outside the cable the magnetic field does not exist, and inside the cable, the magnetic flux density is radial and equal to

$$B(r) = \begin{cases} \dfrac{\mu_0 I r}{2\pi a^2} & \text{for } r \le a \\ \dfrac{\mu_0 I}{2\pi r} & \text{for } r \ge a \end{cases} \tag{3.46}$$

Magnetic Field of a Toroidal Coil

Another commonly used case in inductors and transformers is that of a toroidal coil, Fig. 3.16c. The cross section of the toroid is arbitrary. Assume that the coil is made of N uniformly and densely wound turns with current of intensity I. From the Biot–Savart law, we know that the lines of vector **B** are circles centered on the toroid axis. Also, the magnitude of **B** depends only on the distance r from the axis. Applying Ampere's law yields the following expressions for the magnitude, $B(r)$, of the magnetic flux density vector:

$$B = \begin{cases} 0 & \text{outside the toroid} \\ \dfrac{\mu_0 N I}{2\pi r} & \text{inside the toroid} \end{cases} \tag{3.47}$$

As a numerical example, for $N = 1000$, $I = 2\,\mathrm{A}$, and a mean toroid radius of $r = 10\,\mathrm{cm}$, we obtain $B = 4\,\mathrm{mT}$. This value can be larger if, for example, several layers of wire are wound one on top of each other, so that N is larger. Alternatively, the torus core can be made of a ferromagnetic material, resulting in much larger magnitude of the magnetic flux density inside the core.

Magnetic Field of a Long Solenoid

Assume that the radius r of the toroid becomes very large. Then, at any point inside the toroid, the toroid looks locally as if it were a cylindrical coil, known as a *solenoid* (from a Greek word which, roughly, means "tubelike"), Fig. 3.16d. We conclude that outside an "infinitely long" solenoid the flux density vector is zero. Inside, it is given by Eq. (3.47) with r very large, or since $N' = N/2\pi r$ is the number of turns per unit length of the toroid, i.e., of the solenoid,

$$B = \mu_0 N' I \qquad \text{inside the solenoid (coil)} \tag{3.48}$$

The field inside a very long solenoid is *uniform*, and the expression is valid for *any* cross section of the solenoid. As a numerical example, $N' = 2000$ windings/m and $I = 2\,\mathrm{A}$ result in $B \cong 5\,\mathrm{mT}$.

Magnetic Field of a Planar Current Sheet and Two Parallel Sheets

Consider a large conducting sheet with constant surface current density $\mathbf{J}_s$ at all points, Fig. 3.16e. From the Biot–Savart law, vector $\mathbf{B}$ is parallel to the sheet and perpendicular to vector $\mathbf{J}_s$, and $\mathbf{B}$ is directed in opposite directions on the two sides of the sheet, as indicated in the figure. Applying Ampere's law gives

$$B = \mu_0 \frac{J_S}{2} \qquad \text{for a current sheet} \tag{3.49}$$

For two parallel current sheets with opposite surface currents of the same magnitude (Fig. 3.16f), from the last equation and using superposition we find that the magnetic field outside the sheets is zero, and between two parallel current sheets

$$B = \mu_0 J_S \tag{3.50}$$

Magnetic Field of a Stripline

Equation (3.50) is approximately true if the sheets are not of infinite width, and are close to each other. Such a system is called a *strip line*. Assume that the strip line is filled with a ferrite of relative permeability μ_r. Since $a \gg g$, where a is the finite strip width and g is the distance (gap) between two infinitely long strips, the magnetic field outside the strips can be neglected, and the resulting magnetic flux density inside the strip line is $B = \mu_r \mu_0 I/a$. The magnitude of the magnetization vector in the ferrite is $M = (\mu_r - 1)I/a$. The density of equivalent surface magnetization currents is thus $J_{ms} = (\mu_r - 1)I/a$. These currents have the same direction as the conduction currents in the strips, but are many times greater than the surface current over the strips.

3.3.4. Magnetic Shielding; Magnetic Materials for EMC Testing

Imagine two cavities (air gaps) inside an uniformly magnetized material of relative permeability μ_r. One is a needlelike cavity in the direction of the vector **B**. The other is a thin-disk cavity, normal to that vector. According to boundary conditions, the ratio of magnitudes of the magnetic flux density vectors in the two cavities and that in the surrounding material is equal to $1/\mu_r$ and 1, respectively. We can therefore conclude that the theoretical possibility of reducing the external time-invariant magnetic field by means of "magnetic shielding" is by a factor of $1/\mu_r$. Note, however, that the shielding effect of *conductive* ferromagnetic materials is greatly increased for time-varying fields, due to the skin effect (see chapter on electromagnetic induction). Note that, for EMC/EMI (electromagnetic compatibility and electromagnetic interference) testing, ferrite anechoic chambers are used. These rely on magnetic losses inside ferrite materials and will briefly be discussed in Chapter 4.

3.3.5. Measurements of Basic Properties of Magnetic Materials

The curve $B(H)$ that describes the nonlinear material is usually obtained by measurement. The way this is done is sketched in Fig. 3.17. A thin toroidal core of mean radius R, made of the material we want to measure, has N tightly wound turns of wire, and a cross-sectional area S. If there is a current I through the winding, the magnetic field intensity inside the core is given by

$$H = \frac{NI}{2\pi R} \tag{3.51}$$

From this formula, the magnetic field magnitude for any given current can be calculated. Around the toroidal core there is a second winding, connected to a ballistic galvanometer (an instrument that measures the charge that passes through a circuit). It can be shown that the charge that flows through the circuit is proportional to the change of the magnetic flux, $\Delta Q \propto \Delta\Phi = S\,\Delta B$, and therefore to the change of the B field as well. By changing the current I through the first winding, the curve $B(H)$ can be measured point by point. If the field H is changing slowly during this process, the measured curves are called *static magnetization curves*.

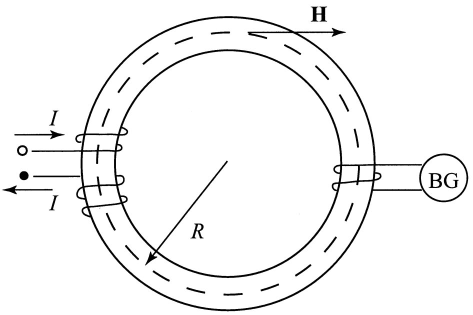


Figure 3.17 Sketch of setup for measurement of magnetization curves.

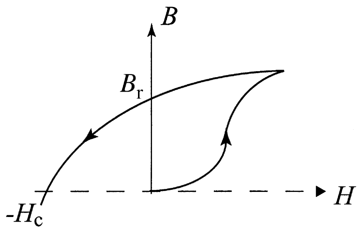


Figure 3.18 Magnetization and demagnetization curves.

Permanent Magnets

If a ferromagnetic body is magnetized (desirably to saturation) and the external magnetic field is switched off, the body remains magnetized, i.e., becomes a permanent magnet. If the body is a thin toroid and the magnetic field is produced by a uniformly wound coil, the magnetization curve is as in Fig. 3.18. When the current after saturation of the toroid is switched off, the operating point moves to the point labeled B_r, known as the *remanent* flux density. If the current is reversed (changes sign), the point moves along the curve to the left of the B axis, sometimes referred to as the demagnetization curve. The magnetic field intensity H_c corresponding to zero B is known as the *coercive* magnetic field. If we cut a piece out of the magnetized toroid with remanent flux density inside it, the operating point will move along the demagnetization curve. A magnetic field will exist in the air gap, and a permanent toroidal magnet is obtained.

Permanent magnets are used in a large variety of applications, one of the most common being data storage on magnetic disks and tapes, in the form of small magnetized patches on thin magnetic films deposited on plastic substrate.

3.3.6. Magnetic Storage

Magnetic materials have been used for storing data since the very first computers. In the 1970s, magnetic core memories were used and an example is shown in Fig. 3.19. The principle of operation of magnetic core memories is an excellent illustration of both magnetostatics and electromagnetic induction. Furthermore, it appears that these components might see a revival for space applications due to the radiation hardness of magnetic materials. In Fig. 3.19a, one bit of the memory is a small ferromagnetic torus, with two wires passing through it. One of the wires is connected to a circuit used for both writing and reading, and the second wire is used only for reading. To write a "1" or a "0", a positive or negative current pulse, respectively, is passed through the first wire. This results in the core magnetized to either B_r or $-B_r$ on the hysteresis curve, respectively. Elements of an entire memory are arranged in matrices, with two wires passing through each torus, Fig. 3.19b. The current passing through each row or column is half of the current needed to saturate the torus, so both the row and the column of the specific bit need to be addressed. The readout process requires electromagnetic induction and will be described in the next chapter.

A common magnetic storage device used today is the hard disk drive in every computer. Information is written to the disk by magnetizing a small piece of the disk surface. As technology is improving, the amount of information that can be stored on a

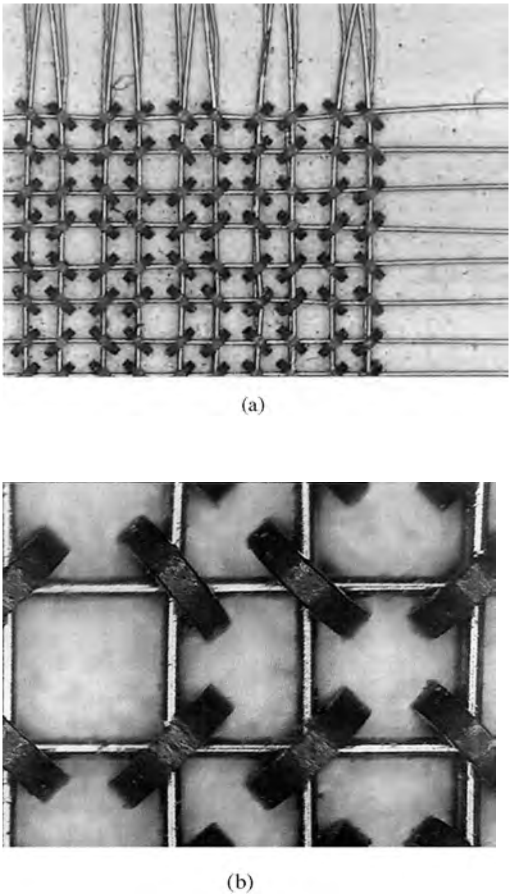


Figure 3.19 (a) A portion of a magnetic core memory matrix and (b) a close-up showing individual memory elements, where each torus represents 1 bit. The wire radius is 35 μm and the core diameter 420 μm.

standard-size hard disk is rapidly growing [6,14]. In 2002, drives with more than 20 Gbytes were readily available in personal computers, while in 1995, a few hundred mega bytes were standard. The development is in the direction of increasing disk capacity and increasing speed (or reducing access time). These two requirements compete with each other, and the engineering solution, as is usually the case, needs to be a compromise.

We will now describe in a simple manner how data are written on the disk. The hard disk itself is coated with a thin coating of ferromagnetic material such as Fe_2O_3. The disk is organized in sectors and tracks, as shown in Fig. 3.20a.

The device that writes data to the disk (and reads data from it) is called a magnetic head. Magnetic heads are made in many different shapes, but all operate essentially according to the same principle. We here describe the operation of the writing process on the example of a simplified head which is easy to understand and is shown in Fig. 3.21. The head is a magnetic circuit with a gap. The gap is in close proximity to the tracks, so there is some leakage flux between the head and the ferromagnetic track.

In the “write” process, a current flows through the windings of the magnetic head, thus creating a fringing magnetic field in the gap. The gap is as small as 5 μm. As the head

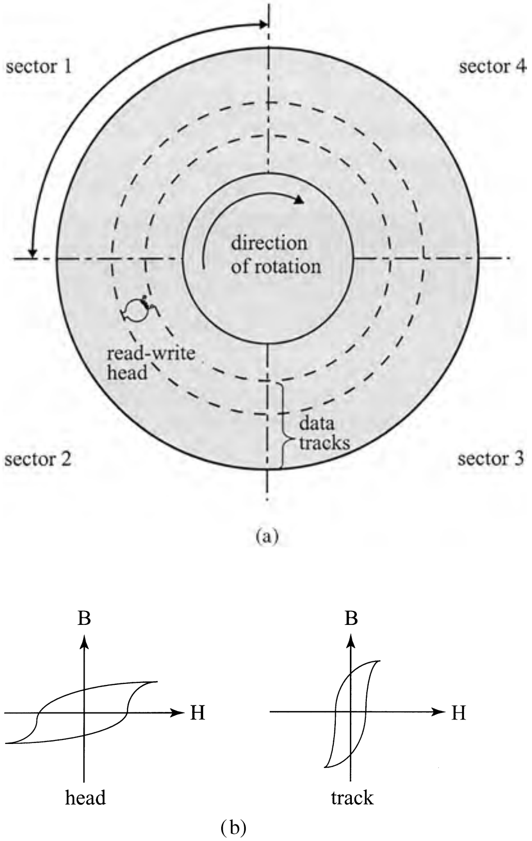


Figure 3.20 (a) Hard disk tracks. (b) Sketch of qualitative shapes of hysteresis curves required for the head and track magnetic materials.

moves along the track (usually the track rotates), the fringing field magnetizes a small part of the track, creating a south and a north pole in the direction of rotation. These small magnets are about 5 μm long and 25 μm wide. A critical design parameter is the height of the head above the track: the head should not hit the track, but it also needs to be as close as possible to maximize the leakage flux that magnetizes the track. Typically, the surface of the track is flat to within several micrometers, and the head follows the surface profile at a distance above it of about 1 μm or less. This is possible because the head aerodynamically flies above the disk surface. The current in the head windings should be strong enough to saturate the ferromagnetic track. If the track is saturated and the remanent flux density of the track material is large, the voltage signal during readout is maximized. The requirements on material characteristics for the head and tracks are completely opposite: the head needs to have a low value of the remanent magnetic flux density, since during readout any remaining B essentially represents noise. In contrast, the track material needs to stay magnetized as long as possible with as high a B as possible. A sketch of the relative desired hysterisis curves is shown in Fig. 3.20b. The principle of readout is an excellent example of electromagnetic induction and is described briefly in the next chapter.

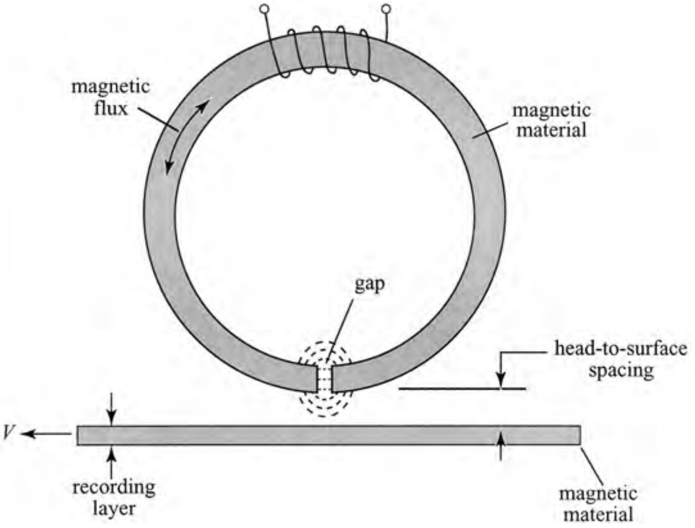


Figure 3.21 The magnetic head aerodynamically flies over the disk surface at a distance above it of only about 1μm while following the surface profile. In the figure, the surface profile is shown as ideally flat, which in practice is not the case.

3.3.7. Magnetic Circuits

Consider a thin toroidal coil of length L, area of cross section S, and with N turns. Assume that the permeability of the core is μ and that a current I is flowing through the coil. Using Eqs. (3.32) and (3.33), the following is obtained

$$\Phi = \frac{NI}{R_m} = \frac{NI}{(L/\mu S)} = \mu \frac{N}{L} IS = \mu N' IS \tag{3.52}$$

This is the same result as that obtained by determining B for the coil using Ampere's law, and $\Phi = B \cdot S$.

The analysis of arbitrarily complex thin linear magnetic circuits is very simple—it is analogous to the analysis of dc electrical circuits. However, real magnetic circuits are neither thin, nor linear. Nevertheless, thin linear magnetic circuits can be used as the basis for approximate analysis of actual magnetic circuits.

Consider a thick, U-shaped core of permeability $\mu_1 \gg \mu_0$, closed by a thick bar of permeability $\mu_2 \gg \mu_0$, as shown in Fig. 3.22. N turns with a current I are wound on the core. The exact determination of the magnetic field in such a case is almost impossible. The first thing we can conclude is that since μ_1,$\mu_2 \gg \mu_0$, the tangential component of the magnetic flux density **B** is much larger in the core than in the air outside it. The normal components of **B** are equal, so the magnetic flux density inside the core is generally much larger than outside the core. Therefore, the magnetic flux can be approximately considered to be restricted to the core. This is never exactly true, so this is the first assumption we are making.

Further, if we assume that Eqs. (3.32) and (3.33) are reasonably accurate if lengths L_1 and L_2 are used as average lengths for the two circuit sections (with their actual cross-sectional areas), we can approximately analyze the circuit using thin-circuit theory. It is instructive to show that the error in doing so is acceptable.

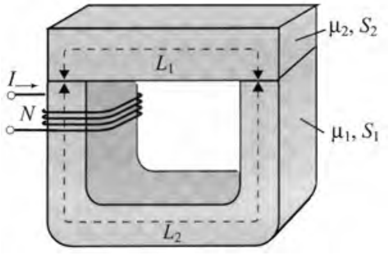


Figure 3.22 A realistic thick magnetic circuit of an inductor.

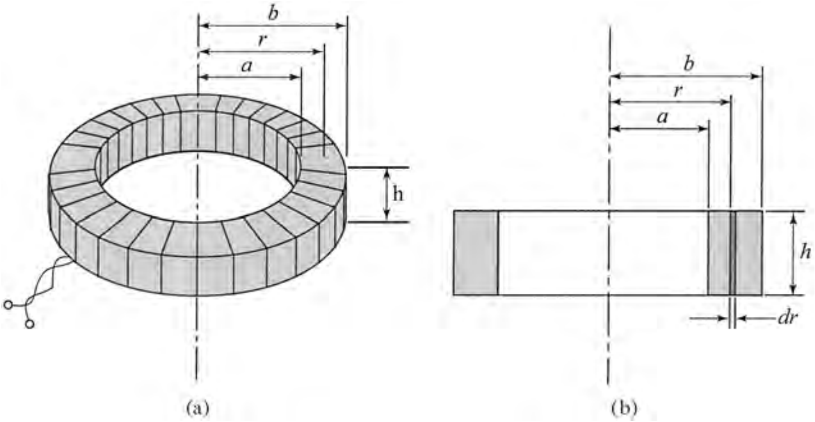


Figure 3.23 (a) A toroidal coil. (b) Cross section of the coil.

A toroidal coil and its cross section are shown in Fig. 3.23. Since the coil has N densely wound turns with a current I, from Ampere's law we find that $B = \mu NI/2\pi r$. The exact value of the magnetic flux through the toroid cross section is

$$\Phi_{\text{exact}} = \frac{\mu NIh}{2\pi}\int_a^b \frac{dr}{r} = \frac{\mu NIh}{2\pi}\ln\frac{b}{a} \tag{3.53}$$

According to Eqs. (3.32) and (3.33), adopting the average length of the toroidal core, the approximate flux is

$$\Phi_{\text{approximate}} = \frac{NI}{R_m} = \frac{NI}{(\pi(a+b))/(\mu(b-a)h)} = \frac{\mu NIh}{2\pi}\frac{2(b-a)}{b+a} \tag{3.54}$$

The relative error is

$$\frac{\Phi_{\text{approximate}} - \Phi_{\text{exact}}}{\Phi_{\text{exact}}} = \frac{2(b/a-1)}{(b/a)\ln(b/a)} - 1 \tag{3.55}$$

which is very small even for quite thick toroids. For example, if $b/a = e = 2.718\ldots$, the error is less than 8%. Therefore, the magnetic flux in the magnetic circuit in Fig. 3.23 can be determined approximately as

$$\Phi \approx \frac{NI}{L_1/(\mu_1 S_1) + L_2/(\mu_2 S_2)} \tag{3.56}$$

If the magnetic material of a circuit cannot be approximated as linear, i.e., there is no equivalent relative permeability, the measured relationship $B(H)$ must be used.

3.3.8. Nuclear Magnetic Resonance (NMR) and Magnetic Resonance Imaging (MRI)

Superconducting loops can carry currents of enormous densities on the order of 1000 A/mm^2 and consequently can be used to make the strongest electromagnets known. Extremely strong superconducting magnets (0.5–30 T) are used in nuclear magnetic resonance (NMR) systems, best known in medical applications as magnetic resonance imaging (MRI). These devices are able to resolve three-dimensional molecule structures. Currently, NMR-based products are used in diverse fields, such as biomedical imaging, human genome research, and pharmaceutical industry [4,5,8,20].

First observed by Felix Bloch and Edward M. Purcell in 1946, the phenomenon which serves as the basis of the NMR technology can be explained as follows [1]. Nuclei of certain common atoms, such as hydrogen and carbon, have a magnetic moment of their own (referred to as *spin*). When in a strong static magnetic field, the atom spins align themselves either against or along the external magnetic field. If, in addition, a radio-frequency magnetic field is applied at exactly the magnetic-field intensity-dependent spin resonant frequency, the spin changes, producing a resonant energy state switching, which results in absorption or emission of energy. Atoms of different elements have different resonance frequencies at which the spin change occurs in the presence of a magnetic field of specified strength. This "signature frequency" allows researchers to identify the atoms and molecules present in the material under test.

Stronger magnetic fields result in increased sensitivity, permitting the analysis of smaller structures and therefore a higher resolution. The increase in the magnetic field strength results in higher concentration of the aligned spins and in higher signature resonant frequency. These two factors give rise to improved resolution by means of a higher signal-to-noise ratio. Finally, since the energy change of the spins through a single scan is very small, a clear, high signal-to-noise ratio image is achieved by superposition of many repeated NMR scans.

In most NMR systems, the strong magnetic field is produced by *superconducting* electromagnets. In some configurations, *hybrid* magnets are used, where an inner layer constructed of a resistive electromagnet is surrounded by a superconducting magnet layer. In both cases, the magnet is commonly placed in the ground, with a conveniently constructed access to the bore. Several key terms are associated with the NMR technology (Dr. Vesna Mitrović, Centre National de la Recherche Scientifique, Grenoble, France, personal communication, 2002; Dr. Mitrović is now with Brown University), and are briefly outlined next.

Spectral resolution of the NMR measurement is expressed in *parts per million* (ppm), with reference to the frequency of the radio signal used for inducing the resonance. The *bore* of the NMR magnet is the hollow part of the NMR device, which holds the material or body under test. The entire NMR system of magnets and coils is cooled in a pool of liquid helium. The *cold bore* structure refers to the configuration where the magnet and the coils are placed directly into the liquid helium, while the “warm bore” configuration has additional layers of vacuum and liquid nitrogen, allowing the space within the bore itself to be at the room temperature. The *shim coils* (or bobbins), the inductive coils strategically placed and current-sourced with respect to the magnet, can be found in all NMR devices and serve for tunable compensation and improvement of magnetic field homogeneity. For detailed specifications and new solutions, the reader is encouraged to read additional information available on the internet sites of the leading manufacturers and research groups: Varian Inc.; Oxford Instruments, UK; National High Magnetic Field Laboratory in Florida, U.S.A.; and Centre National de la Recherche Scientifique in Grenoble, France.

Today, NMR is an essential tool for the discovery and development of pharmaceuticals [8]. Special state-of-the-art sensitive NMR systems with high resolution used in human genome research allow structural analysis to analyze DNA samples found, for example, in protein membranes.

Since the early 1980s, NMR techniques have been used for medical visualization of soft body tissues. This application of NMR is called *magnetic resonance imaging* (MRI), and it is enabled by hydrogen nuclei present in the water and lipid content of animal tissue. Imaging magnets for animal imaging commonly have higher field strengths (3–7 T) than those used for human diagnosis (0.3–1.5 T). MRI provides high-contrast images between different tissues (brain, heart, spleen, etc.) and is sufficiently sensitive to differentiate between normal tissues and those that are damaged or diseased. *Functional MRI* (fMRI) [20] uses higher-field magnets (4 T) to help visualize the activity of the sensory, cognitive and motor system. Figure 3.24 shows an example of an MRI scan of the brain of one of

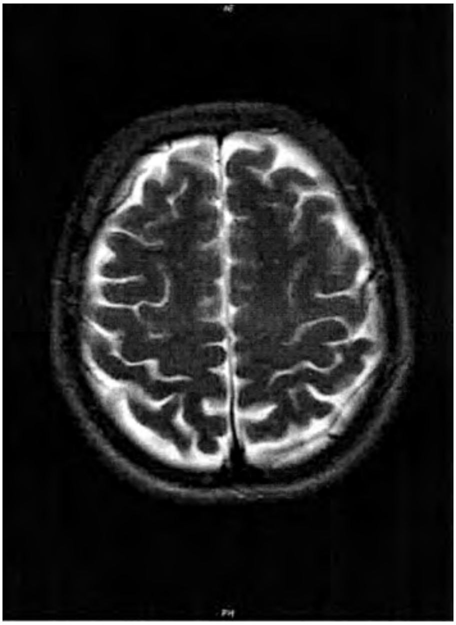


Figure 3.24 An MRI scan of the brain of one of the authors, performed using a GE instrument with a magnetic field flux density of 1.5 T (courtesy University of Colorado Health Science Center, Denver, Colorado).

the authors. At the time of this writing, the main manufacturers or MRI imaging systems are Siemens and General Electric.

REFERENCES AND FURTHER READING

1. Bloch, F.; Hansen, W.W.; Packard, M. The nuclear induction experiment. Phys. Rev. *70*, 474–485.
2. Carter, G.W. *The Electromagnetic Field in Its Engineering Aspects*; American Elsevier: New York, 1967.
3. Cheng, D. K. *Fundamentals of Engineering Electromagnetics*; Addison-Wesley, Reading, MA, 1993, pp. 172–194.
4. Damadian, D.V. Tumor detection by nuclear magnetic resonance. Science, **March**, **1971**.
5. Freeman, R. *A Handbook of Nuclear Magnetic Resonance*; Longman Scientific & Technical: Essex, England, 1988.
6. Grochowski E. The Continuing Evolution of Magnetic Hard Disk Drives. In *Holographic Data Storage*, Coufal H.J., Psaltis D., Sincerbox G.T. Eds.; Springer Series in Optical Sciences, Springer-Verlag; Berlin, 2000; 447–462.
7. Guru, B.S.; Hiziroglu, H.R. *Electromagnetic Field Theory Fundamentals*; PWS Publishing Company: Boston, 1998, pp. 155–187.
8. Gunther, H. *NMR Spectroscopy: Basic Principles*, *Concepts*, *and Applications in Chemistry*, 2nd Ed.; John Wiley & Son: New York, 1995.
9. Harrington, R.F. *Introduction to Electromagnetic Engineering*; McGraw-Hill: New York, 1958.
10. Hoole, S.R.H.; Hoole, R.R.P. *A modern Short Course in Engineering Electromagnetics*; Oxford University Press: New York, 1996, pp. 246–260.
11. Jackson, J.D. *Classical Electrodynamics*; John Wiley & Sons: New York, 1962, pp. 197–201.
12. King, R.W.; Pasad, S. *Fundamental Electromagnetic Theory and Applications*; Prentice Hall: Upper Saddle River, NJ, 1986, pp. 40–48.
13. Kraus, J.D. *Electromagnetics*; McGraw-Hill: New York, 1953, pp. 148–161, 214–238.
14. Mallison, J.C. *Magneto-Resistive and Spin Valve Heads*; Academic Press: San Diego, 2002.
15. Matson, G.B.; Weiner M.W. Spectroscopy. In *Magnetic Resonance Imaging*; Stark, D.D., Bradley, W.G. Jr. Eds.; Chapter 15, Mosby Year Books; St. Louis, 1992, 438–477.
16. Matveev, A.N. *Electricity and Magnetism*, Mir Publishers, Moscow, 1986.
17. Maxwell, J.C. *A Treatise on Electricity and Magnetism*; Dover Publications: New York, 1954, Vols. 1 and 2.
18. Popović, B.D. *Introductory Engineering Electromagnetics*; Addison-Wesley, Reading, MA, 1971, pp. 243–257, 298–322.
19. Popović Z.; Popović, B.D. *Introductory Electromagnetics*; Prentice Hall: NJ, 1999.
20. Rao, S.; Binder, J.; Bandettini, P.; Hammeke, T.; Yetkin, F.; Jesmanowicz, A.; Lisk, L.; Morris, G.; Mueller, W.; Estkowski, L.; Wong, E.; Haughton, V.; Hyde, J. “Functional magnetic resonance imaging of complex human movements.” Neurology **1993**, *43*, 2311–2318.
21. Shen, L.C.; Kong J.A. *Applied Electromagnetism*, 3rd Ed.; PWS Publishing Company: Coston, 1953, pp. 445–462.
22. Slater, J.C.; Frank, H.H. *Electromagnetism*; Dover Publications: New York, 1969, pp. 54–59, 67–70.
23. Silvester, P. *Modern Electromagnetic Fields*; Prentice Hall: Upper Saddle River, NJ, 1968, pp. 172–182.

4

Electromagnetic Induction

Milica Popović
McGill University, Montréal, Quebec, Canada

Branko D. Popović[†]
University of Belgrade, Belgrade, Yugoslavia

Zoya Popović
University of Colorado, Boulder, Colorado, U.S.A.

> To the loving memory of our father, professor, and coauthor. We hope that he would have agreed with the changes we have made after his last edits.
>
> — *Milica and Zoya Popović*

4.1. INTRODUCTION

In 1831 Michael Faraday performed experiments to check whether current is produced in a closed wire loop placed near a magnet, in analogy to dc currents producing magnetic fields. His experiment showed that this could not be done, but Faraday realized that a *time-varying current in the loop was obtained while the magnet was being moved toward it or away from it*. The law he formulated is known as *Faraday's law of electromagnetic induction*. It is perhaps the most important law of electromagnetism. Without it there would be no electricity from rotating generators, no telephone, no radio and television, no magnetic memories, to mention but a few applications.

The phenomenon of electromagnetic induction has a simple physical interpretation. Two charged particles ("charges") at rest act on each other with a force given by Coulomb's law. Two charges moving with *uniform velocities* act on each other with an additional force, the magnetic force. If a particle is *accelerated*, there is another additional force that it exerts on other charged particles, stationary or moving. As in the case of the magnetic force, if only a pair of charges is considered, this additional force is much smaller than Coulomb's force. However, time-varying currents in conductors involve a vast number of accelerated charges, and produce effects significant enough to be easily measurable.

This additional force is of the same *form* as the electric force ($\mathbf{F} = Q\mathbf{E}$). However, other properties of the electric field vector, $\mathbf{E}$ in this case, are different from those of the

[†] Deceased.

electric field vector of static charges. When we wish to stress this difference, we use a slightly different name: the *induced electric field strength.*

The induced electric field and electromagnetic induction have immense practical consequences. Some examples include:

The electric field of electromagnetic waves (e.g., radio waves or light) is basically the induced electric field;

In electrical transformers, the induced electric field is responsible for obtaining higher or lower voltage than the input voltage;

The skin effect in conductors with ac currents is due to induced electric field;

Electromagnetic induction is also the cause of ‘‘magnetic coupling’’ that may result in undesired interference between wires (or metal traces) in any system with time-varying current, an effect that increases with frequency.

The goal of this chapter is to present:

Fundamental theoretical foundations for electromagnetic induction, most importantly Faraday's law;

Important consequences of electromagnetic induction, such as Lentz's law and the skin effect;

Some simple and commonly encountered examples, such as calculation of the inductance of a solenoid and coaxial cable;

A few common applications, such as generators, transformers, electromagnets, etc.

4.2. THEORETICAL BACKGROUND AND FUNDAMENTAL EQUATIONS

4.2.1. The Induced Electric Field

The practical sources of the induced electric field are time-varying currents in a broader sense. If we have, for example, a stationary and rigid wire loop with a time-varying current, it produces an induced electric field. However, a wire loop that changes shape and/or is moving, carrying a *time-constant* current, also produces a time-varying current in space and therefore induces an electric field. Currents equivalent to Ampère's currents in a moving magnet have the same effect and therefore also produce an induced electric field.

Note that in both of these cases there exists, in addition, a time-varying magnetic field. Consequently, a time-varying (induced) electric field is always accompanied by a time-varying magnetic field, and conversely, a time-varying magnetic field is always accompanied by a time-varying (induced) electric field.

The basic property of the induced electric field $\mathbf{E}_{\text{ind}}$ is the same as that of the static electric field: it acts with a force $\mathbf{F} = Q\mathbf{E}_{\text{ind}}$ on a point charge Q. However, the two components of the electric field differ in the work done by the field in moving a point charge around a closed contour. For the static electric field this work is always zero, but for the induced electric field it is not. Precisely this property of the induced electric field gives rise to a very wide range of consequences and applications. Of course, a charge can be situated simultaneously in both a static (Coulomb-type) and an induced field, thus being subjected to a total force

$$\mathbf{F} = Q(\mathbf{E}_{\text{st}} + \mathbf{E}_{\text{ind}}) \tag{4.1}$$

We know how to calculate the static electric field of a given distribution of charges, but how can we determine the induced electric field strength? When a charged particle is moving with a velocity **v** with respect to the source of the magnetic field, the answer follows from the magnetic force on the charge:

$$\mathbf{E}_{\text{ind}} = \mathbf{v} \times \mathbf{B} \qquad (\text{V/m}) \tag{4.2}$$

If we have a current distribution of density **J** (a slowly time-varying function of position) in vacuum, localized inside a volume **v**, the induced electric field is found to be

$$\mathbf{E}_{\text{ind}} = -\frac{\partial}{\partial t}\left(\frac{\mu_0}{4\pi}\int_V \frac{\mathbf{J}\cdot dV}{r}\right) \qquad (\text{V/m}) \tag{4.3}$$

In this equation, r is the distance of the point where the induced electric field is being determined from the volume element dV. In the case of currents over surfaces, $\mathbf{J}(t)\cdot dV$ in Eq. (4.3) should be replaced by $\mathbf{J}_s(t)\cdot dS$, and in the case of a thin wire by $i(t)\cdot dl$.

If we know the distribution of time-varying currents, Eq. (4.3) enables the determination of the induced electric field at any point of interest. Most often it is not possible to obtain the induced electric field strength in analytical form, but it can always be evaluated numerically.

4.2.2. Faraday's Law of Electromagnetic Induction

Faraday's law is an equation for the *total* electromotive force (*emf*) induced in a closed loop due to the induced electric field. This electromotive force is distributed along the loop (not concentrated at a single point of the loop), but we are rarely interested in this distribution. Thus, Faraday's law gives us what is relevant only from the circuit-theory point of view—the *emf* of the Thevenin generator equivalent to all the elemental generators acting in the loop.

Consider a closed conductive contour C, either moving arbitrarily in a time-constant magnetic field or stationary with respect to a system of time-varying currents producing an induced electric field. If the wire segments are moving in a magnetic field, there is an induced field acting along them of the form in Eq. (4.2), and if stationary, the induced electric field is given in Eq. (4.3). In both cases, a segment of the wire loop behaves as an elemental generator of an *emf*

$$de = \mathbf{E}_{\text{ind}}\cdot d\mathbf{l} \tag{4.4}$$

so that the *emf* induced in the entire contour is given by

$$e = \oint_C \mathbf{E}_{\text{ind}}\cdot d\mathbf{l} \tag{4.5}$$

If the *emf* is due to the contour motion only, this becomes

$$e = \oint_C \mathbf{v}\times\mathbf{B}\cdot d\mathbf{l} \tag{4.6}$$

It can be shown that, whatever the cause of the induced electric field (the contour motion, time-varying currents, or the combination of the two), the total *emf* induced in the contour can be expressed in terms of time variation of the magnetic flux through the contour:

$$e = \oint_C \mathbf{E}_{\text{ind}} \cdot d\mathbf{l} = -\frac{d\Phi_{\text{through } C \text{ in } dt}}{dt} = -\frac{d}{dt}\int_S \mathbf{B} \cdot d\mathbf{S} \quad \text{(V)} \tag{4.7}$$

This is *Faraday's law of electromagnetic induction.* The reference direction along the contour, by convention, is connected with the reference direction of the normal to the surface S spanning the contour by the right-hand rule. Note again that the induced *emf* in this equation is nothing but the voltage of the Thévenin generator equivalent to all the elemental generators of electromotive forces $\mathbf{E}_{\text{ind}} \cdot d\mathbf{l}$ acting around the loop. The possibility of expressing the induced *emf* in terms of the magnetic flux alone is not surprising. We know that the induced electric field is always accompanied by a magnetic field, and the above equation only reflects the relationship that exists between the two fields (although the relationship itself is not seen from the equation). Finally, this equation is valid only if the time variation of the magnetic flux through the contour is due either to motion of the contour in the magnetic field or to time variation of the magnetic field in which the contour is situated (or a combination of the two). No other cause of time variation of the magnetic flux will result in an induced *emf.*

4.2.3. Potential Difference and Voltage in a Time-varying Electric and Magnetic Field

The voltage between two points is defined as the line integral of the *total* electric field strength, given in Eq. (4.1), from one point to the other. In electrostatics, the induced electric field does not exist, and voltage does not depend on the path between these points. This is *not* the case in a time-varying electric and magnetic field.

Consider arbitrary time-varying currents and charges producing a time-varying electric and magnetic field, Fig. 4.1. Consider two points, A and B, in this field, and two paths, a and b, between them, as indicated in the figure. The voltage between these two points along the two paths is given by

$$V_{AB\,\text{along}\,a\,\text{or}\,b} = \int_{A\,\text{along}\,a\,\text{or}\,b}^{B} (\mathbf{E}_{\text{st}} + \mathbf{E}_{\text{ind}}) \cdot d\mathbf{l} \tag{4.8}$$

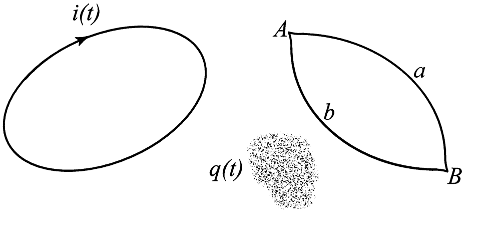


Figure 4.1 An arbitrary distribution of time-varying currents and charges.

The integral between A and B of the static part is simply the potential difference between A and B, and therefore

$$V_{AB\,\mathrm{along}\,a\,\mathrm{or}\,b} = V_A - V_B + \int_{A\,\mathrm{along}\,a\,\mathrm{or}\,b}^{B} \mathbf{E}_{\mathrm{ind}} \cdot d\mathbf{l} \tag{4.9}$$

The potential difference $V_A - V_B$ does not depend on the path between A and B, but the integral in this equation is different for paths a and b. These paths form a closed contour. Applying Faraday's law to that contour, we have

$$e_{\text{induced in closed contour } AaBbA} = \oint_{AaBbA} \mathbf{E}_{\mathrm{ind}} \cdot d\mathbf{l} = \int_{AaB} \mathbf{E}_{\mathrm{ind}} \cdot d\mathbf{l} + \int_{BbA} \mathbf{E}_{\mathrm{ind}} \cdot d\mathbf{l} = -\frac{d\Phi}{dt} \tag{4.10}$$

where Φ is the magnetic flux through the surface spanned by the contour $AaBbA$. Since the right side of this equation is generally nonzero, the line integrals of $\mathbf{E}_{\mathrm{ind}}$ from A to B along a and along b are different. Consequently, *the voltage between two points in a time-varying electric and magnetic field depends on the choice of integration path between these two points.*

This is a very important practical conclusion for time-varying electrical circuits. It implies that, contrary to circuit theory, the voltage measured across a circuit by a voltmeter depends on the shape of the leads connected to the voltmeter terminals. Since the measured voltage depends on the rate of change of magnetic flux through the surface defined by the voltmeter leads and the circuit, this effect is particularly pronounced at high frequencies.

4.2.4. Self-inductance and Mutual Inductance

A time-varying current in one current loop induces an *emf* in another loop. In linear media, an electromagnetic parameter that enables simple determination of this *emf* is the *mutual inductance*.

A wire loop with time-varying current creates a time-varying induced electric field not only in the space around it but also along the loop itself. As a consequence, there is a feedback—the current produces an effect which affects itself. The parameter known as *inductance*, or *self-inductance*, of the loop enables simple evaluation of this effect.

Consider two stationary thin conductive contours C_1 and C_2 in a linear medium (e.g., air), shown in Fig. 4.2. When a time-varying current $i_1(t)$ flows through the first contour, it creates a time-varying magnetic field, as well as a time-varying induced electric field, $\mathbf{E}_{1\,\mathrm{ind}}(t)$. The latter produces an *emf* $e_{12}(t)$ in the second contour, given by

$$e_{12}(t) = \oint_{C_2} \mathbf{E}_{1\,\mathrm{ind}} \cdot d\mathbf{l}_2 \tag{4.11}$$

where the first index denotes the source of the field (contour 1 in this case).

It is usually much easier to find the induced *emf* using Faraday's law than in any other way. The magnetic flux density vector in linear media is proportional to the current

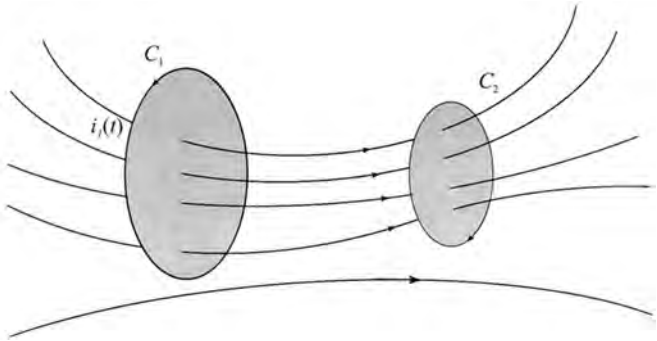


Figure 4.2 Two coupled conductive contours.

that causes the magnetic field. It follows that the flux $\Phi_{12}(t)$ through C_2 caused by the current $i_1(t)$ in C_1 is also proportional to $i_1(t)$:

$$\Phi_{12}(t) = L_{12} \cdot i_1(t) \tag{4.12}$$

The proportionality constant L_{12} is the *mutual inductance* between the two contours. This constant depends only on the geometry of the system and the properties of the (linear) medium surrounding the current contours. Mutual inductance is denoted both by L_{12} or sometimes in circuit theory by M.

Since the variation of $i_1(t)$ can be arbitrary, the same expression holds when the current through C_1 is a dc current:

$$\Phi_{12} = L_{12}I_1 \tag{4.13}$$

Although mutual inductance has no practical meaning for dc currents, this definition is used frequently for the determination of mutual inductance.

According to Faraday's law, the *emf* can alternatively be written as

$$e_{12}(t) = -\frac{d\Phi_{12}}{dt} = -L_{12}\frac{di_1(t)}{dt} \tag{4.14}$$

The unit for inductance, equal to a Wb/A, is called a *henry* (H). One henry is quite a large unit. Most frequent values of mutual inductance are on the order of a mH, μH, or nH.

If we now assume that a current $i_2(t)$ in C_2 causes an induced *emf* in C_1, we talk about a mutual inductance L_{21}. It turns out that $L_{12} = L_{21}$ always. [This follows from the expression for the induced electric field in Eqs. (4.3) and (4.5).] So, we can write

$$L_{12} = \frac{\Phi_{12}}{I_1} = L_{21} = \frac{\Phi_{21}}{I_2} \qquad \text{(H)} \tag{4.15}$$

These equations show that we need to calculate either Φ_{12} or Φ_{21} to determine the mutual inductance, which is a useful result since in some instances one of these is much simpler to calculate than the other.

Note that mutual inductance can be negative as well as positive. The sign depends on the actual geometry of the system and the adopted reference directions along the two loops: if the current in the reference direction of one loop produces a positive flux in the other loop, then mutual inductance is positive, and vice versa. For calculating the flux, the normal to the loop surface is determined by the right-hand rule with respect to its reference direction.

As mentioned, when a current in a contour varies in time, the induced electric field exists everywhere around it and therefore also along its entire length. Consequently, there is an induced *emf* in the contour itself. This process is known as *self-induction.* The simplest (even if possibly not physically the clearest) way of expressing this *emf* is to use Faraday's law:

$$e(t) = -\frac{d\Phi_{\text{self}}(t)}{dt} \tag{4.16}$$

If the contour is in a linear medium (i.e., the flux through the contour is proportional to the current), we define the *self-inductance* of the contour as the ratio of the flux $\Phi_{\text{self}}(\mathrm{t})$ through the contour due to current $i(t)$ in it and $i(t)$,

$$L = \frac{\Phi_{\text{self}}(t)}{i(t)} \qquad (\mathrm{H}) \tag{4.17}$$

Using this definition, the induced *emf* can be written as

$$e(t) = -L\frac{di(t)}{dt} \tag{4.18}$$

The constant L depends only on the geometry of the system, and its unit is again a henry (H). In the case of a dc current, $L = \Phi/I$, which can be used for determining the self-inductance in some cases in a simple manner.

The self-inductances of two contours and their mutual inductance satisfy the following condition:

$$L_{11}L_{22} \geq L_{12}^2 \tag{4.19}$$

Therefore, the largest possible value of mutual inductance is the geometric mean of the self-inductances. Frequently, Eq. (4.19) is written as

$$L_{12} = k\sqrt{L_{11}L_{22}} \qquad -1 \leq k \leq 1 \tag{4.20}$$

The dimensionless coefficient k is called the *coupling coefficient.*

4.2.5. Energy and Forces in the Magnetic Field

There are many devices that make use of electric or magnetic forces. Although this is not commonly thought of, almost any such device can be made in an "electric version" and in a "magnetic version." We shall see that the magnetic forces are several orders of magnitude stronger than electric forces. Consequently, devices based on magnetic forces are much smaller in size, and are used more often when force is required. For example, electric motors in your household and in industry, large cranes for lifting ferromagnetic objects, home bells, electromagnetic relays, etc., all use magnetic, not electric, forces.

A powerful method for determining magnetic forces is based on energy contained in the magnetic field. While establishing a dc current, the current through a contour has to change from zero to its final dc value. During this process, there is a changing magnetic flux through the contour due to the changing current, and an *emf* is induced in the contour. This *emf* opposes the change of flux (see Lentz's law in Sec. 4.3.2). In order to establish the final static magnetic field, the sources have to overcome this *emf*, i.e., to spend some energy. A part (or all) of this energy is stored in the magnetic field and is known as *magnetic energy*.

Let n contours, with currents $i_1(t), i_2(t), \ldots, i_n(t)$ be the sources of a magnetic field. Assume that the contours are connected to generators of electromotive forces $e_1(t), e_2(t), \ldots, e_n(t)$. Finally, let the contours be stationary and rigid (i.e., they cannot be deformed), with total fluxes $\Phi_1(t), \Phi_2(t), \ldots, \Phi_n(t)$. If the medium is linear, energy contained in the magnetic field of such currents is

$$W_m = \frac{1}{2}\sum_{k=1}^{n} I_k \Phi_k \tag{4.21}$$

This can be expressed also in terms of self- and mutual inductances of the contours and the currents in them, as

$$W_m = \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} L_{jk} I_j I_k \tag{4.22}$$

which for the important case of a single contour becomes

$$W_m = \frac{1}{2} I\Phi = \frac{1}{2} LI^2 \tag{4.23}$$

If the medium is ferromagnetic these expressions are not valid, because at least one part of the energy used to produce the field is transformed into heat. Therefore, for ferromagnetic media it is possible only to evaluate the total energy used to obtain the field. If B_1 is the initial magnetic flux density and B_2 the final flux density at a point, energy density spent in order to change the magnetic flux density vector from B_1 to B_2 at that point is found to be

$$\frac{dA_m}{dV} = \int_{B_1}^{B_2} H(t) \cdot dB(t) \qquad (\mathrm{J/m^3}) \tag{4.24}$$

In the case of linear media (see Chapter 3), energy used for changing the magnetic field is stored in the field, i.e., $dA_m = dW_m$. Assuming that the B field changed from zero to some value B, the volume density of magnetic energy is given by

$$\frac{dW_m}{dV} = \int_b^B \frac{B}{\mu} \cdot dB = \frac{1}{2}\frac{B^2}{\mu} = \frac{1}{2}\mu H^2 = \frac{1}{2}BH \qquad (\mathrm{J/m^3}) \tag{4.25}$$

The energy in a *linear medium* can now be found by integrating this expression over the entire volume of the field:

$$W_m = \int_V \frac{1}{2}\mu H^2 \, dV \qquad \text{(J)} \tag{4.26}$$

If we know the distribution of currents in a magnetically homogeneous medium, the magnetic flux density is obtained from the Biot-Savart law. Combined with the relation $d\mathbf{F}_m = I \cdot d\mathbf{l} \times \mathbf{B}$, we can find the magnetic force on any part of the current distribution. In many cases, however, this is quite complicated.

The magnetic force can also be evaluated as a derivative of the magnetic energy. This can be done assuming either (1) the fluxes through all the contours are kept constant or (2) the currents in all the contours are kept constant. In some instances this enables very simple evaluation of magnetic forces.

Assume first that during a displacement dx of a body in the magnetic field along the x axis, we keep the fluxes through all the contours constant. This can be done by varying the currents in the contours appropriately. The x component of the magnetic force acting on the body is then obtained as

$$F_x = -\left(\frac{dW_m}{dx}\right)_{\Phi = \text{const}} \tag{4.27}$$

In the second case, when the currents are kept constant,

$$F_x = +\left(\frac{dW_m}{dx}\right)_{I = \text{const}} \tag{4.28}$$

The signs in the two expressions for the force determine the direction of the force. In Eq. (4.28), the positive sign means that when current sources are producing all the currents in the system ($I = \text{const}$), the magnetic field energy increases, as the generators are the ones that add energy to the system and produce the force.

4.3. CONSEQUENCES OF ELECTROMAGNETIC INDUCTION

4.3.1. Magnetic Coupling

Let a time-varying current $i(t)$ exist in a circular loop C_1 of radius a, Fig. 4.3. According to Eq. (4.3), lines of the induced electric field around the loop are circles centered at the loop axis normal to it, so that the line integral of the induced electric field around a circular contour C_2 indicated in the figure in dashed line *is not zero*. If the contour C_2 is a wire loop, this field acts as a distributed generator along the entire loop length, and a current is induced in that loop.

The reasoning above does not change if loop C_2 is not circular. We have thus reached an extremely important conclusion: *The induced electric field of time-varying currents in one wire loop produces a time-varying current in an adjacent closed wire loop.* Note that the other loop need not (and usually does not) have any physical contact with the first loop. This means that the induced electric field enables transport of energy from one loop to the other through vacuum. Although this coupling is actually obtained by means of the induced electric field, it is known as *magnetic coupling*.

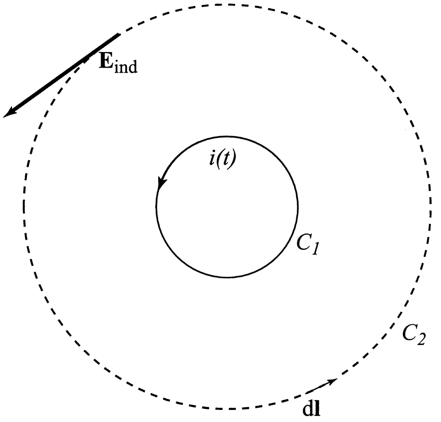


Figure 4.3 A circular loop C_1 with a time-varying current $i(t)$. The induced electric field of this current is tangential to the circular loop C_2 indicated in dashed line, so that it results in a distributed emf around the loop.

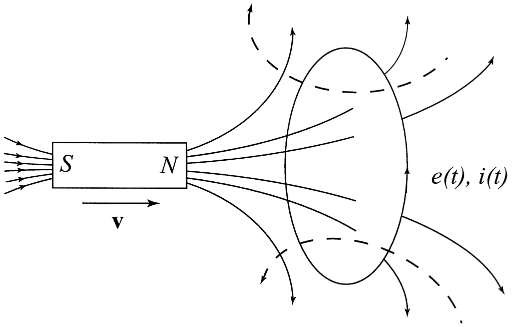


Figure 4.4 Illustration of Lentz's law.

Note that if the wire loop C_2 is not closed, the induced field nevertheless induces distributed generators along it. The loop behaves as an open-circuited equivalent (Thévenin) generator.

4.3.2. Lentz's Law

Figure 4.4 shows a permanent magnet approaching a stationary loop. The permanent magnet is equivalent to a system of macroscopic currents. Since it is moving, the magnetic flux created by these currents through the contour varies in time. According to the reference direction of the contour shown in the figure, the change of flux is positive, $(d\Phi/dt) > 0$, so the induced *emf* is in the direction shown in the figure. The *emf* produces a current through the closed loop, which in turn produces its own magnetic field, shown in the figure in dashed line. As a result, the change of the magnetic flux, caused initially by the magnet motion, is reduced. This is *Lentz's law*: the induced current in a conductive contour tends to decrease the *change* in magnetic flux through the contour. Lentz's law describes a feedback property of electromagnetic induction.

4.3.3. Eddy Currents

A very important consequence of the induced electric field are eddy currents. These are currents induced throughout a solid metal body when the body is situated in a time-varying magnetic (i.e., induced electric) field.

As the first consequence of eddy currents, there is power lost to heat according to Joule's law. Since the magnitude of eddy currents is proportional to the magnitude of the induced electric field, eddy-current losses are proportional to the square of frequency.

As the second consequence, there is a secondary magnetic field due to the induced currents which, following Lentz's law, reduces the magnetic field inside the body. Both of these effects are usually not desirable. For example, in a ferromagnetic core shown in Fig. 4.5, Lentz's law tells us that eddy currents tend to decrease the flux in the core, and the magnetic circuit of the core will not be used efficiently. The flux density vector is the smallest at the center of the core, because there the **B** field of all the induced currents adds up. The total magnetic field distribution in the core is thus nonuniform.

To reduce these two undesirable effects, ferromagnetic cores are made of mutually insulated thin sheets, as shown in Fig. 4.6. Now the flux through the sheets is encircled by much smaller loops, the *emf* induced in these loops is consequently much smaller, and so the eddy currents are also reduced significantly. Of course, this only works if vector **B** is parallel to the sheets.

In some instances, eddy currents are created on purpose. For example, in induction furnaces for melting metals, eddy currents are used to heat solid metal pieces to melting temperatures.

4.3.4. The Skin Effect and the Proximity and Edge Effects

A time-invariant current in a homogeneous cylindrical conductor is distributed uniformly over the conductor cross section. If the conductor is not cylindrical, the time-invariant current in it is not distributed uniformly, *but it exists in the entire conductor*. A time-varying current has a tendency to concentrate near the surfaces of conductors. At very high frequencies, the current is restricted to a very thin layer near the conductor surface, practically on the surfaces themselves. Because of this extreme case, the entire phenomenon of nonuniform distribution of time-varying currents in conductors is known as the *skin effect*.

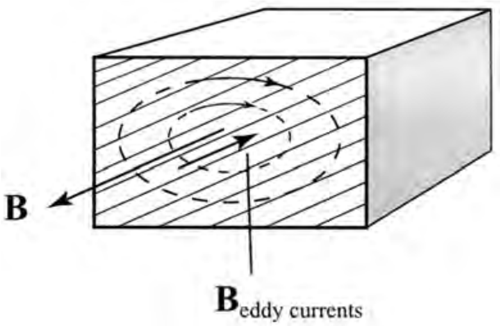


Figure 4.5 Eddy currents in a piece of ferromagnetic core. Note that the total **B** field in the core is reduced due to the opposite field created by eddy currents.

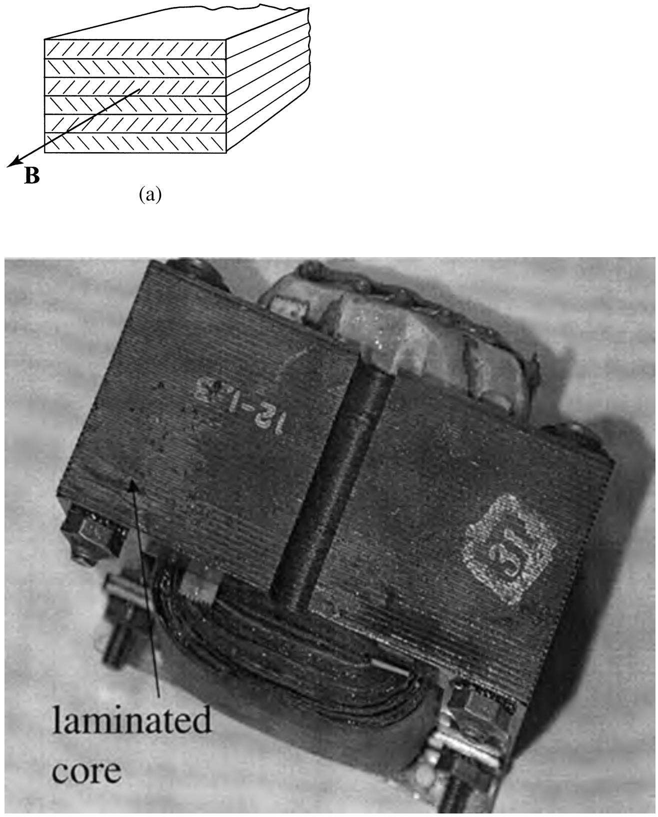


Figure 4.6 A ferromagnetic core for transformers and ac machines consists of thin insulated sheets: (a) sketch of core and (b) photograph of a typical transformer core.

The cause of skin effect is electromagnetic induction. A time-varying magnetic field is accompanied by a time-varying induced electric field, which in turn creates secondary time-varying currents (induced currents) and a secondary magnetic field. The induced currents produce a magnetic flux which opposes the external flux (the same flux that "produced" the induced currents). As a consequence, the total flux is reduced. The larger the conductivity, the larger the induced currents are, and the larger the permeability, the more pronounced the flux reduction is. Consequently, both the total time-varying magnetic field and induced currents inside conductors are reduced when compared with the dc case.

The skin effect is of considerable practical importance. For example, at very high frequencies a very thin layer of conductor carries most of the current. Any conductor (or for that matter, any other material), can be coated with silver (the best available conductor) and practically the entire current will flow through this thin silver coating. Even at power frequencies in the case of high currents, the use of thick solid conductors is not efficient, and bundled conductors are used instead.

The skin effect exists in all conductors, but, as mentioned, the tendency of current and magnetic flux to be restricted to a thin layer on the conductor surface is much more pronounced for a ferromagnetic conductor than for a nonferromagnetic conductor of the same conductivity. For example, for iron at 60 Hz the thickness of this layer is on the order of only 0.5 mm. Consequently, solid ferromagnetic cores for alternating current electric motors, generators, transformers, etc., would result in poor use of the ferromagnetic material and high losses. Therefore, laminated cores made of thin, mutually insulated sheets are used instead. At very high frequencies, ferrites (ferrimagnetic ceramic materials) are used, because they have very low conductivity when compared to metallic ferromagnetic materials.

Consider a body with a sinusoidal current of angular frequency ω and let the material of the body have a conductivity σ and permeability μ. If the frequency is high enough, the current will be distributed over a very thin layer over the body surface, the current density being maximal at the surface (and parallel to it), and decreasing rapidly with the distance z from it:

$$J(z) = J_0 e^{-jz/\delta} \tag{4.29}$$

where

$$\delta = \sqrt{\frac{\omega\mu\sigma}{2}} = \sqrt{\pi\mu\sigma}\sqrt{f} \tag{4.30}$$

The intensity of the current density vector decreases exponentially with increasing z. At a distance δ the amplitude of the current density vector decreases to $1/e$ of its value J_0 at the boundary surface. This distance is known as the *skin depth*. For example, for copper ($\sigma = 57 \times 10^6$ S/m, $\mu = \mu_0$), the skin depth at 1 MHz is only 0.067 mm. For iron ($\sigma = 10^7$ S/m, $\mu_r = 1000$), the skin depth at 60 Hz is 0.65 mm, and for sea water ($\sigma = 4$ S/m, $\mu = \mu_0$), at the same frequency it is 32.5 m. Table 4.1 summarizes the value of skin depth in some common materials at a few characteristic frequencies.

The result for skin depth for iron at power frequencies (50 Hz or 60 Hz), $\delta \cong 5$ mm, tells us something important. Iron has a conductivity that is only about six times less than that of copper. On the other hand, copper is much more expensive than iron. Why do we then not use iron wires for the distribution of electric power in our homes? Noting that there are millions of kilometers of such wires, the savings would be very large. Unfortunately, due to a large relative permeability—iron has very small power-frequency skin depth (a fraction of a millimeter)—the losses in iron wire are large, outweighing the savings, so copper or aluminum are used instead.

Keeping the current intensity the same, Joule losses increase with frequency due to increased resistance in conductors resulting from the skin effect. It can be shown that Joule's losses per unit area are given by

$$\frac{dP_J}{dS} = R_s|H_0|^2 \qquad (\mathrm{W/m^2}) \tag{4.31}$$

Table 4.1 Values of Skin Depth for Some Common Materials at 60 Hz, 1 kHz, 1 MHz, and 1 GHz.

Material	$f = 60$ Hz	$f = 1$ kHz	$f = 1$ MHz	$f = 1$ GHz
Copper	8.61 mm	2.1 mm	0.067 mm	2.11 μm
Iron	0.65 mm	0.16 mm	5.03 μm	0.016 μm
Sea water	32.5 m	7.96 m	0.25 m	7.96 mm
Wet soil	650 m	159m	5.03 m	0.16 m

where H_0 is the complex rms value of the tangential component of the vector **H** on the conductor surface, and R_s is the *surface resistance* of the conductor, given by

$$R_s = \sqrt{\frac{\omega\mu}{2\sigma}} \qquad (\Omega) \tag{4.32}$$

Equation (4.32) is used for determining the attenuation in all metal waveguides, such as two-wire lines (twin-lead), coaxial lines, and rectangular waveguides.

The term *proximity effect* refers to the influence of alternating current in one conductor on the current distribution in another nearby conductor. Consider a coaxial cable of finite length. Assume for the moment that there is an alternating current only in the inner conductor (for example, that it is connected to a generator), and that the outer conductor is not connected to anything. If the outer conductor is much thicker than the skin depth, there is practically no magnetic field inside the outer conductor. If we apply Ampère's law to a coaxial circular contour contained in that conductor, it follows that the induced current on the *inside* surface of the outer conductor is exactly equal and opposite to the current in the inner conductor. This is an example of the proximity effect. If in addition there is normal cable current in the outer conductor, it is the same but opposite to the current on the conductor outer surface, so the two cancel out. We are left with a current over the inner conductor and a current over the inside surface of the outer conductor. This combination of the skin and proximity effects is what is usually actually encountered in practice.

Redistribution of Current on Parallel Wires and Printed Traces

Consider as the next example three long parallel wires a certain distance apart lying in one plane. The three ends are connected together at one and at the other end of the wires, and these common ends are connected by a large loop to a generator of sinusoidal *emf*. Are the currents in the three wires the same? At first glance we should expect them to be the same, but due to the induced electric field they are not: the current intensity in the middle wire will always be smaller than in the other two.

The above example is useful for understanding the distribution of ac current across the cross section of a printed metal strip, such as a trace on a printed-circuit board. The distribution of current across the strip will not be uniform (which it is at zero frequency). The current amplitude will be much greater along the strip edges than along its center. This effect is sometimes referred to as the *edge effect*, but it is, in fact, the skin effect in strip conductors. Note that for a strip line (consisting of two close parallel strips) this effect is very small because the induced electric fields due to opposite currents in the two strips practically cancel out.

4.3.5. Limitations of Circuit Theory

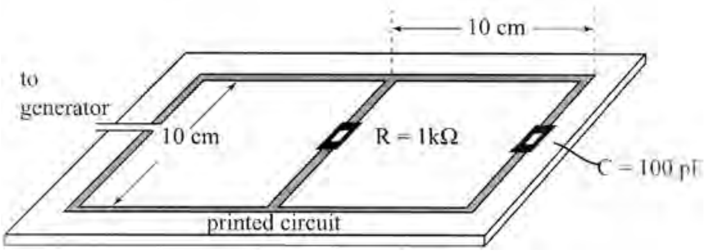
Circuit theory is the basic tool of electrical engineers, but it is approximate and therefore has limitations. These limitations can be understood only using electromagnetic-field theory. We consider here the approximations implicit in Kirchhoff's voltage law (KVL). This law states that the sum of voltages across circuit branches along any closed path is zero and that voltages and currents in circuit branches do not depend on the circuit actual geometrical shape. Basically, this means that this law neglects the induced electric field produced by currents in the circuit branches. This field increases with frequency, so that at a certain frequency (depending on circuit properties and its actual size) the influence of the induced electric field on circuit behavior becomes of the same order of magnitude as that due to generators in the circuit. The analysis of circuit behavior in such cases needs to be performed by electromagnetic analysis, usually requiring numerical solutions.

As a simple example, consider the circuit in Fig. 4.7, consisting of several printed traces and two lumped (pointlike, or much smaller than a wavelength) surface-mount components. For a simple two-loop circuit 10 cm × 20 cm in size, already at a frequency of 10 MHz circuit analysis gives results with errors exceeding 20%. The tabulated values in Fig. 4.7 show the calculated and measured complex impedance seen by the generator at different frequencies.

Several useful practical conclusions can be drawn. The first is that for circuits that contain wires or traces and low-valued resistors, this effect will become pronounced at lower frequencies. The second is that the behavior of an ac circuit always depends on the circuit shape, although in some cases this effect might be negligible. (A complete electromagnetic numerical solution of this circuit would give exact agreement with theory.) This directly applies to measurements of ac voltages (and currents), since the leads of the meter are also a part of the circuit. Sometimes, there is an *emf* induced in the meter leads due to flux through loops formed by parts of the circuit and the leads. This can lead to errors in voltage measurements, and the loops that give rise to the error *emf* are often referred to as *ground loops*.

4.3.6. Superconducting Loops

Some substances have zero resistivity at very low temperatures. For example, lead has zero resistivity below about 7.3 K (just a little bit warmer than liquid helium). This phenomenon is known as *superconductivity*, and such conductors are said to be



Frequency	Calculated Re(Z)	Measured Re(Z)	Calculated Im(Z)	Measured Im(Z)
10 MHz	25 Ω	20 Ω	−150 Ω	−110 Ω
20 MHz	6 Ω	1 Ω	−90 Ω	≈0 Ω
50 MHz	1 Ω	5 Ω	−50 Ω	+180 Ω
100 MHz	≈0 Ω	56 Ω	−15 Ω	+470 Ω

Figure 4.7 Example of impedance seen by the generator for a printed circuit with a surface-mount resistor and capacitor. The table shows a comparison of results obtained by circuit theory and measured values, indicating the range of validity of circuit theory.

superconductors. Some ceramic materials (e.g., yttrium barium oxide) become superconductors at temperatures as "high" as about 70 K (corresponding to the temperature of liquid nitrogen). Superconducting loops have an interesting property when placed in a time-varying magnetic field. The Kirchhoff voltage law for such a loop has the form

$$-\frac{d\Phi}{dt} = 0 \tag{4.33}$$

since the *emf* in the loop is $-d\Phi/dt$ and the loop has zero resistance. From this equation, it is seen that the flux through a superconducting loop remains *constant*. Thus, it is not possible to change the magnetic flux through a superconducting loop by means of electromagnetic induction. The physical meaning of this behavior is the following: If a superconducting loop is situated in a time-varying induced electric field, the current induced in the loop must vary in time so as to produce exactly the same induced electric field in the loop, but in the opposite direction. If this were not so, infinite current would result.

4.4. APPLICATIONS OF ELECTROMAGNETIC INDUCTION AND FARADAY'S LAW

4.4.1. An AC Generator

An ac generator, such as the one sketched in Fig. 4.8, can be explained using Faraday's law. A rectangular wire loop is rotating in a uniform magnetic field (for example, between the poles of a magnet). We can measure the induced voltage in the wire by connecting a voltmeter between contacts C_1 and C_2. Vector **B** is perpendicular to the contour axis. The loop is rotating about this axis with an angular velocity ω. If we assume that at $t = 0$ vector **B** is parallel to vector **n** normal to the surface of the loop, the induced *emf* in the loop is given by

$$e(t) = -\frac{d\Phi(t)}{dt} = \omega abB \sin \omega t = E_{\max} \sin \omega t \tag{4.34}$$

In practice, the coil has many turns of wire instead of a single loop, to obtain a larger induced *emf*. Also, usually the coil is not rotating, but instead the magnetic field is rotating around it, which avoids sliding contacts of the generator.

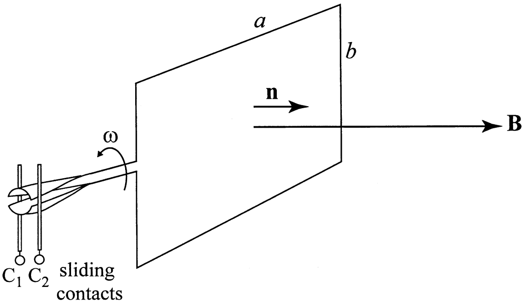


Figure 4.8 A simple ac generator.

4.4.2. Induction Motors

Motors transform electric to mechanical power through interaction of magnetic flux and electric current [1,20,26]. Electric motors are broadly categorized as ac and dc motors, with a number of subclassifications in each category. This section describes the basic operation of induction motors, which are most often encountered in industrial use.

The principles of the polyphase induction motor are here explained on the example of the most commonly used three-phase version. In essence, an induction motor is a transformer. Its magnetic circuit is separated by an air gap into two portions. The fixed *stator* carries the primary winding, and the movable *rotor* the secondary winding, as shown in Fig. 4.9a. An electric power system supplies alternating current to the primary

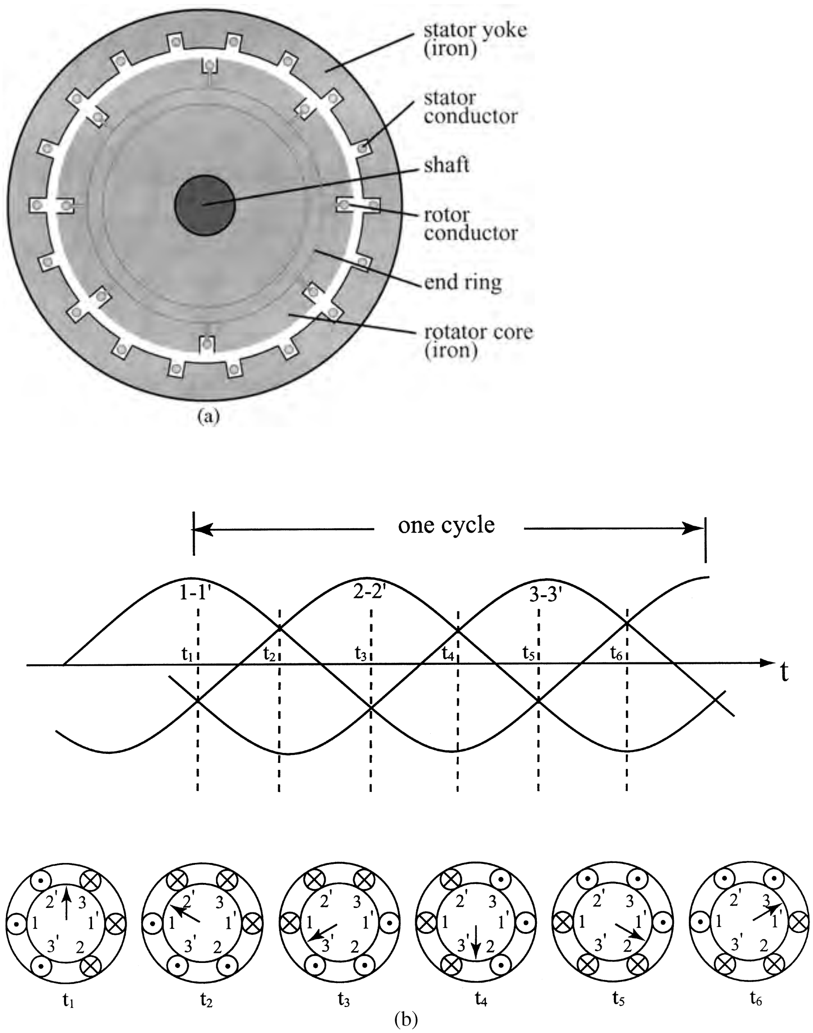


Figure 4.9 (a) Cross section of a three-phase induction motor. 1-1′, 2-2′, and 3-3′ mark the primary stator windings, which are connected to an external three-phase power supply. (b) Time-domain waveforms in the windings of the stator and resulting magnetic field vector rotation as a function of time.

winding, which induces currents in the secondary (short-circuited or closed through an external impedance) and thus causes the motion of the rotor. The key distinguishable feature of this machine with respect to other motors is that the current in the secondary is produced only by electromagnetic induction, i.e., not by an external power source.

The primary windings are supplied by a three-phase system currents, which produce three stationary *alternating* magnetic fields. Their superposition yields a sinusoidally distributed magnetic field in the air gap of the stator, revolving synchronously with the power-supply frequency. The field completes one revolution in one cycle of the stator current, as illustrated in Fig. 4.9b. Thus, the combined effect of three-phase alternating currents with the shown angular arrangement in the stator, results in a *rotating* magnetic field with a constant magnitude and a mechanical angular speed that depends on the frequency of the electric supply.

Two main types of induction motors differ in the configuration of the secondary windings. In squirrel-cage motors, the secondary windings of the rotor are constructed from conductor bars, which are short-circuited by end rings. In the wound-rotor motors, the secondary consists of windings of discrete conductors with the same number of poles as in the primary stator windings.

4.4.3. Electromagnetic Measurement of Fluid Velocity

The velocity of flowing liquids that have a small, but finite, conductivity can be measured using electromagnetic induction. In Fig. 4.10, the liquid is flowing through a flat insulating pipe with an unknown velocity $\mathbf{v}$. The velocity of the fluid is roughly uniform over the cross section of the pipe. To measure the fluid velocity, the pipe is in a magnetic field with a flux density vector **B** normal to the pipe. Two small electrodes are in contact with the fluid at the two ends of the pipe cross section. A voltmeter with large input impedance shows a voltage V when connected to the electrodes. The velocity of the fluid is then given by $v = V/B$.

4.4.4. Measurement of AC Currents

A useful application of the induced electric field is for measurement of a sinusoidal current in a conductor without breaking the circuit (as required by standard current measurement). Figure 4.11 shows a conductor with a sinusoidal current of amplitude I_m and angular frequency ω flowing through it. The conductor is encircled by a flexible thin rubber strip of cross-sectional area S, densely wound along its length with N' turns of wire per unit length. We show that if we measure the amplitude of the voltage between the terminals of the strip winding, e.g., V_m, we can calculate I_m.

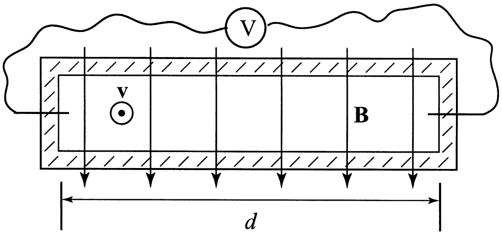


Figure 4.10 Measurement of fluid velocity.

There are $dN = N'dl$ turns of wire on a length dl of the strip. The magnetic flux through a single turn is $\Phi_0 = \mathbf{B} \cdot \mathbf{S}$, and that through dN turns is

$$d\Phi = \Phi_0 \, dN = N'S \, d\mathbf{l} \cdot \mathbf{B} \tag{4.35}$$

The total flux through all the turns of the flexible solenoid is thus

$$\Phi = \oint_C d\Phi = N'S \oint_C \mathbf{B} \cdot d\mathbf{l} = \mu_0 N'S \cdot i(t) \tag{4.36}$$

according to Ampère's law applied to the contour C along the strip. The induced *emf* in the winding is $e = -d\Phi/dt$, so that, finally, the expression for the amplitude of $i(t)$ reads $I_m = V_m/\mu_0 N'S\omega$.

4.4.5. Problems in Measurement of AC Voltage

As an example of the measurement of ac voltage, consider a straight copper wire of radius $a = 1\,\text{mm}$ with a sinusoidal current $i(t) = 1\cos\omega t\,\text{A}$. A voltmeter is connected between points 1 and 2, with leads as shown in Fig. 4.12. If $b = 50\,\text{cm}$ and $c = 20\,\text{cm}$, we will evaluate the voltage measured by the voltmeter for (1) $\omega = 314\,\text{rad/s}$, (2) $\omega = 10^4\,\text{rad/s}$,

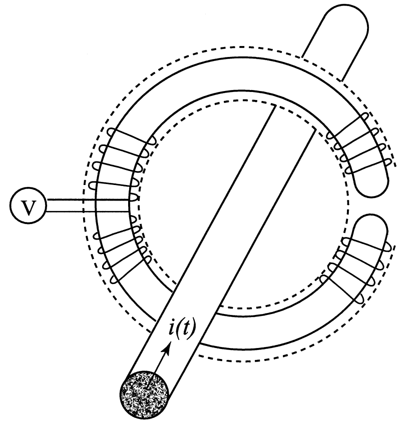


Figure 4.11 A method for measuring ac current in a conductor without inserting an ammeter into the circuit.

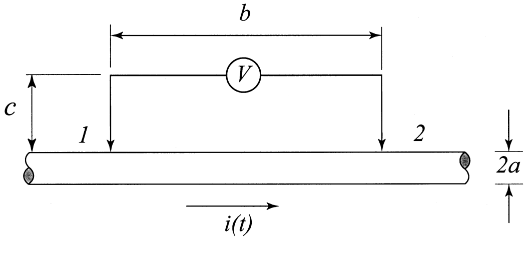


Figure 4.12 Measurement of ac voltage.

and (3) $\omega = 10^6$ rad/s. We assume that the resistance of the copper conductor per unit length, R', is approximately as that for a dc current (which actually is *not* the case, due to skin effect). We will evaluate for the three cases the potential difference $V_1 - V_2 = R'b \cdot i(t)$ and the voltage induced in the leads of the voltmeter.

The voltage measured by the voltmeter (i.e., the voltage between its *very* terminals, and *not* between points 1 and 2) is

$$V_{\text{voltmeter}} = (V_1 - V_2) - e = R'bi - e \tag{4.37}$$

where $R' = 1/\sigma_{\text{Cu}}\pi a^2$, ($\sigma_{\text{Cu}} = 5.7 \times 10^7$ S/m), and e is the induced *emf* in the rectangular contour containing the voltmeter and the wire segment between points 1 and 2 (we neglect the size of the voltmeter). This *emf* is approximately given by

$$e = \frac{\mu_0 b}{2\pi}\frac{di}{dt}\ln\frac{c+a}{a} \tag{4.38}$$

The rms value of the potential difference $(V_1 - V_2)$ amounts to 1.97 mV, and does not depend on frequency. The difference between this potential difference and the voltage indicated by the voltmeter for the three specified frequencies is (1) 117.8 μV, (2) 3.74 μV, and (3) 3.74 V. This difference represents an error in measuring the potential difference using the voltmeter with such leads. We see that in case (2) the relative error is as large as 189%, and that in case (3) such a measurement is meaningless.

4.4.6. Readout of Information Stored on a Magnetic Disk

When a magnetized disk with small permanent magnets (created in the writing process) moves in the vicinity of the air gap of a magnetic head, it will produce time-variable flux in the head magnetic core and the read-and-write coil wound around the core. As a result, an *emf* will be induced in the coil reflecting the magnetization of the disk, in the form of positive and negative pulses. This is sketched in Fig. 4.13. (For the description of the writing process and a sketch of the magnetic head, please see Chapter 3.)

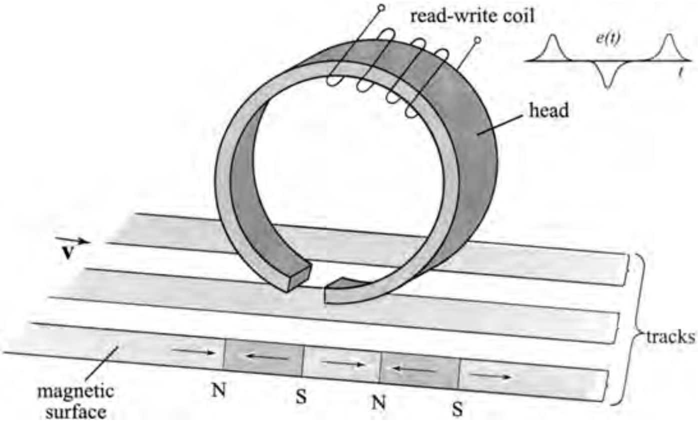


Figure 4.13 A hard disk magnetized through the write process induces a emf in the read process: when the recorded magnetic domains change from south to north pole or vice versa, a voltage pulse proportional to the remanent magnetic flux density is produced. The pulse can be negative or positive.

Historical Note: Magnetic Core Memories

In the readout process of magnetic core memories described in Chapter 3, a negative pulse is passed through circuit 1 in Fig. 4.14. If the core is magnetized to a "1" (positive remanent magnetic flux density of the hysteresis curve), the negative current pulse brings it to the negative tip of the hysteresis loop, and after the pulse is over, the core will remain at the negative remanent flux density point. If, on the other hand, the core is at "0" (negative remanent magnetic flux density of the hysteresis curve), the negative current pulse will make the point go to the negative tip of the hysteresis loop and again end at the point where it started.

While the above described process is occurring, an *emf* is induced in circuit 2, resulting in one of the two possible readings shown in Fig. 4.14. These two pulses correspond to a "1" and a "0." The speed at which this process occurs is about 0.5–5 μs.

4.4.7. Transformers

A transformer is a magnetic circuit with (usually) two windings, the "primary" and the "secondary," on a common ferromagnetic core, Fig. 4.15. When an ac voltage is applied to the primary coil, the magnetic flux through the core is the same at the secondary and induces a voltage at the open ends of the secondary winding. Ampère's law for this circuit can be written as

$$N_1 i_1 - N_2 i_2 = HL \tag{4.39}$$

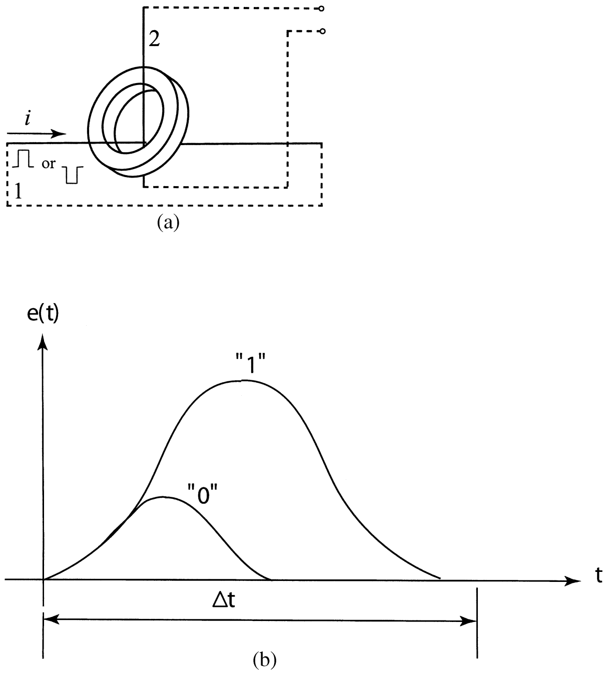


Figure 4.14 (a) A magnetic core memory bit and (b) induced voltage pulses during the readout process.

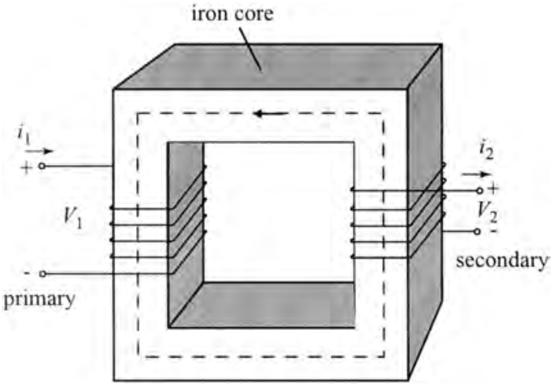


Figure 4.15 Sketch of a transformer with the primary and secondary windings wound on a ferromagnetic core.

where N_1 and N_2 are the numbers of the primary and secondary windings, i_1 and i_2 are the currents in the primary and secondary coils when a generator is connected to the primary and a load to the secondary, H is the magnetic field in the core, and L is the effective length of the core. Since $H = B/\mu$ and, for an *ideal* core, $\mu \to \infty$, both B and H in the ideal core are zero (otherwise the magnetic energy in the core would be infinite). Therefore, for an ideal transformer

$$\frac{i_1}{i_2} = \frac{N_2}{N_1} \tag{4.40}$$

This is the relationship between the primary and secondary currents in an *ideal transformer*. For good ferromagnetic cores, the permeability is high enough that this is a good approximation.

From the definition of magnetic flux, the flux through the core is proportional to the number of windings in the primary. From Faraday's law, the induced *emf* in the secondary is proportional to the number of times the magnetic flux in the core passes through the surface of the secondary windings, i.e., to N_2. (This is even more evident if one keeps in mind that the lines of induced electric field produced by the primary current encircle the core, i.e., going along the secondary winding the integral of the induced electric field is N_2 times that for a single turn.) Therefore, the following can be written for the voltages across the primary and secondary windings:

$$\frac{v_1}{v_2} = \frac{N_1}{N_2} \tag{4.41}$$

Assume that the secondary winding of an ideal transformer is connected to a resistor of resistance R_2. What is the resistance seen from the primary terminals? From Eqs. (4.40) and (4.41),

$$R_1 = \frac{v_1}{i_1} = R_2 \left(\frac{N_1}{N_2}\right)^2 \tag{4.42}$$

Of course, if the primary voltage is sinusoidal, complex notation can be used, and the resistances R_1 and R_2 can be replaced by complex impedances Z_1 and Z_2. Finally, if we assume that in an ideal transformer there are no losses, all of the power delivered to the primary can be delivered to a load connected to the secondary. Note that the voltage in both windings is *distributed*, so that there can exist a relatively high voltage between two adjacent layers of turns. This would be irrelevant if, with increasing frequency, this voltage would not result in increasing capacitive currents and deteriorated transformer performance, i.e., basic transformer equations become progressively less accurate with increasing frequency. The frequency at which a transformer becomes useless depends on many factors and cannot be predicted theoretically.

4.4.8. Induced EMF in Loop Antennas

An electromagnetic plane wave is a traveling field consisting of a magnetic and electric field. The magnetic and electric field vectors are mutually perpendicular and perpendicular to the direction of propagation of the wave. The electric field of the wave is, in fact, an induced (only traveling) electric field. Thus, when a small closed wire loop is placed in the field of the wave, there will be an *emf* induced in the loop. *Small* in this context means much smaller than the wave wavelength, and such a loop is referred to as a *loop antenna.* The maximal *emf* is induced if the plane of the loop is perpendicular to the magnetic field of the wave. For a magnetic field of the wave of root-mean-square (rms) value H, a wave frequency f, and loop area (normal to the magnetic field vector) S, the rms value of the *emf* induced in the loop is

$$emf = \left|\frac{d\Phi}{dt}\right| = 2\pi\mu_0 f \cdot H \cdot S \tag{4.43}$$

4.5. EVALUATION OF MUTUAL AND SELF-INDUCTANCE

The simplest method for evaluating mutual and self-inductance is using Eqs. (4.15) and (4.17), provided that the magnetic flux through one of the contours can be calculated. This is possible in some relatively simple, but practical cases. Some of these are presented below.

4.5.1. Examples of Mutual Inductance Calculations

Mutual Inductance Between a Toroidal Coil and a Wire Loop Encircling the Toroid

In order to find the mutual inductance between a contour C_1 and a toroidal coil C_2 with N turns, Fig. 4.16, determining L_{12} is not at all obvious, because the surface of a toroidal coil is complicated. However, $L_{21} = \Phi_{21}/I_2$ is quite simple to find. The flux $d\Phi$ through the surface $dS = h \cdot dr$ in the figure is given by

$$d\Phi_{21}(r) = B(r) \cdot dS = \frac{\mu_0 N I_2}{2\pi r} h \cdot dr \tag{4.44}$$

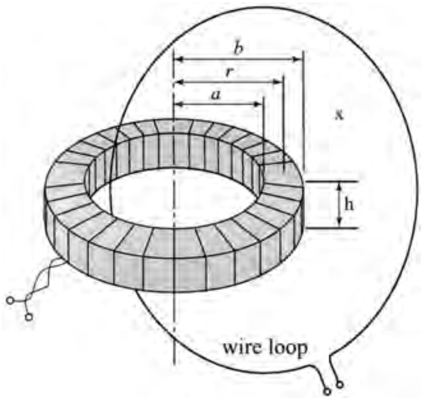


Figure 4.16 A toroidal coil and a single wire loop encircling the toroid.

so that the total flux through C_1, equal to the flux through the cross section of the torus, is

$$\Phi_{21} = \frac{\mu_0 N I_2 h}{2\pi} \int_a^b \frac{dr}{r} = \frac{\mu_0 N I_2 h}{2\pi} \ln\frac{b}{a} \qquad \text{or} \qquad L_{12} = L_{21} = \frac{\mu_0 N h}{2\pi} \ln\frac{b}{a} \tag{4.45}$$

Note that mutual inductance in this case does not depend at all on the shape of the wire loop. Also, if a larger mutual inductance (and thus larger induced *emf*) is required, the loop can simply be wound two or more times around the toroid, to obtain two or more times larger inductance. This is the principle of operation of transformers.

Mutual Inductance Between Two Toroidal Coils

As another example, let us find the mutual inductance between two toroidal coils tightly wound one on top of the other on a core of the form shown in Fig. 4.16. Assume that one coil has N_1 turns and the other N_2 turns. If a current I_2 flows through coil 2, the flux through coil 1 is just N_1 times the flux Φ_{21} from the preceding example, where N should be substituted by N_2. So

$$L_{12} = L_{21} = \frac{\mu_0 N_1 N_2 h}{2\pi} \ln\frac{b}{a} \tag{4.46}$$

Mutual Inductance of Two Thin Coils

Let the mutual inductance of two simple loops be L_{12}. If we replace the two loops by two very thin coils of the same shapes, with N_1 and N_2 turns of very thin wire, the mutual inductance becomes $N_1 N_2 L_{12}$, which is obtained directly from the induced electric field. Similarly, if a thin coil is made of N turns of very thin wire pressed tightly together, its self-inductance is N^2 times that of a single turn of wire.

Mutual Inductance of Two Crossed Two-wire Lines

A two-wire line crosses another two-wire line at a distance *d*. The two lines are normal. Keeping in mind Eq. (4.3) for the induced electric field, it is easily concluded that their mutual inductance is zero.

4.5.2. Inductors and Examples of Self-inductance Calculations

Self-inductance of a Toroidal Coil

Consider again the toroidal coil in Fig. 4.16. If the coil has N turns, its self-inductance is obtained directly from Eq. (4.45): This flux exists through all the N turns of the coil, so that the flux the coil produces through itself is simply N times that in Eq. (4.45). The self-inductance of the coil in Fig. 4.16 is therefore

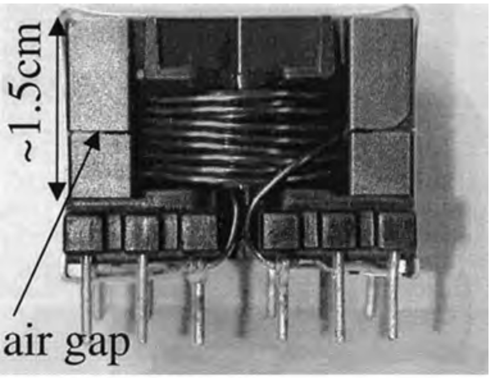
$$L = \frac{\mu_0 N^2 h}{2\pi} \ln \frac{b}{a} \tag{4.47}$$

Self-inductance of a Thin Solenoid

A thin solenoid of length b and cross-sectional area S is situated in air and has N tightly wound turns of thin wire. Neglecting edge effects, the self-inductance of the solenoid is given by

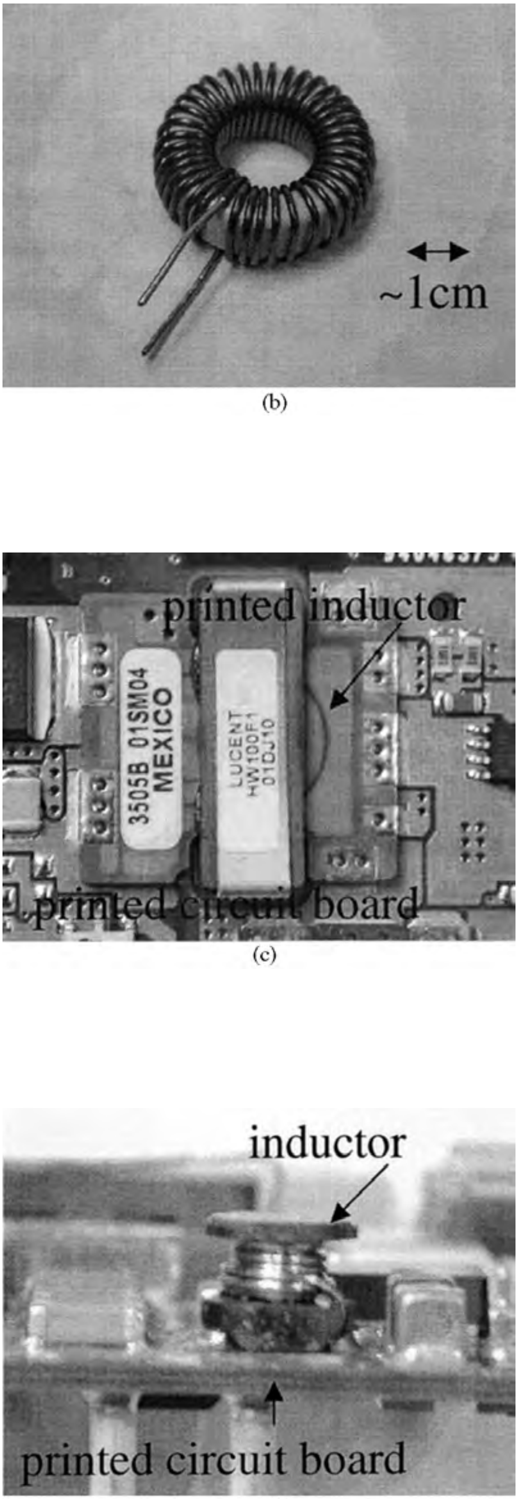
$$L = \frac{\mu_0 N^2 S}{b} \tag{4.48}$$

However, in a practical inductor, there exists mutual capacitance between the windings, resulting in a parallel resonant equivalent circuit for the inductor. At low frequencies, the capacitor is an open circuit, but as frequency increases, the reactance of the capacitor starts dominating. At resonance, the parallel resonant circuit is an open, and beyond that frequency, the inductor behaves like a capacitor. To increase the valid operating range for inductors, the windings can be made smaller, but that limits the current handling capability. Figure 4.17 shows some examples of inductor implementations.



(a)

Figure 4.17 (a) A low-frequency inductor, with $L \approx 1\,\text{mH}$, wound on a core with an air gap, (c) an inductor with a permalloy core with $L \approx 0.1\,\text{mH}$, (c) a printed inductor surrounded by a ferromagnetic core with $L \approx 10\,\mu\text{H}$, (d) small higher frequency inductors with $L \approx 1\,\mu\text{H}$, (e) a chip inductor for surface-mount circuits up to a few hundred MHz with $L \approx 0.1\,\mu\text{H}$, and (f) a micromachined spiral inductor with $L \approx 10$–$20\,\text{nH}$ and a cutoff frequency of 30 GHz ($Q > 50$).

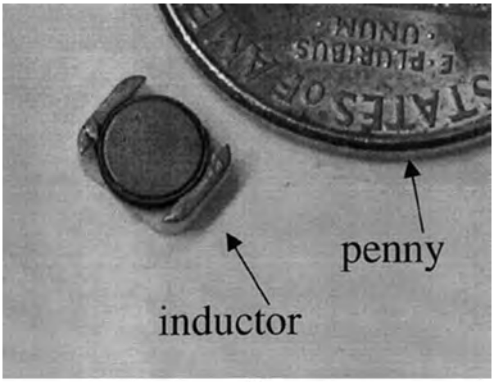


(b)

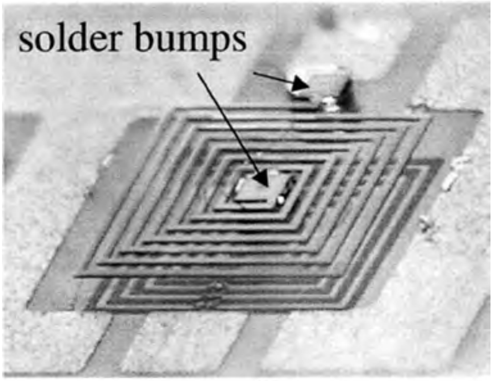
(c)

(d)

Figure 4.17 Continued.



(e)



(f)

Figure 4.17 Continued.

The inductor in Fig. 4.17a is wound on a ferromagnetic core, which has the effect of increasing the inductance by μ_r. The small air gap in the core increases the current handling capability of the inductor. To see why this is true, the magnetic circuit equations (3.31)–(3.33) can be applied, using the approximation that an effective μ_r can be defined for the ferromagnetic. Let the effective length (perimeter of the centerline) of the core be L and that of the air gap $L_0 \ll L$. Then

$$NI = R_{m0}\Phi_0 + R_m\Phi \cong (R_{m0} + R_m)\Phi = \frac{1}{\mu_0} B\left(L_0 + \frac{L}{\mu_r}\right) \tag{4.49}$$

where N is the number of windings of the inductor, Φ is the magnetic flux through the core and, approximately, through the air gap, and B is the corresponding magnetic flux density. For an inductor wound on a core without an air gap, the above expression becomes

$$NI' = R_m\Phi = \frac{L}{\mu_0\mu_r} B \tag{4.50}$$

where I' is the current with no air gap in the core, assuming the number of windings, the flux and the dimensions of the core are kept the same. The ratio of the two currents is

$$\frac{I}{I'} = \frac{L_0 + L/\mu_r}{L/\mu_r} = 1 + \frac{\mu_r L_0}{L} \tag{4.51}$$

It can be seen that for the same windings, flux, size, and core material, a higher current in the windings can be used if an air gap is present. Typical relative permeabilities of ferromagnetic cores are in the several thousands (see Table 3.1), while the air gap length is controlled by insulator (usually mylar) sheets of variable thickness, on the order of a fraction of a millimeter. Almost all the magnetic energy is contained in the air gap, since the magnetic field in the gap is μ_r times larger than in the core. The gap therefore enables both larger inductance values and higher current handling, as long as the ferromagnetic does not saturate.

In moly permalloy materials (see Table 3.1), the relative permeability is smaller than in pure ferromagnetics because the material is made with distributed air gaps (bubbles). An inductor with a permalloy core is shown in Fig. 4.17b. The distributed air regions increase the magnetic energy and therefore the current handling capability. Since the effective μ_r is lower, the inductance values are not as high (in the μH range).

Figure 4.17c shows a printed spiral inductor, whose value is increased by wrapping a core around the printed-circuit board. Such inductors can have values on the order of 0.1 mH. Figure 4.15d shows a small high-frequency inductor (several hundred MHz) with a value on the order of tens of microhenry, and Fig. 4.17e shows a surface-mount inductor with values on the order of 0.1 μH and cutoff frequency in the few hundred megahertz range. Figure 4.17f shows a miniature high-frequency micromachined (MEM) inductor suspended in air in order to reduce capacitance due to the presence of the dielectric, resulting in values of inductance on the order of 10–20 nH with a usable frequency range above 20 GHz [21]. At high frequencies, due to the skin effect, the loss in the inductor becomes large, and values of the Q factor are in the range of Q > 0.

Self-inductance of a Coaxial Cable

Let us find the external self-inductance per unit length of a coaxial cable. We first need to figure out through which surface to find the flux. If the cable is connected to a generator at one end and to a load at the other, the current flows "in" through the inner conductor and flows back through the outer conductor. The flux through such a contour, for a cable of length h, is the flux through the rectangular surface in Fig. 4.18,

$$\Phi = \int_a^b B(r)h \cdot dr = \frac{\mu_0 I h}{2\pi} \ln\frac{b}{a} \tag{4.52}$$

so that the external self-inductance per unit length of the cable is

$$L' = \frac{\mu_0}{2\pi} \ln\frac{b}{a} \tag{4.53}$$

As a numerical example, for $b/a = e = 2.71828\ldots$, $L' = 0.2\,\mu\text{H/m}$. For a common high-frequency coaxial cable RG-55/U cable, $a = 0.5$ mm, $b = 2.95$ mm, and the inductance per unit length is around 3.55 nH/cm.

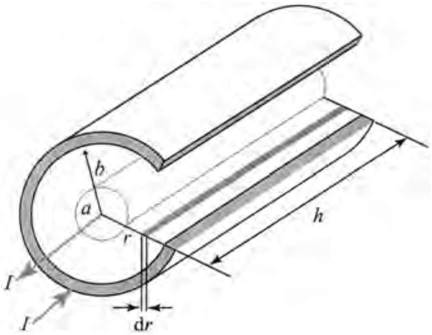


Figure 4.18 Calculating the self-inductance of a coaxial cable.

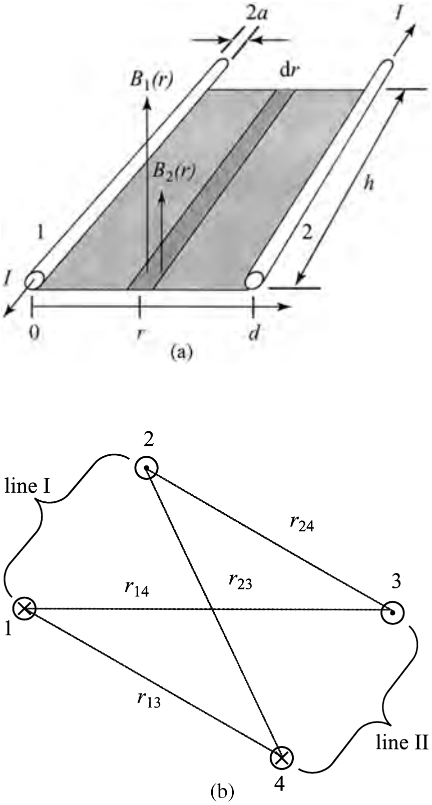


Figure 4.19 (a) Calculating the self-inductance of a thin two-wire line and (b) the mutual inductance between two parallel two-wire lines.

External Self-inductance of a Thin Two-wire Line

A frequently used system for transmission of signals is a thin two-wire line, Fig. 4.19a. Its inductance per unit length is determined as follows. We can imagine that the line is actually a very long rectangular contour (closed with a load at one end, and a generator at the other end), and that we are looking at only one part of it, sketched in the figure.

At a distance r from conductor 1, the current in it produces a magnetic flux density of intensity $B_1(r) = \mu_0 I/2\pi r$, and the current in conductor 2 a magnetic flux density $B_2(r) = \mu_0 I/2\pi(d - r)$. The total flux through a strip of width dr and length h shown in the figure is therefore

$$\Phi = \int_a^{d-a} [B_1(r) + B_2(r)] \cdot h \cdot dr \cong \frac{\mu_0 I h}{\pi} \ln \frac{d}{a} \tag{4.54}$$

since $d \gg a$. The inductance per unit length of the two-wire line is therefore

$$L' = \frac{\mu_0}{\pi} \ln \frac{d}{a} \tag{4.55}$$

As a numerical example, for $d/a = 200$, $L' = 2.12\,\mu\text{H/m}$. We have only calculated the flux through the surface outside of the conductors. The expression for L' above is therefore called the *external self-inductance* of the line. There is also an *internal self-inductance*, due to the flux through the wires themselves (see the example below, Sec. 4.6.3).

Bifilar Coil

To obtain a resistive wire with the smallest self-inductance possible, the wire is bent sharply in the middle and the two mutually insulated halves are pressed tightly together. This results in the smallest external flux possible and, consequently, in the smallest self-inductance. If such a bent wire is wound into a winding, a *bifilar coil* is obtained.

4.5.3. More on Mutual Inductance

Mutual Inductance Between Two Parallel Two-wire Lines

Mutual inductance per unit length of two two-wire lines running parallel to each other, shown in the cross section in Fig. 4.19b, can be obtained by calculating the magnetic flux per unit length due to current in one line through the other. For the reference directions of the two lines the indicated result is

$$L'_{\text{I,II}} = \frac{\Phi'_{\text{I,II}}}{I_{\text{I}}} = \frac{\mu_0}{2\pi} \ln \frac{r_{14} r_{23}}{r_{13} r_{24}} \tag{4.56}$$

Self-inductance and Mutual Inductance of Two Windings over a Toroidal Core

A thin toroidal core of permeability μ, mean radius R, and cross-sectional area S is densely wound with two coils of thin wire, with N_1 and N_2 turns, respectively. The windings are wound one over the other. The self- and mutual inductances of the coils are

$$L_1 = \frac{\mu N_1^2 S}{2\pi R}, \qquad L_2 = \frac{\mu N_2^2 S}{2\pi R}, \qquad L_{12} = \frac{\mu N_1 N_2 S}{2\pi R} \tag{4.57}$$

so that the coupling coefficient $k = 1$.

4.5.4. Neumann's Formula for Inductance Calculations

Neumann's Formula for Mutual Inductance of Two-wire Loops

Starting from the induced electric field due to a thin-wire loop, it is possible to derive a general formula for two thin-wire loops in a homogeneous medium, Fig. 4.20, known as *Neumann's formula*. With reference to Fig. 4.20, it is of the form

$$L_{12} = \frac{\mu_0}{4\pi}\oint_{C_2}\oint_{C_1}\frac{d\mathbf{l}_1 \cdot d\mathbf{l}_2}{r} \tag{4.58}$$

Note that L_{21} would have the same form, except that the order of integration and the dot product of current elements would exchange places. Since this does not affect the result, we conclude that $L_{12} = L_{21}$. Note also that explicit evaluation of the dual integral in Eq. (4.58) can be performed only in rare instances, but it can always be integrated numerically with ease.

Flat Multiconductor Cable

As an example of application of Neumann's formula, consider n narrow coplanar strips that run parallel to each another over a distance d, Fig. 4.21. This is a model of a flat

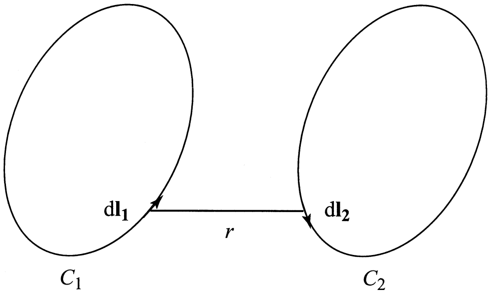


Figure 4.20 Two loops made out of thin wire.

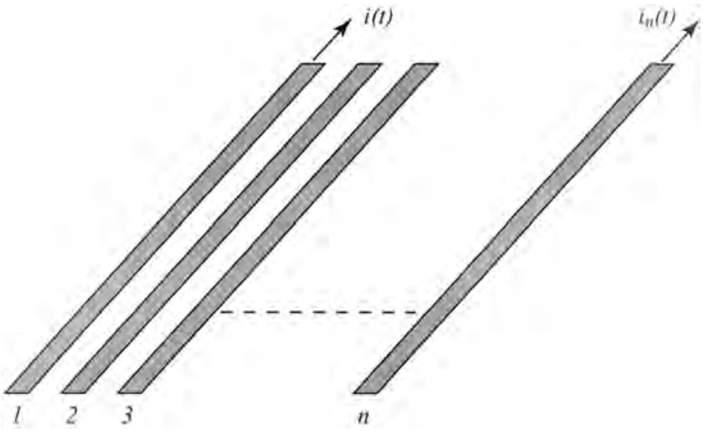


Figure 4.21 A flat multiconductor cable (transmission line).

multiconductor cable (transmission line), such as the ones used to connect a printer to a computer. Let the currents in the strips be $i_1(t), i_2(t), \ldots, i_n(t)$. We wish to determine the *emf* induced, for example, in strip no. 1 by the time-varying current in all the other strips. Although, as explained, this type of coupling is usually referred to as *magnetic coupling*, it is actually an example of mutual coupling by means of an induced electric field.

Note that we need not have closed loops in the Neumann formula—what matters is the induced electric field and the length of the wire in which we determine the induced *emf* (i.e., the line integral of the induced field). The total *emf* in conductor no. 1 induced by the other conductors running parallel to it for a distance d is given by

$$e_1(t) = -\frac{\mu_0}{4\pi}\sum_{j=2}^{n}\frac{di_j(t)}{dt}\left(\int_0^d\int_0^d \frac{dl_1 dl_j}{r_{1j}}\right) \tag{4.59}$$

The elements $dl_1, dl_2, \ldots, dl_n$ are along the center lines of the strips. The integrals can be evaluated explicitly using tables of integrals. Note that the reference direction of currents in all the strips is assumed to be the same.

Neumann's Formula for External Self-inductance of a Wire Loop

Neumann's formula in Eq. (4.58) can be modified to enable the evaluation of external self-inductance. At first glance, one can just consider the case when the two contours, C_1 and C_2, in Neumann's formula overlap, and the self-inductance of a loop results. This is not so, however, because the integral becomes singular and divergent ($1/r$ is zero when elements dl_1 and dl_2 coincide).

To alleviate this problem, assume instead that one loop, e.g., $C = C_1$, is along the axis of the loop and the other, $C' = C_2$, is along the surface of the wire (Fig. 4.22). The distance between line elements of such two contours is never zero, and the integral becomes convergent. The flux computed in this case is the flux that a line current along the loop axis produces through a contour on the wire surface, so this is precisely the external loop inductance. The Neumann formula for the external inductance of a loop is thus

$$L = \frac{\mu_0}{4\pi}\oint_C\oint_{C'}\frac{d\mathbf{l}\cdot d\mathbf{l}'}{r} \tag{4.60}$$

As with Eq. (4.58), it is possible to integrate the dual integral in this equation explicitly only rarely, but it can always be integrated numerically.

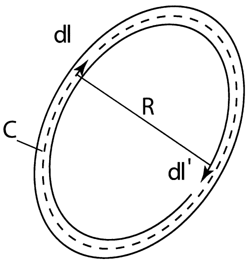


Figure 4.22 A wire loop and two possible contours of integration in the Neumann formula for self-inductance.

4.6. ENERGY AND FORCES IN THE MAGNETIC FIELD: IMPLICATIONS AND APPLICATIONS

4.6.1. Magnetic Energy of Two Magnetically Coupled Contours

In the case of two contours ($n = 2$), Eqs. (4.21) and (4.22) for the magnetic energy of n contours become

$$W_m = \frac{1}{2}(I_1\Phi_1 + I_2\Phi_2) \tag{4.61}$$

and

$$W_m = \frac{1}{2}L_{11}I_1^2 + \frac{1}{2}L_{22}I_2^2 + L_{12}I_1I_2 \tag{4.62}$$

This energy can be smaller or larger than the sum of energies of the two contours when isolated, since L_{12} can be positive or negative.

4.6.2. Losses in Ferromagnetic Materials Due to Hysteresis and Eddy Currents

Let us observe what happens to energy needed to maintain a sinusoidal magnetic field in a piece of ferromagnetic material. The hysteresis curve of the material is shown in Fig. 4.23, and the arrows show the direction in which the operating point is moving in the course of time. According to Eq. (4.24), the energy density that needs to be spent at a point where the magnetic field is H, in order to change the magnetic flux density by dB, is equal to $H\,dB$. In the diagram in Fig. 4.23, this is proportional to the area of the small shaded rectangle. So, the integral of $H\,dB$ is proportional to the sum of all such rectangles as the point moves around the hysteresis curve.

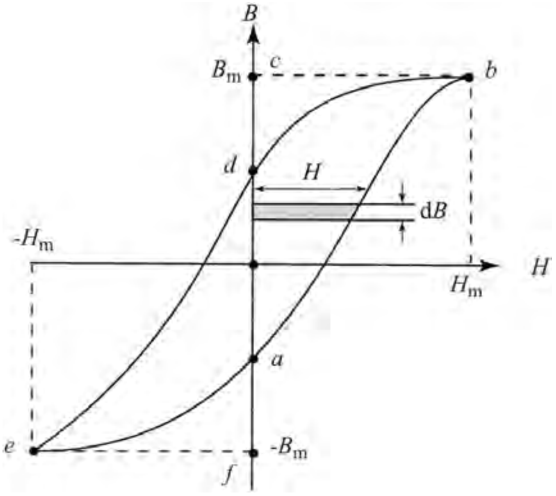


Figure 4.23 Hysteresis curve of a ferromagnetic material.

Starting from point *a* in Fig. 4.23 moving to point *b*, the magnetic field *H* is positive. The increase *dB* is also positive, so *H.dB* is positive, and the energy density needed to move from point *a* to *b* is proportional to the area of the curved triangle *abc* in the figure. From *b* to *d*, *H* is positive, but *B* is decreasing, so that *dB* is negative. Therefore, the product *H.dB* is negative, which means that in this region the energy *used up* on maintaining the field is negative. This in turn means that this portion of the energy is returned back from the field to the sources. The density of this returned energy is proportional to the area of the curved triangle *bdc*. From *d* to *e*, the product *H.dB* is positive, so this energy is spent on maintaining the field, and from *e* to *a*, the product is negative, so this energy is returned to the sources. Therefore, we come to the conclusion that only the energy density proportional to the area of the curved triangles *bcd* and *efa* is returned to the sources. All the rest, which is proportional to the area formed by the hysteresis loop, is lost to heat in the ferromagnetic material. These losses are known as *hysteresis losses*. If the frequency of the field is *f*, the operating point circumscribes the loop *f* times per second. Consequently, *hysteresis losses are proportional to frequency* (and to the volume of the ferromagnetic material if the field is uniform).

If the ferromagnetic material is conductive, there are also eddy-current losses, proportional to the square of frequency. As an example, consider a solenoid with a ferromagnetic core made of thin, mutually insulated sheets. To estimate the eddy-current and hysteresis losses, the *total* power losses are measured at two frequencies, f_1 and f_2, for the same amplitude of the magnetic flux density. The total power losses are found to be P_1 and P_2, respectively. Knowing that hysteresis losses are proportional to frequency, and eddy-current losses to the square of frequency, it is possible to separately determine these losses as follows. First the total losses can be expressed as

$$P = P_{\text{total losses}} = P_{\text{hysterisis losses}} + P_{\text{eddy-current losses}} = Af + Bf^2 \tag{4.63}$$

where A and B are constants. Consequently,

$$P_1 = Af_1 + Bf_1^2 \qquad \text{and} \qquad P_2 = Af_2 + Bf_2^2 \tag{4.64}$$

from which

$$A = \frac{P_1 f_2^2 - P_2 f_1^2}{f_1 f_2 (f_2 - f_1)} \tag{4.65}$$

Ferrite Anechoic Chambers for EMC/EMI Testing

For testing electromagnetic compatibility and interference over a broad frequency range, dimensions of absorber material for adequately low reflections would be very large and impractical. Anechoic chambers made for this purpose have walls made of ferrite material, with very high magnetic losses in the megahertz and gigahertz frequency range. As a result of the high losses, the walls can be made much thinner, and for increased absorption over a broader bandwidth the ferrite tiles are sometimes backed with a dielectric layer of dielectric backed with metal.

4.6.3. Internal Inductance of a Straight Wire at Low Frequencies

The energy of a wire with a current i is distributed outside the wire, as well as inside the wire, since there is a magnetic field both outside and inside the wire. From the energy expression $W_m = Li^2/2$ for a single current contour, we can write

$$L_{\text{internal}} = \frac{2W_{m\ \text{inside conductor}}}{i^2} \qquad \text{and} \qquad L_{\text{external}} = \frac{2W_{m\ \text{outside conductor}}}{i^2} \tag{4.66}$$

Consider a long straight wire of circular cross section and permeability μ. If the current in the wire is assumed to be distributed uniformly (or very nearly so, i.e., we consider low frequencies), according to Ampère's law, the magnetic field inside the wire is equal to $H(r) = Ir/2\pi a^2$. Using Eqs. (4.26) and (4.66), we find that the internal inductance of the wire per unit length is given by

$$L'_{\text{internal}} = \frac{\mu_0}{8\pi} \tag{4.67}$$

Note that the internal inductance does not depend on the radius of the wire.

4.6.4. Total Inductance of a Thin Two-wire Line at Low Frequencies

The total self-inductance per unit length of a thin two-wire line with wires made of a material with permeability μ, radius of the wires a, and distance between the wire axes $d \gg a$ is the sum of its external inductance in Eq. (4.55) and the internal *inductance of both wires*:

$$L' = L'_{\text{internal}} + L'_{\text{external}} = \frac{\mu_0}{\pi}\ln\frac{d}{a} + 2\frac{\mu_0}{8\pi} \tag{4.68}$$

As a numerical example, if $\mu = \mu_0$ and $d/a = 100$, we get $L'_{\text{external}} = 1.84\,\mu\text{H/m}$ and $L'_{\text{internal}} = 0.1\,\mu\text{H/m}$. In this example, the external inductance is much larger than the internal inductance. This is usually the case.

4.6.5. Force of an Electromagnet

As an example of the force formula in Eq. (4.27), the attractive force of an electromagnet, sketched in Fig. 4.24, is evaluated below. The electromagnet is in the shape of a horseshoe, and its magnetic force is lifting a weight W, shown in the figure. This is a magnetic circuit. Let us assume that when the weight W moves by a small amount dx upward, the flux in the magnetic circuit does not change. That means that when the weight is moved upward, the only change in magnetic energy is the *reduction* in energy contained in the two air gaps, due to their decreased length. This energy reduction is

$$-\frac{dW_m}{dx} = \frac{1}{2}\frac{B^2}{\mu_0}2S \cdot dx \tag{4.69}$$

and the force is now equal to

$$F_x = \frac{1}{2}\frac{B^2}{\mu_0}2S = \frac{\Phi^2}{\mu_0}S \tag{4.70}$$

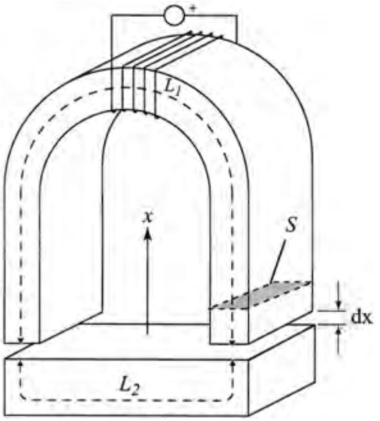


Figure 4.24 Sketch of an electromagnet lifting a ferromagnetic weight.

As a numerical example, let $B = 1\,\mathrm{T}$, and $S = 1000\,\mathrm{cm}^2$. For this case, $F_x = 7.96 \times 10^4\,\mathrm{N}$, which means that this electromagnet can lift a weight of about 8 tons! Such electromagnets are used, for example, in cranes for lifting large pieces of iron.

4.6.6. Comparison of Electric and Magnetic Pressure

The expression for the pressure of magnetic forces on boundary surfaces between materials of different magnetic properties can be obtained starting from Eqs. (4.27) and (4.28). For two magnetic media of permeabilities μ_1 and μ_2, the pressure on the interface, assumed to be directed into medium 1, is given by

$$p = \frac{1}{2}(\mu_2 - \mu_1)\left(H^2_{\tan g} + \frac{B^2_{\mathrm{norm}}}{\mu_1 \mu_2}\right) \tag{4.71}$$

with the reference direction of pressure into medium 1. We know that magnetic flux density of about 1 T is quite large and not easily attainable. Therefore, for $B_{\mathrm{norm}} = 1\,\mathrm{T}$, $H_{\tan g} = 0$, and $\mu_2 \gg \mu_1 = \mu_0$, the largest magnetic pressure that can be obtained is on the order of

$$p_{m,\max} \approx 400{,}000\,\frac{\mathrm{N}}{\mathrm{m}^2}$$

The electric pressure on a metallic conductor in vacuum is given by the expression $p_{e,\max} = (1/2)\varepsilon_0 E^2$. The electric strength of air is about $3 \times 10^6\,\mathrm{V/m}$. This means that the largest electric pressure in air is approximately

$$p_{e,\max} = 0.5(8.86 \times 10^{-12})\,(3 \times 10^6)^2 \simeq 40\,\mathrm{N/m}^2$$

Consequently, the ratio of the maximal magnetic and maximal electric pressure is approximately

$$\frac{p_{m,\max}}{p_{e,\max}} = 10{,}000$$

This is an interesting and important conclusion. Although "electric" and "magnetic" versions of almost any device can be designed using electric and magnetic forces, the magnetic version will require much smaller space for the same amount of power. To get an idea for the order of magnitude of the magnetic and electric pressure, note that a typical car-tire pressure is around 200 kPa = 200,000 N/m^2 (or 30 psi).

4.6.7. High-frequency Resistance and Internal Inductance of a Wire

It is easy to understand from Ohm's law that a metal wire has a resistance at dc given by the resistance for a uniform resistor. As the frequency increases, this resistance changes due to the skin effect, i.e., the redistribution of current across the cross section of the conductor. For a cylindrical wire of radius a, the associated resistance per unit length is given by $R' = R_S/2\pi a$, where $R_S = \sqrt{\omega\mu/2\sigma}$ is the surface resistance of the conductor with conductivity σ at an angular frequency ω and is obtained from assuming the current flows through a cross section determined by skin depth.

At high frequencies, a wire also has magnetic energy stored nonuniformly inside it, and this is associated with internal inductance of the wire per unit length. It can be shows that the reactive power at high frequencies inside a conductor is equal to the power of Joule losses due to the wire surface resistance. The power of heat loss is given by $P'_{\text{heat}} = R_S I^2/2\pi a$, and the inductance per unit length of a cylindrical wire at high frequencies is found from $R_S I^2/2\pi a = \omega L'_{\text{int}} I^2$. Therefore, the resistance and internal inductance per unit length of a cylindrical metal wire at a frequency ω are given by

$$R' = \frac{1}{2\pi a}\sqrt{\frac{\omega\mu}{2\sigma}} \qquad \text{and} \qquad L'_{\text{int}} = \frac{1}{2\pi a\omega}\sqrt{\frac{\omega\mu}{2\sigma}} \tag{4.72}$$

This frequency-variable internal inductance should be added to the external inductance when, e.g., calculating the characteristic impedance of cables at high frequencies.

4.7. SOME INTERESTING EXAMPLES OF ELECTROMAGNETIC INDUCTION

In this section a few interesting examples that the authors encountered in practice, and that they feel might be useful to the reader, are described. In particular, a commonly encountered case is that of signal cross-talk due to a current-carrying wire that passes through a hole in a metal casing (Sec. 4.7.2).

4.7.1. Mutual Inductance Between Monophase Cables Laid on the Bottom of the Sea

Assume we have three single-phase 60-Hz power cables laid at the bottom of the sea, for example, to supply electric power to an island. The cables are spaced by a few hundred meters and are parallel to each other. (Three distant single-phase instead of one three-phase cable are often used for safety reasons: if a ship accidentally pulls and breaks one cable with an anchor, two are left. In addition, usually a spare single-phase cable is laid to enable quick replacement of a damaged one.) If the length of the cables is long (in practice, it can be many kilometers), we might expect very large mutual inductance between these cables, due to the huge loops they form, and, consequently, unbalanced currents in the three cables. The 60-Hz sea water skin depth, however, tells us that there will be practically no mutual inductance between the cables.

4.7.2. Cross Talk Due to Current in Wire Passing Through a Hole in a Metal Casing

An interesting and commonly encountered practical effect occurs when a single wire with a high-frequency current passes through a hole in a metal sheet (e.g., the side of a metal chassis). *High frequency* in this case means that skin depth should be much smaller than the sheet thickness. Consequently, the reasoning is valid, surprisingly, also for power frequencies (60 or 50 Hz) if the sheet is ferromagnetic and its thickness is on the order of 10 mm or greater.

If Ampère's law is applied to a contour encircling the hole so that it is further away from the sheet surface than the skin depth, the line integral of vector **H** is practically zero, because there is no magnetic field so deep in the sheet. This means that the total current encircled by the contour is practically zero, i.e., *that the current induced on the hole surface is practically the same as the current in the wire*. Of course, once it leaves the surface of the hole, this current continues to flow over both sheet surfaces, producing its own magnetic and induced electric field. Consequently, the signal carried by the current through the wire can be transmitted as described to undesirable places, causing unexpected cross talk.

4.7.3. Rough Calculation of Induced Voltages in a Human Body Due to Currents in Power Lines

There is often concern that fields radiated by power lines might be harmful for human health. It is interesting to do a calculation of the induced voltages in the human body that result from currents in power lines. There are two mechanisms by which a voltage can be produced in such a situation: that produced by the electric field, and that induced through electromagnetic induction due to magnetic field variations. This example shows a calculation of the two mechanisms on the example of the human head, assuming that it is the most important, and possibly the most sensitive part of a human body. Assume that a human head is a sphere with a radius of 10 cm and consisting mostly of salty water. For this example, the induced voltages in the head are calculated for the power lines being as close as 20 m from the human head, and they carry 100 A of unbalanced current, Fig. 4.25. For any other input information, the results can be easily scaled.

The magnetic flux density 20 m away from a wire with 100 A of current is $B = \mu_0 I / 2\pi r = 1\,\mu\text{T}$. (How large is this? The earth's dc magnetic field is on average 50 μT on the surface, and as a person moves in this field, some voltage will be induced, but humans are presumably adapted to this effect.) Faraday's law can be used to calculate the induced *emf* around the head due to the calculated value of *B*:

$$\oint_{\text{around head}} \mathbf{E} \cdot d\mathbf{l} = -\frac{\partial}{\partial t} \int_{\text{head cross section}} \mathbf{B} \cdot d\mathbf{S}$$

In complex notation, the above equation becomes $2\pi E_{\text{induced}} = -j\omega B\pi a^2$, where a is the radius of the head. From here, the value of the voltage due to the induced field

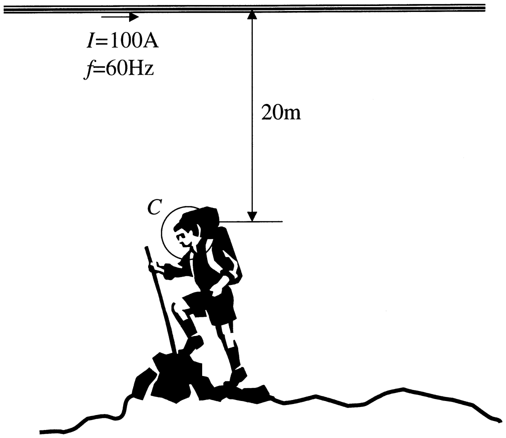


Figure 4.25 Calculating approximately the electromagnetic influence of a power line on the head of a human standing under it.

across a single 10-μm cell in our head is calculated to be about 33 pV for a power-line frequency of 60 Hz.

This is only one component of the effect of power lines on the human. The other is due to the electric field, which depends on the voltage of the power line. A reasonable value for the electric field close to the power line is around $E = 1\,\text{kV/m}$. Salt water has a resistivity ρ of about 1 Ωm, and to find the voltage across a single cell that can be added to the induced voltage, the following reasoning can be made. We first find the charge density σ produced on the head due to the high field, we then find the total charge Q by integration. Assuming a 60-Hz field frequency, this changing charge will produce a current I, and a corresponding current density J. The current density in the nonperfect conductor produces an ohmic voltage drop across a cell. The following equations describe this reasoning, assuming the head is perfectly spherical:

$$\sigma(\theta) = 3\varepsilon_0 E \cos\theta \qquad Q = \int_{\text{head}} \sigma\, dS = \int_0^{\pi/2} \sigma(\theta) 2\pi a \sin\theta \cdot d\theta = 3\pi\varepsilon_0 E a^2$$

$$I = 2\pi f Q = 0.315\,\mu\text{A} \qquad J = \frac{I}{\pi a^2} \qquad V = E \cdot 10\,\mu\text{m} = \rho J \cdot 10\,\mu\text{m} \simeq 100\,\text{pV}$$

Thus the total voltage across a cell in the human head due to a high-voltage line nearby is calculated to be about 133 pV. For comparison, normal neural impulses are much larger: they are spikes with around 100-mV amplitudes, frequency between 1 and 100 Hz, and duration of about a millisecond.

REFERENCES

1. Beaty, H.B.; Kirtley, J.L. Jr. *Electric Motor Handbook*; McGraw-Hill: New York, 1998.
2. Becker, R. *Electromagnetic Fields and Interactions*; Dover Publications: New York, 1982.
3. Bewley, L.V. *Flux Linkages and Electromagnetic Induction*; Dover Publications: New York, 1964.

4. Cheston, W.B. *Elementary Theory of Electromagnetic Fields*; John Wiley & Sons: New York, 1964.
5. Coren, R.L. *Basic Engineering Electromagnetics: An Applied Aprroach*; Prentice Hall: Upper Saddle River, NJ, 1989.
6. Coulson, C.A. *Electricity*, Oliver and Boyd, Edingurgh, 1953.
7. Elliot, R.S. *Electromagnetics*; McGraw-Hill: New York, 1996.
8. Harsanyi, G. *Sensors in Biomedical Applications: Fundamentals, Technology and Applications*; Technomic Pub. Co.: Lancaster, Pa, 2000.
9. Hayt, W.H. Jr. *Engineering Electromagnetics*; McGraw-Hill: New York, 1967.
10. Iskander, M.F. *Electromagnetic Fields and Waves*; Prentice Hall: Upper Saddle River, NJ, 1992.
11. Jordan, E.C.; Balmain, K.G. *Electromagnetic Waves and Radiating Systems*; Prentice Hall: Upper Saddle River, NJ, 1968.
12. Küpfmüler, K. *Einführung in die theoretische Elektrotechnik*; Springer Verlag: Berlin, 1962.
13. Landau, L.; Lifschitz, L. *The Classical Theory of Fields*; Addison-Wesley: Reading, MA, 1951.
14. Marshall, S.V.; DuBroff, R.E.; Skitek G.G. *Electromagnetic Concepts and Applications*; 4th Ed.; Prentice Hall: Upper Saddle River, NJ, 1996.
15. Maxwell, J.C. *A Treatise on Electricity and Magnetism*; Dover Publications: New York, 1954; Vols. 1 and 2.
16. Popović, B.D. *Introductory Engineering Electromagnetics*; Addison-Wesley: Reading, MA, 1971.
17. Popović, R.S. *Hall Effect Devices*; The Adam Hilger Series on Sensors; IOP Publishing, 1991.
18. Popović, Z.; Popović, B.D. *Introductory Electromagnetics*; Prentice Hall: Upper Saddle River, NJ, 1999.
19. Ramo, S.; Whinnery, J.R.; van Duzer, T. *Fields and Waves in Communication Electronics*; John Wiley & Sons: New York, 1965; 1st Ed.; and 1994; 3rd Ed.
20. Ramshaw, R.; van Heeswijk, R.G. *Energy Conversion: Electric Motors and Generators*; Saunders College Publishing: Philadelphia, 1990.
21. Ripka, P., Ed. *Magnetic Sensors and Magnetometers*; Boston: Artech House, 2001.
22. Rutledge, D.B. *Electromagnetics*; Lecture Notes, Caltech, 1990.
23. Schelkunoff, S.A. *Electromagnetic Fields*; Blaisdell Publishing Company: New York, 1963.
24. Smythe, W.R. *Static and Dynamic Electricity*; McGraw-Hill: New York, 1968.
25. Sommerfeld, A. *Electrodynamics*; Academic Press: New York, 1952.
26. "Electric Motors", Encyclopedia Britannica, 2004. Encyclopedia Britannica Online, 16 Feb. 2004. <http://www.search.eb.com/eb/article?eu = 108542>

5
Wave Propagation

Mohammad Kolbehdari
Intel Corporation
Hillsboro, Oregon, U.S.A.

Matthew N. O. Sadiku
Prairie View A&M University
Prairie View, Texas, U.S.A.

Electromagnetic (EM) wave propagation deals with the transfer of energy or information from one point (a transmitter) to another (a receiver) through the media such as material space, transmission line, and waveguide. It can be described using both theoretical models and practical models based on empirical results. Here we describe the free-space propagation model, path loss models, and the empirical path loss formula. Before presenting these models, we first discuss the theoretical basis and characteristics of EM waves as they propagate through material media.

5.1. WAVE EQUATIONS AND CHARACTERISTICS

The EM wave propagation theory can be described by Maxwell's equations [1,2].

$$\nabla \times \mathbf{E}(\mathbf{r},t) = -\frac{\partial}{\partial t}\mathbf{B}(\mathbf{r},t) \tag{5.1}$$

$$\nabla \times \mathbf{H}(\mathbf{r},t) = \frac{\partial}{\partial t}\mathbf{D}(\mathbf{r},t) + \mathbf{J}(\mathbf{r},t) \tag{5.2}$$

$$\nabla \cdot \mathbf{D}(\mathbf{r},t) = \rho(\mathbf{r},t) \tag{5.3}$$

$$\nabla \cdot \mathbf{B}(\mathbf{r},t) = 0 \tag{5.4}$$

In the above equations, the field quantities **E** and **H** represent, respectively, the electric and magnetic fields, and **D** and **B** the electric and magnetic displacements. **J** and ρ represent the current and charge sources. This set of differential equations relates the time and space rates of change of various field quantities at a point in space and time. Furthermore, the position vector **r** defines a particular location in space (x,y,z) at which the field is being measured. Thus, for example,

$$\mathbf{E}(x,y,z,t) = \mathbf{E}(\mathbf{r},t) \tag{5.5}$$

An auxiliary relationship between the current and charge densities, **J** and ρ, called the *continuity equation* is given by

$$\nabla \cdot \mathbf{J}(\mathbf{r},t) = -\frac{\partial}{\partial t}\rho(\mathbf{r},t) \tag{5.6}$$

The constitutive relationships between the field quantities and electric and magnetic displacements provide the additional constraints needed to solve Eqs. (5.1) and (5.2). These equations characterize a given isotropic material on a macroscopic level in terms of two scalar quantities as

$$\mathbf{B} = \mu\mathbf{H} = \mu_0\mu_r\mathbf{H} \tag{5.7}$$

$$\mathbf{D} = \varepsilon\mathbf{E} = \varepsilon_0\varepsilon_r\mathbf{E} \tag{5.8}$$

where $\mu_0 = 4\pi \times 10^{-7}$ H/m (henrys per meter) is the permeability of free space and $\varepsilon_0 = 8.85 \times 10^{-12}$ F/m (farads per meter) is the permittivity of free space. Also, ε_r and μ_r, respectively, characterize the effects of the atomic and molecular dipoles in the material and the magnetic dipole moments of the atoms constituting the medium.

Maxwell's equations, given by Eqs. (5.1) to (5.4), can be simplified if one assumes time-harmonic fields, i.e., fields varying with a sinusoidal frequency ω. For such fields, it is convenient to use the complex exponential $e^{j\omega t}$. Applying the time-harmonic assumption to Eqs. (5.1) to (5.4), we obtain the time-harmonic wave propagation equations

$$\nabla \times \mathbf{E}(\mathbf{r}) = -j\omega\mathbf{B}(\mathbf{r}) \tag{5.9}$$

$$\nabla \times \mathbf{H}(\mathbf{r}) = j\omega\mathbf{D}(\mathbf{r}) + \mathbf{J}(\mathbf{r}) \tag{5.10}$$

$$\nabla \cdot \mathbf{D}(\mathbf{r}) = \rho(\mathbf{r}) \tag{5.11}$$

$$\nabla \cdot \mathbf{B}(\mathbf{r}) = 0 \tag{5.12}$$

The solution of Maxwell's equations in a source free isotropic medium can be obtained by using Eqs. (5.9) and (5.10) and applying Eqs. (5.7) and (5.8) as follows:

$$\nabla \times \mathbf{E}(\mathbf{r}) = -j\omega\mu\mathbf{H}(\mathbf{r}) \tag{5.13}$$

$$\nabla \times \mathbf{H}(\mathbf{r}) = j\omega\varepsilon\mathbf{E}(\mathbf{r}) \tag{5.14}$$

Taking the curl of the Eq. (5.13) and using Eq. (5.14) we get

$$\nabla \times \nabla \times \mathbf{E}(\mathbf{r}) = -j\omega\mu\nabla \times \mathbf{H}(\mathbf{r}) = \omega^2\mu\varepsilon\mathbf{E}(\mathbf{r}) \tag{5.15}$$

Using a vector identity, and noting that $\rho = 0$, we can write Eq. (15) as

$$\nabla^2\mathbf{E}(\mathbf{r}) + \omega^2\mu\varepsilon\mathbf{E}(\mathbf{r}) = 0 \tag{5.16}$$

This relation is called the *wave equation*. For example, the x component of $\mathbf{E}(\mathbf{r})$ is

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)E_x(\mathbf{r}) + \omega^2\mu\varepsilon E_x(\mathbf{r}) = 0 \tag{5.17}$$

5.1.1. Attenuation

If we consider the general case of a lossy medium that is charge free ($\rho = 0$), Eqs. (5.9) to (5.12) can be manipulated to yield Helmholz' wave equations

$$\nabla^2 \mathbf{E} - \gamma^2 \mathbf{E} = 0 \tag{5.18}$$

$$\nabla^2 \mathbf{H} - \gamma^2 \mathbf{H} = 0 \tag{5.19}$$

where $\gamma = \alpha + j\beta$ is the propagation constant, α is the attenuation constant in nepers per meter or decibels per meter, and β is the phase constant in radians per meters. Constants α and β are given by

$$\alpha = \omega \sqrt{\frac{\mu\varepsilon}{2}\left[\sqrt{1 + \left(\frac{\sigma}{\omega\varepsilon}\right)^2} - 1\right]} \tag{5.20}$$

$$\beta = \omega \sqrt{\frac{\mu\varepsilon}{2}\left[\sqrt{1 + \left(\frac{\sigma}{\omega\varepsilon}\right)^2} + 1\right]} \tag{5.21}$$

where $\omega = 2\pi f$ is the angular frequency of the wave and σ is the conductivity of the medium.

Without loss of generality, if we assume that the wave propagates in the z direction and the wave is polarized in the x direction, solving the wave equations in Eqs. (5.18) and (5.19), we obtain

$$E_x = E_0 e^{-\alpha z} \cos\,(\omega t - \beta z) \tag{5.22}$$

$$H_y = \frac{E_0}{|\eta|} e^{-\alpha z} \cos\,(\omega t - \beta z - \theta_\eta) \tag{5.23}$$

where $\eta = |\eta| \angle \theta_\eta$ is the intrinsic impedance of the medium and is given by

$$|\eta| = \frac{\sqrt{\mu/\varepsilon}}{\sqrt[4]{\left[1 + (\sigma/\omega\varepsilon)^2\right]}} \qquad \tan 2\theta_\eta = \frac{\sigma}{\omega\varepsilon} \qquad 0 \le \theta_\eta \le 45^\circ \tag{5.24}$$

Equations (5.22) and (5.23) show that as the EM wave propagates in the medium, its amplitude is attenuated to $e^{-\alpha z}$.

5.1.2. Dispersion

A plane electromagnetic wave can be described as

$$E_x(\mathbf{r}) = E_{x0} e^{-j\mathbf{k}\cdot\mathbf{r}} = E_{x0} e^{-j(k_x x + k_y y + k_z z)} \tag{5.25}$$

where E_{x0} is an arbitrary constant, and $\mathbf{k} = k_x\,\mathbf{a}_x + k_y\,\mathbf{a}_y + k_z\mathbf{a}_z$ is the vector wave and $\mathbf{r} = \mathbf{a}_x x + \mathbf{a}_y\,y + \mathbf{a}_z z$ is the vector observation point. The substitution of the assumed form of the plane wave in Eq. (5.17) yields

$$k_x^2 + k_y^2 + k_z^2 = k^2 = \omega^2 \mu\varepsilon \tag{5.26}$$

This equation is called the *dispersion relation*. It may also be written in terms of the velocity v defined by

$$k = \frac{\omega}{v} \tag{5.27}$$

The other components of $\mathbf{E}(\mathbf{r})$ with the same wave equation also have the same dispersion equation.

The characteristic impedance of plane wave in free space is given by

$$\eta = \frac{|\mathbf{E}|}{|\mathbf{H}|} = \sqrt{\frac{\mu}{\varepsilon}} = \sqrt{\frac{\mu_0}{\varepsilon_0}} = 377\,\Omega \tag{5.28}$$

5.1.3. Phase Velocity

By assuming $k = k_z = \omega\sqrt{\mu\varepsilon}$, the electric field can be described by

$$E(z,t) = E_0 \cos\,(\omega t - k_z z + \varphi) \tag{5.29}$$

For an observer moving along with the same velocity as the wave, an arbitrary point on the wave will appear to be constant, which requires that the argument of the $E(z,t)$ be constant as defined by

$$\omega t - k_z z + \varphi = \text{constant} \tag{5.30}$$

Taking the derivative with respect to the z yields

$$\frac{dz}{dt} = \frac{\omega}{k_z} = v_p \tag{5.31}$$

where v_p is defined as the phase velocity; for free space it is

$$\frac{\omega}{k_z} = \frac{1}{\sqrt{\mu_0\varepsilon_0}} \cong 3 \times 10^8\ \mathrm{m/s} \tag{5.32}$$

which is the velocity of light in free space.

5.1.4. Group Velocity

A signal consisting of two equal-amplitude tones at frequencies $\omega_0 \pm \Delta\omega$ can be represented by

$$f(t) = 2\cos\omega_0 t \cos\Delta\omega t \tag{5.33}$$

which corresponds to a signal carrier at frequency ω_0 being modulated by a slowly varying envelope having the frequency $\Delta\omega$. If we assume that each of the two signals travels along a propagation direction z with an associated propagation constant $k(\omega)$, then the propagation constant of each signal is $k(\omega_0 \pm \Delta\omega)$. An expansion in a first-order Taylor series yields

$$k(\omega_0 \pm \Delta\omega) \cong k(\omega_0) \pm \Delta\omega k^1(\omega_0) \tag{5.34}$$

where

$$k^1(\omega_0) = \frac{dk(\omega)}{d\omega}|_{\omega=\omega_0} \tag{5.35}$$

The substitution of Eq. (5.34) into Eq. (5.33) following some mathematical manipulation yields

$$f(t,z) = 2\cos\omega_0(t - \tau_p)\ \cos\Delta\omega(t - \tau_g) \tag{5.36}$$

where

$$\tau_p = \frac{k(\omega_0)}{\omega_0} z \tag{5.37}$$

and

$$\tau_g = k^1(\omega_0)z \tag{5.38}$$

The quantities τ_p and τ_g are defined as the phase and group delays, respectively. The corresponding propagation velocities are

$$v_p = \frac{z}{\tau_p} \tag{5.39}$$

$$v_g = \frac{z}{\tau_g} \tag{5.40}$$

For a plane wave propagating in a uniform unbounded medium, the propagation constant is a linear function of frequency given in Eq. (5.26). Thus, for a plane wave, phase and group velocities are equal and given by

$$v_p = v_g = \frac{1}{\sqrt{\mu\varepsilon}} \tag{5.41}$$

It is worthwhile to mention that if the transmission medium is a waveguide, $k(\omega)$ is no longer a linear function of frequency. It is very useful to use the ω-k diagram shown in Fig. 5.1, which plots ω versus $k(\omega)$. In this graph, the slope of a line drawn from the origin to the frequency ω_0 gives the phase velocity and the slope of the tangent to the curve at ω_0 yields the group velocity.

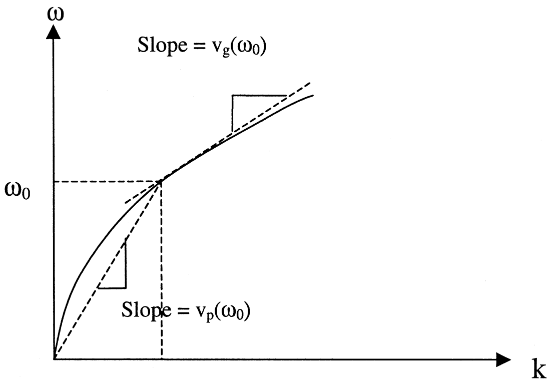


Figure 5.1 ω-k diagram.

5.1.5. Polarization

The electric field of a plane wave propagating in the z direction with no components in the direction of propagation can be written as

$$\mathbf{E}(z) = (\mathbf{a}_x E_{x0} + \mathbf{a}_y E_{y0})e^{-jk_z z} \tag{5.42}$$

By defining

$$E_{x0} = E_{x0}e^{j\varphi_x} \tag{5.43}$$

$$E_{y0} = E_{y0}e^{j\varphi_y} \tag{5.44}$$

we obtain

$$E(z) = (\mathbf{a}_x E_{x0}e^{j\varphi_x} + \mathbf{a}_y E_{y0}^{j\varphi_y})e^{-jk_z z} \tag{5.45}$$

Assuming $A = E_{y0}/E_{x0}$ and $\varphi = \varphi_y - \varphi_x$, and $E_{x0} = 1$, we can write Eq. (5.45) as

$$\mathbf{E}(z) = (\mathbf{a}_x + \mathbf{a}_y A e^{j\varphi})e^{-jk_z z} \tag{5.46}$$

Case I: $A = 0$. $\mathbf{E}(z) = \mathbf{a}_x e^{-jk_z z}$ and $\mathbf{E}(z,t) = \mathbf{a}_x \cos(\omega t - k_z z)$. The movement of the electric field vector in the $z = 0$ plane is along the x axis. This is known as a *linearly polarized wave* along the x axis.

Case II: $A = 1$, $\varphi = 0$.

$$\mathbf{E}(z) = (\mathbf{a}_x + \mathbf{a}_y)e^{-jk_z z} \tag{5.47}$$

and

$$\mathbf{E}(z,t) = (\mathbf{a}_x + \mathbf{a}_y)\cos(\omega t - k_z z) \tag{5.48}$$

This is again a linear polarized wave with the electric field vector at 45 degrees with respect to the x axis.

Case III: $A = 2$, $\varphi = 0$.

$$\mathbf{E}(z,t) = (\mathbf{a}_x + 2\mathbf{a}_y)\ \cos\,(\omega t - k_z z) \tag{5.49}$$

This is again a linear polarized wave with the electric field vector at 63 degrees with respect to the x axis.

Case IV: $A = 1$, $\varphi = \pi/2$.

$$\mathbf{E}(z) = (\mathbf{a}_x + j\mathbf{a}_y)e^{-jk_z z} \tag{5.50}$$

and

$$\mathbf{E}(z,t) = \mathbf{a}_x \cos\,(\omega t - k_z z) - \mathbf{a}_y \sin\,(\omega t - k_z z) \tag{5.51}$$

In this case the electric field vector traces a circle and the wave is defined to be left-handed circularly polarized. Similarly, with $\varphi = -\pi/2$, it is a right-handed circularly polarized wave.

Case VI: $A = 2$ and $\varphi \neq 0$. This is an example of an elliptically polarized wave.

5.1.6. Poynting's Theorem

The relationships between the electromagnetic fields can be described by Poynting's theorem. For an isotropic medium, Maxwell's curl equations can be written as

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t} \tag{5.52}$$

$$\nabla \times \mathbf{H} = \varepsilon \frac{\partial \mathbf{E}}{\partial t} + \mathbf{J} \tag{5.53}$$

where the current density $\mathbf{J}$ can be described as having two components:

$$\mathbf{J} = \mathbf{J}_s + \mathbf{J}_c \tag{5.54}$$

where $\mathbf{J}_c = \sigma\mathbf{E}$ represents conduction current density induced by the presence of the electric fields and $\mathbf{J}_s$ is a source current density that induces electromagnetic fields. The quantity $\mathbf{E} \cdot \mathbf{J}$ has the unit of power per unit volume (watts per unit cubic meter). From Eqs. (5.52) and (5.53) we can get

$$\mathbf{E} \cdot \mathbf{J} = \mathbf{E} \cdot \nabla \times \mathbf{H} - \varepsilon\mathbf{E} \cdot \frac{\partial \mathbf{E}}{\partial t} \tag{5.55}$$

Applying the vector identity

$$\nabla \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot \nabla \times \mathbf{A} - \mathbf{A} \cdot \nabla \times \mathbf{B} \tag{5.56}$$

gives

$$\mathbf{E} \cdot \mathbf{J} = \mathbf{H} \cdot \nabla \times \mathbf{E} - \nabla \cdot (\mathbf{E} \times \mathbf{H}) - \varepsilon \mathbf{E} \cdot \frac{\partial \mathbf{E}}{\partial t} \tag{5.57}$$

Substituting Eq. (5.52) into Eq. (5.57) yields

$$\mathbf{E} \cdot \mathbf{J} = -\mu \mathbf{H} \cdot \frac{\partial \mathbf{H}}{\partial t} - \nabla \cdot (\mathbf{E} \times \mathbf{H}) - \varepsilon \mathbf{E} \cdot \frac{\partial \mathbf{E}}{\partial t} \tag{5.58}$$

Integrating Eq. (5.58) over an arbitrary volume V that is bounded by surface S with an outward unit normal to the surface $\hat{\mathbf{n}}$ shown in Fig. 5.2 gives

$$\iiint_v \mathbf{E} \cdot \mathbf{J}\, dv = \frac{\partial}{\partial t}\left(\iiint_v 1/2\mu|\mathbf{H}|^2\, dv + \iiint_v 1/2\varepsilon|\mathbf{E}|^2\, dv\right) + \oiint_s \hat{\mathbf{n}} \cdot (\mathbf{E} \times \mathbf{H})\, ds \tag{5.59}$$

where the following identity has been used

$$\iiint_v \nabla \cdot \mathbf{A}\, dv = \oiint_s \hat{\mathbf{n}} \cdot \mathbf{A}\, ds \tag{5.60}$$

Equation (5.59) represents the Poynting theorem. The terms $1/2\mu|\mathbf{H}|^2$ and $1/2\varepsilon|\mathbf{E}|^2$ are the energy densities stored in magnetic and electric fields, respectively. The term $\oiint_s \hat{\mathbf{n}} \cdot (\mathbf{E} \times \mathbf{H})\, ds$ describes the power flowing out of the volume V. The quantity $\mathbf{P} = \mathbf{E} \times \mathbf{H}$ is called the *Poynting vector* with the unit of power per unit area. For example, the Poynting theorem can be applied to the plane electromagnetic wave given in Eq. (5.29), where $\varphi = 0$. The wave equations are

$$E_x(z,t) = E_0 \cos(\omega t - k_z z) \tag{5.61}$$

$$H_y(z,t) = \sqrt{\frac{\varepsilon}{\mu}} E_0 \cos(\omega t - k_z z) \tag{5.62}$$

The Poynting vector is in the z direction and is given by

$$P_z = E_x H_y = \sqrt{\frac{\varepsilon}{\mu}} E_0^2 \cos^2(\omega t - k_z z) \tag{5.63}$$

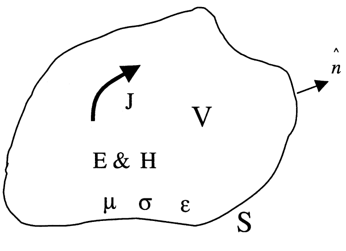


Figure 5.2 A volume V enclosed by surface S and unit vector $\mathbf{n}$.

Applying the trigonometric identity yields

$$P_z = \sqrt{\frac{\varepsilon}{\mu}} E_0^2 \left[\frac{1}{2} + \frac{1}{2} \cos 2\,(\omega t - k_z z) \right] \tag{5.64}$$

It is worth noting that the constant term shows that the wave carries a time-averaged power density and there is a time-varying portion representing the stored energy in space as the maxima and the minima of the fields pass through the region.

We apply the time-harmonic representation of the field components in terms of complex phasors and use the time average of the product of two time-harmonic quantities given by

$$\langle A(t)B(t)\rangle = \tfrac{1}{2}\mathrm{Re}(AB^*) \tag{5.65}$$

where B^* is the complex conjugate of B. The time average Poynting power density is

$$\langle \mathrm{P}\rangle = \tfrac{1}{2}\mathrm{Re}(\mathbf{E} \times \mathbf{H}^*) \tag{5.66}$$

where the quantity $\mathbf{P} = \mathbf{E} \times \mathbf{H}$ is defined as the complex Poynting vector.

5.1.7. Boundary Conditions

The boundary conditions between two materials shown in Fig. 5.3 are

$$E_{t1} = E_{t2} \tag{5.67}$$

$$H_{t1} = H_{t2} \tag{5.68}$$

In the vector form, these boundary conditions can be written as

$$\hat{\mathbf{n}} \times (\mathbf{E}_1 - \mathbf{E}_2) = 0 \tag{5.69}$$

$$\hat{\mathbf{n}} \times (\mathbf{H}_1 - \mathbf{H}_2) = 0 \tag{5.70}$$

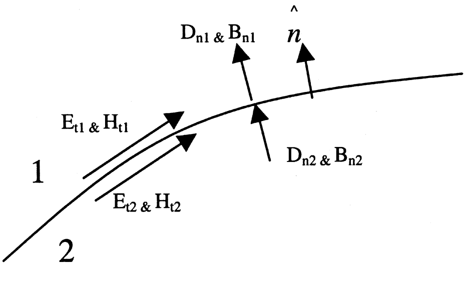


Figure 5.3 Boundary conditions between two materials.

Thus, the tangential components of electric and magnetic field must be equal on the two sides of any boundary between the physical media. Also for a charge- and current-free boundary, the normal components of electric and magnetic flux density are continuous, i.e.,

$$D_{n1} = D_{n2} \tag{5.71}$$

$$B_{n1} = B_{n2} \tag{5.72}$$

For the perfect conductor (infinite conductivity), all the fields inside of the conductor are zero. Thus, the continuity of the tangential electric fields at the boundary yields

$$E_t = 0 \tag{5.73}$$

Since the magnetic fields are zero inside of the conductor, the continuity of the normal magnetic flux density yields

$$B_n = 0 \tag{5.74}$$

Furthermore, the normal electric flux density is

$$D_n = \rho_s \tag{5.75}$$

where ρ_s is a surface charge density on the boundary. The tangential magnetic field is discontinuous by the current enclosed by the path, i.e.,

$$H_t = \mathbf{J}_s \tag{5.76}$$

where $\mathbf{J}_s$ is the surface current density.

5.1.8. Wave Reflection

We now consider the problem of a plane wave obliquely incident on a plane interface between two lossless dielectric media, as shown in Fig. 5.4. It is conventional to define two cases of the problem: the electric field is in the xz plane (parallel polarization) or normal to the xz plane (parallel polarization). Any arbitrary incident plane wave may be treated as a linear combination of the two cases. The two cases are solved in the same manner: obtaining expressions for the incident, reflection, and transmitted fields in each region and matching the boundary conditions to find the unknown amplitude coefficients and angles.

For parallel polarization, the electric field lies in the xz plane so that the incident fields can be written as

$$E_i = E_0(a_x \cos\theta_i - a_z \sin\theta_i)e^{-jk_1(x\sin\theta_i + z\cos\theta_i)} \tag{5.77}$$

$$H_i = \frac{E_0}{\eta_1} a_y e^{-jk_1(x\sin\theta_i + z\cos\theta_i)} \tag{5.78}$$

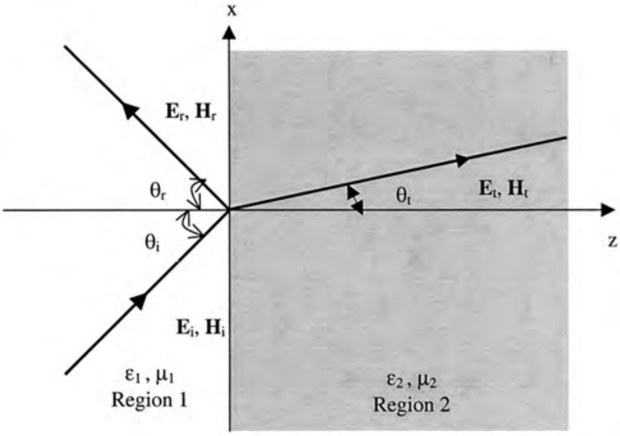


Figure 5.4 A plane wave obliquely incident at the interface between two regions.

where $k_1 = \omega\sqrt{\mu_1\varepsilon_1}$ and $\eta_1 = \sqrt{\mu_1/\varepsilon_1}$. The reflected and transmitted fields can be obtained by imposing the boundary conditions at the interface.

$$E_r = \Gamma_{||}E_0(a_x \cos\theta_r + a_z \sin\theta_r)e^{-jk_1(x\sin\theta_r - z\cos\theta_r)} \tag{5.79}$$

$$H_r = -\frac{\Gamma_{||}E_0}{\eta_1}a_y e^{-jk_1(x\sin\theta_r - z\cos\theta_r)} \tag{5.80}$$

$$E_t = E_0 T_{||}(a_x \cos\theta_t - a_z \sin\theta_t)e^{-jk_2(x\sin\theta_t + z\cos\theta_t)} \tag{5.81}$$

$$H_t = \frac{E_0 T_{||}}{\eta_2}a_y e^{-jk_2(x\sin\theta_t + z\cos\theta_t)} \tag{5.82}$$

where $k_2 = \omega\sqrt{\mu_2\varepsilon_2}$, $\eta_2 = \sqrt{\mu_2/\varepsilon_2}$

$$\theta_r = \theta_i \qquad k_1 \sin\theta_i = k_2 \sin\theta_t \quad \text{(Snell's law)} \tag{5.83}$$

$$\Gamma_{||} = \frac{\eta_2\cos\theta_t - \eta_1\cos\theta_i}{\eta_2\cos\theta_t + \eta_1\cos\theta_i} \tag{5.84}$$

and

$$T_{||} = \frac{2\eta_2\cos\theta_i}{\eta_2\cos\theta_t + \eta_1\cos\theta_i} \tag{5.85}$$

For perpendicular polarization, the electric field is normal to the xz plane. The incident fields are given by

$$E_i = E_0 a_y e^{-jk_1(x\sin\theta_i + z\cos\theta_i)} \tag{5.86}$$

$$H_i = \frac{E_0}{\eta_1}(-a_x \cos\theta_i + a_z \sin\theta_i)e^{-jk_1(x\sin\theta_i + z\cos\theta_i)} \tag{5.87}$$

while the reflected and transmitted fields are

$$E_r = \Gamma_\perp E_0 a_y e^{-jk_1(x\sin\theta_r - z\cos\theta_r)} \tag{5.88}$$

$$H_r = \frac{\Gamma_\perp E_0}{\eta_1}(a_x\cos\theta_r + a_z\sin\theta_r)e^{-jk_1(x\sin\theta_r - z\cos\theta_r)} \tag{5.89}$$

$$E_t = E_0 T_\perp a_y e^{-jk_2(x\sin\theta_t + z\cos\theta_t)} \tag{5.90}$$

$$H_t = \frac{E_0 T_\perp}{\eta_2}(-a_x\cos\theta_t + a_z\sin\theta_t)e^{-jk_2(x\sin\theta_t + z\cos\theta_t)} \tag{5.91}$$

where

$$k_1\sin\theta_i = k_1\sin\theta_r = k_2\sin\theta_t \qquad \text{(Snell's law)} \tag{5.92}$$

$$\Gamma_\perp = \frac{\eta_2\cos\theta_i - \eta_1\cos\theta_t}{\eta_2\cos\theta_i + \eta_1\cos\theta_t} \tag{5.93}$$

and

$$T_\perp = \frac{2\eta_2\cos\theta_i}{\eta_2\cos\theta_i + \eta_1\cos\theta_t} \tag{5.94}$$

5.2. FREE-SPACE PROPAGATION MODEL

The free-space propagation model is used in predicting the received signal strength when the transmitter and receiver have a clear line-of-sight path between them. If the receiving antenna is separated from the transmitting antenna in free space by a distance r, as shown in Fig. 5.5, the power received P_r by the receiving antenna is given by the Friis equation [3]

$$P_r = G_r G_t \left(\frac{\lambda}{4\pi r}\right)^2 P_t \tag{5.95}$$

where P_t is the transmitted power, G_r is the receiving antenna gain, G_t is the transmitting antenna gain, and λ is the wavelength $(= c/f)$ of the transmitted signal. The Friis equation

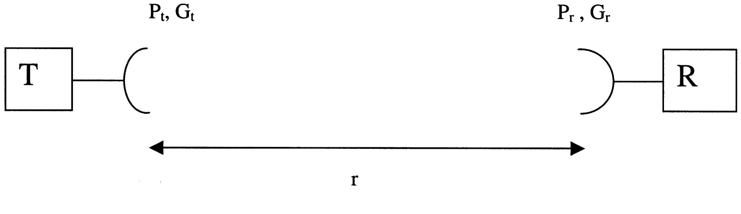


Figure 5.5 Basic wireless system.

relates the power received by one antenna to the power transmitted by the other, provided that the two antennas are separated by $r > 2d^2/\lambda$, where d is the largest dimension of either antenna. Thus, the Friis equation applies only when the two antennas are in the far field of each other. It also shows that the received power falls off as the square of the separation distance r. The power decay as $1/r^2$ in a wireless system, as exhibited in Eq. (5.95), is better than the exponential decay in power in a wired link. In actual practice, the value of the received power given in Eq. (5.95) should be taken as the maximum possible because some factors can serve to reduce the received power in a real wireless system. This will be discussed fully in the next section.

From Eq. (5.95), we notice that the received power depends on the product P_tG_t. The product is defined as the *effective isotropic radiated power* (EIRP), i.e.,

$$\mathrm{EIRP} = P_tG_t \tag{5.96}$$

The EIRP represents the maximum radiated power available from a transmitter in the direction of maximum antenna gain relative to an isotropic antenna.

5.3. PATH LOSS MODEL

Wave propagation seldom occurs under the idealized conditions assumed in Sec. 5.1. For most communication links, the analysis in Sec. 5.1 must be modified to account for the presence of the earth, the ionosphere, and atmospheric precipitates such as fog, raindrops, snow, and hail [4]. This will be done in this section.

The major regions of the earth's atmosphere that are of importance in radio wave propagation are the troposphere and the ionosphere. At radar frequencies (approximately 100 MHz to 300 GHz), the troposphere is by far the most important. It is the lower atmosphere consisting of a nonionized region extending from the earth's surface up to about 15 km. The ionosphere is the earth's upper atmosphere in the altitude region from 50 km to one earth radius (6370 km). Sufficient ionization exists in this region to influence wave propagation.

Wave propagation over the surface of the earth may assume one of the following three principal modes:

Surface wave propagation along the surface of the earth
Space wave propagation through the lower atmosphere
Sky wave propagation by reflection from the upper atmosphere

These modes are portrayed in Fig. 5.6. The sky wave is directed toward the ionosphere, which bends the propagation path back toward the earth under certain conditions in a limited frequency range (below 50 MHz approximately). This is highly dependent on the condition of the ionosphere (its level of ionization) and the signal frequency. The surface (or ground) wave takes effect at the low-frequency end of the spectrum (2–5 MHz approximately) and is directed along the surface over which the wave is propagated. Since the propagation of the ground wave depends on the conductivity of the earth's surface, the wave is attenuated more than if it were propagation through free space. The space wave consists of the direct wave and the reflected wave. The direct wave travels from the transmitter to the receiver in nearly a straight path while the reflected wave is due to ground reflection. The space wave obeys the optical laws in that direct and reflected wave

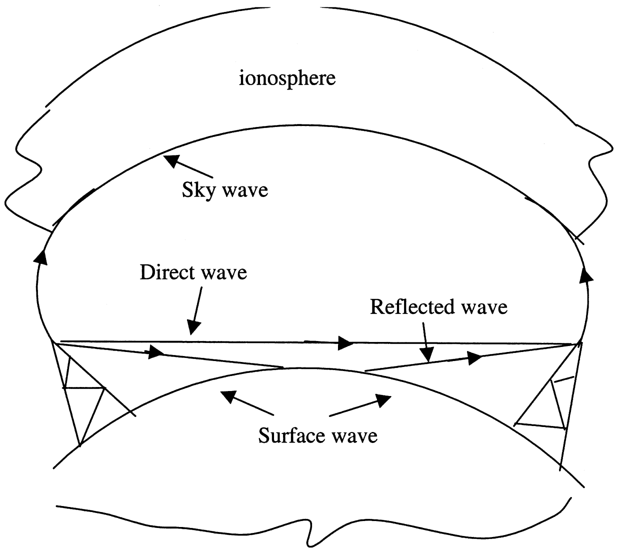


Figure 5.6 Modes of wave propagation.

components contribute to the total wave component. Although the sky and surface waves are important in many applications, we will only consider space wave in this chapter.

In case the propagation path is not in free space, a correction factor F is included in the Friis equation, Eq. (5.74), to account for the effect of the medium. This factor, known as the *propagation factor*, is simply the ratio of the electric field intensity E_m in the medium to the electric field intensity E_o in free space, i.e.,

$$F = \frac{E_m}{E_o} \tag{5.97}$$

The magnitude of F is always less than unity since E_m is always less than E_o. Thus, for a lossy medium, Eq. (5.95) becomes

$$P_r = G_r G_t \left(\frac{\lambda}{4\pi r}\right)^2 P_t |F|^2 \tag{5.98}$$

For practical reasons, Eqs. (5.95) and (5.98) are commonly expressed in logarithmic form. If all the terms are expressed in decibels (dB), Eq. (5.98) can be written in the logarithmic form as

$$P_r = P_t + G_r + G_t - L_o - L_m \tag{5.99}$$

where P = power in dB referred to 1 W (or simply dBW), G = gain in dB, L_o = free-space loss in dB, and L_m loss in dB due to the medium. (Note that $G\,\text{dB} = 10\log_{10} G$.) The free-space loss is obtained directly from Eq. (5.98) as

$$L_o = 20\log\frac{4\pi r}{\lambda} \tag{5.100}$$

while the loss due to the medium is given by

$$L_m = -20 \log |F| \tag{5.101}$$

Our major concern in the rest of this subsection is to determine L_o and L_m for an important case of space propagation that differs considerably from the free-space conditions.

The phenomenon of multipath propagation causes significant departures from free-space conditions. The term *multipath* denotes the possibility of EM wave propagating along various paths from the transmitter to the receiver. In multipath propagation of an EM wave over the earth's surface, two such path exists: a direct path and a path via reflection and diffractions from the interface between the atmosphere and the earth. A simplified geometry of the multipath situation is shown in Fig. 5.7. The reflected and diffracted component is commonly separated into two parts: one *specular* (or coherent) and the other *diffuse* (or incoherent), that can be separately analyzed. The specular component is well defined in terms of its amplitude, phase, and incident direction. Its main characteristic is its conformance to Snell's law for reflection, which requires that the angles of incidence and reflection be equal and coplanar. It is a plane wave, and as such, is uniquely specified by its direction. The diffuse component, however, arises out of the random nature of the scattering surface and, as such, is nondeterministic. It is not a plane wave and does not obey Snell's law for reflection. It does not come from a given direction but from a continuum.

The loss factor F that accounts for the departures from free-space conditions is given by

$$F = 1 + \Gamma \rho_s DS(\theta) e^{-j\Delta} \tag{5.102}$$

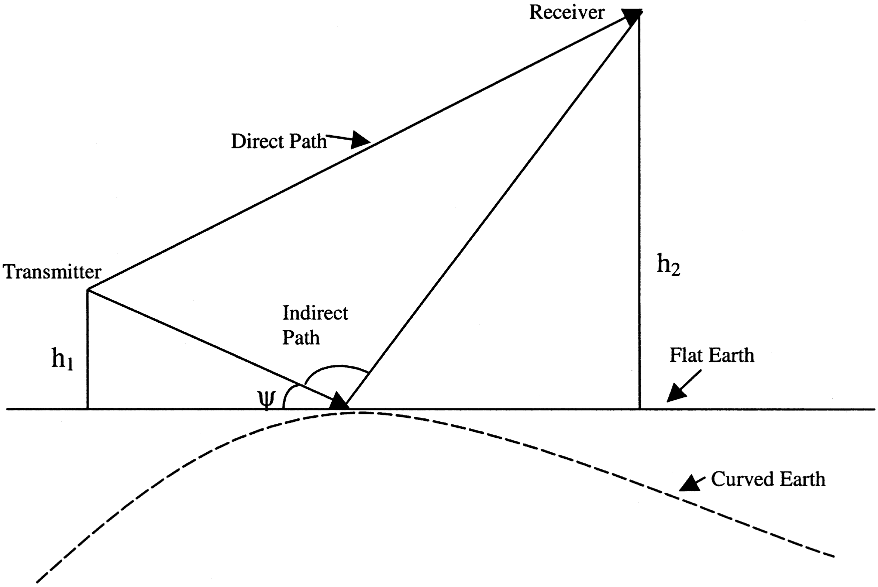


Figure 5.7 Multipath geometry.

where

$\Gamma =$ Fresnel reflection coefficient.
$\rho_s =$ roughness coefficient.
$D =$ divergence factor.
$S(\theta) =$ shadowing function.
$\Delta =$ phase angle corresponding to the path difference.

The Fresnel reflection coefficient Γ accounts for the electrical properties of the earth's surface. Since the earth is a lossy medium, the value of the reflection coefficient depends on the complex relative permittivity ε_c of the surface, the grazing angle ψ, and the wave polarization. It is given by

$$\Gamma = \frac{\sin\psi - z}{\sin\psi + z} \tag{5.103}$$

where

$$z = \sqrt{\varepsilon_c - \cos^2\psi} \qquad \text{for horizontal polarization} \tag{5.104}$$

$$z = \frac{\sqrt{\varepsilon_c - \cos^2\psi}}{\varepsilon_c} \qquad \text{for vertical polarization} \tag{5.105}$$

$$\varepsilon_c = \varepsilon_r - j\frac{\sigma}{\omega\varepsilon_o} = \varepsilon_r - j60\sigma\lambda \tag{5.106}$$

ε_r and σ are, respectively, the dielectric constant and the conductivity of the surface; ω and λ are, respectively, the frequency and wavelength of the incident wave; and ψ is the grazing angle. It is apparent that $0 < |\Gamma| < 1$.

To account for the spreading (or divergence) of the reflected rays due to earth curvature, we introduce the divergence factor D. The curvature has a tendency to spread out the reflected energy more than a corresponding flat surface. The divergence factor is defined as the ratio of the reflected field from curved surface to the reflected field from flat surface. Using the geometry of Fig. 5.8, we get D as

$$D = \left(1 + \frac{2G_1G_2}{a_eG\sin\psi}\right)^{-1/2} \tag{5.107}$$

where $G = G_1 + G_2$ is the total ground range and $a_e = 6370$ km is the effective earth radius. Given the transmitter height h_1, the receiver height h_2, and the total ground range G, we can determine G_1, G_2, and ψ. If we define

$$p = \frac{2}{\sqrt{3}}\left[a_e(h_1 + h_2) + \frac{G^2}{4}\right]^{1/2} \tag{5.108}$$

$$\alpha = \cos^{-1}\left[\frac{2a_e(h_1 - h_2)G}{p^3}\right] \tag{5.109}$$

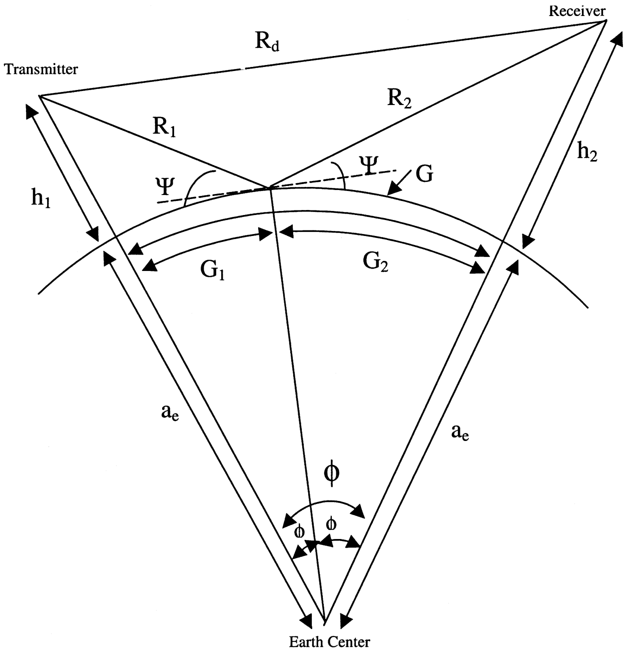


Figure 5.8 Geometry of spherical earth reflection.

and assume $h_1 \leq h_2$ and $G_1 \leq G_2$, using small angle approximation yields [5]

$$G_1 = \frac{G}{2} + p\cos\frac{\pi+\alpha}{3} \tag{5.110}$$

$$G_2 = G - G_1 \tag{5.111}$$

$$\phi_i = \frac{G_i}{a_e}, \qquad i = 1,2 \tag{5.112}$$

$$R_i = \left[h_i^2 + 4a_e(a_e + h_i)\sin^2(\phi_i/2)\right]^{1/2} \qquad i = 1,2 \tag{5.113}$$

The grazing angle is given by

$$\psi = \sin^{-1}\left(\frac{2a_eh_1 + h_1^2 - R_1^2}{2a_eR_1}\right) \tag{5.114}$$

or

$$\psi = \sin^{-1}\left(\frac{2a_eh_1 + h_1^2 + R_1^2}{2(a_e + h_1)R_1}\right) - \phi_1 \tag{5.115}$$

Although D varies from 0 to 1, in practice D is a significant factor at low grazing angle ψ (less than 0.1 %).

The phase angle corresponding to the path difference between direct and reflected waves is given by

$$\Delta = \frac{2\pi}{\lambda}(R_1 + R_2 - R_d) \tag{5.116}$$

The roughness coefficient ρ_s takes care of the fact that the earth surface is not sufficiently smooth to produce specular (mirrorlike) reflection except at very low grazing angle. The earth's surface has a height distribution that is random in nature. The randomness arises out of the hills, structures, vegetation, and ocean waves. It is found that the distribution of the heights of the earth's surface is usually the gaussian or normal distribution of probability theory. If σ_h is the standard deviation of the normal distribution of heights, we define the roughness parameters

$$g = \frac{\sigma_h \sin \psi}{\lambda} \tag{5.117}$$

If $g < 1/8$, specular reflection is dominant; if $g > 1/8$, diffuse scattering results. This criterion, known as the *Rayleigh criterion*, should only be used as a guideline since the dividing line between a specular and a diffuse reflection or between a smooth and a rough surface is not well defined [6]. The roughness is taken into account by the roughness coefficient $(0 < \rho_s < 1)$, which is the ratio of the field strength after reflection with roughness taken into account to that which would be received if the surface were smooth. The roughness coefficient is given by

$$\rho_s = \exp[-2(2\pi g)^2] \tag{5.118}$$

Shadowing is the blocking of the direct wave due to obstacles. The shadowing function $S(\theta)$ is important at low grazing angle. It considers the effect of geometric shadowing—the fact that the incident wave cannot illuminate parts of the earth's surface shadowed by higher parts. In a geometric approach, where diffraction and multiple scattering effects are neglected, the reflecting surface will consist of well-defined zones of illumination and shadow. As there will be no field on a shadowed portion of the surface, the analysis should include only the illuminated portions of the surface. A pictorial representation of rough surfaces illuminated at angle of incidence $\theta(= 90^\circ - \psi)$ is shown in Fig. 5.9. It is evident

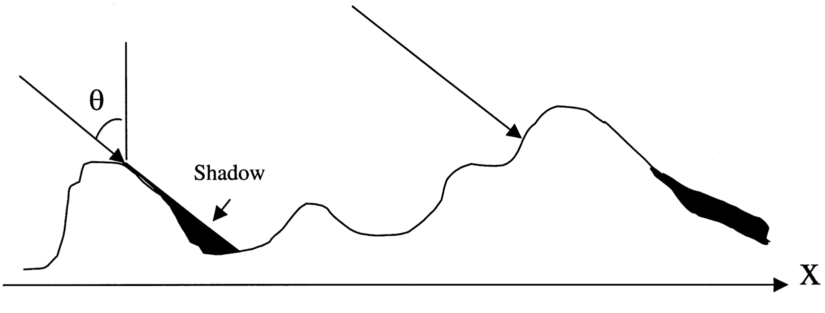


Figure 5.9 Rough surface illuminated at an angle of incidence.

from the figure that the shadowing function $S(\theta)$ equals unity when $\theta = 0$ and zero when $\theta = \pi/2$. According to Smith [7],

$$S(\theta) = \frac{1 - (1/2)\mathrm{erfc}(a)}{1 + 2B} \tag{5.119}$$

where erfc(x) is the complementary error function,

$$\mathrm{erfc}(x) = 1 - \mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt \tag{5.120}$$

and

$$B = \frac{1}{4a}\left[\frac{1}{\sqrt{\pi}} e^{a^2} - a\,\mathrm{erfc}(a)\right] \tag{5.121}$$

$$a = \frac{\cot\theta}{2s} \tag{5.122}$$

$$s = \frac{\sigma_h}{\sigma_l} = \text{rms surface slope} \tag{5.123}$$

In Eq. (5.123), σ_h is the rms roughness height and σ_l is the correlation length. Alternative models for $S(\theta)$ are available in the literature. Using Eqs. (5.103) to (5.123), we can calculate the loss factor in Eq. (5.102). Thus

$$L_o = 20 \log \frac{4\pi R_d}{\lambda} \tag{5.124}$$

$$L_m = -20 \log\left[1 + \Gamma \rho_s D S(\theta) e^{-j\Delta}\right] \tag{5.125}$$

5.4. EMPIRICAL PATH LOSS FORMULA

Both theoretical and experimental propagation models are used in predicting the path loss. In addition to the theoretical model presented in the previous section, there are empirical models for finding path loss. Of the several models in the literature, the Okumura et al. model [8] is the most popular choice for analyzing mobile-radio propagation because of its simplicity and accuracy. The model is based on extensive measurements in and around Tokyo, compiled into charts, that can be applied to VHF and UHF mobile-radio propagation. The medium path loss (in dB) is given by [9]

$$L_p = \begin{cases} A + B\log_{10} r & \text{for urban area} \\ A + B\log_{10} r - C & \text{for suburban area} \\ A + B\log_{10} r - D & \text{for open area} \end{cases} \tag{5.126}$$

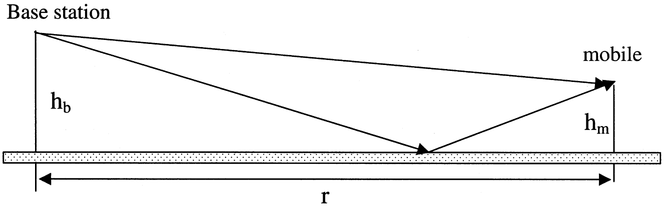


Figure 5.10 Radio propagation over a flat surface.

where r (in kilometers) is the distance between the base and mobile stations, as illustrated in Fig. 5.10. The values of A, B, C, and D are given in terms of the carrier frequency f (in MHz), the base station antenna height h_b (in meters), and the mobile station antenna height h_m (in meters) as

$$A = 69.55 + 26.16\log_{10} f - 13.82\log_{10} h_b - a(h_m) \tag{5.127a}$$

$$B = 44.9 - 6.55\log_{10} h_b \tag{5.127b}$$

$$C = 5.4 + 2\left(\log_{10}\frac{f}{28}\right)^2 \tag{5.127c}$$

$$D = 40.94 - 19.33\log_{10} f + 4.78\left(\log_{10} f\right)^2 \tag{5.127d}$$

where

$$a(h_m) = \begin{cases} 0.8 - 1.56\log_{10} f + \left(1.1\ \log_{10} f - 0.7\right)h_m & \text{for medium/small city} \\ 8.28\left[\log_{10}(1.54h_m)\right]^2 - 1.1 & \text{for } f \geq 200\,\text{MHz} \\ 3.2\left[\log_{10}(11.75h_m)\right]^2 - 4.97 & \text{for } f < 400\,\text{MHz for large city} \end{cases} \tag{5.128}$$

The following conditions must be satisfied before Eq. (5.127) is used: $150 < f < 1500$ MHz; $1 < r < 80$ km; $30 < h_b < 400$ m; $1 < h_m < 10$ m. Okumura's model has been found to be fairly good in urban and suburban areas but not as good in rural areas.

REFERENCES

1. David M. Pozar. *Microwave Engineering*; Addison-Wesley Publishing Company: New York, NY, 1990.
2. Kong, J.A. *Theory of Electromagnetic Waves*; Wiley: New York, 1975.
3. Sadiku, M.N.O. *Elements of Electromagnetics*, 3rd Ed.; Oxford University Press: New York, 2001; 621–623.
4. Sadiku, M.N.O. Wave propagation, In *The Electrical Engineering Handbook*; Dorf, R.C., Ed.; CRC Press: Boca Raton, FL, 1997; 925–937.
5. Blake, L.V. *Radar Range-Performance Analysis*; Artech House: Norwood, MA, 1986; 253–271.
6. Beckman, P.; Spizzichino, A. *The Scattering of Electromagnetic Waves from Random Surfaces*; Macmillan: New York, 1963.

7. Smith, B.G. Geometrical shadowing of a random rough surface. IEEE Trans. Ant. Prog. **1967**, *15*, 668–671.
8. Okumura, Y. et al. Field strength and its variability in VHF and UHF land mobile service. Review of Electrical Communication Lab **Sept./Oct. 1969**, *16*, 825–873.
9. Feher, K. *Wireless Digital Communications*; Prentice-Hall: Upper Saddle River, NJ, 1995; 74–76.

6
Transmission Lines

Andreas Weisshaar
Oregon State University
Corvallis, Oregon, U.S.A.

6.1. INTRODUCTION

A transmission line is an electromagnetic guiding system for efficient point-to-point transmission of electric signals (information) and power. Since its earliest use in telegraphy by Samual Morse in the 1830s, transmission lines have been employed in various types of electrical systems covering a wide range of frequencies and applications. Examples of common transmission-line applications include TV cables, antenna feed lines, telephone cables, computer network cables, printed circuit boards, and power lines. A transmission line generally consists of two or more conductors embedded in a system of dielectric media. Figure 6.1 shows several examples of commonly used types of transmission lines composed of a set of parallel conductors.

The coaxial cable (Fig. 6.1a) consists of two concentric cylindrical conductors separated by a dielectric material, which is either air or an inert gas and spacers, or a foam-filler material such as polyethylene. Owing to their self-shielding property, coaxial cables are widely used throughout the radio frequency (RF) spectrum and in the microwave frequency range. Typical applications of coaxial cables include antenna feed lines, RF signal distribution networks (e.g., cable TV), interconnections between RF electronic equipment, as well as input cables to high-frequency precision measurement equipment such as oscilloscopes, spectrum analyzers, and network analyzers.

Another commonly used transmission-line type is the two-wire line illustrated in Fig. 6.1b. Typical examples of two-wire lines include overhead power and telephone lines and the flat twin-lead line as an inexpensive antenna lead-in line. Because the two-wire line is an open transmission-line structure, it is susceptible to electromagnetic interference. To reduce electromagnetic interference, the wires may be periodically twisted (twisted pair) and/or shielded. As a result, unshielded twisted pair (UTP) cables, for example, have become one of the most commonly used types of cable for high-speed local area networks inside buildings.

Figure 6.1c–e shows several examples of the important class of planar-type transmission lines. These types of transmission lines are used, for example, in printed circuit boards to interconnect components, as interconnects in electronic packaging, and as interconnects in integrated RF and microwave circuits on ceramic or semiconducting substrates. The microstrip illustrated in Fig. 6.1c consists of a conducting strip and a

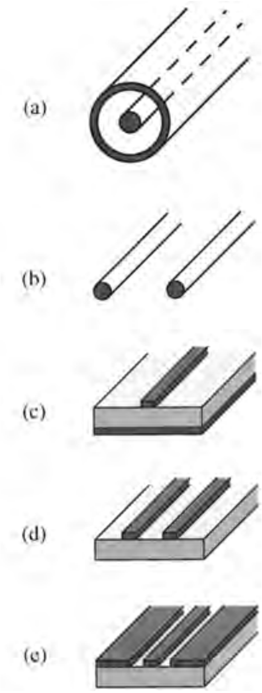


Figure 6.1 Examples of commonly used transmission lines: (a) coaxial cable, (b) two-wire line, (c) microstrip, (d) coplanar stripline, (e) coplanar waveguide.

conducting plane (ground plane) separated by a dielectric substrate. It is a widely used planar transmission line mainly because of its ease of fabrication and integration with devices and components. To connect a shunt component, however, through-holes are needed to provide access to the ground plane. On the other hand, in the coplanar stripline and coplanar waveguide (CPW) transmission lines (Fig. 6.1d and e) the conducting signal and ground strips are on the same side of the substrate. The single-sided conductor configuration eliminates the need for through-holes and is preferable for making connections to surface-mounted components.

In addition to their primary function as guiding system for signal and power transmission, another important application of transmission lines is to realize capacitive and inductive circuit elements, in particular at microwave frequencies ranging from a few gigahertz to tens of gigahertz. At these frequencies, lumped reactive elements become exceedingly small and difficult to realize and fabricate. On the other hand, transmission-line sections of appropriate lengths on the order of a quarter wavelength can be easily realized and integrated in planar transmission-line technology. Furthermore, transmission-line circuits are used in various configurations for impedance matching. The concept of functional transmission-line elements is further extended to realize a range of microwave passive components in planar transmission-line technology such as filters, couplers and power dividers [1].

This chapter on transmission lines provides a summary of the fundamental transmission-line theory and gives representative examples of important engineering applications. The following sections summarize the fundamental mathematical transmission-line equations and associated concepts, review the basic characteristics of transmission lines, present the transient response due to a step voltage or voltage pulse as well as the sinusoidal steady-state response of transmission lines, and give practical application examples and solution techniques. The chapter concludes with a brief summary of more advanced transmission-line concepts and gives a brief discussion of current technological developments and future directions.

6.2. BASIC TRANSMISSION-LINE CHARACTERISTICS

A transmission line is inherently a distributed system that supports propagating electromagnetic waves for signal transmission. One of the main characteristics of a transmission line is the delayed-time response due to the finite wave velocity.

The transmission characteristics of a transmission line can be rigorously determined by solving Maxwell's equations for the corresponding electromagnetic problem. For an "ideal" transmission line consisting of two parallel perfect conductors embedded in a homogeneous dielectric medium, the fundamental transmission mode is a transverse electromagnetic (TEM) wave, which is similar to a plane electromagnetic wave described in the previous chapter [2]. The electromagnetic field formulation for TEM waves on a transmission line can be converted to corresponding voltage and current circuit quantities by integrating the electric field between the conductors and the magnetic field around a conductor in a given plane transverse to the direction of wave propagation [3,4].

Alternatively, the transmission-line characteristics may be obtained by considering the transmission line directly as a distributed-parameter circuit in an extension of the traditional circuit theory [5]. The distributed circuit parameters, however, need to be determined from electromagnetic field theory. The distributed-circuit approach is followed in this chapter.

6.2.1. Transmission-line Parameters

A transmission line may be described in terms of the following distributed-circuit parameters, also called *line parameters*: the inductance parameter L (in H/m), which represents the series (loop) inductance per unit length of line, and the capacitance parameter C (in F/m), which is the shunt capacitance per unit length between the two conductors. To represent line losses, the resistance parameter R (in Ω/m) is defined for the series resistance per unit length due to the finite conductivity of both conductors, while the conductance parameter G (in S/m) gives the shunt conductance per unit length of line due to dielectric loss in the material surrounding the conductors.

The R, L, G, C transmission-line parameters can be derived in terms of the electric and magnetic field quantities by relating the corresponding stored energy and dissipated power. The resulting relationships are [1,2]

$$L = \frac{\mu}{|I|^2} \int_S \mathbf{H} \cdot \mathbf{H}^* ds \tag{6.1}$$

$$C = \frac{\epsilon'}{|V|^2} \int_S \mathbf{E} \cdot \mathbf{E}^* ds \tag{6.2}$$

$$R = \frac{R_s}{|I|^2} \int_{C_1 + C_2} \mathbf{H} \cdot \mathbf{H}^* dl \tag{6.3}$$

$$G = \frac{\omega\epsilon' \tan\delta}{|V|^2} \int_S \mathbf{E} \cdot \mathbf{E}^* ds \tag{6.4}$$

where **E** and **H** are the electric and magnetic field vectors in phasor form, "*" denotes complex conjugate operation, R_s is the surface resistance of the conductors,[†] ϵ' is the permittivity and $\tan\delta$ is the loss tangent of the dielectric material surrounding the conductors, and the line integration in Eq. (6.3) is along the contours enclosing the two conductor surfaces.

In general, the line parameters of a lossy transmission line are frequency dependent owing to the *skin effect* in the conductors and loss tangent of the dielectric medium.[‡] In the following, a lossless transmission line having constant L and C and zero R and G parameters is considered. This model represents a good first-order approximation for many practical transmission-line problems. The characteristics of lossy transmission lines are discussed in Sec. 6.4.

6.2.2. Transmission-line Equations for Lossless Lines

The fundamental equations that govern wave propagation on a lossless transmission line can be derived from an equivalent circuit representation for a short section of transmission line of length Δz illustrated in Fig. 6.2. A mathematically more rigorous derivation of the transmission-line equations is given in Ref. 5.

By considering the voltage drop across the series inductance $L\Delta z$ and current through the shunt capacitance $C\Delta z$, and taking $\Delta z \to 0$, the following fundamental transmission-line equations (also known as *telegrapher's equations*) are obtained.

$$\frac{\partial v(z,t)}{\partial z} = -L\frac{\partial i(z,t)}{\partial t} \tag{6.5}$$

$$\frac{\partial i(z,t)}{\partial z} = -C\frac{\partial v(z,t)}{\partial t} \tag{6.6}$$

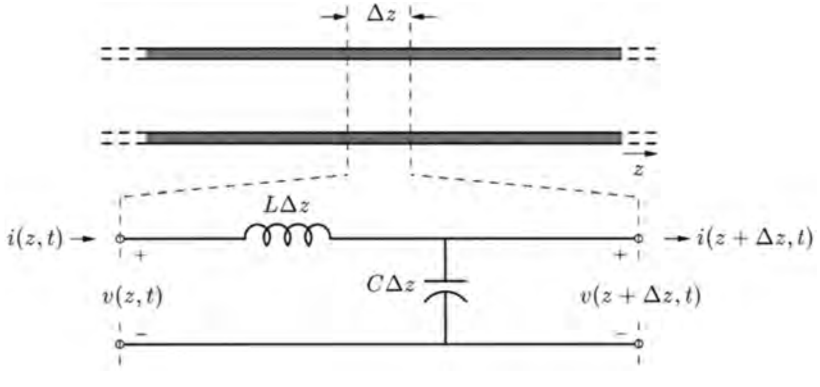


Figure 6.2 Schematic representation of a two-conductor transmission line and associated equivalent circuit model for a short section of lossless line.

[†] For a good conductor the surface resistance is $R_s = 1/\sigma\delta_s$, where the skin depth $\delta_s = 1/\sqrt{\pi f \mu\sigma}$ is assumed to be small compared to the cross-sectional dimensions of the conductor.

[‡] The skin effect describes the nonuniform current distribution inside the conductor caused by the time-varying magnetic flux within the conductor. As a result the resistance per unit length increases while the inductance per unit length decreases with increasing frequency. The loss tangent of the dielectric medium $\tan\delta = \epsilon''/\epsilon'$ typically results in an increase in shunt conductance with frequency, while the change in capacitance is negligible in most practical cases.

The transmission-line equations, Eqs. (6.5) and (6.6), can be combined to obtain a one-dimensional wave equation for voltage

$$\frac{\partial^2 v(z,t)}{\partial z^2} = LC\frac{\partial^2 v(z,t)}{\partial t^2} \tag{6.7}$$

and likewise for current.

6.2.3. General Traveling-wave Solutions for Lossless Lines

The wave equation in Eq. (6.7) has the general solution

$$v(z,t) = v^+\left(t - \frac{z}{v_p}\right) + v^-\left(t + \frac{z}{v_p}\right) \tag{6.8}$$

where $v^+(t - z/v_p)$ corresponds to a wave traveling in the positive z direction, and $v^-(t + z/v_p)$ to a wave traveling in the negative z direction with constant velocity of propagation

$$v_p = \frac{1}{\sqrt{LC}} \tag{6.9}$$

Figure 6.3 illustrates the progression of a single traveling wave as function of position along the line and as function of time.

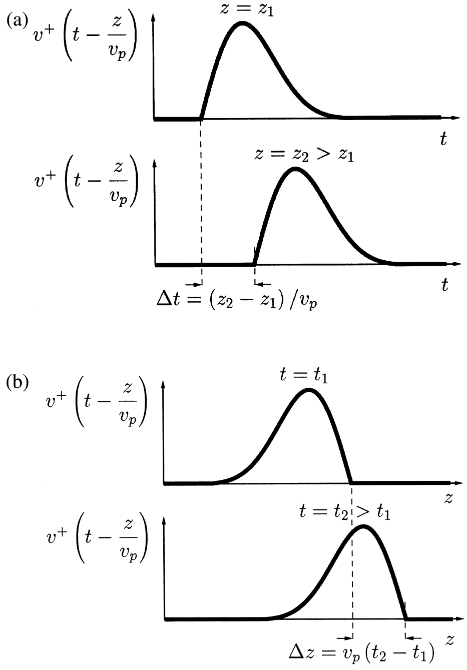


Figure 6.3 Illustration of the space and time variation for a general voltage wave $v^+(t - z/v_p)$: (a) variation in time and (b) variation in space.

A corresponding solution for sinusoidal traveling waves is

$$\begin{aligned} v(z,t) &= v_0^+ \cos\left[\omega\left(t - \frac{z}{v_p}\right) + \phi^+\right] + v_0^- \cos\left[\omega\left(t + \frac{z}{v_p}\right) + \phi^-\right] \\ &= v_0^+ \cos(\omega t - \beta z + \phi^+) + v_0^- \cos(\omega t + \beta z + \phi^-) \end{aligned} \tag{6.10}$$

where

$$\beta = \frac{\omega}{v_p} = \frac{2\pi}{\lambda} \tag{6.11}$$

is the phase constant and $\lambda = v_p/f$ is the wavelength on the line. Since the spatial phase change βz depends on both the physical distance and the wavelength on the line, it is commonly expressed as *electrical distance* (or *electrical length*) θ with

$$\theta = \beta z = 2\pi\frac{z}{\lambda} \tag{6.12}$$

The corresponding wave solutions for current associated with voltage $v(z,t)$ in Eq. (6.8) are found with Eq. (6.5) or (6.6) as

$$i(z,t) = \frac{v^+(t - z/v_p)}{Z_0} - \frac{v^-(t + z/v_p)}{Z_0} \tag{6.13}$$

The parameter Z_0 is defined as the *characteristic impedance* of the transmission line and is given in terms of the line parameters by

$$Z_0 = \sqrt{\frac{L}{C}} \tag{6.14}$$

The characteristic impedance Z_0 specifies the ratio of voltage to current of a single traveling wave and, in general, is a function of both the conductor configuration (dimensions) and the electric and magnetic properties of the material surrounding the conductors. The negative sign in Eq. (6.13) for a wave traveling in the negative z direction accounts for the definition of positive current in the positive z direction.

As an example, consider the coaxial cable shown in Fig. 6.1a with inner conductor of diameter d, outer conductor of diameter D, and dielectric medium of dielectric constant ϵ_r. The associated distributed inductance and capacitance parameters are

$$L = \frac{\mu_0}{2\pi}\ln\frac{D}{d} \tag{6.15}$$

$$C = \frac{2\pi\epsilon_0\epsilon_r}{\ln(D/d)} \tag{6.16}$$

where $\mu_0 = 4\pi \times 10^{-7}$ H/m is the free-space permeability and $\epsilon_0 \approx 8.854 \times 10^{-12}$ F/m is the free-space permittivity. The characteristic impedance of the coaxial line is

$$Z_0 = \sqrt{\frac{L}{C}} = \frac{1}{2\pi}\sqrt{\frac{\mu_0}{\epsilon_0\epsilon_r}}\ln\frac{D}{d} = \frac{60}{\sqrt{\epsilon_r}}\ln\frac{D}{d} \qquad (\Omega) \tag{6.17}$$

and the velocity of propagation is

$$v_p = \frac{1}{LC} = \frac{1}{\sqrt{\mu_0 \epsilon_0 \epsilon_r}} = \frac{c}{\sqrt{\epsilon_r}} \tag{6.18}$$

where $c \approx 30\,\mathrm{cm/ns}$ is the velocity of propagation in free space.

In general, the velocity of propagation of a TEM wave on a lossless transmission line embedded in a homogeneous dielectric medium is independent of the geometry of the line and depends only on the material properties of the dielectric medium. The velocity of propagation is reduced from the free-space velocity c by the factor $1/\sqrt{\epsilon_r}$, which is also called the *velocity factor* and is typically given in percent.

For transmission lines with inhomogeneous or mixed dielectrics, such as the microstrip shown in Fig. 6.1c, the velocity of propagation depends on both the cross-sectional geometry of the line and the dielectric constants of the dielectric media. In this case, the electromagnetic wave propagating on the line is not strictly TEM, but for many practical applications can be approximated as a quasi-TEM wave. To extend Eq. (6.18) to transmission lines with mixed dielectrics, the inhomogeneous dielectric is replaced with a homogeneous dielectric of *effective dielectric constant* ϵ_{eff} giving the same capacitance per unit length as the actual structure. The effective dielectric constant is obtained as the ratio of the actual distributed capacitance C of the line to the capacitance of the same structure but with all dielectrics replaced with air:

$$\epsilon_{\mathrm{eff}} = \frac{C}{C_{\mathrm{air}}} \tag{6.19}$$

The velocity of propagation of the quasi-TEM wave can be expressed with Eq. (6.19) as

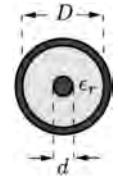
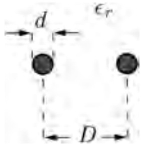
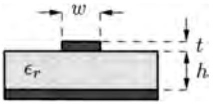
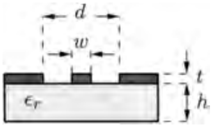
$$v_p = \frac{1}{\sqrt{\mu_0 \epsilon_0 \epsilon_{\mathrm{eff}}}} = \frac{c}{\sqrt{\epsilon_{\mathrm{eff}}}} \tag{6.20}$$

In general, the effective dielectric constant needs to be computed numerically; however, approximate closed-form expressions are available for many common transmission-line structures. As an example, a simple approximate closed-form expression for the effective dielectric constant of a microstrip of width w, substrate height h, and dielectric constant ϵ_r is given by [6]

$$\epsilon_{\mathrm{eff}} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2} \frac{1}{\sqrt{1 + 10h/w}} \tag{6.21}$$

Various closed-form approximations of the transmission-line parameters for many common planar transmission lines have been developed and can be found in the literature including Refs. 6 and 7. Table 6.1 gives the transmission-line parameters in exact or approximate closed form for several common types of transmission lines (assuming no losses).

Table 6.1 Transmission-line Parameters for Several Common Types of Transmission Lines

Transmission line	Parameters
 Coaxial line	$L = \frac{\mu_0}{2\pi}\ln(D/d)$ $C = \frac{2\pi\epsilon_0\epsilon_r}{\ln(D/d)}$ $Z_0 = \frac{1}{2\pi}\sqrt{\frac{\mu_0}{\epsilon_0\epsilon_r}}\ln(D/d)$ $\epsilon_{\text{eff}} = \epsilon_r$
 Two-wire line	$L = \frac{\mu_0}{\pi}\cosh^{-1}(D/d)$ $C = \frac{\pi\epsilon_0\epsilon_r}{\cosh^{-1}(D/d)}$ $Z_0 = \frac{1}{\pi}\sqrt{\frac{\mu_0}{\epsilon_0\epsilon_r}}\cosh^{-1}(D/d)$ $\epsilon_{\text{eff}} = \epsilon_r$
 Microstrip	$\epsilon_{\text{eff}} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2}\frac{1}{\sqrt{1 + 10h/w}}$ $Z_0 = \begin{cases} \frac{60}{\sqrt{\epsilon_{\text{eff}}}}\ln\left(\frac{8h}{w} + \frac{w}{4h}\right) & \text{for } w/h \leq 1 \\ \frac{120\pi}{F\sqrt{\epsilon_{\text{eff}}}} & \text{for } w/h \geq 1 \end{cases}$ $F = w/h + 2.42 - 0.44h/w + (1 - h/w)^6$ $t \to 0$ [6]
 Coplanar waveguide	$\epsilon_{\text{eff}} = 1 + \frac{(\epsilon_r - 1)K(k_1')K(k')}{2K(k_1)K(k)}$ $k_1' = \sqrt{1 - k_1^2} = \frac{\sinh[\pi w/(4h)]}{\sinh[\pi d/(4h)]}$ $k' = \sqrt{1 - k^2} = \sqrt{(1 - (w/d)^2)}$ $Z_0 = \frac{30}{\sqrt{\epsilon_{\text{eff}}}}\frac{K(k')}{K(k)}$ $t \to 0$ [6] ($K(k)$ is the elliptical integral of the first kind)

6.3. TRANSIENT RESPONSE OF LOSSLESS TRANSMISSION LINES

A practical transmission line is of finite length and is necessarily terminated. Consider a transmission-line circuit consisting of a section of lossless transmission line that is connected to a source and terminated in a load, as illustrated in Fig. 6.4. The response of the transmission-line circuit depends on the transmission-line characteristics as well as the characteristics of the source and terminating load. The ideal transmission line of finite

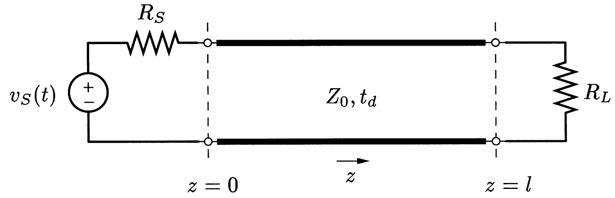


Figure 6.4 Lossless transmission line with resistive Thévenin equivalent source and resistive termination.

length is completely specified by the distributed L and C parameters and line length l, or, equivalently, by its characteristic impedance $Z_0 = \sqrt{L/C}$ and delay time

$$t_d = \frac{l}{v_p} = l\sqrt{LC} \tag{6.22}$$

of the line.* The termination imposes voltage and current boundary conditions at the end of the line, which may give rise to wave reflections.

6.3.1. Reflection Coefficient

When a traveling wave reaches the end of the transmission line, a reflected wave is generated unless the termination presents a load condition that is equal to the characteristic impedance of the line. The ratio of reflected voltage to incident voltage at the termination is defined as *voltage reflection coefficient* ρ, which for linear resistive terminations can be directly expressed in terms of the terminating resistance and the characteristic impedance of the line. The corresponding current reflection coefficient is given by $-\rho$. For the transmission-line circuit shown in Fig. 6.4 with resistive terminations, the voltage reflection coefficient at the termination with load resistance R_L is

$$\rho_L = \frac{R_L - Z_0}{R_L + Z_0} \tag{6.23}$$

Similarly, the voltage reflection coefficient at the source end with source resistance R_S is

$$\rho_S = \frac{R_S - Z_0}{R_S + Z_0} \tag{6.24}$$

The inverse relationship between reflection coefficient ρ_L and load resistance R_L follows directly from Eg. (6.23) and is

$$R_L = \frac{1 + \rho_L}{1 - \rho_L} Z_0 \tag{6.25}$$

*The specification in terms of characteristic impedance and delay time is used, for example, in the standard SPICE model for an ideal transmission line [8].

It is seen from Eq. (6.23) or (6.24) that the reflection coefficient is positive for a termination resistance greater than the characteristic impedance, and it is negative for a termination resistance less than the characteristic impedance of the line. A termination resistance equal to the characteristic impedance produces no reflection ($\rho = 0$) and is called *matched termination*. For the special case of an open-circuit termination the voltage reflection coefficient is $\rho_{\mathrm{oc}} = +1$, while for a short-circuit termination the voltage reflection coefficient is $\rho_{\mathrm{sc}} = -1$.

6.3.2. Step Response

To illustrate the wave reflection process, the step-voltage response of an ideal transmission line connected to a Thévénin equivalent source and terminated in a resistive load, as shown in Fig. 6.4, is considered. The transient response for a step-voltage change with finite rise time can be obtained in a similar manner. The step-voltage response of a *lossy* transmission line with constant or frequency-dependent line parameters is more complex and can be determined using the Laplace transformation [5].

The source voltage $v_S(t)$ in the circuit in Fig. 6.4 is assumed to be a step-voltage given by

$$v_S(t) = V_0 U(t) \tag{6.26}$$

where

$$U(t) = \begin{cases} 1 & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases} \tag{6.27}$$

The transient response due to a rectangular pulse $v_{\mathrm{pulse}}(t)$ of duration T can be obtained as the superposition of two step responses given as $v_{\mathrm{pulse}}(t) = V_0 U(t) - V_0 U(t - T)$.

The step-voltage change launches a forward traveling wave at the input of the line at time $t = 0$. Assuming no initial charge or current on the line, this first wave component presents a resistive load to the generator that is equal to the characteristic impedance of the line. The voltage of the first traveling wave component is

$$v_1^+(z, t) = V_0 \frac{Z_0}{R_S + Z_0} U\left(t - \frac{z}{v_p}\right) = V_1^+ U\left(t - \frac{z}{v_p}\right) \tag{6.28}$$

where v_p is the velocity of propagation on the line. For a nonzero reflection coefficient ρ_L at the termination, a reflected wave is generated when the first traveling wave arrives at the termination at time $t = t_d = l/v_p$. If the reflection coefficients at both the source and the termination are nonzero, an infinite succession of reflected waves results. The total voltage response on the line is the superposition of all traveling-wave components and is given by

$$\begin{aligned} v(z,t) = \frac{Z_0}{R_S+Z_0} V_0 \Bigg[& U\left(t-\frac{z}{v_p}\right) + \rho_L U\left(t-2t_d+\frac{z}{v_p}\right) \\ & + \rho_S\rho_L U\left(t-2t_d-\frac{z}{v_p}\right) + \rho_S\rho_L^2 U\left(t-4t_d+\frac{z}{v_p}\right) \\ & + \rho_S^2\rho_L^2 U\left(t-4t_d-\frac{z}{v_p}\right) + \rho_S^2\rho_L^3 U\left(t-6t_d+\frac{z}{v_p}\right) \\ & + \cdots \Bigg] \end{aligned} \tag{6.29}$$

Similarly, the total current on the line is given by

$$\begin{aligned} i(z,t) = \frac{V_0}{R_S+Z_0} \Bigg[& U\left(t-\frac{z}{v_p}\right) - \rho_L U\left(t-2t_d+\frac{z}{v_p}\right) \\ & + \rho_S\rho_L U\left(t-2t_d-\frac{z}{v_p}\right) - \rho_S\rho_L^2 U\left(t-4t_d+\frac{z}{v_p}\right) \\ & + \rho_S^2\rho_L^2 U\left(t-4t_d-\frac{z}{v_p}\right) - \rho_S^2\rho_L^3 U\left(t-6t_d+\frac{z}{v_p}\right) \\ & + \cdots \Bigg] \end{aligned} \tag{6.30}$$

The reflected wave components on the lossless transmission line are successively delayed copies of the first traveling-wave component with amplitudes appropriately adjusted by the reflection coefficients. Equations (6.29) and (6.30) show that at any given time and location on the line only a finite number of wave components have been generated. For example, for $t = 3t_d$ three wave components exist at the input of the line (at $z = 0$) and four wave components exist at the load (at $z = l$).

Unless both reflection coefficients have unity magnitudes, the amplitudes of the successive wave components become progressively smaller in magnitude and the infinite summations in Eqs. (6.29) and (6.30) converge to the dc values for $t \to \infty$. The steady-state (dc) voltage V_∞ is obtained by summing the amplitudes of all traveling-wave components for $t \to \infty$.

$$\begin{aligned} V_\infty = v(z, t \to \infty) &= \frac{Z_0}{R_S+Z_0} V_0 \{1 + \rho_L + \rho_S\rho_L + \rho_S\rho_L^2 + \rho_S^2\rho_L^2 + \cdots\} \\ &= \frac{Z_0}{R_S+Z_0} V_0 \frac{1+\rho_L}{1-\rho_S\rho_L} \end{aligned} \tag{6.31}$$

The steady-state voltage can also be directly obtained as the dc voltage drop across the load after removing the lossless line, that is

$$V_\infty = \frac{R_L}{R_S+R_L} V_0 \tag{6.32}$$

The steady-state current is

$$I_\infty = \frac{V_0}{R_S + R_L} \tag{6.33}$$

6.3.3. Lattice Diagram

The *lattice diagram* (also called *bounce* or *reflection diagram*) provides a convenient graphical means for keeping track of the multiple wave reflections on the line. The general lattice diagram is illustrated in Fig. 6.5. Each wave component is represented by a sloped line segment that shows the time elapsed after the initial voltage change at the source as a function of distance z on the line. For bookkeeping purposes, the value of the voltage amplitude of each wave component is commonly written above the corresponding line segment and the value of the accompanying current is added below. Starting with voltage $V_1^+ = V_0Z_0/(R_S + Z_0)$ of the first wave component, the voltage amplitude of each successive wave is obtained from the voltage of the preceding wave by multiplication with the appropriate reflection coefficient ρ_L or ρ_S in accordance with Eq. (6.29). Successive current values are obtained by multiplication with $-\rho_L$ or $-\rho_S$, as shown in Eq. (6.30).

The lattice diagram may be conveniently used to determine the voltage and current distributions along the transmission line at any given time or to find the time response at any given position. The variation of voltage and current as a function of time at a given position $z = z_1$ is found from the intersection of the vertical line through z_1 and the sloped line segments representing the wave components. Figure 6.5 shows the first five wave intersection times at position z_1 marked as t_1, t_2, t_3, t_4, and t_5, respectively. At each

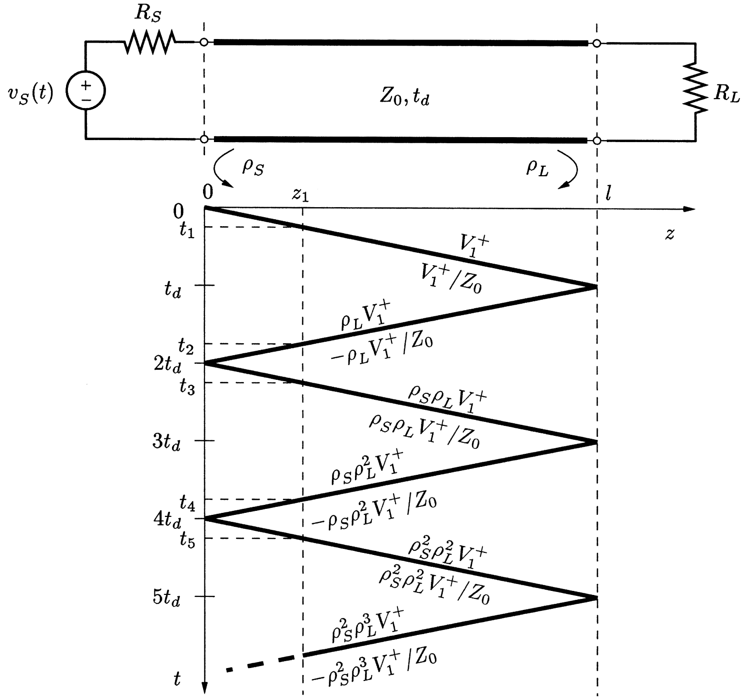


Figure 6.5 Lattice diagram for a lossless transmission line with unmatched terminations.

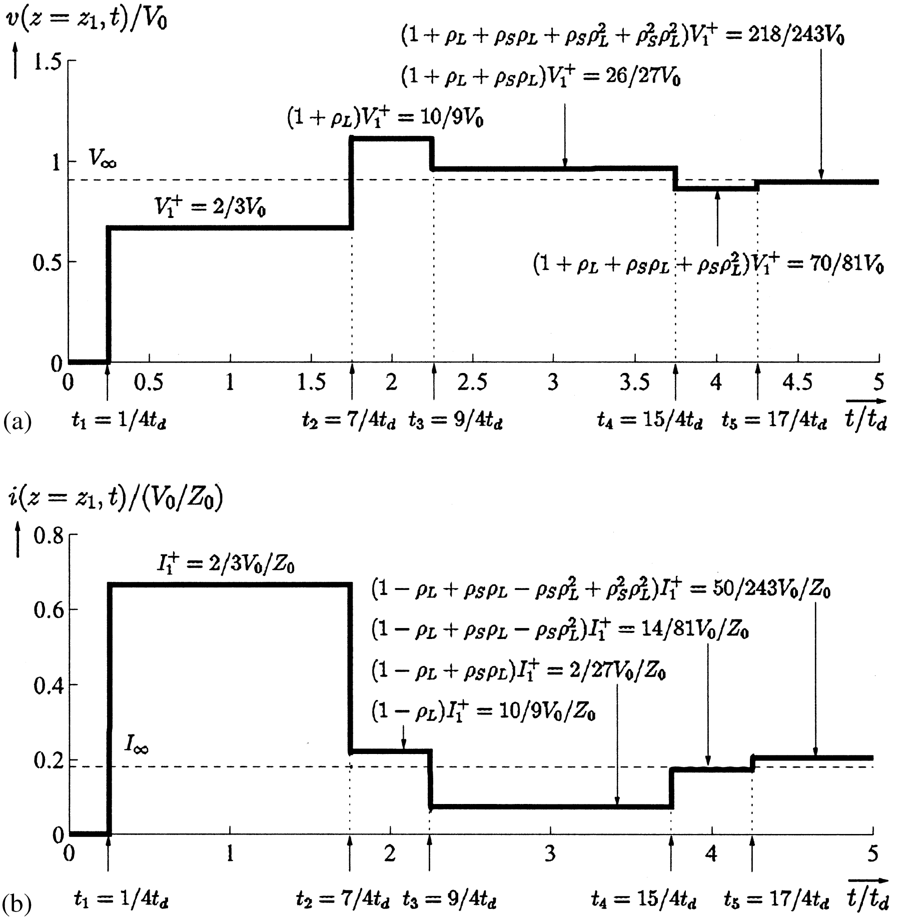


Figure 6.6 Step response of a lossless transmission line at $z = z_1 = l/4$ for $R_S = Z_0/2$ and $R_L = 5Z_0$; (a) voltage response, (b) current response.

intersection time, the total voltage and current change by the amplitudes specified for the intersecting wave component. The corresponding transient response for voltage and current with $R_S = Z_0/2$ and $R_L = 5Z_0$ corresponding to reflection coefficients $\rho_S = -1/3$ and $\rho_L = 2/3$, respectively, is shown in Fig. 6.6. The transient response converges to the steady-state $V_\infty = 10/11\ V_0$ and $I_\infty = 2/11(V_0/Z_0)$, as indicated in Fig. 6.6.

6.3.4. Applications

In many practical applications, one or both ends of a transmission line are matched to avoid multiple reflections. If the source and/or the receiver do not provide a match, multiple reflections can be avoided by adding an appropriate resistor at the input of the line (source termination) or at the end of the line (end termination) [9,10]. Multiple reflections on the line may lead to signal distortion including a slow voltage buildup or signal overshoot and ringing.

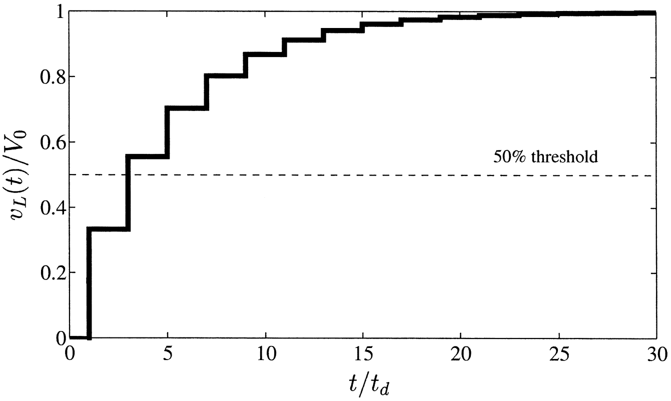


Figure 6.7 Step-voltage response at the termination of an open-circuited lossless transmission line with $R_S = 5Z_0$ ($\rho_S = 2/3$).

Over- and Under-driven Transmission Lines

In high-speed digital systems, the input of a receiver circuit typically presents a load to a transmission line that is approximately an open circuit (unterminated). The step-voltage response of an unterminated transmission line may exhibit a considerably different behavior depending on the source resistance.

If the source resistance is larger than the characteristic impedance of the line, the voltage across the load will build up monotonically to its final value since both reflection coefficients are positive. This condition is referred to as an *underdriven* transmission line. The buildup time to reach a sufficiently converged voltage may correspond to many round-trip times if the reflection coefficient at the source is close to $+1$ (and $\rho_L = \rho_{\text{oc}} = +1$), as illustrated in Fig. 6.7. As a result, the effective signal delay may be several times longer than the delay time of the line.

If the source resistance is smaller than the characteristic impedance of the line, the initial voltage at the unterminated end will exceed the final value (overshoot). Since the source reflection coefficient is negative and the load reflection coefficient is positive, the voltage response will exhibit ringing as the voltage converges to its final value. This condition is referred to as an *overdriven* transmission line. It may take many round-trip times to reach a sufficiently converged voltage (long settling time) if the reflection coefficient at the source is close to -1 (and $\rho_L = \rho_{\text{oc}} = +1$), as illustrated in Fig. 6.8. An overdriven line can produce excessive noise and cause intersymbol interference.

Transmission-line Junctions

Wave reflections occur also at the junction of two tandem-connected transmission lines having different characteristic impedances. This situation, illustrated in Fig. 6.9a, is often encountered in practice. For an incident wave on line 1 with characteristic impedance $Z_{0,1}$, the second line with characteristic impedance $Z_{0,2}$ presents a load resistance to line 1 that is equal to $Z_{0,2}$. At the junction, a reflected wave is generated on line 1 with voltage reflection coefficient ρ_{11} given by

$$\rho_{11} = \frac{Z_{0,2} - Z_{0,1}}{Z_{0,2} + Z_{0,1}} \tag{6.34}$$

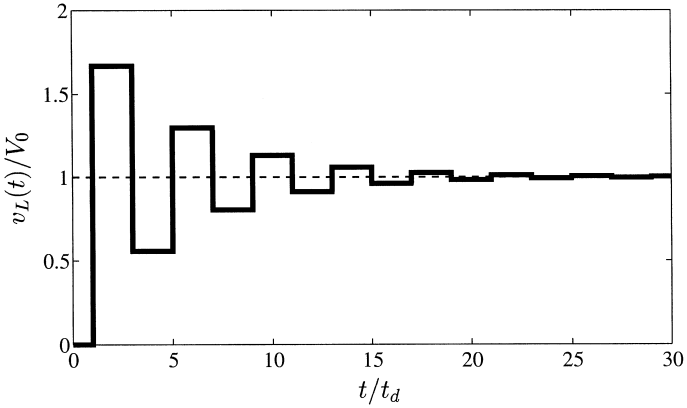


Figure 6.8 Step-voltage response at the termination of an open-circuited lossless transmission line with $R_S = Z_0/5$ ($\rho_S = -2/3$).

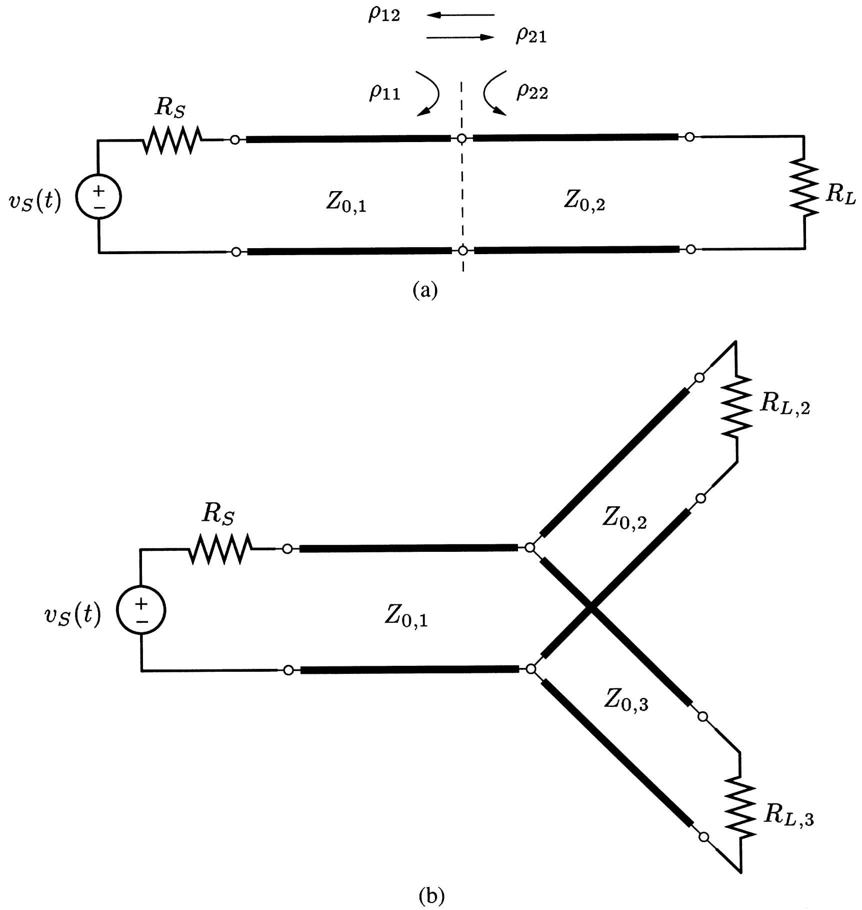


Figure 6.9 Junction between transmission lines: (a) two tandem-connected lines and (b) three parallel-connected lines.

In addition, a wave is launched on the second line departing from the junction. The voltage amplitude of the transmitted wave is the sum of the voltage amplitudes of the incident and reflected waves on line 1. The ratio of the voltage amplitudes of the transmitted wave on line 2 to the incident wave on line 1 is defined as the *voltage transmission coefficient* ρ_{21} and is given by

$$\rho_{21} = 1 + \rho_{11} = \frac{2Z_{0,2}}{Z_{0,1} + Z_{0,2}} \tag{6.35}$$

Similarly, for an incident wave from line 2, the reflection coefficient ρ_{22} at the junction is

$$\rho_{22} = \frac{Z_{0,1} - Z_{0,2}}{Z_{0,1} + Z_{0,2}} = -\rho_{11} \tag{6.36}$$

The voltage transmission coefficient ρ_{12} for a wave incident from line 2 and transmitted into line 1 is

$$\rho_{12} = 1 + \rho_{22} = \frac{2Z_{0,1}}{Z_{0,1} + Z_{0,2}} \tag{6.37}$$

If in addition lumped elements are connected at the junction or the transmission lines are connected through a resistive network, the reflection and transmission coefficients will change, and in general, $\rho_{ij} \leq 1 + \rho_{jj}$ [5].

For a parallel connection of multiple lines at a common junction, as illustrated in Fig. 6.9b, the effective load resistance is obtained as the parallel combination of the characteristic impedances of all lines except for the line carrying the incident wave. The reflection and transmission coefficients are then determined as for tandem connected lines [5].

The wave reflection and transmission process for tandem and multiple parallel-connected lines can be represented graphically with a lattice diagram for each line. The complexity, however, is significantly increased over the single line case, in particular if multiple reflections exist.

Reactive Terminations

In various transmission-line applications, the load is not purely resistive but has a reactive component. Examples of reactive loads include the capacitive input of a CMOS gate, pad capacitance, bond-wire inductance, as well as the reactance of vias, package pins, and connectors [9,10]. When a transmission line is terminated in a reactive element, the reflected waveform will not have the same shape as the incident wave, i.e., the reflection coefficient will not be a constant but be varying with time. For example, consider the step response of a transmission line that is terminated in an uncharged capacitor C_L. When the incident wave reaches the termination, the initial response is that of a short circuit, and the response after the capacitor is fully charged is an open circuit. Assuming the source end is matched to avoid multiple reflections, the incident step-voltage wave is $v_1^+(t) = V_0/2U(t - z/v_p)$. The voltage across the capacitor changes exponentially from the initial voltage $v_{\text{cap}} = 0$ (short circuit) at time $t = t_d$ to the final voltage

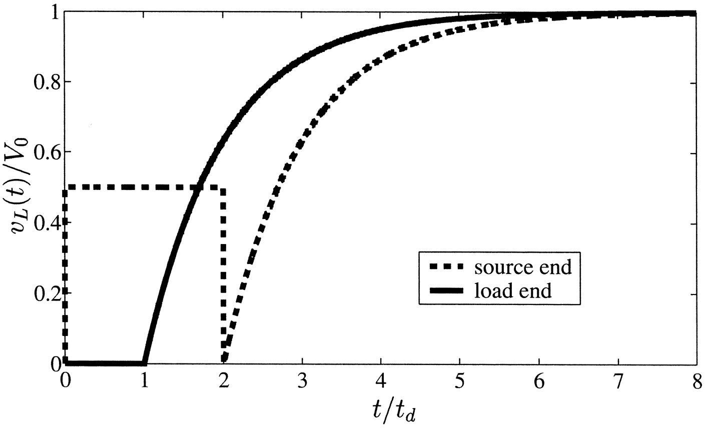


Figure 6.10 Step-voltage response of a transmission line that is matched at the source and terminated in a capacitor C_L with time constant $\tau = Z_0 C_L = t_d$.

$v_{\text{cap}}(t \to \infty) = V_0$ (open circuit) as

$$v_{\text{cap}}(t) = V_0\left[1 - e^{-(t-t_d)/\tau}\right]U(t - t_d) \tag{6.38}$$

with time constant

$$\tau = Z_0 C_L \tag{6.39}$$

where Z_0 is the characteristic impedance of the line. Figure 6.10 shows the step-voltage response across the capacitor and at the source end of the line for $\tau = t_d$.

If the termination consists of a parallel combination of a capacitor C_L and a resistor R_L, the time constant is obtained as the product of C_L and the parallel combination of R_L and characteristic impedance Z_0. For a purely inductive termination L_L, the initial response is an open circuit and the final response is a short circuit. The corresponding time constant is $\tau = L_L/Z_0$.

In the general case of reactive terminations with multiple reflections or with more complicated source voltages, the boundary conditions for the reactive termination are expressed in terms of a differential equation. The transient response can then be determined mathematically, for example, using the Laplace transformation [11].

Nonlinear Terminations

For a nonlinear load or source, the reflected voltage and subsequently the reflection coefficient are a function of the cumulative voltage and current at the termination including the contribution of the reflected wave to be determined. Hence, the reflection coefficient for a nonlinear termination cannot be found from only the termination characteristics and the characteristic impedance of the line. The step-voltage response for each reflection instance can be determined by matching the *I–V* characteristics of the termination and the cumulative voltage and current characteristics at the end of the transmission line. This solution process can be constructed using a graphical technique known as the *Bergeron method* [5,12] and can be implemented in a computer program.

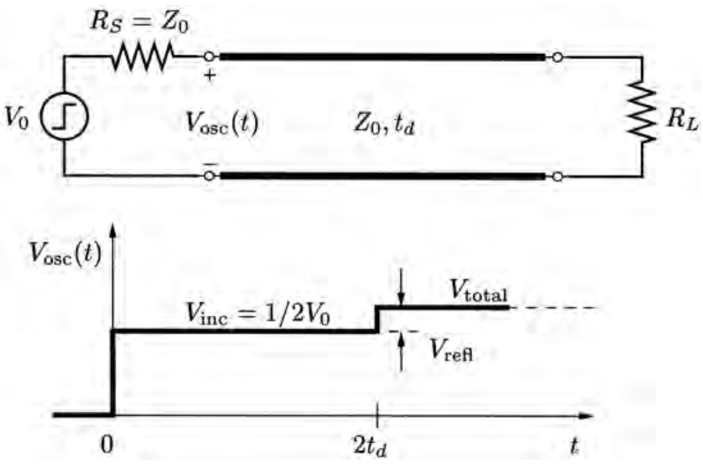


Figure 6.11 Illustration of the basic principle of time-domain reflectometry (TDR).

Time-Domain Reflectometry

Time-domain reflectometry (TDR) is a measurement technique that utilizes the information contained in the reflected waveform and observed at the source end to test, characterize, and model a transmission-line circuit. The basic TDR principle is illustrated in Fig. 6.11. A TDR instrument typically consists of a precision step-voltage generator with a known source (reference) impedance to launch a step wave on the transmission-line circuit under test and a high impedance probe and oscilloscope to sample and display the voltage waveform at the source end. The source end is generally well matched to establish a reflection-free reference. The voltage at the input changes from the initial incident voltage when a reflected wave generated at an impedance discontinuity such as a change in line impedance, a line break, an unwanted parasitic reactance, or an unmatched termination reaches the source end of the transmission line-circuit.

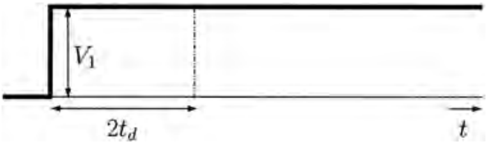
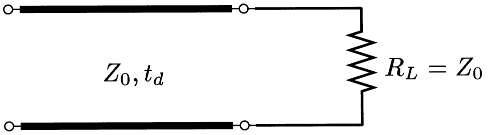
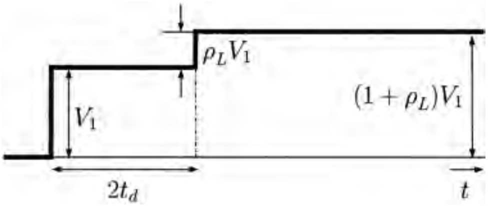
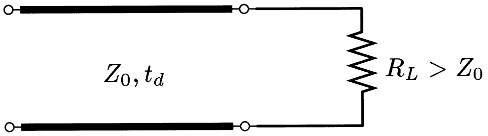
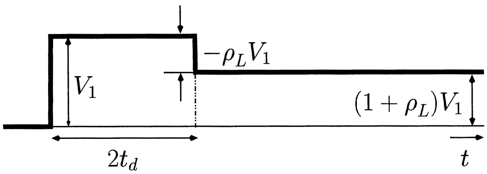
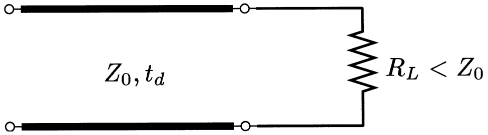
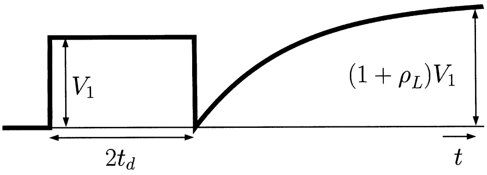
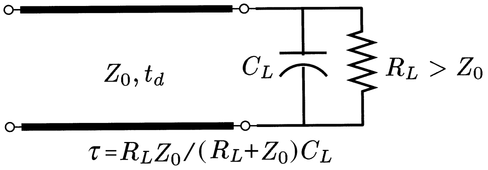
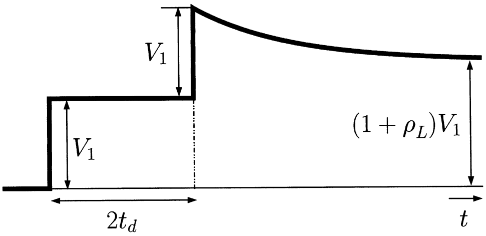
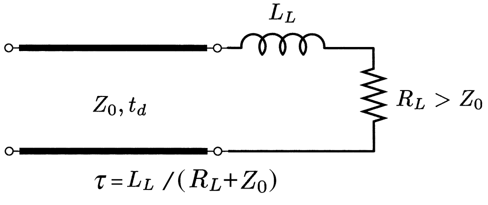
The time elapsed between the initial launch of the step wave and the observation of the reflected wave at the input corresponds to the round-trip delay $2t_d$ from the input to the location of the impedance mismatch and back. The round-trip delay time can be converted to find the distance from the input of the line to the location of the impedance discontinuity if the propagation velocity is known. The capability of measuring distance is used in TDR cable testers to locate faults in cables. This measurement approach is particularly useful for testing long, inaccessible lines such as underground or undersea electrical cables.

The reflected waveform observed at the input also provides information on the type of discontinuity and the amount of impedance change. Table 6.2 shows the TDR response for several common transmission-line discontinuities. As an example, the load resistance in the circuit in Fig. 6.11 is extracted from the incident and reflected or total voltage observed at the input as

$$R_L = Z_0 \frac{1+\rho}{1-\rho} = Z_0 \frac{V_{\text{total}}}{2V_{\text{incident}} - V_{\text{total}}} \tag{6.40}$$

where $\rho = V_{\text{reflected}}/V_{\text{incident}} = (R_L - Z_0)/(R_L + Z_0)$ and $V_{\text{total}} = V_{\text{incident}} + V_{\text{reflected}}$.

Table 6.2 TDR Responses for Typical Transmission-line Discontinuities.

TDR response	Circuit
	
	
	
	 Z_0, t_d C_L $R_L > Z_0$ $\tau = R_L Z_0/(R_L + Z_0)C_L$
	 L_L Z_0, t_d $R_L > Z_0$ $\tau = L_L/(R_L + Z_0)$

The TDR principle can be used to profile impedance changes along a transmission line circuit such as a trace on a printed-circuit board. In general, the effects of multiple reflections arising from the impedance mismatches along the line need to be included to extract the impedance profile. If the mismatches are small, higher-order reflections can be ignored and the same extraction approach as for a single impedance discontinuity can be applied for each discontinuity. The resolution of two closely spaced discontinuities, however, is limited by the rise time of step voltage and the overall rise time of the TDR system. Further information on using time-domain reflectometry for analyzing and modeling transmission-line systems is given e.g. in Refs. 10,11,13–15.

6.4. SINUSOIDAL STEADY-STATE RESPONSE OF TRANSMISSION LINES

The steady-state response of a transmission line to a sinusoidal excitation of a given frequency serves as the fundamental solution for many practical transmission-line applications including radio and television broadcast and transmission-line circuits operating at microwave frequencies. The frequency-domain information also provides physical insight into the signal propagation on the transmission line. In particular, transmission-line losses and any frequency dependence in the R, L, G, C line parameters can be readily taken into account in the frequency-domain analysis of transmission lines. The time-domain response of a transmission-line circuit to an arbitrary time-varying excitation can then be obtained from the frequency-domain solution by applying the concepts of Fourier analysis [16].

As in standard circuit analysis, the time-harmonic voltage and current on the transmission line are conveniently expressed in phasor form using Euler's identity $e^{j\theta} = \cos\theta + j\sin\theta$. For a cosine reference, the relations between the voltage and current phasors, $V(z)$ and $I(z)$, and the time-harmonic space–time-dependent quantities, $v(z,t)$ and $i(z,t)$, are

$$v(z,t) = \mathrm{Re}\{V(z)e^{j\omega t}\} \tag{6.41}$$

$$i(z,t) = \mathrm{Re}\{I(z)e^{j\omega t}\} \tag{6.42}$$

The voltage and current phasors are functions of position z on the transmission line and are in general complex.

6.4.1. Characteristics of Lossy Transmission Lines

The transmission-line equations, (general telegrapher's equations) in phasor form for a general lossy transmission line can be derived directly from the equivalent circuit for a short line section of length $\Delta z \to 0$ shown in Fig. 6.12. They are

$$-\frac{dV(z)}{dz} = (R + j\omega L)I(z) \tag{6.43}$$

$$-\frac{dI(z)}{dz} = (G + j\omega C)V(z) \tag{6.44}$$

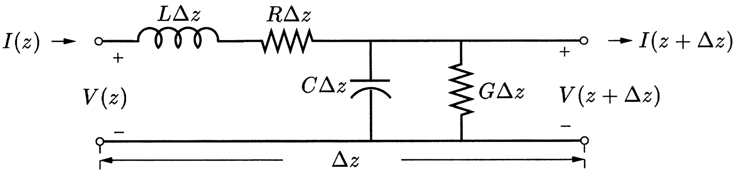


Figure 6.12 Equivalent circuit model for a short section of lossy transmission line of length Δz with R, L, G, C line parameters.

The transmission-line equations, Eqs. (6.43) and (6.44) can be combined to the complex wave equation for voltage (and likewise for current)

$$\frac{d^2V(z)}{dz^2} = (R + j\omega L)(G + j\omega C)V(z) = \gamma^2 V(z) \tag{6.45}$$

The general solution of Eq. (6.45) is

$$V(z) = V^+(z) + V^-(z) = V_0^+ e^{-\gamma z} + V_0^- e^{+\gamma z} \tag{6.46}$$

where γ is the *propagation constant* of the transmission line and is given by

$$\gamma = \alpha + j\beta = \sqrt{(R + j\omega L)(G + j\omega C)} \tag{6.47}$$

and $V_0^+ = |V_0^+|e^{j\phi^+}$ and $V_0^- = |V_0^-|e^{j\phi^-}$ are complex constants. The real time-harmonic voltage waveforms $v(z,t)$ corresponding to phasor $V(z)$ are obtained with Eq. (6.41) as

$$\begin{aligned} v(z,t) &= v^+(z,t) + v^-(z,t) \\ &= |V_0^+|e^{-\alpha z}\cos(\omega t - \beta z + \phi^+) + |V_0^-|e^{\alpha z}\cos(\omega t + \beta z + \phi^-) \end{aligned} \tag{6.48}$$

and are illustrated in Fig. 6.13.

The real part α of the propagation constant in Eq. (6.47) is known as the *attenuation constant* measured in nepers per unit length (Np/m) and gives the rate of exponential attenuation of the voltage and current amplitudes of a traveling wave.* The imaginary part of γ is the *phase constant* $\beta = 2\pi/\lambda$ measured in radians per unit length (rad/m), as in the lossless line case. The corresponding *phase velocity* of the time-harmonic wave is given by

$$v_p = \frac{\omega}{\beta} \tag{6.49}$$

which depends in general on frequency. Transmission lines with frequency-dependent phase velocity are called *dispersive* lines. Dispersive transmission lines can lead to signal distortion, in particular for broadband signals.

The current phasor $I(z)$ associated with voltage $V(z)$ in Eq. (6.46) is found with Eq. (6.43) as

$$I(z) = \frac{V^+}{Z_0}e^{-\gamma z} - \frac{V^-}{Z_0}e^{+\gamma z} \tag{6.50}$$

*The amplitude attenuation of a traveling wave $V^+(z) = V_0^+ e^{-\gamma z} = V_0^+ e^{-\alpha z} e^{-j\beta z}$ over a distance l can be expressed in logarithmic form as $\ln|V^+(z)/V^+(z+l)| = \alpha l$ (nepers). To convert from the attenuation measured in nepers to the logarithmic measure $20\log_{10}|V^+(z)/V^+(z+l)|$ in dB, the attenuation in nepers is multiplied by $20\log_{10} e \approx 8.686$ (1 Np corresponds to about 8.686 dB). For coaxial cables the attenuation constant is typically specified in units of dB/100 ft. The conversion to Np/m is 1 dB/100 ft $\approx$ 0.0038 Np/m.

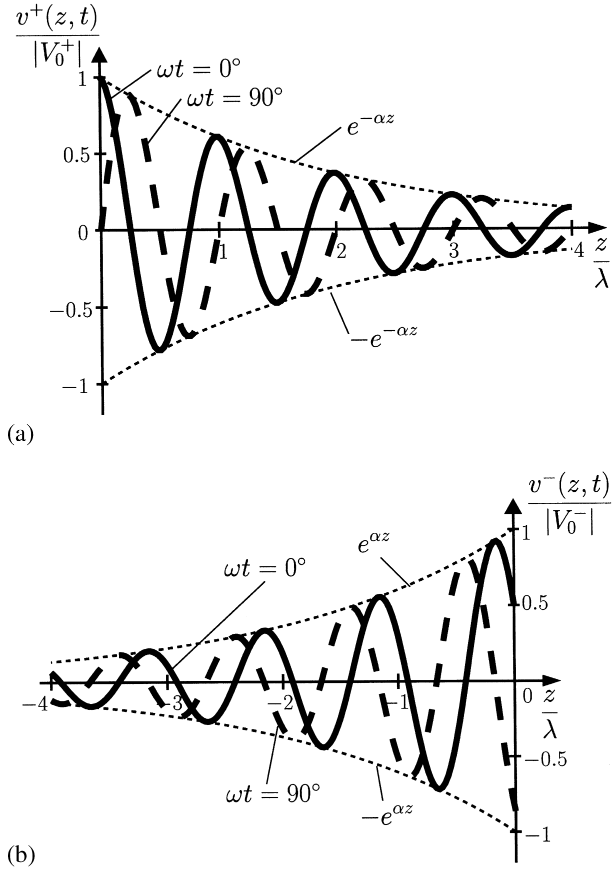


Figure 6.13 Illustration of a traveling wave on a lossy transmission line: (a) wave traveling in $+z$ direction with $\phi^+ = 0$ and $\alpha = 1/(2\lambda)$ and (b) wave traveling in $-z$ direction with $\phi^- = 60°$ and $\alpha = 1/(2\lambda)$.

The quantity Z_0 is defined as the characteristic impedance of the transmission line and is given in terms of the line parameters by

$$Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \tag{6.51}$$

As seen from Eq. (6.51), the characteristic impedance is in general complex and frequency dependent.

The inverse expressions relating the R, L, G, C line parameters to the characteristic impedance and propagation constant of a transmission line are found from Eqs. (6.47) and (6.51) as

$$R + j\omega L = \gamma Z_0 \tag{6.52}$$

$$G + j\omega C = \gamma / Z_0 \tag{6.53}$$

These inverse relationships are particularly useful for extracting the line parameters from experimentally determined data for characteristic impedance and propagation constant.

Special Cases

For a lossless line with $R=0$ and $G=0$, the propagation constant is $\gamma = j\omega\sqrt{LC}$. The attenuation constant α is zero and the phase velocity is $v_p = \omega/\beta = 1/\sqrt{LC}$. The characteristic impedance of a lossless line is $Z_0 = \sqrt{L/C}$, as in Eq. (6.14).

In general, for a lossy transmission line both the attenuation constant and the phase velocity are frequency dependent, which can give rise to signal distortion.* However, in many practical applications the losses along the transmission line are small. For a low loss line with $R \ll \omega L$ and $G \ll \omega C$, useful approximate expressions can be derived for the characteristic impedance Z_0 and propagation constant γ as

$$Z_0 \approx \sqrt{\frac{L}{C}}\left[1 - j\frac{1}{2\omega}\left(\frac{R}{L} - \frac{G}{C}\right)\right] \tag{6.54}$$

and

$$\gamma \approx \frac{R}{2}\sqrt{\frac{C}{L}} + \frac{G}{2}\sqrt{\frac{L}{C}} + j\omega\sqrt{LC} \tag{6.55}$$

The low-loss conditions $R \ll \omega L$ and $G \ll \omega C$ are more easily satisfied at higher frequencies.

6.4.2. Terminated Transmission lines

If a transmission line is terminated with a load impedance that is different from the characteristic impedance of the line, the total time-harmonic voltage and current on the line will consist of two wave components traveling in opposite directions, as given by the general phasor expressions in Eqs. (6.46) and (6.50). The presence of the two wave components gives rise to standing waves on the line and affects the line's input impedance.

Impedance Transformation

Figure 6.14 shows a transmission line of finite length terminated with load impedance Z_L. In the steady-state analysis of transmission-line circuits it is expedient to measure distance on the line from the termination with known load impedance. The distance on the line from the termination is given by z'. The line voltage and current at distance z' from the

*For the special case of a line satisfying the condition $R/L = G/C$, the characteristic impedance $Z_0 = \sqrt{L/C}$, the attenuation constant $\alpha = R/\sqrt{L/C}$, and the phase velocity $v_p = 1/\sqrt{LC}$ are frequency independent. This type of line is called a *distortionless line*. Except for a constant signal attenuation, a distortionless line behaves like a lossless line.

termination can be related to voltage $V_L = V(z' = 0)$ and current $I_L = I(z' = 0)$ at the termination as

$$V(z') = V_L \cosh \gamma z' + I_L Z_0 \sinh \gamma z' \tag{6.56}$$

$$I(z') = V_L \left(\frac{1}{Z_0}\right) \sinh \gamma z' + I_L \cosh \gamma z' \tag{6.57}$$

where $V_L/I_L = Z_L$. These voltage and current transformations between the input and output of a transmission line of length z' can be conveniently expressed in $ABCD$ matrix form as*

$$\begin{bmatrix} V(z') \\ I(z') \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} V(0) \\ I(0) \end{bmatrix} = \begin{bmatrix} \cosh(\gamma z') & Z_0 \sinh(\gamma z') \\ (1/Z_0)\sinh(\gamma z') & \cosh(\gamma z') \end{bmatrix} \begin{bmatrix} V(0) \\ I(0) \end{bmatrix} \tag{6.58}$$

The ratio $V(z')/I(z')$ defines the input impedance $Z_{\text{in}}(z')$ at distance z' looking toward the load. The input impedance for a general lossy line with characteristic impedance Z_0 and terminated with load impedance Z_L is

$$Z_{\text{in}}(z') = \frac{V(z')}{I(z')} = Z_0 \frac{Z_L + Z_0 \tanh \gamma z'}{Z_0 + Z_L \tanh \gamma z'} \tag{6.60}$$

It is seen from Eq. (6.60) that for a line terminated in its characteristic impedance ($Z_L = Z_0$), the input impedance is identical to the characteristic impedance, independent of distance z'. This property serves as an alternate definition of the characteristic impedance of a line and can be applied to experimentally determine the characteristic impedance of a given line.

The input impedance of a transmission line can be used advantageously to determine the voltage and current at the input terminals of a transmission-line circuit as well as the average power delivered by the source and ultimately the average power dissipated in the load. Figure 6.15 shows the equivalent circuit at the input (source end) for the transmission-line circuit in Fig. 6.14. The input voltage V_{in} and current I_{in} are easily

*The $ABCD$ matrix is a common representation for two-port networks and is particularly useful for cascade connections of two or more two-port networks. The overall voltage and current transformations for cascaded lines and lumped elements can be easily obtained by multiplying the corresponding $ABCD$ matrices of the individual sections [1]. For a lossless transmission line, the $ABCD$ parameters are

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}_{\text{lossless line}} = \begin{bmatrix} \cos\theta & jZ_0 \sin\theta \\ (j/Z_0)\sin\theta & \cos\theta \end{bmatrix} \tag{6.59}$$

where $\theta = \beta z'$ is the electrical length of the line segment.

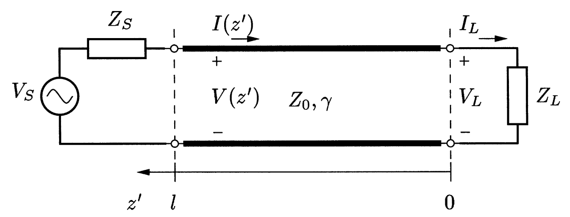


Figure 6.14 Transmission line of finite length terminated in load impedance Z_L.

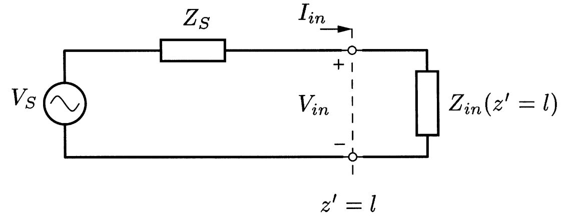


Figure 6.15 Equivalent circuit at the input of the transmission line circuit shown in Fig. 6.14.

determined from the voltage divider circuit. The average power delivered by the source to the input terminals of the transmission line is

$$P_{\text{ave, in}} = \frac{1}{2}\text{Re}\{V_{\text{in}}I^*_{\text{in}}\} \tag{6.61}$$

The average power dissipated in the load impedance $Z_L = R_L + jX_L$ is

$$P_{\text{ave}, L} = \frac{1}{2}\text{Re}\{V_L I^*_L\} = \frac{1}{2}|I_L|^2 R_L = \frac{1}{2}\left|\frac{V_L}{Z_L}\right|^2 R_L \tag{6.62}$$

where V_L and I_L can be determined from the inverse of the *ABCD* matrix transformation Eq. (6.58).* In general, $P_{\text{ave}, L} < P_{\text{ave, in}}$ for a lossy line and $P_{\text{ave}, L} = P_{\text{ave, in}}$ for a lossless line.

Example. Consider a 10 m long low-loss coaxial cable of nominal characteristic impedance $Z_0 = 75\,\Omega$, attenuation constant $\alpha = 2.2\,\text{dB}$ per 100 ft at 100 MHz, and velocity factor of 78%. The line is terminated in $Z_L = 100\,\Omega$, and the circuit is operated at $f = 100\,\text{MHz}$. The *ABCD* parameters for the transmission line are $A = D = -0.1477 + j0.0823$, $B = (-0.9181 + j74.4399)\,\Omega$, and $C = (-0.0002 + j0.0132)\,\Omega^{-1}$. The input impedance of the line is found as $Z_{\text{in}} = (59.3 + j4.24)\,\Omega$. For a source voltage $|V_S| = 10\,\text{V}$ and source impedance $Z_S = 75\,\Omega$, the average power delivered to the input of the line is $P_{\text{ave,in}} = 164.2\,\text{mW}$ and the average power dissipated in the load impedance is $P_{\text{ave},L} = 138.3\,\text{mW}$. The difference of 25.9 mW ($\approx 16\%$ of the input power) is dissipated in the transmission line.

*The inverse of Eq. (6.58) expressing the voltage and current at the load in terms of the input voltage and current is

$$\begin{bmatrix} V_L \\ I_L \end{bmatrix} = \begin{bmatrix} D & -B \\ -C & A \end{bmatrix} \begin{bmatrix} V_{\text{in}} \\ I_{\text{in}} \end{bmatrix} \tag{6.63}$$

Transmission Lines as Reactive Circuit Elements

In many practical transmission-line applications, transmission-line losses are small and often negligible. In particular, short sections of transmission lines used as circuit elements in high-frequency circuits are often assumed to be lossless.

For a lossless line with $\gamma = j\beta$ and terminated in a complex load impedance Z_L, the input impedance is

$$Z_{\text{in}}(\theta) = Z_0 \frac{Z_L + jZ_0 \tan\theta}{Z_0 + jZ_L \tan\theta} \tag{6.64}$$

where $\theta = \beta z' = 2\pi z'/\lambda$ is the electrical distance from the termination. Two particularly important special cases are the short-circuited line with $Z_L = 0$ and the open-circuited line with $Z_L \to \infty$.

The input impedance of an open-circuited lossless transmission line is

$$Z_{\text{oc}} = -jZ_0 \cot\theta = jX_{\text{oc}} \tag{6.65}$$

which is purely reactive. The normalized reactance is plotted in Fig. 6.16a. For small line lengths of less than a quarter wavelength ($\theta < 90°$), the input impedance is purely

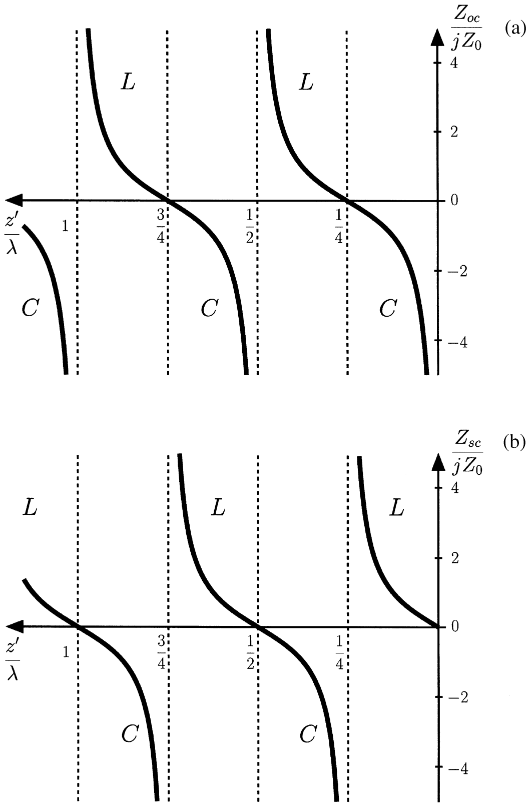


Figure 6.16 Normalized input reactance of a lossless transmission line terminated in (a) an open circuit and (b) a short circuit.

capacitive, as expected. With increasing electrical distance θ, the input reactance alternates every quarter wavelength between being capacitive and inductive. Any reactance value $-\infty < X_{\text{oc}} < +\infty$ can be achieved by appropriately adjusting the electrical length (i.e., by varying the physical length or the frequency (wavelength)). Furthermore, for line lengths corresponding to multiples of a half wavelength, the input impedance is again an open circuit. In contrast, for line lengths corresponding to odd multiples of a quarter wavelength, the input impedance is zero [$Z_{\text{oc}}(z' = \lambda/4 + n\lambda/2) = 0$, $n = 0, 1, 2, \ldots$].

The input impedance of a short-circuited lossless transmission line is also purely reactive and is given by

$$Z_{\text{sc}} = jZ_0 \tan\theta = jX_{\text{sc}} \tag{6.66}$$

Figure 6.16b shows the normalized reactance X_{sc}/Z_0 as a function of electrical length θ. For small line lengths of less than a quarter-wavelength ($\theta < 90^\circ$), the input impedance of a short-circuited line is purely inductive, as expected. The dependence of the input reactance of the short-circuited line on electrical length θ corresponds to that of the open-circuited line with a shift by a quarter wavelength. In particular, for line lengths corresponding to multiples of $\lambda/2$, the input impedance is zero, whereas for line lengths corresponding to odd multiples of $\lambda/4$, the input impedance of a short-circuited lossless line is an open circuit [$Z_{\text{sc}}(z' = \lambda/4 + n\lambda/2) \to \infty$, $n = 0, 1, 2, \ldots$].

An important application of open- and short-circuited transmission lines is the realization of reactive circuit elements for example for matching networks and filters, in particular at microwave frequencies ranging from a few gigahertz to tens of gigahertz.* At these frequencies, ordinary lumped elements become exceedingly small and difficult to realize and fabricate. In contrast, open- and short-circuited transmission-line sections with lengths on the order of a quarter wavelength become physically small enough to be realized at microwave frequencies and can be easily integrated in planar circuit technology. In practice, it is usually easier to make a good short-circuit termination than an open-circuit termination because of radiation from the open end and coupling to nearby conductors.

Example. To illustrate the design of reactive transmission-line segments, an equivalent inductance $L_{\text{eq}} = 5\,\text{nH}$ and an equivalent capacitance $C_{\text{eq}} = 2\,\text{pF}$ are realized

*Open- and short-circuit input impedance measurements for a general lossy transmission line can also be used to determine the transmission-line parameters. From $Z_{\text{oc}} = Z_0 \coth\gamma z'$ and $Z_{\text{sc}} = Z_0 \tanh\gamma z'$ for a lossy line follows

$$Z_0 = \sqrt{Z_{\text{oc}} Z_{\text{sc}}} \tag{6.67}$$

and

$$\tanh\gamma z' = \sqrt{\frac{Z_{\text{sc}}}{Z_{\text{oc}}}} \tag{6.68}$$

However, care should be taken in the extraction of $\gamma = \alpha + j\beta$ from Eq. (6.68) due to the periodicity of the phase term $\beta z'$, which must be approximately known.

at $f = 5\,\text{GHz}$ using a short-circuited 50-Ω microstrip line with effective dielectric constant $\epsilon_{\text{eff}} = 1.89$. From Eq. (6.66) follows

$$L_{\text{eq, sc}} \frac{Z_0 \tan \theta_L}{\omega} \tag{6.69}$$

$$C_{\text{eq, sc}} = -\frac{1}{\omega Z_0 \tan \theta_C} \tag{6.70}$$

The minimum electrical lengths for positive values for L_{eq} and C_{eq} are found as $\theta_L = 72.3^\circ$ ($l_L/\lambda = 0.201$) and $\theta_C = 162.3^\circ$ ($l_C/\lambda = 0.451$). With $\lambda = 4.36\,\text{cm}$ the corresponding physical lengths of the short-circuited microstrip segments are $l_L = 0.88\,\text{cm}$ and $l_C = 1.97\,\text{cm}$.

Complex Reflection Coefficient

The behavior of a terminated line is further examined in terms of incident and reflected waves at the termination. The ratio of the voltage phasors V^- and V^+ at the termination is defined as the voltage reflection coefficient $\Gamma_L = V^-/V^+$ and is given in terms of the load impedance Z_L and characteristic impedance Z_0 as

$$\Gamma_L = |\Gamma_L| e^{j\theta_L} = \frac{Z_L - Z_0}{Z_L + Z_0} \tag{6.71}$$

The load reflection coefficient Γ_L is in general complex. Here, a different symbol than in Eq. (6.23) is used to emphasize the definition of the complex reflection coefficient as ratio of voltage phasors. For a passive load $|\Gamma_L| \leq 1$. If the terminating load impedance equals the characteristic impedance of the line (matched termination), $\Gamma_L = 0$ and $V^- = 0$. For an open-circuit termination, $\Gamma_L = \Gamma_{\text{oc}} = +1$, while for a short-circuit termination, $\Gamma_L = \Gamma_{\text{sc}} = -1$. In general, for a purely reactive termination $Z_L = jX_L$ ($X_L > 0$ or $X_L < 0$) and real characteristic impedance, the magnitude of the reflection coefficient is $|\Gamma_L| = 1$.

Standing Waves

The total voltage and current along a lossless transmission line with $\gamma = j\beta$ can be expressed with reflection coefficient Γ_L at the termination as

$$V(z') = V_0^+ \{e^{+j\beta z'} + \Gamma_L e^{-j\beta z'}\} \tag{6.72}$$

$$I(z') = \frac{V^+}{Z_0} \{e^{+j\beta z'} - \Gamma_L e^{-j\beta z'}\} \tag{6.73}$$

The superposition of the two opposing traveling wave components leads to periodic variations in voltage and current along the line due to constructive and destructive wave interference. The resulting wave interference component is known as a *standing wave*. For an arbitrary termination with reflection coefficient $\Gamma_L = |\Gamma_L| e^{j\theta_L}$, the voltage and current

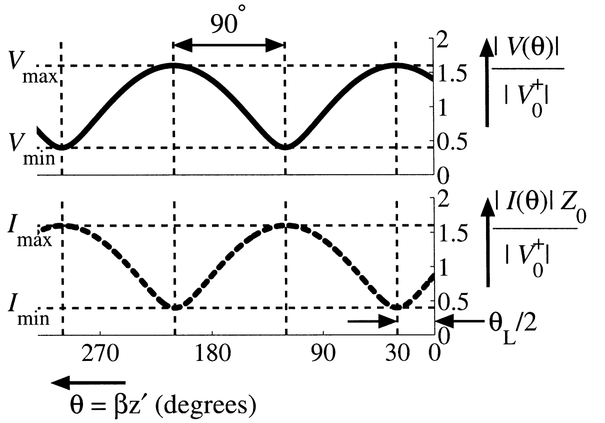


Figure 6.17 Voltage and current standing-wave patterns for a lossless transmission line terminated in a complex load impedance with $\Gamma_L = 0.6\,e^{j60^\circ}$.

standing-wave patterns are given by

$$|V(z')| = |V_0^+|\sqrt{(1+|\Gamma_L|)^2\cos^2\left(\beta z' - \frac{\theta_L}{2}\right) + (1-|\Gamma_L|)^2\sin^2\left(\beta z' - \frac{\theta_L}{2}\right)} \tag{6.74}$$

$$|I(z')| = \frac{|V_0^+|}{Z_0}\sqrt{(1-|\Gamma_L|)^2\cos^2\left(\beta z' - \frac{\theta_L}{2}\right) + (1+|\Gamma_L|)^2\sin^2\left(\beta z' - \frac{\theta_L}{2}\right)} \tag{6.75}$$

Figure 6.17 illustrates the relative voltage and current variations along a lossless transmission line for a general complex load impedance with $\Gamma_L = 0.6e^{j60^\circ}$. In general, the standing-wave pattern on a lossless transmission line is periodic with a period of $\lambda/2$ (or 180° in θ). The voltage magnitude alternates between the maximum and minimum values $V_{\max}$ and $V_{\min}$ given by

$$V_{\max} = (1+|\Gamma_L|)|V_0^+| \tag{6.76}$$

$$V_{\min} = (1-|\Gamma_L|)|V_0^+| \tag{6.77}$$

Similarly, the maximum and minimum current values $I_{\max}$ and $I_{\min}$ are

$$I_{\max} = (1+|\Gamma_L|)\frac{|V_0^+|}{Z_0} = \frac{V_{\max}}{Z_0} \tag{6.78}$$

$$I_{\min} = (1-|\Gamma_L|)\frac{|V_0^+|}{Z_0} = \frac{V_{\min}}{Z_0} \tag{6.79}$$

The locations with maximum voltage can be found from the condition $\beta z' - \theta_L/2 = n\pi$ ($n = 0, 1, 2, \ldots$). The minimum voltages are located a quarter wavelength from the maximum voltages. Locations with current maximum correspond to locations with minimum voltage and vice versa.

The ratio of $V_{\max}$ to $V_{\min}$ is defined as the *standing-wave ratio* (SWR, or S for short) and is given in terms of the reflection coefficient at the termination by F6.

$$\mathrm{SWR} = \frac{V_{\max}}{V_{\min}} = \frac{I_{\max}}{I_{\min}} = \frac{1 + |\Gamma_L|}{1 - |\Gamma_L|} \tag{6.80}$$

The standing-wave ratio is a measure for the amount of mismatch at the termination. The standing-wave ratio for a matched termination is $\mathrm{SWR} = 1$. For an open-circuit, a short-circuit, or a purely reactive termination $\mathrm{SWR} \to \infty$. For a resistive or complex termination, $1 < \mathrm{SWR} < \infty$. In general, SWR varies in the range

$$1 \leq \mathrm{SWR} \leq \infty \tag{6.81}$$

Table 6.3 shows the standing-wave patterns for several special types of terminations. For an open-circuit termination and a purely resistive termination with $R_L > Z_0$, the voltage is maximum at the termination. In contrast, the voltage at the termination is minimum for a short-circuit termination or a purely resistive termination with $R_L < Z_0$. A resistive termination causes a compression in the standing-wave pattern, whereas a reactive termination gives rise to a shift of the voltage maximum away from the termination. For a complex termination as shown in Fig. 6.17 with $\Gamma_L = 0.6e^{j60^\circ}$, the standing-wave pattern is both compressed ($\mathrm{SWR} = 4$) and shifted toward the source side by $\theta_L/2 = +30^\circ$ compared to the open-circuit case.

The standing-wave ratio and the distance from the termination to the nearest voltage maximum can be determined in an experimental setup to find the complex reflection coefficient and, hence, the complex impedance of an unknown termination.* The reflection coefficient magnitude $|\Gamma_L|$ is given in terms of SWR as

$$|\Gamma_L| = \frac{\mathrm{SWR} - 1}{\mathrm{SWR} + 1} \tag{6.82}$$

Example. From standing-wave measurements, the standing-wave ratio is found as $\mathrm{SWR} = V_{\max}/V_{\min} = 5$, the distance between successive voltage minima is 20 cm, and the distance from the termination to the nearest voltage minimum is 4 cm. From Eq. (6.82) follows the magnitude of the reflection coefficient at the termination as $|\Gamma_L| = (5-1)/(5+1) = 2/3$. The wavelength on the line corresponds to twice the distance between successive voltage minima and is $\lambda = 40\,\mathrm{cm}$. The distance from the termination to the closest voltage minimum is $4/40\,\lambda = \lambda/10$ or 36°, and the distance to the nearest voltage maximum is $\lambda/10 + \lambda/4 = 0.35\,\lambda$ or 126°. The phase of the reflection coefficient is $\theta_L = 2 \times 126^\circ = 252^\circ$. The corresponding load impedance is found with $Z_L = Z_0(1 + \Gamma_L)/(1 - \Gamma_L)$ as $Z_L = (0.299 - j0.683)Z_0$.

In most applications, the phase information for the reflection coefficient is not needed. The magnitude of the reflection coefficient directly determines the fraction of

*In practice, it is easier to accurately determine the location of a voltage minimum. The location to the voltage maximum can be obtained from the location of the voltage minimum by adding or subtracting a quarter wavelength.

Table 6.3 Standing-wave Patterns on a Lossless Transmission Line for Special Types of Terminations

<table>
<tr><th>Type of termination</th><th>Standing-wave pattern</th></tr>
<tr><td>

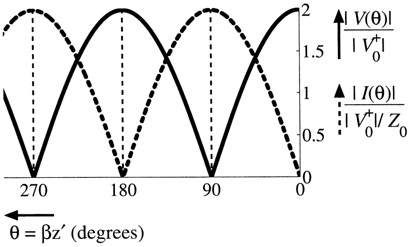
Open circuit

$$|V(z')| = 2|V_0^+||\cos\beta z'|$$

$$|I(z')| = \frac{|V_0^+|}{Z_0}|\sin\beta z'|$$

$$\Gamma_L = +1 \quad \text{SWR} = \infty$$

</td><td>



</td></tr>
<tr><td>

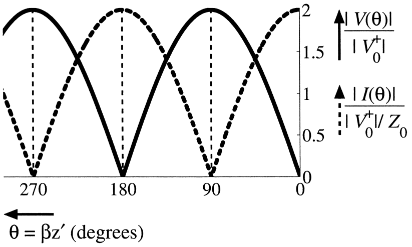
Short circuit

$$|V(z')| = 2|V_0^+||\sin\beta z'|$$

$$|I(z')| = 2\frac{V_0^+}{Z_0}|\cos\beta z'|$$

$$\Gamma_L = -1 \quad \text{SWR} = \infty$$

</td><td>



</td></tr>
<tr><td>

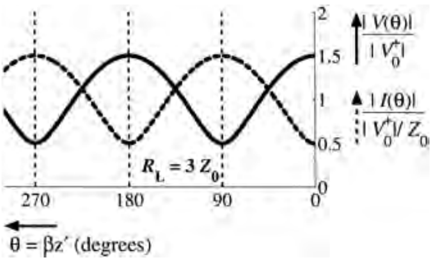
Resistive termination $R_L > Z_0$

$$|V(z')| = \frac{2}{R_L + Z_0}|V_0^+|\sqrt{R_L^2\cos^2\beta z' + Z_0^2\sin^2\beta z'}$$

$$|I(z')| = \frac{2}{R_L + Z_0}\frac{|V_0^+|}{Z_0}\sqrt{Z_0^2\cos^2\beta z' + R_L^2\sin^2\beta z'}$$

$$\Gamma_L = \frac{R_L - Z_0}{R_L + Z_0} > 0 \quad \text{SWR} = \frac{R_L}{Z_0}$$

</td><td>



</td></tr>
<tr><td>

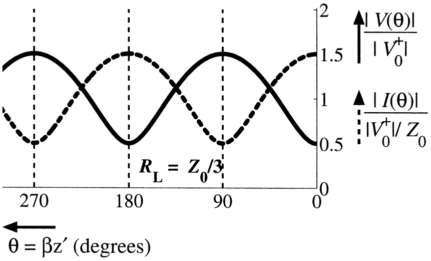
Resistive termination $R_L < Z_0$

$$|V(z')| = \frac{2}{R_L + Z_0}|V_0^+|\sqrt{Z_0^2\cos^2\beta z' + R_L^2\sin^2\beta z'}$$

$$|I(z')| = \frac{2}{R_L + Z_0}\frac{|V_0^+|}{Z_0}\sqrt{R_L^2\cos^2\beta z' + Z_0^2\sin^2\beta z'}$$

$$\Gamma_L = \frac{R_L - Z_0}{R_L + Z_0} < 0 \quad \text{SWR} = \frac{Z_0}{R_L}$$

</td><td>



</td></tr>
<tr><td>

Reactive termination $Z_L = jX_L$

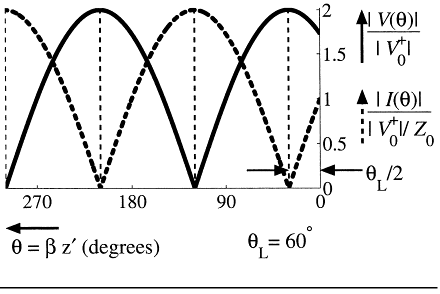
$$|V(z')| = 2|V_0^+|\left|\cos(\beta z' - \theta_L/2)\right|$$

$$|I(z')| = 2\frac{|V_0^+|}{Z_0}\left|\sin(\beta z' - \theta_L/2)\right|$$

$$\Gamma_L = 1e^{j\theta_L}$$

$$\theta_L = 2\tan^{-1}(Z_0/X_L) \quad \text{SWR} \to \infty$$

</td><td>



</td></tr>
</table>

average incident power that is reflected back on the transmission line. With Eqs. (6.72) and (6.73), the net power flow on a lossless transmission line is given by

$$P_{\text{ave}}(z') = \frac{1}{2}\text{Re}\{V(z')I^*(z')\} = \frac{|V_0^+|^2}{2Z_0}\left(1 - |\Gamma_L|^2\right) = P_{\text{ave}}^+\left(1 - |\Gamma_L|^2\right) \tag{6.83}$$

which is independent of position z' on the line. The fraction of average incident power P_{ave}^+ that is reflected is

$$P_{\text{ave}}^- = -|\Gamma_L|^2 P_{\text{ave}}^+ \tag{6.84}$$

The negative sign in Eq. (6.84) indicates the power flow away from the load. Note that the incident power P_{ave}^+ is the combined power due to all forward traveling wave components and thus depends on the load impedance if the source is not matched ($Z_S \neq Z_0$).

In many transmission systems, such as a radio transmitter site, it is critical to monitor the amount of reflected power. The percentage of reflected power can be directly determined from the measured standing-wave ratio. For example, for SWR = 1.5, the magnitude of the reflection coefficient is 0.2, which means that 4% of the incident power is reflected. For a 60-KW transmitter station this would amount to a reflected power of 2400 W.

6.4.3. The Smith Chart

The *Smith chart*, developed by P. H. Smith in 1939, is a powerful graphical tool for solving and visualizing transmission-line problems [17,18]. Originally intended as a graphical transmission-line calculator before the computer age to perform calculations involving complex impedances, the Smith chart has become one of the primary graphical display formats in microwave computer-aided design software and in some commonly used laboratory test equipment, in particular the network analyzer.

The transformation of complex impedance along a transmission line given in Eq. (6.64) is mathematically complicated and lacks visualization and intuition. On the other hand, the reflection coefficient undergoes a simple and intuitive transformation along the transmission line. The reflection coefficient at distance z' from the termination is defined as $\Gamma(z') = V^-(z')/V^+(z')$ and is given in terms of the reflection coefficient at the termination Γ_L by

$$\Gamma(z') = \Gamma_L e^{-j2\beta z'} = |\Gamma_L| e^{j(\theta_L - 2\beta z')} \tag{6.85}$$

The magnitude of the reflection coefficient is unchanged along the lossless line and the phase of the reflection coefficient is reduced by twice the electrical distance from the termination.

The Smith chart combines the simple transformation property of the reflection coefficient along the line with a graphical representation of the mapping of normalized

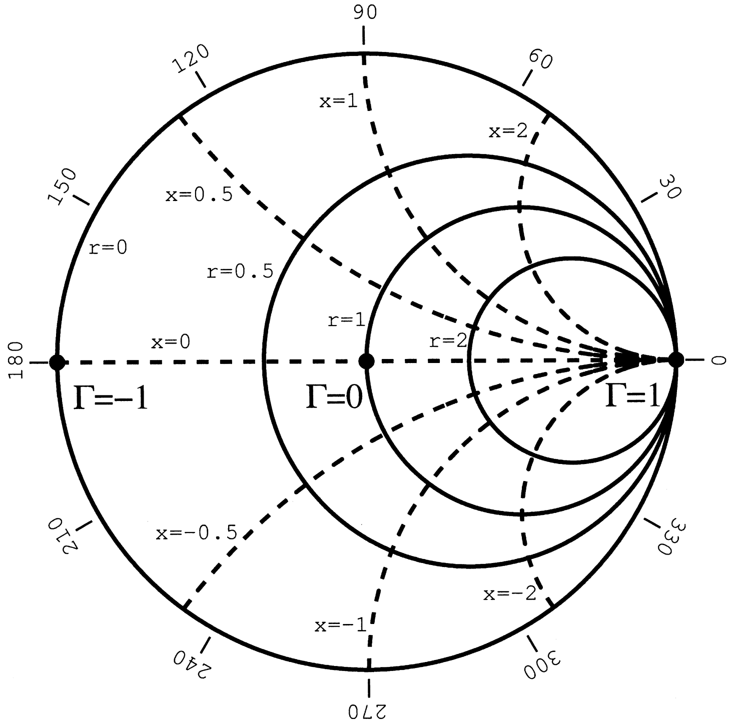


Figure 6.18 Illustration of the basic features of the Smith chart.

impedance to the complex reflection coefficient plane given by

$$\mathsf{z}(z') = \frac{Z_{\text{in}}(z')}{Z_0} = \frac{1 + \Gamma(z')}{1 - \Gamma(z')} \tag{6.86}$$

Here, $\mathsf{z} = \mathsf{r} + j\mathsf{x} = Z/Z_0$ is defined as the normalized impedance with respect to the characteristic impedance of the line. The combination of these two operations in the Smith chart enables the simple graphical determination and visualization of the impedance transformation along a transmission line. Other parameters, such as the standing-wave ratio or the locations of voltage maxima and minima on the line can be simply read off the Smith chart, and more advanced transmission-line calculations and circuit designs can be performed with the Smith chart.

Figure 6.18 illustrates the basic features of the Smith chart. The chart shows a grid of normalized impedance coordinates plotted in the complex plane of the reflection coefficient. The impedance grid consists of a set of circles for constant values of normalized resistance r and a set of circular arcs for constant values of normalized reactance x. Any normalized impedance $\mathsf{z} = \mathsf{r} + j\mathsf{x}$ on a transmission line corresponds to a particular point on or within the unit circle ($|\Gamma| = 1$ circle) in the complex plane of the reflection coefficient. For a matched impedance the $\mathsf{r} = 1$ circle and $\mathsf{x} = 0$ line intersect at the origin of the Smith chart ($\Gamma = 0$). The open-circuit point $\Gamma = 1$ is to the far right, and the short-circuit point $\Gamma = -1$ is to the far left, as indicated in Fig. 6.18. In a real Smith chart, as shown

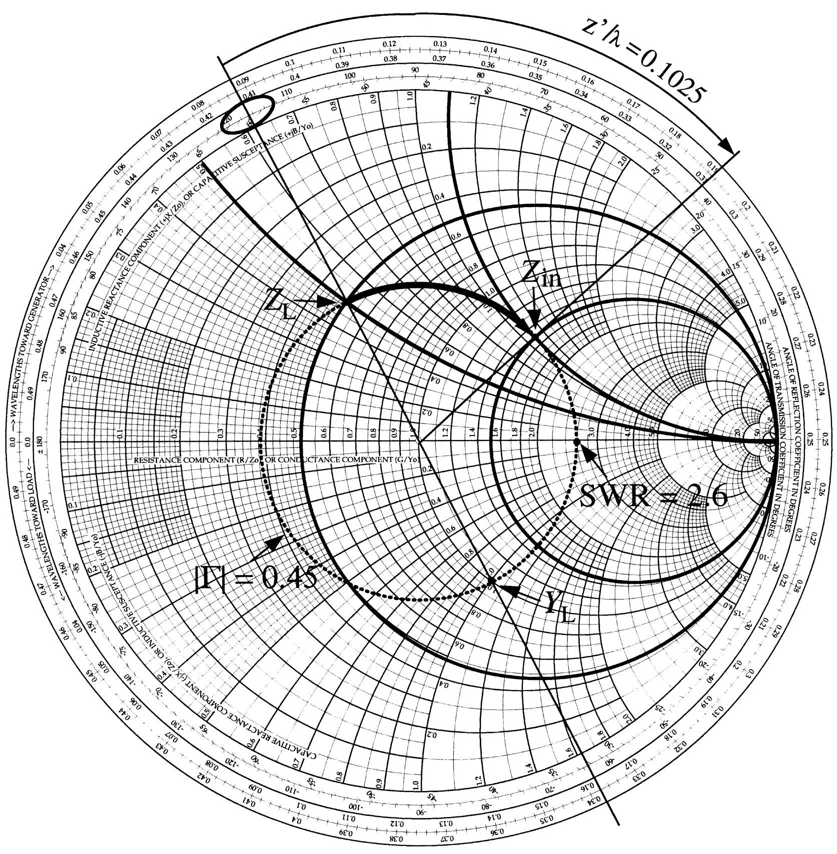

z'/λ =0.1025
Z_in
Z_L
SWR = 2.6
|Γ|=0.45
Y_L
WAVELENGTHS TOWARD GENERATOR
WAVELENGTHS TOWARD LOAD
ANGLE OF REFLECTION COEFFICIENT IN DEGREES
ANGLE OF TRANSMISSION COEFFICIENT IN DEGREES
RESISTANCE COMPONENT (R/Zo) OR CONDUCTANCE COMPONENT (G/Yo)
INDUCTIVE REACTANCE COMPONENT (+jX/Zo) OR CAPACITIVE SUSCEPTANCE (+jB/Yo)
CAPACITIVE REACTANCE COMPONENT (-jX/Zo) OR INDUCTIVE SUSCEPTANCE (-jB/Yo)

Figure 6.19 Smith chart example for $Z_L = (25 + j25)\,\Omega$ and $Z_0 = 50\,\Omega$.

in Fig. 6.19, a fine grid is used for added accuracy, and scales are added to help with the calculation of phase change in reflection coefficient along the transmission line.

Example. To illustrate the use of the Smith chart for transmission-line calculations, consider a lossless line with characteristic impedance $Z_0 = 50\,\Omega$, which is terminated in a complex load impedance $Z_L = (25 + j25)\,\Omega$. The normalized load impedance is $\mathsf{z} = 0.5 + j0.5$ and is shown on the Smith chart as the intersection of the $\mathsf{r} = 0.5$ and $\mathsf{x} = 0.5$ grid circles. The load reflection coefficient can be directly read off from the Smith chart. The radius of the transformation circle through Z_L (relative to the radius of the unit circle $\mathsf{r} = 0$) gives the magnitude of the reflection coefficient as $|\Gamma_L| = 0.45$. The phase of the reflection coefficient is $\theta_L = 116.5°$. The standing-wave ratio on the line corresponds to the normalized maximum impedance $\mathsf{z}_{\max}$ along the line, which is real and lies on the intersection of the transformation circle and the $\mathsf{x} = 0$ line. The standing-wave ratio can be directly read off the Smith chart as SWR = 2.6. For a given electrical length of the line, the input impedance is found by first determining the reflection coefficient at the input through clockwise rotation on the transformation (SWR) circle by twice the electrical length of the line, as given by Eq. (6.85). Assuming an electrical length of $l = 0.1025\,\lambda$, the phase of the reflection coefficient changes by $-2\beta l = -4\pi \times 0.1025$. This amounts to a rotation in *clockwise* direction by about 74°. For convenience, the Smith chart includes scales around its periphery, which can be used to determine the amount of phase rotation directly in units of wavelengths. In this example, the starting value at the load on the rotation scale labeled "toward generator" is 0.088. The end value is $0.088 + 0.1025 = 0.1905$. The phase of the reflection coefficient is read off as $\theta_{\text{in}} = 42.5°$. Finally, the input impedance is obtained as the intersection of the line through the origin with constant phase and the transformation circle. The normalized input impedance is approximately found as $\mathsf{z}_{\text{in}} = 1.5 + j1.1$, or $Z_{\text{in}}(z' = l = 0.1025\lambda) = (75 + j55)\,\Omega$.

In transmission-line problems with parallel-connected elements, it is advantageous to work with admittances rather than impedances. The impedance Smith chart can be conveniently used with normalized admittances $\mathsf{y} = \mathsf{g} + j\mathsf{b} = YZ_0$ by considering the relationship

$$\Gamma = \frac{\mathsf{z} - 1}{\mathsf{z} + 1} = -\frac{\mathsf{y} - 1}{\mathsf{y} + 1} \tag{6.87}$$

where $\mathsf{y} = 1/\mathsf{z}$. This relationship shows that the impedance grid can be directly used as admittance grid with $\mathsf{g} = \text{const}$ circles and $\mathsf{b} = \text{const}$ circular arcs if the reflection coefficient is multiplied by negative one, which amounts to a rotation by 180° on the Smith chart. Then, the open circuit is located at the far left and the short circuit is at the far right. The conversion from normalized impedance coordinates to normalized admittance coordinates given by $\mathsf{y} = 1/\mathsf{z}$ can be simply achieved on the Smith chart by a 180° rotation along the transformation (SWR) circle. For example, for the normalized load impedance $\mathsf{z} = 0.5 + j0.5$, the normalized load admittance is found as $\mathsf{y} = 1/\mathsf{z} = 1 - j$, as indicated in Fig. 6.19.

6.4.4. Impedance Matching

In many transmission-line applications, it is desirable to match the load impedance to the characteristic impedance of the line and eliminate reflections in order to maximize the power delivered to the load and minimize signal distortion and noise.* Reducing or eliminating reflections from the load is particularly important in high-power RF transmission systems to also minimize hot spots along a transmission line (e.g., the feed line between the transmitter and the antenna) that are caused by standing waves and not exceed the power-handling capabilities of the transmission line. Excessive reflections can also damage the generator, especially in high-power applications.

In practice, the impedance of a given load is often different from the characteristic impedance Z_0 of the transmission line, and an additional impedance transformation network is needed to achieve a matched load condition. Figure 6.20 illustrates the basic idea of matching an arbitrary load impedance Z_L to a transmission line. The matching network is designed to provide an input impedance looking into the network that is equal to Z_0 of the transmission line, and thus eliminate reflections at the junction between the transmission line and the matching network. The matching network is ideally lossless so

*In general, impedance matching can be done at the load or the source end, or at both ends of the transmission line. For a matched source, maximum power is delivered to the load when it is matched to the transmission line and power loss on the line is minimized. For a given source impedance Z_S, maximum power transmission on a lossless line is achieved with conjugate matching at the source ($Z_{\text{in}} = Z_S^*$) [1].

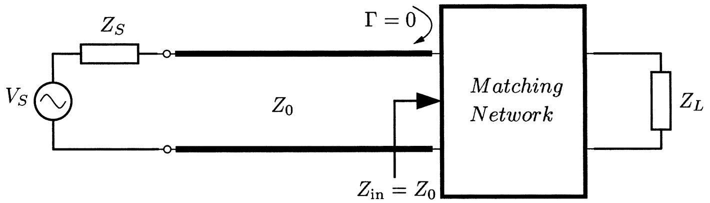


Figure 6.20 General illustration of impedance matching at the termination.

that all incident power on the line ends up being dissipated in the load. A lossless matching network may consist of lumped reactive elements or reactive transmission-line elements (stubs) at higher frequencies and/or cascaded transmission-line sections of appropriate length.

A matching network requires at least two adjustable parameters, such as a lumped series element and a lumped shunt element, each with adjustable reactance value, to independently transform the real and imaginary parts of the load impedance. Because the elements in the matching network are frequency dependent, the exact matching condition is generally achieved only at a single design frequency. For other frequencies, the reflection coefficient will be sufficiently small only over a narrow bandwidth about the design frequency. Larger matching bandwidths may be achieved if more independent elements are used in the matching network.

Many different design choices of matching networks are available. The selection of a particular matching network may depend on a number of factors including realizability in a given technology, required bandwidth, simplicity, occupied space, tunability of the matching network, and cost of implementation. In the following, two common matching methods using sections of lossless transmission lines are described to further illustrate the concept of impedance matching.

Quarter-wave Transformer

A lossless transmission line of length $l = \lambda/4$ has a special simplified impedance transformation property, which can be advantageously used for impedance matching. With Eq. (6.64), the input impedance of a lossless transmission line of length $l = \lambda/4$ and characteristic impedance $Z_{0,\mathrm{T}}$ that is terminated with load impedance Z_L is

$$Z_{\text{in}}|_{l=\lambda/4} = \frac{Z_{0,\mathrm{T}}^2}{Z_L} \tag{6.88}$$

In particular, any purely resistive load impedance $Z_L = R_L$ is transformed into a resistive input impedance given by $R_{\text{in}} = Z_{0,\mathrm{T}}^2/R_L$. Hence, a quarter-wave section of a transmission line can be directly used to match a purely resistive load impedance R_L to a line with characteristic impedance Z_0 if the characteristic impedance $Z_{0,\mathrm{T}}$ of the quarter-wave section is given by

$$Z_{0,\mathrm{T}} = \sqrt{R_L Z_0} \tag{6.89}$$

For example, to match a half-wave dipole antenna with input impedance $Z_L \approx 73\,\Omega$ to a twin-lead cable with $Z_0 = 300\,\Omega$, the characteristic impedance of the quarter-wave transformer should be $Z_{0,\mathrm{T}} = \sqrt{73\,\Omega \cdot 300\,\Omega} \approx 148\,\Omega$.

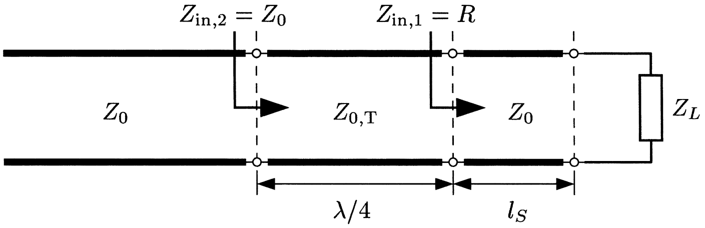


Figure 6.21 Impedance matching of a complex load using a quarter-wave transformer.

If the load impedance is complex, it is necessary to first transform the complex impedance to a real impedance. This can be accomplished with a section of transmission line of appropriate length l_s between the load and the quarter-wave transform, as illustrated in Fig. 6.21. A transmission line can transform any complex load impedance with $|\Gamma_L| < 1$ to a resistive impedance at the locations with either voltage maximum or voltage minimum. The transmission-line transformation of a complex load to a real impedance is best illustrated on the Smith chart. For example, consider a complex load consisting of a parallel combination of $R_L = 125\,\Omega$ and $C_L = 2.54\,\text{pF}$. At the design frequency $f_0 = 1\,\text{GHz}$, the load impedance is $Z_L = (25 - j50)\,\Omega$. The normalized load impedance $\mathsf{z}_L = 0.5 - j$ for $Z_0 = 50\,\Omega$ is shown on the Smith chart in Fig. 6.22. The transformation circle intersects the $\mathsf{x} = 0$ grid line at $r_{\min} \approx 0.24$ and $r_{\max} = 1/r_{\min} \approx 4.2$. The distance to the closest location with real input impedance ($\mathsf{z}_{\text{in}} = r_{\min}$) is found as $l_s = 0.135\,\lambda$. The input impedance at this location is $Z_{\text{in},1} = R = r_{\min}Z_0 \approx 12\,\Omega$, and the characteristic impedance of the quarter-wave transformer is found as $Z_{0,\text{T}} = \sqrt{RZ_0} \approx 24.5\,\Omega$. The second solution with real input impedance is at the voltage maximum with $R = r_{\max}Z_0 \approx 210\,\Omega$ and $l_s = 0.135\,\lambda + 0.25\,\lambda = 0.385\,\lambda$. The corresponding characteristic impedance of the quarter-wave transformer is $Z_{0,\text{T}} = \sqrt{RZ_0} \approx 102.5\,\Omega$. Typically, the solution with the shortest line length l_s is chosen unless it is difficult to realize the characteristic impedance of the corresponding quarter-wave transformer.

Figure 6.23 shows the response of the matching network as a function of frequency. The matching network gives an exact match (SWR = 1) at the design frequency $f_0 = 1\,\text{GHz}$. The bandwidth defined here as the frequency band around the center frequency with SWR ≤ 1.5 is about 100 MHz or 10%. The standing-wave ratio response without matching network is also shown in Fig. 6.23 for comparison.

The bandwidth of the matching network can be increased, for example, by cascading multiple quarter-wave sections (multisection quarter-wave transformer) with smaller impedance steps per section giving an overall more gradual impedance transformation [1]. This type of matching network can be easily implemented in planar transmission line technology, such as microstrip, where the characteristic impedance can be changed continuously by varying the line width or spacing.

Stub Matching

In another common impedance matching technique, a reactive element of appropriate value is connected either in series or in parallel to the transmission line at a specific distance from the load. The reactive element can be realized as open- or short-circuited

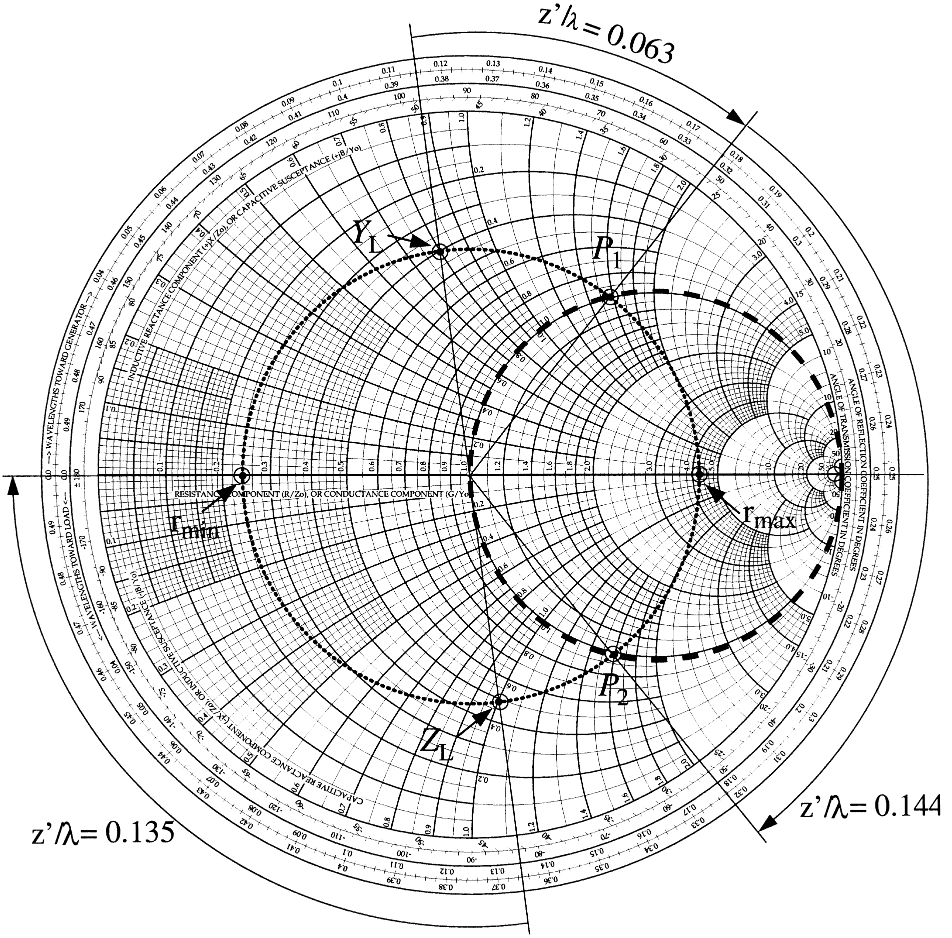

z'/λ = 0.063
z'/λ= 0.135
z'/λ= 0.144
Y_L
P_1
P_2
Z_L
r_{min}
r_{max}
RESISTANCE COMPONENT (R/Zo), OR CONDUCTANCE COMPONENT (G/Yo)
INDUCTIVE REACTANCE COMPONENT (+jX/Zo), OR CAPACITIVE SUSCEPTANCE (+jB/Yo)
CAPACITIVE REACTANCE COMPONENT (-jX/Zo), OR INDUCTIVE SUSCEPTANCE (-jB/Yo)
WAVELENGTHS TOWARD GENERATOR
WAVELENGTHS TOWARD LOAD
ANGLE OF REFLECTION COEFFICIENT IN DEGREES
ANGLE OF TRANSMISSION COEFFICIENT IN DEGREES

Figure 6.22 Graphical illustration on the Smith chart of quarter-wave matching and shunt (stub) matching of a complex load impedance $Z_L/Z_0 = 0.5 - j$.

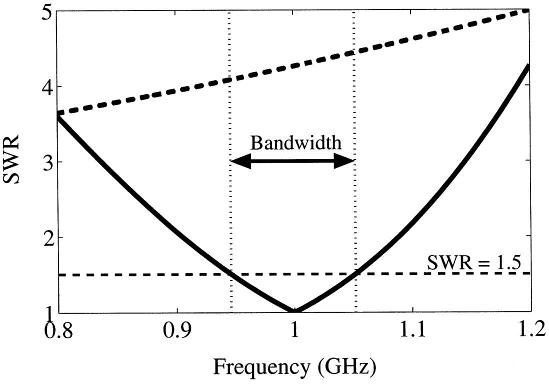


Figure 6.23 SWR = 1.5 bandwidth of an example matching network using a quarter-wave transformer. Also shown with a dashed line is the response without the matching network.

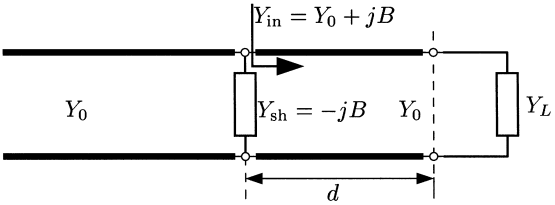


Figure 6.24 Matching network with a parallel shunt element.

stub or as lumped inductor or capacitor element. The two design parameters of a stub matching network are the distance from the termination at which the reactive element is connected, and the stub length needed to realize the required reactance.

The general matching procedure with a single reactive element or a stub is demonstrated for a parallel (shunt) configuration with shunt admittance element Y_{sh} connected at distance d from the termination, as illustrated in Fig. 6.24. For shunt connections it is more convenient to work with admittances than with impedances. The transmission line transforms the load admittance $Y_L = 1/Z_L$ to an input admittance $Y_{\text{in}} = G + jB$ at distance d from the termination. In the first step of the matching procedure, distance d is selected such that the real part of the input admittance is matched as $G = Y_0$, and the nonzero input susceptance B is determined. In the second step, a reactive shunt element with admittance $Y_{\text{sh}} = -jB$ is added to cancel out susceptance B in the input admittance. The summation of shunt admittance and input admittance of the line yields a matched total admittance $Y_0 = 1/Z_0$.

The shunt matching procedure is further illustrated on the Smith chart shown in Fig. 6.22. The same normalized load impedance $\mathsf{z}_L = 0.5 - j$ as in the previous matching network example is assumed. The corresponding normalized load admittance is found from the Smith chart as $\mathsf{y}_L = 0.4 + j0.8$. The transformation circle with $|\Gamma| = \text{const}$ intersects the $\mathsf{g} = 1$ circle at two points labeled as P_1 and P_2 satisfying the condition $\mathsf{y}_{\text{in}} = 1 + jb$. Any complex load admittance with $|\Gamma_L| < 1$ can be transformed by a transmission line of appropriate length to a point on the $\mathsf{g} = 1$ circle. The normalized input admittances at points P_1 and P_2 are $\mathsf{y}_{\text{in},1} = 1 + j1.58$ and $\mathsf{y}_{\text{in},2} = 1 - j1.58$, respectively. The distance from the termination to point P_1 on the line with matched real part of the input admittance is found as $d_1 = 0.063\,\lambda$. The distance to P_2 is $d_2 = d_1 + 0.144\,\lambda = 0.207\,\lambda$. The normalized input susceptance $b_1 = 1.58$ at position P_1 is capacitive and needs to be canceled with an inductive shunt element with normalized admittance $\mathsf{y}_{\text{sh}} = -j1.58$. The required shunt element may be realized with a lumped inductor or an open- or short-circuited stub of appropriate length. Similarly, matching position P_2 with $\mathsf{y}_{\text{in},2} = 1 - j1.58$ requires a capacitive shunt element to cancel the susceptance. The capacitive shunt admittance may be realized with a lumped capacitor or an open- or short-circuited stub of appropriate length.

6.5. FURTHER TOPICS OF TECHNOLOGICAL IMPORTANCE AND FUTURE DIRECTIONS

In this section, further transmission-line topics of technological importance are briefly discussed and current developments and future directions are outlined.

6.5.1. Coupled Lines

Transmission-line circuits often consist of multiple parallel conductors that are in close proximity to each other. Examples of multiconductor transmission-line systems include multiphase power lines, telephone cables, and data bus lines on the printed-circuit board (PCB) of a digital system. Due to the proximity of the conductors, the time-varying electromagnetic fields generated by the different transmission lines interact, and the lines become capacitively and inductively coupled. The propagation characteristics of coupled lines depend not only on the line parameters of the individual lines but also on the mutual distributed capacitance and inductance parameters.

The capacitive and inductive coupling between transmission lines often leads to adverse effects in a transmission system. As an example, coupling between closely spaced lines (interconnects) in digital systems can lead to unwanted crosstalk noise and generally sets an upper limit in interconnection density (see e.g. Refs. 10 and 19). On the other hand, electromagnetic coupling between adjacent lines can be used to advantage to realize a variety of components for microwave circuits such as filters, directional couplers, and power dividers [1]. Recently, there has also been increased interest in the realization of compact three-dimensional embedded passive components for RF and mixed-signal modules, and new compact designs using coupled lines have been demonstrated (e.g., Ref. 20). A general overview of coupled transmission-line theory and its application to cross-talk analysis and design of passive microwave components is given, e.g., in Ref. 5.

6.5.2. Differential Lines

A differential line can be considered as a special case of two symmetric coupled lines. A differential line consists of two closely spaced symmetric signal conductors that are driven with identical signals of opposite polarity with respect to a common ground reference (differential signaling). The main advantages of differential lines include an increased immunity to common-mode noise and the localized ground references at the input and output of the line. In particular, the net return current in the ground conductor of a differential transmission line is ideally zero, which helps to eliminate or reduce the effects of nonideal current return paths with finite resistance and inductance. As a disadvantage, differential lines require more conductor traces and generally need to be carefully routed to avoid conversion between differential- and common-mode signals. Because of the advantageous properties of differential lines compared to regular (single-ended) lines, however, differential lines are increasingly being used for critical signal paths in high-speed analog and digital circuits (see, e.g., Refs. 10 and 19). Differential circuit architectures are also being employed in parts of RF circuits because of their superior noise-rejection properties [21].

6.5.3. Chip- and Package-level Interconnects

Transmission lines or electrical interconnects are present at various levels of an electronic system ranging from cabling to printed-circuit board level to chip packaging to chip level. The electrical interconnections in an electronic package constitute the electrical interface between the chip (or a set of chips packaged in a module) and the rest of the electronic system. The package interconnections can generally be represented by a combination of lumped R, L, C elements and nonuniform coupled transmission lines. In some advanced high-performance packages the interconnections are realized in form of a miniature printed wiring board with several levels of metalization. The electronic package may significantly influence the electrical performance of an integrated circuit; hence, the package characteristics should be included in the design phase of the integrated circuit. The co-design of the integrated circuit and package has recently been pursued for both digital and RF integrated circuits as well as for system-on-a-chip solutions.

At the chip level, interconnects in VLSI and RF integrated circuits usually behave as lumped or distributed RC circuits because of the large series resistance of the metalization. With increasing clock frequencies, however, the distributed series inductance becomes more and more significant. As a result, inductance effects cannot be neglected in some of the longer on-chip interconnects in present-day high-performance VLSI circuits [22]. On-chip interconnects with nonnegligible inductance exhibit transmission-line behavior and need to be modeled as lossy transmission lines rather than RC lines.

6.5.4. CAD Modeling of Transmission Lines

The development of dispersive single and coupled transmission-line models for computer-aided design (CAD) tools is an active area of research in both industry and academia. In general, the line parameters of a transmission line are frequency dependent because of conductor loss (including skin and proximity effects), substrate loss, and dispersion induced by inhomogeneous dielectric substrates. The frequency-dependent transmission-line parameters, however, cannot be represented directly in a time-domain simulator environment such as SPICE. Several approaches for modeling lossy dispersive transmission lines have been developed including (1) convolution with the impulse response of the lossy transmission line, (2) synthesis of the frequency-dependent line parameter in terms of ideal lumped elements and controlled sources for a short line section, and (3) mathematical macromodels obtained with model-order reduction (MOR) techniques resulting in an approximation of the transmission-line characteristics with a finite number of pole-residue pairs. Other areas of current and future interest include the efficient extraction of the line parameters (or parasitics) and the cosimulation of the electromagnetic, thermal, and mechanical phenomena in an electronic system. A review of the methodologies for the electrical modeling of interconnects and electronic packages is given in Ref. 23. Modeling of coupled transmission lines–interconnects based on model-order reduction is further described in Ref. 24.

REFERENCES

1. Pozar, D.M. *Microwave Engineering*, 2nd Ed.; Wiley: New York, 1998.
2. Collin, R.E. *Field Theory of Guided Waves*, 2nd Ed.; IEEE Press: New York, 1991.
3. Cheng, D.K. *Field and Wave Electromagnetics*, 2nd Ed.; Addison-Wesley: Reading, MA, 1990.
4. Ramo, S.; Whinnery, J.R; Van Duzer, T. *Fields and Waves in Communication Electronics*, 3rd Ed.; Wiley: New York, 1993.
5. Magnusson, P.C.; Alexander, G.C.; Tripathi, V.K.; Weisshaar, A. *Transmission Lines and Wave Propagation*, 4th Ed.; CRC Press: Boca Raton, FL, 2001.
6. Hoffmann, R.K. *Handbook of Microwave Integrated Circuits*; Artech House: Norwood, MA, 1987.
7. Wadell, B.C. *Transmission Line Design Handbook*; Artech House: Norwood, MA, 1991.

8. Nagel, L.W. SPICE: A computer program to simulate semiconductor circuits, Tech. Rep. ERL-M520, Univ. California, Berkeley, May 1975.
9. Johnson, H.W.; Graham, M. *High-Speed Digital Design: A Handbook of Black Magic*; Prentice-Hall: Englewood Cliffs, NJ, 1993.
10. Hall, S.H.; Hall, G.W.; McCall, J.A. *High-Speed Digital System Design: A Handbook of Interconnect Theory and Design Practices*; Wiley: New York, 2000.
11. Freeman, J.C. *Fundamentals of Microwave Transmission Lines*; Wiley: New York, 1996.
12. DeFalco, J.A. Reflection and cross talk in logic circuit interconnections. IEEE Spectrum **July 1970**, 44–50.
13. Oliver, B.M. Time-domain reflectometry. Hewlett-Packard **Feb**. **1964**, *15* (6), 1–7.
14. Inan, U.S.; Inan, A.S. *Engineering Electromagnetics*; Addison-Wesley: Reading, MA, 1998.
15. Jong, J.M.; Tripathi, V.K. Equivalent circuit modeling of interconnects from time domain measurements. IEEE Trans. Comp., Pack., Manufact. Technol. **Feb**. **1993**, *16* (1), 119–126.
16. Lathi, B.P. *Linear Systems and Signals*; Oxford University Press: New York, 2002.
17. Smith, P.H. Transmission line calculator. Electronics **Jan**. **1939**, *12* (1), 29–31.
18. Smith, P.H. An improved transmission-line calculator. Electronics **Jan**. **1944**, *17* (1), 130, 318.
19. Dally, W.J.; Poulton, J.W. *Digital Systems Engineering*; Cambridge University Press: New York, 1998.
20. Settaluri, R.K.; Weisshaar, A; Tripathi, V.K. Design of compact multilevel folded-line bandpass filters. IEEE Trans. Microwave Theory Tech. **Oct. 2001**, *49* (10), 1804–1809.
21. Lee, T.H. *The Design of CMOS Radio-Frequency Integrated Circuits*; Cambridge University Press: New York, 1998.
22. Deutsch, A. When are transmission-line effects important for on-chip interconnections? IEEE Trans. Microwave Theory Tech. **Oct**. **1997**, *45* (10), 1836–1846.
23. Ruehli, A.E.; Cangellaris, A.C. Progress in the methodologies for the electrical modeling of interconnects and electronic packages. Proc. IEEE **May 2001**, *89* (5), 740–771.
24. Achar, R.; Nakhla, M. Simulation of high-speed interconnects. Proc. IEEE **May 2001**, *89* (5), 693–728.

7
Waveguides and Resonators

Kenneth R. Demarest
The University of Kansas
Lawrence, Kansas, U.S.A.

7.1. INTRODUCTION

Any structure that transports electromagnetic waves can be considered as a *waveguide*. Most often, however, this term refers to either metal or dielectric structures that transport electromagnetic energy without the presence of a complete circuit path. Waveguides that consist of conductors and dielectrics (including air or vacuum) are called *metal waveguides*. Waveguides that consist of only dielectric materials are called *dielectric waveguides*.

Metal waveguides use the reflective properties of conductors to contain and direct electromagnetic waves. In most cases, they consist of a long metal cylinder filled with a homogeneous dielectric. More complicated waveguides can also contain multiple dielectrics and conductors. The conducting cylinders usually have rectangular or circular cross sections, but other shapes can also be used for specialized applications. Metal waveguides provide relatively low loss transport over a wide range of frequencies—from RF through millimeter wave frequencies.

Dielectric waveguides guide electromagnetic waves by using the reflections that occur at interfaces between dissimilar dielectric materials. They can be constructed for use at microwave frequencies, but are most commonly used at optical frequencies, where they can offer extremely low loss propagation. The most common dielectric waveguides are optical fibers, which are discussed elsewhere in this handbook (Chapter 14: Optical Communications).

Resonators are either metal or dielectric enclosures that exhibit sharp resonances at frequencies that can be controlled by choosing the size and material construction of the resonator. They are electromagnetic analogs of lumped resonant circuits and are typically used at microwave frequencies and above. Resonators can be constructed using a large variety of shaped enclosures, but simple shapes are usually chosen so that their resonant frequencies can be easily predicted and controlled. Typical shapes are rectangular and circular cylinders.

7.2. MODE CLASSIFICATIONS

Figure 7.1 shows a uniform waveguide, whose cross-sectional dimensions and material properties are constant along the waveguide (i.e., z) axis. Every type of waveguide has an

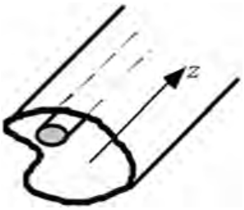


Figure 7.1 A uniform waveguide with arbitrary cross section.

infinite number of distinct electromagnetic field configurations that can exist inside it. Each of these configurations is called a *mode*. The characteristics of these modes depend upon the cross-sectional dimensions of the conducting cylinder, the type of dielectric material inside the waveguide, and the frequency of operation.

When waveguide properties are uniform along the z axis, the phasors representing the forward-propagating (i.e., $+z$) time-harmonic modes vary with the longitudinal coordinate z as **E**, H $\propto e^{-\gamma z}$, where the $e^{j\omega t}$ phasor convention is assumed. The parameter γ is called the *propagation constant* of the mode and is, in general, complex valued:

$$\gamma = \alpha + j\beta \tag{7.1}$$

where $j = \sqrt{-1}$, α is the modal attenuation constant, which controls the rate of decay of the wave amplitude, β is the phase constant, which controls the rate at which the phase of the wave changes, which in turn controls a number of other modal characteristics, including wavelength and velocity.

Waveguide modes are typically classed according to the nature of the electric and magnetic field components that are directed along the waveguide axis, E_z and H_z, which are called the *longitudinal components*. From Maxwell's equations, it follows that the *transverse components* (i.e., directed perpendicular to the direction of propagation) are related to the longitudinal components by the relations [1]

$$E_x = -\frac{1}{h^2}\left(\gamma\frac{\partial E_z}{\partial x} + j\omega\mu\frac{\partial H_z}{\partial y}\right) \tag{7.2}$$

$$E_y = -\frac{1}{h^2}\left(\gamma\frac{\partial E_z}{\partial y} - j\omega\mu\frac{\partial H_z}{\partial x}\right) \tag{7.3}$$

$$H_x = -\frac{1}{h^2}\left(-j\omega\varepsilon\frac{\partial E_z}{\partial y} + \gamma\frac{\partial H_z}{\partial x}\right) \tag{7.4}$$

$$H_y = -\frac{1}{h^2}\left(j\omega\varepsilon\frac{\partial E_z}{\partial x} + \gamma\frac{\partial H_z}{\partial y}\right) \tag{7.5}$$

where,

$$h^2 = k^2 + \gamma^2 \tag{7.6}$$

$k = 2\pi f\sqrt{\mu\varepsilon}$ is the *wave number* of the dielectric, $f = \omega/2\pi$ is the operating frequency in Hz, and μ and ε are the permeability and permittivity of the dielectric, respectively. Similar expressions for the transverse fields can be derived in other coordinate systems, but regardless of the coordinate system, the transverse fields are completely determined by the spatial derivatives of longitudinal field components across the cross section of the waveguide.

Several types of modes are possible in waveguides.

TE modes: Transverse-electric modes, sometimes called *H modes*. These modes have $E_z = 0$ at all points within the waveguide, which means that the electric field vector is always perpendicular (i.e., transverse) to the waveguide axis. These modes are always possible in metal waveguides with homogeneous dielectrics.

TM modes: Transverse-magnetic modes, sometimes called *E modes*. These modes have $H_z = 0$ at all points within the waveguide, which means that the magnetic field vector is perpendicular to the waveguide axis. Like TE modes, they are always possible in metal waveguides with homogeneous dielectrics.

EH modes: These are hybrid modes in which neither E_z nor H_z is zero, but the characteristics of the transverse fields are controlled more by E_z than H_z. These modes usually occur in dielectric waveguides and metal waveguides with inhomogeneous dielectrics.

HE modes: These are hybrid modes in which neither E_z nor H_z is zero, but the characteristics of the transverse fields are controlled more by H_z than E_z. Like EH modes, these modes usually occur in dielectric waveguides and in metal waveguides with inhomogeneous dielectrics.

TEM modes: Transverse-electromagnetic modes, often called *transmission-line modes*. These modes can exist only when more than one conductor with a complete dc circuit path is present in the waveguide, such as the inner and outer conductors of a coaxial cable. These modes are not considered to be waveguide modes.

Both transmission lines and waveguides are capable of guiding electromagnetic signal energy over long distances, but waveguide modes behave quite differently with changes in frequency than do transmission-line modes. The most important difference is that waveguide modes can typically transport energy only at frequencies above distinct cutoff frequencies, whereas transmission line modes can transport energy at frequencies all the way down to dc. For this reason, the term *transmission line* is reserved for structures capable of supporting TEM modes, whereas the term *waveguide* is typically reserved for structures that can only support waveguide modes.

7.3. MODAL FIELDS AND CUTOFF FREQUENCIES

For all uniform waveguides, E_z and H_z satisfy the scalar wave equation at all points within the waveguide [1]:

$$\nabla^2 E_z + k^2 E_z = 0 \tag{7.7}$$

$$\nabla^2 H_z + k^2 H_z = 0 \tag{7.8}$$

where ∇^2 is the Laplacian operator and k is the wave number of the dielectric. However, for $+z$ propagating fields, $\partial()/\partial z = -\gamma()$, so we can write

$$\nabla_{\mathrm{t}}^2 E_z + h^2 E_z = 0 \tag{7.9}$$

and

$$\nabla_t^2 H_z + h^2 H_z = 0 \tag{7.10}$$

where h^2 is given by Eq. (7.5) and ∇_t^2 is the transverse Laplacian operator. In Cartesian coordinates, $\nabla_t^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$. When more than one dielectric is present, E_z and H_z must satisfy Eqs. (7.9) and (7.10) in each region for the appropriate value of k in each region.

Modal solutions are obtained by first finding general solutions to Eqs. (7.9) and (7.10) and then applying boundary conditions that are appropriate for the particular waveguide. In the case of metal waveguides, $E_z = 0$ and $\partial H_z/\partial p = 0$ at the metal walls, where p is the direction perpendicular to the waveguide wall. At dielectric–dielectric interfaces, the E- and H-field components tangent to the interfaces must be continuous. Solutions exist for only certain values of h, called *modal eigenvalues*. For metal waveguides with homogeneous dielectrics, each mode has a single modal eigenvalue, whose value is independent of frequency. Waveguides with multiple dielectrics, on the other hand, have different modal eigenvalues in each dielectric region and are functions of frequency, but the propagation constant γ is the same in each region.

Regardless of the type of waveguide, the propagation constant γ for each mode is determined by its modal eigenvalue, the frequency of operation, and the dielectric properties. From Eqs. (7.1) and (7.6), it follows that

$$\gamma = \alpha + j\beta = \sqrt{h^2 - k^2} \tag{7.11}$$

where h is the modal eigenvalue associated with the dielectric wave number k. When a waveguide has no material or radiation (i.e., leakage) loss, the modal eigenvalues are always real-valued. For this case, γ is either real or imaginary. When $k^2 > h^2$, $\alpha = 0$ and $\beta > 0$, so the modal fields are propagating fields with no attenuation. On the other hand, when $k^2 < h^2$, $\alpha > 0$, and $\beta = 0$, which means that the modal fields are nonpropagating and decay exponentially with distance. Fields of this type are called *evanescent fields*. The frequency at which $k^2 = h^2$ is called the *modal cuttoff frequency* f_c. A mode operated at frequencies above its cutoff frequency is a propagating mode. Conversely, a mode operated below its cutoff frequency is an evanescent mode.

The dominant mode of a waveguide is the one with the lowest cutoff frequency. Although higher order modes are often useful for a variety of specialized uses of waveguides, signal distortion is usually minimized when a waveguide is operated in the frequency range where only the dominant mode is propagating. This range of frequencies is called the *dominant rang*e of the waveguide.

7.4. PROPERTIES OF METAL WAVEGUIDES

Metal waveguides are the most commonly used waveguides at RF and microwave frequencies. Like coaxial transmission lines, they confine fields within a conducting shell, which reduces cross talk with other circuits. In addition, metal waveguides usually exhibit lower losses than coaxial transmission lines of the same size. Although they can be constructed using more than one dielectric, most metal waveguides are simply metal pipes filled with a homogeneous dielectric—usually air. In the remainder of this chapter, the term *metal waveguides* will denote self-enclosed metal waveguides with homogeneous dielectrics.

Metal waveguides have the simplest electrical characteristics of all waveguide types, since their modal eigenvalues are functions only of the cross-sectional shape of the metal cylinder and are independent of frequency. For this case, the amplitude and phase constants of any allowed mode can be written in the form:

$$\alpha = \begin{cases} h\sqrt{1 - \left(\dfrac{f}{f_c}\right)^2} & \text{for } f < f_c \\ 0 & \text{for } f > f_c \end{cases} \tag{7.12}$$

and

$$\beta = \begin{cases} 0 & \text{for } f < f_c \\ h\sqrt{\left(\dfrac{f}{f_c}\right)^2 - 1} & \text{for } f > f_c \end{cases} \tag{7.13}$$

where

$$f_c = \frac{h}{2\pi\sqrt{\mu\varepsilon}} \tag{7.14}$$

Each mode has a unique modal eigenvalue h, so each mode has a specific cutoff frequency. The mode with the smallest modal eigenvalue is the dominant mode. If two or more modes have the same eigenvalue, they are *degenerate modes*.

7.4.1. Guide Wavelength

The distance over which the phase of a propagating mode in a waveguide advances by 2π is called the *guide wavelength* λ_g. For metal waveguides, β is given by Eq. (7.13), so λ_g for any mode can be expressed as

$$\lambda_g = \frac{2\pi}{\beta} = \frac{\lambda}{\sqrt{1-(f_c/f)^2}} \tag{7.15}$$

where $\lambda = (f\sqrt{\mu\varepsilon})^{-1}$ is the wavelength of a plane wave of the same frequency in the waveguide dielectric. For $f \gg f_c$, $\lambda_g \approx \lambda$. Also, $\lambda_g \to \infty$ as $f \to f_c$, which is one reason why it is usually undesirable to operate a waveguide mode near modal cutoff frequencies.

7.4.2. Wave Impedance

Although waveguide modes are not plane waves, the ratio of their transverse electric and magnetic field magnitudes are constant throughout the cross sections of the metal waveguides, just as for plane waves. This ratio is called the modal *wave impedance* and has the following values for TE and TM modes [1]:

$$Z_{\text{TE}} = \frac{E_T}{H_T} = \frac{j\omega\mu}{\gamma} = \frac{\eta}{\sqrt{1-(f_c/f)^2}} \tag{7.16}$$

and

$$Z_{\mathrm{TM}} = \frac{E_T}{H_T} = \frac{\gamma}{j\omega\varepsilon} = \eta\sqrt{1-(f_c/f)^2} \tag{7.17}$$

where E_T and H_T are the magnitudes of the transverse electric and magnetic fields, respectively, and $\eta = \sqrt{\mu/\varepsilon}$ is the intrinsic impedance of the dielectric. In the limit as $f \to \infty$, both Z_{TE} and Z_{TM} approach η. On the other hand, as $f \to f_c$, $Z_{\mathrm{TE}} \to \infty$ and $Z_{\mathrm{TM}} \to 0$, which means that the transverse electric fields are dominant in TE modes near cutoff and the transverse magnetic fields are dominant in TM modes near cutoff.

7.4.3. Wave Velocities

The phase and group velocities of waveguide modes are both related to the rates of change of the modal propagation constant β with respect to frequency. The phase velocity u_p is the velocity of the phase fronts of the mode along the waveguide axis and is given by [1]

$$u_p = \frac{\omega}{\beta} \tag{7.18}$$

Conversely, the group velocity is the velocity at which the amplitude envelopes of narrowband, modulated signals propagate and is given by [1]

$$u_g = \frac{\partial\omega}{\partial\beta} = \left(\frac{\partial\beta}{\partial\omega}\right)^{-1} \tag{7.19}$$

Unlike transmission-line modes, where β is a linear function frequency, β is not a linear function of frequency for waveguide modes; so u_p and u_g are not the same for waveguide modes. For metal waveguides, it is found from Eqs. (7.13), (7.18), and (7.19) that

$$u_p = \frac{u_{\mathrm{TEM}}}{\sqrt{1-(f_c/f)^2}} \tag{7.20}$$

and

$$u_g = u_{\mathrm{TEM}}\sqrt{1-(f_c/f)^2} \tag{7.21}$$

where $u_{\mathrm{TEM}} = 1/\sqrt{\mu\varepsilon}$ is the velocity of a plane wave in the dielectric.

Both u_p and u_g approach u_{TEM} as $f \to \infty$, which is an indication that waveguide modes appear more and more like TEM modes at high frequencies. But near cutoff, their behaviors are very different: u_g approaches zero, whereas u_p approaches infinity. This behavior of u_p may at first seem at odds with Einstein's theory of special relativity, which states that energy and matter cannot travel faster than the vacuum speed of light c. But this result is not a violation of Einstein's theory since neither information nor energy is conveyed by the phase of a steady-state waveform. Rather, the energy and information are transported at the group velocity, which is always less than or equal to c.

7.4.4. Dispersion

Unlike the modes on transmission lines, which exhibit differential propagation delays (i.e., *dispersion*) only when the materials are lossy or frequency dependent, waveguide modes are always dispersive, even when the dielectric is lossless and walls are perfectly conducting. The pulse spread per meter Δt experienced by a modulated pulse is equal to the difference between the arrival times of the lowest and highest frequency portions of the pulse. Since the envelope delay per meter for each narrow-band components of a pulse is equal to the inverse of the group velocity at that frequency, we find that the pulse spreading Δt for the entire pulse is given by

$$\Delta t = \left.\frac{1}{u_g}\right|_{\max} - \left.\frac{1}{u_g}\right|_{\min} \tag{7.22}$$

where $1/u_g|_{\max}$ and $1/u_g|_{\min}$ are the maximum and minimum inverse group velocities encountered within the pulse bandwidth, respectively. Using Eq. (7.21), the pulse spreading in metal waveguides can be written as

$$\Delta t = \frac{1}{u_{\mathrm{TEM}}}\left(\frac{1}{\sqrt{1-(f_c/f_{\min})^2}} - \frac{1}{\sqrt{1-(f_c/f_{\max})^2}}\right) \tag{7.23}$$

where $f_{\min}$ and $f_{\max}$ are the minimum and maximum frequencies within the pulse 3-dB bandwidth. From this expression, it is apparent that pulse broadening is most pronounced when a waveguide mode is operated close to its cutoff frequency f_c.

The pulse spreading specified by Eq. (7.23) is the result of *waveguide dispersion*, which is produced solely by the confinement of a wave by a guiding structure and has nothing to do with any frequency-dependent parameters of the waveguide materials. Other dispersive effects in waveguides are *material dispersion* and *modal dispersion*. Material dispersion is the result of frequency-dependent characteristics of the materials used in the waveguide, usually the dielectric. Typically, material dispersion causes higher frequencies to propagate more slowly than lower frequencies. This is often termed *normal dispersion*. Waveguide dispersion, on the other hand, causes the opposite effect and is often termed *anomalous dispersion*.

Modal dispersion is the spreading that occurs when the signal energy is carried by more than one waveguide mode. Since each mode has a distinct group velocity, the effects of modal dispersion can be very severe. However, unlike waveguide dispersion, modal dispersion can be eliminated simply by insuring that a waveguide is operated only in its dominant frequency range.

7.4.5. Effects of Losses

There are two mechanisms that cause losses in metal waveguides: dielectric losses and metal losses. In both cases, these losses cause the amplitudes of the propagating modes to decay as e^{-az}, where α is the attenuation constant, measured in units of Nepers per meter. Typically, the attenuation constant is considered as the sum of two components: $\alpha = \alpha_d + \alpha_c$, where α_d and α_c are the attenuation constants due to dielectric and metal losses alone, respectively. In most cases, dielectric losses are negligible compared to metal losses, in which case $\alpha \approx \alpha_c$.

Often, it is useful to specify the attenuation constant of a mode in terms of its decibel loss per meter length, rather than in Nepers per meter. The conversion formula between the two unit conventions is

$$\alpha\ (\mathrm{dB/m}) = 8.686 \times \alpha\ (\mathrm{Np/m}) \tag{7.24}$$

Both unit systems are useful, but it should be noted that α must be specified in Np/m when it is used in formulas that contain the terms of the form $e^{-\alpha z}$.

The attenuation constant α_d can be found directly from Eq. (7.11) simply by generalizing the dielectric wave number k to include the effect of the dielectric conductivity σ. For a lossy dielectric, the wave number is given by $k^2 = \omega^2\mu\varepsilon(1 + \sigma/j\omega\varepsilon)$, where σ is the conductivity of the dielectric, so the attenuation constant α_d due to dielectric losses alone is given by

$$\alpha_d = \mathrm{Re}\left(\sqrt{h^2 - \omega^2\mu\varepsilon\left(1 + \frac{\sigma}{j\omega\varepsilon}\right)}\right) \tag{7.25}$$

where Re signifies "the real part of" and h is the modal eigenvalue.

The effect of metal loss is that the tangential electric fields at the conductor boundary are no longer zero. This means that the modal fields exist both in the dielectric and the metal walls. Exact solutions for this case are much more complicated than the lossless case. Fortunately, a perturbational approach can be used when wall conductivities are high, as is usually the case. For this case, the modal field distributions over the cross section of the waveguide are disturbed only slightly; so a perturbational approach can be used to estimate the metal losses except at frequencies very close to the modal cutoff frequency [2].

This perturbational approach starts by noting that the power transmitted by a waveguide mode decays as

$$P = P_0 e^{-2\alpha_c z} \tag{7.26}$$

where P_0 is the power at $z = 0$. Differentiating this expression with respect to z, solving for α_c, and noting that dP/dz is the negative of the power loss per meter P_L, it is found that

$$a_c = \frac{1}{2}\frac{P_L}{P} \tag{7.27}$$

Expressions for α_c in terms of the modal fields can be found by first recognizing that the transmitted power P is integral of the average Poynting vector over the cross section S of the waveguide [1]:

$$P = \frac{1}{2}\mathrm{Re}\left(\int_S \mathbf{E} \times \mathbf{H}^* \cdot \mathbf{ds}\right) \tag{7.28}$$

where "*" indicates the complex conjugate, and "·" and "×" indicate the dot and cross products, respectively.

Similarly, the power loss per meter can be estimated by noting that the wall currents are controlled by the tangential H field at the conducting walls. When conductivities are high, the wall currents can be treated as if they flow uniformly within a skin depth of the surface. The resulting expression can be expressed as [1]

$$P_L = \frac{1}{2} R_s \oint_C |H|^2 dl \tag{7.29}$$

where $R_s = \sqrt{\pi f \mu / \sigma}$ is the surface resistance of the walls (μ and σ are the permeability and conductivity of the metal walls, respectively) and the integration takes place along the perimeter of the waveguide cross section.

As long as the metal losses are small and the operation frequency is not too close to cuttoff, the modal fields for the perfectly conducting case can be used in the above integral expressions for P and P_L. Closed form expressions for α_c for rectangular and circular waveguide modes are presented later in this chapter.

7.5. RECTANGULAR WAVEGUIDES

A rectangular waveguide is shown in Fig. 7.2, consisting of a rectangular metal cylinder of width a and height b, filled with a homogenous dielectric with permeability and permittivity μ and ε, respectively. By convention, it is assumed that $a \geq b$. If the walls are perfectly conducting, the field components for the TE$_{mn}$ modes are given by

$$E_x = H_0 \frac{j\omega\mu}{h_{mn}^2} \frac{n\pi}{b} \cos\left(\frac{m\pi}{a}x\right) \sin\left(\frac{n\pi}{b}y\right) \exp(j\omega t - r_{mn}z) \tag{7.30a}$$

$$E_y = -H_0 \frac{j\omega\mu}{h_{mn}^2} \frac{m\pi}{a} \sin\left(\frac{m\pi}{a}x\right) \cos\left(\frac{n\pi}{b}y\right) \exp(j\omega t - r_{mn}z) \tag{7.30b}$$

$$E_z = 0 \tag{7.30c}$$

$$H_x = H_0 \frac{\gamma_{mn}}{h_{mn}^2} \frac{m\pi}{a} \sin\left(\frac{m\pi}{a}x\right) \cos\left(\frac{n\pi}{b}y\right) \exp(j\omega t - r_{mn}z) \tag{7.30d}$$

$$H_y = H_0 \frac{\gamma_{mn}}{h_{mn}^2} \frac{n\pi}{b} \cos\left(\frac{m\pi}{a}x\right) \sin\left(\frac{n\pi}{b}y\right) \exp(j\omega t - r_{mn}z) \tag{7.30e}$$

$$H_z = H_0 \cos\left(\frac{m\pi}{a}x\right) \cos\left(\frac{n\pi}{b}y\right) \exp(j\omega t - r_{mn}z) \tag{7.30f}$$

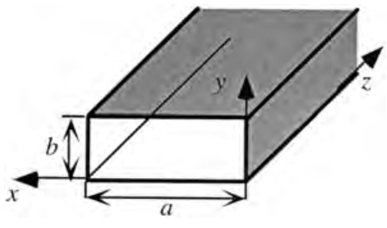


Figure 7.2 A rectangular waveguide.

The modal eigenvalues, propagation constants, and cutoff frequencies are

$$h_{mn} = \sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2} \tag{7.31}$$

$$\gamma_{mn} = \alpha_{mn} + j\beta_{mn} = j(2\pi f)\sqrt{\mu\varepsilon}\sqrt{1 - \left(\frac{f_{c_{mn}}}{f}\right)^2} \tag{7.32}$$

$$f_{c_{mn}} = \frac{1}{2\sqrt{\mu\varepsilon}}\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2} \tag{7.33}$$

For the TE$_{mn}$ modes, m and n can be any positive integer values, including zero, so long as both are not zero.

The field components for the TM$_{mn}$ modes are

$$E_x = -E_0\left(\frac{\gamma_{mn}}{h_{mn}^2}\right)\left(\frac{m\pi}{a}\right)\cos\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right)\exp(j\omega t - r_{mn}z) \tag{7.34a}$$

$$E_y = -E_0\frac{\gamma_{mn}}{h_{mn}^2}\frac{n\pi}{b}\sin\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right)\exp(j\omega t - r_{mn}z) \tag{7.34b}$$

$$E_z = E_0\sin\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right)\exp(j\omega t - r_{mn}z) \tag{7.34c}$$

$$H_x = E_0\frac{j\omega\varepsilon}{h_{mn}^2}\frac{n\pi}{b}\sin\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right)\exp(j\omega t - r_{mn}z) \tag{7.34d}$$

$$H_y = -E_0\frac{j\omega\varepsilon}{h_{mn}^2}\frac{m\pi}{a}\cos\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right)\exp(j\omega t - r_{mn}z) \tag{7.34e}$$

$$H_z = 0 \tag{7.34f}$$

where the values of h_{mn}, γ_{mn}, and $f_{c_{mn}}$ are the same as for the TE$_{mn}$ modes [Eqs. (7.31)–(7.33)]. For the TM$_{mn}$ modes, m and n can be any positive integer value except zero.

The dominant mode in a rectangular waveguide is the TE$_{10}$ mode, which has a cutoff frequency of

$$f_{c_{10}} = \frac{1}{2a\sqrt{\mu\varepsilon}} \tag{7.35}$$

The modal field patterns for this mode are shown in Fig. 7.3. Table 7.1 shows the cutoff frequencies of the lowest order rectangular waveguide modes (referenced to the

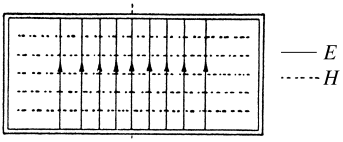


Figure 7.3 Field configuration for the TE$_{10}$ (dominant) mode of a rectangular waveguide. (Adapted from Ref. 2 with permission.)

Table 7.1 Cutoff Frequencies of the Lowest Order Rectangular Waveguide Modes for $a/b = 2.1$.

f_c/f_{c10}	Modes
1.0	TE_{10}
2.0	TE_{20}
2.1	TE_{01}
2.326	TE_{11}, TM_{11}
2.9	TE_{21}, TM_{21}
3.0	TE_{30}
3.662	TE_{31}, TM_{31}
4.0	TE_{40}

Frequencies are Referenced to the Cutoff Frequency of the Dominant Mode.

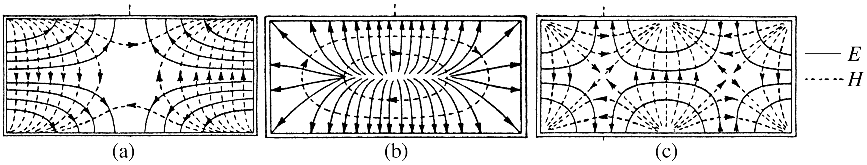


Figure 7.4 Field configurations for the TE_{11}, TM_{11}, and the TE_{21} modes in rectangular waveguides. (Adapted from Ref. 2 with permission.)

cutoff frequency of the dominant mode) when $a/b = 2.1$. The modal field patterns of several lower order modes are shown in Fig. 7.4.

The attenuation constants that result from metal losses alone can be obtained by substituting the modal fields into Eqs. (7.27)–(7.29). The resulting expressions are [3]

$$\alpha_{mn} = \frac{2R_s}{b\eta\left(1 - h_{mn}^2/k^2\right)^{1/2}}\left[\frac{h_{mn}^2}{k^2}\left(1 + \frac{b}{a}\right) + \frac{b}{a}\left(\frac{\varepsilon_{0m}}{2} - \frac{h_{mn}^2}{k^2}\right)\left(\frac{n^2ab + m^2a^2}{n^2b^2 + m^2a^2}\right)\right] \quad \text{TE modes} \tag{7.36}$$

and

$$\alpha_{mn} = \frac{2R_s}{b\eta\left(1 - h_{mn}^2/k^2\right)^{1/2}}\left(\frac{n^2b^3 + m^2a^3}{n^2b^2a + m^2a^3}\right) \quad \text{TM modes} \tag{7.37}$$

where $R_s = \sqrt{\pi f \mu/\sigma}$ is the surface resistance of the metal, η is the intrinsic impedance of the dielectric (377 Ω for air), $\varepsilon_{0m} = 1$ for $m = 0$ and 2 for $m > 0$, and the modal eigenvalues h_{mn} are given by Eq. (7.31). Figure 7.5 shows the attenuation constant for several lower order modes as a function of frequency. In each case, losses are highest at frequencies near the modal cutoff frequencies.

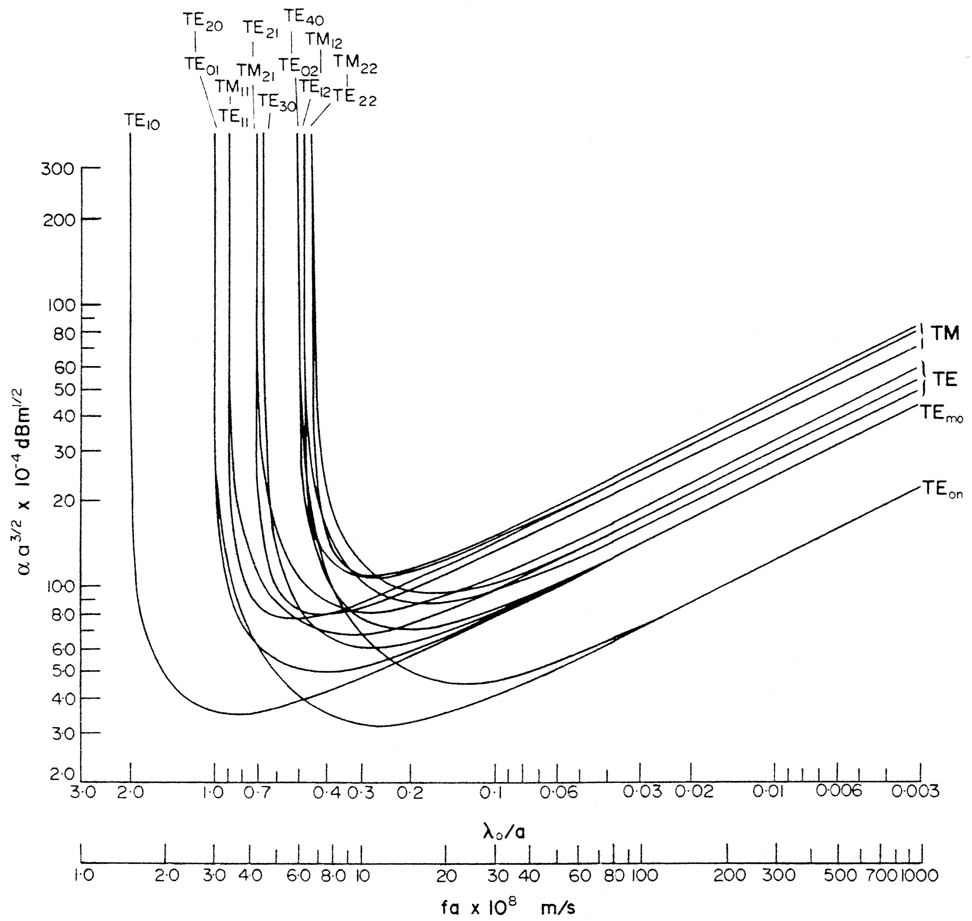


Figure 7.5 The attenuation constant of several lower order modes due to metal losses in rectangular waveguides with $a/b = 2$, plotted against normalized wavelength. (Adapted from Baden Fuller, A.J. *Microwaves*, 2nd Ed.; Oxford: Pergamon Press Ltd., 1979, with permission.)

7.6. CIRCULAR WAVEGUIDES

A circular waveguide with inner radius a is shown in Fig. 7.6, consisting of a rectangular metal cylinder with inside radius a, filled with a homogenous dielectric. The axis of the waveguide is aligned with the z axis of a circular-cylindrical coordinate system, where ρ and ϕ are the radial and azimuthal coordinates, respectively. If the walls are perfectly conducting, the equations for the TE$_{nm}$ modes are

$$E_\rho = H_0 \frac{j\omega\mu n}{h_{nm}^2 \rho} J_n(h_{nm}\rho) \sin n\phi \exp(j\omega t - \gamma_{nm} z) \tag{7.38a}$$

$$E_\phi = H_0 \frac{j\omega\mu}{h_{nm}} J_n'(h_{nm}\rho) \cos n\phi \exp(j\omega t - \gamma_{nm} z) \tag{7.38b}$$

$$E_z = 0 \tag{7.38c}$$

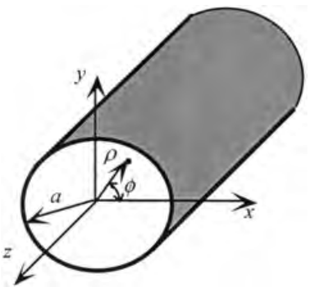


Figure 7.6 A circular waveguide.

$$H_\rho = -H_0 \frac{\gamma_{nm}}{h_{nm}} J_n'(h_{nm}\rho)\cos n\phi \exp(j\omega t - \gamma_{nm} z) \tag{7.38d}$$

$$H_\phi = H_0 \frac{\gamma_{nm} n}{h_{nm}^2 \rho} J_n(h_{nm}\rho)\sin n\phi \exp(j\omega t - \gamma_{nm} z) \tag{7.38e}$$

$$H_z = H_0 J_n(h_{nm}\rho)\cos n\phi \exp(j\omega t - \gamma_{nm} z) \tag{7.38f}$$

where n is any positive valued integer, including zero and $J_n(x)$ and $J_n'(x)$ are the regular Bessel function of order n and its first derivative [4,5], respectively, and μ and ε are the permeability and permittivity of the interior dielectric, respectively. The allowed modal eigenvalues h_{nm} are

$$h_{mn} = \frac{p_{nm}'}{a} \tag{7.39}$$

Here, the values p_{nm}' are roots of the equation

$$J_n'(p_{nm}') = 0 \tag{7.40}$$

where m signifies the mth root of $J_n'(x)$. By convention, $1 < m < \infty$, where $m = 1$ indicates the smallest root. Also for the TE modes,

$$\gamma_{nm} = \alpha_{nm} + j\beta_{nm} = j(2\pi f)\sqrt{\mu\varepsilon}\sqrt{1 - \left(\frac{f_{c_{nm}}}{f}\right)^2} \tag{7.41}$$

$$f_{c_{nm}} = \frac{p_{nm}'}{2\pi a\sqrt{\mu\varepsilon}} \tag{7.42}$$

The equations that define the TM$_{nm}$ modes in circular waveguides are

$$E_\rho = -E_0 \frac{\gamma_{nm}}{h_{nm}} J_n'(h_{nm}\rho)\cos n\phi \exp(j\omega t - \gamma_{nm} z) \tag{7.43a}$$

$$E_\phi = E_0 \frac{\gamma_{nm} n}{h_{nm}^2 \rho} J_n(h_{nm}\rho)\sin n\phi \exp(j\omega t - \gamma_{nm} z) \tag{7.43b}$$

$$E_z = E_0 J_n(h_{nm}\rho)\cos n\phi \exp(j\omega t - \gamma_{nm}z) \tag{7.43c}$$

$$H_\rho = -E_0 \frac{j\omega\varepsilon n}{h_{nm}^2 \rho} J_n(h_{nm}\rho)\sin n\phi \exp(j\omega t - \gamma_{nm}z) \tag{7.43d}$$

$$H_\phi = -E_0 \frac{j\omega\varepsilon}{h_{nm}} J_n'(h_{nm}\rho)\cos n\phi \exp(j\omega t - \gamma_{nm}z) \tag{7.43e}$$

$$H_z = 0 \tag{7.43f}$$

where n is any positive valued integer, including zero. For the TM$_{nm}$ modes, the values of the modal eigenvalues are given by

$$h_{nm} = \frac{p_{nm}}{a} \tag{7.44}$$

Here, the values p_{nm} are roots of the equation

$$J_n(p_{nm}) = 0 \tag{7.45}$$

where m signifies the mth root of $J_n(x)$, where $1 < m < \infty$. Also for the TM modes,

$$\gamma_{nm} = \alpha_{nm} + j\beta_{nm} = j(2\pi f)\sqrt{\mu\varepsilon}\sqrt{1 - \left(\frac{f_{c_{nm}}}{f}\right)^2} \tag{7.46}$$

$$f_{c_{nm}} = \frac{p_{nm}}{2\pi a\sqrt{\mu\varepsilon}} \tag{7.47}$$

The dominant mode in a circular waveguide is the TE$_{11}$ mode, which has a cutoff frequency given by

$$f_{c_{11}} = \frac{0.293}{a\sqrt{\mu\varepsilon}} \tag{7.48}$$

The configurations of the electric and magnetic fields of this mode are shown in Fig. 7.7. Table 7.2 shows the cutoff frequencies of the lowest order modes for circular waveguides, referenced to the cutoff frequency of the dominant mode. The modal field patterns of several lower order modes are shown in Fig. 7.8.

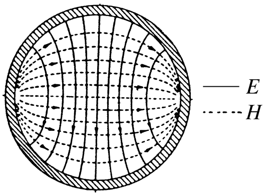


Figure 7.7 Field configuration for the TE$_{11}$ (dominant) mode in a circular waveguide. (Adapted from Ref. 2 with permission.)

Table 7.2 Cutoff Frequencies of the Lowest Order Circular Waveguide Modes.

f_c/f_{c11}	Modes
1.0	TE_{11}
1.307	TM_{01}
1.66	TE_{21}
2.083	TE_{01}, TM_{11}
2.283	TE_{31}
2.791	TM_{21}
2.89	TE_{41}
3.0	TE_{12}

Frequencies are Referenced to the Cutoff Frequency of the Dominant Mode.

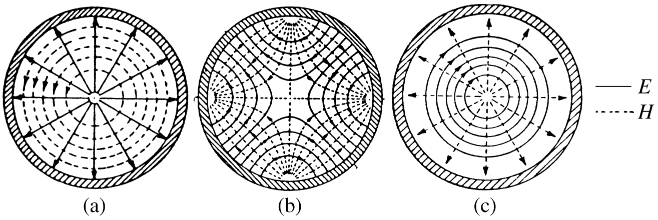


Figure 7.8 Field configurations of the TM_{01}, TE_{01}, and TE_{21} modes in a circular waveguide. (Adapted from Ref. 2 with permission.)

The attenuation constants that result from metal losses alone can be obtained by substituting the modal fields into Eqs. (7.27)–(7.29). The resulting expressions are [3]

$$\alpha_{nm} = \frac{R_s}{a\eta\left[1-(p'_{nm}/ka)^2\right]^{1/2}}\left[\frac{(p'_{nm})^2}{a^2k^2}+\frac{n^2}{(p'_{nm})^2-n^2}\right] \quad \text{TE modes} \tag{7.49}$$

and

$$\alpha_{nm} = \frac{R_s}{a\eta\left[1-(p_{nm}/ka)^2\right]^{1/2}} \quad \text{TM modes} \tag{7.50}$$

Figure 7.9 shows the metal attenuation constants for several circular waveguide modes, each normalized to the surface resistance R_s of the walls. As can be seen from this figure, the TE_{0m} modes exhibit particularly low loss at frequencies significantly above their cutoff frequencies, making them useful for transporting microwave energy over large distances.

7.7. COAXIAL-TO-WAVEGUIDE TRANSITIONS

When coupling electromagnetic energy into a waveguide, it is important to ensure that the desired mode is excited and that reflections back to the source are minimized, and

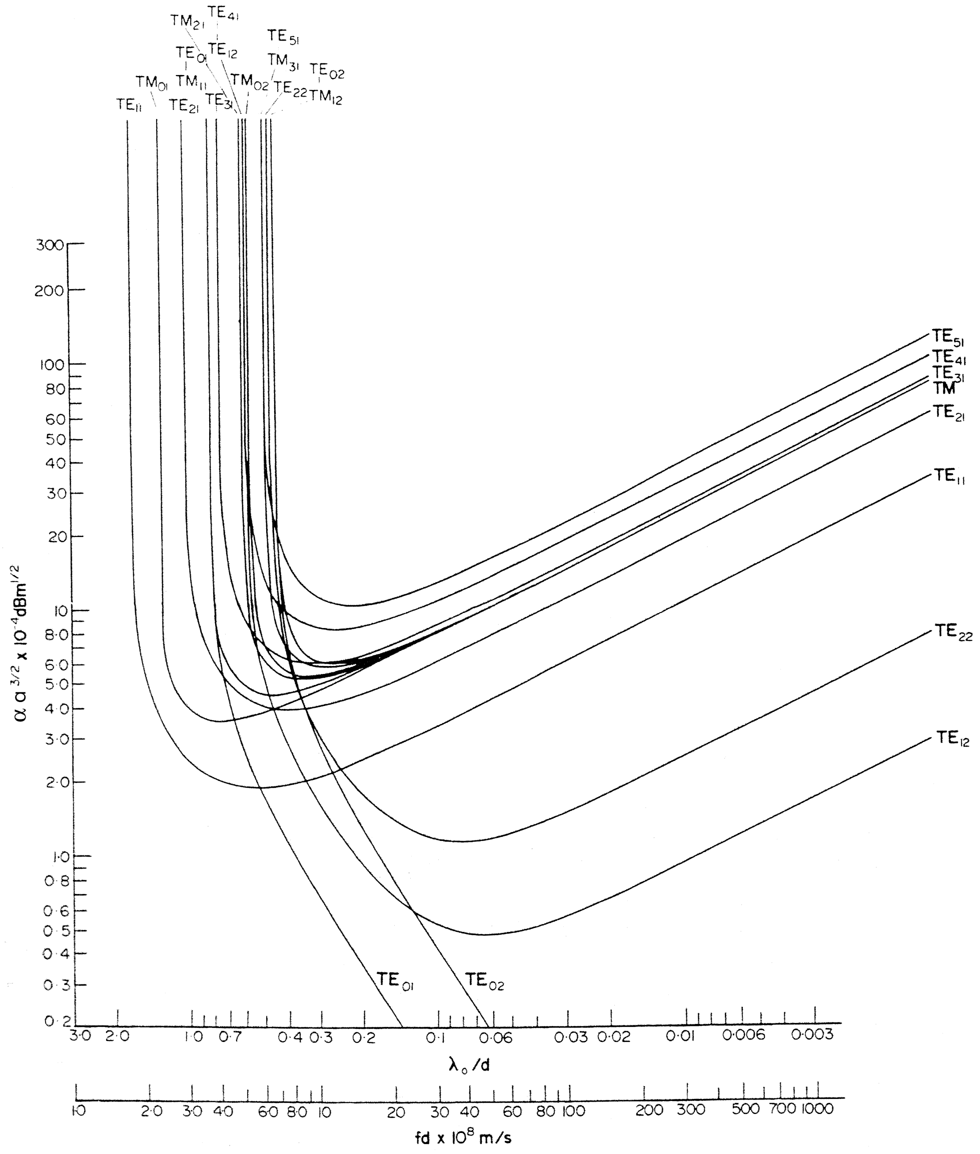


Figure 7.9 The attenuation constant of several lower order modes due to metal losses in circular waveguides with diameter *d*, plotted against normalized wavelength. (Adapted from Baden Fuller, A.J. *Microwaves*, 2nd Ed.; Oxford: Pergamon Press Ltd., 1979, with permission.)

that undesired higher order modes are not excited. Similar concerns must be considered when coupling energy from a waveguide to a transmission line or circuit element. This is achieved by using launching (or coupling) structures that allow strong coupling between the desired modes on both structures.

Figure 7.10 shows a mode launching structure launching the TE$_{10}$ mode in a rectangular waveguide from a coaxial transmission line. This structure provides good coupling between the TEM (transmission line) mode on the coaxial line and the

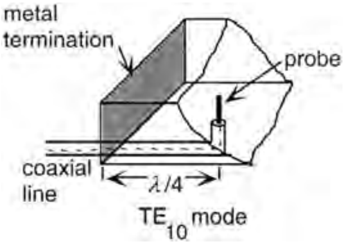


Figure 7.10 Coaxial-to-rectangular waveguide transition that couples the coaxial line to the TE_{10} waveguide mode.

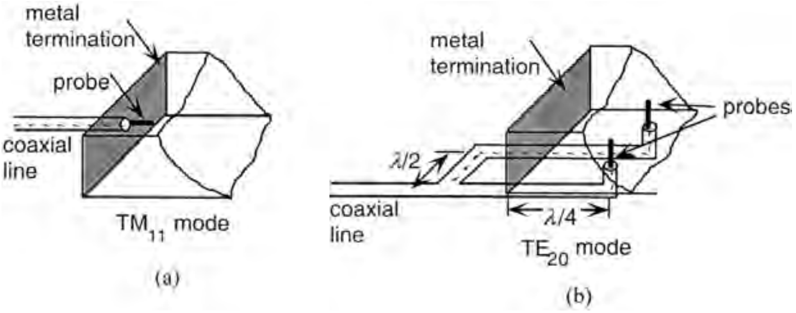


Figure 7.11 Coaxial-to-rectangular transitions that excite the TM_{11} and TM_{12} modes.

TE_{10} mode. The probe extending from inner conductor of the coaxial line excites a strong vertical electric field in the center of the waveguide, which matches the TE_{10} modal E field. The distance between the probe and the short circuit back wall is chosen to be approximately $\lambda_g/4$, which allows the backward-launched fields to reflect off the short circuit and arrive in phase with the fields launched toward the right.

Launching structures can also be devised to launch higher order modes. Mode launchers that couple the transmission line mode on a coaxial cable to the TM_{11} and TE_{20} waveguide modes are shown in Fig. 7.11.

7.8. COMPARATIVE SURVEY OF METAL WAVEGUIDES

All waveguides are alike in that they can propagate electromagnetic signal energy via an infinite number of distinct waveguide modes. Even so, each waveguide type has certain specific electrical or mechanical characteristics that may make it more or less suitable for a specific application. This section briefly compares the most notable features of the most common types: rectangular, circular, elliptical, and ridge waveguides.

Rectangular waveguides are popular because they have a relatively large dominant range and moderate losses. Also, since the cutoff frequencies of the TE_{10} and TE_{01} modes are different, it is impossible for the polarization direction to change when a rectangular waveguide is operated in its dominant range, even when nonuniformities such as bends and obstacles are encountered. This is important when feeding devices such as antennas, where the polarization of the incident field is critical.

Circular waveguides have a smaller dominant range than rectangular waveguides. While this can be a disadvantage, circular waveguides have several attractive features. One of them is their shape, which allows the use of circular terminations and connectors, which are easier to manufacture and attach. Also, circular waveguides maintain their shapes reasonably well when they are bent, so they can be easily routed between the components of a system. Circular waveguides are also used for making rotary joints, which are needed when a section of waveguide must be able to rotate, such as for the feeds of revolving antennas. Another useful characteristic of circular waveguides is that some of their higher order modes have particularly low loss. This makes them attractive when signals must be sent over relatively long distances, such as for the feeds of microwave antennas on tall towers.

An elliptical waveguide is shown in Fig. 7.12a. As might be expected by their shape, elliptical waveguides bear similarities to both circular and rectangular waveguides. Like circular waveguides, they are easy to bend. The modes of elliptical waveguides can be expressed in terms of Mathieu functions [6] and are similar to those of circular waveguides, but exhibit different cutoff frequencies for modes polarized along the major and minor axes of the elliptical cross section of the waveguide. This means that unlike circular waveguides, where the direction of polarization tends to rotate as the waves pass through bends and twists, modal polarization is much more stable in elliptical waveguides. This property makes elliptical waveguides attractive for feeding certain types of antennas, where the polarization state at the input to the antenna is critical.

Single and double ridge waveguides are shown in Fig. 7.12b and c, respectively. The modes of these waveguides bear similarities to those of rectangular guides, but can only be derived numerically [7]. Nevertheless, the effect of the ridges can be seen by realizing that they act as a uniform, distributed capacitance that reduces the characteristic impedance of the waveguide and lowers its phase velocity. This reduced phase velocity results in a lowering of the cutoff frequency of the dominant mode by a factor of 5 or higher, depending upon the dimensions of the ridges. Thus, the dominant range of a ridge waveguide is much greater than that of a standard rectangular waveguide. However, this increased frequency bandwidth is obtained at the expense of increased loss and decreased power handling capacity. The increased loss occurs because of the concentration of current flow on the ridges, with result in correspondingly high ohmic losses. The decreased power handling capability is a result of increased E-field levels in the vicinity of the ridges, which can cause breakdown (i.e., arcing) in the dielectric.

Waveguides are also available in a number of constructions, including rigid, semi-rigid, and flexible. In applications where it is not necessary for the waveguide to bend, rigid construction is always the best since it exhibits the lowest loss. In general, the more flexible the waveguide construction, the higher the loss.

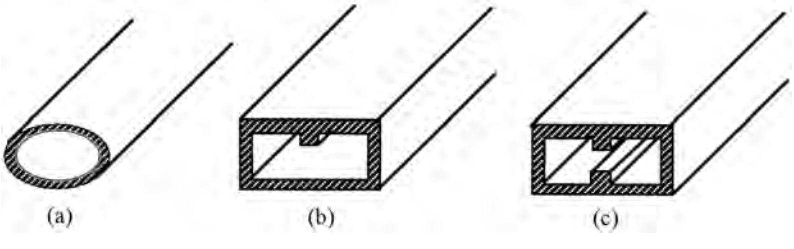


Figure 7.12 (a) Elliptical, (b) single-ridge, and (c) double-ridge waveguides.

7.9. CAVITY RESONATORS

Resonant circuits are used for a variety of applications, including oscillator circuits, filters and tuned amplifiers. These circuits are usually constructed using lumped reactive components at audio through RF frequencies, but lumped components become less desirable at microwave frequencies and above. This is because at these frequencies, lumped components either do not exist or they are too lossy.

A more attractive approach at microwave frequencies and above is to construct devices that use the constructive and destructive interferences of multiply reflected waves to cause resonances. These reflections occur in enclosures called *cavity resonators*. Metal cavity resonators consist of metallic enclosures, filled with a dielectric (possibly air). Dielectric resonators are simply a solid block of dielectric material, surrounded by air. Cavity resonators are similar to waveguides in that they both support a large number of distinct modes. However, resonator modes are usually restricted to very narrow frequency ranges, whereas each waveguide mode can exist over a broad range of frequencies.

7.9.1. Cylindrical Cavity Resonators

A cylindrical cavity resonator is shown in Fig. 7.13, consisting of a hollow metal cylinder of radius a and length d, with metal end caps. The resonator fields can be considered to be combinations of upward- and downward-propagating waveguide modes. If the dielectric inside the resonator is homogeneous and the conducting walls are lossless, the TE fields are

$$E_\rho = H_0 \frac{j\omega\mu n}{h_{nm}^2 \rho} J_n(h_{nm}\rho) \sin n\phi \left[A^+ e^{-j\beta_{nm}z} + A^- e^{j\beta_{nm}z}\right] e^{j\omega t} \tag{7.51a}$$

$$E_\phi = H_0 \frac{j\omega\mu}{h_{nm}} J_n'(h_{nm}\rho) \cos n\phi \left[A^+ e^{-j\beta_{nm}z} + A^- e^{j\beta_{nm}z}\right] e^{j\omega t} \tag{7.51b}$$

$$H_\rho = -H_0 \frac{\gamma_{nm}}{h_{nm}} J_n'(h_{nm}\rho) \cos n\phi \left[A^+ e^{-j\beta_{nm}z} - A^- e^{j\beta_{nm}z}\right] e^{j\omega t} \tag{7.51c}$$

$$H_\phi = H_0 \frac{\gamma_{nm} n}{h_{nm}^2 \rho} J_n(h_{nm}\rho) \sin n\phi \left[A^+ e^{-j\beta_{nm}z} - A^- e^{j\beta_{nm}z}\right] e^{j\omega t} \tag{7.51d}$$

$$H_z = H_0 J_n(h_{nm}\rho) \cos n\phi \left[A^+ e^{-j\beta_{nm}z} + A^- e^{j\beta_{nm}z}\right] e^{j\omega t} \tag{7.51e}$$

Here, the modal eigenvalues are $h_{nm} = p_{nm}'/a$, where the values of p_{nm}' are given by Eq. (7.40). To insure that E_ρ and E_ϕ vanish at $z = \pm d/2$, it is required that $A^- = A^+$ (even

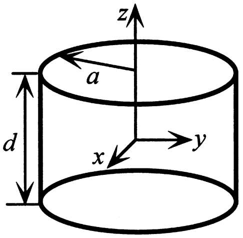


Figure 7.13 A cylindrical cavity resonator.

modes) or $A^- = -A^+$ (odd modes) and that β_{nm} be restricted to the values $l\pi/d$, where $l = 0, 1, \ldots$. Each value of l corresponds to a unique frequency, called a *resonant frequency*. The resonant frequencies of the TE_{nml} modes are

$$f_{nml} = \frac{1}{2\pi\sqrt{\mu\varepsilon}}\sqrt{\left(\frac{p'_{nm}}{a}\right)^2 + \left(\frac{l\pi}{d}\right)^2} \qquad (\mathrm{TE}_{nml}\text{ modes}) \tag{7.52}$$

In a similar manner, the TM fields inside the resonator are of the form

$$E_\rho = -E_0 \frac{\gamma_{nm}}{h_{nm}} J'_n(h_{nm}\rho)\cos n\phi\left[A^+ e^{-j\beta_{nm}z} + A^- e^{j\beta_{nm}z}\right]e^{j\omega t} \tag{7.53a}$$

$$E_\phi = E_0 \frac{\gamma_{nm} n}{h_{nm}^2 \rho} J_n(h_{nm}\rho)\sin n\phi\left[A^+ e^{-j\beta_{nm}z} + A^- e^{j\beta_{nm}z}\right]e^{j\omega t} \tag{7.53b}$$

$$E_z = E_0 J_n(h_{nm}\rho)\cos n\phi\left[A^+ e^{-j\beta_{nm}z} - A^- e^{j\beta_{nm}z}\right]e^{j\omega t} \tag{7.53c}$$

$$H_\rho = -E_0 \frac{j\omega\varepsilon n}{h_{nm}^2 \rho} J_n(h_{nm}\rho)\sin n\phi\left[A^+ e^{-j\beta_{nm}z} - A^- e^{j\beta_{nm}z}\right]e^{j\omega t} \tag{7.53d}$$

$$H_\phi = -E_0 \frac{j\omega\varepsilon}{h_{nm}} J'_n(h_{nm}\rho)\cos n\phi\left[A^+ e^{-j\beta_{nm}z} - A^- e^{j\beta_{nm}z}\right]e^{j\omega t} \tag{7.53e}$$

where the modal eigenvalues are $h_{nm} = p_{nm}/a$, where the values of p_{nm} are given by Eq. (7.45). Here, E_ϕ must vanish at $z = \pm d/2$, so it is required that $A^- = A^+$ (even modes) or $A^- = -A^+$ (odd modes) and that β_{nm} be restricted to the values $l\pi/d$, where $l = 0, 1, \ldots$. The eigenvalues of the TM_{nm} modes are different than the corresponding TE modes, so the resonant frequencies of the TM_{nml} modes are also different:

$$f_{nml} = \frac{1}{2\pi\sqrt{\mu\varepsilon}}\sqrt{\left(\frac{p_{nm}}{a}\right)^2 + \left(\frac{l\pi}{d}\right)^2} \qquad (\mathrm{TM}_{nml}\text{ modes}) \tag{7.54}$$

Figure 7.14 is a resonant mode chart for a cylindrical cavity, which shows the resonant frequencies of the lowest order modes as a function of the cylinder radius to length ratio. Here, it is seen that the TE_{111} mode has the lowest resonant frequency when $a/d < 2$, whereas the TM_{010} mode has the lowest resonant frequency when $a/d > 2$.

An important characteristic of a resonant mode is its quality factor Q, defined as

$$Q = 2\pi f \times \frac{\text{average energy stored}}{\text{power loss}} \tag{7.55}$$

At resonance, the average energies stored in the electric and magnetic fields are equal, so Q can be expressed as

$$Q = \frac{4\pi f_o W_e}{P_L} \tag{7.56}$$

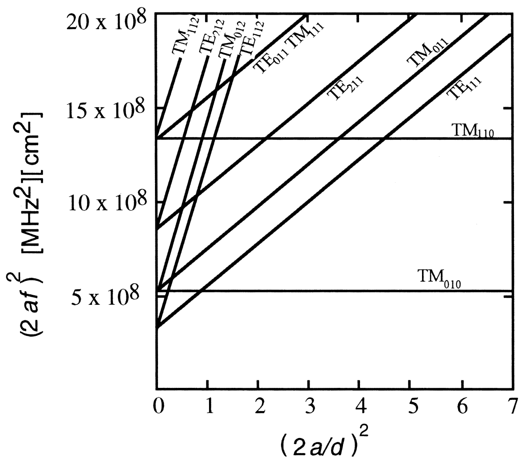


Figure 7.14 Resonant mode chart for cylindrical cavities. (Adapted from Collin, R. *Foundations for Microwave Engineering*; McGraw-Hill, Inc.: New York, 1992, with permission.)

where W_e is the time-average energy stored in electric field and P_L is the time-average dissipated power at resonance. This is the same definition for the quality factor as is used for lumped-element tuned circuits [8]. Also as in lumped circuits, the quality factor Q and the 3-dB bandwidth (BW) of a cavity resonator are related by

$$\mathrm{BW} = \frac{2\pi f_o}{Q} \quad [\mathrm{Hz}] \tag{7.57}$$

where f_o is the resonant frequency of the cavity.

The losses in metal resonators are nearly always dominated by the conduction losses in the cylinder walls. Similar to the way in which waveguide losses are evaluated, this power loss can be evaluated by integrating the tangential H fields over the outer surface of the cavity:

$$\begin{aligned} P_L &= \frac{R_s}{2}\oint_s H_{\tan}^2\, ds \\ &= \frac{R_s}{2}\left\{\int_0^{2\pi}\int_0^d \left[|H_\phi(\rho = a)|^2 + |H_z(\rho = a)|^2\right] a\, d\phi\, dz \right. \\ &\quad \left. + 2\int_0^a\int_0^{2\pi}\left[|H_\rho(z = 0)|^2 + |H_\phi(z = 0)|^2\right]\rho\, d\rho\, d\phi\right\} \end{aligned} \tag{7.58}$$

where R_s is the surface resistance of the conducting walls and the factor 2 in the second integral occurs because the losses on the upper and lower end caps are identical. Similarly, the energy stored in the electric field is found by integrating the electric energy density throughout the cavity.

$$W_e = \frac{\varepsilon}{4}\int_0^a\int_0^{2\pi}\int_{-d/2}^{d/2}\left(|E_\rho^2| + |E_\phi^2| + |E_z^2|\right)\rho\, d\rho\, d\phi\, dz \tag{7.59}$$

Using the properties of Bessel functions, the following expressions can be obtained for TE_{nml} modes [9]:

$$Q\frac{\delta}{\lambda_o} = \frac{\left[1-\left(n/p'_{nm}\right)^2\right]\left[(p'_{nm})^2+(l\pi a/d)^2\right]^{3/2}}{2\pi\left[\left(p'_{nm}\right)^2+(2a/d)(l\pi a/d)^2+\left(nl\pi a/p'_{nm}d\right)(1-2a/d)\right]} \qquad \mathrm{TE}_{nml}\ \text{modes} \qquad (7.60)$$

where $\delta = 1/\sqrt{\pi f \mu \varepsilon}$ is the skin depth of the conducting walls and λ_o is the free-space wavelength. Similarly, for TM_{nml} modes [9],

$$Q\frac{\delta}{\lambda_o} = \begin{cases} \dfrac{p_{nm}}{2\pi(1+2a/d)} & l=0 \\[2ex] \dfrac{\left[p_{nm}^2+(l\pi a/d)^2\right]^{1/2}}{2\pi(1+2a/d)} & l>0 \end{cases} \qquad \mathrm{TM}_{nml}\ \text{modes} \qquad (7.61)$$

Figure 7.15 shows the Q values of some of the lowest order modes as a function of the of the cylinder radius-to-length ratio. Here it is seen that the TE_{012} has the highest Q, which makes it useful for applications where a sharp resonance is needed. This mode also has the property that $H_\phi = 0$, so there are no axial currents. This means that the cavity endcaps can be made movable for tuning without introducing additional cavity losses.

Coupling between metal resonators and waveguiding structures, such as coaxial cables and waveguides, can be arranged in a variety of ways. Figure 7.16 shows three possibilities. In the case of Fig. 7.16a, a coaxial line is positioned such that the E field of the desired resonator mode is tangential to the center conductor probe. In the case of Fig. 7.16b, the loop formed from the coaxial line is positioned such that the H field of the desired mode is perpendicular to the plane of the loop. For waveguide to resonator coupling, an aperture is typically placed at a position where the H fields of both the cavity and waveguide modes have the same directions. This is shown in Fig. 7.16c.

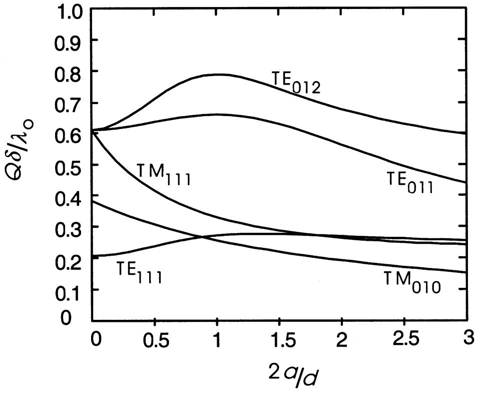


Figure 7.15 Q for cylindrical cavity modes. (Adapted from Collin, R. *Foundations for Microwave Engineering*; McGraw-Hill, Inc.: New York, 1992, with permission.)

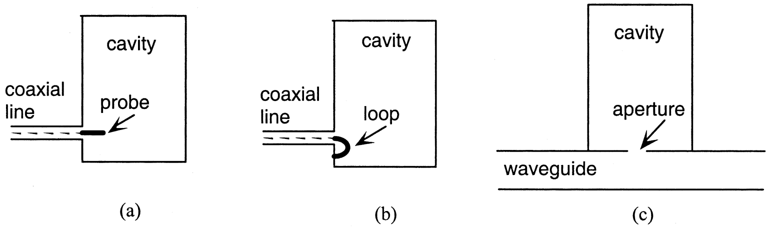


Figure 7.16 Coupling to methods for metal resonators. (a) probe coupling, (b) loop coupling, (c) aperture coupling.

7.9.2. Dielectric Resonators

A resonant cavity can also be constructed using a dielectric cylinder. Like metal cavity resonators, dielectric resonators operate on the principle of constructive interference of multiply reflected waves, but dielectric resonators differ in that some fringing or leakage of the fields occur at the dielectric boundaries. Although this fringing tends to lower the resonator Q values, it has the advantage that it allows easier coupling of energy into and out of these resonators. In addition, the high dielectric constants of these resonators allow them to be made much smaller than air-filled cavity resonators at the same frequencies. A number of dielectric materials are available that have both high dielectric constants, low loss-tangents ($\tan\delta$), and high temperature stability. Typical examples are barium tetratitanate ($\varepsilon_r = 37$, $\tan\delta = 0.0005$) and titania ($\varepsilon_r = 95$, $\tan\delta = 0.001$).

Just as in the case of metal cavity resonators, the modes of dielectric resonators can be considered as waveguide modes that reflect back and forth between the ends of the cylinder. The dielectric constants of dielectric resonators are usually much larger than the host medium (usually air), so the reflections at the air–dielectric boundaries are strong, but have polarities that are opposite to those obtained at dielectric–conductor boundaries. These reflections are much like what would be obtained if a magnetic conductor were present at the dielectric interface. For this reason, the TE modes of dielectric resonators bear similarities to the TM modes of metal cavity resonators, and vice versa.

An exact analysis of the resonant modes of a dielectric resonator can only be performed numerically, due to the difficulty of modeling the leakage fields. Nevertheless, Cohn [10] has developed an approximate technique that yields relatively accurate results with good physical insight. This model is shown in Fig. 7.17. Here, a dielectric cylinder of radius a, height d, and dielectric constant ε_r is surrounded by a perfectly conducting magnetic wall. The magnetic wall forces the tangential H field to vanish at $\rho = a$, which greatly simplifies analysis, but also allows fields to fringe beyond endcap boundaries.

The dielectric resonator mode that is most easily coupled to external circuits (such as a microstrip transmission line) is formed from the sum of upward and downward TE_{01} waves. Inside the dielectric ($|z| < d/2$), these are

$$H_z = H_0 J_o(k_\rho \rho)\left(A^+ e^{-j\beta z} + A^- e^{j\beta z}\right)e^{j\omega t} \tag{7.62a}$$

$$H_\rho = -H_0 \frac{j\beta}{k_\rho} J_o'(k_\rho \rho)\left(A^+ e^{-j\beta z} - A^- e^{j\beta z}\right)e^{j\omega t} \tag{7.62b}$$

$$E_\phi = H_0 \frac{j\omega\mu_o}{k_\rho} J_o'(k_\rho \rho)\left(A^+ e^{-j\beta z} + A^- e^{j\beta z}\right)e^{j\omega t} \tag{7.62c}$$

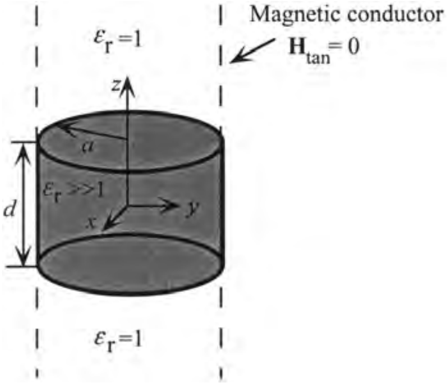


Figure 7.17 Magnetic conductor model of dielectric resonator.

where

$$\beta=\sqrt{\varepsilon_r k_o^2-k_\rho^2} \tag{7.63}$$

and $k_o=2\pi f\sqrt{\mu_o\varepsilon_o}$ is the free-space wave number. The value of k_ρ is set by the requirement that H_z vanishes at $\rho=a$, so

$$k_\rho a=p_{01}=2.4048 \tag{7.64}$$

Symmetry conditions demand that either $A^+=A^-$ (even modes) or $A^+=-A^-$ (odd modes).

The same field components are present in the air region ($|z|>d/2$), where there are evanescent fields which decay as $e^{-\alpha|z|}$, where the attenuation constant α is given by

$$\alpha=\sqrt{k_\rho^2-k_o^2} \tag{7.65}$$

Requiring continuity of the transverse electric and magnetic fields at the cylinder endcaps $z=\pm d/2$ yields the following resonance condition [11]:

$$\beta d=2\tan^{-1}\left(\frac{\alpha}{\beta}\right)+l\pi \tag{7.66}$$

where l is an integer. Using Eqs. (7.63) and (7.65), Eq. (7.66) can be solved numerically for k_o to obtain the resonant frequencies. The lowest order mode (for $l=0$) exhibits a less-than-unity number of half-wavelength variations along the axial coordinate z. For this reason, this mode is typically designated as the $\mathrm{TE}_{01\delta}$ mode.

An even simpler formula, derived empirically from numerical solutions, for the resonant frequency of the $\mathrm{TE}_{01\delta}$ mode is [12]

$$f_{\mathrm{GHz}}=\frac{34}{a_{\mathrm{mm}}\sqrt{\varepsilon_r}}\left(\frac{a}{d}+3.45\right) \tag{7.67}$$

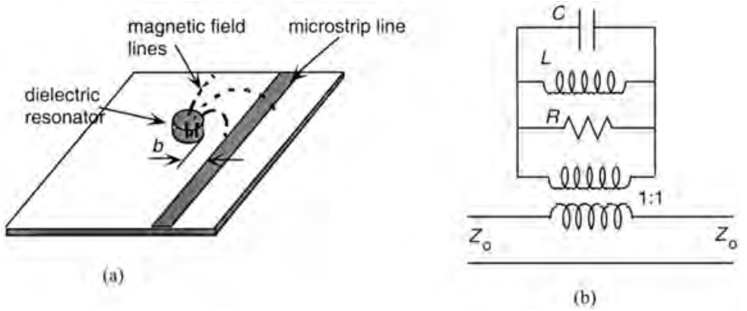


Figure 7.18 (a) Dielectric resonator coupled to a microstrip line and (b) the equivalent circuit.

where a_{mm} is the cylinder radius in millimeters. This formula is accurate to roughly 2% for the range $0.5 < a/d < 2$ and $30 < \varepsilon_r < 50$.

Dielectric resonators typically exhibit high Q values when low-loss dielectrics are used. In that case, radiation loss is the dominant loss mechanism, and typical values for the unloaded Q range from 100 to several thousand. For situations where higher Q values are required, the resonator can be placed in a shielding box. Care should be taken that the distance between the box and the resonator is large enough so that the resonant frequency of the resonator is not significantly affected.

Figure 7.18a shows a dielectric resonator that is coupled to a microstrip transmission line. Here, it is seen that the magnetic fields lines generated by the microstrip line couple strongly to the fringing magnetic field of the $\mathrm{TE}_{01\delta}$ mode. The amount of coupling between the the microstrip line and the resonator is controlled by the offset distance b between the resonator and the line.

The equivalent circuit that the resonator presents to the microstrip line is shown in Fig. 7.18b. In this model, the resonator appears as a parallel resonant circuit, coupled to the microstrip like through a 1:1 transformer. The resonator's resonant frequency f_o and unloaded Q are related to the lumped circuit parameters by the relations

$$f_o = \frac{1}{2\pi\sqrt{LC}} \tag{7.68}$$

$$Q = 2\pi f_o RC \tag{7.69}$$

The effect of the coupling between the resonator and the transmission line is to decrease the circuit Q. The larger the coupling, the smaller the overall Q. The coupling g between the resonator and the transmission line is defined as the ratio of the unloaded Q to the external Q. When both the source and load sides of the transmission line are terminated in matched loads, the external load presented to the resonator is $2Z_o$, so

$$g = \frac{Q}{Q_{\mathrm{ext}}} = \frac{\omega_o RC}{\omega_o(2Z_o)C} = \frac{R}{2Z_o} \tag{7.70}$$

where Z_o is the characteristic impedance of the transmission line. In practice, g can be determined experimentally by measuring the reflection coefficient Γ seen from the source end of the transmission line when both the source and load are matched to the transmission line. At resonance, the load seen by the source is $Z_o + R$, so the reflection coefficient is:

$$\Gamma = \frac{(Z_o + R) - Z_o}{(Z_o + R) + Z_o} = \frac{g}{1 + g} \tag{7.71}$$

Equations (7.68)–(7.71) can be used to uniquely determine the lumped parameters that a given resonator presents to a transmission line.

REFERENCES

1. Demarest, K.R. *Engineering Electromagnetics*; Prentice Hall: Upper Saddle River, NJ, 1998. Chapter 13, 508–563.
2. Marcuvitz, N. *Waveguide Handbook*, 2nd Ed., Peter Peregrinus Ltd.: London, 1986; 24–25.
3. Collin, R. *Foundations for Microwave Engineering*, 2nd Ed.; McGraw-Hill, Inc.: New York, 1992; 189.
4. Pozar, D. *Microwave Engineering*, 2nd Ed.; Wiley: New York, 2005, 683–686.
5. Abramowitz, M.; Stegun, I. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*; Wiley: New York, 1972, 358–374.
6. Kretzschmar, J. Wave propagation in hollow conducting elliptical waveguides. IEEE Transactions on Microwave Theory and Techniques **1970**, *18*: 547–554.
7. Montgomery, J. On the complete eigenvalue solution of ridged waveguide. IEEE Transactions on Microwave Theory and Techniques **1971**, *19*, 457–555.
8. Collin, R. *Foundations for Microwave Engineering*, 2nd Ed.; McGraw-Hill, Inc.: New York, 1992; 325–330.
9. Collin, R. *Foundations for Microwave Engineering*; McGraw-Hill, Inc.: New York, 1992; 506–507.
10. Cohn, R. Microwave bandpass filters containing high-Q dielectric resonators. IEEE Transaction on Microwave Theory and Techniques **1968**, *16*, 218–227.
11. Kajfez, D.; Guillon, P. *Dielectric Resonators*; Artech House, Inc.: Dedham, MA, 1986; 126–132.
12. Kajfez, D.; Guillon, P. *Dielectric Resonators*; Artech House, Inc.: Dedham, MA, 1986; 3.
13. Collin, R. *Foundations for Microwave Engineering*; McGraw-Hill, Inc.: New York, 1992; 508.

FURTHER INFORMATION

There are many textbooks and handbooks that cover the subject of waveguides in great detail. In addition to the references cited above, others include

Baden Fuller, A.J. *Microwaves: An Introduction to Microwave Theory and Techniques*; Pergamon Press: Oxford [England]; New York, 1990.

Cronin, N. *Microwave and Optical Waveguides*; Institute of Physics: Bristol; Philadelphia, 1995.

Liao, S. *Microwave Devices and Circuits*, 3rd Ed.; Prentice Hall: Englewood Cliffs, N.J., 1990.

Pozar, D. *Microwave Engineering*; Wiley: New York, 1998.

Carpentier, M.; Smith, B. *Microwave Engineering Handbook*; Van Nostrand Reinhold: New York, 1993.

Abramowitz, M.; Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*; Wiley: New York, 1972.

Lewin, L. *Theory of Waveguides*; Wiley: New York, 1975.

Collin, R.; Field, R.E. *Theory of Guided Waves*; 2nd Ed., IEEE Press: Piscataway, NJ, 1991.

Gardiol, F. *Introduction to Microwaves*; Artech House, Inc.: Dedham, MA, 1984.

Ramo, S.; Whinnery, J.; Van Duzer, J. *Fields and Waves in Communication Electronics*; Wiley: New York, 1994.

8
Antennas: Fundamentals

David Thiel
Griffith University
Nathan, Queensland, Australia

8.1. INTRODUCTION TO RADIATION

Electromagnetic radiation is one of the principal forms of conveying information from one point to another—from person to person, computer to computer, telephone to telephone and broadcast radio station to radio receiver. The radiation used in these communications systems usually lies in the frequency range from extremely low frequencies (ELF) to optical and ultraviolet (UV) frequencies. For example, ELF radiation (frequency band 3 Hz to 3 kHz) is used in through-earth propagation and telephone modems. Optical and UV frequencies are commonly used with optical fibers and sometimes in open-air links. Electromagnetic radiation can be trapped and directed along conductive wires (transmission lines), dielectric filled conducting pipes (wave guides), and in dielectric pipes sheathed with dielectric materials with a lower dielectric constant (optical fibers).

In many cases it is desirable to have a wireless EM link so that the radiation is unguided and will generally follow a line-of-sight path (i.e., a geometrical optics path). In the radio-frequency (RF)–microwave-frequency range, antennas are often used to launch and focus the radiation to a limited beam width so that the signal to noise ratio at the receiver is maximum and the interference to other wireless links in the same frequency band is minimized. An antenna is therefore a device that converts confined radiation from a transmission line or waveguide into an unguided but directed electromagnetic wave in the ambient medium (often, but not always, air).

While electromagnetic waves can propagate along the interface between media (e.g., surface waves) and in waveguides (e.g., TE and TM waves) in such a way that the electric and magnetic fields are not perpendicular to the direction of propagation, in most cases, the free-space radiation is in the form of a transverse electromagnetic (TEM) wave. For example, for a TEM wave, if we choose the electric field vector $\mathbf{E}$ to lie in the direction of the z axis, the magnetic field vector $\mathbf{H}$ to lie in the x direction, then the direction of propagation can be defined by the wave vector $\mathbf{k}$, which lies in the y direction. This is shown in Fig. 8.1.

The electric field strength $\mathbf{E}$ has the units of V/m, the magnetic field strength $\mathbf{H}$ has units A/m, so the power density of the electromagnetic wave is given by the

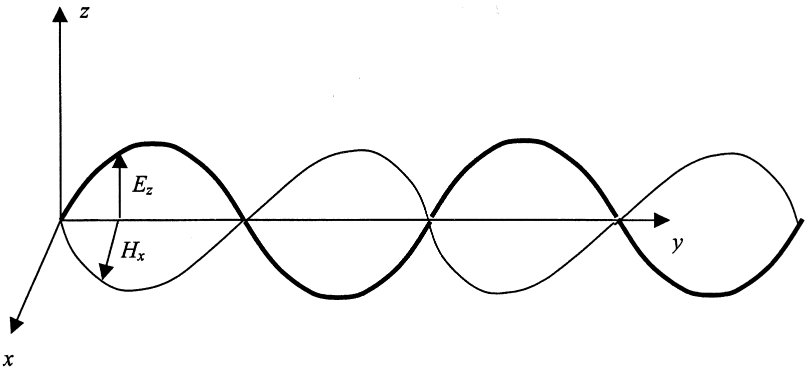


Figure 8.1 TEM wave with axis definition.

Poynting vector **S**, where

$$\mathbf{S} = \mathbf{E} \times \mathbf{H} \tag{8.1}$$

S has the units of watts per square meter and so is a measure of the power density of the radiation.

The radiation is described as being linearly polarized if the direction of the **E** field remains constant along the path of propagation.

In a TEM wave, it is possible for the direction of **E** to vary continuously in the xz plane perpendicular to the direction of radiation. In this case the radiation scribes an ellipse or circle as it propagates, and the radiation is called *elliptical* or *circular polarization*. Simple vector addition can show that elliptically polarized radiation has a linearly polarized component and a linearly polarized wave has a circularly polarized component.

The electric field vector **E** defines the force exerted on a charged particle in the presence of a TEM wave. If the charged particles are free to move, e.g., as electrons on the surface of a good conductor, a current is induced on the wire, and this can be detected and processed using standard electronic circuit techniques. Clearly, if a linearly polarized TEM wave has an **E** component in the z direction, then a straight wire in the z direction will have maximum current induced, whereas a wire directed in the x or y direction will have zero current. Thus a simple straight wire can be used to detect the presence on an electromagnetic wave, and so a wire is the basic form of a linearly polarized antenna.

The receiving characteristics of an antenna are identical to its transmitting characteristics; thus, descriptions of the properties of an antenna are equally valid in terms of the reception characteristics and transmission characteristics. This property is described in terms of the reciprocity principle for a communications link in which the transmitting and receiving antennas can be exchanged and the signal strength into the receiver is unchanged providing there is no media boundary in the vicinity of the antennas.

8.2. ANTENNA TERMINOLOGY

There are many different requirements for antenna systems. In broadcast applications (e.g., radio, television), it is desirable that the transmitted radiation can be detected over a large area. In point to point applications (e.g., fixed microwave link, communications with a fixed earth station and a geostationary satellite), it is desirable to confine the transmitted radiation to a small angle. In mobile communications applications, a point-to-point communications link is required, but the location of one point can move continuously during the transmission. In the case of multiple in–multiple out systems (MIMO), the location of the antennas can be quite varied. Recently there is increased interest in ad hoc radio networks or unplanned networks that self-assemble using smart antennas.

In designing a communications system, it is necessary to calculate the radiation strength at the receiver to ensure adequate signal-to-noise ratio for the correct interpretation of the signals received. This is called a *link budget calculation*. Therefore, it is necessary to specify the directional characteristics of the antenna in such a way that the power received by the target receiver can be calculated from the power delivered to the input terminals of the transmitting antenna.

It is convenient to specify the principal radiation direction in terms of a spherical polar coordinate system centered on the transmitting antenna. In Fig. 8.2, the principal radiation direction (main beam direction) of the antenna is (ϕ_0,θ_0), and the strength of the radiation in other directions is plotted as a three-dimensional surface. In this polar plot, the distance from the origin of the coordinate system (the phase center of the antenna) and a point on the surface represents the radiation field strength in that particular direction when measured in the far field, i.e., some considerable distance from the antenna. The spread in the radiation field can be defined in terms of the angular displacement from the principal direction of radiation where the field strength falls to one half the power (or −3 dB) in the principal direction. These half power points define the two beam widths $\Delta\theta$ and $\Delta\phi$ as shown in Fig. 8.2.

The directional characteristics can be described by two-dimensional and three-dimensional radiation patterns, in which the relative signal strength is plotted as a function of angle. The field strength is usually plotted in dBi. This is the gain in dB relative to an isotropic radiator having the same power output. As the radiation pattern represents the three-dimensional gain as a function of two angular directions (θ and ϕ), it is common to define a plane (e.g., the $\theta = 90°$ plane) and plot the signal strength for all angular positions

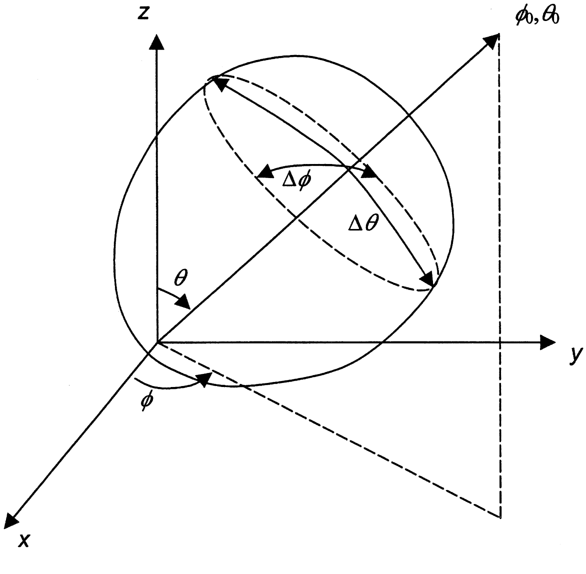


Figure 8.2 Main beam direction definition and beam width.

in that plane (e.g., all values of ϕ for the $\theta = 90°$ case). Usually that plane includes the origin of the coordinate system or the phase center of the antenna. In some applications such as an antenna located just above an infinite ground plane, the cut-plane for the radiation pattern includes the maximum radiation gain, which is elevated from the ground plane. In this case the radiation pattern is taken at a fixed elevation angle above the phase center of the antenna. In Fig. 8.2, this elevated radiation pattern would be located at $\theta = \theta_0$ for all values of ϕ.

Figure 8.3 is an example of a two-dimensional radiation pattern in which the −3 dB beam width is defined. The front-to-back ratio *FB* of an antenna is another important characteristic and is defined in terms of the ratio of the field strength in the direction (θ_0, ϕ_0) to the field strength in the opposite direction $(180° - \theta_0, \phi_0 + 180°)$. *FB* is usually defined in dB.

The directivity of an antenna is the ratio of the power density in the main beam to the average power density (i.e., total radiated power divided by 4π) [1]. The larger the value of the directivity, the more directional is the antenna. The directivity is always greater than 1.

The antenna efficiency is similar to the directivity but also includes losses in the antenna structure (e.g., the effect of finite conductivity, dielectric losses and sometimes even the impedance mismatch with the transmission feed line).

Antenna gain is the ratio of the radiation intensity in the main beam to the radiation intensity in every direction assuming that all radiated power is evenly distributed in all possible directions [1].

An antenna can also be described in terms of a circuit element connected to a transmission line. The input impedance of an antenna Z_a has both real and imaginary parts—the real part R_a relates to the loss of energy due to the radiated field and material losses; the imaginary part X_a relates to the inductive or capacitive load that the antenna structure presents. Maximum power transfer is achieved when the antenna input impedance is equal to the characteristic impedance Z_0 of the transmission feed line across the frequency range of interest. That is,

$$Z_a = R_a + jX_a = Z_0 \tag{8.2}$$

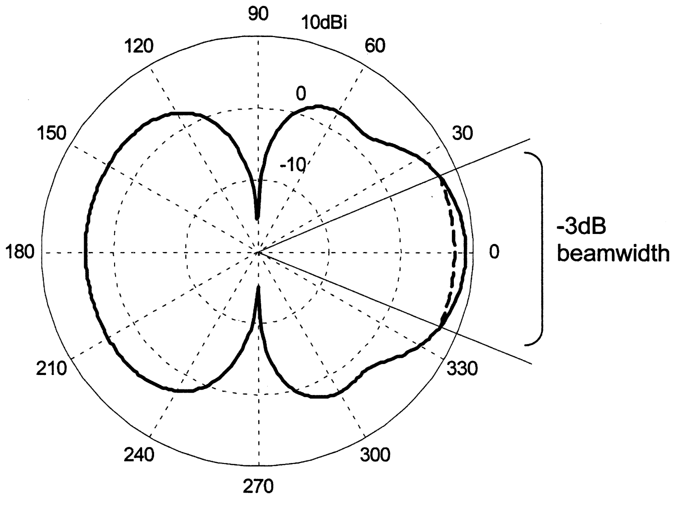


Figure 8.3 Typical two-dimensional radiation pattern illustrating the 3-dB beamwidth.

If there is an impedance mismatch, then there is reflection of the signal back into the transmission line. This is commonly described in terms of the scattering parameter S_{11} and can be determined using Eq. (8.3)

$$S_{11} = 20\log_{10}\left|\frac{Z_a - Z_0}{Z_a + Z_0}\right| \tag{8.3}$$

The resonant frequency f_0 of the antenna can be defined as that frequency where the reactance of the antenna is zero [1]. This can be shown to be true when the S_{11} value is a minimum. The frequency bandwidth of an antenna is commonly defined as the frequency range where the S_{11} value is less than $-10\,\text{dB}$. In numerical terms, this definition implies that the antenna constitutes an impedance mismatch that reflects less than 10% of the power back into the transmission line.

It is possible and sometimes desirable to define the resonant frequency of an antenna in terms of the radiation pattern or antenna gain rather than the impedance. This approach allows the designer to devote most attention to the radiation characteristics of the antenna rather than the impedance matching. It is possible to construct impedance matching circuits to reduce the impact of S_{11} on the link budget. Quarter wavelength chokes are one matching technique used with coaxial cables and microstrip lines [2,3].

The power delivered to the transmission line connected to the feed point of a receiving antenna has been extracted from the radiation falling on the antenna. The radiated field strength is measured in watts per square meter so that one can define the effective area of an antenna illuminated by the incoming radiation, even when the physical area of the antenna structure is very small. In some cases, such as a parabolic dish antenna or other aperture antennas, the antenna area is obvious. For wire antenna structures it is not so obvious, and the effective antenna area must be calculated from the antenna gain assuming uniform radiation is incident over the surface of the antenna [1,2].

Note that in receiving the power from an incoming radio wave, currents are excited in the antenna, which, in turn, cause the receiving antenna to radiate. Thus the maximum energy harvested from an antenna is one-half the energy falling on the antenna.

8.3. SIMPLE ANTENNA STRUCTURES

From the principle of reciprocity, it is possible to describe antennas in terms of their transmission or reception characteristics. In this section we will focus on reception characteristics—that is, the conversion of an incoming TEM wave to a current on a transmission line.

A linearly polarized TEM wave with an electric field component parallel to a conducting wire will induce a current to flow in the wire. This current is maximized if the wire forms part of a resonant circuit at the frequency of the incoming radiation. Thus a straight wire in air having length $l = \lambda/2$, with a transmission line connected to its center point has a fundamental resonance frequency f given by the equation

$$f = \frac{nc}{2l} \tag{8.4}$$

where $n = 1$. There are additional resonant frequencies for positive integer values of n.

At the resonant frequency, the current in the antenna is a standing wave. The RMS current along the length of the antenna element is one-half of a sinusoid with maximum current in the center and zero current on the ends. The voltage distribution on the antenna is approximately one half cosinusoidal so the impedance of the antenna at the center feed point is maximized. As noted before, there is maximum power transfer from the antenna to the transmission line when the antenna impedance is identical to the characteristic impedance of the transmission line. This can be achieved by using standard transmission line matching techniques, by adjusting the feed position along the wire, or by adjusting the inductive load on the antenna by changing the wire thickness or the wire length.

A short thin wire has a cosinusoidal radiation pattern in the plane containing the antenna wires (see Fig. 8.4). This is referred to as the *E-plane radiation pattern* as it lies parallel to the **E** field vector of the radiation from the antenna. There is no preferred direction in the plane perpendicular to the antenna wire, the so-called *H-plane radiation pattern* as the structure is completely symmetric about this line. The input resistance R_a for an electrically short dipole wire (i.e., $l \ll \lambda/2$) is given by [2,3]

$$R_a = 20\pi^2 \left(\frac{l}{\lambda}\right)^2 \tag{8.5}$$

The corresponding radiation pattern is found in all major antenna textbooks [2–5]. Also, the antenna impedance is linearly related to the length of the element provided the inequality

$$l < \frac{\lambda}{3}$$

remains valid.

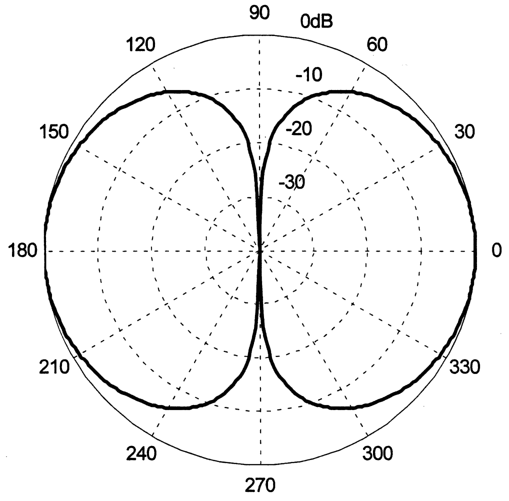


Figure 8.4 E-plane radiation pattern of a Hertzian dipole in dB. The gain has been normalized to 0 dB.

For a resonant straight-wire antenna, $l = \lambda/2$, the radiation pattern is still dependent on θ only and is given by

$$E(\theta, \phi) = \frac{\cos(\sin(\theta)\pi/2))}{\sin\theta} \tag{8.6}$$

and the antenna impedance

$$Z = 73 + 42.5j\,\Omega$$

This antenna is called a *half-wave dipole* and its radiation pattern is shown in Fig. 8.4 [2–5]. If the length of the antenna is reduced slightly, the imaginary part of the impedance can be reduced to zero and the antenna resonates with an input impedance which has a real component only [3].

A long straight wire with length l has a radiation pattern given by

$$E(\theta) = E_0 \frac{\cos(\cos(\theta)kl/2) - \cos(kl/2)}{\sin\theta} \tag{8.7}$$

where the wave number $k = 2\pi/\lambda$. Note that when the size of a radiating structure exceeds λ in one or more dimensions, the radiation pattern has side lobes and nulls. This is illustrated in Fig. 8.5 for a number of center-fed thin-wire antennas with different lengths.

A wire antenna located in the vicinity of a ground plane has its radiation pattern and impedance influenced by the ground plane because currents are induced to flow in the conductor. The simplest approach to understanding this type of antenna structure is to imagine that the ground plane can be replaced by an image antenna element which is located equidistant below the plane. This is illustrated in Fig. 8.6. The vertical current components are in phase with the source currents and the horizontal current components

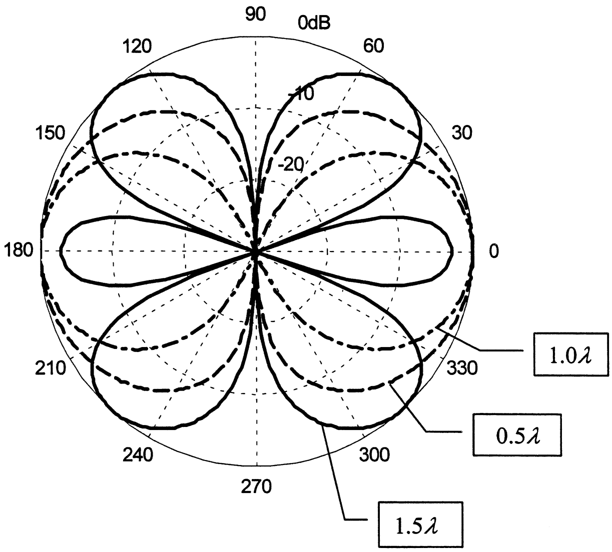


Figure 8.5 Radiation patterns for a number of thin-wire dipole antennas. The antennas lengths are 0.5 λ, 1.0 λ, and 1.5 λ as shown. All gains have been independently normalized to 0 dB.

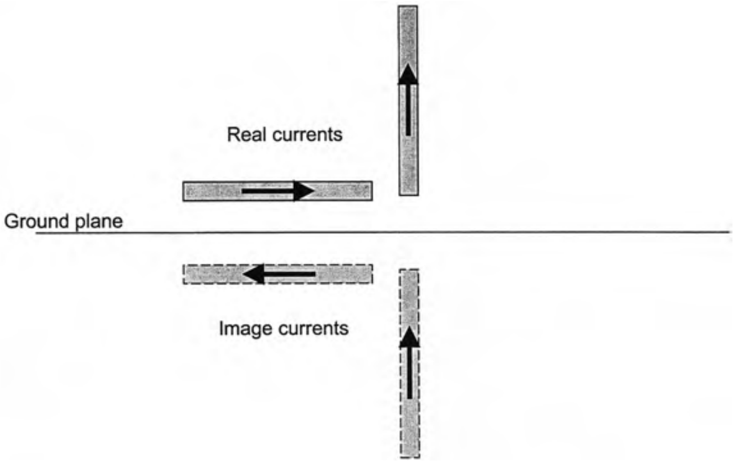


Figure 8.6 Current image elements reflected in the perfectly conducting ground plane of infinite extent. Note that currents normal to the ground plane have an image current in the same direction whereas horizontal currents have an image current in the opposite direction.

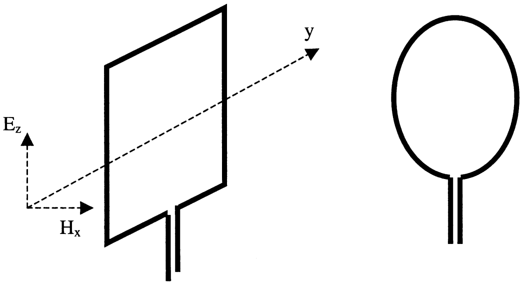


Figure 8.7 Simple loop antenna structures with balanced transmission lines.

are 180 degrees out of phase with the driven element. Thus a vertical wire element of length $\lambda/4$ with one end located on the ground plane has the radiation pattern of a half wave dipole in the hemisphere above the plane. The input impedance of this element is one-half that of the half-wave dipole. This antenna configuration is referred to as a quarter-wave monopole [2–5].

An alternative approach to constructing radiating structures is to use a conducting loop of wire. This can be considered to react to the magnetic field component of a TEM wave. The **H** field component of the radiation drives currents to circulate in the loop. Two simple, single turn, loop antenna structures are illustrated in Fig. 8.7. The current induced in a loop antenna can be increased by increasing the number of turns of wire in the loop structure. When the circumference p of the loop is very much smaller than λ, one obtains maximum response (i.e., the principal radiation direction) when **H** of the TEM wave is perpendicular to the plane of the loop (the yz plane in Fig. 8.7). This antenna is linearly polarized.

Larger sized loops with $p > \lambda$, are mainly used as folded dipole structures (see Fig. 8.8), which have the directionality of the equivalent dipole antenna but with a modified input impedance [2,3].

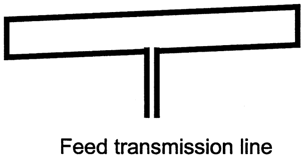


Figure 8.8 Folded dipole loop antenna.

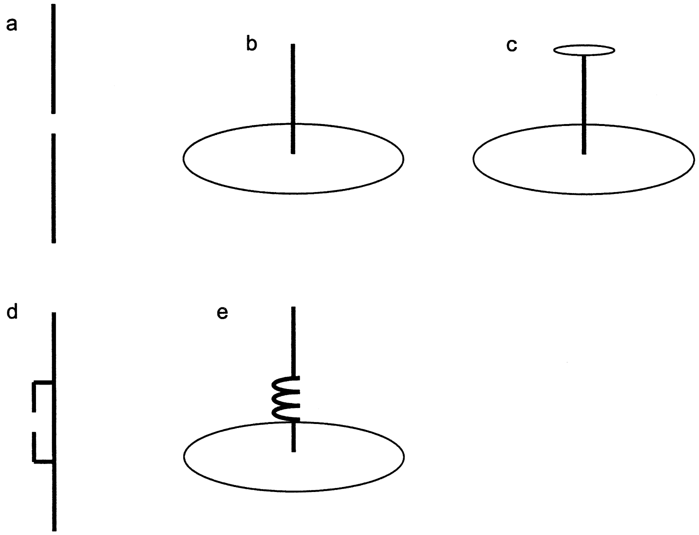


Figure 8.9 Variations on a straight wire antenna. (a) center-fed dipole antenna, (b) end-fed monopole antenna above a ground plane, (c) capacitive loaded monopole antennas, (d) T-feed dipole antennas, and (e) coil-loaded monopole antenna.

All other wire antennas can be described as variations of one of these basic antenna types: the wire monopole on a ground plane, a wire dipole in free space, a wire loop in free space, or a half loop attached to a ground plane. For example, Fig. 8.9 illustrates a number of variations of a straight wire antenna. Figure 8.9a is a center-fed straight thin-wire dipole. Figure 8.9b is a straight thin-wire monopole located on a perfectly conducting ground plane. The feed point of the antenna is at the base of the straight wire. Figure 8.9c is a capacitively loaded thin wire antenna (sometimes called a *capacitive plate antenna* [2]). The conductive disk at the top of the antenna is used to alter the input impedance of the antenna [2]. Figure 8.9d is a T-match configuration for a dipole antenna [2]. In this case the input impedance seen by the transmission line is modified by the feed position on the main antenna element. Figure 8.9e illustrates an end-fed monopole antenna on a ground plane which has a wire coil located partway along its length. This coil has the effect of providing a delay line between the lower straight wire element (usually $\lambda/4$) and the upper straight wire element (usually $\lambda/2$) to provide more gain in the horizontal direction. There are many other variations to straight wire, end-fed, monopole antennas on ground planes and center-fed dipole antennas. The gains of these antennas can be further improved through the use of reflecting planes, corner reflectors, and parasitic elements [2–5].

In order to respond to circularly polarized radiation, two half-wave dipoles oriented perpendicular to each other and perpendicular to the direction of the radiation will have maximum response to circular polarized radiation when one is fed 90 degrees out of phase with the other. When the phase shifter is deployed in the feed line of the other dipole, the sense of the circular polarization is reversed; i.e., right-hand circular polarization becomes left-hand circular polarization. Figure 8.10a shows a simple planar spiral antenna in which the arms of a dipole antenna have been shaped to respond to circularly polarized radiation. This antenna is fed by a balanced transmission line at the two terminals in the center of the antenna. The geometry of this antenna corresponds to the straight-line approximation to an Archimedean spiral [3]. The principal radiation direction is out of the plane. Figure 8.10b shows a helical antenna on a ground plane designed to respond to circularly polarized radiation [2]. The geometry of the helix (i.e., the radius, the number of turns, the total length of the wire, the diameter of the wire, and the pitch of the spiral) has a significant effect on the directional characteristics of the antenna [2,3].

One can increase the effective length of a wire antenna by embedding it in dielectric material (see Fig. 8.11a). A thin coating of dielectric on an end-fed monopole element results in the launching of a trapped surface wave mode in the dielectric. The effective wavelength for this trapped mode $\lambda_g < \lambda$, and the length of the resonant antenna is effectively reduced. The resonance condition required that the length of the monopole is approximately $\lambda_g/4$. The size reduction is dependent on the relative permittivity of the dielectric and the thickness of the coating [5].

In Fig. 8.9c, a top-loaded monopole is illustrated. If the top plate is sufficiently large compared to λ, then a waveguide mode is set up between the top plate and the ground plane. When this propagating wave reaches the end of the parallel plate waveguide, i.e., the edge of the top plate, energy is reflected back toward the source, and a standing wave can be set up between the two plates. Some energy, however, leaks past this termination and is launched as a linearly polarized TEM wave normal to the plane of the patch. This is the basis of a patch antenna element [2–5]. For a simple linearly polarized patch antenna, one can image that the two ends of the patch where the current is zero, are effectively two parallel radiating slots. As the two slots radiate in phase at the resonant frequency, the calculation of the radiation pattern is based on double slit interference where the separation distance between the two slits is approximately the length of the patch.

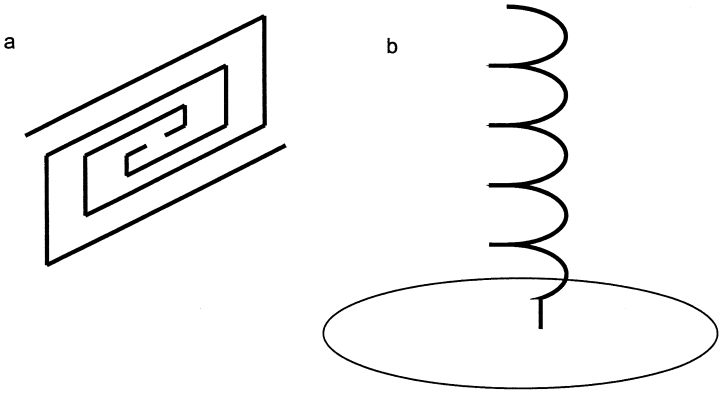


Figure 8.10 Circularly polarized wire antennas. (a) center-fed planar spiral antenna and (b) end-fed helical antenna above a ground plane.

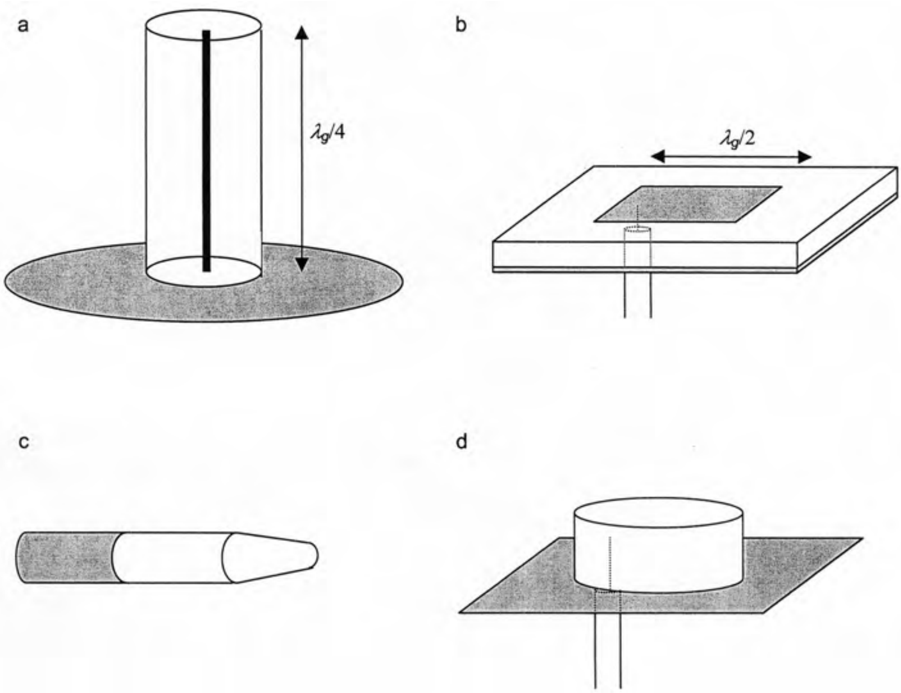


Figure 8.11 Antenna structures incorporating dielectric materials: (a) embedded monopole antenna on a ground plane, (b) patch antennas with coaxial feed probe, (c) dielectric rod antenna as the end of a circular waveguide, and (d) cylindrical dielectric resonator antenna on a ground plane with a coaxial feed probe.

These two edges of the patch are referred to as the *radiating edges*. The other two edges present no significant discontinuity to the current and so do not radiate. Note that with the probe fed patch antenna illustrated in Fig. 8.11b, there are two possible current directions. If the patch is square, then the position of the feed will determine which horizontal direction will provide the best impedance match to the transmission feed line. If the feed probe is located in the center of the patch in one direction, then the current will be maximized for this direction and the antenna impedance at this point is very small. A feed position offset from the center point along one axis can be chosen to provide a near perfect match to a standard coaxial transmission line.

If the patch is rectangular, then there are two possible resonant conditions at different frequencies. The radiation at the two frequencies will be linearly polarized but in orthogonal directions. The effectiveness of the antenna depends on the position of the feed point and the S_{11} of the impedance match.

Patch antenna structures (i.e., a single patch or multiple patches) can be manufactured using standard printed-circuit board photolithographic techniques. The relative permittivity of the substrate material controls the wavelength in the parallel plate waveguide. If the length of the patch is equal to $\lambda_g/2$ where λ_g is the wavelength of the radiation in the waveguide, the patch resonates and the launching efficiency of the antenna is high. The thickness and the relative permittivity of the substrate both have a significant effect on the bandwidth of the antenna. If the substrate is too thick, then the radiation efficiency is low because a trapped surface-wave mode can propagate through the substrate even when there is no metalization on the upper surface [2–5].

It is possible to launch a circularly polarized TEM wave from a square patch antenna through the use of two orthogonal feed points to create four radiating edges. Alternatively it is possible to use a single feed point by altering the overall shape of the patch with internal slots, truncated corners, etc., to make the propagation in the wave guide circular. Alternatively triangular, circular or other patch shapes can be used. Patch antennas can also be driven using microstrip lines and aperture coupled resonating elements.

A resonant patch can be achieved using a $\lambda_g/4$-long patch terminated along one radiating end with a short-circuit plane connected directly to the ground. In this case, it is important that the current is shorted along the entire length of the patch. This is a slightly more complex problem when constructing these patches.

Another class of antenna uses shaped dielectrics to either guide the radiation as a propagating wave along an elongated structure (a traveling wave dielectric antenna) or as a leakage of radiation from a resonating, compact dielectric structure (see Fig. 8.11c and d, respectively). In the first case, radiation from a waveguide is directed into the shaped dielectric rod where it escapes. The length and the shape of the dielectric material determine the characteristics of the radiation. Commonly, the dielectric material is several wavelengths long.

Figure 8.11d is referred to as a *dielectric resonator antenna* (DRA) where a probe feed is directed through a ground plane into a dielectric cylinder, hemisphere, or cube. The high-permittivity material acts as a resonator with some radiation leaking from the surfaces. The higher the relative permittivity of the dielectric material, the narrower is the impedance bandwidth of the DRA and the smaller is the structure.

The radiation characteristics of all antenna structures can be modified by the close proximity of undriven conducting and dielectric materials. A conducting resonant wire ($l = \lambda/2$) located close to a radiating half-wave dipole acquires an induced current, which means that the element (termed a *parasitic element*) contributes to the impedance of the antenna and its radiation pattern [5]. This effect is termed *mutual coupling*. Effective parasitic elements can be constructed from wires, patches and slots providing they are close to resonance. It is possible to design single feed, multiband antennas with two or more parasitic elements that resonate at frequencies different to the driven element. It is also common to use parasitic elements with resonance points close to that of the center frequency. This allows the total antenna to radiate with an enhanced impedance bandwidth. A Yagi-Uda antenna [2–5] in its traditional form made from wire antenna elements, consists of a center-fed half-wave dipole with a slightly longer "reflector" parasitic element behind it and a number of slightly shorter "director" parasitic elements in the direction of propagation (see Fig. 8.12). This type of antenna has increased bandwidth and narrower beamwidth when compared to a half-wave dipole in isolation.

Planar antennas can be fabricated using printed-circuit board photolithographic processes, which greatly reduces the cost of fabricating large arrays in addition to providing antennas which are conformal with the surface of the support structure. Conformal antennas are often desirable for aesthetic reasons and can be mounted discretely presenting minimal wind resistance. They can also be sealed from the environment using a waterproof coating. If the relative permittivity of the coating is greater than one, then the coating will decrease the resonant frequency of the antenna.

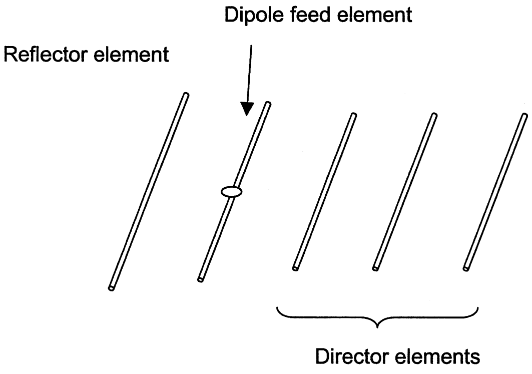


Figure 8.12 Five-element Yagi Uda antenna.

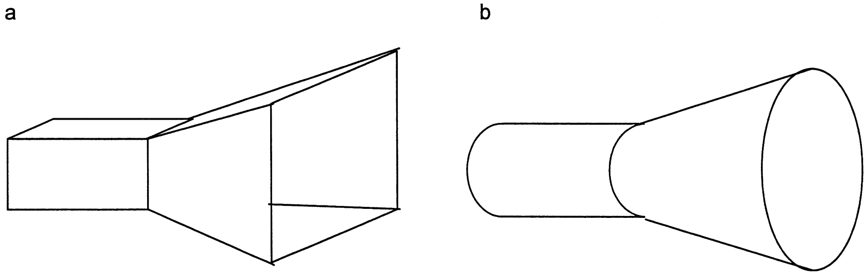


Figure 8.13 Aperture antennas connected to waveguides: (a) pyramidal horn and (b) circular horn antenna.

Another basic form of radiating system is an aperture antennas. These antennas are fed with a waveguide, and the antenna unit consists of a fitting at the end of the waveguide for impedance matching and directivity. Commonly the antenna is a rigid metallic fitting which increases the aperture size from that of the original waveguide. The flare can be in the form of a pyramidal horn or a circular horn (Fig. 8.13a and b, respectively). To reduce the front-to-back ratio, the inside surface of the horn can be treated with slots or dielectric coating so that the currents flowing on the inside of the horn do not flow on the outside surface. The presence of current on the outer side of the horn can greatly increase the back lobe and so decrease the front-to-back ratio of the antenna.

The final category of antenna we will discuss is reflector antennas. In this case, any of the antennas described previously may be placed at the focus of a conducting (and so reflecting) parabolic section. The radiation from such a combination is predominantly a parallel beam with the side-lobe levels and the nulls being principally determined by the size of the reflector. If the feed antenna is located along the axis of the parabolic surface it will lie in the path of the radiation. This effect is referred to as *feed blockage*, and the gain of the antenna is decreased. For this reason, the feed antenna at the focus should be as small as possible or offset from the main beam of the antenna.

All antennas have a radiation pattern that can be described by a function $F(\theta,\phi)$ and input impedance Z_a. It is possible to improve the directivity and gain of the antenna by using a number of identical elements all oriented in the same direction. Such arrangements are referred to as *antenna arrays*, and a number of common array configurations are discussed in the next section.

8.4. ANTENNA ARRAYS AND PATTERN SYNTHESIS

The gain of an antenna system can be increased significantly if many antennas are positioned in a simple, regular, geometrical configuration such as a straight line, a circle, or a plane (Fig. 8.14). By varying the amplitude and phase of the currents I_n, in the

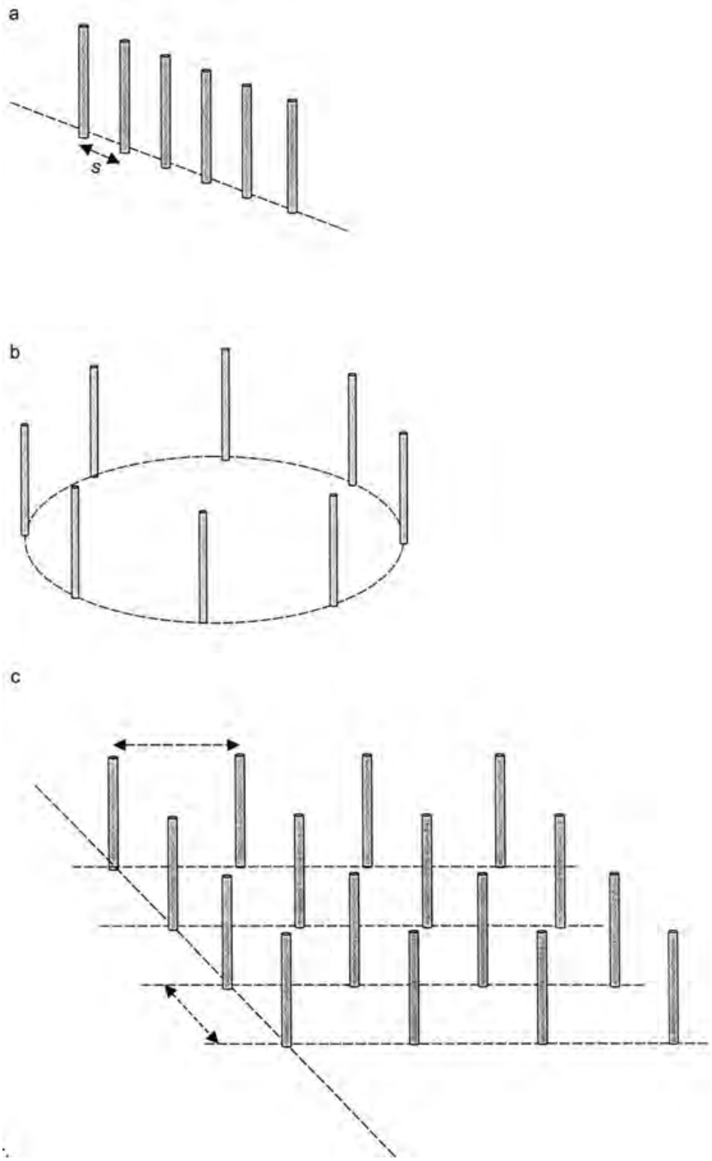


Figure 8.14 Regular monopole array structures: (a) linear array, (b) circular array, and (c) regular planar array.

elements of the array, the direction of the radiation can be altered. As an illustration, consider an equally spaced (along the z axis) linear array of N elements (Fig. 8.14a). Each element in the array is fed with the same current magnitude I with a regular stepped increase in phase of the elements ξ. Assume that each element has directional characteristics $F(\theta,\phi)$ and the antennas are sufficiently far apart ($s > \lambda$) to ensure that mutual coupling between adjacent elements is sufficiently small to ensure they radiate independently.

The radiation pattern $E_{tot}(\theta,\phi)$ for the array is the sum of the phase shifted fields from each element. Thus we can write

$$\begin{aligned} E_{tot}(\theta,\phi) &= IF(\theta,\phi) + (\theta,\phi)e^{jks\cos\theta+\xi} + IF(\theta,\phi)e^{j2ks\sin\theta+2\xi} + \cdots IF(\theta,\phi)e^{jNks\cos\theta+N\xi} \\ &= F(\theta,\phi)\,\frac{\sin N\varphi}{N\sin\varphi} = F(\theta,\phi)F_A(\theta,\phi)I \end{aligned} \tag{8.8}$$

where $\varphi = ks\cos\theta + \xi$.

The radiation pattern of the array is the product of the element radiation pattern $F(\theta,\phi)$ and the array factor $F_A(\theta,\phi)$ while the radiation intensity is directly proportional to the current magnitude I.

While F_A may be nonzero at a particular angular location (θ,ϕ), if this is the position of a null in the element factor $F(\theta,\phi)$, then clearly there will be a null in $E_{tot}(\theta,\phi)$ at this same angular position.

In extreme cases, the element spacing s can be a significant portion of the circumference of the earth. This type of array has a very narrow beam width and is referred to as an *interferometer array*. The most difficult task in these large arrays is ensuring phase coherence between the spaced receivers.

From the array analysis given above, a change in ξ will change the θ direction of the array. This in turn will change the position of the nulls and usually changes the beam width of the main beam. In a two-dimensional array of antennas, the direction of the main beam can be changed in both θ and ϕ directions using appropriate phase shifts. Careful control over the amplitude and the phase of the current in each element individually can improve the side-lobe levels and the gain of the antenna system. This requires significant computation if the number of elements in the array is large.

Note that when the elements in the array are too close, i.e., s is too small, then mutual coupling can restrict the size of the phase difference between adjacent elements regardless of the feed voltages. This can result in unexpected nulls in the radiation pattern. This effect is referred to as *scan blindness*.

The analysis used to derive the directional characteristics of an array can also be used in reverse; that is, if the radiation pattern is known, it is possible to calculate the magnitude and phase of the currents required on each element to generate such a pattern [2]. An alternative view of this process is to recognize that the far field radiation pattern is given by the two-dimensional Fourier transform of the spatial distribution of the antenna element currents. The current distribution in the array is the inverse Fourier transform of the far field radiation pattern.

Unfortunately this antenna design process is not simple as the far field pattern is of infinite extent and the phase distribution of the fields is unknown. This means an exact solution cannot be obtained and an iterative optimization procedure is required to obtain the best solution based on a defined cost function [5].

This optimization technique has been used to arrange appropriate satellite coverage of the more densely populated areas in the continents of the world. The array of waveguide manifolds and horns can be quite complex involving a large number of apertures with appropriate phase shifting waveguide lengths and power splitters and combiners.

8.5. SMART ANTENNAS

Electronically controlled antennas have been described as smart antennas. In such a system, the main beam direction, resonant frequency (or frequencies for multiband antennas), and null positions are altered electronically to ensure optimum signal-to-noise/interference (SNI) ratio. It is possible to achieve this control through the use of programmable attenuators and phase shifters in phased arrays or the feed position and resonant status of parasitic elements in switched parasitic antennas [5].

In the section on phased arrays, it was demonstrated that the main beam position is controlled by the magnitude and phase of the current in each element. Programmable phase shifters can be constructed using p.i.n. diodes or micro-electro-mechanical radio frequency (MEM RF) switches [6,7] to change the electrical length of the feed transmission line or by applying voltages to electrically active ferrite materials in waveguides. Note that these technologies have some level of insertion loss and the precise phase shift is difficult to control during manufacture. This usually means that the system, once constructed, must be calibrated precisely. Figure 8.15 shows the layout of a serpentine digital phase shifter based on open- and short-circuited switches. If N is the number of bits in the phase shifter, then there are 2^N switch positions, and the lengths of the transmission line elements are

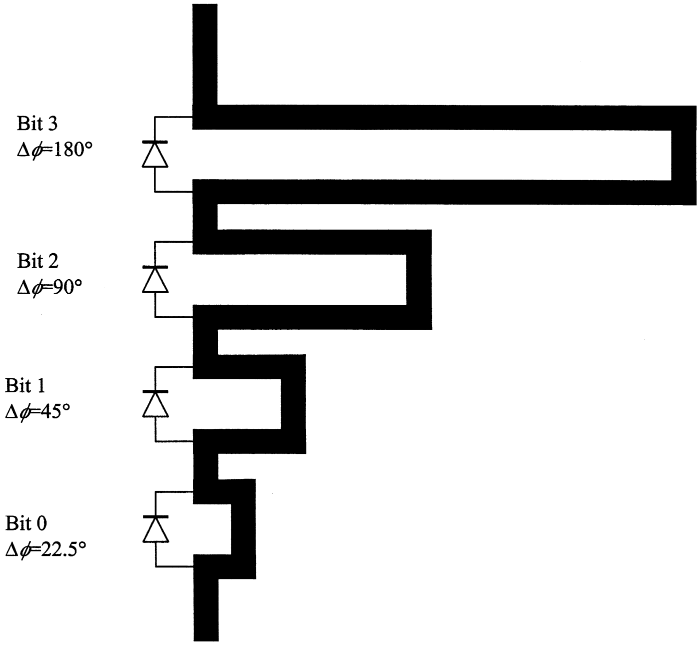


Figure 8.15 Serpentine digital phase shifter electrically controlled by p.i.n. diodes that short-circuit the microstrip line with an applied voltage.

$360n/N$ degrees, where $n = 1, 2, \ldots, N - 1$. A change in the least significant binary bit corresponds to a $360/N$-degree phase shift. A similar technique can be used to fabricate a digital attenuator in which different resistive loads are switched in and out of the transmission line.

Switched parasitic antenna principles can be illustrated simply using a 2- and 3-dipole antenna system [5]. In the switched active, switched parasitic antenna (SASPA), an RF switch is used to change the position of the feed from one dipole to the other (see Figure 8.16a). The inactive element is short circuited at its center to provide a reflector element. This is shown as F/S in Fig. 8.16a. In this way the main beam direction is changed by 180 degrees (Fig. 8.17). Note that the F/S combination in one element is reversed in the other (S/F).

The fixed active, switched parasitic antenna (FASPA) has two symmetrically located parasitic antennas located either side of the active element marked as F in Fig. 8.16b.

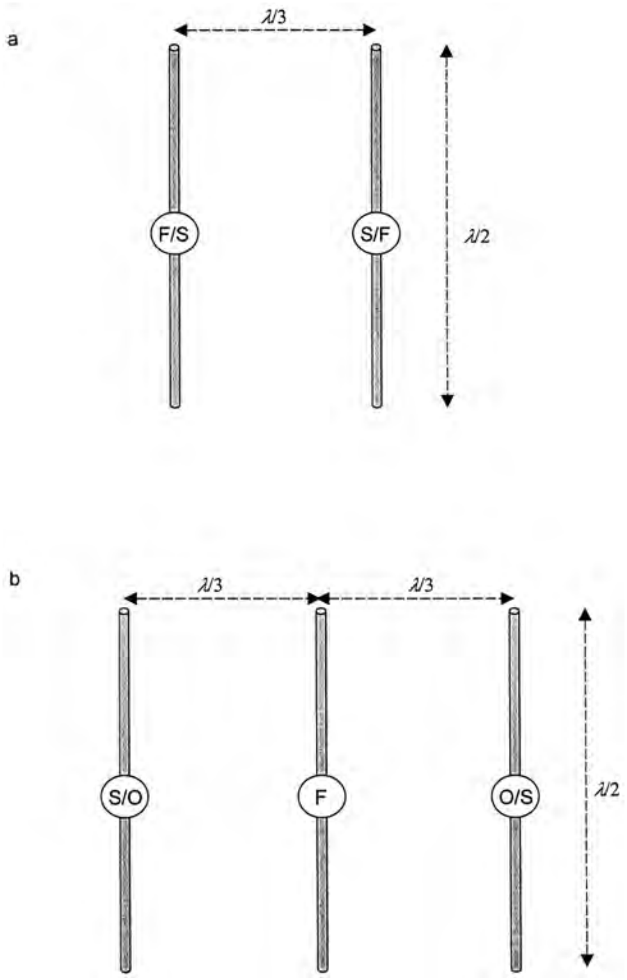


Figure 8.16 Switched parasitic antennas. (a) SASPA using two dipole elements: F/S allows the center of the element to be switched between an RF feed point and a short-circuit (parasitic) element. (b) FASPA using a fixed feed element (F) and two parasitic elements that can be switched between short and open circuit alternatively (S/O).

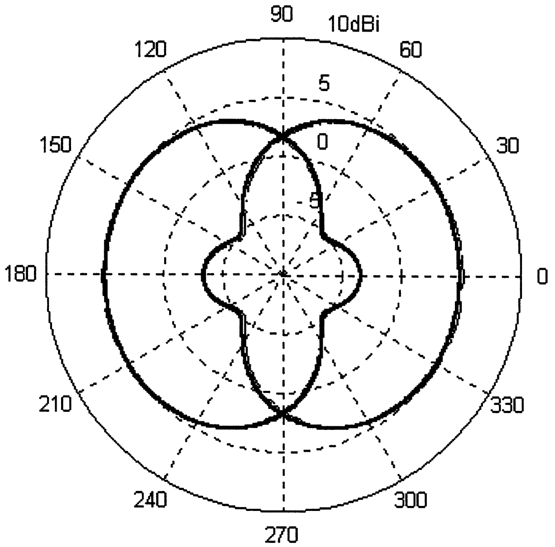


Figure 8.17 Radiation pattern for switched parasitic antennas shown in Fig. 8.16.

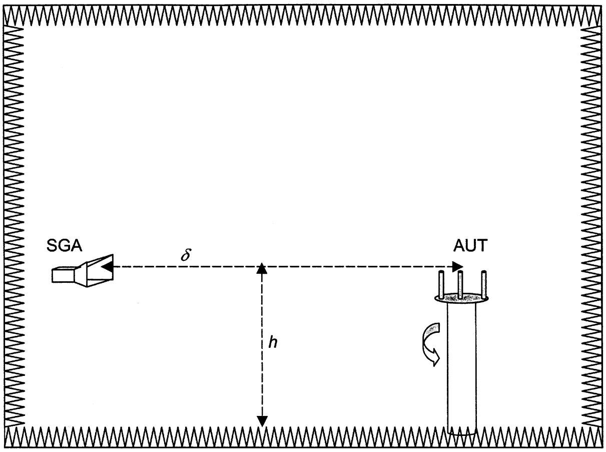


Figure 8.18 Anechoic chamber test facility for measuring antenna radiation patterns. The walls are lined with pyramidal absorber. The antenna under test (AUT) is rotated around a vertical axis while illuminated by the standard gain antenna (SGA).

One parasitic element switch is set to open circuit and the other to short circuit. This is shown as S/O and O/S in Fig. 8.16b. The radiation is directed away from the short-circuit element. When the switch settings are reversed, the principal direction is reversed. Note that in both cases (i.e., FASPA and SASPA), the input impedance of the feed point is independent of the direction of the main beam, and the beam width and null position remains constant relative to the main beam. The switching can be achieved using active devices (e.g., FET or p.i.n. diodes) or passive switches (e.g., RF MEMs switches). While the switches are driven by a DC voltage, the impedance of the switch at open and short circuit must provide sufficient RF isolation. This can limit the frequency of operation of switched parasitic antennas.

It is possible to arrange a continuous rather than discrete switching operation by changing the capacitive load at the center of the parasitic elements. This can be achieved using variable reactive loads [8]. This allows for a more continuous scan of the beam, but with potentially increased variation of the input impedance of the antenna.

Smart antennas must be controlled using a computer or microprocessor in a feedback loop. The SNI ratio into the detector needs to be determined during a scan of all possible main beam directions (a global scan), and then the signal source tracked using a dithering procedure (i.e., by checking the SNI in adjacent main beam positions) together with prediction techniques should the source or receiving antenna be moving. The response time of the smart antenna is dependent on the switching speed, the detector settling speed and the time required to verify the identification of the required source [5].

Two-dimensional phased arrays can have a main beam variation in both θ and ϕ directions, while the switched parasitic antennas are most commonly limited to control in one angular direction. The SNI is sometimes achieved with a large beamwidth antenna by positioning a sharp null in the direction of a strong interfering noise source rather than simply seeking the strongest signal. In a 2D phased array, the search procedure can be quite complex and the computation time required for the calculations may be significant. This is an active research area for digital processing specialists. Direction finding algorithms include the MUSIC technique used for multiple source location and tracking.

8.6. ANTENNA MEASUREMENTS

Following antenna fabrication, the radiation $F(\phi,\theta)$ and impedance $Z(f)$ descriptions of an antenna system usually must be verified experimentally. While there is now an extensive number of computer modeling packages and techniques that allow the accurate calculation of these parameters [3], variations in machining tolerances, the effects of finite conductivity and dielectric loss, and increasing problems associated with the spurious generation of intermodulation frequencies in adjacent frequency bands ensure that antenna testing remains very important.

When planning antenna measurements, there are several important factors to be considered [9]. The antenna performance must be isolated from its supporting structures unless they are to be part of the final installation environment. This commonly means that all objects (both conductive and dielectric) must be located away from the antenna by several Fresnel zones [1]. The first Fresnel zone is defined as the region of space in which the direct radiation path and any possible reflection path differ in length by less than $\lambda/2$. The second zone has reflections from a distance greater than $\lambda/2$ and less than λ. Higher order Fresnel zones follow this definition with the reflection path length increasing by $\lambda/2$ for the next Fresnel zone. Clearly, objects located in the near field of the radiating structure will influence both $F(\phi,\theta)$ and $Z(f)$. This general principle applies to ground reflections, side wall and ceiling reflections, in addition to the effects of feed cables to both the transmitter and receiver, and the antennas themselves. For this reason, for an antenna with a maximum aperture dimension of $D > \lambda$, the separation distance between the two antennas d must be greater than that given by the equation [9]

$$d > \frac{2D^2}{\lambda} \tag{8.9}$$

noting that D must be the maximum aperture dimension for both antennas. Secondly, the antennas should be at the same height h above a reflecting ground plane where [9]

$$h > 4D \tag{8.10}$$

if the antenna is to be rotated about a vertical axis (see Fig. 8.18).

Given these two restrictions, antenna measurements are generally made in an open range (i.e., no obstacles apart from the ground for many Fresnel zones) or in a chamber lined with EM absorbing materials—an anechoic chamber. A number of such materials are available commercially including pyramidal cones made from carbon impregnated foam [2] and flat ceramic tiles. Both are limited to particular frequency bands. In the case of radiation patterns with very large differences between the lobes and nulls (e.g., high-gain antennas), the finite reflections from these absorbing tiles can affect results significantly.

The gain of an antenna can be determined using two standard gain antennas (SGA) and the antenna under test (AUT), or two identical AUT's. The procedure requires the feed cables to both the transmitting antenna and the receiving antenna to be disconnected from the antennas and shorted together. The vector network analyzer is then calibrated for zero insertion loss and the received power P_{sc} noted for every frequency of interest. The two identical antennas (SGAs or AUTs) are separated by distance d_1 where $d_1 > d$ and d is defined by Eq. (8.9), and the received power P is again noted. The free-space path, loss P_L is given by the expression

$$P_L = 20\log\frac{2\pi d_1}{\lambda} \tag{8.11}$$

If all values are in dB, the gain of the identical antennas G_i in dB is given by the expression

$$G_i = \frac{P + P_L - P_{sc}}{2} \tag{8.12}$$

In the standard gain horn measurement, the gain of the AUT (G_a) is measured from a power measurement taken from the SGA $\Longleftrightarrow$ AUT measurement P_d using the following calculation:

$$G_a = P + P_L - P_{sc} - G_{sc} \tag{8.13}$$

where G_{sc} is the gain of the SGA. These gain determinations include the antenna mismatch S_{11} for both antennas. A simple, yet very sensitive, test to verify that the antennas are identical is to compare the $S_{11}(f)$ for both antennas.

The input impedance of the antenna is calculated with the antenna in free space or in an anechoic chamber. Initially the effects of the cable must be removed from the calculation. This requires a three-point calibration procedure across the frequency range of interest. A high quality short-circuit, open-circuit, and matched load termination must be sequentially applied to the vector network analyzer (VNA) and the reflected field strength (amplitude and phase) noted at every frequency. These measurements can be designated Γ_s, Γ_o, and Γ_0, respectively. The AUT is then attached to the end of the cable, and Γ_a measured. The impedance of the antenna is then determined mathematically.

The accuracy of the measurement depends strongly on the quality of the loads used in the calibration procedure in addition to the noise environment surrounding the antenna [2].

Commonly the calibration procedures are an intrinsic part of the operation of a VNA, and the user is prompted for the appropriate connections during the calibration procedure.

REFERENCES

1. *IEEE Standard Definitions of Terms for Antennas*; IEEE Press: Piscataway, NJ, IEEE Std 145-1993.
2. Balanis, C.A. *Antenna Theory Analysis and Design*; 2nd Ed.; Wiley: New York, 1997.
3. Stutzman, W.L.; Thiele, G.A. *Antenna Theory and Design*; 2nd Ed.; Wiley: New York, 1998.
4. Kraus, J.D. *Antennas*; 2nd Ed.; McGraw-Hill: New York, 1988.
5. Thiel, D.V.; Smith, S.A. *Switched Parasitic Antennas for Cellular Communications*; Artech House: Boston MA, 2001.
6. Brown, E.R. RF-MEMs switches for reconfigurable integrated circuits. IEEE Trans. Microwave Theory and Techniques **1998**, *46*, 1868–1880.
7. Goldsmith, C.L.; Yao, Z.; Eshelman, S.; Denniston, D. Performance of low-loss RF-MEMs capacitive switches. IEEE Microwave Guided Wave Lett. **1998**, *22*, 269–271.
8. Harrington, R.F. Reactively controlled directive arrays. IEEE Trans. Antennas Propag. **1978**, *26*, 390–395.
9. IEEE Antenna Standards Committee. IEEE standard test procedures for antennas. ANSI/IEEE Std 149-1979.

9
Antennas: Representative Types

David R. Jackson, Jeffery T. Williams, and Donald R. Wilton
University of Houston
Houston, Texas, U.S.A.

In Chapter 8, an overview of basic antenna terminology and antenna properties was given, including a discussion of concepts that are common to all antennas. In the present chapter, the discussion is focused on the specific properties of several representative classes of antennas that are commonly used. These include microstrip antennas, broadband antennas, phased arrays, traveling and leaky-wave antennas, and aperture antennas. Of course, in a single chapter, it is impossible to cover all of the types of antennas that are commonly used, or even to adequately cover all of the design aspects of any one type of antenna. However, this chapter should provide enough information about the five types of antennas that are discussed here to allow the reader to obtain a basic overview of the fundamental properties of these major classes of antennas and to see how the properties vary from one class to another. The references that are provided can be consulted to obtain more detailed information about any of these types of antennas or to learn about other types of antennas not discussed here.

9.1. MICROSTRIP ANTENNAS

9.1.1. Introduction

Microstrip antennas are one of the most widely used types of antennas in the microwave frequency range, and they are often used in the millimeter-wave frequency range as well [1–3]. (Below approximately 1 GHz, the size of a microstrip antenna is usually too large to be practical, and other types of antennas such as wire antennas dominate.) Also called *patch* antennas, microstrip patch antennas consist of a metallic patch of metal that is on top of a grounded dielectric substrate of thickness h, with relative permittivity and permeability ε_r and μ_r as shown in Fig. 9.1 (usually $\mu_r = 1$). The metallic patch may be of various shapes, with rectangular and circular being the most common, as shown in Fig. 9.1. Most of the discussion in this section will be limited to the rectangular patch, although the basic principles are the same for the circular patch. (Many of the CAD formulas presented will apply approximately for the circular patch if the circular patch is modeled as a square patch of the same area.) Various methods may be used to feed the patch, as discussed below.

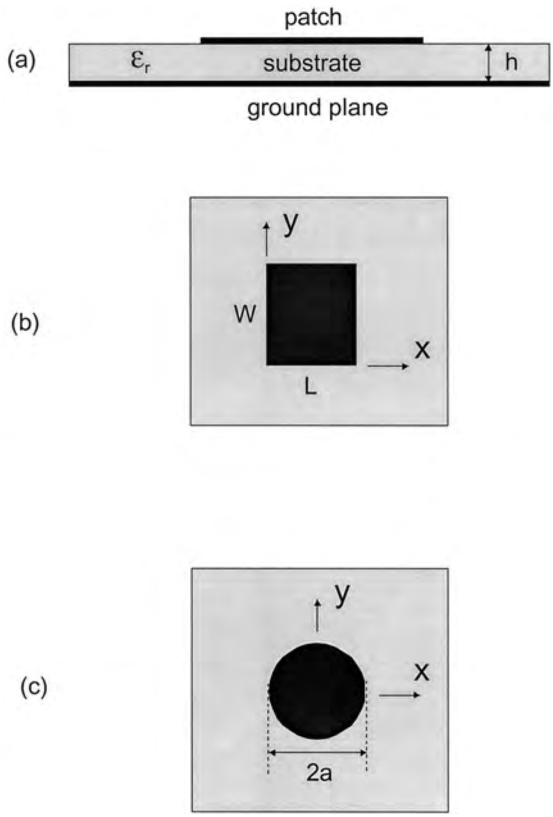


Figure 9.1 Geometry of microstrip patch antenna: (a) side view showing substrate and ground plane, (b) top view showing rectangular patch, and (c) top view showing circular patch.

One advantage of the microstrip antenna is that it is usually low profile, in the sense that the substrate is fairly thin. If the substrate is thin enough, the antenna actually becomes "conformal," meaning that the substrate can be bent to conform to a curved surface (e.g., a cylindrical structure). A typical substrate thickness is about $0.02\lambda_0$. The metallic patch is usually fabricated by a photolithographic etching process or a mechanical milling process, making the construction relatively easy and inexpensive (the cost is mainly that of the substrate material). Other advantages include the fact that the microstrip antenna is usually lightweight (for thin substrates) and durable.

Disadvantages of the microstrip antenna include the fact that it is usually narrowband, with bandwidths of a few percent being typical. Some methods for enhancing bandwidth are discussed later, however. Also, the radiation efficiency of the patch antenna tends to be lower than those of some other types of antennas, with efficiencies between 70% and 90% being typical.

9.1.2. Basic Principles of Operation

The metallic patch essentially creates a resonant cavity, where the patch is the top of the cavity, the ground plane is the bottom of the cavity, and the edges of the patch form the sides of the cavity. The edges of the patch act approximately as an open-circuit boundary condition. Hence, the patch acts approximately as a cavity with perfect electric conductor on the top and bottom surfaces, and a perfect "magnetic conductor" on the sides. This point of view is very useful in analyzing the patch antenna, as well as in understanding its behavior. Inside the patch cavity the electric field is essentially z directed and independent of the z coordinate. Hence, the patch cavity modes are described by a double index (m, n). For the (m, n) cavity mode of the rectangular patch in Fig. 9.1b, the electric field has the form

$$E_z(x, y) = A_{mn} \cos \frac{m\pi x}{L} \cos \frac{n\pi y}{W} \tag{9.1}$$

where L is the patch length and W is the patch width. The patch is usually operated in the (1, 0) mode, so that L is the resonant dimension, and the field is essentially constant in the y direction. The surface current on the bottom of the metal patch is then x directed, and is given by

$$J_{sx}(x) = A_{10}\left(\frac{\pi/L}{j\omega\mu_0\mu_r}\right)\sin\left(\frac{\pi x}{L}\right) \tag{9.2}$$

For this mode the patch may be regarded as a wide microstrip line of width W, having a resonant length L that is approximately one-half wavelength in the dielectric. The current is maximum at the center of the patch, $x = L/2$, while the electric field is maximum at the two "radiating" edges, $x = 0$ and $x = L$. The width W is usually chosen to be larger than the length ($W = 1.5\,L$ is typical) to maximize the bandwidth, since the bandwidth is proportional to the width. [The width should be kept less than twice the length, however, to avoid excitation of the (0, 1) mode.]

At first glance, it might appear that the microstrip antenna will not be an effective radiator when the substrate is electrically thin, since the patch current in Eq. (9.2) will be effectively shorted by the close proximity to the ground plane. If the modal amplitude A_{10} were constant, the strength of the radiated field would in fact be proportional to h. However, the Q of the cavity therefore increases as h decreases (the radiation Q is inversely proportional to h). Therefore, the amplitude A_{10} of the modal field at resonance is inversely proportional to h. Hence, the strength of the radiated field from a resonant patch is essentially independent of h, if losses are ignored. The resonant input resistance will likewise be nearly independent of h. This explains why a patch antenna can be an effective radiator even for very thin substrates, although the bandwidth will be small.

9.1.3. Feeding Techniques

The microstrip antenna may be fed in various ways. Perhaps the most common is the direct probe feed, shown in Fig. 9.2a for a rectangular patch, where the center conductor of a coaxial feed line penetrates the substrate to make direct contact with the patch. For linear polarization, the patch is usually fed along the centerline, $y = W/2$. The feed point location at $x = x_f$ controls the resonant input resistance. The input resistance is highest when the patch is fed at the edge and smallest (essentially zero) when the patch is fed at the center ($x = L/2$). Another common feeding method, preferred for planar fabrication, is the direct-contact microstrip feed line, shown in Fig. 9.2b. An inset notch is used to control the resonant input resistance at the contact point. The input impedance seen by the

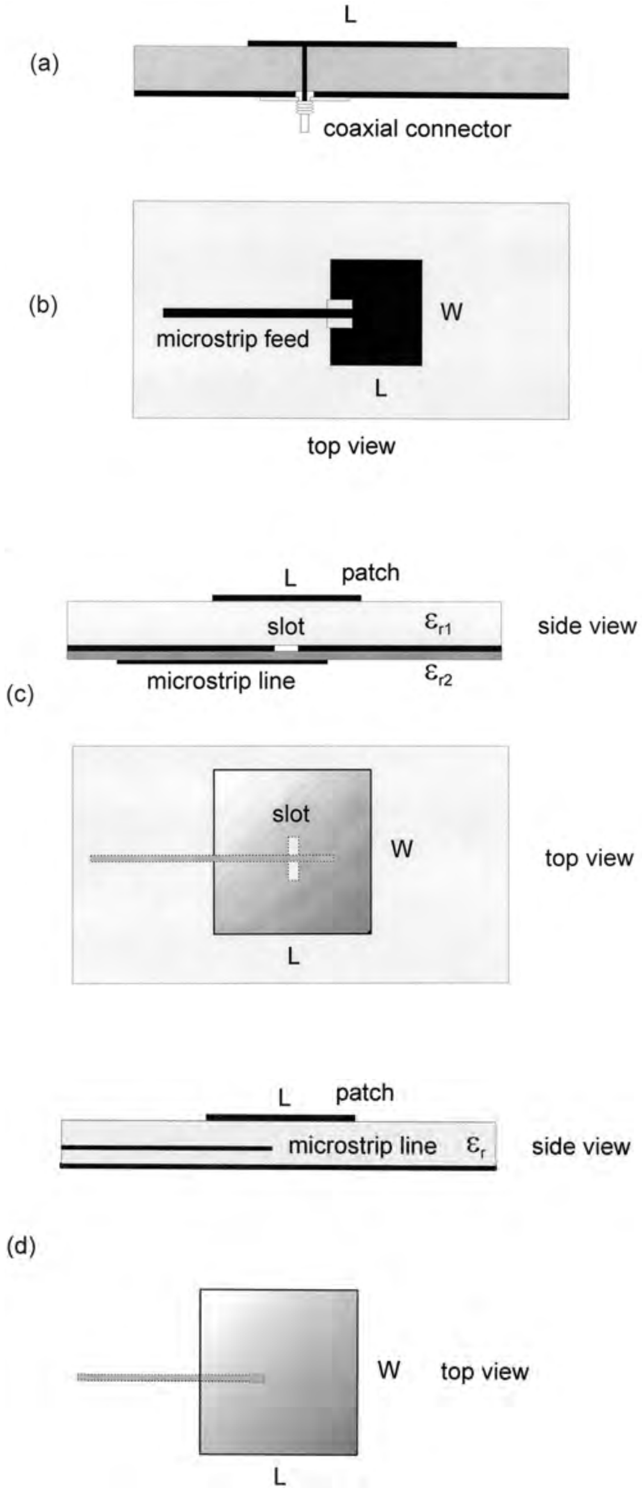


Figure 9.2 Common feeding techniques for a patch antenna: (a) coaxial probe feed, (b) microstrip line feed, (c) aperture-coupled feed, and (d) electromagnetically coupled (proximity) feed.

microstrip line is approximately the same as that seen by a probe at the contact point, provided the notch does not disturb the modal field significantly.

An alternative type of feed is the aperture-coupled feed shown in Fig. 9.2c. In this scheme, a microstrip line on a back substrate excites a slot in the ground plane, which then excites the patch cavity. This scheme has the advantage of isolating the feeding network from the radiating patch element. It also overcomes the limitation on substrate thickness imposed by the feed inductance of a coaxial probe, so that thicker substrates and hence higher bandwidths can be obtained. Using this feeding technique together with a foam substrate, it is possible to achieve bandwidths greater than 25% [4].

Another alternative, which has some of the advantages of the aperture-coupled feed, is the "electromagnetically coupled" or "proximity" feed, shown in Fig. 9.2d. In this arrangement the microstrip line is on the same side of the ground plane as the patch, but does not make direct contact. The microstrip line feeds the patch via electromagnetic (largely capacitive) coupling. With this scheme it is possible to keep the feed line closer to the ground plane compared with the direct feed, in order to minimize feed line radiation. However, the fabrication is more difficult, requiring two substrate layers. Another variation of this technique is to have the microstrip line on the same layer as the patch, with a capacitive gap between the line and the patch edge. This allows for an input match to be achieved without the use of a notch.

9.1.4. Resonance Frequency

The resonance frequency for the (1, 0) mode is given by

$$f_0 = \frac{c}{2L_e\sqrt{\varepsilon_r}} \tag{9.3}$$

where c is the speed of light in vacuum. To account for the fringing of the cavity fields at the edges of the patch, the length, the effective length L_e is chosen as

$$L_e = L + 2\Delta L \tag{9.4}$$

The Hammerstad formula for the fringing extension is [1]

$$\Delta L/h = 0.412\left[\frac{(\varepsilon_{\mathrm{eff}} + 0.3)((W/h) + 0.264)}{(\varepsilon_{\mathrm{eff}} - 0.258)((W/h) + 0.8)}\right] \tag{9.5}$$

where

$$\varepsilon_{\mathrm{eff}} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2}\left(1 + 10\frac{h}{W}\right)^{-1/2} \tag{9.6}$$

9.1.5. Radiation Patterns

The radiation field of the microstrip antenna may be determined using either an "electric current model" or a "magnetic current model." In the electric current model, the current

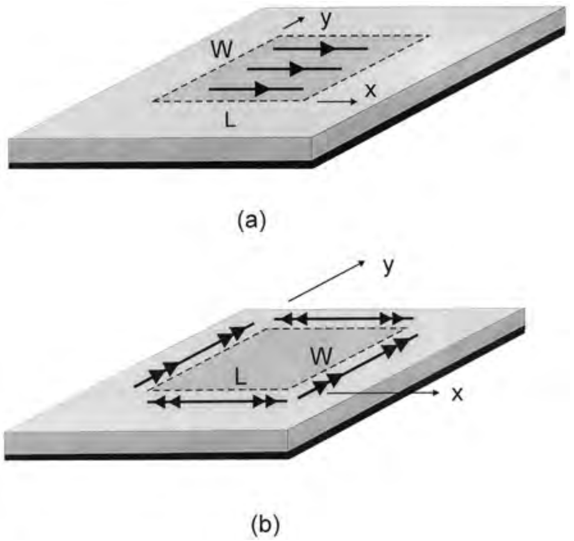


Figure 9.3 Models that are used to calculate the radiation from a microstrip antenna (shown for a rectangular patch): (a) electric current model and (b) magnetic current model.

in Eq. (9.2) is used directly to find the far-field radiation pattern. Figure 9.3a shows the electric current for the $(1, 0)$ patch mode. If the substrate is neglected (replaced by air) for the calculation of the radiation pattern, the pattern may be found directly from image theory. If the substrate is accounted for, and is assumed infinite, the reciprocity method may be used to determine the far-field pattern [5].

In the magnetic current model, the equivalence principle is used to replace the patch by a magnetic surface current that flows on the perimeter of the patch. The magnetic surface current is given by

$$\mathbf{M}_s = -\hat{\mathbf{n}} \times \mathbf{E} \tag{9.7}$$

where $\mathbf{E}$ is the electric field of the cavity mode at the edge of the patch and $\hat{\mathbf{n}}$ is the outward pointing unit-normal vector at the patch boundary. Figure 9.3b shows the magnetic current for the $(1, 0)$ patch mode. The far-field pattern may once again be determined by image theory or reciprocity, depending on whether the substrate is neglected [5]. The dominant part of the radiation field comes from the "radiating edges" at $x = 0$ and $x = L$. The two nonradiating edges do not affect the pattern in the principal planes (the E plane at $\phi = 0$ and the H plane at $\phi = \pi/2$), and have a small effect for other planes.

It can be shown that the electric and magnetic current models yield exactly the same result for the far-field pattern, provided the pattern of each current is calculated in the presence of the substrate at the resonant frequency of the patch cavity mode [5]. If the substrate is neglected, the agreement is only approximate, with the largest difference being near the horizon.

According to the electric current model, accounting for the infinite substrate, the far-field pattern is given by [5]

$$E_i(r,\theta,\phi) = E_i^h(r,\theta,\phi)\left(\frac{\pi WL}{2}\right)\left[\frac{\sin\left(k_y W/2\right)}{k_y W/2}\right]\left[\frac{\cos(k_x L/2)}{(\pi/2)^2-(k_x L/2)^2}\right] \tag{9.8}$$

where

$$k_x = k_0 \sin\theta\cos\phi \tag{9.9}$$

$$k_y = k_0 \sin\theta\sin\phi \tag{9.10}$$

and E_i^h is the far-field pattern of an infinitesimal (Hertzian) unit-amplitude x-directed electric dipole at the center of the patch. This pattern is given by [5]

$$E_\theta^h(r,\theta,\phi) = E_0 \cos\phi\, G(\theta) \tag{9.11}$$

$$E_\phi^h(r,\theta,\phi) = -E_0 \sin\phi F(\theta) \tag{9.12}$$

where

$$E_0 = \left(\frac{-j\omega\mu_0}{4\pi r}\right)e^{-jk_0 r} \tag{9.13}$$

$$F(\theta) = \frac{2\tan(k_0 hN(\theta))}{\tan(k_0 hN(\theta)) - j(N(\theta)/\mu_r)\sec\theta} \tag{9.14}$$

$$G(\theta) = \frac{2\tan(k_0 hN(\theta))cos\theta}{\tan(k_0 hN(\theta)) - j[\varepsilon_r/N(\theta)]\cos\theta} \tag{9.15}$$

and

$$N(\theta) = \sqrt{n_1^2 - \sin^2(\theta)} \tag{9.16}$$

$$n_1 = \sqrt{\varepsilon_r\mu_r} \tag{9.17}$$

The radiation patterns (E and H plane) for a rectangular patch antenna on an infinite nonmagnetic substrate of permittivity $\varepsilon_r = 2.2$ and thickness $h/\lambda_0 = 0.02$ are shown in Fig. 9.4. The patch is resonant with $W/L = 1.5$. Note that the E-plane pattern is broader than the H-plane pattern. The directivity is approximately 6 dB.

9.1.6. Radiation Efficiency

The radiation efficiency of the patch antenna is affected not only by conductor and dielectric losses, but also by surface-wave excitation—since the dominant TM_0 mode of the grounded substrate will be excited by the patch. As the substrate thickness decreases, the effect of the conductor and dielectric losses becomes more severe, limiting the efficiency. On the other hand, as the substrate thickness increases, the surface-wave power

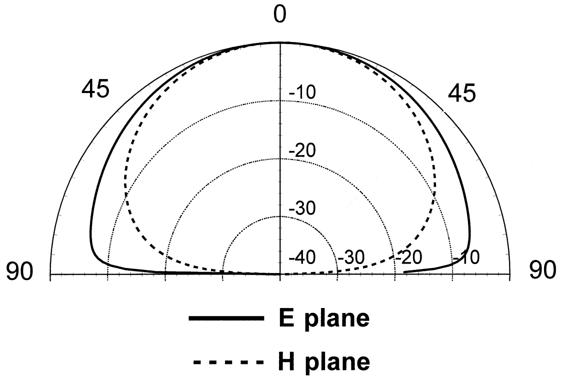


Figure 9.4 The radiation patterns for a rectangular patch antenna on an infinite substrate of permittivity $\varepsilon_r = 2.2$ and thickness $h/\lambda_0 = 0.02$. The patch is resonant with $W/L = 1.5$. The E-plane (xz plane) and H-plane (yz plane) patterns are shown.

increases, thus limiting the efficiency. Surface-wave excitation is undesirable for other reasons as well, since surface waves contribute to mutual coupling between elements in an array and also cause undesirable edge diffraction at the edges of the ground plane or substrate, which often contributes to distortions in the pattern and to back radiation. For an air (or foam) substrate there is no surface-wave excitation. In this case, higher efficiency is obtained by making the substrate thicker, to minimize conductor and dielectric losses (making the substrate too thick may lead to difficulty in matching, however, as discussed above). For a substrate with a moderate relative permittivity such as $\varepsilon_r = 2.2$, the efficiency will be maximum when the substrate thickness is approximately $0.02\lambda_0$.

The radiation efficiency is defined as

$$e_r = \frac{P_{\text{sp}}}{P_{\text{total}}} = \frac{P_{\text{sp}}}{P_{\text{sp}} + P_{\text{c}} + P_{\text{d}} + P_{\text{sw}}} \tag{9.18}$$

where P_{sp} is the power radiated into space, and the total input power P_{total} is given as the sum of P_{c} is the power dissipated by conductor loss, P_{d} is the power dissipated by dielectric loss, and P_{sw} is the surface-wave power. The efficiency may also be expressed in terms of the corresponding Q factors as

$$e_r = \frac{Q_{\text{total}}}{Q_{\text{sp}}} \tag{9.19}$$

where

$$\frac{1}{Q_{\text{total}}} = \frac{1}{Q_{\text{sp}}} + \frac{1}{Q_{\text{sw}}} + \frac{1}{Q_{\text{d}}} + \frac{1}{Q_{\text{c}}} \tag{9.20}$$

The dielectric and conductor Q factors are given by

$$Q_{\rm d} = \frac{1}{\tan\delta} \tag{9.21}$$

$$Q_{\rm c} = \frac{1}{2}\eta_0\mu_r\frac{k_0 h}{R_s} \tag{9.22}$$

where $\tan\delta$ is the loss tangent of the substrate and R_s is the surface resistance of the patch and ground plane metal (assumed equal) at radian frequency $\omega = 2\pi f$, given by

$$R_s = \sqrt{\frac{\omega\mu_0}{2\sigma}} \tag{9.23}$$

where σ is the conductivity of the metal.

The space-wave Q factor is given approximately as [6]

$$Q_{\rm sp} = \frac{3}{16}\left(\frac{\varepsilon_r}{pc_1}\right)\left(\frac{L}{W}\right)\left(\frac{1}{h/\lambda_0}\right) \tag{9.24}$$

where

$$c_1 = 1 - \frac{1}{n_1^2} + \frac{2/5}{n_1^4} \tag{9.25}$$

and

$$p = 1 + \frac{a_2}{10}(k_0W)^2 + \left(a_2^2 + 2a_4\right)\left(\frac{3}{560}\right)(k_0W)^4 + c_2\left(\frac{1}{5}\right)(k_0L)^2 + a_2c_2\left(\frac{1}{70}\right)(k_0W)^2(k_0L)^2 \tag{9.26}$$

with $a_2 = -0.16605$, $a_4 = 0.00761$, and $c_2 = -0.0914153$.

The surface-wave Q factor is related to the space-wave Q factor as

$$Q_{\rm sw} = Q_{\rm sp}\left(\frac{e_r^{\rm sw}}{1 - e_r^{\rm sw}}\right) \tag{9.27}$$

where $e_r^{\rm sw}$ is the radiation efficiency accounting only for surface-wave loss. This efficiency may be accurately approximated by using the radiation efficiency of an infinitesimal dipole on the substrate layer [6], giving

$$e_r^{\rm sw} = \frac{1}{1 + (k_0h)(3/4)(\pi\mu_r)(1/c_1)\left(1 - 1/n_1^2\right)^3} \tag{9.28}$$

A plot of radiation efficiency for a resonant rectangular patch antenna with $W/L = 1.5$ on a nonmagnetic substrate of relative permittivity $\varepsilon_r = 2.2$ or $\varepsilon_r = 10.8$ is shown in Fig. 9.5. The conductivity of the copper patch and ground plane is assumed to be $\sigma = 3.0 \times 10^7$ [S/m]

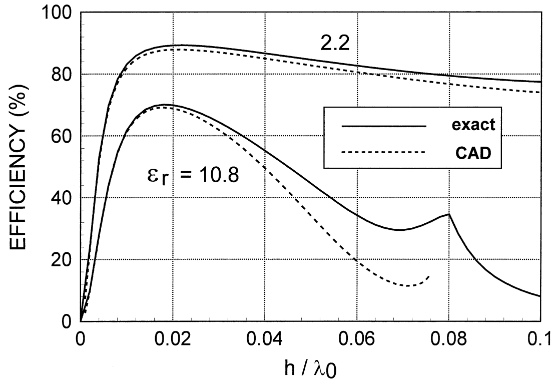


Figure 9.5 Radiation efficiency (%) for a rectangular patch antenna versus normalized substrate thickness. The patch is resonant at 5.0 GHz with $W/L = 1.5$ on a substrate of relative permittivity $\varepsilon_r = 2.2$ or $\varepsilon_r = 10.8$. The conductivity of the copper patch and ground plane is $\sigma = 3.0 \times 10^7$ S/m and the dielectric loss tangent is $\tan\delta = 0.001$. The exact efficiency is compared with the result of the CAD formula [Eq. (9.19) with Eqs. (9.20)–(9.28))].

and the dielectric loss tangent is taken as $\tan\delta = 0.001$. The resonance frequency is 5.0 GHz. [The result is plotted versus normalized (electrical) thickness of the substrate, which does not involve frequency. However, a specified frequency is necessary to determine conductor loss.] For $h/\lambda_0 < 0.02$, the conductor and dielectric losses dominate, while for $h/\lambda_0 < 0.02$, the surface-wave losses dominate. (If there were no conductor or dielectric losses, the efficiency would approach 100% as the substrate thickness approaches zero.)

9.1.7. Bandwidth

The bandwidth increases as the substrate thickness increases (the bandwidth is directly proportional to h if conductor, dielectric, and surface-wave losses are ignored). However, increasing the substrate thickness lowers the Q of the cavity, which increases spurious radiation from the feed, as well as from higher-order modes in the patch cavity. Also, the patch typically becomes difficult to match as the substrate thickness increases beyond a certain point (typically about $0.05\lambda_0$). This is especially true when feeding with a coaxial probe, since a thicker substrate results in a larger probe inductance appearing in series with the patch impedance. However, in recent years considerable effort has been spent to improve the bandwidth of the microstrip antenna, in part by using alternative feeding schemes. The aperture-coupled feed of Fig. 9.2c is one scheme that overcomes the problem of probe inductance, at the cost of increased complexity [7].

Lowering the substrate permittivity also increases the bandwidth of the patch antenna. However, this has the disadvantage of making the patch larger. Also, because the Q of the patch cavity is lowered, there will usually be increased radiation from higher-order modes, degrading the polarization purity of the radiation.

By using a combination of aperture-coupled feeding and a low-permittivity foam substrate, bandwidths exceeding 25% have been obtained. The use of stacked patches (a parasitic patch located above the primary driven patch) can also be used to

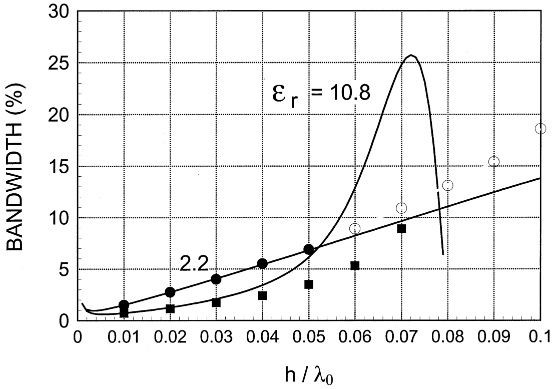


Figure 9.6 Bandwidth (%) for a rectangular patch antenna versus normalized substrate thickness. The parameters are the same as in Fig. 9.5. The exact bandwidth (SWR < 2.0) from a cavity model calculation is compared with the result of the CAD formula [Eq. (9.29)]. The exact calculation assumes a feed location $x_0 = L/4$, $y_0 = W/2$. The exact result is shown with a solid line, and the CAD results are shown with the discrete data points. For the low-permittivity substrate, the hollow dots indicate that the reactance does not go to zero at any frequency. For these cases the resonance frequency is defined as the frequency that minimizes the reactance, and the corresponding minimum reactance value is subtracted from the impedance at each frequency in order to define the SWR bandwidth.

increase bandwidth even further, by increasing the effective height of the structure and by creating a double-tuned resonance effect [8].

A CAD formula for the bandwidth (defined by SWR < 2.0) is

$$BW = \frac{1}{\sqrt{2}}\left(\tan\delta + \frac{R_s}{\pi\eta_0\mu_r}\frac{1}{h/\lambda_0} + \frac{16}{3}\frac{pc_1}{\varepsilon_r}\frac{h}{\lambda_0}\frac{W}{L}\frac{1}{e_r^{\mathrm{sw}}}\right) \tag{9.29}$$

where the terms have been defined in the previous section on radiation efficiency. The result should be multiplied by 100 to get percent bandwidth. Note that neglecting conductor and dielectric loss yields a bandwidth that is proportional to the substrate thickness h.

Figure 9.6 shows calculated and measured bandwidth for the same patch in Fig. 9.5. It is seen that bandwidth is improved by using a lower substrate permittivity and by making the substrate thicker.

9.1.8. Input Impedance

Several approximate models have been proposed for the calculation of input impedance for a probe-fed patch. These include the transmission line method [9], the cavity model [10], and the spectral-domain method [11]. These models usually work well for thin substrates, typically giving reliable results for $h/\lambda_0 < 0.02$. Commercial simulation tools using FDTD, FEM, or MoM can be used to accurately predict the input impedance for any substrate thickness. The cavity model has the advantage of allowing for a simple physical CAD model of the patch to be developed, as shown in Fig. 9.7. In this model the patch cavity is modeled as a parallel RLC circuit, while the probe inductance is

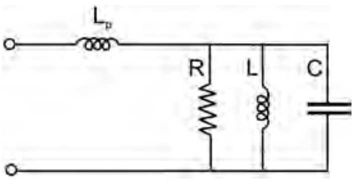


Figure 9.7 CAD model for the input impedance of a coaxial probe-fed microstrip antenna, operating near the resonance frequency.

modeled as a series inductor. The input impedance of this circuit is approximately described by

$$Z_{\text{in}} \approx jX_f + \frac{R}{1 + j2Q(f/f_0 - 1)} \tag{9.30}$$

where f_0 is the resonance frequency, R is the input resistance at the resonance of the RLC circuit (where the input resistance of the patch is maximum), $Q = Q_{\text{total}}$ is the quality factor of the patch cavity (9.20), and $X_f = \omega L_p$ is the feed (probe) reactance of the coaxial probe. A CAD formula for the input resistance R is

$$R = R_{\text{edge}} \cos^2\left(\frac{\pi x_0}{L}\right) \tag{9.31}$$

where the input resistance at the edge is

$$R_{\text{edge}} = \frac{(4/\pi)(\mu_r \eta_0)(L/W)(h/\lambda_0)}{\tan\delta + (R_s/(\pi\eta_0\mu_r))(1/(h/\lambda_0)) + (16/3)(pc_1/\varepsilon_r)(W/L)(h/\lambda_0)\left(1/e_r^{hed}\right)} \tag{9.32}$$

A CAD formula for the feed reactance due to the probe is

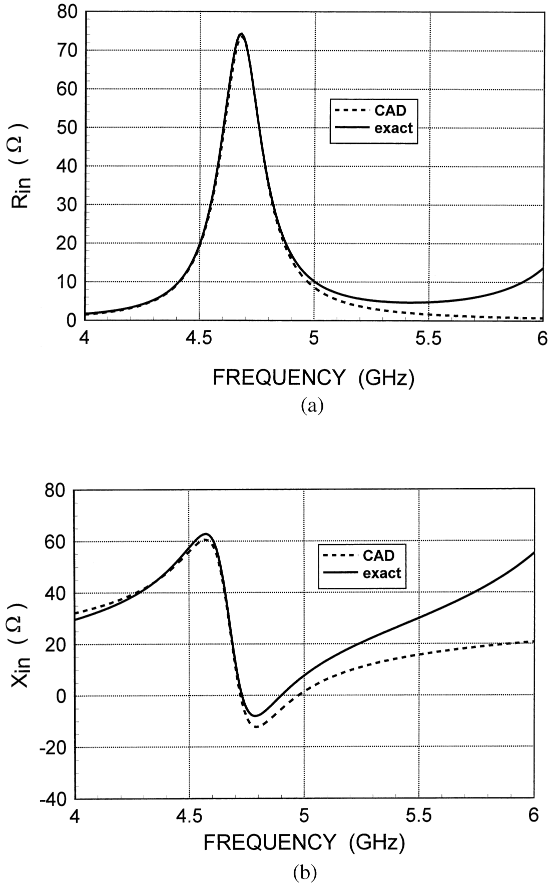
$$X_f = \frac{\eta_0}{2\pi}\mu_r(k_0 h)\left[-\gamma + \ln\frac{2}{\sqrt{\varepsilon_r\mu_r}(k_0 a)}\right] \tag{9.33}$$

where $\gamma = 0.577216$ is Euler's constant.

Figure 9.8 shows a comparison of the input impedance obtained from the simple CAD model Eq. (9.30) with that obtained by a more accurate cavity model analysis. At the resonance frequency, the substrate thickness is approximately $0.024\lambda_0$. Near the resonance frequency, the simple CAD model gives results that agree quite well with the cavity model.

9.1.9. Improving Performance

Much research has been devoted to improving the performance characteristics of the microstrip antenna. To improve bandwidth, the use of thick low-permittivity (e.g., foam) substrates can give significant improvement. To overcome the probe inductance associated with thicker substrates, the use of capacitive-coupled feeds such as the top-loaded probe [12] or the L-shaped probe [13] shown in Fig. 9.9a and b may be used. Alternatively, the aperture-coupled fed shown in Fig. 9.2c may be used, which also has the advantage of



(a)

(b)

Figure 9.8 Input impedance versus frequency for a rectangular coaxial probe-fed patch antenna. The results from the CAD model in Fig. 9.7 are compared with those obtained by a cavity-model analysis: (a) input resistance and (b) input reactance. $L = 2.0\,\text{cm}$ and $W/L = 1.5$. The feed probe is located at $x_0 = L/4$, $y_0 = W/2$ and has a radius of 0.05 cm. The substrate has a permittivity of $\varepsilon_r = 2.2$ and a thickness of 0.1524 cm.

eliminating spurious probe radiation. To increase the bandwidth even further, a stacked patch arrangement may be used, in which a parasitic patch is stacked above the driven patch [8]. This may be done using either a probe feed or, to obtain even higher bandwidths, an aperture-coupled feed (Fig. 9.9c). The bandwidth enhancement is largely due to the existence of a double resonance and, to some extent, to the fact that one of the radiators is further from the ground plane. Bandwidths as large as one octave (2:1 frequency band) have been obtained with such an arrangement. By using a diplexer feed to split the feeding signal into two separate branches, and feeding two aperture-coupled stacked patches with different center frequencies, bandwidths of 4:1 have been obtained [14]. Parasitic patches may also be placed on the same substrate as the driven patch, surrounding the driven patch. A pair of parasitic patches may be coupled to the radiating edges, the nonradiating edges, or all four edges [15]. This planar arrangement saves vertical height and allows for easier fabrication, although the substrate area occupied by the antenna to be larger, and there may be more variation of the radiation pattern across

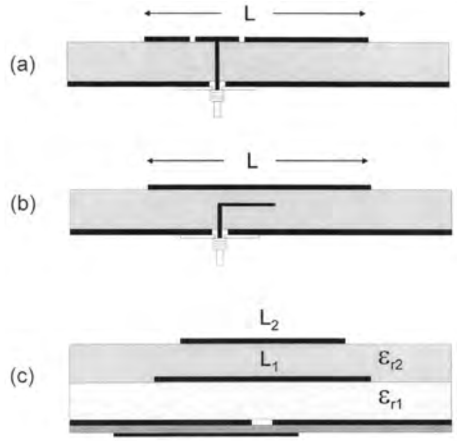


Figure 9.9 Some schemes for improving bandwidth: (a) probe with a capacitive top loading, (b) L-shaped probe, (c) stacked patches, and (d) aperture-coupled stacked patches.

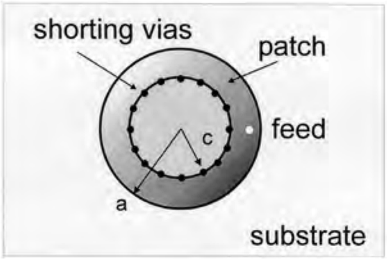


Figure 9.10 The "reduced-surface-wave" microstrip antenna. This antenna excites less surface-wave and lateral radiation than does a conventional microstrip antenna. The antenna consists of a circular patch of radius *a* that has a short-circuit boundary (array of vias) at an inner radius *c*.

the frequency band since the current distribution on the different patches changes with frequency. Broadbanding may also be achieved through the use of slots cut into the patch, as in the U-slot patch design [16]. This has the advantage of not requiring multiple layers or increasing the size of the patch as with parasitic elements.

Another variation of the microstrip antenna that has been introduced recently is the "reduced surface wave" microstrip antenna shown in Fig. 9.10 [17]. This design is a variation of a circular patch, with an inner ring of vias that create a short-circuit inner boundary. By properly selecting the outer radius, the patch excites very little surface-wave field and also only a small amount of lateral (horizontally propagating) radiation. The inner short-circuit boundary is used to adjust the dimensions of the patch cavity (between the inner and outer boundaries) to make the patch resonant. The reduced surface-wave and lateral radiation result in less edge diffraction from the edges of the supporting ground plane, giving smoother patterns in the front-side region and less radiation in the back-side region. Also, there is less mutual coupling between pairs of such antennas, especially as the separation increases. The disadvantage of this antenna is that it is physically fairly large, being about $0.60\lambda_0$ in diameter, regardless of the substrate permittivity.

9.2. BROADBAND ANTENNAS

9.2.1. Introduction

Until relatively recently, broadband antennas (for the purpose of this discussion, broadband suggests bandwidths of approximately an octave, 2:1, or more) have been predominately employed in radar and tracking applications and in specialized broadband communications systems. However, with the move to digital modulation and spread spectrum coding schemes over multiple frequency bands in modern communication systems, the need for broadband antennas has increased rapidly. There are many ways to achieve wideband antenna performance. Typically, however, antennas that provide broadband coverage fall into one of two categories: multiband elements and arrays that simultaneously cover multiple "spot" (narrow) bands, and naturally broadband (quasi-frequency independent) radiators. The focus of this discussion will be on the latter. In addition, the antenna designs considered will be primarily for RF and microwave applications; however, many of the designs can be used at lower and higher frequencies. The discussion will be limited to outlining the general properties and operation of the most common broadband antenna elements; helical, spiral, and log-periodic antennas.

9.2.2. Helical Antenna

Helical antennas, or helixes, are relatively simple structures with one, two, or more wires each wound to form a helix, usually backed by a ground plane or shaped reflector and driven with an appropriate feed [18–20]. The most common design is a single wire (monofilar helix), backed by a ground and fed with a coaxial line, as shown in Fig. 9.11. For this typical helix geometry, L is the axial length, D is the diameter, S is the inner winding spacing, C is the circumference, α is the pitch angle (defined as the angle between a tangent line to the helix wire and the plane perpendicular to the axis of the helix), and a is the radius of the helix wire. The helix has N turns. In general, the radiation properties of the helical antenna are associated with the electrical size of the structure, where the input impedance is more sensitive to the pitch and wire size. Helical antennas have two predominant radiation modes, the normal (broadside) mode and the axial (end-fire) mode. The normal mode occurs when C is small compared to a wavelength and the axial mode occurs when C is on the order of a wavelength. For most applications, the axial mode is used. Hence, the following discussion will focus on the end-fire mode of operation for a helical antenna.

The radiation pattern for the axial (end-fire) mode is characterized by a major lobe along the axial direction. The polarization along this direction is elliptical, and when appropriately designed ($3/4\lambda_0 < C < 4/3\lambda_0$, $S \approx \lambda_0/4$), good circular polarization (CP) can be obtained. The handedness of the radiation is determined by twist of the helix. If wound as a right-handed (RH) screw, the polarization of the radiated field is RH. If wound as a left-handed (LH) screw, the polarization of the radiated field is LH. The helix shown in Fig. 9.11 is LH. The helix is characterized by an approximately real input impedance over a slightly less than 2:1 bandwidth. The value of this impedance ranges between 100 and 200 Ω. For axial mode helix antennas with $C/\lambda_0 \approx 1$ [18],

$$R_{IN} \approx 140\frac{C}{\lambda_0} \quad [\Omega] \tag{9.34}$$

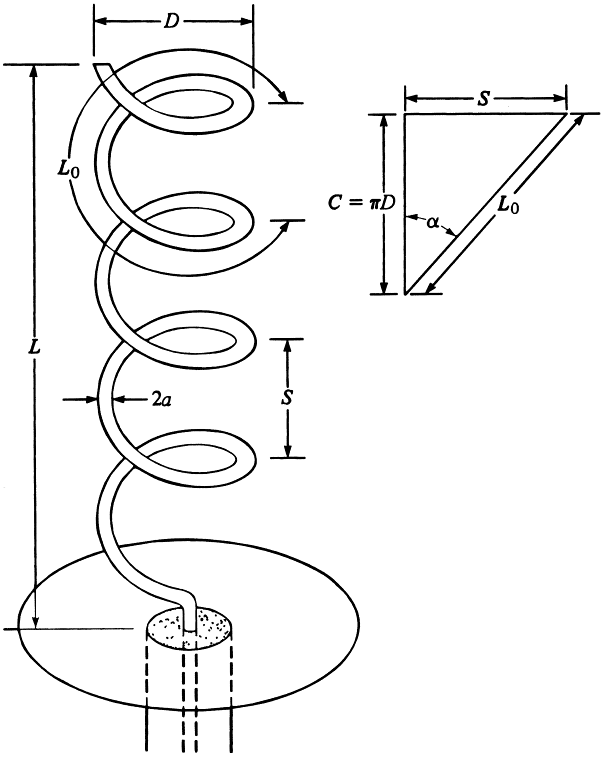


Figure 9.11 Monofilar helical antenna with ground plane [38].

Since most of the feeding coaxial lines have a characteristic impedance significantly less than R_{IN} (typically, 50 Ω), helical antennas are not usually fed directly by the coax as depicted in Fig. 9.11. A variety of techniques have been developed to match the higher impedance helix with the feed coax, including varying the pitch and the diameter of the helix wire at the feed to essentially form a tapered matching section [20]. However, the most common matching technique is to move the coax feed off the axis of the helix and insert a microstrip matching transformer between the coax feed and the beginning of the helix.

The radiation pattern for an axial-mode helix is approximated by treating the helix as a linear-end-fire array of one-wavelength circumference loop antennas with spacing S [21]. The progressive phase shift between the elements corresponds to the shift associated with a traveling wave along the helical wire. The interelement phase shift (δ) is given by

$$\delta \approx -\frac{\omega \ell}{v_p} = -\frac{k_0 L_0}{v_p/c} \tag{9.35}$$

where k_0 is the free-space wavenumber, L_0 is the length along one turn of the helix, c is the speed of light in free space, and v_p phase velocity of the traveling wave current along the helix. For an axial mode helix,

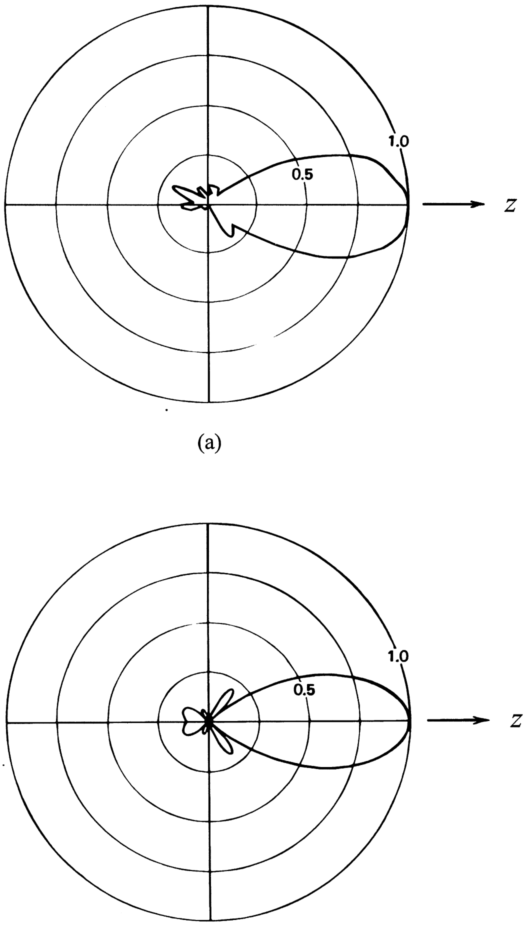
$$\frac{v_p}{c} \approx \frac{L_0/\lambda_0}{S/\lambda_0 + (2N+1)/2N} \tag{9.36}$$

The traveling wave current is a slow wave with respect to c. The current along the helical wire decays away from the feed due to radiation; however, a simplifying assumption used to approximate the radiation pattern is to assume the amplitude of the current on each loop is the same. Hence, the circularly polarized electric field radiated by an axial-mode helix is approximated as

$$E(\theta) \approx A\cos\theta \frac{\sin(N\psi/2)}{N\sin(\psi/2)} \tag{9.37}$$

where $\psi = k_0 S\cos\theta + \delta$. An example of the pattern for an axial-mode helical antenna is shown in Fig. 9.12. The directivity of a helix is approximated by [18]

$$D \approx 15\left(\frac{C}{\lambda_0}\right)^2 \frac{L}{\lambda_0} = 15\left(\frac{C}{\lambda_0}\right)^2 \frac{NS}{\lambda_0} \tag{9.38}$$



(a)

(b)

Figure 9.12 Radiation pattern for an axial-mode helical antenna ($N = 10$, $C = \lambda_0, \alpha = 13°$, $f = 8\,\text{GHz}$): (a) measured and (b) theory [25].

To achieve the desired radiation and polarization characteristics the length of the helical antenna needs to be sufficiently large to ensure that the outward propagating traveling wave on the helix is attenuated to the point that the reflected wave at the end of the helix is small. Typically, this is achieved by making $L > \lambda_0/2$. Gradually reducing the radius of the helix turns near the end of the antenna also has the effect of reducing the reflection of the outward traveling wave and helps flatten the impedance characteristics of the antenna [22]. Ideally, the polarization of a helical antenna is circular; however, this is only true for an infinite helix. The polarization of a finite helix is actually elliptical, with an axial ratio approximated by [18]

$$\mathrm{AR} = \frac{2N+1}{2N} \tag{9.39}$$

Thus, as the number of turns is increased (the length of antenna increased for a fixed pitch), the polarization approaches circular.

Multifilar helical antennas typically are slightly more directive and have better axial ratios compared to equivalent monofilar helixes. The feeds for these structures, however, are more complicated [19].

9.2.3. Frequency Independent Antennas

For bandwidths much larger than an octave, (quasi-) frequency independent antennas are typically used. The design of frequency independent antennas is based upon the knowledge that the impedance and radiation properties of an antenna are associated with the electrical dimensions of the structure (dimensions expressed in wavelengths). Hence, if an arbitrary scaling of the antenna structure results in the same structure, possibly rotated about the vertex, the electrical properties of the antenna will be independent of frequency [23]. Such antennas can be described solely by angular specifications. Antennas that can be described on conical surfaces and by equiangular spiral curves satisfy this angle requirement. Theoretically, the structure must be infinite in extent and emanate from a point vertex to be truly frequency independent. In practice, the operating bandwidth of these antennas are limited at low frequencies by the outer dimensions of the structure and how the structure is terminated along the outside boundary. The currents on these antennas tend to decay rapidly away from the center of the structure, particularly beyond the active or radiation region (the portion of the antenna near the radius $r = \lambda/2\pi$) [24]. Thus, if the structure is appropriately truncated beyond the radiation region where the currents are relatively low, the performance of the antenna is not adversely affected. The high-frequency limit of these antennas is dictated by the inside dimensions and the precision of the antenna near the vertex. This is commonly the feed region of these antennas.

It is also interesting to note that the input impedance of a self-complementary antenna is frequency independent. A self-complementary planar structure (as defined by Babinet's principle) is one that remains the same, with the exception of a rotation about the vertex, when the metallic and nonmetallic regions on the antenna surface are interchanged. In this case the input impedance for both structures is the same. If the structures are infinite, the input impedance is independent of frequency and, from Babinet's principle, equal to $188.5\,\Omega$ [25]. The input impedance for a truncated self-complementary antenna is typically slightly less than this value but in practice can be made relatively constant over a wide band. While self-complementary antennas do have a frequency independent input impedance, being self-complementary is not a necessary requirement for frequency independent performance. Many frequency independent antenna designs are not self-complementary.

9.2.4. Spiral Antennas

Broadband spiral antennas can be realized on either planar or conical surfaces. Planar spirals are typically bidirectional, radiating near circularly polarized fields on both the front and back sides. To eliminate back-side radiation planar spirals are backed by an absorber-filled cavity (more typical) or a conducting ground plane (less typical). Conical spirals are unidirectional, radiating along the direction of the apex, thereby eliminating the requirement for a backing cavity. The most common spiral designs have two arms; however, to improve pattern symmetry and for direction-finding and tracking systems, multiarm (typically four-arm) spirals are used [26].

The planar equiangular spiral antenna (log-spiral) is defined by the equiangular spiral curve shown in Fig. 9.13 [27]

$$r = \rho = \rho_i e^{a(\phi-\phi_i)} \qquad \text{for } \theta = \frac{\pi}{2} \tag{9.40}$$

where ρ is the radial distance from the vertex in the $\theta = \pi/2$ plane, ϕ is the angle from the x axis, ρ_i and ϕ_i are the coordinates to the start of the spiral curve, and a is the flare rate of the curve. As shown in the figure, the flare rate is related to the pitch angle of the spiral (ψ) by

$$\tan\psi = \frac{1}{a} \tag{9.41}$$

Note that beyond the starting point this curve is defined solely by the angle ϕ.

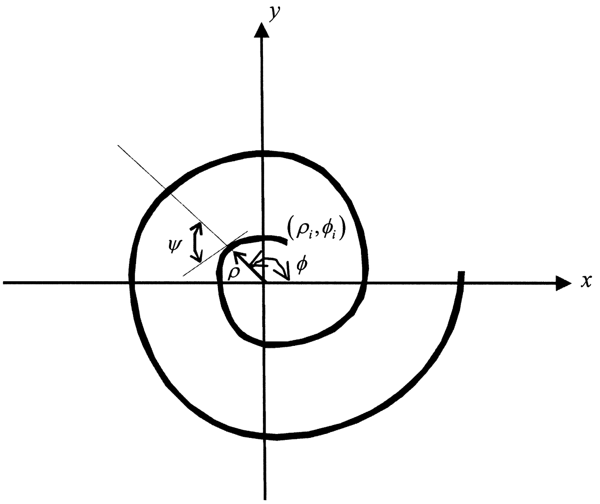


Figure 9.13 Equiangular spiral curve.

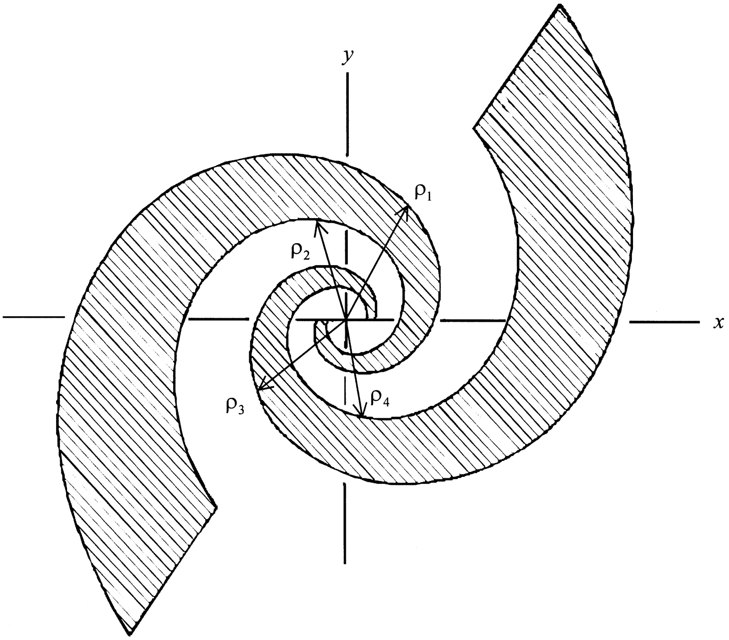


Figure 9.14 Planar two-arm equiangular spiral antenna.

A log-spiral antenna defined using Eq. (9.40) is shown in Fig. 9.14. The first arm, which begins along the x axis in this example, is defined by the edges

$$\begin{aligned} &\text{Outer edge:} \quad && \rho_1 = \rho_0 e^{a\phi} \\ &\text{Inner edge:} \quad && \rho_2 = \rho_0 e^{a(\phi-\delta)} \end{aligned} \tag{9.42}$$

where

$$\frac{\rho_2}{\rho_1} = e^{-a\delta} < 1 \tag{9.43}$$

The second arm is realized by simply rotating the first arm by π radians. The expressions for the edges of this arm are obtained by multiplying the expressions in Eq. (9.42) by $e^{-a\pi}$. This two-arm structure is self-complementary when $\delta = \pi/2$. The scaling ratio for an equiangular spiral,

$$\tau = \frac{\rho(\phi)}{\rho(\phi + 2\pi)} = e^{-2\pi a} \tag{9.44}$$

should typically be between 0.1 and 0.9. Optimum performance for the equiangular spiral is obtained when the number of turns is between 1.5 and 3. A complementary structure, a spiral slot antenna has similar electrical properties and in many cases is easier to physically implement.

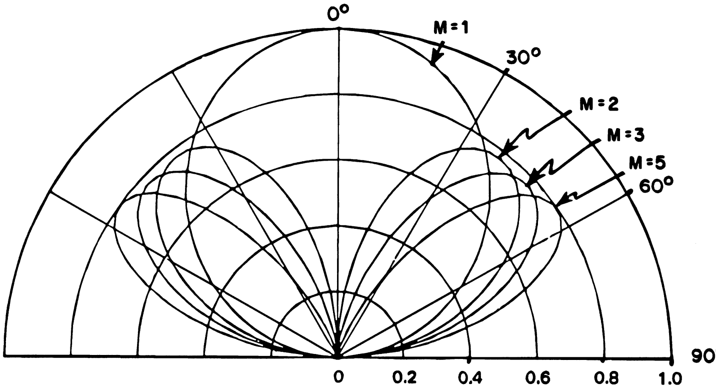


Figure 9.15 Representative radiation pattern for the $M = 1$, 2, 3, and 5 modes of a planar two-arm equiangular spiral antenna.

An N-arm spiral antenna with discrete rotational symmetry characterized by the angle $2\pi/N$ radians supports N normal modes. For the Mth mode ($M \in [1, N]$), the spiral arms are excited with equal amplitudes and a progressive phase shift equal to $e^{-j2\pi M/N}$. Qualitatively, the predominant radiation for the Mth mode is from the active (radiation) region, the region on the spiral near the radius $r = M\lambda/2\pi$. In the active region, the phasing of the currents along the adjacent arms are such that they essentially form an annular ring of traveling-wave current M wavelengths in circumference [24]. All the modes are bidirectional, and all but the $M = 1$ mode have nulls broadside to the antenna. The $M = 1$ mode has single lobes broadside to the front and back sides and is approximately characterized by $\cos\theta$. An example of these patterns is shown in Fig. 9.15. If fed appropriately, the radiation is nearly perfectly circularly polarized (CP) along the axis of the antenna, degrading as the observation angle moves away from broadside. In addition, since physically scaling the equiangular structure is equivalent to a rotation in ϕ, the radiation patterns rotate with frequency. Only at frequencies scaled by τ are the scaled structures congruent and the patterns identical, assuming the arms are appropriately fed and terminated.

Spiral antennas are inherently balanced structures, thereby requiring a balanced feed. Since the antenna is typically fed with a 50 Ω coax and the input impedance for the spiral is on the order of 150–180 Ω, broadband impedance transformers and baluns are required. While beyond the scope of this discussion, the balun design is critical to exciting purely the desired mode on the spiral at all frequencies of operation [26]. If additional modes are excited, the result is typically a degradation in the impedance and CP performance, along with squint in the radiation pattern. In addition, if the currents are not sufficiently attenuated through the active region of the antenna, the residual energy they carry must be either radiated or dissipated to prevent reflections from the end of the arms and the consequent opposite-handed reradiation. Tapering of the ends of the arms and use of resistive–absorbing loads along the outer portion of the antenna are techniques used to damp these residual currents.

Although not strictly frequency independent because the structure is not defined solely by angles, the archimedean spiral design is widely used. Archimedean spirals have bandwidths over 10 : 1 and excellent polarization and pattern characteristics [26].

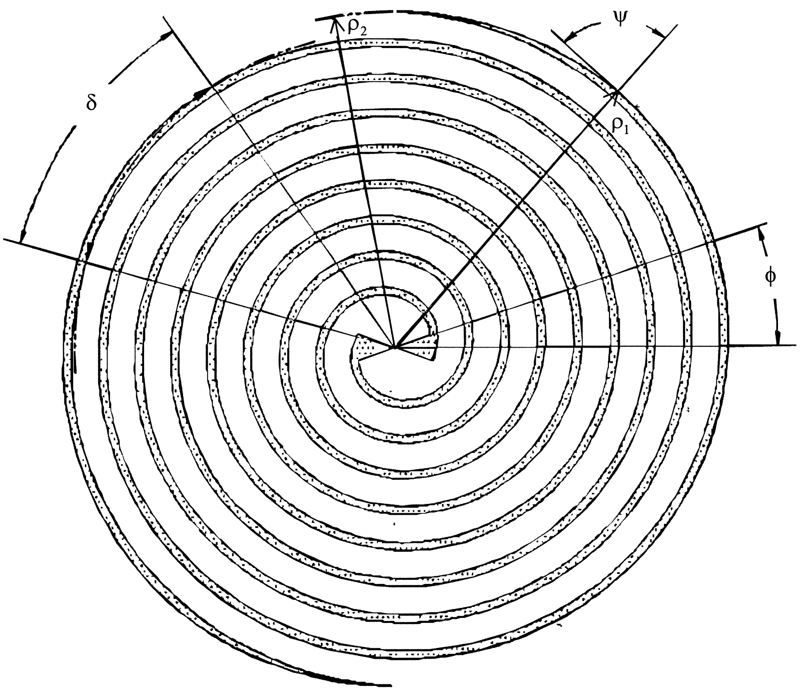


Figure 9.16 Planar two-arm Archimedean spiral antenna.

An example of an Archimedean spiral is shown in Fig. 9.16. The centerline of the antenna is defined by

$$\rho = a\phi \tag{9.45}$$

The first arm, which begins along the x-axis in this example, is defined by the edges

$$\begin{aligned} &\text{Outer edge:} \qquad \rho_1 = a\left(\phi + \frac{\delta}{2}\right) \\ &\text{Inner edge:} \qquad \rho_2 = a\left(\phi - \frac{\delta}{2}\right) \end{aligned} \tag{9.46}$$

where the pitch of the spiral arms is given by

$$\tan\psi = \frac{\rho}{a} \tag{9.47}$$

Again, additional arms are realized by successively rotating this arm by the angle $2\pi/N$. The arm width (W) and centerline spacing between adjacent arms (S) are given by

$$\begin{aligned} W &= a\delta \sin\psi \\ S &= 2\pi a \sin\psi \end{aligned} \tag{9.48}$$

For a two-arm Archimedean, if $W/S = 1/4$, the structure is self-complementary. This is equivalent to $\delta = \pi/2$. The radiation properties of the Archimedean spiral are very similar to the log spiral.

As mentioned earlier, planar spiral antennas are bidirectional. To eliminate the back-side radiation, the antenna is usually backed by an absorber-filled cavity [27]. If appropriately designed, the cavity has little affect on the front-side pattern and impedance properties of the antenna, but it does reduce the efficiency of the structure to less than 50%. In addition, these cavities are rather deep. For low profile applications and to improve the efficiency of the antenna, printed spirals—spiral antennas printed on a relatively thin conductor-backed dielectric substrate—have been developed [28]. These printed spirals are typically Archimedean and are much narrower band than their cavity-backed counterparts, with typical bandwidths on the order of 2:1. The arms of these antennas essentially form microstrip lines; hence, the currents are more tightly bound to the structure than in other spiral designs. As a result, care must be taken to appropriately load the outer portions of the arms of the printed spirals in order to attenuate residual currents that propagate through the active region of the antenna [29]. In addition, the antenna cannot be made too large because of perturbing radiation from higher order modes, i.e., modes with active regions of circumference $(M + pN)\lambda$, where $p = 1, 2, 3, \ldots$ These higher order modes are excited by the residual currents. This size requirement also places limits on the operating bandwidth of the antenna.

Another technique used to realize unidirectional patterns is to place the equiangular spiral on a conical form, as shown in Fig. 9.17 [30]. The result, a conical equiangular spiral antenna, has a broad, single-lobed, circularly polarized pattern along the direction of the apex of the cone. On the conical form, with half angle θ_0, the equiangular spiral curve is defined as

$$r = r_i e^{(a \sin\theta_0)(\phi - \phi_i)} \qquad \text{for } \theta = \theta_0 \tag{9.49}$$

As shown in the figure, the flare rate is related to the pitch angle of the spiral by Eq. (9.41). The first arm is defined by the edges

$$\begin{aligned} &\text{Outer edge:} \qquad r_1 = r_0 e^{(a \sin\theta_0)\phi} \\ &\text{Inner edge:} \qquad r_2 = r_0 e^{(a \sin\theta_0)(\phi - \delta)} \end{aligned} \tag{9.50}$$

As before, a two-arm conical spiral is self-complementary when $\delta = \pi/2$. It is useful to note that the front-to-back ratio of the unidirectional pattern increases with pitch angle and decreases with increasing cone angle (θ_0) [31].

9.2.5. Log-periodic Antennas

If a particular antenna structure has the property that it is equal to itself after being increased in scale by a factor $1/\tau$, then the antenna will have the same electrical properties at the frequencies f and τf. As such,

$$\frac{f_L}{f_H} = \tau \qquad f_L < f_H \tag{9.51}$$

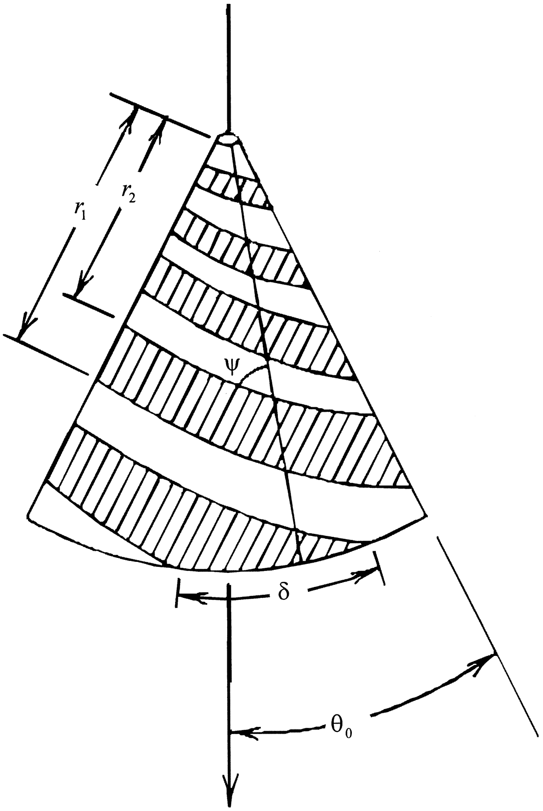


Figure 9.17 Conical two-arm equiangular spiral antenna.

Taking the logarithm of both sides of (9.51),

$$\log f_H = \log f_L + \log \frac{1}{\tau} \tag{9.52}$$

Hence, the electrical properties of the antenna are periodic when plotted on a log-frequency scale, with a period of $\log(1/\tau)$. Antennas that are based upon this principle are called *log-periodic* antennas. If these properties are relatively constant over the range $(f, \tau f)$, then they will be relatively constant for all frequencies, and the antenna will be quasi-frequency independent. Note that the equiangular spiral antenna is technically a log-periodic antenna (thus, its commonly referred to as a *log-spiral*) since the structure is unchanged when scaled by Eq. (9.44). The only variation in the electrical properties for an infinite log-spiral over the range $(f, \tau f)$ is a rotation of the radiation pattern.

The number of log-periodic antenna designs in use is too large to cover in this discussion [26]. Rather the general operating principles of a common structure will be given. Consider the planar trapezoidal-tooth log-periodic antenna shown in Fig. 9.18. The two elements are rotated versions of each other and connected to the feed at their respective vertex. The antenna is balanced; hence, a balanced feed must be used, usually with an impedance matching transformer. The number of teeth on each side of the center

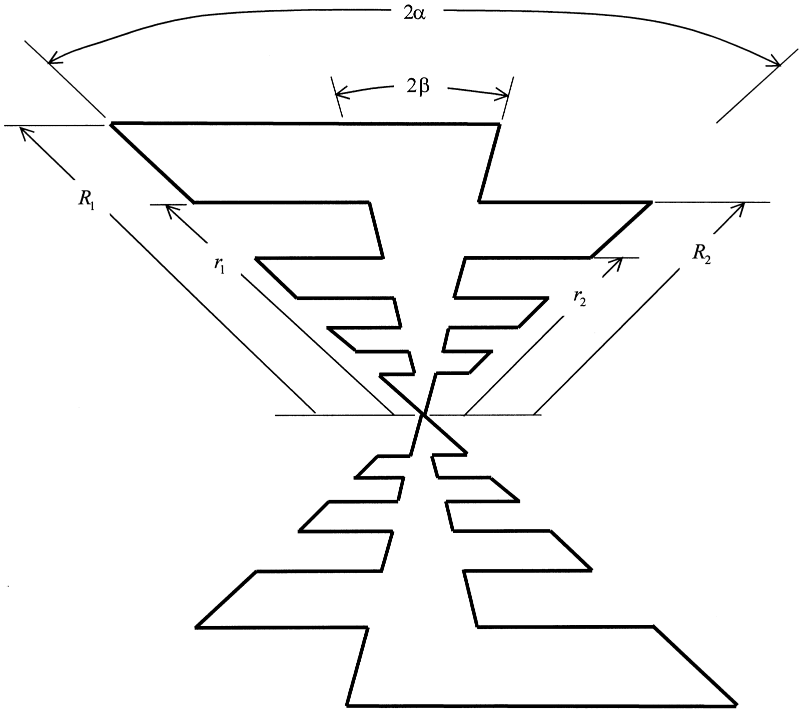


Figure 9.18 Planar trapezoidal-tooth log-periodic antenna.

strip (angular dimension 2β) should be the same and stagger spaced, as shown in the figure. Defining R_n as the distance from the vertex to the outer edge of the nth tooth ($n = 1$ for outer tooth, even n on one side of the strip, odd n on the other) and r_n as the distance from the vertex to the inner edge of the nth tooth, the scaling ratio for the structure is defined as

$$\tau = \frac{R_{n+2}}{R_n} < 1 \tag{9.53}$$

and the width of each tooth is characterized by

$$\varepsilon = \frac{r_n}{R_n} < 1 \tag{9.54}$$

Typically, $r_n = R_{n+1}$. With respect to the logarithm of frequency, the input impedance has a period $\log(1/\tau)$; however, given the staggered teeth and rotational symmetry of the arms, the pattern has a period $2\log(1/\tau)$.

For this structure, the angular center strips act as a transmission line, carrying current from the feed to the effective dipoles formed by the opposing teeth. Most of the radiation from the antenna occurs in the region where the effective dipole length on each arm is approximately $\lambda_0/2$. This is the active region of the antenna. The radiation pattern is bidirectional for the planar structure, essentially that of two-parallel dipoles, and it is polarized along the direction of the teeth. Unidirectional log-periodic antennas can be formed by bending the planar elements at the feed to form a wedge, similar in concept to the conical spiral. The low-frequency limit of the antenna occurs when the longest tooth is approximately $\lambda_0/4$. The high-frequency limit is established by the shortest teeth.

9.3. TRAVELING AND LEAKY-WAVE ANTENNAS

9.3.1. Introduction

Traveling-wave antennas are a class of antennas that use a traveling wave on a guiding structure as the main radiating mechanism [32,33]. They possess the advantage of simplicity, as no complicated feed network is required as in, for example, a typical array antenna.

Traveling-wave antennas fall into two general categories, slow-wave antennas and fast-wave antennas, which are usually referred to as *leaky-wave antennas* [34,35]. In a slow-wave antenna, the guided wave is a slow wave, meaning a wave that propagates with a phase velocity that is less than the speed of light in free space. Such a wave does not fundamentally radiate by its nature, and radiation occurs only at discontinuities (typically the feed and the termination regions). The propagation wave number of the traveling wave is therefore a real number (ignoring conductor or other losses). Because the wave radiates only at discontinuities, the radiation pattern physically arises from two equivalent sources, one at the beginning and one at the end of the structure. This makes it difficult to obtain highly directive single-beam radiation patterns. However, moderately direct patterns having a main beam near end fire can be achieved, although with a significant side-lobe level. For these antennas there is an optimum length depending on the desired location of the main beam.

By contrast, the wave on a leaky-wave antenna is a fast wave, with a phase velocity greater than the speed of light. This type of wave radiates continuously along its length, and hence the propagation wave number is complex [36,37], consisting of both a phase and an attenuation constant. Highly directive beams at an arbitrary specified angle can be achieved with this type of antenna, with a low side-lobe level. The phase constant of the-wave controls the beam angle, while the attenuation constant controls the beam width. The aperture distribution can also be easily tapered to control the side-lobe level or beam shape.

Leaky-wave antennas can be divided into two important categories, uniform and periodic, depending on the type of guiding structure [34]. A uniform structure has a cross section that is uniform (constant) along the length of the structure, usually in the form of a waveguide that has been partially opened to allow radiation to occur. The guided wave on the uniform structure is a fast wave, and thus radiates as it propagates.

A periodic leaky-wave antenna structure is one that consists of a uniform structure that supports a slow (nonradiating) wave that has been periodically modulated in some fashion. Since a slow wave radiates at discontinuities, the periodic modulations (discontinuities) cause the wave to radiate continuously along the length of the structure. From a more sophisticated point of view, the periodic modulation creates a guided wave that consists of an infinite number of space harmonics (Floquet modes). Although the main ($n=0$) space harmonic is a slow wave, one of the space harmonics (usually the $n=-1$) is designed to be a fast wave and hence a radiating wave.

9.3.2. Slow-wave Antennas

A variety of guiding structures can be used to support a slow wave. Examples include wires in free space or over a ground plane, helixes, dielectric slabs or rods, corrugated conductors, etc. Many of the basic principles of slow-wave antennas can be illustrated by considering the case of a long wire antenna, shown in Fig. 9.19. The current on the wire is taken as

$$I(z') = I_0 e^{-j\beta z'} \qquad 0 < z' < L \tag{9.55}$$

where it is assumed that $\beta \geq k_0$. The far-field pattern is equal to the pattern of an infinitesimal unit amplitude electric dipole in the z direction multiplied by an array factor term, which is expressed in terms of the Fourier transform of the wire current. The far-zone electric field is polarized in the θ direction and is given by

$$E_\theta = \left(\frac{j\omega\mu_0}{4\pi r} e^{-jk_0 r} \sin\theta\right) \int_0^L I(z') e^{jk_0 z' \cos\theta} dz' \tag{9.56}$$

Substituting the current expression, Eq. (9.55), and performing the integral yields

$$E_\theta = \left(\frac{j\omega\mu_0}{4\pi r} e^{-jk_0 r} \sin\theta\right) I_0 \left[\frac{e^{jL(k_0\cos\theta - \beta)}}{j(k_0\cos\theta - \beta)} - \frac{1}{j(k_0\cos\theta - \beta)}\right] \tag{9.57}$$

Equation (9.56) indicates that the radiation comes (by superposition) from the entire length of the radiating current. The form of Eq. (9.57), on the other hand, makes it clear that the radiation may also be interpreted as coming from the two ends of the current segment. Both points of view are correct, since Eqs. (9.56) and (9.57) are mathematically equivalent. For calculation purposes, it is convenient to write Eq. (9.57) as

$$E_\theta = \left(\frac{j\omega\mu_0}{4\pi r} e^{-jk_0 r} \sin\theta\right) (I_0 L) e^{-j(L/2)(\beta/k_0 - \cos\theta)} \operatorname{sinc}\left[\frac{k_0 L}{2}\left(\frac{\beta}{k_0} - \cos\theta\right)\right] \tag{9.58}$$

where $\operatorname{sinc} x \equiv (\sin x)/x$.

The $\operatorname{sinc} x$ function is maximum at $x = 0$, which corresponds to an angle θ_0 in "invisible space," since $\cos\theta_0 = \beta/k_0 > 1$. This explains why it is not possible to obtain a narrow single-beam pattern with this type of current. Although the array factor (the sinc term) is maximum at end fire ($\theta = 0$), the presence of the $\sin\theta$ term from the element

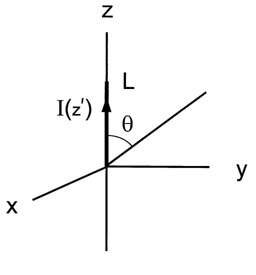
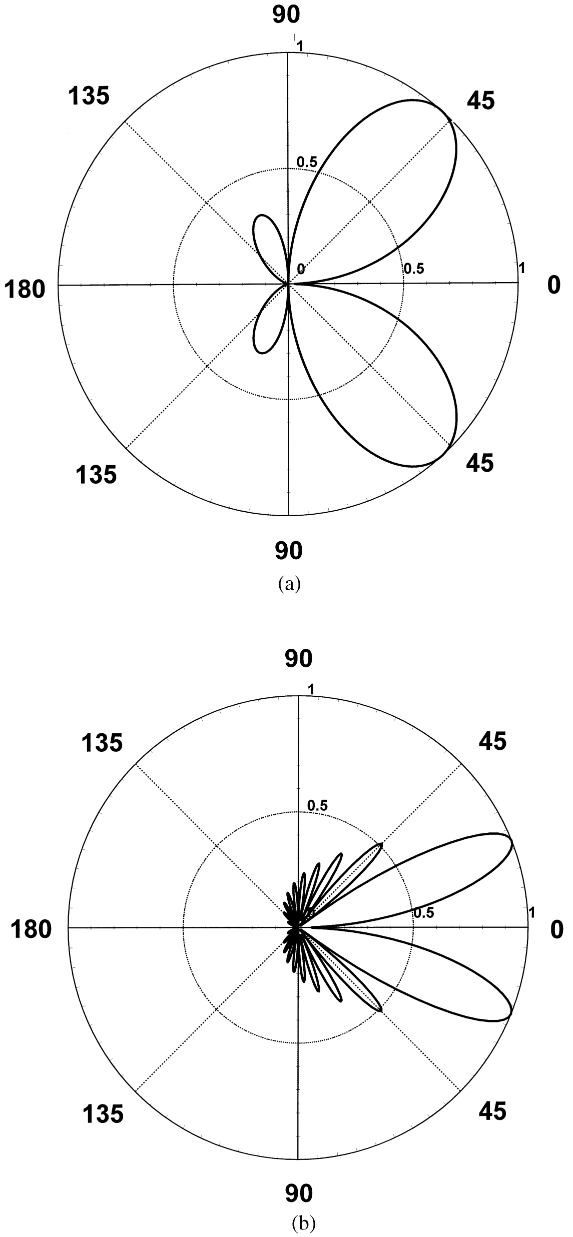


Figure 9.19 A traveling-wave current $I(z')$ on a wire, existing from $z = 0$ to $z = L$.

pattern of the infinitesimal dipole results in a main beam that has a maximum shifted away from end fire.

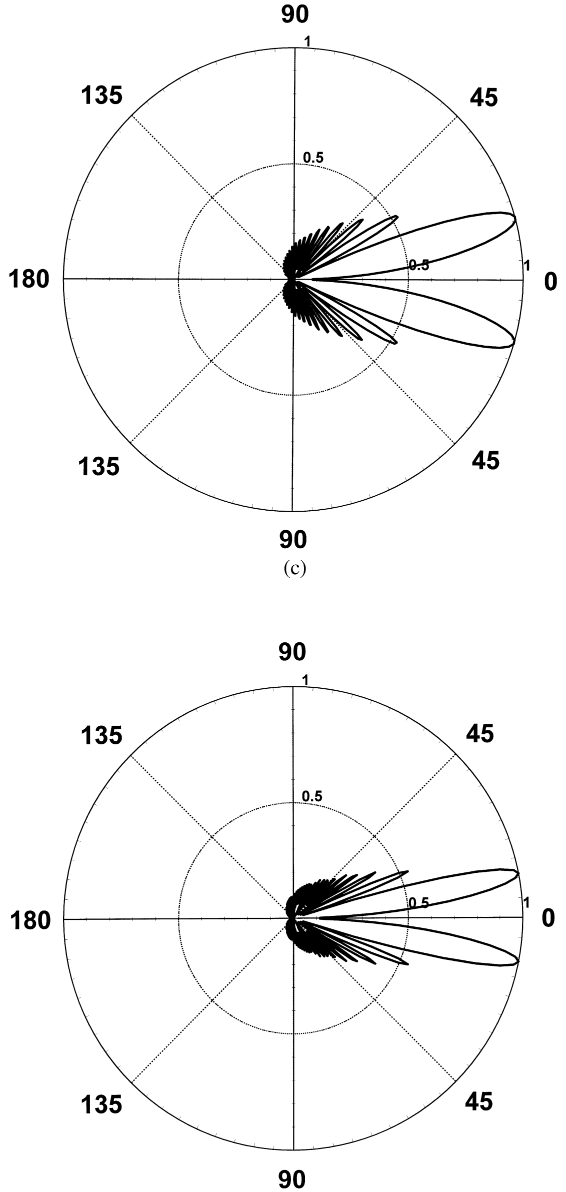
Figure 9.20 shows patterns for the practical case $\beta/k_0 = 1$, for several lengths of current. It is seen that a longer current results in a "main beam" with a maximum closer to end fire, which is moderately more directive. However, the number of lobes in the pattern also increases with increasing current length. An independent control of the beam angle



(a)

(b)

Figure 9.20 Far-field radiation patterns showing $E_\theta(\theta)$ for the traveling-wave current in Fig. 9.19, with $\beta = k_0$: (a) $L = 1.0\lambda_0$, (b) $L = 5.0\lambda_0$, (c) $L = 10.0\lambda_0$, and (d) $L = 20.0\lambda_0$.



(c)

(d)

Figure 9.20 Continued.

and the beam width is not possible. Hence, the antenna length must be chosen in accordance with the desired angle of maximum radiation. An approximate formulas for the angle of maximum radiation is

$$\theta_0 = \cos^{-1}\left(1 - \frac{0.371}{L/\lambda_0}\right) \tag{9.59}$$

Subsequent maxima occur at $\theta = \theta_m\,(m = 1, 2, 3, \ldots)$ given by [38]

$$\theta_m = \cos^{-1}\left(1 - \frac{p_m}{L/\lambda_0}\right) \tag{9.60}$$

where $p_m = 1.465, 2.48, 3.485, 4.495, 5.5, \ldots,$ for $m = 1, 2, 3, \ldots$. Note that as m increases, p_m approaches $(2m + 1)/2$.

The angle at which the nth null away from end fire appears is [38]

$$\theta_n = \cos^{-1}\left(1 - \frac{n}{L/\lambda_0}\right) \tag{9.61}$$

A practical arrangement for producing a traveling wave current is the horizontal wire at a height h over the earth (modeled as a perfect conductor), shown in Fig. 9.21. The wire and earth form (via image theory) a two-wire (twin-line) transmission line with separation $2h$ between the wires. A termination of the wire to the earth with a load resistance equal to twice the characteristic impedance Z_0 of the twin line will minimize reflections from the end. For this antenna the pattern of Eq. (9.58) is modified by multiplying by a factor $2j\sin(k_0 h\cos\theta)$. The pattern then becomes

$$\begin{aligned} E_\theta = &\left(-\frac{j\omega\mu_0}{4\pi r}e^{-jk_0 r}\cos\theta\cos\phi\right)(I_0 L)e^{-j(L/2)(\beta/k_0-\cos\theta_x)} \\ &\times 2j\sin(k_0 h\cos\theta)\mathrm{sinc}\left[\frac{k_0 L}{2}\left(\frac{\beta}{k_0} - \cos\theta_x\right)\right] \end{aligned} \tag{9.62}$$

$$\begin{aligned} E_\phi = &\left(\frac{j\omega\mu_0}{4\pi r}e^{-jk_0 r}\sin\phi\right)(I_0 L)e^{-j(L/2)(\beta/k_0-\cos\theta_x)} \\ &\times 2j\sin(k_0 h\cos\theta)\mathrm{sinc}\left[\frac{k_0 L}{2}\left(\frac{\beta}{k_0} - \cos\theta_x\right)\right] \end{aligned} \tag{9.63}$$

where $\cos\theta_x = \sin\theta\cos\phi$. Note that in the coordinate system of Fig. 9.21, the antenna radiates both polarizations. The electric field would be polarized in the θ direction if the coordinate system were rotated to align the z axis with the wire.

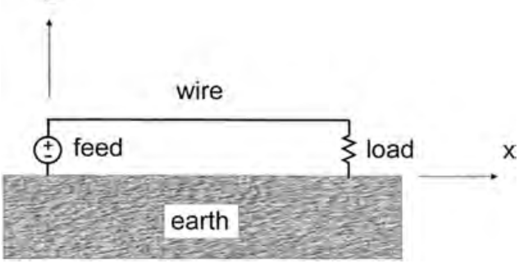


Figure 9.21 A practical traveling-wave antenna consisting of a long horizontal wire over the earth, terminated by a matched load resistor.

9.3.3. Leaky-wave Antennas

As mentioned previously, a leaky-wave antenna (LWA) support a fast wave that radiates continuously along the length of the structure. The two types, uniform and periodic, are considered separately.

Uniform Structures

A typical example of a uniform leaky-wave antenna is a rectangular waveguide with a longitudinal slot, as shown in Fig. 9.22 [39]. Another variation on the design would be an array of closely spaced rectangular or circular slots in the waveguide wall instead of a long longitudinal slot [39]. Although, technically speaking, the periodic structure would not be a uniform structure, it could be modeled as such, provided the slots are closely spaced so that radiation comes only from the fundamental fast waveguide mode, and not a higher order Floquet mode (as for the periodic leaky-wave antennas discussed in the next section). The simple structure in Fig. 9.22 illustrates the basic properties common to all uniform leaky-wave antennas.

The fundamental TE_{10} waveguide mode is a fast wave, with $\beta < k_0$. In particular,

$$\frac{\beta}{k_0} = \sqrt{1 - \left(\frac{\pi}{k_0 a}\right)^2} \tag{9.64}$$

The fast-wave property of the aperture distribution in the slot causes the antenna to radiate similar to a phased array, with the angle of the beam from the waveguide axis given approximately by [37]

$$\cos\theta_0 \approx \frac{\beta}{k_0} \tag{9.65}$$

The radiation causes the wave number k_z of the propagating mode within the open waveguide structure to become complex, so that $k_z = \beta - j\alpha$. The constants β and α are

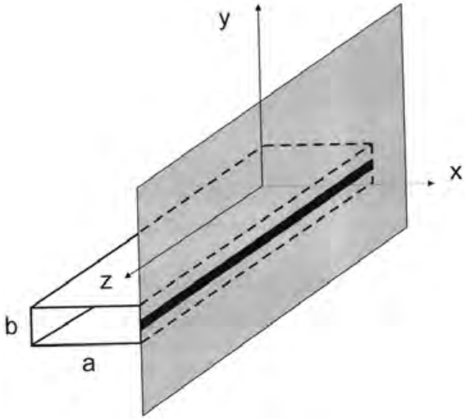


Figure 9.22 A leaky-wave antenna consisting of a rectangular waveguide with a long longitudinal slot in the narrow wall of the waveguide. An infinite ground plane surrounds the slot.

referred to as the *phase* and *attenuation* constants, respectively. The phase constant controls the beam angle, and this can be varied by changing the frequency.

From image theory, the radiation from this structure in the region $x > 0$ is essentially due to a magnetic line current in free space of the form

$$K(z') = A\exp(-jk_z z') \tag{9.66}$$

that exists from $z' \in (0, \infty)$. The normalized pattern shape for the far-zone field $E_\phi(\theta)$ has the shape [37]

$$R(\theta) = \left|E_\phi(\theta)\right| = \left|\frac{\sin(\theta)}{(k_z/k_0) - \cos\theta}\right| \tag{9.67}$$

The attenuation constant controls the beam width of the pattern. An approximate formula for the beamwidth, measured between half-power points, is

$$BW = 2\csc\theta_0 \frac{\alpha}{k_0} \tag{9.68}$$

As is typical for a uniform LWA, the beam cannot be scanned too close to broadside ($\theta_0 = 90^\circ$), since this corresponds to the cutoff frequency of the waveguide. In addition, the beam cannot be scanned too close to end fire ($\theta_0 = 0^\circ$) since this requires operation at frequencies significantly above cutoff, where higher-order modes can propagate, at least for an air-filled waveguide. The $\sin\theta$ term in Eq. (9.67) also limits the end-fire scan. Scanning is limited to the forward quadrant only ($0 < \theta_0 < \pi/2$), for a wave traveling in the positive z direction.

This one-dimensional (1D) leaky-wave aperture distribution results in a "fan beam" having a narrow beam in the H plane (xz plane) with a beam width given by Eq. (9.67), and a broad beam in the E plane (xy plane). A pencil beam can be created by using an array of such 1D radiators.

H-plane patterns for the case $\beta/k_0 = 0.7071$ and $\alpha/k_0 = 0.1$ and 0.01 are shown in Fig. 9.23. This particular value of β corresponds to a beam angle of 45°. It is seen that, in accordance with Eq. (9.68), the pattern corresponding to the smaller α value has a much smaller beamwidth. Unlike the slow-wave structure, a very narrow beam can be created at any angle by choosing a sufficiently small value of α.

One interesting property of the leaky-wave antenna is the exponentially growing or "improper" nature of the near field surrounding the aperture region [36]. To understand this, consider an infinite line source that extends over the entire z axis, having the form of Eq. (9.66). In the air region surrounding the line source, the electric vector potential would have the form [40]

$$F_z = A\frac{\varepsilon_0}{4j} H_0^{(2)}(k_\rho \rho) e^{-jk_z z} \tag{9.69}$$

where

$$k_\rho = \left(k_0^2 - k_z^2\right)^{1/2} \tag{9.70}$$

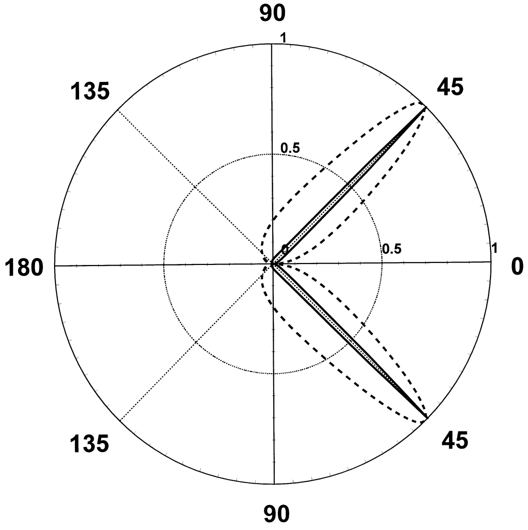


Figure 9.23 H-plane patterns for a leaky-wave line source of magnetic current existing in the semi-infinite region $0 < z < \infty$. The phase constant of the current wave is $\beta/k_0 = 0.7071$. Results are shown for two different values of the attenuation constant, $\alpha/k_0 = 0.1$ (dashed line) and 0.01 (solid line).

To further simplify this expression, consider large radial distances ρ from the z axis, so that the Hankel function may be asymptotically approximated, yielding

$$F_z = A\frac{\varepsilon_0}{4j}\sqrt{\frac{2j}{\pi k_\rho \rho}}e^{-jk_\rho \rho}e^{-jk_z z} \tag{9.71}$$

The wavenumber k_z is in the fourth quadrant of the complex plane. Therefore the radial wavenumber k_ρ from Eq. (9.70) must lie within either the first or third quadrants. Assuming that $\beta < k_0$, the physical choice is the one for which $\mathrm{Re}(k_\rho) > 0$, corresponding to an outward radiating wave. Hence k_ρ is within the first quadrant, and therefore $\mathrm{Im}(k_\rho) > 0$. That is, the wave field exponentially grows with radial distance away from the axis. For a leaky wave existing over the entire z axis, the radiation condition at infinity would be violated. However, for the semi-infinite line source existing over the region $(0, \infty)$ (corresponding to a leaky-wave antenna with a practical feed), the field surrounding the source grows only within an angular region defined by the leakage angle, as shown in Fig. 9.24 [36]. Outside this region the field decays rapidly. (In Fig. 9.24 the strength of the field is indicated by the closeness of the radiation arrows.)

A control of the beam shape may be achieved by tapering the slot width, so that the slot width w, and hence the attenuation constant α, is now a function of z. Suppose that it is desired to achieve a amplitude taper $A(z)$ in the line source amplitude. Approximately, the power radiated per unit length $P_L(z)$ is proportional to $|A(z)|^2$. The attenuation constant is related to $P_L(z)$ and to the power $P(z)$ flowing down the waveguide as [40]

$$\alpha(z) = \frac{P_L(z)}{2P(z)} = -\frac{1}{2P(z)}\frac{dP(z)}{dz} \tag{9.72}$$

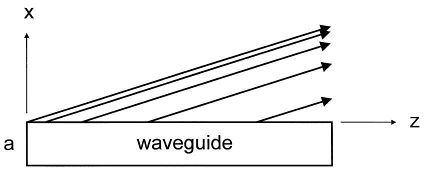


Figure 9.24 An illustration of the near-field behavior of a leaky wave on a guiding structure that begins at $z=0$ (illustrated for the leaky-wave antenna of Fig. 9.22). The rays indicate the direction of power flow in the leaky-wave field, and the closeness of the rays indicates the field amplitude. This figure illustrates the exponential growth of the leaky-wave field in the x direction out to the leakage boundary.

Consider a finite length of radiating aperture, extending from $z=0$ to $z=L$, with a terminating load at $z=L$ that absorbs all remaining power. After some manipulations, the formula for α can be cast into a form involving the desired aperture function $A(z)$ and the radiation efficiency e_r, defined as the power radiated divided by the total input power (the radiation efficiency is less than unity because of the load at the end). The result is [32]

$$\alpha(z)=\frac{(1/2)A^2(z)}{(1/e_r)\int_0^L A^2(z)dz-\int_0^Z A^2(z)dz} \tag{9.73}$$

A typical design would call for a 90% radiation efficiency ($e_r=0.9$).

Equation (9.73) implies that the attenuation constant must become larger near the output (termination) end of the structure, and hence the loading (e.g., slot width) must become larger. In a practical design the loading would typically also be tapered to zero at the input (feed) end to ensure a gradual transition from the nonleaky to the leaky section of waveguide.

Periodic Structures

This type of leaky-wave antenna consists of a fundamentally slow-wave structure that has been modified by periodically modulating the structure in some fashion. A typical example is a rectangular waveguide that is loaded with a dielectric material and then modulated with a periodic set of slots, as shown in Fig. 9.25. Many of the features common to periodic leaky-wave antennas may be discussed by consideration of this simple structure.

It is assumed that the relative permittivity of the filling material is sufficiently high so that the TE_{10} mode is a slow wave over the frequency region of interest. This will be the case provided

$$\varepsilon_r>1+\left(\frac{\pi}{k_0a}\right)^2 \tag{9.74}$$

for all values of k_0 in the range of interest. The waveguide mode is thus a nonradiating slow wave. However, because of the periodicity, the modal field of the periodically loaded waveguide is now in the form of a Floquet mode expansion [33],

$$E(x,y,z)=f(x,y)\sum_{n=-\infty}^{\infty}A_ne^{-jk_{zn}z} \tag{9.75}$$

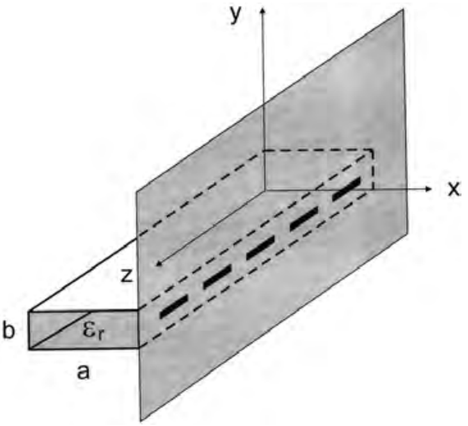


Figure 9.25 A periodic leaky-wave antenna consisting of a rectangular waveguide that is filled with a dielectric material and loaded with a periodic array of longitudinal slots in the narrow wall of the waveguide. An infinite ground plane surrounds the slots. The periodicity is *d*.

where

$$k_{zn} = k_{z0} + \frac{2\pi n}{d} \tag{9.76}$$

is the wave number of the *n*th Floquet mode or space harmonic. The zeroth wave number $k_{z0} = \beta - j\alpha$ is usually chosen to be the wave number that approaches the wave number of the closed waveguide when the loading (slot size) tends to zero. The wave number k_{z0} is then termed the propagation wave number of the guided wave.

Leakage (radiation per unit length of the structure) will occur provided one the space harmonics (usually the $n = -1$ space harmonic) is a fast wave, so that $-k_0 < \beta_{-1} < k_0$, where $\beta_{-1} = \beta - 2\pi/d$. By choosing the period *d* appropriately, the beam can be aimed from backward end fire to forward end fire. The beam will scan as the frequency changes, moving from backward end fire to forward end fire. If one wishes to have single-beam scanning over the entire range, the $n = -2$ space harmonic must remain a slow backward wave ($\beta_{-2} < -k_0$), while the fundamental space harmonic remains a slow forward wave ($\beta > k_0$) as the -1 space harmonic is scanned from backward to forward end fire. These design constraints result in the condition [41]

$$\varepsilon_r > 9 + \left(\frac{d}{a}\right)^2 \tag{9.77}$$

where *a* is the larger waveguide dimension.

One difficulty encountered in the scanning of periodic leaky-wave antennas is that the beam shape degrades as the beam is scanned through broadside. This is because the point $\beta_{-1} = 0$ corresponds to $\beta d = 2\pi$. This is a “stop band” of the periodic structure, where all reflections from the slot discontinuities add in phase back to the source [33]. At this point a perfect standing wave is set up within each unit cell of the structure, and the attenuation constant drops to zero. To understand this, consider the simple model of a transmission line (modeling the waveguide) periodically loaded with shunt loads

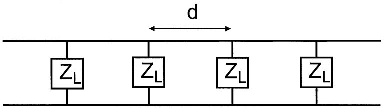


Figure 9.26 A simple approximate transmission line model for the periodic leaky-wave antenna in Fig. 9.7.

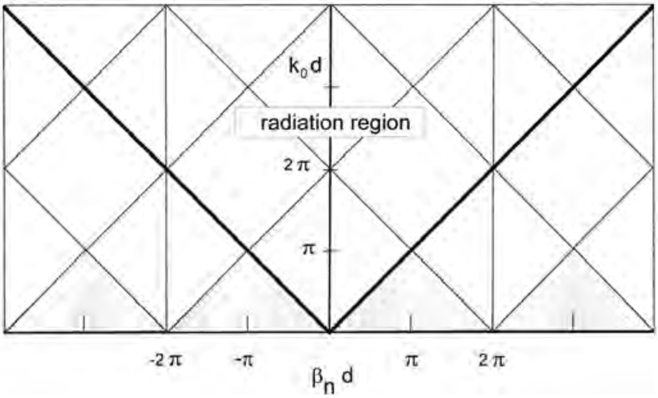


Figure 9.27 Brillouin, or $k-\beta$, diagram that is used for the physical explanation and interpretation of leakage from periodic structures. This diagram is a plot of k_0a versus βa, where a is the period and β is the phase constant of the guided mode (the phase constant of the $n=0$ space harmonic).

(modeling the slots), as shown in Fig. 9.26. When the electrical distance between the loads becomes one wavelength (corresponding to $\beta d = 2\pi$), the total input admittance at any load location becomes infinite (a short circuit). The power absorbed by the loads (radiated by the slots) therefore becomes zero. There are various ways in which the stop-band effect can be minimized. One method is to introduce two radiating elements per unit cell, spaced a distance $d/4$ apart within each cell [42]. At the stop-band point where $\beta d = 2\pi$, the electrical distance between the adjacent elements within the unit cell will be $\pi/2$. The round-trip phase delay between the two elements will then be 180°, which tends to minimize the effects of the reflection from the pair of elements.

When designing, analyzing, and interpreting periodic leaky-wave antennas, a useful tool is the $k-\beta$ or Brillouin diagram [33]. This is a plot of k_0d versus β_nd, as shown in Fig. 9.27. The darker lines on the diagram indicate boundaries where the $n=-1$ space harmonic will be radiating at backward end fire and forward end fire. The shaded regions (the regions inside the lower triangles) are the bound-wave regions [33]. For points in these regions, all of the space harmonics are slow (nonradiating) waves. For a point outside the bound-wave triangles, there must be at least one space harmonic that is a radiating fast wave.

Two-dimensional Leaky-wave Antennas

A broadside pencil beam, or a scanned conical beam, may be obtained by using a two-dimensional (2D) LWA, which supports a radially propagating cylindrical leaky wave instead of a 1D linearly propagating wave. One example of such a structure is the leaky

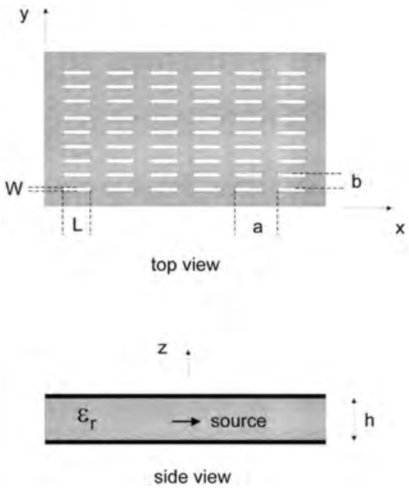


Figure 9.28 A two-dimensional leaky-wave antenna consisting of a periodic array of slots in a top plane, over a grounded dielectric slab. This structure acts as a leaky parallel-plate waveguide that is operating in the first higher order waveguide mode. The structure is excited by a simple source such as a horizontal electric dipole.

parallel-plate waveguide antenna shown in Fig. 9.28, which turns the first higher-order parallel-plate waveguide mode into a leaky mode by allowing radiation to occur through the slots [43]. (Although there is a periodic arrangement of slots, the structure is acting as a uniform leaky parallel-plate waveguide, due to the close spacing of the slots. Another design variation would use a high-permittivity dielectric layer instead of the slotted plate [44].) A simple source such as a horizontal dipole may be used to excite the radial leaky mode. The height h is chosen according to the desired beam angle θ_p. The radial waveguide mode is designed to be a fast wave with a phase constant

$$\beta = \left[k_0^2\varepsilon_r - \left(\frac{\pi}{h}\right)^2\right]^{1/2} \tag{9.78}$$

The beam angle θ_p is measured from broadside and is related to the phase constant as $\beta = k_0 \sin\theta_p$. Solving for the plate separation yields

$$\frac{h}{\lambda_0} = \frac{0.5}{\sqrt{\varepsilon_r - \sin^2\theta_p}} \tag{9.79}$$

An example of a beam produced by such antenna is shown in Fig. 9.29 at a frequency of 12 GHz, using an air substrate. Patterns are shown for three different values of the substrate thickness h, demonstrating how the beam scans from broadside when the substrate thickness is increased. The three thicknesses chosen correspond to a scan angle of $\theta_p = 0°$, 15°, and 30°. The excitation is taken as a simple horizontal y-directed electric dipole in the middle of the air region, directly below the center slot. The structure is assumed to be infinite in the horizontal directions. Near the beam peak, a pencil beam is obtained with nearly equal beam widths in the E and H planes. Further details may be found in Ref. 43.

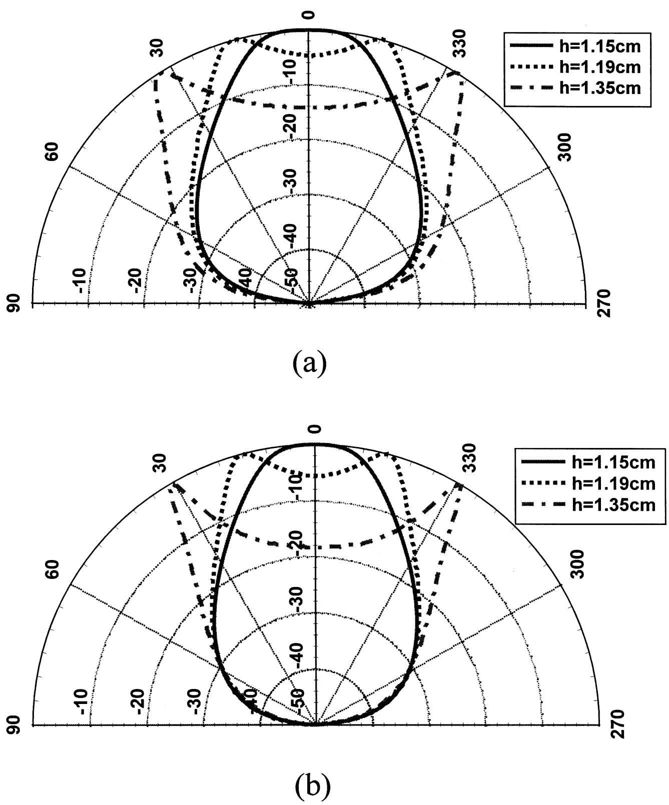


Figure 9.29 Radiation patterns for two-dimensional leaky-wave antenna shown in Fig. 9.28. Patterns are shown for three different values of the substrate thickness ($h = 1.15$ cm, 1.19 cm, 1.35 cm) to illustrate how the beam changes from a pencil beam at broadside to a scanned conical beam as the substrate thickness increases. An air substrate is assumed, and the frequency is 12 GHz: (a) E-plane patterns and (b) H-plane patterns. $l = 0.8$ cm, $w = 0.05$ cm, $a = 1.0$ cm, $b = 0.3$ cm. The y-directed source dipole is in the middle of the air substrate, directly below one of the slots.

9.4. APERTURE ANTENNAS

9.4.1. Introduction

There are classes of antennas that have a physical aperture through which the structure radiates electromagnetic energy. Examples are horn antennas, slotted-waveguide antennas, and open-ended waveguide. In addition, many antennas are more conveniently represented for analysis or qualitative understanding by equivalent apertures. In this section, a general analytical treatment of radiation from apertures is summarized. This approach is used to show the basic characteristics of some common aperture antennas. Also, a discussion of reflector antennas, in the context of an equivalent aperture, is presented.

9.4.2. Radiation from Apertures

Consider a general radiator, as shown in Fig. 9.30. Using the equivalence principle [40], the sources in the closed region S can be removed and equivalent surface-current densities

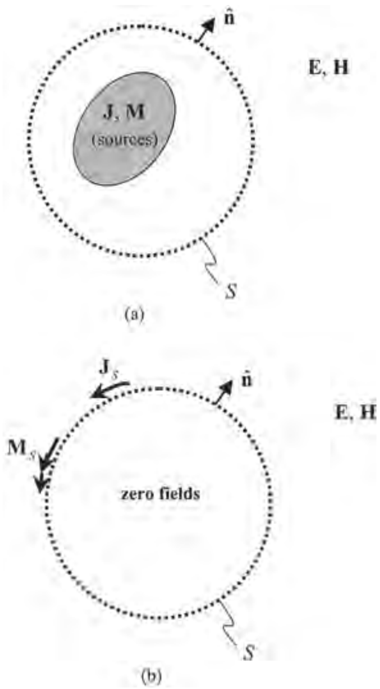


Figure 9.30 Equivalent aperture representation for a general radiator: (a) original problem and (b) equivalent problem.

($\mathbf{J}_S$, $\mathbf{M}_S$) placed on the equivalent aperture surface S. If the fields inside S are assumed to be zero, the equivalent current densities on S are given by

$$\mathbf{J}_S = \hat{\mathbf{n}} \times \mathbf{H}_A \qquad \mathbf{M}_S = \mathbf{E}_A \times \hat{\mathbf{n}} \tag{9.80}$$

where $\mathbf{E}_A$ and $\mathbf{H}_A$ are the fields on S produced by the original sources. These equivalent currents produce the same fields as the original sources in the region outside of S. In the far field the radiated electric field is given by

$$\mathbf{E}(\mathbf{J}_S, \mathbf{M}_S) \approx -j\omega - \hat{\mathbf{r}} \times \hat{\mathbf{r}} \times A - j\omega\sqrt{\mu\varepsilon}\,\mathbf{F} \times \hat{\mathbf{r}} \tag{9.81}$$

where only the θ and ϕ components are used. The far-field potential vectors are given by

$$\begin{aligned} \mathbf{A} &\approx \frac{e^{-jkr}}{4\pi r}\int_S \mathbf{J}_S(\hat{\mathbf{r}}')e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'}\,dS' \\ \mathbf{F} &\approx \frac{e^{-jkr}}{4\pi r}\int_S \mathbf{M}_S(\hat{\mathbf{r}}')e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'}\,dS' \end{aligned} \tag{9.82}$$

Substituting Eq. (9.80) into Eq. (9.82) yields [25]

$$\begin{aligned}
\mathbf{A} &\approx \frac{e^{-jkr}}{4\pi r}\hat{\mathbf{n}} \times \int_S \mathbf{H}_A(\hat{\mathbf{r}}')e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'}dS' = \frac{e^{-jkr}}{4\pi r}\hat{\mathbf{n}} \times \mathbf{Q} \\
\mathbf{F} &\approx -\frac{e^{-jkr}}{4\pi r}\hat{\mathbf{n}} \times \int_S \mathbf{E}_A(\hat{\mathbf{r}}')e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'}dS' = -\frac{e^{-jkr}}{4\pi r}\hat{\mathbf{n}} \times \mathbf{P}
\end{aligned} \tag{9.83}$$

where

$$\begin{aligned}
\mathbf{Q} &= \int_S \mathbf{H}_A(\hat{\mathbf{r}}')e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'}dS' \\
\mathbf{P} &= \int_S \mathbf{E}_A(\hat{\mathbf{r}}')e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'}dS'
\end{aligned} \tag{9.84}$$

In many practical antenna problems, the radiating sources lie in a half space ($z < 0$) and an equivalent aperture surface can be defined in the xy plane ($\hat{\mathbf{n}} = \hat{\mathbf{z}}$). Hence, for this case Eqs. (9.81), (9.83), and (9.84) reduce to the following expressions for the radiated electric fields for $z > 0$:

$$\begin{aligned}
E_\theta &\approx j\omega\mu\frac{e^{-jkr}}{4\pi r}\left[\sqrt{\frac{\varepsilon}{\mu}}\big(P_x\cos\phi + P_y\sin\phi\big) + \cos\theta\big(Q_y\cos\phi - Q_x\sin\phi\big)\right] \\
E_\phi &\approx j\omega\mu\frac{e^{-jkr}}{4\pi r}\left[\cos\theta\sqrt{\frac{\varepsilon}{\mu}}\big(P_y\cos\phi - P_x\sin\phi\big) - \big(Q_x\cos\phi + Q_y\sin\phi\big)\right]
\end{aligned} \tag{9.85}$$

where

$$\begin{aligned}
\mathbf{Q} &= \int_S \mathbf{H}_A(x', y')e^{jk(x'\sin\theta\cos\phi + y'\sin\theta\sin\phi)}dx'dy' = \int_S \mathbf{H}_A(\rho', \phi')e^{jk\rho'\sin\theta\cos(\phi-\phi')}\rho'd\rho'd\phi' \\
\mathbf{P} &= \int_S \mathbf{E}_A(x', y')e^{jk(x'\sin\theta\cos\phi + y'\sin\theta\sin\phi)}dx'dy' = \int_S \mathbf{E}_A(\rho', \phi')e^{jk\rho'\sin\theta\cos(\phi-\phi')}\rho'd\rho'd\phi'
\end{aligned} \tag{9.86}$$

Notice that the integrals in Eq. (9.86) are simply two-dimensional Fourier transforms of the aperture fields.

For many planar aperture antennas, a physical aperture is cut into a conducting ground plane (xy plane). For this discussion, a physical aperture is a slot or hole in a conductor through which radiated electromagnetic waves emanate. The components of the electric field tangent to the ground plane are equal to zero except in the aperture; hence, $\mathbf{M}_S$ is nonzero only in the aperture. In general, $\mathbf{J}_S$ is nonzero over the entire xy plane. In addition, for many directive antennas, the fields in the xy plane are often approximated as zero except in the aperture (even when no ground plane is present). Since the fields in the $z < 0$ half space are assumed to be zero, it is usually convenient to replace this half space with a perfect electric conductor. As a result, using image theory, the equivalent $\mathbf{J}_S$ sources in the xy plane are shorted out and the equivalent $\mathbf{M}_S$ sources double in strength [40]. Therefore, in Eq. (9.86) $\mathbf{H}_A \to 0$, $\mathbf{E}_A \to 2\mathbf{E}_A$, and the integration over the entire xy

plane reduces to integration only over the support of $\mathbf{M}_S$ (aperture surface S_A, where $\mathbf{E}_A$ is nonzero). The directivity of a planar aperture is given by [25]

$$D = \frac{4\pi}{\lambda^2} \frac{\left| \int_{S_A} \mathbf{E}_A \, dS' \right|^2}{\int_{S_A} |\mathbf{E}_A|^2 \, dS'} = \frac{4\pi}{\lambda^2} A_e \tag{9.87}$$

where A_e is the effective aperture area.

In the next few sections, this aperture analysis will be used to understand the basic radiation properties of the most common aperture-type antennas.

9.4.3. Electrically Small Rectangular Slot

Consider the rectangular slot aperture shown in Fig. 9.31, where it will be assumed that L, $W \ll \lambda_0$. A common approximation for small slots is to assume a uniform aperture electric field distribution

$$\mathbf{E}_A \approx \begin{cases} E_0 \hat{\mathbf{y}} & |x| \leq L/2, |y| \leq W/2 \\ 0 & \text{otherwise} \end{cases} \tag{9.88}$$

Using this in Eq. (9.85) yields the following expressions for the radiated fields:

$$E_\theta \approx j\omega\sqrt{\mu\varepsilon} \frac{e^{-jkr}}{2\pi r} E_0(WL) \sin\phi \left\{ \frac{\sin[(kW/2)\sin\theta\cos\phi]}{(kW/2)\sin\theta\cos\phi} \frac{\sin[(kL/2)\sin\theta\sin\phi]}{(kL/2)\sin\theta\sin\phi} \right\}$$

$$E_\phi \approx j\omega\sqrt{\mu\varepsilon} \frac{e^{-jkr}}{2\pi r} E_0(WL) \cos\theta\cos\phi \left\{ \frac{\sin[(kW/2)\sin\theta\cos\phi]}{(kW/2)\sin\theta\cos\phi} \frac{\sin[(kL/2)\sin\theta\sin\phi]}{(kL/2)\sin\theta\sin\phi} \right\} \tag{9.89}$$

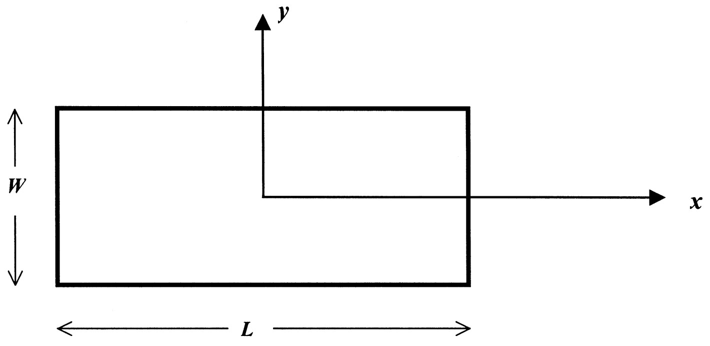


Figure 9.31 Electrically small rectangular slot aperture (L, $W \ll \lambda_0$).

If L, $W \ll \lambda_0$, Eq. (9.89) reduces to

$$\begin{aligned} E_\theta &\approx j\omega\sqrt{\mu\varepsilon}\frac{e^{-jkr}}{2\pi r}E_0 A_p \sin\phi \\ E_\phi &\approx j\omega\sqrt{\mu\varepsilon}\frac{e^{-jkr}}{2\pi r}E_0 A_p \cos\theta\cos\phi \end{aligned} \tag{9.90}$$

where $A_p = WL$ is the physical area of the aperture. For a uniform aperture field $A_e = A_p$; thus, the directivity is

$$D = \frac{4\pi}{\lambda^2}A_p \tag{9.91}$$

9.4.4. Rectangular Horn Antenna

Horn antennas are common high-frequency antennas for moderately high gain applications and applications where exact knowledge of the gain is required (theoretical calculations of the gain are very accurate). There are a number of different horn designs, including those with rectangular apertures (pyramidal, sectoral) and circular apertures (conical) [45,46]. Most use only a single waveguide mode to form the aperture distribution; however, multimode and hybrid-mode horns designs are also used for many specialized applications. In this section the focus will be single-mode horns with rectangular apertures and emphasis will be given to the pyramidal horns since they are the most commonly used horn designs.

Consider the pyramidal horn shown in Fig. 9.32. The horn is excited by the dominant TE_{10} mode of the feeding rectangular waveguide. The electric field distribution in the aperture of the antenna (xy plane) results from this mode propagating from the feed waveguide to the aperture of the horn. The aperture field appears to emanate from the TE_{10} fields at the apex of the horn. In the aperture, the transverse amplitude variation of TE_{10} electric field is preserved; however, the uniform phase (with respect to z) of the exciting mode is not maintained in the aperture since the wave has to propagate different distances to the various points in the aperture. This phase variation from the apex is given by [25]

$$e^{-jk_0(R-R_1)x}e^{-jk_0(R-R_2)y} \tag{9.92}$$

For relatively long horns, $A/2 \ll R_1$ and $B/2 \ll R_2$, the following approximations are commonly used:

$$\begin{aligned} R - R_1 &\approx \frac{1}{2}\frac{x^2}{R_1} \\ R - R_2 &\approx \frac{1}{2}\frac{y^2}{R_2} \end{aligned} \tag{9.93}$$

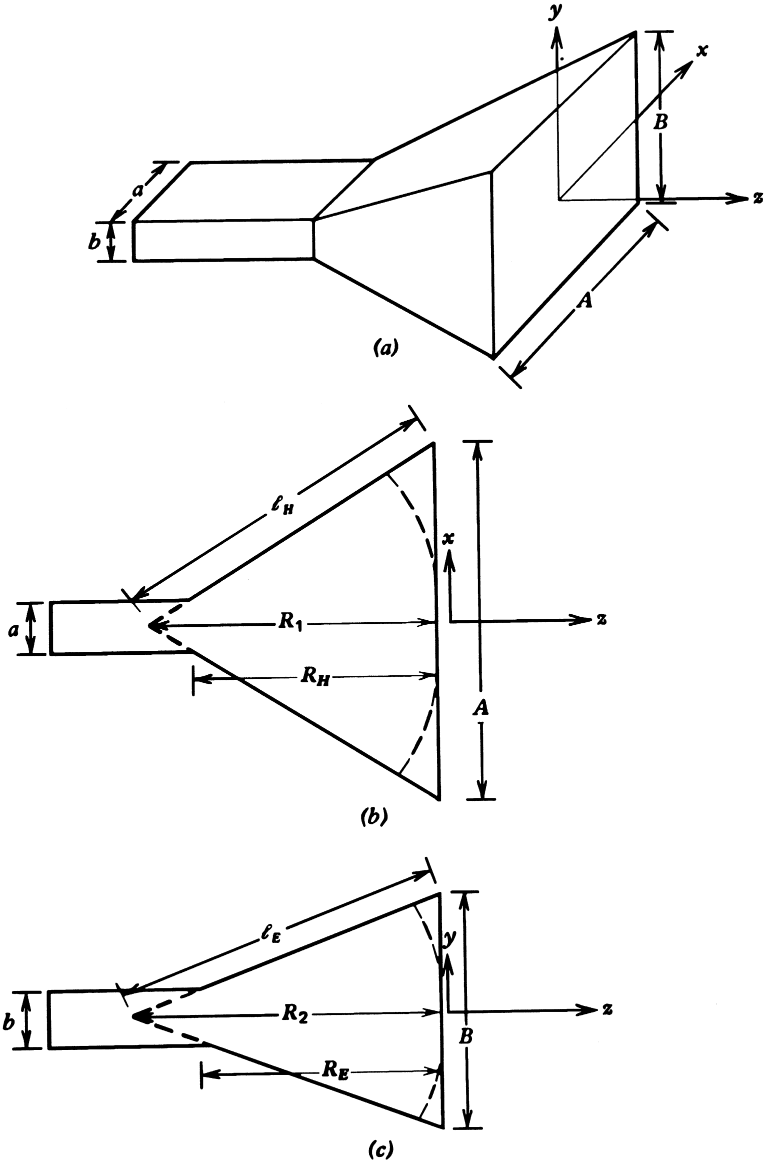


Figure 9.32 Pyramidal horn antenna: (a) perspective view, (b) H-plane (xz plane) cross section, and (c) E-plane (yz plane) cross section.

This leads to an aperture distribution given by

$$\mathbf{E}_A = \hat{\mathbf{y}}E_A = \hat{\mathbf{y}}\cos\frac{\pi x}{A}e^{-j(k_0/2R_1)x^2}e^{-j(k_0/2R_2)y^2} \tag{9.94}$$

Substituting this aperture distribution into Eq. (9.86) yields an integral that can be performed in closed form in the principle planes of the antenna; however, the results are rather complicated in form, involving Fresnel integrals [45]. Plots of the E-plane and H-plane patterns for various horn flares are shown in Fig. 9.33.

The directivity for a horn can be calculated using Eq. (9.87). If there is no phase variation across the aperture (idealized case) the effective aperture area is

$$A_e = \frac{8}{\pi^2} A_p \qquad \text{(uniform aperture phase)} \tag{9.95}$$

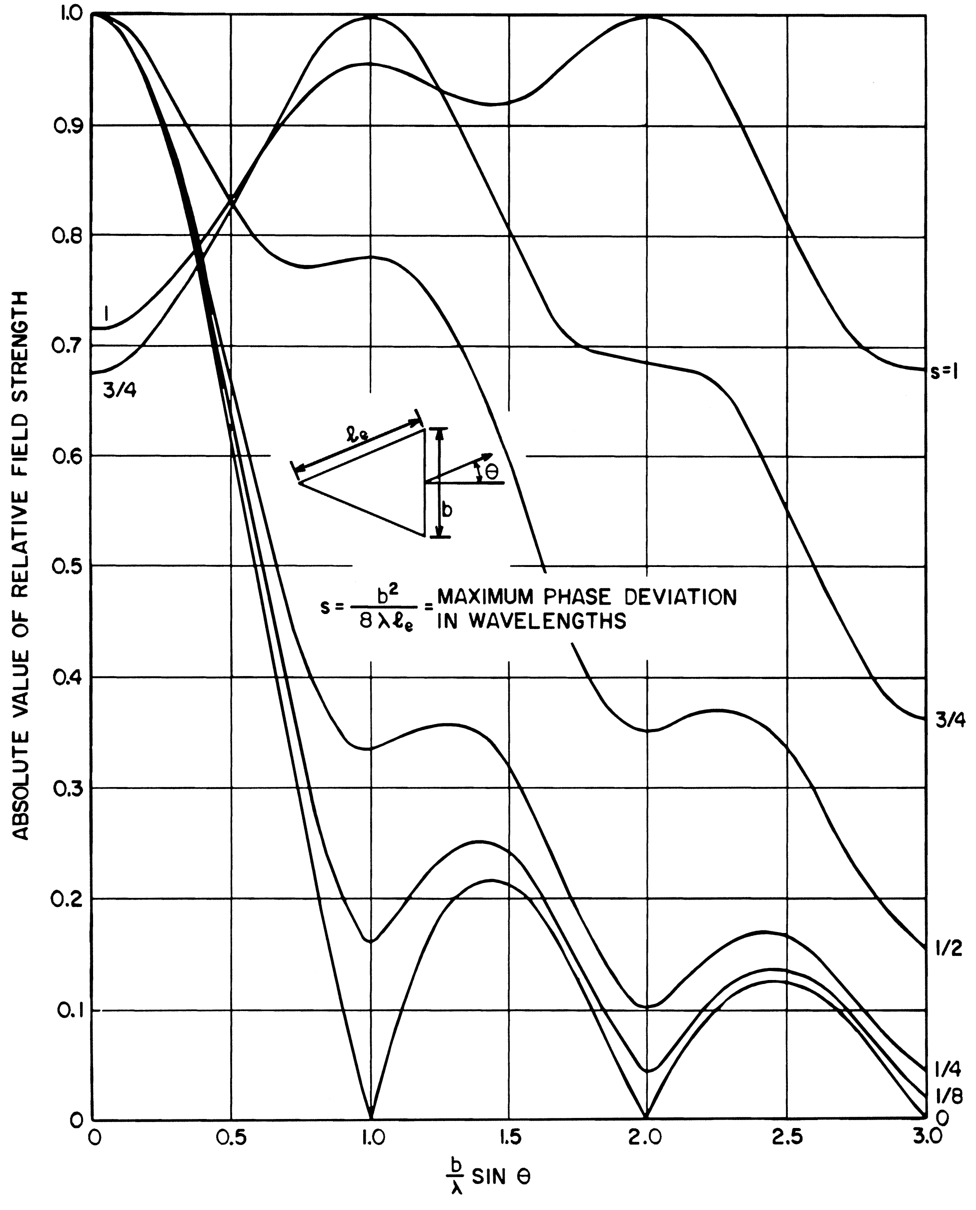
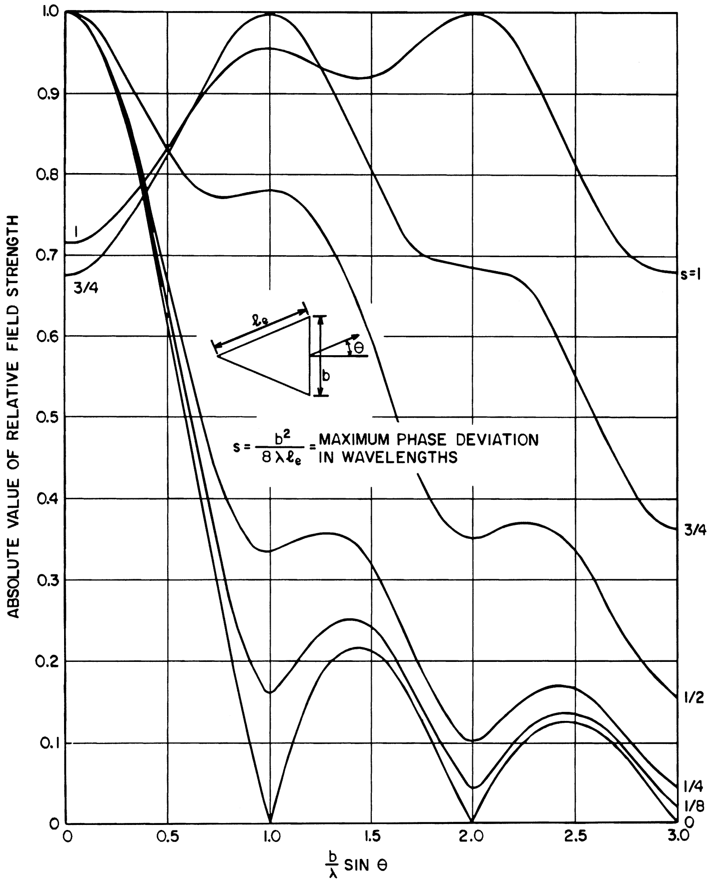


Figure 9.33 Universal radiation patterns for a rectangular horn antenna: (a) E plane and (b) H plane.



(a)

Figure 9.33 Continued.

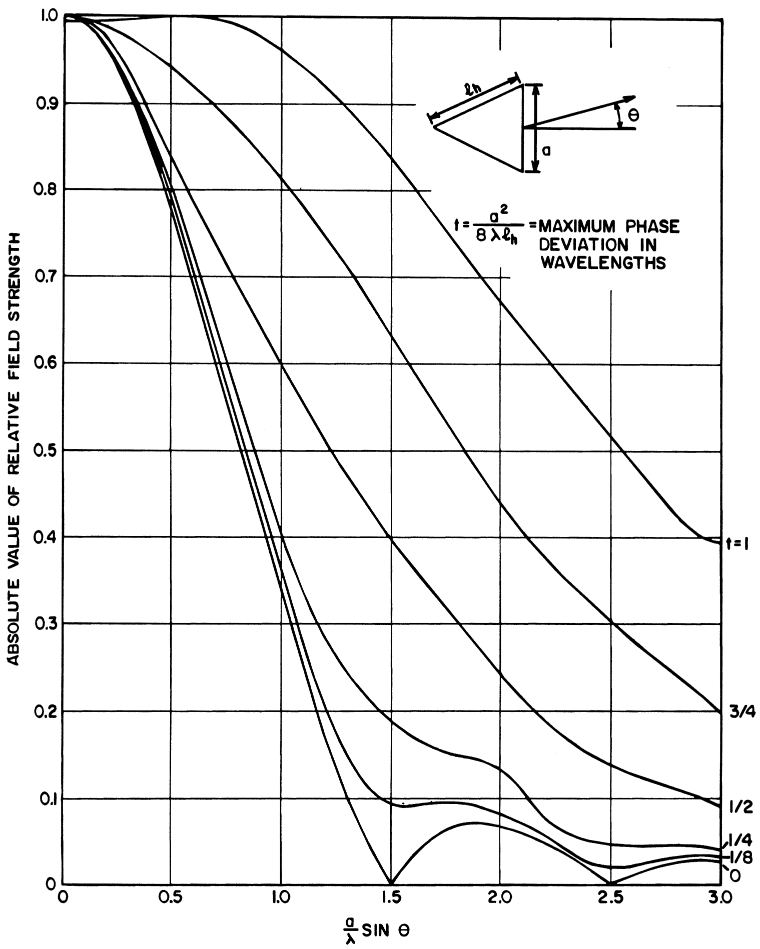
For an optimum pyramidal horn design, a design that produces the maximum directivity along boresight in the E and H planes ($A \approx \sqrt{3\lambda\ell_H}$ and $B \approx \sqrt{2\lambda\ell_E}$) [45], the effective aperture area is

$$A_e = \frac{1}{2} A_p \qquad \text{(optimum pyramidal design)} \tag{9.96}$$

Typically, horn antennas have effective aperture areas that are 40–80% of the physical aperture. Another accurate approach to determine the directivity of a pyramidal horn is to use normalized directivity curves for E- and H-plane sectoral horns [25] as

$$D \approx \frac{\pi}{32}\left(\frac{\lambda}{A} D_E\right)\left(\frac{\lambda}{B} D_H\right) \tag{9.97}$$

where the normalized sectoral-horn directivities (terms in parenthesizes) are shown in Fig. 9.34.



(b)

Figure 9.33 Continued.

9.4.5. Reflector Antennas

For high-gain antenna applications, applications requiring gains of 30 dB or more, reflector antennas are by far the most widely used. The design of these antennas is relatively complex, well beyond the scope of this discussion. In addition, the sheer number of reflector types is too numerous to summarize here [47,48]. However, it is very useful in understanding and designing reflector antennas to think of them in terms of aperture antennas, where the feed and reflector combination establish an aperture distribution that is radiated using the methods described earlier. The objective of this short discussion will therefore be limited to how to calculate an aperture electric field distribution for a simple parabolic reflector antenna.

Consider the parabolic reflector antenna shown in Fig. 9.35. The parabolic reflector is shaped such that the lengths of all ray paths from the feed to the reflector and then to the aperture plane (xy plane) are equal to twice the focal length ($2f$). As such, if the phase of the radiation pattern for the feed antenna is a constant, then the phase distribution in the aperture plane will be a constant. The description for the parabolic surface of the reflector is given by [25]

$$r' = f \sec^2 \frac{\theta'}{2} = \frac{4f^2 + \rho'^2}{4f} \tag{9.98}$$

The displacement from the focal point to any point in the aperture plane is

$$\rho' = r' \sin\theta' = 2f \tan\frac{\theta'}{2} \tag{9.99}$$

Using a geometrical optics or ray argument, it can be readily demonstrated that the amplitude distribution in the aperture plane is a function of the radiation pattern of the

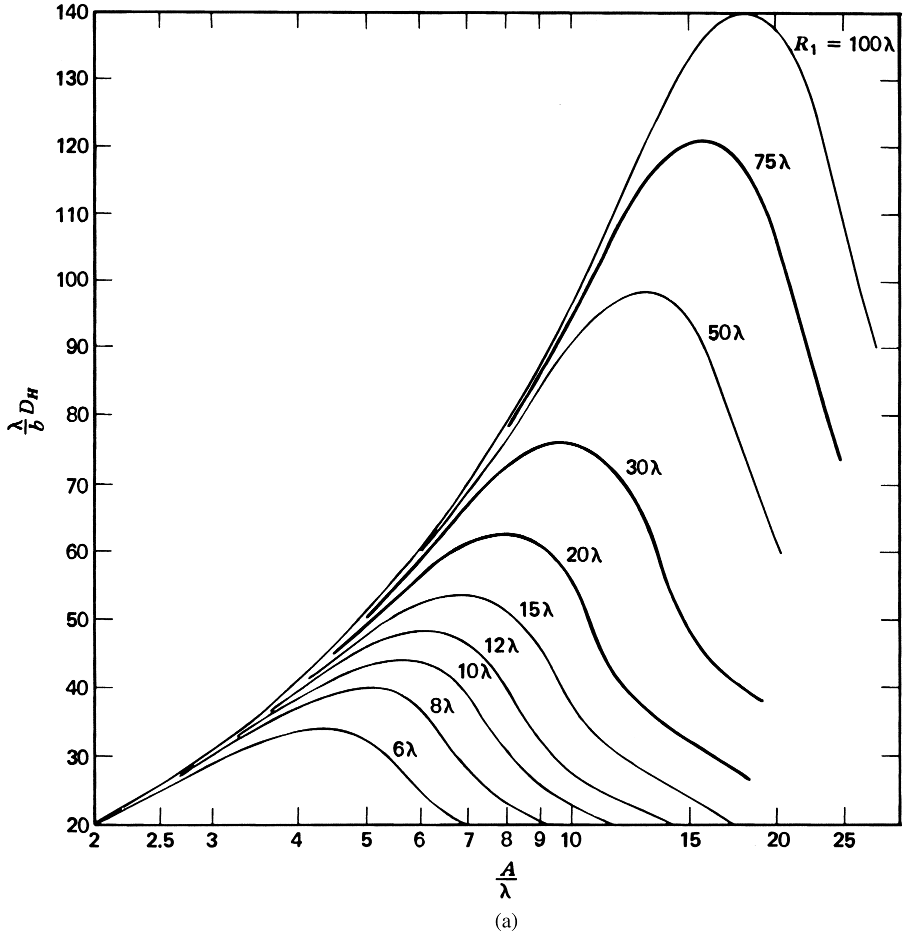


Figure 9.34 Universal directivity curves for rectangular sectoral horn antennas: (a) E plane and (b) H plane.

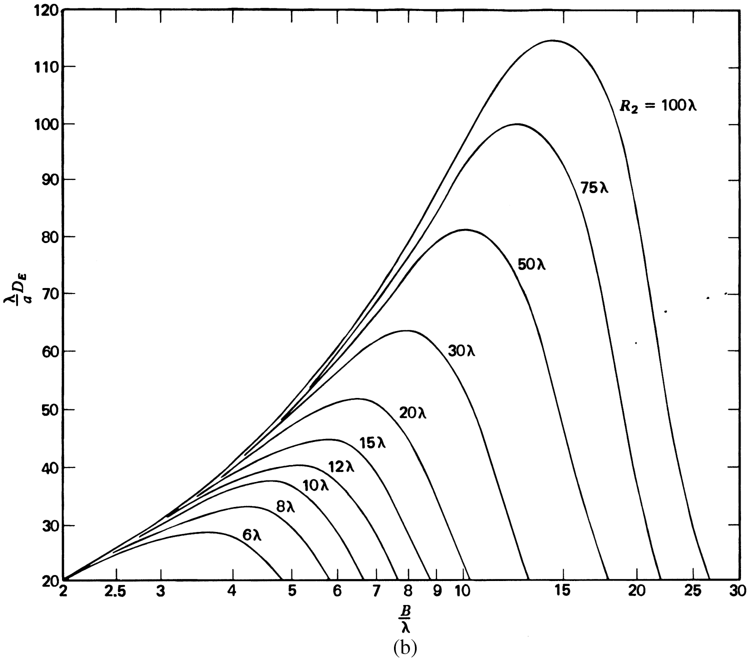


Figure 9.34 Continued.

feed antenna as

$$\mathbf{E}_A(\theta', \phi') = E_0 \frac{F(\theta', \phi')}{r'} \hat{\mathbf{l}}_A \tag{9.100}$$

where $F(\theta', \phi')$ is the normalized pattern of the feed antenna, $\hat{\mathbf{l}}_A$ is a unit vector in the direction of the aperture electric field given by

$$\hat{\mathbf{l}}_A = 2(\hat{\mathbf{n}} \cdot \hat{\mathbf{l}}_F)\hat{\mathbf{n}} - \hat{\mathbf{l}}_F \tag{9.101}$$

$\hat{\mathbf{l}}_F$ is a unit vector in the direction of the electric field radiated by the feed antenna, and $\hat{\mathbf{n}}$ is the unit normal to the surface of the reflector. Finally, the radiation field from the reflector antenna is determined by substituting Eqs. (9.98)–(9.101) into Eqs. (9.85) and (9.86).

9.5. PHASED ARRAYS

9.5.1. Array Far Fields

Phased arrays are arrays of antenna elements for which a radiation beam may be scanned electronically. We first examine the far-field pattern, i.e., the beam characteristics of the array. Consider a planar array of elements uniformly spaced in the $z = 0$ plane. The element locations may be defined via lattice vectors $\mathbf{d}_1$ and $\mathbf{d}_2$, as shown in Fig. 9.36a. Scanning of the array pattern is accomplished by introducing a constant progressive phase

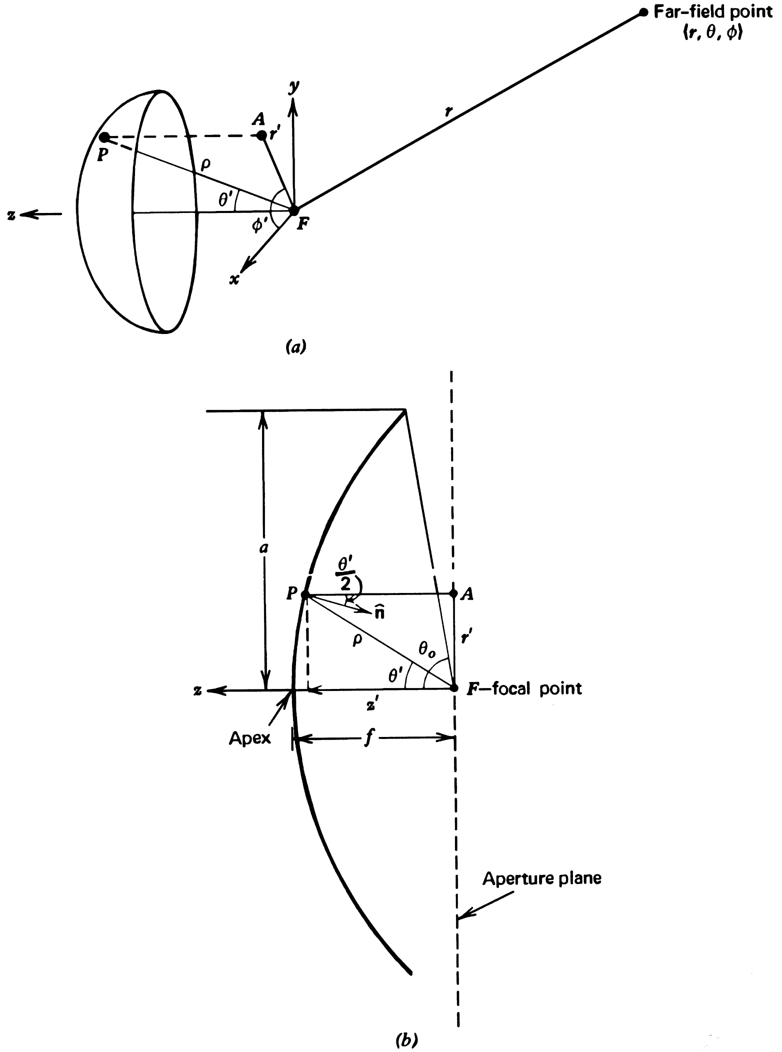


Figure 9.35 Parabolic reflector antenna: (a) perspective view and (b) cross-sectional view.

shift between elements. To determine the radiation pattern, we make the usual assumption, valid for arrays of a large number of elements, that the induced or equivalent current on the (m, n)th array element, $\mathbf{J}_{mn}(\mathbf{r})$, is identical to that of the $(0, 0)$th reference element, $\mathbf{J}_{0,0}(\mathbf{r})$, except for a positive, real amplitude factor a_{mn} and a phase shift ϕ_{mn}:

$$\mathbf{J}_{mn}(\mathbf{r} + m\mathbf{d}_1 + n\mathbf{d}_2) = a_{mn}\mathbf{J}_{0,0}(\mathbf{r})e^{j\phi_{mn}} \tag{9.102}$$

Hence the vector potential in the far field is given by the superposition

$$\begin{aligned}\mathbf{A}(\mathbf{r}) &= \frac{\mu_0}{4\pi r}e^{-jkr}\left(\sum_{m,n} a_{mn}e^{j[k\hat{\mathbf{r}}\cdot(m\mathbf{d}_1+n\mathbf{d}_2)+\phi_{mn}]}\right)\int_{S_{\text{ref}}} \mathbf{J}_{0,0}(\mathbf{r}')e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'}\,dS' \\ &= \mathbf{A}_{\text{ref}}(\mathbf{r})\mathrm{AF}(\hat{\mathbf{r}})\end{aligned} \tag{9.103}$$

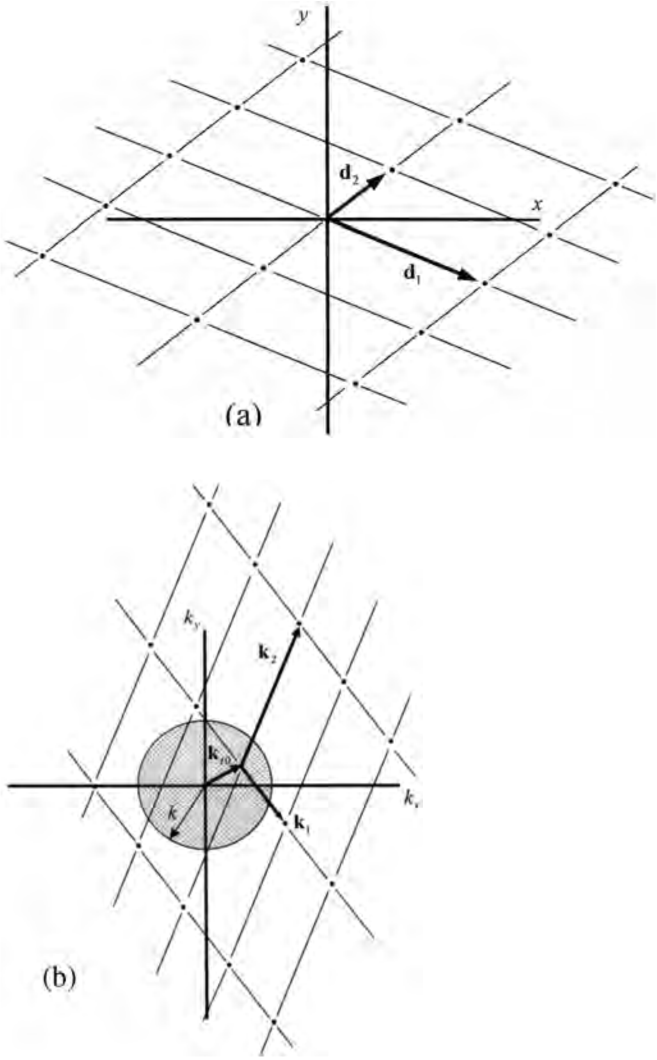


Figure 9.36 (a) The element lattice is defined by the lattice vectors $\mathbf{d}_1$ and $\mathbf{d}_2$. (b) The grating lobe lattice is shown translated by $\mathbf{k}_{t0}$, the phasing needed to scan the array beam within the circle representing the visible region. Lattice vectors $\mathbf{d}_1$ and $\mathbf{d}_2$ are usually chosen such that no grating lobes appear within the visible region as $\mathbf{k}_{t0}$ varies over the desired scan range.

where

$$\mathbf{A}_{\text{ref}}(\mathbf{r}) = \frac{\mu_0}{4\pi r} e^{-jkr} \int_{S_{\text{ref}}} \mathbf{J}_{0,0}(\mathbf{r}') e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'} dS' \tag{9.104}$$

The far-zone electric field is proportional to the components of the vector potential tangent to the far-field sphere:

$$\mathbf{E}(\mathbf{r}) = -j\omega\Big(A_\theta(\mathbf{r})\hat{\boldsymbol{\theta}} + A_\phi(\mathbf{r})\hat{\boldsymbol{\phi}} \Big) \tag{9.105}$$

The angle dependent factor for the reference element,

$$\begin{aligned}\mathbf{F}_{\text{ref}}(\hat{\mathbf{r}}) &= -\frac{4\pi r e^{jkR}}{j\omega\mu_0}\mathbf{E}_{\text{ref}}(\hat{\mathbf{r}}) \\ &= (\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}) \cdot \int_{S_{\text{ref}}} \mathbf{J}_{0,0}(\mathbf{r}')e^{jk\hat{\mathbf{r}}\cdot\mathbf{r}'}dS' \end{aligned} \tag{9.106}$$

is defined as the *element pattern* of the array. The element pattern incorporates, in principle, all the element's directional radiation, polarization, and mutual coupling characteristics. The array factor,

$$\mathrm{AF}(\hat{\mathbf{r}}) = \sum_{m,n} a_{mn} e^{j\left[k\hat{\mathbf{r}}\cdot(m\mathbf{d}_1+n\mathbf{d}_2)+\phi_{mn}\right]} \tag{9.107}$$

accounts for effects due to the array element configuration and excitation. The unit vector in the observation direction (θ, ϕ) is

$$\hat{\mathbf{r}} = \sin\theta\cos\phi\,\hat{\mathbf{x}} + \sin\theta\sin\phi\,\hat{\mathbf{y}} + \cos\theta\,\hat{\mathbf{z}} \tag{9.108}$$

To scan the array to a prescribed beam-pointing angle (θ_0, ϕ_0), the phase factor is chosen so that all contributions to the array factor add in phase in that direction,

$$\phi_{mn} = -k\hat{\mathbf{r}}_0 \cdot (m\mathbf{d}_1 + n\mathbf{d}_2) \tag{9.109}$$

where the unit vector in the beam direction is

$$\hat{\mathbf{r}}_0 = \sin\theta_0\cos\phi_0\hat{\mathbf{x}} + \sin\theta_0\sin\phi_0\hat{\mathbf{y}} + \cos\theta_0\hat{\mathbf{z}} \tag{9.110}$$

Defining observation and phasing wave vectors, $\mathbf{k} = k\hat{\mathbf{r}}$ and $\mathbf{k}_0 = k\hat{\mathbf{r}}_0$, respectively, the array factor may thus be succinctly written as

$$\mathrm{AF}(\hat{\mathbf{r}}) = \sum_{m,n} a_{mn} e^{j(\mathbf{k}-\mathbf{k}_0)\cdot(m\mathbf{d}_1+n\mathbf{d}_2)} \tag{9.111}$$

Components of the wavevectors in the plane of the array are more conveniently expressed in terms of the so-called *grating lobe lattice wave vectors*,

$$\mathbf{k}_1 = -\frac{2\pi\left(\hat{\mathbf{z}} \times \mathbf{d}_2\right)}{A} \qquad \mathbf{k}_2 = \frac{2\pi\left(\hat{\mathbf{z}} \times \mathbf{d}_1\right)}{A} \tag{9.112}$$

that are *biorthogonal* to the configuration space lattice vectors, i.e.,

$$\begin{aligned} &\mathbf{k}_1 \cdot \mathbf{d}_1 = 2\pi \qquad \mathbf{k}_1 \cdot \mathbf{d}_2 = 0, \\ &\mathbf{k}_2 \cdot \mathbf{d}_1 = 0 \qquad \mathbf{k}_2 \cdot \mathbf{d}_2 = 2\pi, \\ &A = \hat{\mathbf{z}} \cdot (\mathbf{d}_1 \times \mathbf{d}_2) = \text{element area} \end{aligned} \tag{9.113}$$

From these properties, it is clear that if the components of the wave vectors transverse to the array normal (denoted by subscript "t") satisfy

$$\mathbf{k}_t = \mathbf{k}_{t0} + p\mathbf{k}_1 + q\mathbf{k}_2 \tag{9.114}$$

in Eq. (9.111) for integer values (p, q) and for any observable angles in the so-called *visible region*, $|k\hat{\mathbf{r}}| < k$, then in addition to the main beam, $(p = 0,\ q = 0)$, additional maxima called *grating lobes* of the array factor appear in the visible region, Fig. 9.36b. These undesired maxima thus appear whenever there exist nonvanishing integers (p, q) satisfying

$$\begin{aligned}&\left|\mathbf{k}_{t0} + p\mathbf{k}_1 + q\mathbf{k}_2\right| < k \quad \text{or}\\ &\left|k\left(\sin\theta_0\cos\phi_0\,\hat{\mathbf{x}} + \sin\theta_0\sin\phi_0\,\hat{\mathbf{y}}\right) + \frac{2\pi\hat{\mathbf{z}}}{A}\times(q\mathbf{d}_1 - p\mathbf{d}_2)\right| < k\end{aligned} \tag{9.115}$$

Equation (9.115) is used in phased array design to determine the allowable element spacings such that no grating lobes are visible when the array is scanned to the boundaries of its prescribed scan range in the angles (θ_0, ϕ_0) and at the highest frequency of array operation. The scan wave vector is shown in Fig. 9.37 and a graphical representation of these grating lobe conditions is depicted in Fig. 9.38. Scanning past one of the grating lobe circle boundaries overlapping the visible region allows a grating lobe into the visible region. For rectangular lattices, it is usually sufficient to check this condition only along the principal scan planes of the array. Triangular element lattices, however, often permit slightly larger spacings (hence larger element areas and fewer elements for a given gain requirement) than rectangular spacings [49]. For one-dimensional (linear) arrays, Eq. (9.115) reduces to

$$\frac{d}{\lambda} < \frac{1}{1 + |\sin\theta_0|} \tag{9.116}$$

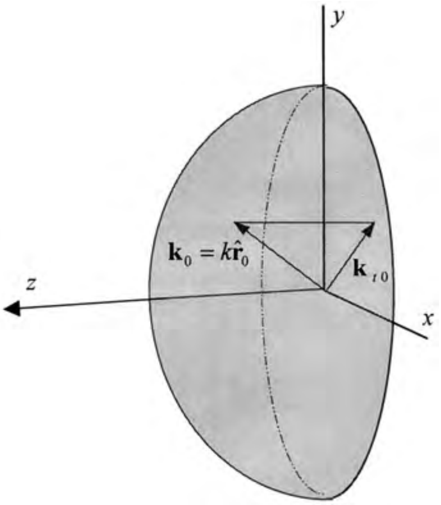


Figure 9.37 The vector $\mathbf{k}_0$ is in the desired scan direction; its projection, $\mathbf{k}_{t0}$, yields the transverse phase gradient required to scan the array to this direction.

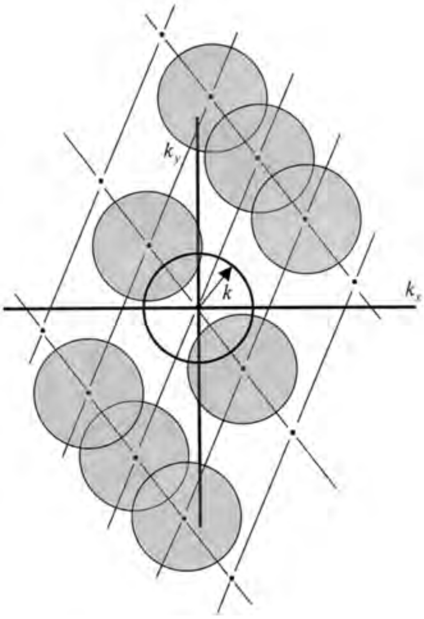


Figure 9.38 Grating-lobe lattice. The circle centers correspond to grating-lobe positions when the main beam is at broadside. As the array main beam is scanned within the visible region (the central unshaded circle), the grating lobes scan within the shaded circles. Grating-lobes enter the visible region when the main beam is scanned to a point in a shaded region overlapping the visible region circle.

where θ_0 is the maximum scan angle measured from the array axis normal. For rectangular element spacings, Eq. (9.116) may also be used to independently determine element spacings in the two orthogonal array planes. Note that for such arrays, grating lobes cannot appear if the spacings are less than $\lambda/2$ in each dimension; for 1λ spacings, grating lobes are at end fire and *always* appear in the visible region as the array is scanned. The above conditions merely ensure that the *peaks* of the grating lobe beams do not appear at end fire, i.e. in the plane of the array, when the array is scanned to a boundary of its scan coverage volume. To ensure that no part of a grating lobe beam appears in the visible region, slightly smaller element spacings than determined by the above procedure must be used. Not only the pattern but also the element impedance matching characteristics of an array may deteriorate rapidly at the onset of a grating lobe.

9.5.2. Array Pattern Characteristics

To examine array pattern characteristics more closely, we specialize to a rectangular array of $M \times N$ elements arranged in a rectangular lattice,

$$\mathbf{d}_1 = d_x\hat{\mathbf{x}} \qquad \mathbf{d}_2 = d_y\hat{\mathbf{y}} \qquad \mathbf{k}_1 = \frac{2\pi}{d_x}\hat{\mathbf{x}}, \qquad \mathbf{k}_2 = \frac{2\pi}{d_y}\hat{\mathbf{y}} \tag{9.117}$$

The excitation of such an array is often *separable*, $a_{mn} = a_m b_n$, so that the array factor is also separable:

$$\mathrm{AF}(\hat{\mathbf{r}}) = \mathrm{AF}_x(\psi_x)\mathrm{AF}_y(\psi_y) \tag{9.118}$$

where

$$\psi_x = kd_x(\sin\theta\cos\phi - \sin\theta_0\cos\phi_0) \qquad \psi_y = kd_y(\sin\theta\sin\phi - \sin\theta_0\sin\phi_0) \tag{9.119}$$

and the array factor is the product of two linear array factors,

$$\mathrm{AF}_x(\psi_x) = \sum_n a_n e^{jn\psi_x} \qquad \mathrm{AF}_y(\psi_y) = \sum_m b_m e^{jm\psi_y} \tag{9.120}$$

If the array is uniformly excited, $a_m = 1/M, b_n = 1/N$, and centered with respect to the coordinate origin with M elements along the x dimension and N elements along the y dimension, the array factors can be summed in closed form, yielding

$$\mathrm{AF}_x(\psi_x) = \frac{\sin(M\psi_x/2)}{M\sin(\psi_x/2)} \qquad \mathrm{AF}_y(\psi_y) = \frac{\sin\left(N\psi_y/2\right)}{N\sin\left(\psi_y/2\right)} \tag{9.121}$$

Thus the array factor is a product of linear array factors of the form

$$\mathrm{AF}(\psi) = \frac{\sin(N\psi/2)}{N\sin(\psi/2)} \tag{9.122}$$

where

$$\psi = kd(\sin\alpha - \sin\alpha_0) \tag{9.123}$$

and α is the observation angle measured from a normal to the equivalent linear array axis; the linear array is assumed to be scanned to an angle α_0 from the array normal. The magnitude of $\mathrm{AF}(\psi)$ versus the parameter ψ for several values of N appears in Fig. 9.39. Since the array patterns are symmetric about $\psi = 0$ and periodic with period 2π, it suffices to plot them on the interval $(0, \pi)$. The pattern has a main beam at $\psi = 0$ with grating lobes at $\psi = \pm p2\pi = 1, 2, \ldots$. The scan angle and element spacing determine whether the grating lobes are visible. Between the main beam and first grating lobe, the pattern has $N - 1$ zeros at $\psi = p2\pi/N, p = 1, 2, \ldots, N - 1$, and $N-2$ side lobes with peaks located approximately at $\psi = (p + 1/2)2\pi/N, p = 1, 2, \ldots, N - 2$. For large N, the first side-lobe level is independent of N and equal to that of a continuous, uniform aperture distribution, -13.2 dB. Indeed, in the vicinity of the main beam at $\psi \approx 0$, for large N we have

$$\mathrm{AF}(\psi) \approx \frac{\sin(N\psi/2)}{N(\psi/2)} \tag{9.124}$$

i.e., the array factor approximates the pattern of a uniform, continuous line source of length Nd. This observation holds for other array distributions as well: if the a_n are chosen

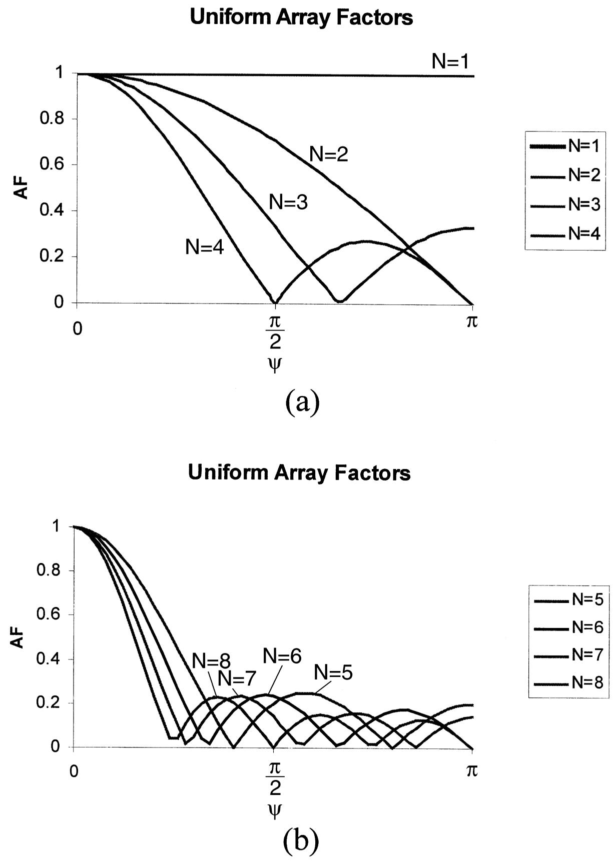


Figure 9.39 Array factors of uniform arrays with (a) $N = 1, 2, 3, 4$ and (b) $N = 5, 6, 7, 8$ elements.

as equi-spaced samples of a continuous distribution, then for large N the pattern in the vicinity of the main beam approaches that associated with the underlying sampled continuous distribution. Thus beam characteristics for continuous distributions may be used to estimate the corresponding quantities for array factors. Table 9.1 [50,51] lists approximate half-power beam widths, first side-lobe levels, and aperture efficiencies for large N for arrays whose element amplitude distributions are discrete samples of the listed continuous distributions. One of the more useful array distributions is the Taylor distribution [52], which allows one to choose $\bar{n}$ side lobes of specified and equal level on either side of the main beam, the remaining side lobes following the behavior of a uniform array pattern. Desirable features of the Taylor distribution are that it produces a physically realizable pattern that has essentially the narrowest beam possible for a given side-lobe level. Figures 9.40 and 9.41 show the pattern and corresponding element weights a_n, respectively, for a Taylor distribution with $\bar{n} = 6$ and a specified 20-dB side-lobe level. A convenient graphical method exists for determining a polar plot of the equivalent line source radiation pattern vs. the observation angle α for a given scan angle α_0. As illustrated for the uniform aperture distribution in Fig. 9.42, the pattern is first plotted vs. the parameter ψ and then projected onto a polar plot of diameter $2kd$ centered at $\psi = -kd \sin \alpha_0$. As the scan angle α_0 changes, the projected beam "scans," and the

Table 9.1 Pattern Characteristics of Various Linear Array Aperture Distributions [50,51]

Aperture distribution $\lvert z_n \rvert < 1,\ n = 1, 2, \ldots, N$ $z_n = \dfrac{2(n-1) - N}{N}$	Aperture efficiency η_a	Half-power beam-width (degrees) L = array length	Maximum side-lobe level (dB below maximum)
Uniform: $a_n = 1$	1.00	51λ/L	13.2
Cosine:			
$a_n = \cos^n\left(\dfrac{\pi z_n}{2}\right)$			
$n = 0$	1	51λ/L	13.2
$n = 1$	0.81	69λ/L	23
$n = 2$	0.667	83λ/L	32
$n = 3$	0.575	95λ/L	40
Parabolic: $a_n = 1 - (1 - \Delta)z_n^2$			
$\Delta = 1.0$	1.0	51λ/L	13.2
$\Delta = 0.8$	0.994	53λ/L	15.8
$\Delta = 0.5$	0.970	56λ/L	17.1
$\Delta = 0$	0.833	66λ/L	20.6
Triangular: $a_n = 1 - \lvert z_n \rvert$	0.75	73λ/L	26.4
Cosine-squared on a pedestal:			
$a_n = 0.33 + 0.66\cos^2\left(\dfrac{\pi z_n}{2}\right)$	0.88	63λ/L	25.7
$a_n = 0.08 + 0.92\cos^2\left(\dfrac{\pi z_n}{2}\right)$	0.74	76.5λ/L	42.8

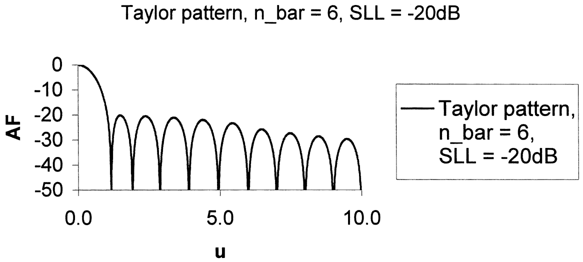


Figure 9.40 Pattern of a continuous Taylor aperture distribution: $\bar{n} = 6$, SLL $=$ -20 dB.

dependence of grating lobe onset on scan angle and element spacing becomes apparent. Further, the approximate broadening of the main beam with scan angle by a factor $1/\cos\alpha_0$ also becomes apparent. For a line source of omnidirectional elements, the actual pattern is obtained by rotating the pattern of the figure about the horizontal axis to form a conical beam.

For separable rectangular aperture distributions, such a graphical representation is not practical to construct, yet it is convenient for visualization. We first imagine that

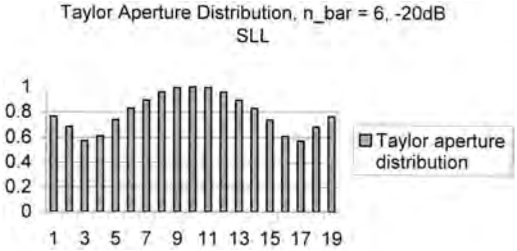


Figure 9.41 Sampled values of continuous Taylor aperture distribution for a 19-element array: $\bar{n} = 6$, SLL $= -20\,\text{dB}$.

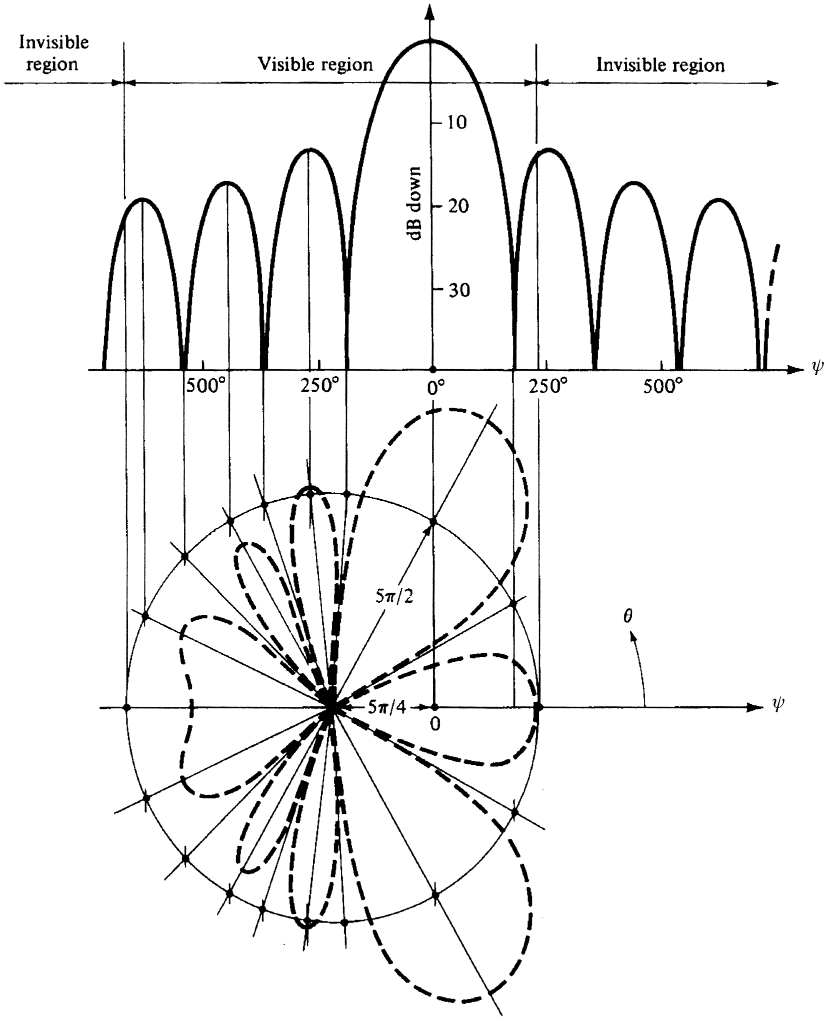


Figure 9.42 Projection of linear array pattern to obtain polar pattern. The location of the center of the polar pattern is determined by the interelement phase shift. The radius of the polar pattern, and hence the visible region, is determined by the frequency. The angle θ in the figure is $\pi/2 - \alpha$. (This material is used by permission of John Wiley and Sons, Inc., *Antenna Theory: Analysis and Design*, C. Balanis, 1997.)

the product $\mathrm{AF}_x(\psi_x)\mathrm{AF}_y(\psi_y)$ is plotted vs. ψ_x/d_x and ψ_y/d_y, where, because of the separability property, pattern cuts for any constant ψ_x/d_x or ψ_y/d_y are identical within a scaling factor. A hemisphere of radius k is then centered at $\psi_x/d_x = -k\sin\theta_0\cos\phi_0$, $\psi_y/d_y = -k\sin\theta_0\sin\phi_0$, and the array-factor pattern multiplied by the vector-valued element pattern is projected onto the hemisphere to obtain the three-dimensional radiation pattern.

9.5.3. Array Gain

The array directivity is usually defined in terms of the *scan element pattern*. Since the directivity of an array changes with its scan and with the mismatch to its feed line, it is convenient to incorporate these effects into the definition of an element gain by referring it to the power *available* rather than the power *input* to the array. Consequently the scan element pattern (gain) is defined as

$$g_{\mathrm{scan}}(\hat{\mathbf{r}}_0) = \frac{4\pi r^2 \left|\mathbf{E}_{\mathrm{scan}}(\hat{\mathbf{r}}_0)\right|^2}{2\eta_0 P_{\mathrm{avail}}} \tag{9.125}$$

where $\mathbf{E}_{\mathrm{scan}}(\hat{\mathbf{r}}_0)$ is the far-zone electric field radiated by a single element in the direction $\hat{\mathbf{r}}_0$ with all other elements terminated in their generator impedance and P_{avail} is the power available to the element. Note that the quantity in Eq. (9.125) is defined in such a way that it is readily measurable. Since the radiated field is proportional to the excitation coefficient a_{mn}, and the total power available at each element is proportional to its square, the array gain is given by

$$G_{\mathrm{array}}(\hat{\mathbf{r}}_0) = \frac{4\pi r^2 \left|\mathbf{E}_{\mathrm{scan}}(\hat{\mathbf{r}}_0)\right|^2}{2\eta_0 P_{\mathrm{avail}}} \frac{\left|\sum\limits_{m,n} a_{mn}\right|^2}{\sum\limits_{m,n} |a_{mn}|^2} = N_{\mathrm{tot}}\, g_{\mathrm{scan}}(\hat{\mathbf{r}}_0)\eta_a \tag{9.126}$$

where

$$\eta_a = \frac{\left|\sum\limits_{m,n} a_{mn}\right|^2}{N_{\mathrm{tot}} \sum\limits_{m,n} |a_{mn}|^2}$$
$$\left[= \frac{\left|\sum\limits_{m} a_m\right|^2 \left|\sum\limits_{n} a_n\right|^2}{\left(M \sum\limits_{m} |a_m|^2\right)\left(N \sum\limits_{n} |a_n|^2\right)} = \eta_{xa}\eta_{ya} \qquad \text{for separable aperture distributions} \right] \tag{9.127}$$

is the *aperture efficiency*, and N_{tot} is the total number of array elements. The equivalent linear array aperture efficiencies, η_{xa}, η_{ya}, that can be used for separable planar apertures are tabulated for a number of common aperture distributions in Table 9.1.

In the following we assume, without loss of generality, that the array excitation distribution is uniform, $a_{mn} = 1$, and relate two different forms of excitation. We define the *isolated element pattern*, $\mathbf{E}_{\text{iso}}(\hat{\mathbf{r}}_0)$, as the far-field element pattern resulting when the terminals of the element are driven by a current source I_{iso}, and all other elements are open-circuited at their terminals. For many elements such as dipoles, the open-circuited elements support negligible currents and hence the isolated element pattern is essentially that of a single element with the remaining array elements removed: $\mathbf{E}_{\text{iso}}(\hat{\mathbf{r}}_0) \approx \mathbf{E}_{\text{single}}(\hat{\mathbf{r}}_0)$. Since the array radiation may be expressed as an appropriate superposition over *either* the scan element pattern or the isolated element pattern, we have

$$N_{\text{tot}}\mathbf{E}_{\text{scan}}(\hat{\mathbf{r}}_0) = N_{\text{tot}}\frac{I_{\text{scan}}(\hat{\mathbf{r}}_0)\mathbf{E}_{\text{iso}}(\hat{\mathbf{r}}_0)}{I_{\text{iso}}} \tag{9.128}$$

where $I_{\text{scan}}(\hat{\mathbf{r}}_0)/I_{\text{iso}}$ is the ratio of current at the terminals of each element under scan conditions with all elements excited to the terminal current of the isolated element.

The power available per element is given by

$$P_{\text{avail}} = \frac{|V_g|^2}{8R_g} = \frac{|I_{\text{scan}}(\hat{\mathbf{r}}_0)|^2 |Z_{\text{scan}}(\hat{\mathbf{r}}_0) + Z_g|^2}{8R_g} \tag{9.129}$$

where V_g is the generator voltage, $Z_g = R_g + jX_g$ is its internal impedance, and

$$Z_{\text{scan}}(\hat{\mathbf{r}}_0) = Z_{\text{iso}} + \sum_{\substack{m,n;\\ m=n\neq 0}} Z_{mn} e^{-jk\hat{\mathbf{r}}_0\cdot(m\mathbf{d}_1+n\mathbf{d}_2)} \tag{9.130}$$

For a large array, the scan impedance $Z_{\text{scan}}(\hat{\mathbf{r}}_0)$ is essentially that of an infinite array. In Eq. (9.130), $Z_{\text{iso}} = R_{\text{iso}} + jX_{\text{iso}}$ is the isolated element input impedance and Z_{mn} is the mutual impedance between elements separated by $m\mathbf{d}_1 + n\mathbf{d}_2$ in the array lattice.

The isolated element gain, in terms of quantities previously defined, is

$$g_{\text{iso}}(\hat{\mathbf{r}}_0) = \frac{4\pi\left(r^2 |\mathbf{E}_{\text{iso}}(\hat{\mathbf{r}}_0)|^2 / 2\eta_0\right)}{(1/2)|I_{\text{iso}}|^2 R_{\text{iso}}} \tag{9.131}$$

Combining Eqs. (9.125), (9.128), (9.129), and (9.131), we have

$$g_{\text{scan}}(\hat{\mathbf{r}}_0) = \frac{4R_g R_{\text{iso}} g_{\text{iso}}(\hat{\mathbf{r}}_0)}{|Z_{\text{scan}}(\hat{\mathbf{r}}_0) + Z_g|^2} \tag{9.132}$$

Defining a conjugate reflection coefficient,

$$\Gamma_*(\hat{\mathbf{r}}_0) = \frac{Z^*_{\text{scan}}(\hat{\mathbf{r}}_0) - Z_g}{Z_{\text{scan}}(\hat{\mathbf{r}}_0) + Z_g} \tag{9.133}$$

Eq. (9.132) can be written as

$$g_{\mathrm{scan}}(\hat{\mathbf{r}}_0) = \frac{R_{\mathrm{iso}}}{R_{\mathrm{scan}}(\hat{\mathbf{r}}_0)} g_{\mathrm{iso}}(\hat{\mathbf{r}}_0)\left(1 - \left|\Gamma_*(\hat{\mathbf{r}}_0)\right|^2\right) \tag{9.134}$$

This result shows how the scan element pattern and the isolated element pattern are linked through the element mismatch and scan element resistance. If the array is assumed matched at angle $\hat{\mathbf{r}}_{\mathrm{match}}$, i.e., $\Gamma_*\left(\hat{\mathbf{r}}_{\mathrm{match}}\right) = 0, \quad R_{\mathrm{scan}}(\hat{\mathbf{r}}_{\mathrm{match}}) = R_g$, then Eq. (9.134) may be expressed as

$$g_{\mathrm{scan}}(\hat{\mathbf{r}}_0) = \frac{R_{\mathrm{g}} g_{\mathrm{scan}}(\hat{\mathbf{r}}_{\mathrm{match}})}{R_{\mathrm{scan}}(\hat{\mathbf{r}}_0)} \frac{g_{\mathrm{iso}}(\hat{\mathbf{r}}_0)}{g_{\mathrm{iso}}(\hat{\mathbf{r}}_{\mathrm{match}})} \left(1 - \left|\Gamma_*(\hat{\mathbf{r}}_0)\right|^2\right) \tag{9.135}$$

Equation (9.134) is particularly convenient for use when the isolated element pattern is almost identical to the single element pattern, i.e., when the currents on the unexcited element essentially vanish when their terminals are *opened.* But this is not the case for some elements, such as slots and patch antennas. For such elements, *shorting* the unexcited elements terminals in an isolated element pattern renders the pattern essentially the same as the single element pattern. Repeating the derivation under the assumption that the isolated pattern is that of a singly excited element with all others terminated in *short* circuits leads to

$$g_{\mathrm{scan}}(\hat{\mathbf{r}}_0) = \frac{G_{\mathrm{iso}}}{G_{\mathrm{scan}}(\hat{\mathbf{r}}_0)} g_{\mathrm{iso}}(\hat{\mathbf{r}}_0)\left(1 - \left|\tilde{\Gamma}_*(\hat{\mathbf{r}}_0)\right|^2\right) \tag{9.136}$$

where G_{iso} and $G_{\mathrm{scan}}(\hat{\mathbf{r}}_0)$ are the isolated and scan conductances, respectively, and a new conjugate reflection coefficient is defined as

$$\tilde{\Gamma}_*\left(\hat{\mathbf{r}}_0\right) = \frac{Y_g - Y^*_{\mathrm{scan}}\left(\hat{\mathbf{r}}_0\right)}{Y_g + Y_{\mathrm{scan}}\left(\hat{\mathbf{r}}_0\right)} \tag{9.137}$$

Hansen [53] points out that the approximate result frequently appearing in the literature,

$$g_{\mathrm{scan}}(\hat{\mathbf{r}}_0) = \frac{4\pi A_{\mathrm{ref}} \cos\theta_0}{\lambda^2}\left(1 - \left|\Gamma(\hat{\mathbf{r}}_0)\right|^2\right) \tag{9.138}$$

where A_{ref} is the area of a unit cell and $\Gamma\left(\hat{\mathbf{r}}_0\right)$ is the terminal reflection coefficient, holds only when no grating lobes or higher-order feed modes are present and the elements are thin and straight.

In many arrays, dielectric slabs or substrates on ground planes are used, and these may excite a surface wave when the transverse scan wave number equals the surface wave propagation constant. The energy in the surface wave is not available for radiation, and hence the effect on the directivity is seen as a “blind spot,” i.e., a large reflection coefficient or sharp dip in the directivity pattern. Similarly, the onset of a grating lobe may also represent a reduction in element directivity and hence a large reflection coefficient.

9.5.4. Array Elements

Almost any electrically small radiator may be used in a phased array, but because of their low cost and/or ease of fabrication, typical elements include dipoles, slots, open-ended waveguides, patches, and notch antennas [53]. Because of strong mutual coupling effects in the array environment, it cannot be expected that the behavior of these elements resembles that of the corresponding isolated element. To account for these coupling effects, calculations of element active impedances are often performed for elements in *infinite* array environments. Unfortunately, array elements spaced about one-half wavelength apart often do not act as if in an infinite array environment unless they are located about 10 elements away from the array edges. Thus in a 60 × 60 array only about 44% of the elements satisfy this condition. Nevertheless, the infinite array active impedance is frequently used to predict element behavior in the phased array environment.

Wire or *printed-circuit dipoles* are often used at low frequencies. The dipole arms are usually about a half wavelength in length and located approximately a quarter wavelength above a ground plane to direct radiation into the forward direction. Bending the arms of the dipole towards the ground plane can increase the angular coverage; using thicker elements tends to increase both the bandwidth and reduce mutual coupling. Printed-circuit dipoles [54] are popular because of the ease with which they may be fabricated; two such dipoles placed orthogonal to one another and fed with a 90°phase difference can provide circular polarization.

Slots milled in the sides of waveguide walls are often used in applications where high power or accurate control of manufacturing for low side-lobe level designs is required. Unless the slots couple into another waveguide or transmission line containing phase shifters, the interelement phase is not generally controllable along the guide dimension. However, if the slots are cut into the narrow wall of the guide, a series of slotted waveguides may be stacked closely together to avoid grating lobes, and phasing between the stacked guides may be introduced to effect scanning in the orthogonal plane. Narrow wall slots cut perpendicular to the waveguide edges do not couple energy from the guide, and hence the slots must be tilted, the angle of tilt determining the degree of coupling. Alternating the tilt angle of adjacent slots alternates the sign of the coupling so that in-phase excitation results when the slots are placed $\lambda_g/2$ apart, where λ_g is the guide wavelength. The slot coupling should be small so as to minimize reflections; therefore a certain fraction of the energy at the input to the guide remains after the last slot. This remaining energy must be absorbed in a matched load, thus reducing the antenna efficiency. For arrays with tapered aperture distributions and many slots, where it is relatively easy to use slots with small coupling, it may be possible to reduce this loss to as low as 1% or 2%. In any case, slot array design is often a tradeoff between dealing with slot maximum coupling limitations and expending energy in the matched load at the end of the line.

Open-ended waveguides are often used as radiators because the waveguides can easily accommodate bulky phase shifters. The resultant structures are not only relatively easy to analyze and match but are also mechanically strong and capable of handling high power. Furthermore, they are suitable for applications requiring flush mounting. The waveguides may also be loaded with dielectrics to reduce the element size. This may be needed both to avoid grating lobes and to provide sufficient space to assemble the structure. Dielectric slabs are often used to match the guides to free space, but they may also support the propagation of surface waves along the array face.

Patch antennas consisting of a thin metallic layer bonded to a grounded substrate are popular array elements because of their ease of fabrication, light weight, low profile, and ability to conform to a planar or curved surface [54–56]. They may be either probe fed by extending the center conductor of a coaxial line through the ground plane and attaching it to the patch, by a microstrip line coupled either directly or indirectly (proximity coupled) to the patch, or aperture coupled to the feeding microstrip line below the aperture in the ground plane. Microstrip patch elements may have strong mutual coupling and narrow bandwidth.

Flared notch antennas may be thought of as slots in a ground plane that are flared to form a one-dimensional horn shape. The gradual curve allows for a broadband match for these elements [57].

9.5.5. Phased Array Feed and Beam-forming Systems

A major concern in phased array design is distributing radiated energy to or collecting received energy from the array elements. Since feed networks for transmission are generally reciprocal, they can, of course, also serve as beam-forming networks for receiving functions. But received energy from phased arrays is frequently formed into multiple, simultaneous beams and this capability is not generally needed for transmitting. For this reason, and because it is often desirable to isolate transmission and receive functions, beam-forming networks are often separated from the transmission feed network. In modern phased arrays, multiple beams are typically generated by digital rather than analog processing, and hence we concentrate on feed networks for generating single transmitting beams. Most phased arrays can be classified as parallel, series, space, or active aperture fed systems. Hybrid systems may employ one type of feed system along one dimension of a planar array and another along the other dimension.

Parallel feed systems are often called *corporate feed* systems because of their resemblance to a corporate organization chart. As illustrated in Fig. 9.43, it is desirable to employ feed networks with a branching network of hybrid junctions that absorb reflections from the elements. These prevent reflections from being reradiated by the array and resulting in pattern deterioration. Corporate feeds for microstrip patch arrays often employ only power dividers, however. Corporate feed systems are generally simpler to design than series feed systems since each element excitation can be controlled independently and all transmitter-to-element paths are equal in length and hence involve the same phase differences and path losses.

Series feed systems couple energy at periodically spaced locations along a guiding system such as a waveguide, as shown in Fig. 9.44. The electrical length between tap positions is frequency dependent and causes series fed arrays to naturally scan slightly with frequency. This tendency is purposely enhanced in frequency scanning arrays by folding the feed line between tap points to further increase the electrical length—at the expense of further lowering the system's bandwidth. Directional couplers may be used to isolate reflections that occur at the many tap locations, and element locations may be chosen to differ slightly from λ_g or $\lambda_g/2$ spacings so that reflections do not add in phase, resulting in

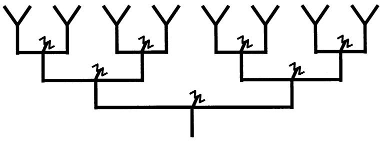


Figure 9.43 Corporate feed for an eight-element array with hybrids at each feed-line junction.

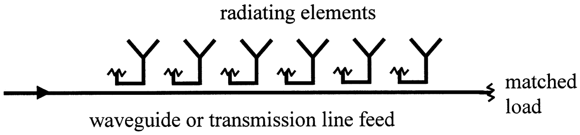


Figure 9.44 Series feed array using directional couplers.

a large input VSWR. The design of series feed systems is iterative [58], as may be seen by considering a common configuration: An array of slots milled into the narrow wall of a waveguide. The slot angle relative to the guide edges determines the degree of coupling for each slot, and there is limit to the maximum practical coupling, especially since the slot conductances should be small to minimize reflections. The aperture distribution a_n, number of slots N, and maximum allowed conductance eventually determine $r = P_{\text{load}}/P_{\text{in}}$, the fraction of the input power remaining in the guide after the last slot, which must be dissipated in a matched load at the end of the feed line. The iterative design process begins with an assumed value for r. Neglecting reflections, the fractional power dissipated by slot n is proportional to $F_n = a_n^2$. Beginning at the load end of the feed line, the design proceeds to determine all the slot conductances for the assumed r. For slots with low reflections on a lossless line, the normalized conductances can be approximately determined from [58]

$$g_n = \frac{F_n}{[1/(1-r)] \sum_{m=1}^{N} F_m - \sum_{m=1}^{n} F_m} \tag{9.139}$$

where element N is nearest the load. If a slot conductance is generated that exceeds the allowed maximum, then r must be increased and the procedure repeated; if no conductance values near the maximum allowable are found, then r may be reduced to increase the array efficiency. The final slot conductances then determine the slot angles. The approach may also be generalized to account for a line attenuation factor between elements [53]. In a series fed array, it is not unusual to dissipate 5–10% of the array input energy in the terminating load.

Lens arrays are fed by illuminating the elements of a receiving aperture by an optical feed system and then retransmitting the received signal after passing it through phase shifters that serve not only to scan the beam but also to correct for the differing path lengths taken from the optical feed to the transmitting aperture. The primary feed is often a single horn antenna or a cluster of horns forming a monopulse system. The feed system's name reflects the fact that the phase shifters act as a lens to collimate the transmitted beam—or since the system is reciprocal, to focus it on receive. To reduce the adverse effect that reflections from the lens have on the VSWR of the optical feed, the feed is often offset. An advantage of the approach is its relative simplicity and low cost. A disadvantage is the fact that antenna elements are required on both faces of the array.

Reflect arrays are similar to lens arrays, except that there is only one array face containing antenna elements. Energy from the optical feed is collected by the elements at the aperture, passes through phase shifters, is reflected by a short circuit, passes back through the phase shifters and is reradiated as a scanned beam plane at the aperture plane. Since the signal passes through the phase shifters twice, the phase shift settings are only half the total needed and the phase shifters must be of the reciprocal type. Lens and reflect arrays share not only the advantages, but also the disadvantages of other optically fed systems. One has less control over the array aperture distribution, which is controlled primarily by feed pattern. Spillover is also a concern for both feed types, and reflect arrays therefore generally employ an offset feed to reduce feed blockage.

None of the feed systems described have the power handling capabilities of *active-aperture* systems in which a transmit–receive module is associated with—and possibly contains—each array element. The element module combines transmit–receive switches, solid-state transmitter and receiver amplifiers, and phase shifters. Feed losses are thus practically eliminated, and, since phase shifters may be located at the transmitter front end, they are not required to handle high power. On receive, not only is the signal-to-noise ratio unaffected by feed or other losses, but also digital beam combining of the receiver outputs may be used to control receive patterns, including adaptively controlling sidelobe levels and null positions.

Monolithic integrated phased arrays attempt to lower the cost per element and increase the reliability and repeatability of a phased array by combining many elements, their transmitter and receive functions, as well as beamforming and array control functions, on a single package. So-called *brick* configurations combine elements common to a row and use the depth dimension to accommodate array components and feed structures. *Tile* configurations combine a number of elements in the same plane with various array components located in separate, parallel layers.

9.5.6. Electronic Beamsteering

The phase of a signal traveling through a wave guiding section of length ℓ with cutoff frequency f_c (in a TEM line, $f_c = 0$) is

$$\varphi = 2\pi\ell\sqrt{\mu\varepsilon\left(f^2 - f_c^2\right)} \tag{9.140}$$

To change this phase so as to scan an array beam, one may change

1. The line length by switching between different line lengths (diode phase shifters)
2. The permeability μ (ferrite phase shifters), or less frequently, the permittivity ε (ferroelectric and plasma phase shifters)
3. The frequency f (frequency scanning)
4. The guide cutoff frequency.

All these approaches have been used, including changing the guide cutoff frequency by mechanically changing the guide dimension, though the latter cannot be considered *electronic beam steering*.

At frequencies below S band, diode phase shifters generally have less loss than ferrite phase shifters; above S band, ferrite phase shifters are usually preferred. Phase shifters are usually digitally controlled and provide quantized values of phase shift, the resolution being determined by the smallest bit of the phase shifter. An N bit phase shifter provides phase shifts between 0°and 360° of phase in steps of $360°/2^N$. Since the desired phase variation is linear whereas achievable phase settings are discrete, the phase error due to quantization is a periodic ramp function. This periodic phase error produces *quantization* *grating lobes*, the *n*th one of whose magnitude, normalized to the peak of the array factor, is given by

$$\frac{1}{|n|2^N} \qquad n = \pm 1, \pm 2, \ldots \tag{9.141}$$

for an *N* bit phase shifter. The angular locations of these quantization lobes depend on the scan angle and number of phase bits, and most will not appear in the visible region of the pattern. Quantization errors also result in beam pointing errors and reduced gain. The "round-off error" in phase due to quantization may be randomized by introducing a known but random phase offset for each element; the resulting element phase errors are incoherent and therefore spread out the grating lobe energy as a random background quantization noise added to the designed array pattern. Thus quantization grating lobes are eliminated at the expense of raising the RMS sidelobe level [59].

Diode phase shifters generally operate as switches that change the electrical length of a signal path according to whether the switch, usually a PIN diode, is in the open or closed state. The two most commonly used types are switched and hybrid-coupled phase shifters. As illustrated in Fig. 9.45, the switched type uses diodes to switch different line lengths into the feed line to obtain the required phase shift. In a hybrid-coupled phase shifter, one bit of which is shown in Fig. 9.46, the states of the two diodes in the branch lines of the hybrid are the same and determine varying reflection point locations along the lines.

Ferrite phase shifters of the latching type operate by changing the magnetization state, and hence the insertion phase, of a ferrite toroid. These phase shifters are commonly employed in waveguide feeds, and usually consist of several cascaded sections of differing lengths—each length representing a different phase shifter bit—to obtain the desired phasing. As shown in Fig. 9.47, the magnetization is controlled by a current pulse provided by a wire threading each toroidal bit. Dielectric spacers between bits provide matching. The pulse drives the core into saturation and the remanence flux provides the magnetization required without need for a holding current. The resulting permeability change provides the phase shift. Such phase shifters are nonreciprocal and hence must be switched between the transmit and receive modes; for this reason, they cannot be used in reflect arrays since they cannot be switched quickly enough during the short time between the passage of incident and reflected pulses, whether operating in transmit or receive modes.

Frequency scanning systems do not require phase shifters but instead rely on changing the frequency to affect the electrical length, and hence the phase, between

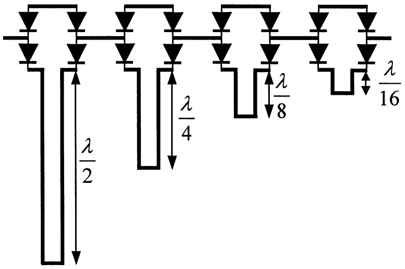


Figure 9.45 A four bit, digitally switched diode phase shifter. An *N*-bit phase shifter provides a phase increment of $360°/2^N$ and requires $4N$ PIN diodes.

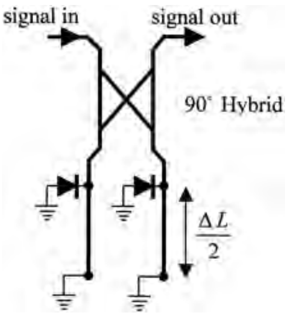


Figure 9.46 A phase bit using a hybrid coupler. Changing the diode states from conducting to nonconducting changes the round-trip path length to the reflection point by ΔL. An N bit phase shifter provides a phase increment of $360°/2^N$ and requires $2N$ PIN diodes.

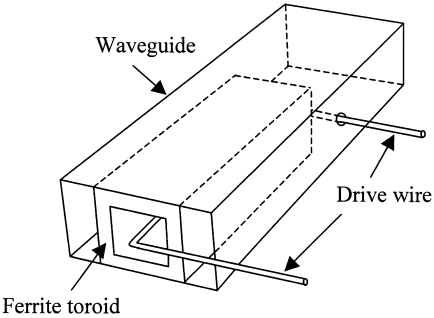


Figure 9.47 A single bit of a ferrite phase shifter in a waveguide. A current pulse in the drive wire saturates the ferrite core; reversing the direction of the pulse reverses the magnetization, thereby changing the phase shift of a signal traversing the core. Different phase-shift bit values are produced by cascading toroids of varying lengths.

elements excited by a traveling wave series feed [60]. At the expense of array bandwidth, the scan sensitivity to frequency (i.e., the change in scan angle per unit change in frequency) is enhanced by folding the feed line to increase the electrical length between elements. The resulting feeds are often called *sinuous* or *serpentine* feeds. If the main beam points at broadside at a frequency f_0, then at a frequency f the scan angle θ_0 is given by

$$\sin\theta_0 = \frac{L}{df}\left(\sqrt{f^2 - f_c^2} - \sqrt{f_0^2 - f_c^2}\right) \tag{9.142}$$

where L is the element separation distance as measured along a mean path inside a waveguide feed with guide cutoff frequency f_c and d is the actual element separation. Equation (9.142) also applies to TEM line feeds with $f_c = 0$, but, with less scan sensitivity, they are used less often. The factor L/d, which is often called the *wraparound* or *wrap-up factor*, controls the sensitivity. At broadside, reflections from element mismatches in the feed, though small, add in phase and may result in a large VSWR at the feed line input. For this reason, some frequency scan systems cover only an angular sector from a few degrees off broadside to nearly end fire. Because of their inherently low bandwidth characteristics, frequency-scanning systems are not often used in modern phased array systems.

9.5.7. Mutual Coupling

Not only is mutual coupling always present in phased arrays, but it is also responsible for most of their unique characteristics. Though it should not be neglected in array design, it is the parameter that is most difficult to obtain. Several general principles concerning mutual coupling are obvious, however [53]:

1. Coupling decreases with distance between elements.
2. Coupling between elements is strongest when their maximum radiation directions are aligned along their line of separation. For linear dipole elements in free space, this occurs, e.g., when they are parallel rather than collinear.
3. Coupling is smaller between large, highly directional elements such as horns.
4. Coupling is stronger between elements when their substrates support surface waves.

Many qualitative array effects can be discerned from the analysis of Wheeler [61] using an infinite sheet current model. The model has also been extended to include dielectric substrates [62] and demonstrates many of the most important effects of element spacing, polarization, scan angle, and substrates. A related concept is the grating-lobe series [63]. These and even more realistic models approach the problem from an infinite-array point of view. If an array is large and the taper is gradual, the interaction between central elements of the array may be approximated by those of an infinite array of like elements; the analysis is then reduced to that of a unit or reference cell of the array. Usually one determines the *scan impedance* or *scan reflection coefficient* of the reference element in the presence of an infinite number of elements similarly excited but with a fixed interelement phase shift. The calculation often requires numerical methods. In principle, integration of the resulting quantity over the phase shift variables on a unit cell of the grating lobe lattice yields the interaction between elements for a singly-excited element. This result forms the basic connection between infinite array analysis and analysis from the opposite extreme—element-by-element analysis.

Element-by-element analyses are necessary for small-to-moderate size arrays and benefit from knowledge of the interaction between isolated pairs of elements. For example, a convenient approximate form for the mutual impedance between linear dipoles in echelon is available [53,64] and is easily extended to slots in a ground plane. Limited data for mutual coupling between horns, open-ended waveguides, microstrip patches, and other array elements are also available or may be computed numerically.

Blind angles are angles at which the scan reflection coefficient is near unity or, equivalently, the scan element pattern is near zero. They may be interpreted, respectively, as angles for which higher modes cancel with the dominant mode in the element, or as angles that allow coupling to a leaky mode on the array [65]. The leaky mode is essentially a surface wave that is supported by the periodic array structure, which is leaky due to radiation from one of the space harmonics of the mode (the phase constant of the radiating space harmonic corresponds to the wave number of the array phasing at the blindness angle). If the periodic loading effect is not very strong, the phase constant of the leaky wave may be approximated as that of the corresponding surface wave. Surface waves have wave numbers $\beta_{\mathrm{sw}} > k$, and if circles of radius β_{sw} are added to and centered on the grating lobe lattice, their intersections with the visible region circle locate possible angles where blind spots can occur [66,67].

Controlling mutual coupling is of utmost importance in phased array design, and several attempts have been made to either reduce or compensate for coupling effects. Grating lobe series analysis shows that close element spacing reduces the variation of reactance since grating lobes, which primarily affect reactance, are pushed further into the invisible region. H-plane baffles placed between rows of slots or dipoles have also been used to significantly reduce impedance variations in the two principal scan planes [68]. Approaches used on open-end waveguide arrays have included the control of multimode excitation in the unit cell by dielectric loading [69] and by adding irises [70]. Other alternatives include the use of slot arrays with parasitic monopoles [71] and of dielectric sheets. Several such sheets in cascade have been used to form a wave filter placed sufficiently far in front of the array so as not to affect impedance, but to improve the scan element pattern [72].

REFERENCES

1. Garg, R.; Bhartia, P.; Bahl, I.; Ittipiboon, A. *Microstrip Antenna Design Handbook*; Artech House: Norwood, MA, 2000.
2. Lee, K.F. Ed. *Advances in Microstrip and Printed Antennas*; Wiley: New York, 1997.
3. Pozar, D.M.; Schaubert, D.H. *Microstrip Antennas: The Analysis and Design of Microstrip Antennas and Arrays*; IEEE Press, Piscataway, NJ, 1995.
4. Gardiol, F.E. *Broadband Patch Antennas*; Artech House: Norwood, MA, 1995.
5. Jackson, D.R.; Williams, J.T. A comparison of CAD models for radiation from rectangular microstrip patches. Int. J Microwave Millimeter-Wave CAD **April 1991**, *1* (2), 236–248.
6. Jackson, D.R.; Long, S.A.; Williams, J.T.; Davis, V.B. Computer-aided design of rectangular microstrip antennas. In: *Advances in Microstrip and Printed Antennas*; Lee, K.F. Ed.; Wiley, New York, 1997; Chap. 5.
7. Pozar, D.M. A reciprocity method of analysis for printed slot and slot-coupled microstrip antennas, IEEE Trans. Antennas Propagation **Dec. 1986**, *AP-34*, 1439–1446.
8. Kumar, G.; Ray, K.P. Broadband Microstrip Antennas; Artech House, Norwood, MA, 2003.
9. Pues, H.; Van de Capelle, A. Accurate transmission-line model for the rectangular microstrip antenna, Proc. IEEE, Vol. 131, Pt. H, No. 6, pp. 334–340, Dec. 1984.
10. Richards, W.F.; Lo, Y.T.; Harrison, D.D. An improved theory of microstrip antennas with applications, IEEE Trans. Antennas Propagation **Jan. 1981**, *AP-29*, 38–46.
11. Pozar, D.M. Input impedance and mutual coupling of rectangular microstrip antennas. IEEE Trans. Antennas Propagation **Nov. 1982**, *AP-30*, 1191–1196.
12. Prior, C.J.; Hall, P.S. Microstrip disk antenna with short-circuited annular ring. Electronics Lett. **1985**, *21*, 719–721.
13. Guo, Y.-X.; Mak, C.-L.; Luk, K.-M.; Lee, K.-F. Analysis and design of L-probe proximity fed patch antennas, IEEE Trans. Antennas Propagation **Feb. 2001**, *AP-49*, 145–149.
14. Ghorbani, K.; Waterhouse, R.B. Ultrabroadband printed (UBP) antenna, IEEE Trans. Antennas Propagation **Dec. 2002**, *AP-50*, 1697–1705.
15. Kumar, G.; Gupta, K.C. Nonradiating edges and four edges gap coupled multiple resonator broadband microstrip antennas. IEEE Trans. Antennas Propagation **Feb. 1985**, *AP-33*, 173–178.
16. Weigand, S.; Huff, G.H.; Pan, K.H.; Bernhard, J.T. Analysis and design of broadband single-layer rectangular U-slot microstrip patch antennas. IEEE Trans. Antennas Propagation **March 2003**, *51*, 457–468.

17. Jackson, D.R.; Williams, J.T.; Bhattacharyya, A.K.; Smith, R.; Buchheit, S.J.; Long, S.A. Microstrip patch designs that do not excite surface waves. IEEE Trans. Antennas Propagation **Aug. 1993**, *41*, 1026–1037.
18. Kraus, J.D. The helical antenna, In *Antennas*; McGraw Hill: NY, 1950; Chap. 7.
19. Adams, A.T.; Greenough, R.F.; Walkenburg, R.K.; Mendelovicz, A.; Lumjiak, C. The quadrifilar helix antenna. IEEE Trans. Antennas Propagation *22* (3), 173–178, 1074.
20. King, H.E.; Wong, J.L. Helical antennas, In *Antenna Engineering Handbook*; 3rd Ed.; Johnson, R.C. Ed.; McGraw-Hill: NY, 1993.
21. Elliott, R.S. *Antenna Theory and Design*; Prentice Hall: Englewood-Cliffs, NJ, 1981.
22. Wong, J.L.; King, H.E. Broadband quasi-tapered helical antennas. IEEE Trans. Antennas Propagation **1979**, *27* (1), 72–78.
23. Rumsey, V.H. *Frequency Independent Antennas*; Academic Press: NY, 1966.
24. Corzine, R.G.; Mosko, J.A. *Four Arm Spiral Antennas*; Artech House: Norwood, MA, 1990.
25. Stutzman, W.L.; Thiele, G.A. *Antenna Theory and Design*; Wiley, NY, 1981.
26. DuHamel, R.H.; Scherer, J.P. Frequency independent antennas, In *Antenna Engineering Handbook*; 3rd Ed.; Johnson, R.C. Ed.; McGraw-Hill: NY, 1993.
27. Dyson, J.D. The equiangular spiral antenna. IRE Trans. Antennas Propagation **1959**, *7* (2), 181–187.
28. Wang, J.J.H.; Tripp, V.K. Design of multi-octave spiral-mode microstrip antennas. IEEE Trans. Antennas Propagation **1991**, *39* (3), 332–335.
29. Champagne, N.J.; Williams, J.T.; Wilton, D.R. Resistively loaded printed spiral antennas. Electromagnetics **1994**, *14* (3–4), 363–395.
30. Dyson, J.D. The characteristics and design of the conical log-spiral antenna. IEEE Trans. Antennas Propagation **1965**, *13* (7), 488–499.
31. Yeh, Y.S.; Mei, K.K. Theory of conical equiangular-spiral antenna—part II: current distributions and input impedances. IEEE Trans. Antennas Propagation **1968**, *16* (1), 14–21.
32. Walter, C.H. *Traveling Wave Antennas*; McGraw-Hill: New York, 1965.
33. Hessel, A. General characteristics of traveling-wave antennas, In *Antenna Theory* Part 2; Colin, R.E.; Zucher, F.J., Eds.; McGraw-Hill: New York, 1969; Chap 19.
34. Tamir, T. Leaky-wave antennas, In *Antenna Theory*, Part 2; Colin, R.E.; Zucher, F.J. Eds.; McGraw-Hill: New York, Chap 20, 1969.
35. Oliner, A.A. Leaky-wave antennas, In *Antenna Engineering Handbook*; 3rd ed.; Hansen, R.C. Ed.; McGraw-Hill: New York, Chap 10, 1993.
36. Tamir, T.; Oliner, A.A. Guided complex waves, part I: fields at an interface. Proc. Inst. Elec. Eng., Vol. 110, pp. 310–324, Feb. 1963.
37. Tamir, T.; Oliner, A.A.; Guided complex waves, part II: relation to radiation patterns, Proc. Inst. Elec. Eng., vol. 110, pp. 325–334, Feb. 1963.
38. Balanis, C.A. *Antenna Theory*; Wiley: New York, 1997.
39. Goldstone, L.O.; Oliner, A.a. Leaky-wave antennas I: rectangular waveguide. IRE Trans. Antennas Propagation **Oct. 1959**, *AP-7*, 307–319.
40. Harrington, R.F. *Time Harmonic Electromagnetic Fields*; McGraw-Hill: New York, 1963.
41. Guglielmi M.; Boccalone, G. A novel theory for dielectric-inset waveguide leaky-wave antennas, IEEE Trans. Antennas Propagation **April 1991**, *AP-39*, 497–504.
42. Guglielmi, M.; Jackson, D.R. Broadside radiation from periodic leaky-wave antennas. IEEE Trans. Antennas Propagation **Jan. 1993**, *41*, 31–37.
43. Zhao, T.; Jackson, D.R.; Williams, J.T. Radiation characteristics of a 2D periodic slot leaky-wave antenna. IEEE AP-S/URSI Intl. Symp. Digest, pp. 482-485, San Antonio, TX, June 16–21, 2002.
44. Jackson, D.R.; Oliner, A.A. A leaky-wave analysis of the high-gain printed antenna configuration. IEEE Trans. Antennas Propagation **July 1988**, *36*, 905–910.
45. Love, A.W. Ed., *Electromagnetic Horn Antennas*; IEEE Press: NY, 1976.
46. Love, A.W. Horn antennas, In *Antenna Engineering Handbook*; 3rd Ed.; Johnson, R.C. Ed.; McGraw-Hill: NY, 1993.

47. Clarricoats, P.J.B.; Poulton, G.T. High-efficiency microwave reflector antennas—a review," Proc. IEEE, Vol. 65, No. 10, pp. 1470–1504, 1977.
48. Kelleher, K.S.; Hyde, G. Reflector antennas, In *Antenna Engineering Handbook*; 3rd Ed.; Johnson, R.C. Ed.; McGraw-Hill: NY, 1993.
49. Sharpe, E.D. "A triangular arrangement of planar-array elements that reduces the number needed." IEEE Trans. Antennas Propagation **Mar. 1961**, *AP-9*, 126–129.
50. Silver, S. *Microwave Antenna Theory and Design, M.I.T. Radiation Laboratory Series*; McGraw-Hill: New York, 1949; Vol. 12.
51. Skolnik, M.I. *Introduction to Radar Systems*; McGraw-Hill: New York, 2001.
52. Taylor, T.T. Design of line source antennas for narrow beamwidth and low side lobes. IRE Trans. **1955**, *AP-7*, 16–28.
53. Hansen, R.C. *Phased Array Antennas*; Wiley: New York, 1998.
54. Carver, K.R.; Mink, J.W. Microstrip antenna technology. IEEE Trans. Antennas Propagation **Jan. 1981**, *AP-29*, 2–24.
55. Liu, C.; Hessel, A.; Shmoys, J. Performance of probe-fed microstrip-patch element phased arrays. IEEE Trans. Antennas Propagation **Nov. 1988**, *AP-36*, 1501–1509.
56. Pozar, D.M.; Schaubert, D.H. Analysis of an infinite array of rectangular microstrip patches with idealized probe feeds. IEEE Trans. Antennas Propagation **1984**, *AP-32*, 1101–1107.
57. Mailloux, R.J. *Phased Array Antenna Handbook*; Artech House: Norwood, MA, 1994.
58. Dion, A. Nonresonant slotted arrays. IRE Trans. Antennas Propagation **Oct. 1958**, *AP-6*, 360–365.
59. Buck, G.J. Quantization and reflection lobe dispersion, In *Phased Array Antennas*; Oliner, A.A.; Knittel, G.H. Eds.; Artech House: Norwood, MA, 1972.
60. Ajioka, J.S. Frequency scan antennas, In *Antenna Engineering Handbook*; 3rd Ed.; Johnson, R.C. Ed.; Mc-Graw Hill: New York, 1993.
61. Wheeler, H.A. Simple relations derived from a phased-array antenna made of an infinite current sheet. IEEE Trans. Antennas Propagation **Jul. 1965**, *AP-13*, 506–514.
62. Pozar, D.M. General relations for a phased array of printed antennas derived from infinite current sheets. IEEE Trans. Antennas Propagation **May 1985**, *AP-33*, 498–504.
63. Wheeler, H.A. The grating-lobe series for the impedance variation in a planar phased-array antenna. IEEE Trans. Antennas Propagation **Nov. 1966**, *AP-14*, 707–714.
64. Hansen, R.C.; Brunner, G. Dipole mutual impedance for design of slot arrays. Microwave J. **Dec. 1979**, *22*, 54–56.
65. Knittel, G.H.; Hessel, A.; Oliner, A.A. Element pattern nulls in phased arrays and their relation to guided waves, Proc. IEEE, Vol. 56, pp. 1822–1836, Nov. 1968.
66. Frazita, R.F. Surface-wave behavior of a phased array analyzed by the grating-lobe series. IEEE Trans. Antennas Propagation **Nov. 1967**, *AP-15*, 823–824.
67. Pozar, D.M.; Schaubert, D.H. Scan blindness in infinite phased arrays of printed dipoles. IEEE Trans. Antennas Propagation. June 1984, *AP-32*, 602–610.
68. Edelberg, S.; Oliner, A.A. Mutual coupling effects in large antenna arrays: part I-slot arrays. IRE Trans. **May 1960**, *AP-8*, 286–297.
69. Tsandoulas, G.N.; Knittel, G.H. The analysis and design of dual-polarization square-waveguide phased arrays. IEEE Trans. Antennas Propagation **Nov. 1973**, *AP-21*, 796–808.
70. Lee, S.W.; Jones, W.R. On the suppression of radiation nulls and broadband impedance matching of rectangular waveguide phased arrays. IEEE Trans. Antennas Propagation **Jan. 1971**, *AP-19*, 41–51.
71. Clavin, A.; Huebner, D.A.; Kilburg, F.J. An improved element for use in array antennas. IEEE Trans. Antennas Propagation **July 1974**, *AP-22*, 521–526.
72. Munk, B.A.; Kornbau, T.W.; Fulton, R.D. Scan independent phased arrays. Radio Sci. **Nov.–Dec. 1979**, *14*, 979–990.

10
Electromagnetic Compatibility

Christos Christopoulos
University of Nottingham,
Nottingham, England

10.1. SIGNIFICANCE OF EMC TO MODERN ENGINEERING PRACTICE

The term *electromagnetic compatibility* (EMC) stands for the branch of engineering dealing with the analysis and design of systems that are compatible with their electromagnetic environment. It may be claimed that there are two kinds of engineers—those who have EMC problems and those who will soon have them. This statement illustrates the impact of EMC on modern engineering practice.

Interference problems are not new. Since the beginning of radio engineers noticed the difficulties encountered when trying to make ground connections to the chassis of different systems and the onset of whistling noise attributed to atmospheric conditions. All these are manifestations of electromagnetic interference (EMI) and demonstrate the need to design systems which are compatible with their electromagnetic environment.

There are two aspects to EMC. First, systems must be designed so that they do not emit significant amounts of unintended electromagnetic (EM) radiation into their environment. This aspect is described as *emission* and may be divided in turn into *conducted* and *radiated* emission. Second, systems must be capable of operating without malfunction in their intended environment. This aspect is described as *immunity*, or alternatively, as *susceptibility*. Hence, all EMC analysis and design techniques aim to address these two aspects using circuit-based and field-based experimental, analytical, and numerical techniques.

It is important to realize why EMC has become so important in recent years. As is usual in such cases, there are several reasons:

Modern design relies increasingly on the processing of digital signals, i.e., signals of a trapezoidal shape with very short rise and fall times. This gives them a very broad frequency spectrum and thus they are more likely to interfere with other systems.

Most modern designs rely on clocked circuits with clock frequencies exceeding 2 GHz. This implies very short transition times (see above) and also the presence of several harmonics well into the microwave region. Such a broad spectrum makes it inevitable that some system resonances will be excited forming efficient antennas for radiating EM energy into the environment and coupling to other systems.

Voltage levels for switching operations have steadily decreased over the years from hundreds of volts (vacuum tubes) to a few volts in modern solid-state devices. This makes systems more susceptible to even small levels of interference.

We make a much greater use of the EM spectrum as, for instance, with mobile phones and other communication services.

Equipment is increasingly constructed using small cabinets made out of various plastics and composites in contrast to traditional design, which used metal (a good conductor) as the primary constructional material. This trend meets the need for lighter, cheaper, and more aesthetically pleasing products. However, poor conductors are not good shields for EM signals, thus exacerbating emission and susceptibility problems.

Miniaturization is the order of the day, as smaller, lighter mobile systems are required. This means close proximity between circuits and thus greater risk of intrasystem interference (cross talk).

We rely increasingly on electronics to implement safety critical functions. Examples are, antilock break systems for cars, fly-by-wire aircraft, etc. It is, therefore, imperative that such circuits be substantially immune to EMI and hence malfunction.

We might add here military systems that use electronics substantially and are continuously exposed to very hostile EM environments either naturally occurring (e.g., lightning) or by deliberate enemy action (e.g., jamming).

These points illustrate the engineering need to design electromagnetically compatible systems. International standardization bodies have recognized for many years the need to define standards and procedures for the certification of systems meeting EMC requirements. The technical advances outlined above have given a new impetus to this work and have seen the introduction of international EMC standards covering most aspects of interference control and design. These are the responsibility of various national standard bodies and are overseen by the International Electrotechnical Commission (IEC) [1].

The impact of EMC is thus multifaceted. The existence of EMC design procedures which adhere to international standards, ensures that goods may be freely moved between states and customers have a reasonable expectation of a well engineered, reliable and safe product. However, meeting EMC specifications is not cost free. The designer needs to understand how electromagnetic interactions affect performance, and implement cost effective remedies. A major difficulty in doing this is the inherent complexity of EM phenomena and the lack of suitably qualified personnel to do this work. This is a consequence of the fact that for several decades most engineers focused on digital design and software developments with little exposure to EM concepts and radio-frequency (RF) design. In this chapter we aim to describe how EM concepts impact on practical design for EMC and thus assist engineers wishing to work in this exciting area. It is also pointed out that modern high-speed electronics have to cope in addition to EMC also with signal integrity (SI) issues. The latter is primarily concerned with the propagation of fast signals in the compact nonuniform environment of a typical multilayer printed-circuit board (PCB). At high clock rates the distinction between EMC and SI issues is somewhat tenuous as the two are intricately connected. Thus most material presented in this chapter is also relevant to SI.

We emphasize predictive EMC techniques rather than routine testing and certification as the art in EMC is to ensure, by proper design, that systems will meet specifications without the need for extensive reengineering and modification. It is in this area that electromagnetics has a major impact to make. It is estimated that up to10% of the cost of a new design is related to EMC issues. This proportion can be considerably higher if proper EM design for EMC has not been considered at the start of the design process. The interested reader can access a number of more extensive books on EMC and SI. The EMC topic is also taken up in my own *Principles and Techniques of Electromagnetic Compatibility* [2]. A general text on SI is Ref. 3. Other references are given in the following sections.

We start in the following section with a brief survey of useful concepts from EM field theory, circuits, and signals as are adapted for use in EMC studies. There follow sections on coupling mechanisms, practical engineering remedies to control EMI and EMC standards and testing. We conclude with an introduction of some new concepts and problems which are set to dominate EMC studies in the years to come.

10.2. USEFUL CONCEPTS AND TECHNIQUES FROM ELECTROMAGNETICS, SIGNALS, AND CIRCUITS

In this section we summarize useful concepts for EMC. Most readers will be familiar with this material but may find it still useful as it is presented in a way that is useful to the EMC engineer.

10.2.1. Elements of EM Field Theory

Most EMC standards and specifications are expressed in terms of the electric field. There are cases where the magnetic field is the primary consideration (e.g., shielding at low frequencies) but these are the minority. In emission studies, the electric field strength is specified at a certain distance from the equipment under test (EUT). These distances are typically, 1 m (for some military specifications), 3 m, 10 m, and 30 m. Measurements or calculations at one distance are then extrapolated to estimate the field at another distance, assuming far-field conditions. This implies an extrapolation law of $1/r$, where r is the distance. This is only accurate if true far-field conditions are established and this can only be guaranteed if the extrapolation is done from estimates of the field taken at least a wavelength away from the EUT. This is not always the case, but the practice is still followed, thus introducing considerable errors in field estimates.

In EMC work electric fields are normally expressed in decibels relative to some reference. A commonly employed reference is 1 μV/m. Thus an electric field E in V/m can be expressed in dBμV/m

$$E\ \mathrm{dB\mu V/m} = 20\log\left(\frac{E}{1\times 10^{-6}}\right) \tag{10.1}$$

Thus, an electric field of 10 mV/m is equal to 80 dBμV/m. Typical emission limits specified in various standards range between 30 and 55 dBμV/m. Similar principles apply when the magnetic field H is expressed normally to a reference of 1 μA/m.

A lot of reliance is placed in EMC analysis on quasistatic concepts. This is due to the desire of designers to stay with familiar circuit concepts and also to the undeniable complexity of working with EM fields at high frequencies. Strictly speaking, quasistatic concepts apply when the physical size of the system D is much smaller than the shortest wavelength of interest $D \ll \lambda$. This is often the case but care must be taken before automatic and indiscriminate use of this assumption is made. Assuming that the quasistatic assumption is valid, we can then talk about the capacitance and inductance of systems and have a ready-made approach for the calculation of their values. Important in many EMC calculations is therefore the extraction of the L and C parameters of systems so that a circuit analysis can follow. This is, in general, much simpler than a full-field analysis and is to be preferred provided accuracy does not suffer.

There are many ways to extract parameters using a variety of computational electromagnetic (CEM) techniques. Whenever an analytical solution is not available [4], CEM techniques such as the finite element method (FEM), the method of moments (MoM), finite-difference time-domain (FDTD) method, and the transmission-line modeling (TLM) method may be employed [5–8]. All such calculations proceed as follows.

A model of the system is established normally in two-dimensions (2D) to obtain the per unit length capacitance. The systems is electrically charged and the resulting electric field is then obtained. The voltage difference is calculated by integration and the capacitance is then finally obtained by dividing charge by the voltage difference. If for instance the parameters of a microstrip line are required, two calculations of the capacitance are done. First with the substrate present and then with the substrate replaced by air. The second calculation is used to obtain the inductance from the formula $L = 1/(c^2 C_0)$, where c is the speed of light ($= 3 \times 10^8$ m/s) and C_0 is the capacitance obtained with the substrate replaced by air. This approach is justified by the fact that the substrate does not normally affect magnetic properties. If this is not the case (the substrate has relative magnetic permeability other than one), then a separate calculation for L must be done by injecting current I into the system, calculating the magnetic flux Φ linked, and thus the inductance $L = \Phi/I$. It is emphasized again that when quasistatic conditions do not apply, the concept of capacitance is problematic as the calculation of voltage is not unique (depends on the path of integration). Similar considerations apply to inductance. At high frequencies, therefore, where the wavelength gets comparable with the size of systems, full-field solutions are normally necessary. This increases complexity and requires sophisticated modeling and computational capabilities. The reader is referred to Ref. 2 for a more complete discussion of the relationship between circuit and field concepts.

In EMC work it is important to grasp that what is crucial is not so much the visible circuit but stray, parasitic, components. This is where an appreciation of field concepts can assist in interpretation and estimation of relevant parameters and interactions. The reason that parasitic components are so important is that they affect significantly the flow of common mode currents. This is explained in more detail further on in this section. Particular difficulties in EMC studies are encountered at high frequencies. Here the quasistatic approximation fails and full-field concepts must be employed. At high frequencies, fields are generally not guided by conductors and spread out over considerable distances. Before we focus on high-frequency problems we state more clearly the range of applicability of the various models used to understand electrical phenomena. Generally, electrical problems fit into three regimes:

1. When the size of a system is smaller than a wavelength in all three dimensions, then it may be adequately represented by lumped component equivalent circuits. Solution techniques are those used in circuit analysis. This is the simplest case, and it is preferred whenever possible.

2. When a system is smaller than a wavelength in two dimensions and comparable or larger to a wavelength in the third dimension, then the techniques of transmission-line analysis can be used. These are based on distributed parameter equivalent circuits.
3. When a system is electrically large in all three dimensions, then full-field calculations must be employed based on the full set of Maxwell's equations.

Clearly, the last case offers the most general solution, and it is the most complex to deal with. In this case it is normally necessary to employ numerical techniques such as those described in [5–8].

We focus here on some of the most useful EM concepts that are necessary to understand the high-frequency behavior of systems. At high frequencies EM energy is transported in a wavelike manner. This is done either in the form of guided waves as in a transmission line or in free space as from a radiating antenna. Taking for simplicity the case of wave propagation in one dimension z, then the electric field has only a y component and the magnetic field an x component. The electric field behavior is described by the wave equation

$$\frac{\partial^2 E_y}{\partial x^2} = \frac{1}{u^2}\frac{\partial^2 E_y}{\partial t^2} \tag{10.2}$$

where, u is the velocity of propagation in the medium concerned,

$$u = \frac{1}{\sqrt{\mu\varepsilon}} \tag{10.3}$$

In the case of propagation in free space, u is equal to the speed of light. An identical equation describes magnetic field behavior. Transport of EM energy after a few wavelengths away from radiating structures, such as the various interconnects, wiring, etc., in electrical systems takes place in accordance to Eq. (10.2). In the so-called far field E and H are transverse to each other and their magnitudes are related by the expression,

$$H = \frac{E}{\eta} \tag{10.4}$$

where, η is the intrinsic impedance of the medium. In the case of free space

$$\eta = \sqrt{\frac{\mu_0}{\varepsilon_0}} = 377\,\Omega \tag{10.5}$$

In EMC calculations it is customary to calculate the magnetic field from the electric field using Eqs. (10.4) and (10.5). This is however only accurate if plane wave conditions apply and this is generally true at a distance exceeding approximately a wavelength away from the radiator. In the near field, the field retains some of the character of the radiating structure that produced it. If the radiator is in the form of a dipole, where voltage differences are accentuated, then the electric field is higher than would be expected for plane wave conditions and the wave impedance is larger than 377 Ω. If however, the radiating structure is in the form of a loop, where currents are accentuated, then the impedance of the medium is smaller than 377 Ω, and the magnetic field predominates. In either case, in the far field the wave impedance settles at 377 Ω.

For a short dipole, the magnitude of the wave impedance as a function of the distance r away from it is given by the formula

$$|Z_w| = \eta \frac{\sqrt{1 + 1/(\beta r)^6}}{1 + 1/(\beta r)^2} \tag{10.6}$$

where, $\beta r = 2\pi r/\lambda$. It is clear that as $r \gg \lambda$, the wave impedance tends to η.

As an example, we give here the formulas for the field near a very short (Hertzian) dipole.

The configuration is shown in Fig. 10.1 with the components in spherical coordinates.

$$\begin{aligned}
E_\vartheta &= \frac{j\omega\mu}{4\pi}(I\Delta l)\frac{e^{-j\beta r}}{r}\sin\vartheta\left[1 + \frac{1}{j\beta r} + \frac{1}{(j\beta r)^2}\right]\\
E_r &= \frac{j\omega\mu}{2\pi}(I\Delta l)\frac{e^{-j\beta r}}{r}\cos\vartheta\left[\frac{1}{j\beta r} + \frac{1}{(j\beta r)^2}\right]\\
E_\varphi &= 0\\
H_r &= H_\vartheta = 0\\
H_\varphi &= \frac{(I\Delta l)}{4\pi}\frac{e^{-j\beta r}}{r}\sin\vartheta\left[j\beta + \frac{1}{r}\right]
\end{aligned} \tag{10.7}$$

where, $\beta = 2\pi/\lambda$ is the phase constant, I is the current, and Δl is the length of the short dipole. It is clear from these formulas that the field varies with the distance r from the dipole in a complex manner. This is particularly true when r is small (near field) when all terms in the right-hand side of Eq. (10.7) are of significant magnitude. In the far field ($r \gg \lambda$), the field simplifies significantly,

$$\begin{aligned}
E_\vartheta &\simeq j\eta\frac{\beta(I\Delta l)}{4\pi}\frac{e^{-j\beta r}}{r}\sin\vartheta\\
H_\varphi &\simeq j\frac{\beta(I\Delta l)}{4\pi}\frac{e^{-j\beta r}}{r}\sin\vartheta
\end{aligned} \tag{10.8}$$

We notice that in far field only two components of the field remain which are orthogonal to each other, and they both decay as $1/r$. This is characteristic of a radiation field.

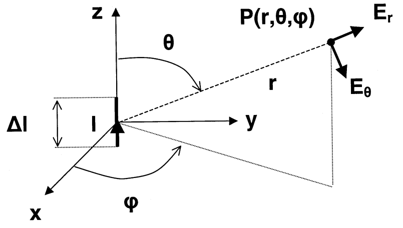


Figure 10.1 Coordinates used for calculating the field components of a very short dipole.

In complex systems with numerous radiating wire segments field behavior is very complex and it can only be studied in detail with powerful modeling tools. Similar formulas apply for short loops. Formulas for radiation from antennas may be found in Ref. 9 and other similar texts on antenna theory.

10.2.2. Treatment of Signals and Sources

The study and characterization of the EMC behavior of systems require an understanding of the nature of electrical signals encountered in engineering practice. One can classify signals in several ways depending on the criterion selected.

Many signals employed during normal engineering work are deterministic in nature, that is, their evolution in time can be precisely predicted. However, in many cases of signals with noise, we cannot predict precisely their time evolution. We call these signals random or stochastic. We can however make precise statements about them which are true in the statistical sense. The study of random signals requires sophisticated tools which are beyond the scope of this chapter. For a brief introduction see Ref. 2 and for a fuller treatment see Ref. 10. We will limit our discussion here to deterministic signals.

Some signals consist of essentially a single frequency (monochromatic or narrowband). A signal that occupies a very narrow band in the frequency spectrum, persists for a long period in time. A typical example is a steady-state sinusoidal signal. Other signals occupy a wide band of frequencies and therefore persist for relatively short periods in time. Typical examples are pulses of the kind found in digital circuits.

Whatever the nature of the signal, we can represent it as the weighted sum of a number of basis functions. A very popular choice of basis functions are harmonic functions, leading to representation of signals in terms of Fourier components [11]. For a periodic signal we obtain a Fourier series and for an aperiodic signal a Fourier transform. As an example, we give the Fourier series components of a signal of great engineering importance—the pulse train shown in Fig.10.2. For this signal the period is T, the duty cycle is τ/T and the transition time (rise and fall time) is τ_r. This trapezoidal-shaped pulse is a good representation of pulses used in digital circuits. The Fourier spectrum of this signal is given below:

$$|A_n| = 2V_0 \frac{\tau - \tau_r}{T} \left| \frac{\sin[\pi n(\tau - \tau_r)/T]}{\pi n(\tau - \tau_r)/T} \right| \left| \frac{\sin(\pi n \tau_r/T)}{\pi n \tau_r/T} \right| \tag{10.9}$$

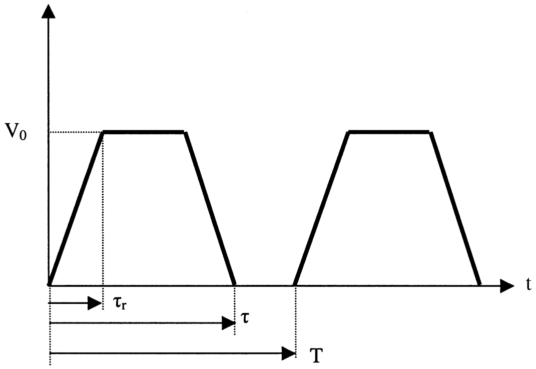


Figure 10.2 Typical trapezoidal pulse waveform.

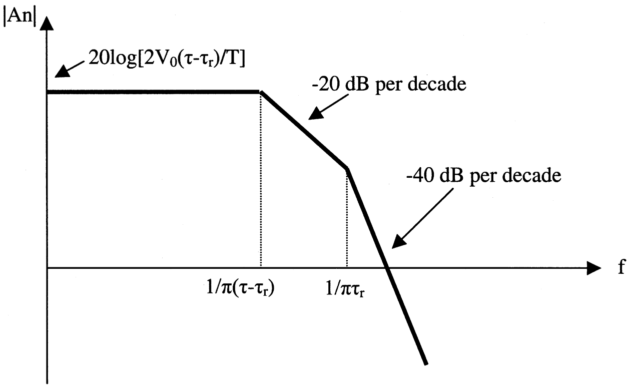


Figure 10.3 Envelope of the amplitude spectrum of the waveform in Fig. 10.2.

where, $n = 0, 1, 2, \ldots$, and $A_0 = 2V_0(\tau - \tau_r)/T$. Equation (10.9) represents a spectrum of frequencies, all multiples if $1/T$, with amplitudes which are modulated by the (sin x)/x functions. Three terms may be distinguished: a constant term $2V_0(\tau - \tau_r)/T$ independent of frequency, a term of magnitude 1 up to frequency $1/[\pi(\tau - \tau_r)]$ thereafter decreasing by 20 dB per decade of frequency, and term of magnitude 1 up to frequency $1/\pi\tau_r$ thereafter decreasing by 20 dB per decade. The envelope of the amplitude spectrum for a trapezoidal pulse train is shown in Fig. 10.3. The shorter the transition time, the higher the frequency at which the amplitude spectrum starts to decline. Short rise times imply a very wide spectrum of frequencies.

It is customary in EMC to study the behavior of systems as a function of frequency. However, increasingly, other techniques are used to speed up experimentation and analysis where a system is excited by short pulses. The former case is referred to as analysis in the frequency domain (FD) and the latter as analysis in the time domain (TD). The two domains are related by the Fourier transform as explained further in the next subsection.

Commonly encountered sources of EMI are characterized as far as possible using standard signal waveforms [2]. Amongst naturally occuring EMI sources most prominent is lightning [12,13] because of its wide spectrum and wide geographical coverage. A general background noise level due to a variety of cosmic sources exists, details of which may be found in Ref. 14. There is also a range of man-made sources including radio transmitters [15,16], electroheat equipment [17], digital circuits and equipment of all kinds [18], switched-mode power supplies and electronic drives [19,20], electrostatic discharge [21], and for military systems NEMP [22,23]. A survey of general background levels of man-made noise may be found in Ref. 24. Reference 25 describes the methodology to be used to establish the nature and severity of the EM environment on any particular site.

10.2.3. Circuit Analysis for EMC

As already mentioned lumped circuit component representation of systems and hence circuit analysis techniques are used whenever possible in EMC. For the serious student of EMC familiarity with the relationship of circuit and field concepts is very useful. As soon as it has been established that a circuit representation of a system is adequate, normal circuit analysis techniques may be employed [26]. In general circuits can be studied in two ways.

First, the frequency response may be obtained. The source signal is analysed into its Fourier components $V_{\text{in}}(j\omega)$, and the output is then obtained from the frequency response or transfer function $H(j\omega)$ of the circuit,

$$V_{\text{out}}(j\omega) = H(j\omega)V_{\text{in}}(j\omega) \tag{10.10}$$

Full-field analysis in the FD is based on the same principles, but the transfer function is much more complex and often cannot be formulated in a closed form.

Second, the problem may be formulated in the time domain whereby the system is characterized by its response to an impulse, by the so-called *impulse response* $h(t)$. The response to any source signal $v(t)$ is then given by the convolution integral,

$$v_{\text{out}}(t) = \int_{\infty}^{\infty} v_{\text{in}}(\tau)h(t-\tau)\,d\tau \tag{10.11}$$

Full-field analysis in the TD is done in the same way but the impulse response is a much more complex function which often cannot be formulated in a closed form. In linear systems Eqs. (10.10) and (10.11) are equivalent formulations as the two response functions $H(j\omega)$ and $h(t)$ are Fourier transform pairs. However, in nonlinear systems, where the principle of superposition does not apply, only the time domain approach can be employed. Full-field solvers broadly reflect these limitations.

Simple nonlinear circuits are used in EMC to implement various detector functions (e.g., peak and quasi-peak detectors). For a discussion of detector functions, see Refs. 27 and 28.

10.3. IMPORTANT COUPLING MECHANISMS IN EMC

In every EMC problem we may distinguish three parts as shown in Fig. 10.4. These are the source of EMI, the victim of EMI and a coupling path. If at least one of these three parts is missing then we do not have an EMC problem. In the previous section we have discussed some of the sources and circuits which may be victims to interference. In the present section we focus on the coupling mechanisms responsible for EMI breaching the gap between source and victim. A comprehensive treatment of this extensive subject is beyond the scope of this chapter. The interested reader is referred to comprehensive texts on EMC such as [2,29–31]. We will however present here the essential principles of EM coupling.

10.3.1. Penetration Through Materials

In many systems the outer skin (e.g., aircraft) or enclosure (e.g., equipment cabinet) forms part of an EM shield which contributes to the reduction of emission and susceptibility

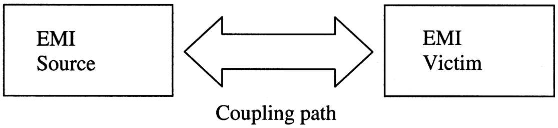


Figure 10.4 Source, coupling path, and victim of EMI.

problems. A perfectly conducting shield without apertures or penetrations would be an ideal shield for all but low-frequency magnetic fields. However such an ideal is difficult to approach in practice. Invariably, shields are not perfectly conducting and have several openings and through wire connections. In this subsection, we focus on penetration through the walls of a shield due to its finite electrical conductivity.

In this and other shielding problems it is important to use the concept of *shielding effectiveness* (SE). SE is defined as the ratio in dB of the field without and with the shield.

$$\mathrm{SE} = 20\log\left|\frac{E_0}{E_t}\right| \tag{10.12}$$

A similar expression is used for the magnetic shielding effectiveness.

The SE of canonical shapes such as spheres, cylinders made out of various materials may be calculated analytically. Of particular relevance in practical applications is the SE due to the material itself at low frequencies and particularly to the magnetic field. Taking as an example a very long cylinder of inner radius a and wall thickness D, the ratio of incident to transmitted longitudinal magnetic field (low-frequency, displacement current neglected) is given by the formula [2,32–34],

$$\frac{H_i}{H_t} = \cosh\gamma D + \frac{\gamma a}{2\mu_r}\sinh\gamma D \tag{10.13}$$

where, γ is the propagation constant inside the wall material $\gamma=(1+j)/\delta$ and δ is the skin depth. The skin depth is given by the formula,

$$\delta = \sqrt{\frac{2}{\omega\mu\sigma}} \tag{10.14}$$

In this expression μ is the magnetic permeability of the wall material and σ is the electrical conductivity. For such configurations, shielding for both electric and magnetic fields can be understood by the two simple equivalent circuits shown in Fig. 10.5. For a thin-walled spherical cell and low frequencies the parameters shown in Fig. 10.5 are given approximately by, $C=3\varepsilon_0 a/2$, $L=\mu_0 a/3$, $R=1/\sigma D$. Study of this circuit gives

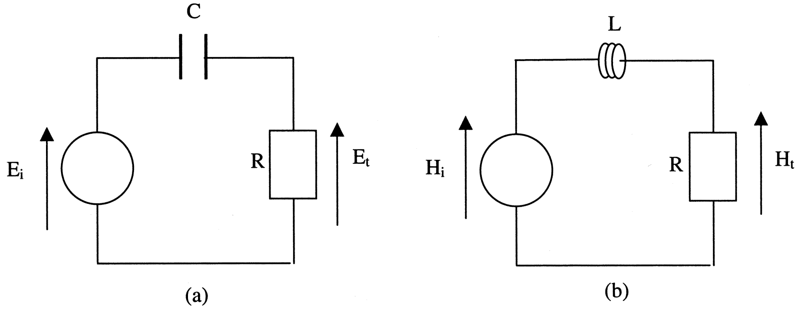


Figure 10.5 Circuit analogs for SE (a) for electric and (b) for magnetic fields.

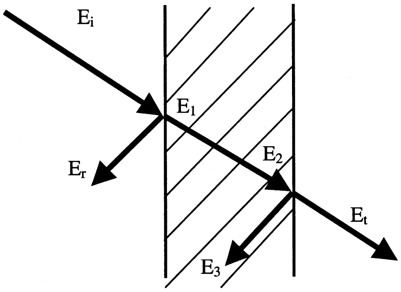


Figure 10.6 Wave approach to penetration through walls.

a good insight into some of the problems encountered with shielding. In each case subscript i indicates the incident field and t the transmitted field inside the structure. Hence a high voltage across R in the equivalent circuits indicates poor shielding. Examining electric field shielding first, we observe that at low frequencies (LF) shielding is very good (C has a very high impedance at LF). Hence it is relatively easy to shield against LF electric fields. In contrast, shielding of magnetic field at LF is very difficult (L has a very low impedance at LF). The shielding of LF magnetic fields requires special arrangements based on forcing it to divert into very high permeability (low reluctance) paths. As a general comment, reductions in R improve shielding. Hence any slots and/or obstructions on the surface of the shield must be placed in such a way that they do not obstruct the flow of eddy currents (thus keeping R low). Further formulas for diffusive penetration through shields for some canonical shapes may be found in Ref. 35.

Another approach to diffusive shielding is based on the wave approach [2,36,37]. This approach is depicted in Fig. 10.6 where an incident electric field E_i is partially reflected from the wall (E_r), partially penetrates (E_1), reaches the other side of the wall after some attenuation (E_2), suffers a partial internal reflection (E_3), and part of it is transmitted into the inner region (E_t). Component E_3 suffers further reflections (not shown) which contribute further to the transmitted wave. In complex problems, numerical solutions are necessary which employ special thin-wall formulations which also allow for inhomogenuities and anisotropies [38,39].

It should be emphasized that although we have discussed shielding here by illustrating penetration from an outer region to an inner region, the reverse process follows the same rules (equivalence principle).

10.3.2. Penetration Through Apertures

A major route for penetration of EM radiation is through apertures. By this we mean any hole, opening, ventilation grid, imperfect joint which breaches the continuity of the conducting shield. It is normally the case that apertures form the major route for radiation breaching a shield. Aperture penetration may be tackled in different ways depending on circumstances. These are based on small hole theory, simple analytical formulations for slots, intermediate level tools, and full numerical models. We examine each approach below.

1. For holes that are electrically small we first calculate the electric field E_{sc} at the position of the whole assuming that the aperture has been replaced by a

perfect conductor (short-circuit electric field). The presence of the aperture is then represented by placing an equivalent dipole inside the wall, where the aperture is again replaced by a perfect conductor. The dipole moment of the dipole is [40]

$$p_e = 2\varepsilon\alpha_e E_{\mathrm{sc}} \tag{10.15}$$

where α_e is the hole electric polarizability [41]. As an example, the polarizability of a round hole of diameter d is $\alpha_e = d^3/12$. The inner field can then be obtained by using antenna theory or any other suitable technique.

2. Alternative formulations have appeared in the literature where calculations of shielding effectiveness have been made for simple commonly encountered apertures. Particularly well known is the SE of a slot of length ℓ [29]:

$$\mathrm{SE} = 20\log\frac{\lambda}{2\ell} \tag{10.16}$$

If the length of the slot is 1/10 of the wavelength then SE = 14 dB. Such performance at 1 GHz implies slot lengths smaller than 3 cm. Clearly the shorter the length the higher the SE. For the same area of aperture it is better to have several smaller apertures rather than one large one. The formula above for N apertures modifies to

$$\mathrm{SE} = 20\log\frac{\lambda}{2\ell\sqrt{N}} \tag{10.17}$$

Equations (10.16) and (10.17) do not take into account either the width of the slot or the presence of a resonant equipment enclosure hence they may result in large errors in SE estimates.

3. Intermediate level tools can make good estimates of SE with a minimum of computational effort and are thus a compromise between accuracy and computational efficiency. The basic configuration is given in Fig. 10.7a and the intermediate level model in Fig. 10.7b [42–45]. The model of penetration through the aperture and propagation in the cabinet is broken down to three components:

   a. First, the incident field is represented by a simple Thevenin equivalent circuit, where the impedance is the intrinsic impedance of free space.
   b. Second, the aperture is represented by two halves of a coplanar strip line, shorted at both ends [46].
   c. Third, the cabinet is represented by a shorted waveguide with an impedance and propagation constant that take account of the first resonant mode.

   The three models are combined to form the complete model shown in Fig. 10.7b. This is relatively simple model to manipulate. The SE for the electric field is simply given in terms of the equivalent circuit parameters,

$$\mathrm{SE} = 20\log\frac{V_0}{2V(z)} \tag{10.18}$$

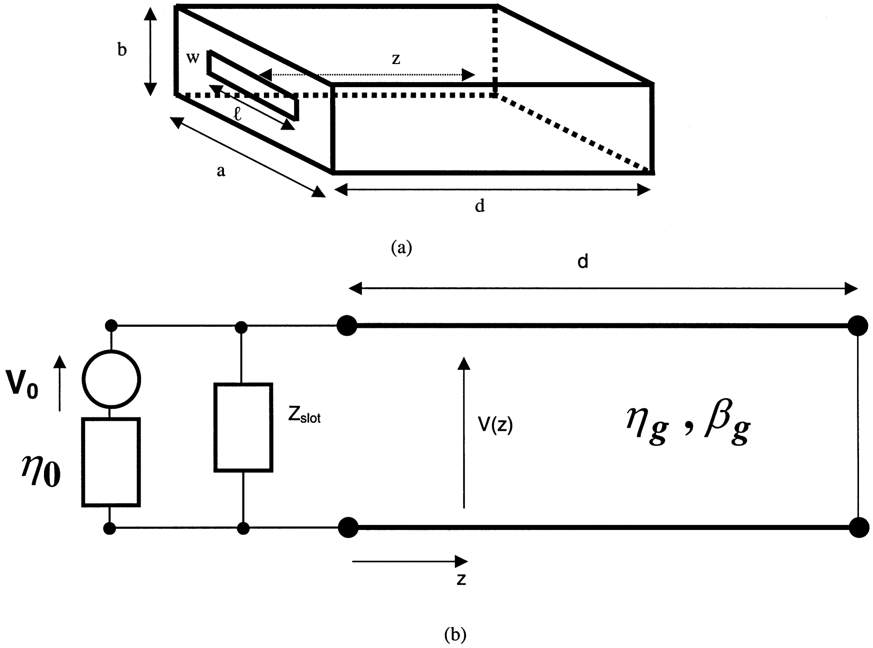


Figure 10.7 Intermediate level model (b) for the SE of a cabinet (a).

SE for the magnetic field is similarly obtained by replacing in Eq. (10.18) voltage by current. Typical results are shown in Fig. 10.8 and they illustrate several important points. At some frequencies, corresponding to cabinet resonances, the SE is negative implying that the presence of the cabinet results in field enhancement. The presence of the cabinet is of major significance in the calculation of SE. The SE has a different value depending on the point chosen to calculate it. Even away from resonances, the simple formula Eq. (10.16) is in considerable error. The introduction of PCBs and other loads inside the cabinet affects SE primarily near resonances. The method of including contents is explained in detail in Refs. 44 and 45. Application of these formulations in industrial problems may be found in Ref. 47.

4. The cabinet and its apertures may be described using one of the full-field solvers described in Refs. 5–8. For the case of a small number of electrically large apertures this process is straightforward [48]. However, in the case of complex and extensive ventilation grids the computational effort required in describing and meshing a large three-dimensional problem is excessive. In such cases, techniques have been developed to calculate SE using full-field models with embedded digital signal algorithms describing the grid of apertures [49–50]. Full-field calculation of SE in densely loaded cabinets, with several apertures, is still a very demanding computational task.

10.3.3. Conducted Penetrations

Conducted penetrations are another major means of introducing EMI into systems. Conducted penetrations may consist of power, control and communication cables which

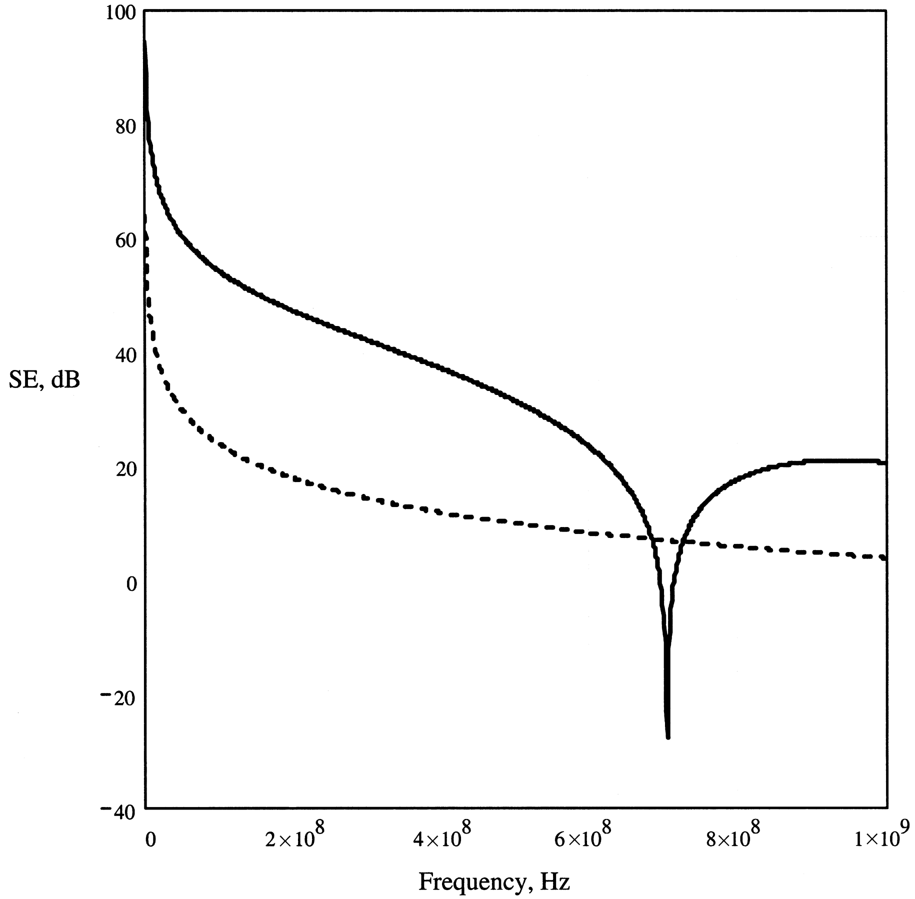


Figure 10.8 Electric field SE of a cabinet ($0.3 \times 0.12 \times 0.3\,\text{m}^3$, slot $0.1\,\text{m} \times 5\,\text{mm}$, $z = 0.15\,\text{m}$). Intermediate model (solid curve), Eq. (10.16) (broken curve).

may be shielded or unshielded. In addition, conducting pipes used for bringing services (water, air, etc.) into buildings and equipment form another route for EMI. Due to the variety of configurations it is difficult to offer general advice and general-purpose models for estimating the level of interference and thus ensuring EMC. We show in Fig. 10.9 in schematic form a penetration of a conductor through a barrier wall without a dc connection between the conductor and the conducting wall [2]. The approach to modeling this penetration is as follows:

We first estimate the coupling of the external field in the portion of the conductor which is the outer region. This can be conveniently done by using antenna theory and working out the coupling of the field to a monopole antenna (conductor above wall) [51,52]. This coupling is represented by the equivalent antenna components V_a and R_a. At the point of entry through the wall we introduce the barrier capacitance C_b to represent high-frequency displacement currents flowing between the conductor and the wall. C_b may be estimated or calculated from a full-field model of the region around the penetration. In the inner region we assume that the conductor is terminated by an equivalent resistance R (or impedance if appropriate), representing in the case of a terminated conductor the

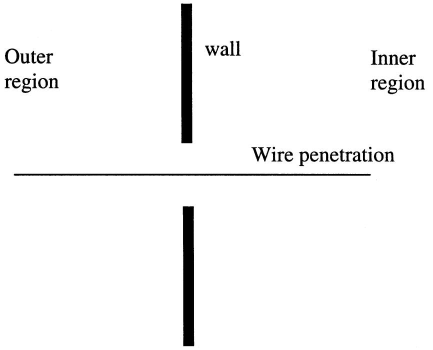


Figure 10.9 Wire penetration through a wall.

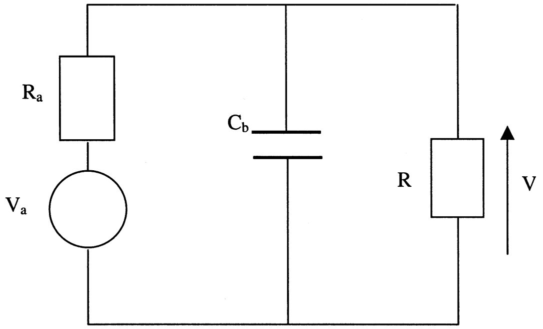


Figure 10.10 Circuit model of penetration for configuration in Fig. 10.9.

actual resistance of the termination, or, in the case of a floating conductor its radiation resistance. The complete approximate circuit is shown in Fig. 10.10. From this circuit we can calculate the voltage across R when the capacitor is present and when it is absent,

$$V_{\text{with } C} = \frac{V_{\text{without } C}}{1 + j\omega RC_b/(1 + R/R_a)} \tag{10.19}$$

From this equation it is clear that at high frequencies the barrier capacitance affords a degree of shielding. A more elaborate arrangement is shown in Fig. 10.11 where a feed through capacitance is shown. This example illustrates the need to use the appropriate model in the prediction of SE. At high frequencies it is not appropriate to use a lumped barrier capacitance as in Fig. 10.10. Taking as a measure of effectiveness the ratio V/I in Fig. 10.11 and treating the feed through capacitor as a short transmission line of length l and characteristic impedance Z_0 we obtain,

$$\frac{V}{I} = \frac{Z_0}{j\sin(2\pi\ell/\lambda)} \tag{10.20}$$

At low frequencies, the impedance in Eq. (10.20) reduces to the impedance of the barrier capacitance. However, when the length of the feed through capacitor approaches

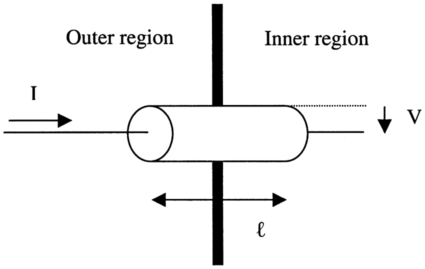


Figure 10.11 A coaxial wire penetration.

the wavelength, the impedance can have very large values. When the length is equal to half the wavelength, the impedance tends to infinity. This illustrates the care that must be taken when constructing models to estimate EMI. Another illustration of this problem is the model required when the penetration in Fig. 10.9 is modified by connecting the conductor to the wall at the entry point using a short length of wire ("pigtail" connection). In such a case it is essential to include in the model the inductance of the pigtail. This then makes it clear that at high frequencies, where the inductive impedance of the pigtail is large, the effectiveness of the connection to the wall is severely reduced. Matters can be improved if a 360° connection of the conductor to the wall is made. In all the cases illustrated above the calculation of the appropriate parameters to include in computations is not a simple matter. Although estimates can normally be made, a full characterization requires full-field EM calculations and the extraction from these of the required parameters. Further discussion of the treatment of wire penetration may be found in Refs. 52 and 53.

An important aspect of EMC analysis and design is the propensity of cables, which are used extensively as interconnects, to pick up and emit EM radiation. Cables with braided shields do not afford complete protection—a certain amount of radiation penetrates. This is traditionally described in terms of a transfer impedance relating the electric field parallel to the inner surface of the shield to the current flowing in the outer surface,

$$Z_T = \frac{E}{I} \tag{10.21}$$

For solid shields of thickness D and inner radius a, the transfer impedance can be calculated analytically [2] and is given by

$$\frac{|Z_T|}{R_{\text{dc}}} \simeq \begin{cases} 1 & D \ll \delta \\ 2\sqrt{2}\dfrac{D}{\delta}e^{-D/\delta} & D \gg \delta \end{cases} \tag{10.22}$$

where, δ is the skin depth, and $R_{\text{dc}} = 1/(2\pi a \sigma D)$ is the dc resistance of the shield.

The situation is much more complex for braided shields. Here, account must be taken of the small holes between strands, incomplete contact between strands, differences in spacing, etc. [54,55]. This is done by adding to a modified Z_T an additional reactive term to account for magnetic field coupling through holes and other imperfections in the shield

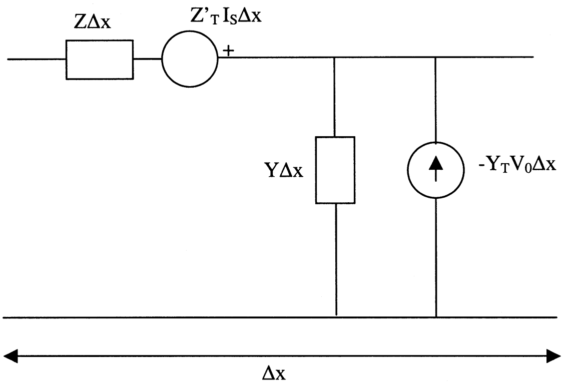


Figure 10.12 Model of a cable segment including transfer impedance and admittance.

Table 10.1 Magnitude of Cable Transfer Impedance (Typical Values in mΩ/m)

Cable type	0.1 MHz	1 MHz	10 MHz	100 MHz
URM102	5	3	4	23
URM43	11	25	158	1585
UR91	2	1	2.5	14
UR67	6	12	55	142
RG62	9	17	100	
RG228	3	0.5	1	
RG22	1	0.04	0.06	

$$Z'_T = Z_T + j\omega M' \tag{10.23}$$

In a similar manner, electric field penetration due to coupling through the capacitance to the inner conductor may be accounted for by a transfer admittance

$$Y_T = j\omega C' \tag{10.24}$$

Further details may be found in Refs. 56 and 57. An equivalent circuit of propagation in a cable taking into account coupling through the shield is shown in Fig. 10.12, where the voltage source accounts for magnetic filed coupling (normally the most significant coupling term), and the current source for electric field coupling (negligible in a well-constructed shield). Some typical values of transfer impedance for commercially available cables are shown in Table 10.1.

10.3.4. Radiation and Cross Talk

An important consideration affecting both EMC and SI is the coupling between adjacent circuits which are in the near field of each other (cross talk), and coupling over large distances through radiation either in the form of emission from circuits, or in the form of coupling of external fields onto circuits. First we tackle near-field coupling (cross talk), and then we examine far-field radiative coupling.

At low frequencies, coupling in the near field can be understood in terms of mutual capacitance and inductance between circuits. A simple approach to this problem is shown in Fig. 10.13. We limit the treatment to static, low-frequency solutions. Capacitive coupling in the case of the configuration in Fig. 10.13a gives the voltage induced on the second conductor due to a voltage on the first conductor as

$$V_2 = \frac{C_{12}}{C_{12} + C_2} V_1 \tag{10.25}$$

where the capacitance in this expression are as marked in the figure. If the second conductor is shielded, but the shield is not grounded, the shield potential will be

$$V_s = \frac{C_{1s}}{C_{1s} + C_s} V_1 \tag{10.26}$$

and since there is no current flowing through C_{2s}, wire 2 will rise to the same potential as the shield. If the shield is grounded, then $V_2 = V_{\rm s} = 0$. This provides electrostatic shielding. A similar calculation can be done for inductive coupling where the capacitive components shown in Fig. 10.13 are now replaced by inductive components. For the case shown in Fig. 10.13 the voltage induced on conductor 2 due a current flowing in conductor 1 is,

$$V_2 = M_{12}\frac{dI_1}{dt} \tag{10.27}$$

where M_{12} is the mutual inductance between the two conductors. If a shield is added as shown in Fig. 10.13, which is floating or connected to ground at only one point, then the induced voltage remains unchanged as given by Eq. (10.27). Only if the shield is connected to ground at both ends will there be a reduction of the induced voltage. Details may be found in Refs. 2 and 29.

The situation becomes considerably more complex as the frequency increases and the length of the conductor becomes comparable to the wavelength. The configuration is shown in Fig. 10.14 with terminations added. The problem is posed as follows: one conductor (the "generator wire" G) is driven by a source. What will be the induced

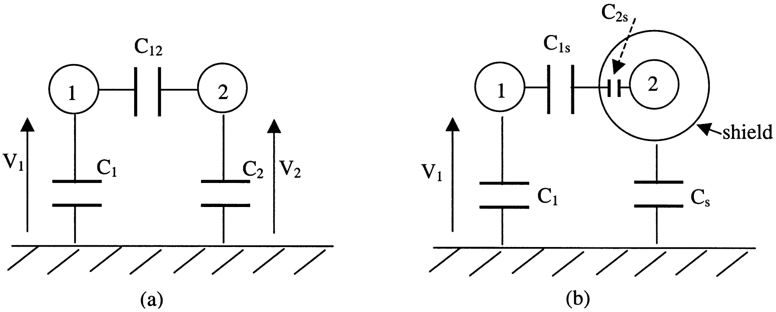


Figure 10.13 Simple electrostatic coupling models between (a) two wires and (b) two wires with shield.

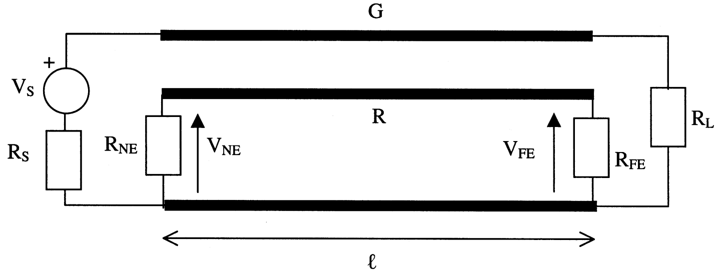


Figure 10.14 Three-wire model for studying cross talk.

voltage on the "receiver wire" R at the end near the source [near end (NE)] and far from the source [far end (FE)]? A full treatment of this problem is given in Refs. 30 and 58. We summarize here the main conclusions. For this configuration the per unit length parameter matrices are

$$\begin{aligned}[L] &= \begin{bmatrix} L_G & M \\ M & L_R \end{bmatrix} \\ [C] &= \begin{bmatrix} c_G + c_M & -c_M \\ -c_M & c_R + c_M \end{bmatrix} = \begin{bmatrix} C_G & -C_M \\ -C_M & C_R \end{bmatrix}\end{aligned} \tag{10.28}$$

The near-end and far-end cross-talk voltages are

$$\begin{aligned}V_{\mathrm{NE}} = \frac{S}{\mathrm{Den}}\Bigg[&\frac{R_{\mathrm{NE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega M\ell\left(C + \frac{j2\pi\ell/\lambda}{\sqrt{1-k^2}}\alpha_{L_G} S\right) I_{G_{\mathrm{dc}}} \\ &+ \frac{R_{\mathrm{NE}} R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega C_M \ell\left(C + \frac{j2\pi\ell/\lambda}{\sqrt{1-k^2}} \frac{1}{\alpha_{L_G}} S\right) V_{G_{\mathrm{dc}}}\Bigg]\end{aligned} \tag{10.29}$$

$$V_{\mathrm{FE}} = \frac{S}{\mathrm{Den}}\left(-\frac{R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega M\ell I_{G_{\mathrm{dc}}} + \frac{R_{\mathrm{NE}} R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega C_M \ell V_{G_{\mathrm{dc}}}\right) \tag{10.30}$$

where,

$$\begin{aligned}\mathrm{Den} &= C^2 - S^2\omega^2\tau_G\tau_R\left[1 - k^2\frac{\left(1 - \alpha_{S_G}\alpha_{L_R}\right)\left(1 - \alpha_{L_G}\alpha_{S_R}\right)}{\left(1 + \alpha_{S_R}\alpha_{L_R}\right)\left(1 + \alpha_{S_G}\alpha_{L_G}\right)}\right] + j\omega CS(\tau_G + \tau_R) \\ C &= \cos\beta\ell \\ S &= \frac{\sin\beta\ell}{\beta\ell} \\ k &= \frac{M}{\sqrt{L_G L_R}} = \frac{C_M}{\sqrt{C_G C_R}} \le 1\end{aligned}$$

$$
\begin{aligned}
\tau_G &= \frac{L_G \ell}{R_S + R_L} + C_G \ell \frac{R_S R_L}{R_S + R_L} \\
\tau_R &= \frac{L_R \ell}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} + C_R \ell \frac{R_{\mathrm{NE}} R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} \\
V_{G_{\mathrm{dc}}} &= \frac{R_L}{R_S + R_L} V_S \\
I_{G_{\mathrm{dc}}} &= \frac{V_S}{R_S + R_L} \\
Z_{C_G} &= \sqrt{\frac{L_G}{C_G}} \qquad Z_{C_R} = \sqrt{\frac{L_R}{C_R}} \\
\alpha_{S_G} &= \frac{R_S}{Z_{C_G}} \qquad \alpha_{L_G} = \frac{R_L}{Z_{C_G}} \qquad \alpha_{S_R} = \frac{R_{\mathrm{NE}}}{Z_{C_R}} \qquad \alpha_{L_R} = \frac{R_{\mathrm{FE}}}{Z_{C_R}}
\end{aligned}
$$

These equations are exact within the limitations of transmission line theory, i.e., TEM approximation is valid and that radiation from the line is negligible. If we assume that the lines are electrically short ($\ell \ll \lambda$) and that they are weakly coupled ($k \ll 1$), then the equations simplify to

$$
\begin{aligned}
V_{\mathrm{NE}} &= \frac{1}{\mathrm{Den}} \left[\frac{R_{\mathrm{NE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega M \ell I_{G_{\mathrm{dc}}} + \frac{R_{\mathrm{NE}} R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega C_M \ell V_{G_{\mathrm{dc}}} \right] \\
V_{\mathrm{NE}} &= \frac{1}{\mathrm{Den}} \left[- \frac{R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega M \ell I_{G_{\mathrm{dc}}} + \frac{R_{\mathrm{NE}} R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega C_M \ell V_{G_{\mathrm{dc}}} \right] \qquad (10.31) \\
\mathrm{Den} &\simeq (1 + j\omega\tau_G)(1 + j\omega\tau_R)
\end{aligned}
$$

For small frequencies ($\omega\tau \ll 1$), Den → 1 and the expressions above simplify further to

$$
\begin{aligned}
V_{\mathrm{NE}} &= \frac{R_{\mathrm{NE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega M \ell I_{G_{\mathrm{dc}}} + \frac{R_{\mathrm{NE}} R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega C_M \ell V_{G_{\mathrm{dc}}} \\
V_{\mathrm{FE}} &= - \frac{R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega M \ell I_{G_{\mathrm{dc}}} + \frac{R_{\mathrm{NE}} R_{\mathrm{FE}}}{R_{\mathrm{NE}} + R_{\mathrm{FE}}} j\omega C_M \ell V_{G_{\mathrm{dc}}}
\end{aligned}
\qquad (10.32)
$$

Both sets of simplified Eqs. (10.31) and (10.32) consist of two terms. The first term indicates inductive coupling and the second capacitive coupling. A study of these expressions permits the following general conclusions to be drawn:

Inductive coupling dominates for low-impedance loads.

Capacitive coupling dominates for high-impedance loads.

Coupling is proportional to frequency; hence, the faster the rate of change of the driving source, the higher the cross-talk levels.

Capacitive components at the near and far end are of the same magnitude and sign.

Inductive components are unequal and of opposite sign; hence, it is possible to choose the terminations to eliminate far-end cross talk.

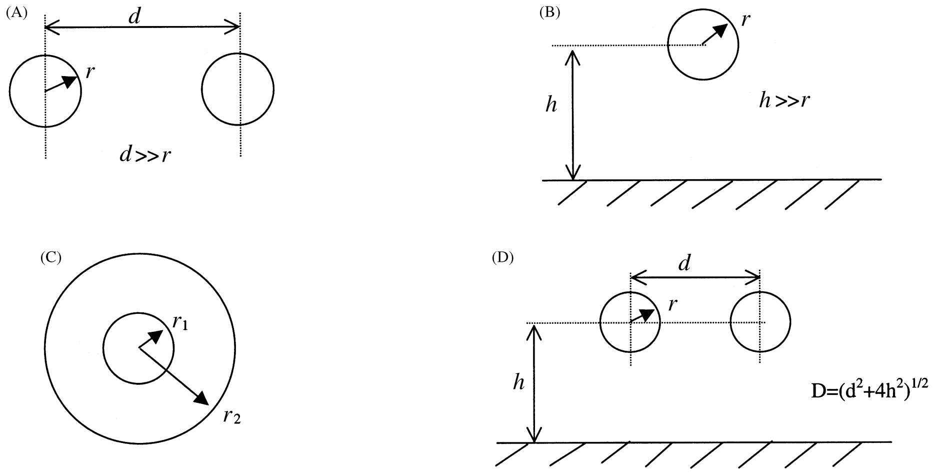
Cross talk at very high frequencies, where the TEM approximation is not valid and where there is substantial radiation from the line, can only be studied by numerical techniques. Another aspect of propagation on multiconductor lines which affects EMC and SI is the presence of more than one mode of propagation. In general, in a system of n lines there are $n-1$ modes of propagation. These modes may travel at different velocities and recombine at loads and discontinuities to produce distorted signals that contribute to noise. An introduction to modal propagation is given in Ref. 2 and a more complete treatment in [59,60]. Formulas for the calculation of parameters of some typical lines are given in Table 10.2.

Far-field radiative coupling refers to the coupling of external EM radiation onto circuits and the reverse effect of emission of EM radiation from circuits.

A typical problem is the calculation of voltages induced on interconnects subject to incident plane waves. A typical configuration is shown in Fig. 10.15a. There are three equivalent formulations to this problem [61]. According to the approach described in Ref. 62, the coupling to the field is described by two equivalent sources representing

Table 10.2 Electrical Parameters of Some Common Configurations

Two parallel wires (A)	$C = \frac{\pi\varepsilon}{\ln(d/r)}$ F/m	$L = \frac{\mu}{\pi} \ln\frac{d}{r}$ H/m
Wire above ground (B)	$C = \frac{2\pi\varepsilon}{\ln(2h/r)}$ F/m	$L = \frac{\mu}{2\pi} \ln\frac{2h}{r}$ H/m
Coaxial cable (C)	$C = \frac{2\pi\varepsilon}{\ln(r_2/r_1)}$ F/m	$L = \frac{\mu}{2\pi} \ln\frac{r_2}{r_1}$ H/m
Two wires above ground (D)	$[C] = \begin{bmatrix} c_{11}+c_{12} & -c_{12} \\ -c_{21} & c_{22}+c_{21} \end{bmatrix}$	$[L] = \frac{\mu}{2\pi} \begin{bmatrix} \ln\frac{2h}{r} & \ln\frac{D}{d} \\ \ln\frac{D}{d} & \ln\frac{2h}{r} \end{bmatrix}$
	$c_{11} = c_{22} = A\ln\frac{2h}{r} \qquad c_{12} = c_{21} = A\ln\frac{D}{d}$ $A = \frac{2\pi\varepsilon}{[\ln(2h/r)]^2 - [\ln(D/d)]^2}$	



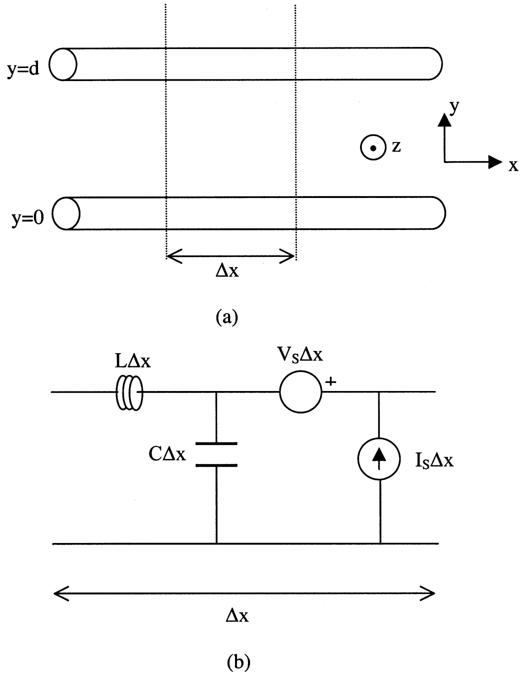


Figure 10.15 Coupling of an external field onto a two-wire line (a) and circuit model (b).

incident electric and magnetic field. The relevant modified transmission-line equations [63] are shown below:

$$\begin{aligned}\frac{dV(x)}{dx} &= -j\omega LI(x) + V_s(x)\\ \frac{dI(x)}{dx} &= -j\omega CV(x) + I_s(x)\end{aligned} \tag{10.33}$$

where L and C are the per unit length inductance and capacitance of the line and V_s and I_s are equivalent sources given by

$$\begin{aligned}V_s(x) &= j\omega\mu\int_0^d H_z^i(x,y)\,dy\\ I_s(x) &= -j\omega C\int_0^d E_y^i(x,y)\,dy\end{aligned} \tag{10.34}$$

H^i and E^i represent the incident field components. A complete treatment for different types of incident field may be found in Refs. 61–63. As an illustration we show the induced current at the two terminations of a line subject to end-fire excitation as shown in Fig. 10.16.

$$\begin{aligned}I_{\mathrm{NE}} &= j\frac{dE_0}{D}\sin\beta\ell\left(1+\frac{Z_L}{Z_C}\right)\\ I_{\mathrm{FE}} &= \frac{dE_0}{2D}\left(1-\frac{Z_S}{Z_C}\right)(1-\cos 2\beta\ell + j\sin 2\beta\ell)\end{aligned} \tag{10.35}$$

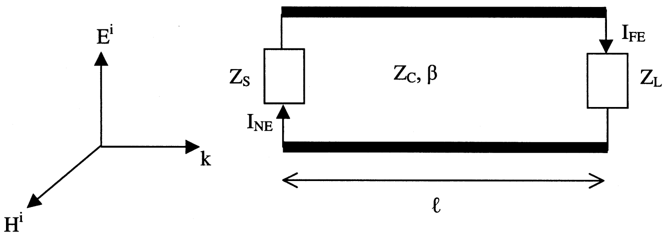


Figure 10.16 Configuration for the study of end-fire coupling.

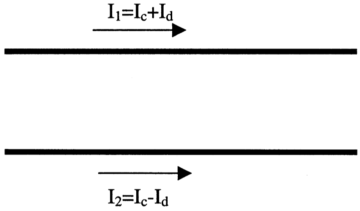


Figure 10.17 Common- and differential-mode currents on a two-wire line.

where, E_0 is the magnitude of the incident electric field and

$$D = \cos\,\beta\ell\,(Z_S + Z_L) + j\sin\,\beta\ell\left(Z_C + \frac{Z_S Z_L}{Z_C}\right)$$

Similar results for other types of excitation may be found in the references given. The reverse problem, namely, the emission of radiation from interconnects, is also important. Analytical techniques rely on a calculation of the currents that flow in an interconnect and then using this information together with antenna theory to obtain the radiated field. Of crucial importance in such calculations is the correct estimation of the current flowing in interconnects. In complex practical configurations, this current is not normally simply that calculated by transmission-line theory (differential current I_d). There is in addition a current component, which is due to a variety of mechanisms such as stray currents to nearby structures, which is described as common mode current I_c. The total current is the superposition of these two currents. The situation is shown schematically in Fig. 10.17. Only if $I_c = 0$ is the total current $I_1 = -I_2$. As the differential current components on the two wires are equal and opposite any radiation from them decays quickly with distance. In contrast, common mode current are in the same direction and make additive contributions to the radiated field. Although in general $I_c \ll I_d$ the contribution of the common mode current to radiation can dominate. Useful formulas for estimating the maximum electric field at a distance d from two parallel wires (separation s, length ℓ) are given below [30]:

$$\begin{aligned} E_{D,\max} &= 1.316 \times 10^{-14}\frac{|I_d|f^2\ell s}{d} \\ E_{C,\max} &= 1.257 \times 10^{-6}\frac{|I_c|f\ell}{d} \end{aligned} \tag{10.36}$$

where f is the frequency.

At high frequencies, numerical solutions are generally necessary. In some cases analytical solutions may be obtained [31].

10.4. PRACTICAL TECHNIQUES FOR THE CONTROL OF INTERFERENCE

Practical design to achieve EMC involves a series of measures to reduce emissions at source, a reduction in the efficiency of coupling paths, and improvements in immunity. Many of these techniques have been mentioned in the previous section.

As already pointed out, fast rise and fall times introduce a very wide spectrum of frequencies. Hence the slowest logic family should be used compatible with operational needs; otherwise, it is difficult to achieve EMC. Similarly, a system should be designed with the narrowest bandwidth to minimize the risk of becoming victim to noise. Many advanced systems use spread-spectrum techniques to address EMC and security problems [64]. Proper software design can also contribute to the immunity of systems to interference by including error checking and correction routines. Shielding and grounding of systems must be well thought out early in the design stage and closely monitored throughout the lifetime of a product. Whenever possible balanced cables should be used as this minimizes the flow of common-mode currents and the attendant problems of high emissions. Matching of interconnects to minimize reflections should be used whenever possible. Particular attention should be paid to the choice and installation of connectors as they tend to be the weak link in an otherwise well designed system. In critical systems, isolation techniques could be used (isolation transformers, optical links) to break interference paths. Filtering of cables on entry and exit from equipment must be considered. In particularly severe environments nonlinear limiting devices should be used to absorb high-energy pulses prior to further attenuation by filters, etc.

Positioning of circuits to minimize interference should be done carefully and segregation techniques to keep apart systems likely to interfere with each other should be considered early in the design phase. Whatever measures are taken during design to ensure EMC, they must be monitored throughout the lifetime of the product. This so-called *management of EMC* is problematic, yet it is an important aspect to consider during planning and design.

Several texts offer more detailed treatment of EMC design at the system and board level and should be consulted by designers [2,65,66].

10.5. EMC STANDARDS AND TECHNIQUES

EMC is the subject of an extensive set of international standards and national legislation to ensure that all products conform to a set of norms. Particularly important in this regard was the European EMC Directive in 1989 which gave an impetus to EMC design and certification. One can distinguish civilian standards, military standards, and company standards. Civilian standards cover a vast range of products and originate from International bodies (IEC, ITU etc.), large economic blocks, e.g., FCC from the United States and CEN from the European Community, etc. There has been a convergence of limits and procedures set by the different standards organizations, but small differences still persist in some areas. Military standards are set by the United States (MIL-STD-461D) [67], the United Kingdom (DEF-STAN 59-41) [68] and other countries. Large companies often impose a set of standards for internal use and for dealing with suppliers, which are designed to ensure that any national and international standards are comfortably met. There is a large range of standards available. Some standards are described as generic, i.e., they address general principles and set general limits. An example is the EN 50081 EMC Generic Emission Standards and the EN 50082, which addresses immunity. In addition, there are many product specific standards which apply to specific classes of equipment, e.g., IT equipment. In Table 10.3 we give contact addresses for the main standards bodies from which the interested reader can get up to date information.

A specialist area, which is beyond the scope of this chapter, is the setting of safe limits for human exposure to EM fields. This is important both for exposure in industrial environments and for exposure by the general public due to radio transmitters, mobile phones etc. The interested reader should consult specialist sources in this area [19,69,70].

Standards specify a test procedure and limits for emission and immunity for each class of equipment. Tests may be performed in different environments (open-area test site [71], screened or anechoic room [72], mode-stirred chamber [73], GTEM cell [74]) depending on the standard chosen. A typical test arrangement is shown in Fig. 10.18. The EUT is placed at the specified distance *D* from the receiving antenna Rx. The EUT is oriented for maximum received signal. Rx may be height scanned to obtain the

Table 10.3 Contact Points for EMC Standards and Codes of Practice.

International Electrotechnical Commission (IEC),
3, rue de Varembe, PO Box 131, CH-1211, Geneva, Switzerland, pubinfor@iec.ch

International Organization for Standardization (ISO). As above, central@iso.ch

European Committee for Standardization (CEN),
36, rue de Stassart, B-1050 Brussels, Belgium, infodesk@cenorm.be

CENELEC: European Committee for Electrotechnical Standardization,
35, rue de Stassart, B-1050 Brussels, Belgium, general@cenelec.be

European Telecommunications Standards Institute (ETSI),
F-06921 Sophia Antipolis Cedex, France, Infocentre@etsi.fr

British Standards Institution (BSI),
389 Chiswick High Road, London, W4 4AL, UK, info@bsi-global.com

CIGRE: International Council for Large Electric Systems,
21, rue d'Artois, 75008 Paris, France, www.cigre.org

Institute of Electrical and Electronic Engineers (IEEE),
3 Park Avenue, New York 10016, USA, http://standards.ieee.org

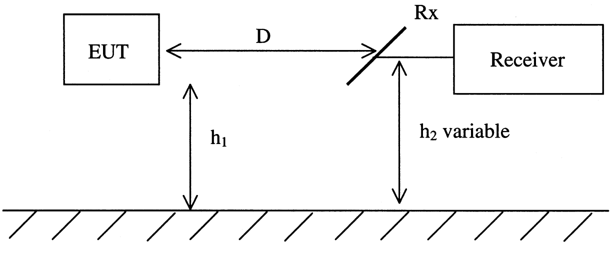


Figure 10.18 A typical arrangement for EMC testing.

maximum signal. The electric field is calculated from the measured voltage and the antenna factor AF

$$E\mathrm{dB}\,\mu\mathrm{V/m} = V_{\mathrm{rec}}\,\mathrm{dB}\mu\mathrm{V} + \mathrm{AF}\,\mathrm{dB} \tag{10.37}$$

The receiver characteristics are specified in standards (peak or quasipeak detectors [27,28]). The measured electric field over the entire frequency range is then compared with the limits specified in standards. As an example, the emission limit according to CISPR 22 measured at an open-area test site and at a distance of 30 m from the EUT is 30 dBμV/m (30–230 MHz) and 37 dBμV/m (245 MHz–1 GHz) for class A equipment (equipment for use in industrial, commercial, and business premises). The same limits apply for equipment in residential use (class B) but measured at a distance of 10 m.

Most test environments are imperfect in some way and care must be taken when taking and interpreting measurements. In screened rooms, one is confronted by the presence of room resonances. For a room of dimensions a, b, c resonances occur at,

$$f\,\mathrm{MHz} = 150\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2 + \left(\frac{p}{c}\right)^2} \tag{10.38}$$

where m, n, and p are integers (no more than one can be zero). Resonances make measurements difficult to interpret. Radiation damping material can be added to remove or damp resonances. This is very difficult especially at low frequencies.

In open are test sites, of major importance is the problem of ground reflections. This is illustrated in Fig. 10.19a. The signal from a transmitter Tx gets to the receiver through the direct path and a reflection from the ground. The two signal paths can be explained in terms of image theory as shown for vertical and horizontal polarizations in Fig. 10.19b,c respectively. The electrical path for the direct and image rays is different. The total signal at Rx is the superposition of these two rays and thus may vary substantially. This explains the requirement in standards that a height scan is done during tests to find the maximum field at Rx.

GTEM cells cannot be made large enough to handle very large equipment and the EUT must be small enough to avoid affecting too much the field profile inside the cell.

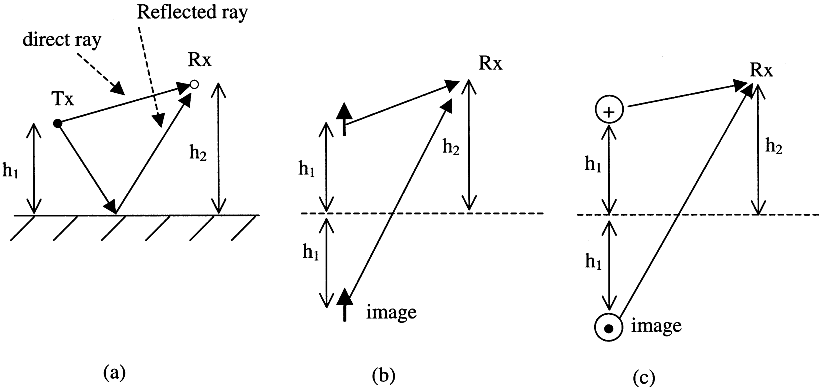


Figure 10.19 (a) Direct and reflected rays from Tx to Rx and images for (b) vertical and (c) horizontal polarization.

The above descriptions refer to radiated emissions above 30 MHz. Below this frequency conducted emission tests are specified which consist of measuring the noise voltage across a specified mains impedance. This is provided by a line impedance stabilizing network (LISN) placed at the connection point of the EUT to the power supply. As an illustration, the CISPR 22 standard for conducted emission specifies 73 dBμV (0.5–30 MHz) for class A equipment.

10.6. OUTSTANDING PROBLEMS AND FUTURE TRENDS

There are several areas where improvements are required and further developments should be expected. In the area of testing there is a need to improve methods and procedures so that tests made in a particular environment can be referred to another environment. This will reduce the number of tests required and improve repeatability of measurements.

EMC tests at frequencies above 1 GHz are tedious and prolonged as it is difficult to scan accurately such a vast range. The directivity of emissions from equipment tends to be higher at higher frequencies. Different standards and approaches to testing may thus be required to rationalize procedures and do more meaningful tests. This is becoming increasingly important as high-speed electronics are becoming more prevalent. Along the same lines, as clock frequencies get higher EMC and SI issue become more closely related and therefore design and predictive tools should be integrated so that these two aspects may be treated concurrently throughout the design phase.

Further major developments are required to improve predictive capabilities and thus to allow iterative EMC design to take place at all stages of design. The tools required will be computer based and will incorporate sophisticated models to handle efficiently the complexities of modern systems.

Another aspect of analysis and synthesis methods in EMC is the way in which uncertainty (manufacturing and component tolerances) is dealt with. Standards, test procedures, and CAD tools are essentially based on deterministic models of systems. With increasing complexity and higher frequencies, consideration should be given to statistical techniques in the EMC characterization and design of systems [75].

There is considerable pressure on the electromagnetic spectrum to accommodate new services such as Bluetooth and TETRA [76]. The impact of these trends on interference and on general background noise levels should be considered so that the EM spectrum remains a well-managed resource for the benefit of all.

REFERENCES

1. International Electrotechnical Commission (IEC). Email: pubinfor@iec.ch
2. Christopoulos, C. *Principles and Techniques of Electromagnetic Compatibility*; CRC Press: Boca Raton, 1995.
3. Young, B. *Digital Signal Integrity*; Prentice Hall PTR: Upper Saddle River, NJ, 2001.
4. Walker, C.S. *Capacitance, Inductance, and Cross-talk Analysis*; Artech House: Boston, Norwood, MA, 1990.
5. Volakis, J.L.; Chaterjee, A.; Kempel, L.C. *Finite Element Method in Electromagnetics*; IEEE Press: New York, 1998.
6. Harrington, R.F. *Field Computation by Moment Methods*; Macmillan: New York, 1968.
7. Taflove, A. *Computational Electrodynamics: The Finite-Difference Time-Domain Method*; Artech House: Norwood, MA, 1995.

8. Christopoulos, C. *The Transmission-Line Modeling Method: TLM*; IEEE Press: New York, 1995.
9. Balanis, C. *Antenna Theory: Analysis and Design*, 2nd Ed.; Wiley Interscience: New York, 1996.
10. Papoulis, A.; Pillai, S.U. *Probability, Random Variables and Stochastic Processes*, 4th Ed.; McGraw Hill: New York, 2002.
11. Papoulis, A. *Signal Analysis*; McGraw Hill: New York, 1977.
12. Uman, M.A.; Krider, E.P. A review of natural lightning: experimental data and modeling. IEEE Trans. EMC **1982**, *24*, 79–112.
13. Gardner, R.L.; Baker, L.; Baum, C.E.; Andersh, D.J. Comparison of lightning with public domain HEMP waveforms on the surface of an aircraft. Proc. 6th Zurich Symp. on EMC, Zurich, 1985; pp. 175–180.
14. ITU Report 670-1. Worldwide minimum external noise levels, 0.1 Hz to 100 GHz, Dusseldorf, 1990, Annex to Vol. 1.
15. Davenport, E.M.; Frank, P.J.; Thomson, J.M. Prediction of field strengths near HF transmitters. Radio Electron. Eng. **1983**, *53*, 75–80.
16. Independent Expert Group on Mobile Phones (IEGMP). Mobile Phones and Health, http://www.iegmp.org.uk, 2000.
17. Sugiura, A.; Okamura, M. Evaluation of interference generated by microwave ovens. Proc. 7th Zurich Symp. EMC, 1987; pp. 267–269.
18. Koga, R.; Wade, O.; Hiraoka, T.; Sano, H. Estimation of electromagnetic impulse noise radiated from a digital circuit board. Proc. Int. Conf. EMC, Nagoya, 1989, pp. 389–393.
19. Ran, L.; Gokani, S.; Clare, J.C.; Bradley, K.J.; Christopoulos, C. Conducted electromagnetic emissions in induction motor drives—Parts I and II. IEEE Trans. Power Electronics **1998**, *4*, 757–776.
20. Working Group 36.04. Guide on EMC in power Plants and Substations, Paris: CIGRE, 1997
21. Ma, M.T. How high is the level of EM fields radiated by an ESD. Proc. 8th Int. Zurich EMC Symp., Zurich, 1989, pp. 361–365.
22. Longmire, C.L. On the electromagnetic pulse produced by nuclear explosions. IEEE Trans. EMC **1978**, *29*, 3–13.
23. Gardner, R.L.; Baker, L.; Baum, C.E.; Andersh, D.J. Comparison of lightning with public domain HEMP waveforms on the surface of an aircraft. Proc. 6th Int. Zurich EMC Symp., Zurich, 1985, pp. 175–180.
24. ITU Report 258-5. Man-made Radio Noise. Dusseldorf, 1963-1990, Annex to Vol. VI, 1990.
25. IEEE recommended Practice for an Electromagnetic Site Survey (10 kHz–10 GHz), IEEE Std. 473, 1985.
26. Suprynowitz, V.A. *Electrical and Electronics Fundamentals*; West Publ. Company, St Paul (MN), 1987.
27. Geselowitz, D.B. Response of ideal radio noise meter to continuous sine wave, recurrent impulses, and random noise. IRE Trans. Radio Interference **1961**, 2–11.
28. CISPR Publ. 16. Specification for radio interference measuring apparatus and measurement methods.
29. Ott, H.W. *Noise Reduction Techniques in Electronic Systems*, 2nd Ed.; Wiley Interscience: New York, 1988.
30. Paul, C.R. *Introduction to Electromagnetic Compatibility*; Wiley Interscience: New York, 1992.
31. Tesche, F.M.; Ianoz, M.V.; Karlsson, T. *EMC Analysis Methods and Computational Models*; Wiley Interscience: New York, 1997.
32. Cooley, W.W. Low-frequency shielding effectiveness of nonuniform enclosures. IEEE Trans. EMC **1968**, *10*, 34–43.
33. King, L.V. Electromagnetic shielding at radio frequencies. Philos. Mag. **1993**, *15*, 201–223.
34. Thomas, A.K. Magnetic shielding enclosure design in the dc and VLF region. IEEE Trans. EMC **1968**, *10*, 142–152.
35. Lee, K.S.H. Electromagnetic shieding. In *Recent Advances in Electromagnetic Theory*; Kritikos, H.N., Jaggard, D.L., Eds.; Springer-Verlag: New York, 1990.

36. Schelkunoff, S.A. *Electromagnetic Waves*; Van Nostrand: Toronto, 1943.
37. Kaden, H. *Wirbelstrome und Schirmung in der Nachrichtentechnik*, 2nd Ed.; Springer-Verlag: New York, 1959.
38. Trenkic, V.; Duffy, A.P.; Benson, T.M.; Christopoulos, C. Numerical simulation of penetration and coupling using the TLM method. Proc. EMC Symp., Rome, 1994, pp. 321–326.
39. Trenkic, V.; Christopoulos, C.; Benson, T.M. Numerical simulation of polymers and other materials for electronic shielding applications. Proc. Polymat 94, London, 1994, 384–387.
40. Bethe, H.A. Theory of diffraction by small holes. Phys. Rev. **1944**, *66*, 163–182.
41. Cohn, S.B. Electric polarizability of apertures of arbitrary shape. Proc. IRE, 1952, pp. 1069–1071.
42. Robinson, M.P.; Turner, J.D.; Thomas, D.W.P.; Dawson, J.F.; Ganley, M.D.; Marvin, A.C.; Porter, S.J.; Benson, T.M.; Christopoulos, C. Shielding effectiveness of a rectangular enclosure with a rectangular aperture. Electronics Lett. **1996**, *32*, 1559–1560.
43. Sewell, P.; Turner, J.D.; Robinson, M.P.; Thomas, D.W.P.; Benson, T.M.; Christopoulos, C.; Dawson, J.F.; Ganley, M.D.; Marvin, A.C.; Porter, S.J. Comparison of analytic numerical and approximate models for shielding effectiveness with measurements. IEE Proc.-Sci. Meas. Technol. **1998**, *145*, 61–66.
44. Thomas, D.W.P.; Denton, A.; Konefal, T.; Benson, T.M.; Christopoulos, C.; Dawson, J.F.; Marvin, A.C.; Porter, S.J. Characterization of the shielding effectiveness of loaded equipment cabinets. EMC York 99, IEE Conf. Publ. 464, York, 1999, pp. 89–94.
45. Thomas, D.W.P.; Denton, A.; Konefal, T.; Benson, T.M.; Christopoulos, C.; Dawson, J.F.; Marvin, A.C.; Porter, S.J.; Sewell, P. Model of the EM fields inside a cuboidal enclosure populated by conducting planes or printed-circuit boards. IEEE Trans. EMC **2001**, *43*, 161–169.
46. Gupta, K.C.; Garg, R.; Bahl, I.J. *Microstrip Lines and Slotlines*; Artech House: Norwood, MA, 1979, Chap. 7.
47. De Smedt, R.; De Moerloose, J.; Criel, S.; De Zutter, D.; Olyslager, F.; Laermans, E.; Wallyn, W.; Lietaert, N. Approximate simulation of the shielding effectiveness of a rectangular enclosure with a grid wall. Proc. IEEE Int. Conf. EMC, Denver, 1998, pp. 1030–1034.
48. Kraft, C.H. Modeling leakage through finite apertures with TLM. Proc. IEEE Int. Symp. EMC, Chicago, 1994, pp. 73–76.
49. Podlozny, V.; Paul, J.; Christopoulos, C. Efficient calculation of the shielding effectiveness of equipment cabinets in full-field numerical models. Proc. EMC Europe 2002, Sorrento, 2002, pp. 853–857.
50. Podlozny, V.; Christopoulos, C.; Paul, J. Efficient description of fine features using digital filters in time-domain computational electromagnetics. IEE Proc. Sci. Meas. Technol. **2002**, *149*, 254–257.
51. Tang, T.G.; Tieng, Q.M.; Gunn, M.W. Equivalent circuit of a dipole antenna using frequency-independent lumped elements. IEEE Trans. AP **1993**, *41*, 100–103.
52. Thomas, D.W.P.; Denton, A.; Benson, T.M.; Christopoulos, C.; Paul, J.; Konefal, T.; Dawson, J.F.; Marvin, A.C.; Porter, S.J. Electromagnetic coupling to an enclosure via a wire penetration. Proc. IEEE Int. Symp. EMC, Montreal, 2001, pp. 183–188.
53. Sarto, M.S.; Scarlatti, A. Combined FDTD-TL modelling of a transmission line crossing a metallic box. Proc. 4th European Symp. EMC, Brugge, 2000, pp. 239–244.
54. Vance, E.F. Shielding effectiveness of braided-wire shields. IEEE Trans. EMC **1975**, *17*, 71–77.
55. Casey, K.F. EMP coupling through cable shields. IEEE Trans. EMC **1978**, *20*, 100–106.
56. Hoeft, L.O.; Hofstra, J.S.; Peel, R.J. Experimental evidence for purpoising coupling and optimization in braided cables. Proc. 8th Zurich Int. EMC Symp., Zurich, 1989, pp. 505–509.
57. Benson, F.A.; Cudd, P.A.; Tealby, J.M. Leakage from coaxial cables. IEE Proc. **1992**, *A 139*, 285–302.
58. Paul, C.R. Solution of the transmission-line equations for three conductor lines in homogeneous media. IEEE Trans. EMC **1978**, *20*, 216–222.
59. Djordjevic, A.R.; Sarkar, T.K.; Harrington, R.F. Time-domain response of multiconductor transmission lines. Proc. IEEE **1987**, *75*, 743–764.

60. Paul, C.R. Frequency response of multiconductor transmission lines illuminated by an electromagnetic field. IEEE Trans. EMC **1976**, *18*, 183–190.
61. Nucci, C.A.; Rachidi, F. On the contribution of the electromagnetic field components in field-to-transmission line interaction. IEEE Trans. EMC **1995**, *37*, 505–508.
62. Abraham, R.T.; Paul, C.R. Basic EMC technology advancement for C^3 systems—coupling of EM fields onto transmission lines. RADC-TR-82-286, Vol. IVA, 1982.
63. Paul, C.R. *Analysis of Multiconductor Transmission Lines*; Wiley Interscience: New York, 1994.
64. Dixon, R.C. *Spread Spectrum Systems*; Wiley Interscience: New York, 1984.
65. Weston, D.A. *Electromagnetic Compatibility—Principles and Applications*; Marcel Dekker: New York, 1991.
66. Williams, T. *EMC for Systems and Installations*; Oxford: Newnes, 1999.
67. Department of Defence, Washington, DC. Requirement for the Control of Electromagnetic Interference Emissions and Susceptibility. MIL-STD-461D, 1993.
68. Ministry of Defence, Glasgow, UK. Electromagnetic Compatibility. DEF-STAN 59-41, 1988.
69. International Commission on Nonionizing Radiation Protection (ICNIRP). Guidelines on limits of exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz). Health Phys **1998**, *74*, 494–522.
70. Repacholi, M.H. Assessment of the health effects of EMF exposure. Radio Sci. Bull. **2002**, *301*, 14–24.
71. Smith, A.A. Standard-site method for determining antenna factors. IEEE Trans. EMC **1982**, *24*, 316–322.
72. Christopoulos, C.; Paul, J.; Thomas, D.W.P. Absorbing materials and damping of screened rooms for EMC testing. Proc. Int. Symp. EMC, Tokyo, 1999, pp. 504–507.
73. Ma, M.T. Understanding reverberating chambers as an alternative facility for EMC testing. J. Electromagnetic Waves Appl. **1988**, *2*, 339–351.
74. Garbe, H.; Hansen, D. The GTEM cell concept: application of this new test environment to emission and susceptibility measurements. Proc. 7th Int. Conf. EMC, York, 1990, pp. 152–156.
75. Holland, R.; St John, R. *Statistical Electromagnetics*; Taylor and Francis: Philadelphia, PA, 1999.
76. Wesolowski, K. *Mobile Communication Systems*; Wiley Interscience: New York, 2002.

11
Radar

Levent Sevgi
DOGUS University
Istanbul, Turkey

Electronic sensors are being used in a variety of applications in our modern life, from security and defense to public health, education to transportation, science to sports. The sensors may be electromagnetic, acoustic, thermal, chemical, biological, etc. A radar (an acronym for *radio detection and ranging*) is commonly used for an electromagnetic sensor. In this chapter, fundamentals of radar are presented. Starting from the historical background, the theory, the signal environment, the radar equation, and applications are outlined.

11.1. INTRODUCTION AND HISTORICAL BACKGROUND

Radar is about using electromagnetic waves to detect the presence of objects and to extract as much information as possible from the interaction of electromagnetic waves with objects. The concept can be traced back to the pioneering studies on radio transmission and reception; to the works of Hertz in 1886, Hulsmeyer in 1903, and Marconi in 1922. Radar development studies accelerated in the United Kingdom, France, Germany, and the United States during 1935–1940 and particularly in the United States during 1940–1945. The period 1950–1960 corresponds to the introduction of new techniques in radar applications, especially coherent techniques, such as Doppler processing and pulse compression. The principles, technology, and applications of radar were publicized by fundamental books, such as those written by Skolnik and Barton (see, for example, Refs. 1–3, their latter editions) during 1960–1970. The solid-state technology, integrated circuits, microprocessors, etc., accelerated its development during 1970–1980, and finally, the period 1980–1990 corresponds to the mature age of the radar theory and technology. A brief historical overview is given in Ref. 4.

As given in the applicable IEEE standard [5], a radar is

> A device for transmitting electromagnetic signals and receiving echoes from objects of interests (targets) within its volume of coverage. Presence of a target is revealed by detection of its echo or its transponder reply. Additional information about a target provided by a radar includes one or more of the following: distance (range), by the elapsed time between transmission of the signal and reception of the return signal; direction, by use of directive antenna patterns; rate of change of range, by measurement of Doppler shift; description or classification of target, by analysis of echoes and their variation with time.

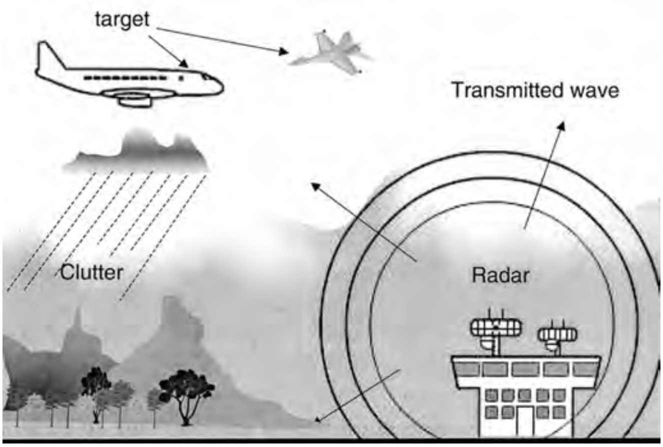


Figure 11.1 A typical radar scenario.

This simple and clear definition of radar shows that

Radar is a device that transmits and receives electromagnetic signals.
There are objects of interest (targets) and noninterest (clutter, interference, jammer).
Information is extracted from the echo signal initially by detection.
Target information includes, but not limited to range, range rate (velocity), direction, description, etc.

A typical radar scenario is pictured in Fig. 11.1, where an airport surveillance radar is in operation. The scenario includes two air targets (a fighter and a commercial airplane), mountains and trees on the ground, clouds in the sky, rain, etc. The radar transmitted signal interacts with the environment, and its receiver receives echoes from possible targets, unwanted echoes from mountains, trees, clouds in the sky, rain, etc. The total received echo is a signal, which contains signatures of different components, generally categorized as target, noise, clutter, and interference.

11.2. TERMS AND CONCEPTS

A radar *target* is the object of interest that is embedded in noise and clutter together with interfering signals. *Noise* is a floor signal which limits the smallest signal that can be measured in the receiver. Noise is present in all electronic circuits, although it is often quite small compared with useful signals. *Clutter* is a radar (background) echo or group of echoes from ground, sea, rain, birds, chaff, etc., that is operationally unwanted in the situation being considered. There is no single definition for clutter, and clutter or target may interchange depending on the duty of the radar. For example, an echo from rain is clutter for an airport surveillance radar but is the target for a weather radar. Similarly, ground echo is clutter for a ground surveillance radar but is itself the target (useful signal) for a ground imaging radar.

Radars can be classified according to purpose, application, type, installation, operating frequency, transmit waveform, receiver processing techniques, etc., as illustrated in Table 11.1 Major purposes of a radar may be detection, tracking, classification, identification, surveillance, imaging, or guidance, as listed.

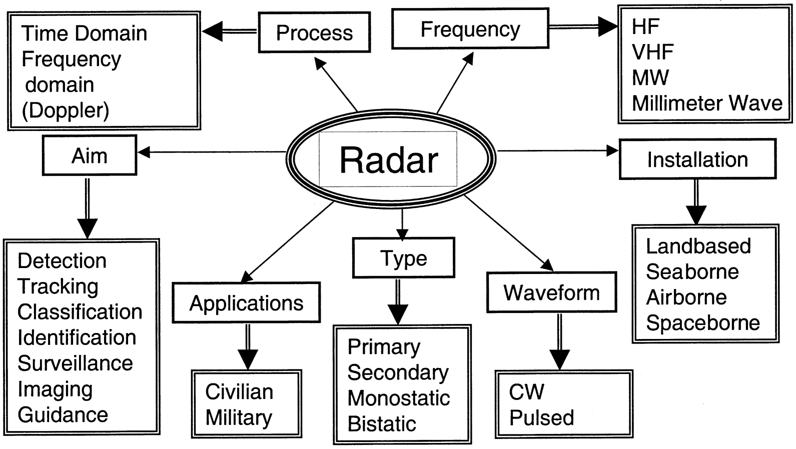


Table 11.1 Types of Radars According to Different Parameters

Target detection is the ability to distinguish target at the receiver. The total radar echo signal at the receiver consists of the target (wanted signal), the noise (unwanted, uncorrelated signal), the clutter (unwanted but correlated echoes from unwanted targets), and the interference (unintentional such as radio and TV broadcast signals and/or intentional jamming signals).

Target tracking is the process of following the moving target continuously, i.e., to monitor its range, direction, velocity, etc. Tracking may be done mechanically (i.e., by steering the receiver antenna in a way to hold the target inside the receiver beam) or electronically (i.e., by digital beam steering techniques at the receiver processor). It may be single-target tracking or multitarget tracking, where the latter requires target discrimination.

Target classification is to distinguish certain types of targets and group them according to certain characteristics called *features*. For example, to group the detected sea surface targets into frigates, boats, tankers, or air targets into fighters, cargo planes, etc., is the process of classification. Grouping them as military and civilian targets is also a form of classification. Possible distinguishing features may be their size, speed, onboard electronic devices, electromagnetic reflectivity, maneuvers, etc.

Target identification is the process of finding out "who" the target is. This knowledge of a particular radar return signal that is from a specific target may be obtained by determining size, shape, timing, position, maneuvers, or rate of change of any of these parameters by means of coded responses through secondary radar or by electronic counter measures (i.e., by listening to and recording active communication and radar systems onboard of the target).

Imaging radars are microwave (MW) radars, which can provide high resolution in range and cross-range to obtain "radar picture" of surface and air targets and earth's surface. They are range profiling (RP) radars, synthetic aperture radars (SAR), and inverse synthetic aperture radars (ISAR).

An RP radar is a high-resolution active instrument, which has range resolution cell sizes much smaller than typical dimensions of the observed targets so that multireturns from different range cells along the target can be used to have a longitudinal reflectivity profile.

SAR is an airborne or spaceborne active instrument that produces high-resolution imagery of surface targets and earth's surface (ocean and terrain), which achieves its mission by tracing the target via the motion of the platform. The high cross-range resolution is obtained via a synthetic aperture or, equivalently, via Doppler processing. The term *synthetic aperture* refers to the distance the radar travels during data collection for Doppler processing.

ISAR is a land-based and/or airborne active instrument that produces high-resolution imagery of surface and air targets, and it uses the motion of the target as information.

Surveillance is systematic observation of a region (aerospace, surface, or subsurface areas) by a different number of different sensors, primarily for the purpose of detecting, tracking, classifying, and identifying activities of interest. Surveillance may be air to air (A/A), air to ground (A/G), air to surface (A/S), surface to air (S/A), surface to surface (S/S), etc.

Basic radar applications are listed in Table 11.2. They may be grouped in two ways: monostatic/bistatic and primary/secondary. A monostatic radar is a radar where its transmitter and receiver are colocated. When the transmitter and the receiver sites are separated, the radar is said to be bistatic. In a primary radar system, subsystem, or mode of operation, the return signals are the echoes of its own emitted signals obtained by reflection from the target. On the other hand, a secondary radar extracts target information from a target transmit signal sent by any IFF (identify friend or foe) transponder. It is a radar technique or mode of operation in which the return signals are obtained from a transponder or a repeater carried by the target.

Radars may be installed at fixed locations (land based) or mobile on a truck (land based), aircraft (airborne), ship (seaborne), or satellite or space shuttle (spaceborne). They may use signals with different frequencies (such as HF, 3–30 MHz; VHF, 30–300 MHz; or microwave (MW), 300 MHz up to tens of GHz) with different waveforms [continuous wave (CW), frequency modulated CW (FMCW), FM interrupted CW (FMICW),

Table 11.2 Some Applications of Current Radars

Civil Appl.	Military Appl.
Air traffic control and flight management Intelligent traffic management systems Precision approach and landing Vessel traffic management (harbors, waterways, straits) Navigation and collision avoidance Weather radar and ocean monitoring Search and rescue Ground surveillance and intruder alarms Ground probing and subsurface imaging Vehicle speed sensors and altimeters Wide-area surveillance Multifunction	Land, ocean, and air surveillance Detection and tracking Classification and identification ballistic Missile defense Airborne early warning Fire control and missile guidance Mortar and artillary location Search and rescue operations Ground probing and subsurface Detection simulation and modeling Multifunction

pulsed, etc.]. Their receiver processor may perform detection either in the time domain (TD) or in the frequency domain (FD).

CW radars are simple and occupy minimal spread in the frequency spectrum. Its transmitted power level is much less than the peak power level of a pulsed radar. It is used to measure the speed of a target (i.e., a traffic radar) by using the Doppler effect (i.e., the shift in the frequency of CW signal caused by the radial speed of a target moving toward or away from the radar). It is also used to detect moving targets in a region (e.g., an intruder alarm). On the other hand, it is not capable of measuring the range of a target, unless the CW signal is frequency modulated. In the frequency modulated CW (FMCW) radars, the frequency of the CW wave is periodically modulated by applying a frequency shift that varies linearly with time in the range of $f_0 \pm f_m$, where f_0 is the frequency of the carrier wave and f_m is the frequency deviation. CW and FMCW radars are also attractive because of their low-level transmitter power requirements, which are within the capability of current solid-state power amplifiers (low cost). However, CW and FMCW radars are bistatic since their transmitter (with a relatively high-level transmitted power) and receiver (with a relatively low-level echo signal) are on at the same time. This results in a direct arrival of the transmitted signal with noise to the receiver, where it competes with the target echo.

Frequency is a basic radar parameter that determines not only the design and construction of a radar but also the application and performance. Table 11.3 lists military and commercial radar bands. Although there are many in the table, we group radar frequencies mainly into three; HF radars, VHF radars, and MW radars, according to EM propagation characteristics and target interaction properties in these regions. The MW radars find wide areas of applications, some of which are listed in Table 11.4 together with the assigned frequency ranges.

Pulsed radars find more applications than CW radars; therefore, most of the chapter is devoted to pulsed radar systems. Unless otherwise stated, the information included here refers to pulsed radars.

A typical block diagram of a radar is shown in Fig. 11.2, which consists of a transmitter block and a receiver block. The carrier signal is generated from a local oscillator and modulated by a suitable radar waveform (best suited for the operational purposes) in the transmitter, and this signal is transmitted via the transmit antenna system. All the echoes are received by the receive antenna system and processed first in the receiver processor unit [which includes all electronic processing stages, such as filtering, amplifying, mixing, and analog-to-digital converting (ADC)] and then in the data processor unit (the hardware units where digital data is processed by computer algorithms). Finally, the output is displayed as graphics in the video display unit.

Table 11.3 Military and Commercial Radar Bands

VHF	30–300 MHz	Very high frequency	138–144 MHz 890–942 MHz
S	2–4 GHz		2.3–2.5 GHz
C	4–8 GHz		5.25–5.925 GHz
Ku	12–18 GHz		13.4–14.0 GHz
K	18–27 GHz		24.05–24.25 GHz
mm	40–300 GHz	Millimeter waves	33.4–36.0 GHz

Table 11.4 Frequency Ranges Assigned to Certain Radar Applications

Frequency range	Spectrum allowance
1.35–1.4 GHz	Military comms/radar
1.435–1.535 GHz	L-band telemetry
2.45–2.69 GHz	Commercial comms/radar
2.9–3.7 GHz	Miscellaneous radar
4.2–4.4 GHz	Radar altimeter
5.25–5.925 GHz	Miscellaneous radars
8.5–10.55 GHz	Miscellaneous radars
9.3–9.5 GHz	Weather radar and maritime navigation radar
13.25–14 GHz	Miscellaneous radars and satellite comms
15.7–17.7 GHz	Miscellaneous radars
24.25–25.25 GHz	Navigation radar
33.4–36 GHz	Miscellaneous radar

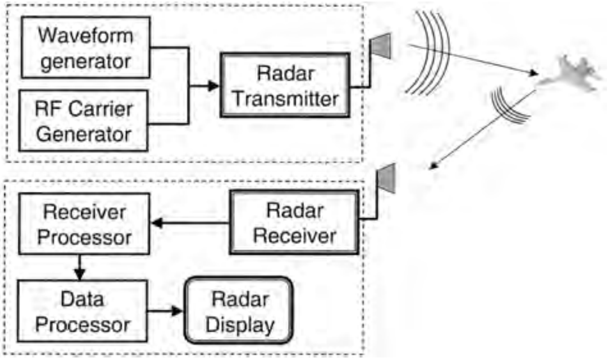


Figure 11.2 A simple block diagram of a radar.

11.2.1. Resolution and Accuracy

Radar is a measuring device that measures target's range (distance between the radar and the target), range rate (velocity of the target), direction (angular position of the target), and reflected power [radar cross section (RCS) of the target]. Because of imperfections in any measuring instrument, some amount of error will always be introduced. The errors of a radar are characterized by two performance parameters: resolution and accuracy. *Resolution* is the radar's ability to distinguish two targets in close proximity of each other, mostly in a three-dimensional space: (1) range, (2) bearing, and (3) velocity (Doppler). *Accuracy* is the ability of the radar to measure the true value (i.e., the true range, velocity, direction, etc.) to within some stated error specification, and intuition tells us that it must be related to the received power level of the target (i.e., sharpness of the target signal above the noise level). Any measurement made in a gaussian type noise environment and with a signal-to-noise ratio (SNR), using a system with a basic resolution Δ, will have an rms error, δ, which can generally be expressed as $\delta = \Delta/(2\mathrm{SNR})^{1/2}$ [1]. It has to be noted that this definition of accuracy corresponds to measurement errors which are bounded by one standard deviation. Hence, assuming a gaussian-like noise distribution, one can conclude that the probabilities for the actual result to be within δ and 2δ vicinity of its measured value is approximately 70% and 95%, respectively.

11.2.2. PRF and Maximum Range

Ideal transmitted and received time series of a pulsed radar (without carrier signal) are pictured together in Fig. 11.3. Here, pulses marked as 1, 2, 3,..., are the transmitted pulses and the rest are the received pulses. When the transmitter is on during the transmission of a pulse having a pulse width τ, the receiver is off (to secure its sensitive electronic components from high transmitted power effects). This is repeated every T seconds, which is called the *pulse repetition interval* (PRI), inverse of which is equal to the pulse repetition frequency ($\mathrm{PRF} = 1/T$). The ratio τ/T is called the *duty factor*. During the time interval $(T - \tau)$, the transmitter is off and the receiver is on to receive any possible target echoes. The range (R) of the target is measured from the time delay (t_d) between the transmit and received signals as

$$R = \frac{ct_d}{2}\ \mathrm{m} \tag{11.1}$$

where c is the speed of light ($c = 3 \times 10^8$ m/s). The factor 2 in Eq. (11.1) arises because the distance traveled by the signal is $2R$, i.e., to the target and back. The maximum useful range is determined by the PRI (or PRF) as

$$R_{\max} = \frac{cT}{2} = \frac{c}{2\mathrm{PRF}}\ \mathrm{m} \tag{11.2}$$

Usually, the radar receiver samples the received time echo signal every τ seconds and each sample represents a distance ΔR called a *range bin* or *range gate*:

$$\Delta R = \frac{c\tau}{2}\ \mathrm{m} \tag{11.3}$$

which is also called a *range resolution*. The number of range bins N is then equal to the maximum range divided by the range resolution ($N = R_{\max}/\Delta R$). The narrower the pulse width τ, the better the range resolution, and in turn, the higher the number of range bins. On the other hand, the narrower the pulse width, the wider the bandwidth B [Hz] of the signal (i.e., $\tau \sim 1/B$). It should be noted that there are different definitions for both pulse width and bandwidth of a signal when it is not an ideal rectangular pulse. Here, they

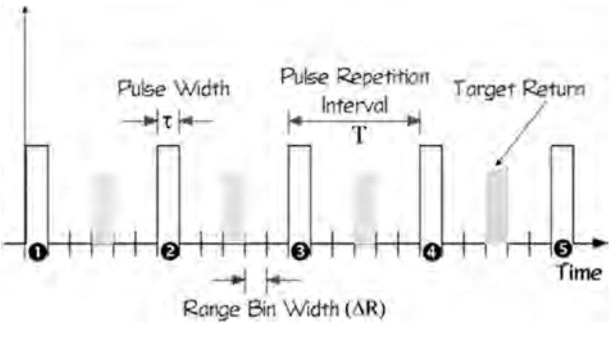


Figure 11.3 Radar transmit and receive pulse definitions.

are both defined as the distance between half-power points of the pulse in TD and of the spectrum in FD. For a pulse width τ, the range accuracy can be given as [1]

$$\delta R \cong \frac{c\tau}{2\sqrt{2 \times \mathrm{SNR}}} = \frac{c}{2B\sqrt{2 \times \mathrm{SNR}}}\ \mathrm{m} \tag{11.4}$$

11.2.3. Pulse Integration and Doppler Frequency

A radially approaching (receding) target causes a slight increase (decrease) in the carrier frequency of the radar that is proportional to its radial speed. This is called *Doppler effect* and is used in radar systems to detect and/or discriminate targets in FD. Targets that cannot be discriminated in TD, because of strong unwanted echoes, may easily be discriminated in FD by using radial velocity differences. Radars that use the Doppler effect are called *pulse Doppler* or *moving-target-indicator* (MTI) radars.

Measurements of the position of a target can be made quickly (with a unique received pulse), but it takes some time to estimate velocities and to distinguish differences in velocities of targets. The smaller the velocity differences, the longer the time needed to estimate. This is clearly understood from the Fourier theory. It is known that analytical Fourier transform is defined for infinite time ($-\infty < t < \infty$) and infinite frequency ($-\infty < f < \infty$) range. Observing a signal for an infinite duration in time yields a zero frequency resolution, that is, one can get frequency information from an infinite time series at any particular frequency. Similarly, one needs to know infinite frequency range behavior to rebuild the signal *exactly* in TD. As for all other discrete real signal processing cases, radar signals have a finite duration in time. From a finite duration time series (let's call it observation time T_{obs}) with Δt sampling interval, one can obtain FD response via discrete Fourier transform (DFT), or mostly fast fourier transform (FFT), with a maximum frequency $f_{\max}$ and a minimum frequency resolution Δf as

$$f_{\max} = \frac{1}{2\,\Delta t}\mathrm{Hz} \qquad \text{and} \qquad \Delta f = \frac{1}{T_{\mathrm{obs}}}\mathrm{Hz} \tag{11.5}$$

It is obvious from Eq. (11.5) that one needs to obverse the target (illuminate the target, collect consecutive pulses at the receiver and maintain a time series) longer if a better frequency resolution is required. This process is called *integration* and is done to enhance detectability, to reduce measurement errors, to improve resolution or some other performances. The length of time taken to make an observation with a radar is called the *integration time*. Coherent (incoherent) integration is the process of collecting consecutive pulses in TD, where the receiver is tuned to the same carrier frequency with the transmitter with (without) phase locking to it. In coherent radars, a complex time series is formed (with target echo amplitude and phase) and Fourier transformed (via FFT), and detection is done in FD, where a moving target with a radial speed component appears as an impulse like signal along the frequency axis, far from the zero frequency. This Doppler shift of a target depends on the radial velocity v_r and the carrier signal wavelength λ_0 of the radar as

$$f_d = \frac{2v_r}{\lambda_0}\mathrm{Hz} \tag{11.6}$$

Finally, the velocity accuracy of a target can be calculated as [1]

$$\delta v_r \cong \frac{\Delta f_d}{\sqrt{2 \times \mathrm{SNR}}} \approx \frac{\lambda}{2T_{\mathrm{int}}\sqrt{2 \times \mathrm{SNR}}}\ \mathrm{m/s} \tag{11.7}$$

11.2.4. Angular and Elevation Scan

Location of a target with a radar system requires determining its range (radial distance), azimuth (angular position), and height. In a pulsed radar, range information is extracted by directly time gating in the receiver. Azimuth and elevation information is obtained by the scanning characteristics of the receive antenna system, as either a mechanical scanning or an electronic scanning system. In mechanical scanning (see Fig. 11.4), the directive antenna with a narrow beam characteristics is rotated mechanically with a constant speed, which is adequate to illuminate each angular sector for a while and receive the number of required echoes. If the antenna has beam widths of $\Delta\varphi$ (rad) and $\Delta\theta$ (rad) in azimuth and elevation, respectively, the radar azimuth and elevation resolutions will be $\Delta\varphi$ and $\Delta\theta$, respectively. Its azimuth and elevation accuracies are

$$\delta\varphi \cong \frac{\Delta\varphi}{\sqrt{2 \times \mathrm{SNR}}}\,\mathrm{rad} \qquad \delta\theta \cong \frac{\Delta\theta}{\sqrt{2 \times \mathrm{SNR}}}\,\mathrm{rad} \tag{11.8}$$

respectively.

MW radars with mechanical scanning use dish type (parabolic reflector) antennas and their directivity (i.e., lossless) gain is given as [2]

$$G = \frac{4\pi}{\Delta\varphi\,\Delta\theta} \tag{11.9}$$

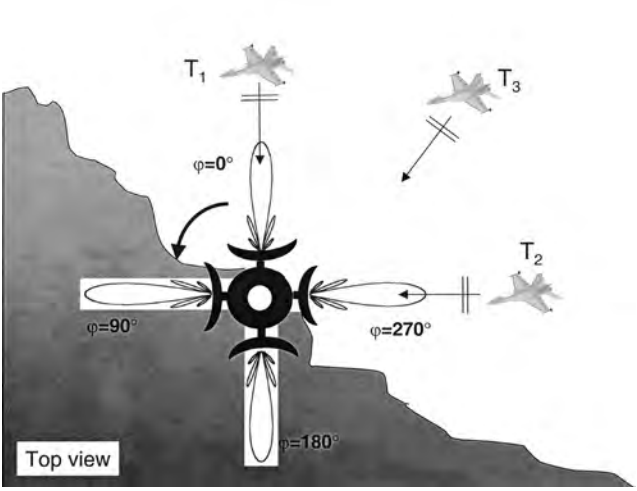


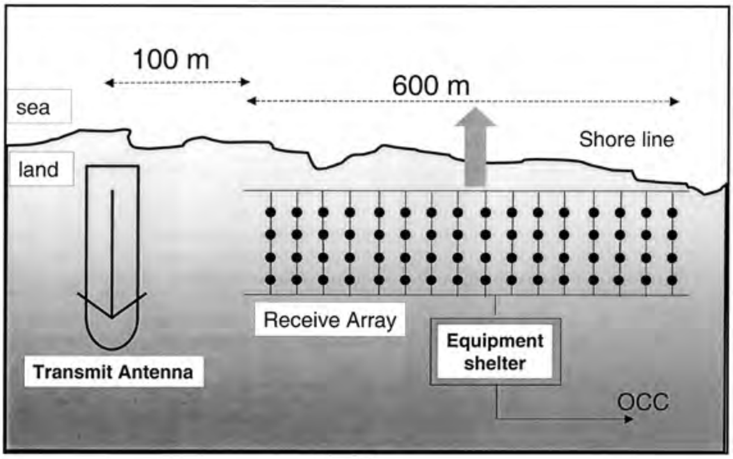
Figure 11.4 Mechanical scanning.

For a solid beam width $\Omega = \varepsilon \Delta\varphi\, \Delta\theta$, the number of beams required to scan the hemisphere (4π steradians being the entire sphere) is

$$\text{Number of beams} = \frac{2\pi}{\Delta\varphi\, \Delta\theta} = \frac{G}{2} \tag{11.10}$$

Mechanical scanning is used mostly in MW radars, especially in tracking radars, where the target's path is determined and its route is predicted. Although tracking can be carried out using range, angle, or Doppler information, angular tracking is the characteristic feature in tracking radars. It should be noted that all surveillance radars may also be considered as tracking systems (to some extent) since they keep track of many targets simultaneously. The process of mechanical tracking is called *track-while-scan*.

In HF and most of VHF radars, mechanical scanning is not possible because of the large apertures of the receiving antenna systems; therefore, electronic scanning is the only way to obtain angular information of a target. A typical transmit and receive antenna site of a HF surface wave (HFSW) radar (that operates in the frequency range of 3–6 MHz) is pictured in Fig. 11.5a, where a 24 × 4 vertical monopole array over the earth's surface is used as the receive antenna system [6,7]. Here, the aim is to cover a coastal region, so a quadlet (four element) end-fire array is used as the receive array element. Its mission is to direct radar energy toward the ocean with a high front-to-back ratio. As narrow as 4°–5° azimuth beam widths are obtained by using 16 to 24 quadlets and as much as 100°–120° azimuthal coverage can be obtained. The width and length of this receive antenna array is 300 m × 600 m, which may extend to more than a kilometer. In a HFSW radar, angular locations of the targets are obtained solely by digital beam forming (i.e., electronic scanning). Typical beams formed digitally (with computer simulations [6]) are given in Fig. 11.5b. As easily seen from the figure, the shapes of the antenna beams are quite different and become distorted as the beams leave the boresight azimuthally. For example, it is easy to locate the angular position of target T_1 in beam A in the figure, since the main beam is far stronger than the other lobes (side and back lobes), where echo signals from T_2 and T_3 are easily suppressed. The angular discrimination is not that good in beam B (for 30° angular



(a)

Figure 11.5 (a) A transmit and receive antenna site of an HFSW radar and (b) electronic scanning.

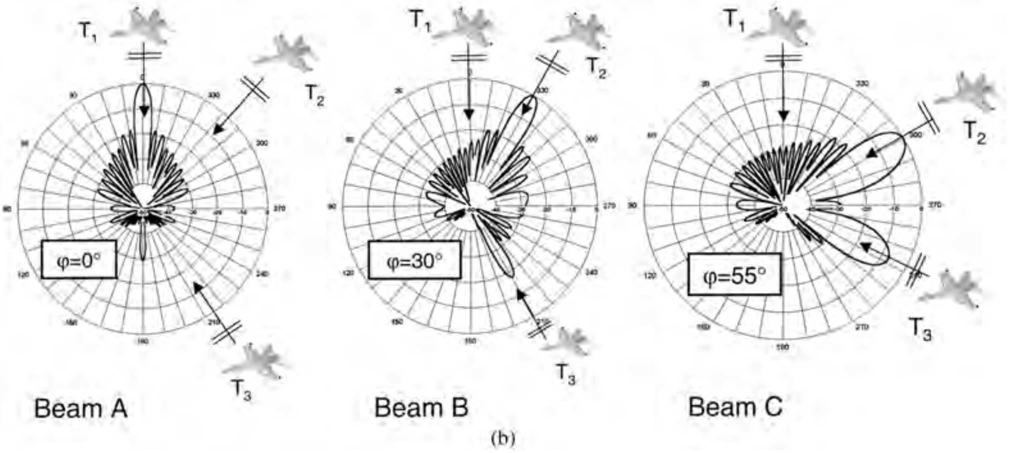


Figure 11.5 Continued.

scan) and gets worse in beam C (for 55° angular scan), since strong unwanted side and back lobes appear, where targets T_2 and T_3 may appear to be in the same angular direction.

Both mechanical and electronic scanning have advantages and limitations, which forces the radar engineer to do optimization depending on the data flow and processing speed in the radar computer. Good coverage and good resolution are payoffs. Users usually desire to have a radar with good coverage (long in range and wide in azimuth/ elevation) and good resolution (narrow range and angular resolutions), which means a high number of beams to scan or form. The higher the number of beams, the higher the scan rate and the lower the dwell time (pulse integration period).

11.2.5. Analog-to-Digital Conversion (ADC) Process

The power of today's radar systems comes from both its electronic subsystems and intelligent software. The return echo signal is processed by high-speed computers via powerful algorithms, and this is accomplished first by digitizing (sampling) the time signal via ADC. Sampling a signal in TD makes its spectrum periodic in FD. Mathematically, it corresponds to multiplying the time signal by an infinite extend impulse train. A Fourier transform of an infinite-extent impulse train with impulse separation τ is another infinite-extent impulse train with a separation in the FD of $1/\tau$. Also, multiplication in TD corresponds to convolution in FD, which makes the spectrum periodic. Finally, if a signal reconstruction in TD is required a low-pass filtration in FD (that is equivalent to multiplication of the discrete time signal with a Zinc function in TD) is applied, which are illustrated in Fig. 11.6. This is a well-established theory (sampling theory) in digital signal processing.

ADC translates the input voltage of the receiver to binary numbers that computer hardware can process. The radar receiver may have very large echoes from nearby huge targets (and often from clutter) and at the same time very weak echoes from distant small targets. Therefore, a radar receiver must have a high *dynamic range* (the dynamic range of a receiver is the range between maximum and minimum detectable signals), which is an obvious consequence of the two-way path loss. The higher the dynamic range the greater the number of bits in the ADC. In most radar receiver systems, the echo

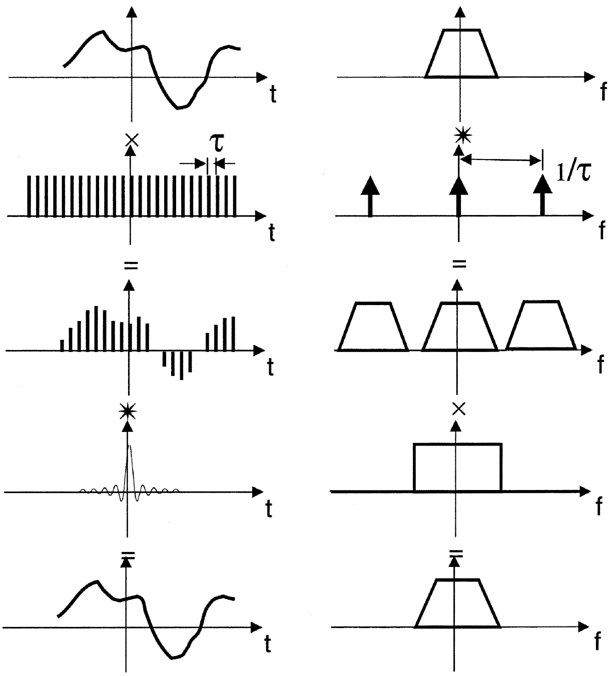


Figure 11.6 Sampling and DFT effects (×: multiplication, *: convolution).

signal undergoes some electronic and digital processes, such as RF filtering, IF converting and amplifying, and finally video (baseband) filtering and information extracting. At which stage ADC will be used depends on the speeds of the ADC. With today's ADC speeds (which reach up to Gbit/s) ADC may be used almost anywhere in the radar receiver.

11.2.6. Range–Doppler Ambiguities

Equation (11.2) relates the maximum range of a radar to its PRF. The lower the PRF (the higher the PRI), the longer the maximum range. If longer ranges (longer T) with good resolution (shorter τ) are to be obtained, the radar needs to operate at very high peak powers, since the transmitted average and peak powers, P_t and P_p, are related via the duty factor as

$$P_t = P_p \frac{\tau}{T} \text{ W} \tag{11.11}$$

It is not easy to obtain high peak powers in solid-state radar technology; therefore, low PRF radars with high resolution are not easily realizable.

What happens if longer ranges (longer than $R_{\max}$ determined by T) are to be covered? As illustrated in Fig. 11.7, the receiver may not distinguish whether the second received echo (Rx 2) belongs to the first transmitted pulse (Tx 1) or the second one (Tx 2). In this case the received echo amplitude may give a clue (as the range of a target increases its echo amplitude decreases because of propagation losses) but there may still be an ambiguity. For example, a huge target (with strong RCS) at range 3 may have the same order echo amplitude with a small target (with weak RCS) at range 2.

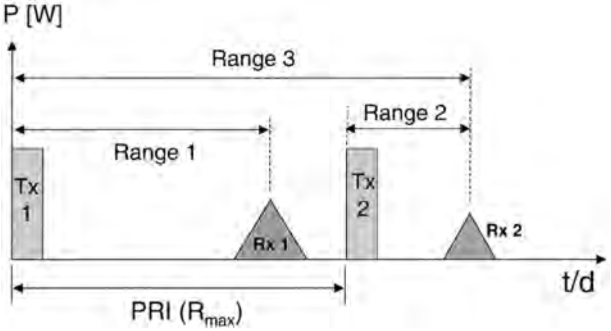


Figure 11.7 PRF, maximum range, and ambiguities.

PRF (i.e., $1/T$) also determines maximum target speed in Doppler FD, since the spectrum become periodic with PRF. The lower the PRF, the lower the maximum target speed in a Doppler radar. Therefore, designing low PRF radar to avoid range ambiguity problems causes ambiguities in the Doppler domain for today's high speed targets. Also, targets with radial velocities of integer multiple of radar PRF cannot be detected. These are called *blind velocities* (v_b) and are calculated as [2]

$$v_b = \text{PRF}\,\frac{n\lambda}{2}\,\text{m/s} \tag{11.12}$$

where n is an integer and λ is the radar carrier wavelength ($\lambda = c/f_0$).

High PRF radars are ambiguous in range; low PRF radars are ambiguous in velocity; medium PRF radars are ambiguous in both range and velocity. There are methods to overcome ambiguity problems without changing PRF, but they are not generally applicable. For example, one method is to label each pulse (transmitting them with different frequencies, phases, polarizations, etc.) so that echoes can easily be assigned to their transmitted pulses. But, this does not work for targets when their RCS fluctuates from pulse to pulse. Labeling the transmitted pulse also causes difficulties in Doppler processing. Another way of overcoming ambiguities is to use a burst of pulses with variable PRFs, which in turn may result in the occurrence of blind ranges and blind speeds at different places.

11.2.7. Pulse Compression and Matched Filter

The detectability of an echo is dependent on the total energy of the radar signal. The higher the transmitted energy, the longer the range and the higher the detectability. The aim of a radar is (1) to detect the target, (2) to measure its range and velocity as accurately as possible, and (3) to discriminate targets in range, angle, and velocity, which are not compatible with each other. For example, transmitted energy must be increased to increase detectability, which may be achieved by using a longer pulse duration τ. On the contrary, to increase range resolution, a short pulse duration τ (i.e., a large bandwidth) must be used, which also results in poor Doppler resolution. To achieve both, short pulses with high energy must be used, which requires very large peak powers. Very large peak powers are not desirable for long-term operation with minimal cost or may not be available in the solid-state design. The solution is pulse compression, that is, to use long

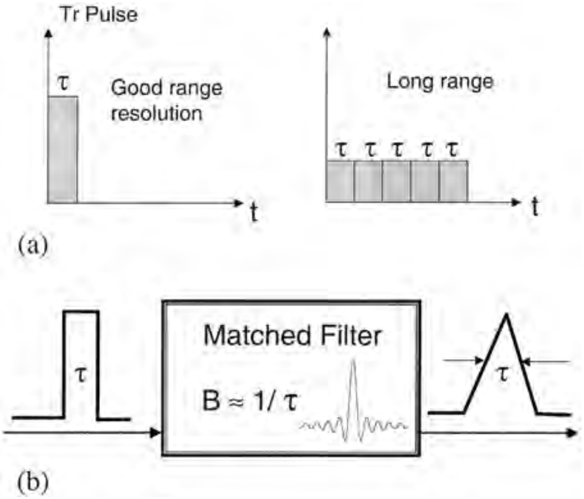


Figure 11.8 (a) Pulse compression and (b) matched filter.

pulses during transmission for long range coverage and better detection performance (also for good Doppler resolution) requiring a reasonable peak power. During reception, short pulses are needed to achieve a high-range resolution. This can be accomplished by designing a suitable waveform, normally by using frequency modulation and by using a correlation receiver, the matched filter (Fig. 11.8).

The radar echo at the receiver consists of a possible target, noise, clutter, and interference signals. The higher the target signal among the others, the better the detection process. If one assumes the noise floor as the threshold signal, the fundamental criterion that determines the ability to detect a target is the signal-to-noise ratio (SNR). As will be discussed later, noise power depends on the receiver bandwidth and is independent of the radar signal. Therefore, for a fixed noise power SNR may be maximized by using a correlation receiver (whose frequency domain implementation is the matched filter). As shown in Fig. 11.8, the matched filter is a filter, whose transfer bandwidth is equal to the radar pulse bandwidth. Physically, this corresponds to minimizing the noise power, which results in maximization of SNR. Geometrically, it corresponds to sharpening the peak of the received pulse (as shown in the figure) which in turn maximizes the SNR.

11.2.8. CFAR Detection and Decision Making

At the radar receiver, target echoes are contaminated with other unwanted echoes (such as noise and clutter); therefore, a threshold decision is required for target detection. Constant-false-alarm-rate (CFAR) detection in a radar is a technique to extract target signals from the noise and clutter, which, in general, are all random processes [8]. As a random signal, the total radar echo (target, noise, and clutter) fluctuates with time. A random variable can be represented by a probability density function (PDF) of its amplitude fluctuation, plus its frequency content, the power spectrum function. The important parameters of a PDF are mean and median values, standard deviation, and cumulative probability function.

A typical radar echo is pictured in Fig. 11.9. Here, analog and digitized (sampled) radar echoes are given on top and bottom, respectively. The vertical axis is the power

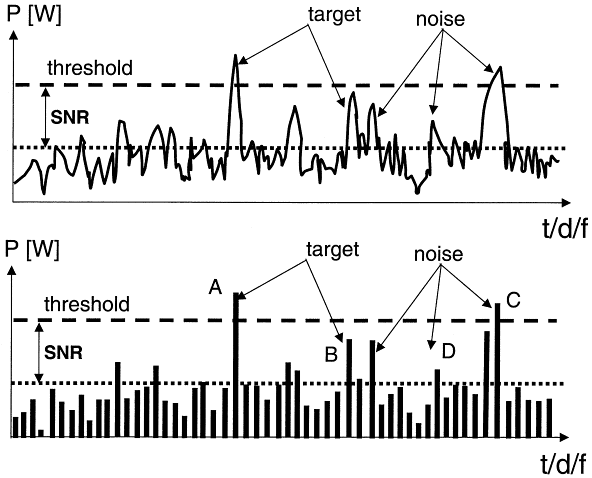


Figure 11.9 CFAR detection and decision making.

or voltage at the receiver, the horizontal axis may be time, range, or frequency (velocity). The horizontal dotted line represents the mean value of the noise floor. The horizontal dashed line is the CFAR detection threshold, which is well above the noise floor determined by a given SNR value. The main problem in CFAR detection is to determine the most suitable SNR value, which allows best decision making among the following four cases:

1. Target is present, and its echo is above the threshold (*true detection*).
2. Target is present, but its echo is below the threshold (*missed detection*).
3. Target is not present, but an echo is above the threshold (*false alarm*).
4. Target is not present, and no echo is above the threshold (*no action*).

For example, peak A, B, C, and D in Fig. 11.9 correspond to a true detection, a missed detection, a false alarm, and no action cases, respectively. The reliability of the CFAR detection unit depends on understanding the stochastic characteristics of random processes target, noise, and clutter, which are discussed in the following sections.

11.2.9. False Alarm Probability and ROC

For any threshold setting, there will be a corresponding probability that noise alone (and/or clutter) may exceed this threshold. When this happens, the radar will erroneously report a detected target at the corresponding position. The associated probability is therefore called *the false alarm probability*, P_{fa}. In a similar way, the probability of making a correct detection P_d can be associated for each chosen threshold. Since they both depend on the same threshold setting, it is clear that they are somehow related to each other.

If one assumes that at decision making instants the probability density functions of noise and noise plus target associated with the amplitude (y) distributions are $p_n(y)$ and $p_{s+n}(y)$, respectively, these probabilities are determined as

$$P_{\mathrm{fa}} = \int_{Y}^{\infty} p_n(y)\,dy \qquad P_d = \int_{Y}^{\infty} p_{s+n}(y)\,dy \tag{11.13}$$

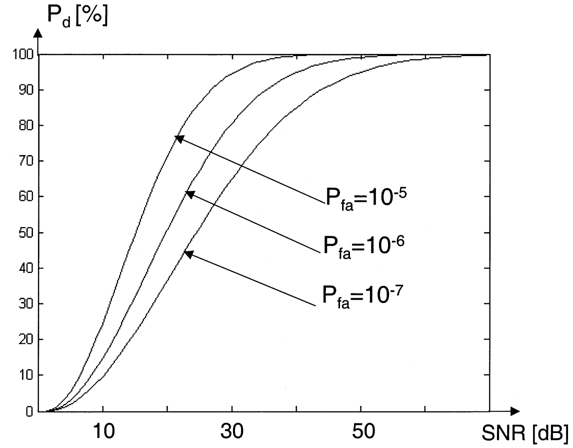


Figure 11.10 A typical radar ROC graph.

Above relations indicate that

Increasing the detection threshold Y will decrease both P_{fa} and P_d.
Decreasing the detection threshold Y will increase both P_{fa} and P_d.

which show that there is always a trade-off between improving detections and reducing false alarms. Since both P_{fa} and P_d change with the detection threshold level, it is possible to calculate a value of SNR which is required to achieve a certain detection probability P_d, for a given value of P_{fa}. In practice, P_{fa} is chosen first and then P_d is calculated or determined by using receiver operating characteristics (ROC) of the radar receiver. A typical ROC is illustrated in Fig. 11.10, where detection probability vs. SNR are plotted for different false alarm rates. Very low false alarm probabilities are used in radar systems as shown in the figure: as low as $P_{\text{fa}} = 10^{-6}$ to 10^{-8}. If the noise amplitude distribution is assumed to be gaussian, the ROC of a radar can be obtained analytically. If not, or if clutter determines CFAR detection threshold, then these graphs must be prepared either experimentally or numerically.

The determination of the ROC of a radar depends on a number of factors, including the statistical behavior of the target fluctuations and of the noise plus clutter, the number of pulse integrated coherently and incoherently. All these are extensively studied and well documented in the literature [9].

11.3. PROPAGATION ENVIRONMENT AND PATH LOSS PREDICTION

Radars operate above the earth's surface through the atmosphere and understanding EM scattering properties and propagation effects are critical [10–14]. A typical propagation environment is pictured in Fig. 11.11. Here, surface (on land and ocean) and air targets, land-based and airborne radars, which may operate at different frequencies from HF, VHF to MWs are illustrated. The propagation scenario may include the spherical earth's surface with nonflat terrain (mountains, valleys, etc.), rough surfaces (e.g., ocean surface or land irregularities and vegetation) and atmospheric variations (e.g., clouds, rain), all of which behave quite differently in different frequency regimes.

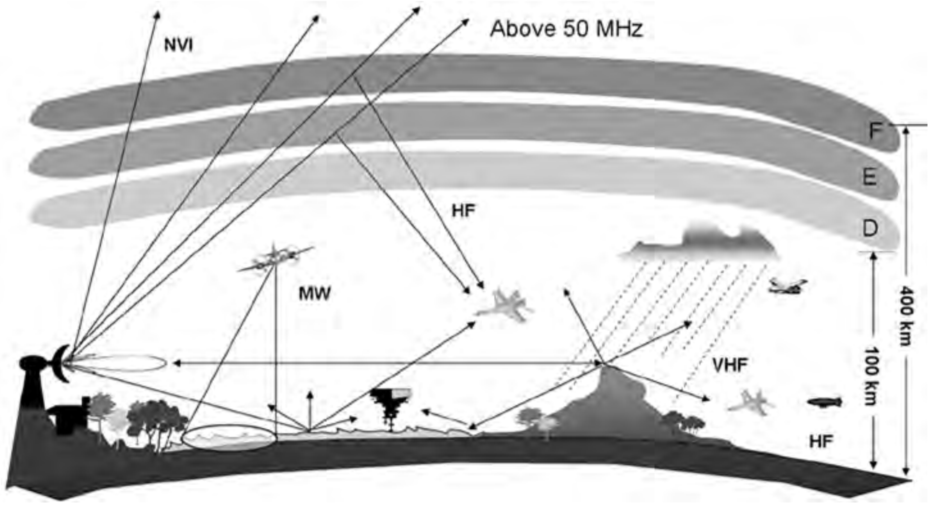


Figure 11.11 Radio wave propagation environment.

In general, propagation occurs via ground waves and sky waves. Ground waves have three components: direct waves, ground-reflected waves, and surface waves. Sky waves use high altitude atmospheric layers (i.e., D, E, and F layers of the atmosphere, called the *ionosphere*). The model environment is a spherical earth with various ground characteristics, above which exists a radially inhomogeneous atmosphere. Since the radar and the target may be anywhere on or above the ground, the propagation model may have different canonical features. The physical characteristics of propagation depend on many parameters, such as the operating frequency, medium parameters, transmitter and receiver locations, and the geometry (boundary conditions, BC) between them.

Depending on the mission and allowed frequency bands, propagation characteristics may be totally different. In Table 11.4, some of the frequency ranges and assigned missions are listed. The mission may be to defend a military base or airport against airborne attacks, to assist aircraft landing in a critical region surrounded by high hills, to assist in weather or oceanographic surveillance, a maritime navigation along a narrow waterway, or to measure the speed of a vehicle (police Doppler radar), height of a vehicle (altimeter), etc. The mission may also be to prepare agricultural maps (SAR), to identify ships (ISAR) or subsurface imaging (GPR). Propagation requirements for these missions are quite different from each other. Some applications need a range of hundreds of kilometers, others work over less than a meter. Generally speaking, radars may be classified into three groups in terms of their propagation characteristics: HF radars, VHF radars, and MW radars (Infrared and optical sensors are not mentioned here). Some of the propagation characteristics in terms of this classification can be listed as follows [10,11]:

For A/A and/or S/A microwave radars, it is essential to understand free-space path loss, first. This is not actual loss in fact and is nothing but a decrease in power density because of the spherical wave spread. *Free-space path loss* is defined as the power density of a unit power isotropic radar transmitter captured by an isotropic (unit gain) receive antenna with an effective aperture of $A_{e=}\lambda^2/4\pi$ and given by

$$Lfree = \left(\frac{4\pi d}{\lambda}\right)^2 = \left(\frac{4\pi df}{c}\right)^2 \tag{11.14}$$

where d, f, and λ are the distance between the transmitter and receiver, the frequency, and the wavelength of the radar carrier, respectively. By using d in kilometers and f in MHz, Eq. (11.14) can be rearranged in logarithmic form as

$$Lfree = 32.4 + 20 \log_{10} d_{\text{km}} + 20 \log_{10} f_{\text{MHz}}\,\text{dB} \tag{11.15}$$

For A/G, A/S, and S/S microwave radars the total propagation loss is more than the free-space path loss and may include

Absorption loss (due to atmospheric, ionospheric gases and precipitation)
Earth's curvature loss (due to the spherical nature of the ground, which diverges the waves)
Reflection, multipath, and scattering loss (due to ground irregularities)
Diffraction loss (due to the nonflat terrain along the propagation path)
Refractivity loss (due to ducting, guiding, and antiguiding effects of atmospheric layers, especially in the first few kilometers above the ground)
Depolarization loss (due to tropo-scattering effects), etc.

These are defined as *additional losses*, and the *total path loss* is usually defined as the free-space path loss plus additional losses.

For MW radars only line-of-sight (LOS) propagation is possible. *Line of sight* is defined as the distance when there is no obstacle between the transmitter and the target. Within LOS is the interference region where ground waves have all three components. Beyond the LOS is the diffraction region, and propagation is possible only by means of surface waves and/or sky waves (via ionospheric reflections).

Since S/S MW radars are limited by the LOS, the radar needs elevated platforms to overcome the earth's curvature effects. The LOS (in kilometers) of a MW radar with platform height h (m) can be calculated via [2]

$$\text{LOS} = 4.12\sqrt{h}\ \text{km} \tag{11.16}$$

For example, a S/S MW radar on a 40-m tower may cover ranges up to only 25–26 km.

For S/S, HF radars the frequency range of interest is between 3–30 MHz depending on the area of surveillance and the mode of propagation. Frequencies between 3–6 MHz are used for wide-area ocean surveillance (up to 400–500 km in range), while 5–15 MHz are good for 10–30 km (may reach up to 25 MHz for ranges of a few kilometers) when the surface-wave mode of operation is used. On the other hand, sky-wave mode of propagation may also be used for S/S wide-area surveillance up to a few thousands of kilometers.

For S/S, SWHF radars the transmitter and the receiver are close to surface; so direct and ground-reflected waves cancel each other, and only surface wave remains. The earth's electrical parameters are important in reaching longer ranges.

Table 11.5 Typical Ground Characteristics at MF and HF

Ground	σ (S/m)	ε_r
Sea	5.0	80.0
Medium ground	0.01	15.0
Poor ground	0.001	7.0

Sea surface is a good conductor, but ground is a poor conductor at these frequencies. For example, with the same transmitter and receiver characteristics, a 5-MHz signal, which reaches the 400-km range over the sea can only reach up to 40–50 km in range over poor ground. Typical electrical parameters (σ, conductivity, ε_r, relative permittivity) of ground are listed in Table 11.5 [3].

In the lower HF band (3–10 MHz), propagation well beyond the LOS is possible via surface waves. Surface waves are hardly excited and coupled to the surface, when the transmitter is many wavelengths (e.g., 3λ–4λ) above the ground. When excited and coupled, they exponentially decay with height. Surface waves are rarely used above 10–15 MHz.

At higher HF frequencies (15–30 MHz), beyond the LOS coverage is only possible by means of the sky waves. The lower layer of the ionosphere (D layer) absorbs EM waves and causes extra loss. The higher layers (E and F) bend EM waves toward the earth's surface, act as a reflecting upper boundary and form a kind of earth–ionosphere waveguide. This waveguide can be used at most up to the frequencies of 45–50 MHz. Beyond 50-MHz EM waves are not bent and escape into outer space.

At lower VHF frequencies (50–150 MHz), propagation beyond the LOS is still possible by means of diffracted EM wave components (typically, 5–20 km beyond obstacles).

At upper VHF frequencies and above, (i.e., for frequencies 200 MHz and above) propagation is limited by the LOS, because surface waves are negligible at these frequencies.

Ground-wave propagation through atmosphere (up to nearly 100 GHz) is affected by oxygen and water-vapor molecules. The air can be considered as a nondispersive medium and can be represented by its refractive index ($n = \sqrt{\varepsilon_r}$). The refractivity of the propagation medium should be well understood, since nonflat terrain and/or earth's curvature may also be implemented via refractivity in most of the analytical as well as numerical approaches. Refractive index of the air is very close to unity (e.g., 1.000320); therefore, it is customary to use the refractivity N, defined as

$$N = (n - 1) \times 10^6 \tag{11.17}$$

N is dimensionless, but is measured in "N units" for convenience. N depends on the pressure P (mbar), the absolute temperature T (K) and the partial pressure of water vapor e (mbar) as [12]

$$N = 77.6\frac{P}{T} + 3.73 \times 10^5 \frac{e}{T^2} \tag{11.18}$$

which is valid in earth–troposphere waveguides and can be used in ground-wave propagation modeling. If the refractive index were constant, radio waves would propagate in straight lines. Since n decreases with height, radio waves are bent downward toward the earth, so that the radio horizon lies further away than the optical horizon (i.e., LOS). It should be noted that the radio horizon effect is taken into account either by using N with the effective earth radius a_e or by introducing a fictitious medium where N is replaced by the modified refractivity M [12],

$$M = N + \frac{x}{a} \times 10^6 = N + 157x \tag{11.19}$$

with the height x given in kilometers. In Eq. (11.19)

$$a = 6378\ \text{km} \qquad \frac{10^6}{a} = 157 \qquad \text{and} \qquad \frac{\partial M}{\partial x} = \frac{\partial N}{\partial x} + 157 \tag{11.20}$$

For the standard atmosphere (i.e., for a vertical linearly decreasing refractive index), N decreases by about 40 Nunit/km, while M increases by about 117 Nunit/km. Subrefraction (superrefraction) occurs when the rate of change in N with respect to height (i.e., $\partial N/\partial x$) is less (more) than 40 Nunit/km.

A linearly decreasing (increasing) vertical refractive index variation forces the waves to be trapped near the earth's surface while propagating. Similar effects are also caused by concave and convex surfaces. Therefore, there is an analogy between refractive index and surface geometry in terms of propagation effects. By using this analogy, earth's curvature effect is included into the refractive index of the air. Earth's curvature effect is equivalent to a vertical refractivity gradient of 157 Nunit/km (i.e., linear vertical increasing refractivity profile).

Characteristic features of the EM wave propagation are graphically illustrated in Figs. 11.12–11.14. In Fig. 11.12, both refractivity and nonflat terrain effects are shown at 30 MHz as a range–height field strength color map, when the transmitter is on the ground [11]. The ducting effect is clearly observed in the second graph under a trilinear refractivity variation. In Fig. 11.13, surface-wave path loss vs. range is plotted at three different frequencies along the ocean surface with a 50-km length island at a distance of 250 km from the radar. The radar transmitter and the receiver heights are zero. The dashed curves are for a homogeneous path. At 100 kHz, the island hardly affects the smooth path loss variation. But in the HF band, the path loss increases substantially over land, and signal recovery is observed behind the island. Finally, normalized vertical fields (propagation factor) vs. height are plotted in Fig. 11.14. This corresponds to propagation over a spherical earth with standard atmosphere with a 3-GHz transmitter that is located 31 m above the surface. The five curves correspond to ranges of 20, 30, 40, 50, and 60 km, respectively. The ground is assumed to be a perfect electric conductor (PEC). A 10-dB scale is also given in the figure. The 31-m transmitter height places the first four range curves into the region of interference between the direct and ground reflected waves (within LOS) as evident in the plots. At the 60-km range, heights up to 200 m are in the shadow region below the LOS and no interference is observed [11].

11.4. RADAR EQUATION

A derivation of radar equation is best understood with the case of an isotropic transmitting antenna. If the average transmitted power is P_t (W), then the power density

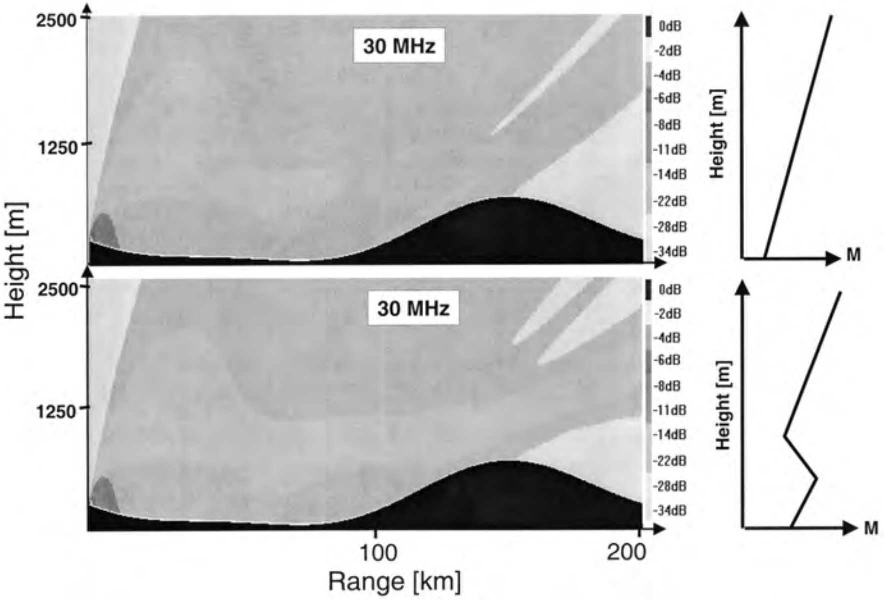


Figure 11.12 Propagation over earth's surface.

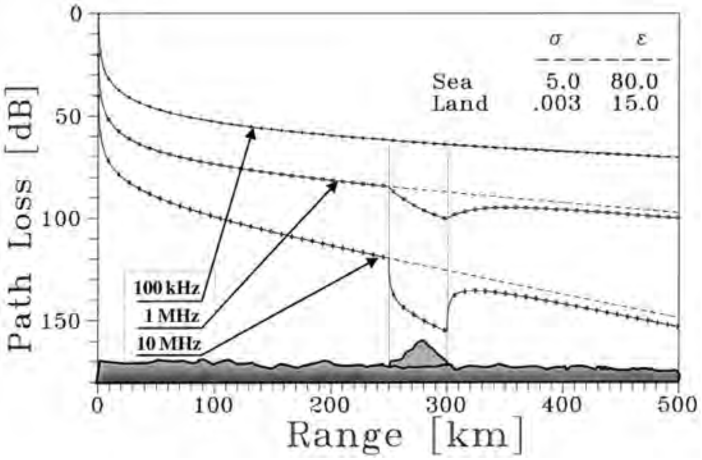


Figure 11.13 Surface-wave path loss vs. range.

at the target located at a distance R (m) is

$$\mathrm{PD} = \frac{P_t}{4\pi R^2}\ \mathrm{W/m^2} \tag{11.21}$$

If one uses a directive antenna, the power density at the target is given by

$$\mathrm{PD} = \frac{P_t}{4\pi R^2} \times G_t\ \mathrm{W/m^2} \tag{11.22}$$

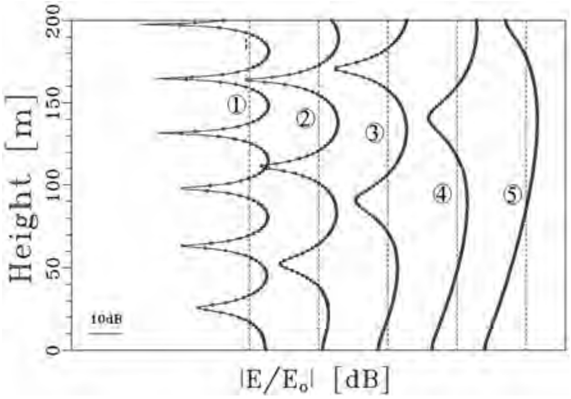


Figure 11.14 Propagation factor vs. height at 3 GHz.

where G_t is the transmit antenna gain in the direction of the target. The transmitted signal interacts with the target and is reflected back to the radar. The power density of this echo signal back at the radar is

$$\mathrm{PD} = \frac{P_t}{4\pi R^2} \times G_t \times \sigma \times \frac{1}{4\pi R^2}\ \mathrm{W/m^2} \tag{11.23}$$

where σ (m^2) is the radar cross section of the target. The echo is received by the receive antenna of the radar and the power received is then given by

$$P_r = \frac{P_t}{4\pi R^2} \times G_t \times \sigma \times \frac{1}{4\pi R^2} \times A_e\,\mathrm{W} \tag{11.24}$$

where A_e (m^2) is the effective receiving antenna aperture. The effective antenna aperture depends on the operating frequency and the antenna gain, mostly as

$$A_e = \frac{G_r \lambda^2}{4\pi}\ \mathrm{m^2} \tag{11.25}$$

where G_r is the receive antenna gain in the target direction. As given in Eq. (11.11), range is inversely proportional to power. If the minimum detectable (received) power $P_{\min}$ is defined as

$$P_{\min} = P_t \times G_t \times G_r \times \sigma \times \frac{\lambda^2}{(4\pi)^3 R^4}\ \mathrm{W} \tag{11.26}$$

then maximum range can be obtained as

$$R_{\max} = \left[\frac{P_t G_t G_r \sigma \lambda^2}{(4\pi)^3 P_{\min}}\right]^{1/4}\ \mathrm{m} \tag{11.27}$$

which is called the simplest form of *radar equation*. If Eq. (11.26) is rewritten in terms of free-space path loss L_{free}

$$P_{\min} = \left[\frac{P_t \times G_t \times G_r \times \sigma \times 4\pi}{\lambda^2 L_{\text{free}}^2}\right] \text{W} \tag{11.28}$$

is obtained. It should be remembered that the basic radar equation is derived for S/A and/or A/A surveillance (i.e., in free space). When A/S or S/S is of interest surface wave attenuation factor A is introduced and may be combined with the free-space path loss, yielding a total one way propagation loss, L_p as [3]

$$L_p = \left(\frac{4\pi R}{\lambda A}\right)^2 \tag{11.29}$$

The signal at the radar receiver fluctuates for a variety of reasons, which makes the total radar echo a random process. The total radar echo contains targets, noise, clutter, and other interfering signals (such as jamming signals, intentional radio and communication broadcast signals). The minimum required target signal is defined via signal-to-threshold ratio, where the threshold is usually determined either by noise or by clutter. Therefore, the radar equation is generally given as signal-to-noise ratio (SNR) or signal-to-clutter ratio (SCR).

11.4.1. Noise-Limited Detection

Electronic circuits and receivers are affected by a variety of noise sources. Typical noise sources and their frequency variations are illustrated in Fig. 11.15. Here, the horizontal axis is the frequency in log scale and the vertical axis is the electric field in dBμV/m for

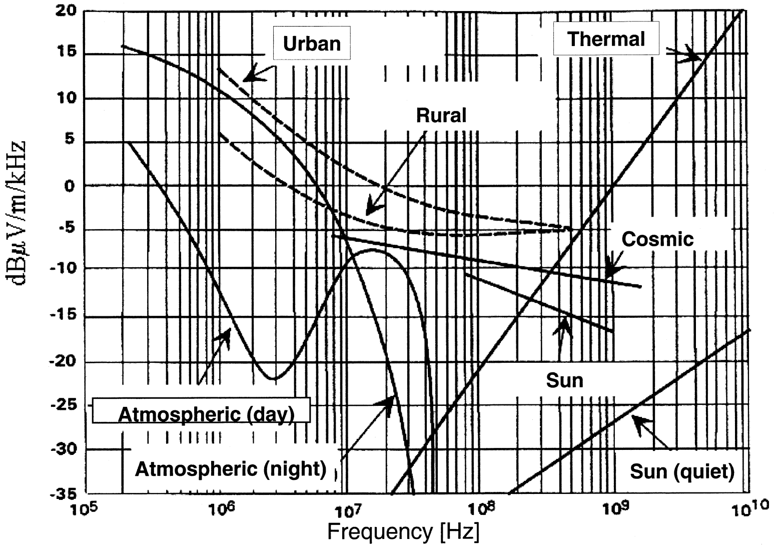


Figure 11.15 Characteristic noise types and frequency regions.

a 1-kHz bandwidth. In terms of radar engineering, noise can be classified into two groups: thermal noise (also called *internal noise*, caused by electronic devices themselves) and the environmental noise (atmospheric, cosmic, man-made, etc.). As shown in the figure, thermal noise dominates the others for frequencies above a few hundred MHz. Therefore, noise limited radar detection can be grouped into thermal noise limited detection (e.g., MW radars) and environmental noise limited detection (e.g., HF radars).

Thermal noise (which is also called *white noise*) is directly proportional to the receiver bandwidth and can be calculated as

$$N_t = kTB\,\mathrm{W} \tag{11.30}$$

where B (Hz) is the noise bandwidth, k is Boltzmann's constant ($k = 1.38\,10^{-23}$ J/K), and T is the temperature in Kelvin. Thermal noise can be considered to be white noise (i.e., gaussian amplitude distribution and flat power spectral density, $S(f) = kT$). Since $kT = -204$ dBW/Hz at 300K, thermal noise can also be calculated as

$$N_t = -204 + B_{dB_{Hz}}\,\mathrm{dB} \tag{11.31}$$

On the other hand, the environmental noise (daytime and nighttime atmospheric noise and man-made noise in rural and urban areas) determines the noise floor at HF frequencies. As shown in the figure, nighttime atmospheric noise level is much higher than the daytime level, especially at lower HF frequencies, which degrades the performances of HF radars at night. Environmental noise can be expressed by the empirical formula [6]

$$N_e = \left(N_0 - 12.6\ln\frac{f_{MHz}}{3}\right)B\,\mathrm{dBW} \tag{11.32}$$

where, N_0 is equal to man-made noise level of the selected radar site (typically, -136, -148, -164 dBW/Hz in residental, rural, or remote sites, respectively) and the term inside the parentheses is the noise density given as dB/Hz.

As a result, radar equation for a noise limited detection case can be written as

$$\mathrm{SNR} = \frac{P_t \times G_t \times G_r \times \sigma \times 4\pi}{\lambda^2 L_p^2 N} \tag{11.33}$$

where N is N_t or N_e, depending on whether the radar is a MW (thermal noise limited) or an HF (environmental noise limited) radar, respectively. It is important to note that noise is present at the input of the radar receiver, is range and operating frequency independent, and is proportional to the receiver bandwidth. The narrower the receiver bandwidth, the lower the noise threshold and higher the SNR. It should be noted that Eq. (11.33) is a basic radar equation and may change due to the usage of peak power, coherent and/or incoherent integration, etc.

In order to find out the maximum range from Eq. (11.33), the first step is to determine the detection probability and the false alarm rate. These two determine the minimum SNR, which can be extracted form the ROC of the radar receiver prepared for different target types, number of coherent and incoherent integration, etc. Once SNR is determined, $R_{\max}$ can be calculated from Eq. (11.33) together with Eq. (11.29).

11.4.2. Clutter-Limited Detection

Unlike noise, which is present inherently in the receiver, clutter is a target like echo signal that comes from many small scatterers, such as rain droplets, birds, ocean waves, terrain irregularities, vegetation, aurora, meteors, etc.; therefore, it is a radar parameter and range dependent. When clutter power dominates over noise (i.e., when SNR $\gg$ SCR) the radar is said to operate in a clutter limited condition. In this case SCR is calculated as

$$\mathrm{SCR} = \frac{\sigma_t}{\sigma_c} \tag{11.34}$$

where, σ_t and σ_c are the RCS of a target and clutter, respectively. Clutter may occur as distributed clutter, which increases with radar resolution, and as point clutter, which does not. Clutter characteristics and surface and volume clutter calculations are discussed in the next section.

11.5. RADAR SIGNAL ENVIRONMENT

A total radar echo usually consists of (1) target, (2) noise, (3) clutter, and (4) interference signals, all of which randomly fluctuate with time. This means, a radar signal environment is a stochastic environment. Usually, the target signal is embedded within a background (noise + clutter + interference), its power level is much less than the others and it is extremely difficult to extract it. The process of extracting useful information (generally the target) from the total echo is called (stochastic) *signal processing* and performed via powerful, intelligent algorithms. The power of these algorithms arise from the physical understanding of the target, noise, clutter, and interference signals.

11.5.1. Target

A radar target is characterized by its EM reflectivity. This reflectivity is called the *radar cross section* (RCS) σ and is defined as

$$\sigma = \mathrm{RCS} = \lim_{R\to\infty} 4\pi R^2 \frac{|E_s|^2}{|E_i|^2}\ \mathrm{m}^2 \tag{11.35}$$

where R, E_i, and E_s are the distance, electric fields of the illuminating and target-scattered waves, respectively. Spherical coordinates are of interest in RCS calculations, so both incident and scattered fields may be either E_θ or E_φ. Also, RCS is a far field concept and has to be measured and/or simulated at ranges sufficiently far from the radar transmitter. In other words, the illuminating wave has to be a plane wave. σ has the dimension of area in m^2, or in dB m^2 (referred to 1 m^2).

RCS of a target is classified according to the types of radars (monostatic and bistatic radars), and the polarization of the transmitter and the receiver. In most of practical cases, RCS of a target is mentioned for the four cases listed in Table 11.6, when RCS of a target is of interest these parameters should be given:

Angle of incidence (θ_i,φ_i) and angle of scatter (θ_s,φ_s)
Incident and scattered field polarizations

Table 11.6 Co- and Cross-Polarized RCS Cases

Linear[a]	VV	VH
	HH	HV
Circular[b]	RL	RR
	LR	LL

[a] V: vertical, H: horizontal.
[b] R: right, L: left.

Frequency and target geometry (size)
RCS value (in m^2 or dB)

Depending on the radar operating frequency (i.e., wavelength) and size of a target, RCS of a radar target falls into one of three characteristic regimes (where qualitative as well as quantitative differences occur) [15];

Low frequencies (*Rayleigh region*) where target dimensions (l) are much less than the radar wavelength ($l \ll \lambda$). In this region a radar target acts as a point reflector, and its RCS is proportional to the fourth power of frequency ($\sigma \approx f^4$).

Medium frequencies (*resonance region*) where target dimensions and the radar wavelength are of the same order ($l \approx \lambda$). In this region, the target contributes to its RCS as a whole, (which is called *bulk RCS*), therefore mathematical RCS calculations are almost impossible for targets with complex geometries. Fortunately, there are powerful time and frequency domain RCS tools in this regime [16]. Aircraft and ships have dimensions (typically tens of meters) that put them in the resonant scattering regime for HF radars, for example (where radar wavelengths are also of the order of tens of meters).

High frequencies (*optical region*) where target dimensions are very large compared to the radar wavelength ($l \gg \lambda$). For example, aircraft and ships mentioned above fall in this region for the microwave radars (where radar wavelengths are of the order of centimeters). In this region, RCS is roughly the same size as the real area of target. Local characteristics within the area of illumination usually dominate the target RCS and high-frequency asymptotic techniques, such as geometric optics (GO), geometric theory of diffraction (GTD), physical optics (PO), physical theory of diffraction (PTD), and uniform theory of diffraction (UTD), can be applied [17] in this regime.

Spatial and temporal RCS fluctuations differ from target to target. The fluctuations may be slow (correlated within seconds, e.g., a tanker ship navigating on a calm sea) or fast (correlated only in miliseconds, e.g., a maneuvering high speed fighter aircraft); the fluctuating RCS values may be distributed or may accumulate around a few dominant scatterers. These are categorized in terms of probability distribution functions and are grouped into three as

Steady targets like mountains, buildings, etc. (TYPE 0)
Group of small scatterers (without dominant contribution) (slow, TYPE 1 and fast, TYPE 2)
Group of scatterers with a few dominant ones (slow, TYPE 3 and fast, TYPE 4)

These types are also called *Swerling types* (SW) and determine radar waveform as well as echo integration process in the radar receiver. For example, since SW 1 and SW 3 targets (which have a decorrelation time of seconds) are stationary from pulse-to-pulse (with PRI of miliseconds) pulse-to-pulse coherent integration can be applied. Target echo fluctuations act upon detectability factor and, in general, are considered a loss factor. They are included as a fluctuation loss in the radar equation. Fluctuation loss is small (large) for low (high) P_d and heavy for SW 1 and SW 2 type targets.

In general, RCS of a target cannot be given in terms of simple mathematical functions, and either powerful numerical simulation methods or real measurements are used to obtain RCS values of different targets. On the other hand, simple relations can be used for simple geometries in optical RCS regime. For example, the RCS of a perfectly electrical conductor square plate is given as

$$\mathrm{RCS} = \frac{4\pi a^2}{\lambda^2}\ \mathrm{m}^2 \qquad \text{(for vertical illumination)} \tag{11.36}$$

where a is the edge dimension and λ is the radar wavelength. At 3 GHz (with S-band radar), a 1-m^2 PEC square plate yields nearly 1200 m^2 (~31 dB) RCS and increases to more than 10,000 m^2 (~40 dB) at 10 GHz (X-band radar).

RCS of a target may be given as a figure (or table), where, for example, RCS vs. frequency is given, or, as radial plots, where mono- and/or bistatic RCS variations at different frequencies are plotted. A typical example is given in Fig. 11.16, for a navy frigate model calculated via the FDTD technique [15,16].

11.5.2. Noise

Sensitivity of a radar receiver shows the lowest signal level that is measurable with the device (which is called floor signal). Usually, the receiver sensitivity, i.e., the floor signal, is determined by either internal or external electromagnetic disturbances. Electromagnetic

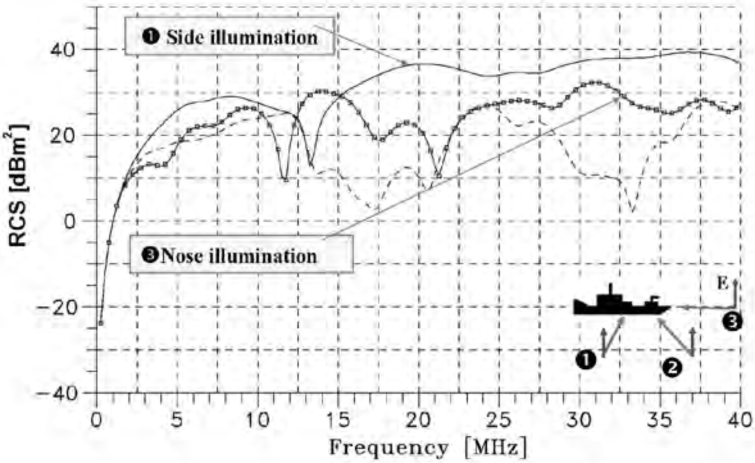


Figure 11.16 RCS vs. frequency of a 45 m-long, 20 m-wide navy frigate model.

disturbances caused by internal and/or external, man-made or natural are generally called noise. For example, man-made sources, such as power lines, electric razors, hair-dry machines, etc., and, natural sources, such as, lightning, electrical storms, galactic effects, etc. are typical external noise sources. On the other hand, power amplifiers, mixers, diodes, and transistors are some of the internal noise sources, because of random collisions of the electrons. Various noise sources and their effective frequency ranges are shown in Fig. 11.15.

Roughly speaking, noise can be classified into internal and external in most of the radar systems. Internal noise, which is also called *thermal noise* (as explained in Sec. 11.4.1), generally limits the detection threshold in MW radars (typically at frequencies 100 MHz and above). External noise determines the detection threshold at HF and partially VHF radars.

The amount of noise that is present at the input of a radar receiver depends on various factors, such as receiver bandwidth, antenna radiation pattern, antenna side lobes, or objects of illumination. In practical radar receivers, the noise level is found to be more than the level at the input of the receiver by a factor known as *noise figure*. The ratio of the noise present at the output of the radar receiver to the noise due to thermal effects alone is called the *receiver's noise figure*. Noise figure is also an important parameter that determines the maximum range in the radar equation given above.

Noise is a random signal and must be handled in a stochastic manner. Its amplitude distribution characteristics, average value, deviation, frequency characteristics (power spectrum), etc. must be well understood when dealing with detection theory. The distinguishing characteristic of the noise is that it is a pulse-to-pulse uncorrelated signal (usually called as white noise or gaussian noise). Signal correlation in time domain and power spectrum in frequency domain forms a Fourier pair. An uncorrelated (a delta type) function in time domain corresponds to a constant value in frequency domain. Therefore, the larger the receiver bandwidth, the higher the noise level in radar receivers.

The sensitivity of a radar is determined by its ability to maximize the SNR of the received echo. In other words, noise must be minimized while amplifying the target signal. The probability that a noise spike will reach a certain level at the output terminals of the receiver is given by a Rayleigh distribution function. The probability of a target signal imbedded in a noise is given by a gaussian distribution. Using these two distributions (under noise limited detection conditions), detection probabilities and false alarm rates can be calculated mathematically.

Noise elimination in a radar receiver can be achieved by integration, either in predetection or postdetection stages [2].

11.5.3. Clutter

Clutter is a word used to describe all unwanted echoes in a radar receiver. Clutter can be characterized as a distributed nondirectional source. Depending on the "mission" of a radar, what is clutter in one application may not be so in another. For example, a radar designed to detect aircraft includes the echoes from land, sea, clouds, rain, birds, insects, etc., which are all called *clutter*. On the other hand, aircraft is one of the clutters for an HF radar designed for oceanographic surveillance. Similarly, backscatter echoes from land can degrade the performance of many radars (as land clutter) but represent the target of interest for a ground-mapping radar.

Unwanted echoes usually occur as distributed clutter, as surface clutter (such as land and sea echoes) or as volume clutter (such as rain, chaff). Because of its distributed nature, clutter is characterized in terms of RCS density, rather than the RCS as described for conventional radar targets (ships, aircraft, etc.). For surface clutter, the average RCS density σ_0 , the RCS per unit area, is given by the ratio

$$\sigma_0 = \frac{\sigma_c}{A_c} \mathrm{m}^2/\mathrm{m}^2 \tag{11.37}$$

where σ_c is the RCS of the area A_c. Similarly, volume clutter RCS density is given as the average RCS of a unit volume V_c as

$$\eta_0 = \frac{\sigma_c}{V_c} \mathrm{m}^{-1} \tag{11.38}$$

where σ_c is the RCS of a unit volume V_c. The values of both surface and volume RCS densities depend on many factors, such as, the type of terrain observed, the direction of illumination and observation, radar wavelength, polarization. For example, the ocean RCS density depends on ocean wave height, wind characteristics (direction and speed), and wave direction. Ground RCS density depends on soil type, surface roughness, foliage cover, etc. Typical RCS densities, for example, for ocean are −80 dB at S band for 0.1° grazing angle for vertical polarization, but increases to −45 dB at X band for 3° grazing angle for the same polarization. On the other hand, values up to −30 dB and −35 dB are used for HF frequencies.

The geometry of surface clutter is pictured in Fig. 11.17. With the parameters mentioned in the figure (i.e., pulse length τ, range R, azimuth beam width φ, elevation angle θ) the clutter RCS can be given as

$$\sigma_c = \sigma_0 R\varphi \frac{c\tau}{2} \sec\theta \, \mathrm{m}^2 \tag{11.39}$$

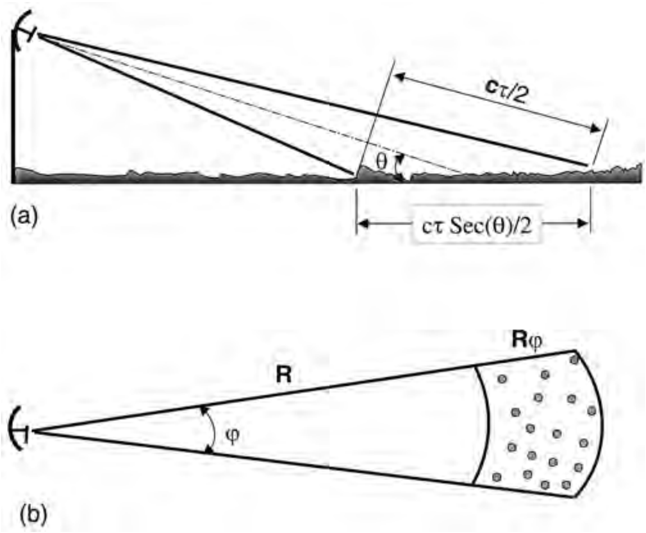


Figure 11.17 Geometry of a radar clutter: (a) elevation view and (b) plan view.

and the signal-to-clutter ratio in Eq. (11.34) reduces to

$$\mathrm{SCR} = \frac{\sigma_t}{\sigma_0 R\varphi(c\tau/2)\sec\theta} \tag{11.40}$$

Ocean and land clutters are basic surface clutters that determine the performances of surveillance radars. They are also stochastic processes and are characterized by temporal as well as spatial distribution characteristics. Both are pulse-to-pulse coherent signals, so elimination by just averaging (postdetection integration as done for noise) is not possible. Usually, clutter elimination is performed in FD. Land clutter occupies a very low frequency range around zero Doppler frequency. On the other hand, ocean clutter has different Doppler characteristics at different radar frequencies.

A typical example is given in Fig. 11.18a, where (synthetically produced) Doppler characteristics of the ocean clutter are pictured for a HFSW radar. Ocean clutter is the result of the interaction of the radiated electromagnetic wave with ocean waves [6,7]. The dominant contribution is produced by scatter from ocean waves having a wavelength half that of the radar wavelength and moving radially to and away from the radar site. This first-order resonant scatter results in two dominant peaks called *Bragg lines* [1,2]. Ocean waves are trochoidal and Bragg resonant scatter will also occur at harmonics of the principal wavelength. These result in second order peaks in the spectrum. Another source of second-order scatter is the interaction between crossing ocean waves. If these crossing ocean waves generate a third ocean wave, with a wavelength equal to one-half the radar wavelength, then Bragg resonance scatter will occur. It is this condition that leads to an increase in the continuum level between the Bragg lines in the Doppler spectrum and is referred to as the second-order continuum. The energy contained within the second-order continuum is related to the sea state and hence surface wind speed and duration.

In Fig. 11.18a, the operating frequency is 5 MHz. This yields the dominant Bragg frequencies of ± 0.228 Hz that are normalized to ±1 Hz. For this operating frequency, the blind velocities (they correspond to ocean waves speed resonating at Bragg

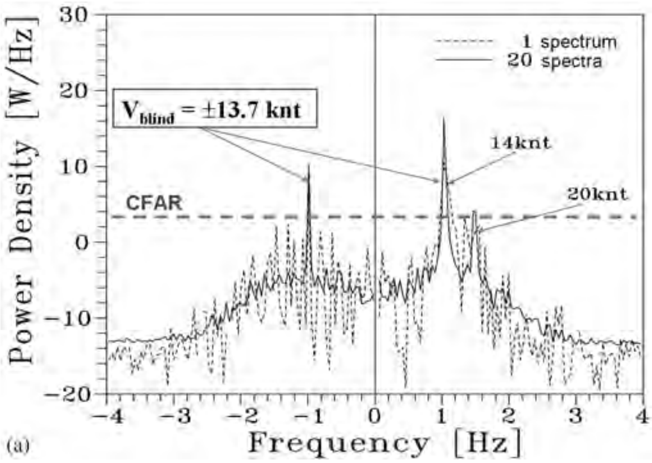


Figure 11.18 (a) Typical (synthetic) HF ocean spectra at 5 MHz. (b) Typical (real) HF ocean spectra at 3.6 MHz.

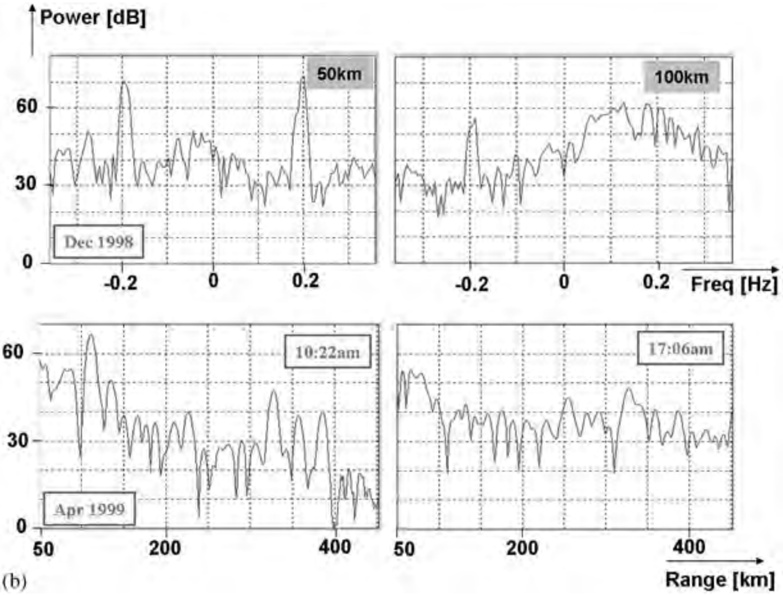


Figure 11.18 Continued.

frequencies) are ±13.7 kn (1 kn ≈ 0.5 m/s). The clutter to noise ratio is 30 dB. Two surface targets with 20 dB and 15 dB signal to noise ratios and 14 kn and 20 kn radial velocities, respectively, are included in the spectrum. Dashed and solid lines correspond to one spectrum and average of consecutive 20 spectra, respectively. The reduction in noise floor by spectrum averaging is clearly observed in the figure. The dominant Bragg returns at ±1 Hz, second-order continuum and the target with 20 kn velocity are clearly seen in the figure. The other target with 14 kn radial velocity is obscured by the dominant Bragg return at 1 Hz.

Usually, the real Doppler spectra are not as pure as the simulated ones, which means detection decisions are really hard to give. Typical real Doppler spectra recorded between 1998 and 1999 is given in Fig. 11.18b (recorded at Cape Bonavista with a HFSW radar operated at 3.6 MHz). On top, Doppler spectra at two different radial ranges along a chosen beam is shown. At the bottom, range profile at fixed radar beam, at two different times are plotted. A threshold level of 20 dB below the thermal noise floor is chosen for the vertical axis for the peak signal power level. The complexity of ocean clutter spectrum is clearly observed in these plots.

There are many differences between ocean and land clutter. When compared to ocean clutter, land clutter is less time dependent, but backscatter from land is significantly greater than ocean clutter in most cases.

11.5.4. Interference

Interfering signals received within the total radar echo may be classified into different groups:

Intentional broadcast and communication systems and their harmonics
Local and remote cochannel signals

Intentional jamming signals (ECM, electronic counter measures)
Unintentional multipath arrivals (e.g., ground reflected target echo or signals from mutual interactions between different nearby targets)
Scattered signals from atmospheric discontinuities, etc.

These interfering signals have different characteristics, which determine the techniques that may be applied in interference cancellation. For example, radio broadcast signals may be narrow band and direction dependent, therefore, the interfering signal may appear in one azimuth beam along all the range gates and may not in nearby beams. Applying a correlation process between the cells of different beams and/or along range cells may be an effective interference cancellation technique. On the other hand, jamming signals are high power, broad band signals; therefore, angular and/or range correlation may not be effective. In this case, correlation in the frequency domain may be a solution.

Cochannel interference, coming from local or remote sources, may also be a problem in the detection process. Local interfering signals are generally from known sources and interference can be avoided by choosing alternate frequencies. Interference from distance sources poses a more serious problem in that it is more random in time and frequency.

Interfering signals depend also on the type of the radar in operation. For example, ionosphere is a severe interference path for HFSW radars. Not all the energy emitted by the HFSW radar propagates along the surface. Some energy is directed upward and, as with short-wave radio, may, under certain conditions, reflect from the ionosphere. In some cases, the energy reflected from the ionosphere returns to the radar. This signal may be viewed as multipath clutter or self-interference. Ionospheric self-interference may be divided into two main categories, specifically, *near vertical incidence (NVI) clutter* and *range folded clutter*. With NVI clutter, the HFSW radar signal travels vertically from the radar and is reflected from an ionospheric layer directly back to the radar. Range folded clutter occurs when the signal is directed at an angle other than vertical. After reflecting from the ionosphere the signal travels outward whereupon it reflects from the sea or land and returns along the same path, or via the surface wave. Given the geometry of the problem, the total path length of the returned signal places it at a range outside the system maximum set by the PRI. In effect, the HFSW radar will receive returns from previous pulses while collecting data from the current transmit pulse. NVI self-interference appears at a narrow band of ranges corresponding to the height of an ionospheric layer. One option for combating NVI self-interference is frequency agility. By increasing the frequency, the layer-critical frequency will be exceeded and the HFSW radar signal will penetrate through the layer. Similarly, the HFSW radar can be operated during the daytime at a frequency that does not support skywave propagation.

11.6. PARAMETER SELECTION FOR SURVEILLANCE RADAR

Radar surveillance is to maintain cognizance of selected traffic within a selected area, such as an airport terminal area, air route, critical mountains inside a military conflict region, coastal regions for offshore security, and waterways or narrow straits for vessel traffic management. This may be achieved with a single radar or may require a group of sensors with different types and numbers. One typical scenario is pictured in Fig. 11.19.

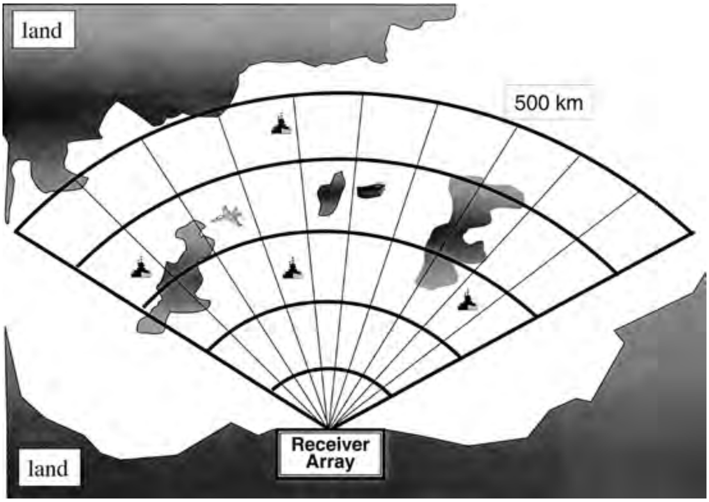


Figure 11.19 SWHF radar coastal coverage.

Suppose a wide area is to be monitored as given in the figure. Basic requirements and fundamental parameters may be listed as follows:

The requirement may be a tactical coverage for military purposes in a high-conflict ocean area or cruise and tanker traffic monitoring in a heavy traffic region. There may be beautiful islands, fishing regions, and/or petroleum drilling regions. The area may be along an international high-density surface transportation route. Depending on the scenario types and number of radars, their locations may be quite variable. For example, real time, continuous monitoring is essential for most of military purposes. On the other hand, off-line monitoring with a few hours update may be adequate for monitoring illegal fishing or cruise transportation, etc.

In the scenario in Fig. 11.19, suppose the region, up to 500 km in range and 120 degrees in azimuth is to be covered and both surface and air targets are to be monitored continuously. This is a typical scenario for countries with long coastal regions. The United Nations Convention on the Law of the Sea (UNCLOS) gives coastal nations sovereign rights over 200 nautical miles (nm) of sea known as the Exclusive Economic Zone (EEZ). In return countries are required to establish and maintain Administration, Law Enforcement, and Environmental Protection over this new frontier that is many times larger than their previous 12 nm territorial limits.

What sensors are available to monitor these typical regions? Traditional land-based MW radars are limited to operate within LOS. Even by elevating the radar platform the maximum range is limited to 50–60 km. The EEZ can be covered by a number of airborne radars, but these provide only a snap shot in time of activity within the EEZ. Sky-wave high-frequency (HF) radars can be used for this purpose, but they need large installations, are expensive and detection of surface targets is still limited. Satellites have neither the spatial nor the temporal resolution to provide the necessary level of real-time surveillance.

An optimal solution may be an integrated maritime surveillance (IMS) system that uses multitype multisensors. Effective surveillance also requires the integration of data from a number of complementary sensors. The primary sensor for this scenario may be one or two HFSW radars that are capable of tracking both surface and airborne targets at ranges in excess of 200 nm. The radar data may be enhanced with target identification obtained from automatic identification systems, such as Automatic Dependent Surveillance (ADS) systems [6,7], IFF, as well as information obtained from patrol vessels, communications, mandatory reporting procedures, etc. This information is associated with the radar data to provide a complete, real time, picture of activities within the EEZ.

Once types and number of radars are determined, power requirements, receive and transmit antenna systems, optimal radar waveform, signal and data processing hardware, and software, display utilities, etc., are to be taken into account for the types of targets to be monitored.

11.7. HFSW RADAR-BASED WIDE-AREA SURVEILLANCE

The region in Fig. 11.19 may be covered with one SWHF, located along a shore as pictured. It should be long range (up to 500 km) radar with wide azimuth coverage capability. HFSW radars use the lower end of HF frequency band (3–6 MHz) to provide the required coverage. A typical HFSW radar site is pictured in Fig. 11.5a. A broad band transmit antenna is located 100–200 m away from the receiver array that consist of 16 subarrays [7]. The arrays are located parallel to the shoreline with a clear field of view of the desired coverage area and may be located on a beach or cliff. The receive array yields a nominal beam width, at boresight, of approximately 5–10 degrees. For optimum performance the radar system should be located at an electrically quite area (where environmental noise level is as low as required) as defined by International Radio Consultative Committee (CCIR) [6,7].

The antenna system for an HFSW radar is designed to satisfy a number of criteria. The transmit antenna must provide a high gain over the specified band. Energy must be distributed equally and only over the desired surveillance area. The receive array must be parallel to the shore line, have high and equal array gain over the entire surveillance area with minimum sensitivity to signals arriving from other directions. Both transmit and receive arrays must also provide a deep, broad, null at NVI. When predicting the performance of both the transmit and receive arrays the effect of local site topography and ground conditions must be considered. Operating at the low end of the HF band requires that the receive array occupies a significant shoreline area, with the aperture of the array inversely proportional to frequency. For example, at 3 MHz, a 5° azimuthal beam width requires an array aperture of approximately 1 km.

The HFSW detects targets in three-dimensional space: range, bearing, and velocity (Doppler):

1. The *range resolution* is directly proportional to the bandwidth of the transmitted waveform. HFSW radar typically operates with a maximum bandwidth of 10–20 kHz, for which the resolution is about 8–15 km (note that several MHz bandwidths are used in MW radars where a few centimeters of range resolutions are of interest).

2. The *azimuthal resolution* is directly proportional to the aperture size of the antenna. The aperture is measured in terms of the radar wavelength. For an antenna array with element separation of half a wavelength, 16-element to 24-element antenna arrays produce beam widths at boresight of approximately 5–10 degrees. This translates to a cross range of 50–60–km at a distance of 400–500 km. As the beam is steered away from boresight, the beam width increases (azimuth beam width is usually less than a degree in MW radars).
3. The *velocity resolution* is directly proportional to the coherent integration time. HFSW radar employs three simultaneous integration intervals corresponding to 20 s for air targets, 164 s for ship targets and 1200 s for near stationary targets. These correspond to velocity resolutions of 4, 0.5, and 0.06 kn, respectively (CIT of miliseconds are used in MW radars).

Even though the resolution capabilities of HFSW radars seem at first glance to be rather moderate, in practice it is not a serious issue since two targets can be resolved provided that they are separable in one of the three dimensions. The probability of having two targets in close proximity in all three dimensions is not high. In the event that two targets are not resolvable, the radar will track either the larger or the composite return of the two until such time that they can be resolved.

Another performance parameter is the accuracy of the estimate of target position. Although the resolution capabilities of HFSW radars are moderate, an accuracy of better than one-tenth of the basic resolution can be achieved with even moderate signal-to-noise ratios.

A typical HFSW radar receiver is pictured in Fig. 11.20. The receiver obtains echoes from the operational area and two-step gating is applied: range gating followed by digital beam forming. As shown in the figure time histories of data of N × M resolution cells are accumulated (coherently integrated) and Fourier transformed (via FFT), and detection is achieved by Doppler processing. Target detection using a CFAR algorithm, follows beamforming. Different CFAR variants are used for surface and air targets as well as constant-velocity and manoeuvring targets. Because of the complexity of the real signal environment, it is desirable to employ adaptive parameters in the detection process to accommodate clutter, noise, and interference levels that vary from a CIT to CIT as well as from cell to cell.

HFSW radars are coherent radars and detection is based on the SNR at a given Doppler where the noise bandwidth is determined by the inverse of the coherent integration time. A consequence of the much reduced noise bandwidth is that the PD, even at a relatively low SNR, is extremely high and for the noise-limited case, the PFA at a given Doppler bin is very low. The number of false alarms for a given CIT is equal to the probability of false alarm multiplied by the number of independent range/azimuth Doppler cells (approximately a few million). Consequently even a moderate probability of false alarm translates into a high number of false detections. These false detections are dominated by ionospheric clutter and are characterized as rings of detections corresponding to the vertical range of the various ionospheric layers. The high number of high false alarm associated with HFSW radars is typical as illustrated in Fig. 11.21. Here, detections associated with 30 consecutive CITs (approximately 75 min) have been plotted on a range-azimuth scale. The figure is derived from real data collected in February 1999 [7]. The 10 to 20 surface target tracks among the large number of false detections (as many as thousands) may be observed in the figure. The spoking of the

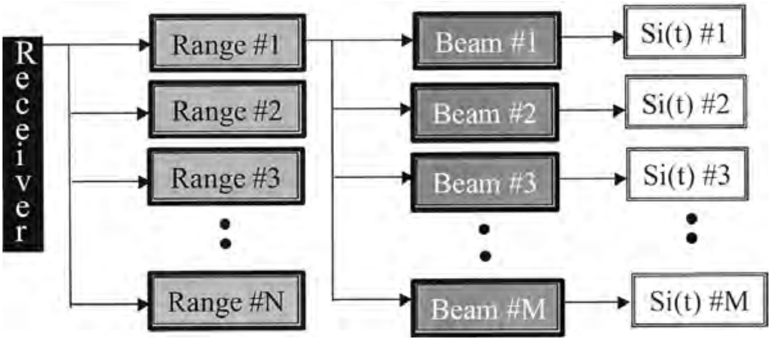


Figure 11.20 SWHF Receiver stages.

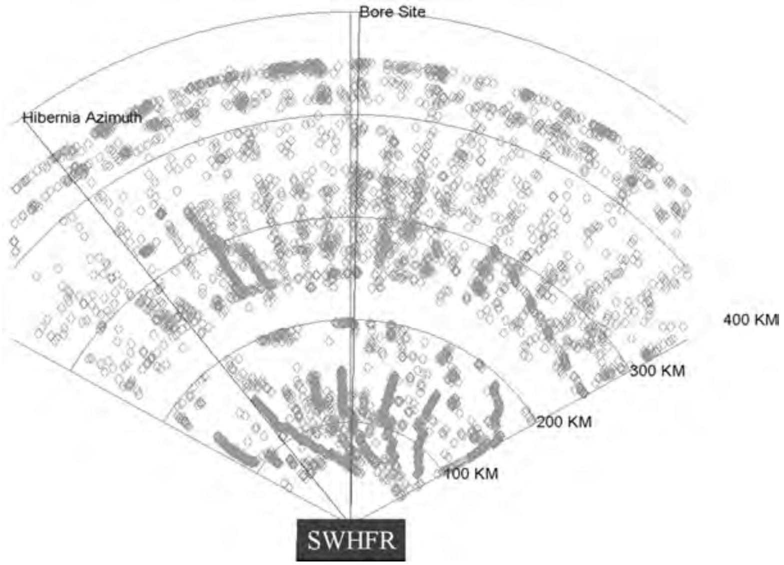


Figure 11.21 Detection of a SWHFR in 30 CIT.

data can be attributed to detections that occurred in a single beam. This high false alarm rate is almost unavoidable if the number of missed detections is to be minimized. Therefore, the Tracker must be designed to accommodate these false alarms but ensure that they do not propagate through the system to generate false tracks.

For a target that travels at a constant velocity within a CIT interval, the echo is characterized by an impulse in the Doppler spectrum. Surface targets appear in the vicinity of the Bragg peaks, while air targets are generally far removed from this sea clutter region and are typically detected against a noise background. Separate, optimized, detection processes are used to accommodate both the clutter and noise limited detection scenarios.

A typical HFSW radar picture is given in Fig. 11.22, which was obtained in Cape Race on June 9, 1999 (around noon). The data set are for days when a ground truthing aircraft was used to verify both the targets under track and to search for any

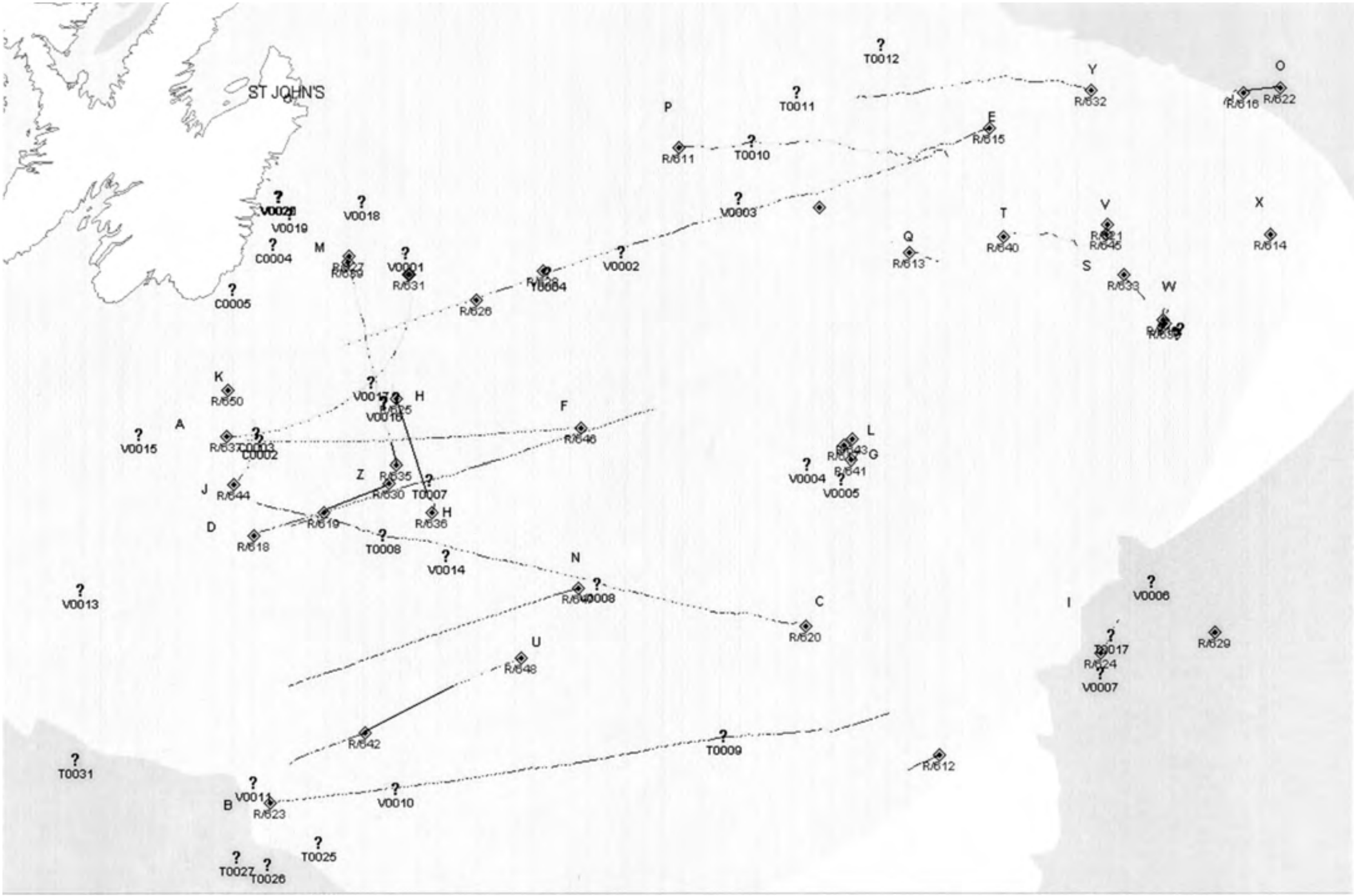


Figure 11.22 Typical (real) HFSWR system picture (Cape Race, June 2000).

targets not seen by the radar. In the figures a "?" corresponds to an independently observed target. The suffix V indicates a visual observation and the suffix T represents an airborne radar observation. It can be observed that the HFSWR successfully tracked all targets observed by the aircraft. Those targets that were tracked by the radar but were not observed by the aircraft entered the coverage area after the aircraft had finished searching that zone.

11.8. CONCLUSION

Radars are electronic sensors that extract information from EM wave–object interaction. From simple detections back in 1940s to today's complex multisensor integrated systems, radars have had a great impact on modern life. They are used as monitoring, guiding, controlling, etc., instruments in a variety of applications. Since electronic sensing information will always be required, radar will continue to play essential roles in modern societies in the future.

REFERENCES

1. Skolnik, M.I. *Introduction to Radar Systems*; McGraw Hill: New York, 1985.
2. Skolnik, M.I. *Radar Handbook*; McGraw Hill: New York, 1990.
3. Barton, D.K. *Modern Radar System Analysis*; Artech House: Norwood, MA, 1988.
4. Skolnik, M.I. Fifty years of radar. Proc. IEEE **1985**, *73*.
5. IEEE Standard Radar Definitions, std-686-1990.
6. Sevgi, L.; Ponsford, A.M.; Chan, H.C. An integrated maritime surveillance system based on surface wave HF radars, Part I—Theoretical background and numerical simulations. IEEE Antennas Propagation Mag. **2001**, *43*(4), 28–43.
7. Ponsford, A.M.; Sevgi, L.; Chan H.C. An integrated maritime surveillance system based on surface wave HF radars, Part II—Operational Status and System Performance. IEEE Antennas Propagation Mag. **2001**, *43*(5), 52–63.
8. Popoulis, A. *Probability, Random Variables and Stochastic Processes*; McGraw Hill: New York, 1985.
9. Farina, A. (Ed.). *Optimized Radar Processors*; IEE Publication: London, U.K., 1987.
10. Sevgi, L.; Akleman, F.; Felsen, L.B. Ground wave propagation modeling: problem-matched analytical formulations and direct numerical techniques. IEEE Antennas Propagation Mag. **Feb. 2002**, *44*(1), 55–75.
11. Sevgi, L.; Felsen, L.B. A new algorithm for ground wave propagation problems based on a hybrid ray-mode approach. lnt. J. Numerical Modeling **1998**, *11*(2), 8–103.
12. Hall, M.P.M.; Barclay, L.W.; Hewitt, M.T. *Propagation of Radiowaves*; IEEE Publication: London, U.K., 1996.
13. Fock, V.A. *Electromagnetic Diffraction and Propagation Problems*; Pergamon: Oxford, 1965.
14. Kerr, D.E. (Ed.). *The Propagation of Short Radio Waves, Radiation Lab. Series*; McGraw-Hill: New York, 1951.
15. Shaeffer, J.F.; Tuley, M.T.; Knot E.F. *Radar Cross Section*; Artech House: Norwood, MA, 1985.
16. Sevgi, L. Target reflectivity and RCS interaction in integrated maritime surveillance systems based on surface wave HF radar radars. IEEE Antennas Propagation Mag. **2001**, *43*(1), 36–51.
17. Balanis, C.A. *Advanced Engineering Electromagnetics*; Wiley: New York, 1989.

Other Fundamental Radar Books

18. Meikle, H.D. *Modern Radar Systems*; Artech House: Norwood, MA, 2001.
19. Sullivan, R.J. *Microwave Radar: Imaging and Advanced Concepts*; Artech House: Norwood, MA, 2001.
20. Ince, N.; Topuz, E.; Panayirci, E.; Isik C. *Principles of Integrated Maritime Surveillance Systems*; Kluwer Academic: Boston, 2000.
21. Kingsley, S.; Quegan, S. *Understanding Radar Systems*; McGraw Hill Co.: London, U.K., 1992.
22. Nathanson, F.E. *Radar Design Principles*; McGraw Hill: New York, 1991.
23. Eaves, J.L.; Reedy, E.K. (Eds.). *Principles of Modern Radar*; Van Nostrand Reinhold: New York, 1987.
24. Eaves, L.; Reedy, E.K. *Principles of Modern Radars*, Van Nostrand Reinhold: New York, 1987.
25. Meeks, M.L. *Radar Propagation at Low Altitudes*; Artech House: Norwood, MA, 1982.

12
Wireless Communication Systems

Nathan Blaunstein
Ben-Gurion University of the Negev
Beer Sheva, Israel

12.1. INTRODUCTION

This chapter provides a basic tutorial on key aspects of wireless communication systems. Because the optical communication systems are also "wireless," we must declare at the outset that such systems are not a subject of this chapter despite the fact that radio waves and optical waves are both independent parts of the electromagnetic spectrum. Therefore, this chapter will focus on antennas as transducers of radio waves, radio propagation in the wireless communication channels with emphasis on land communication channels. Specific propagation models for various land environments (rural, forested, hilly, built-up) will be presented with a view to understanding the main propagation characteristics in such environments, such as path loss, and slow and fast fading effects. The cellular concept for wireless systems and a strategy for cell design will also be discussed briefly.

12.1.1. Definition of the Wireless Communication System

Although wire-based communication systems, such as telephony which connects each telephone with a central operator station through a pair (or more) of copper wires or cables (usually called the *local* or *subscriber loops*), have been successfully employed for more than a century, during recent decades wireless communication systems have been developed to satisfy continually increasing demands for personal, local, mobile and satellite communications, by enhancing and replacing wire-media loops with wireless communication media.

Depending on the specific application of wireless communications, these media include the subsoil layers, water and ground surface, atmosphere, ionosphere, and cosmic space. We will treat each medium as a radio propagation channel across which radio signals are sent. These signals are created by transmitting antennas whose dimensions, as we will show later, are approximately the same as the wavelength of the radio signal generated. According to the Huygens principle, these waves are similar to the light rays in optics, which can be generated by a light bulb or a spot light. These radio waves are captured by the receiving antenna connected to a receiver.

Thus, a basic wireless communication system consists of a transmitter (T), a receiver (R), and the radio propagation channel (Fig. 21.1 according to Ref. 1). As follows from

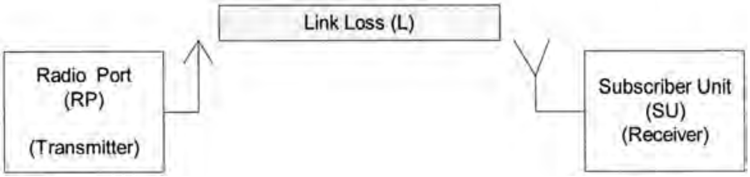


Figure 12.1 The simple scheme of three main independent electronic and electromagnetic design tasks related to the wireless communication channels.

the simple scheme depicted in Fig. 12.1, there are *three* main independent electronic and electromagnetic design tasks related to this communication system. The *first* task is the specification of the electronic equipment that controls all operations within the transmitter, including the transmitting antenna operation. The radio propagation channel, denoted as a *second* element in the scheme presented in Fig. 12.1, plays a separate independent role. Its main output characteristics depend on the conditions of radio wave propagation in the various operational environments. The *third* task concerns the same operations and signals, but for the receiver, with its own peculiarities. For both kinds of antennas, the transmitting and the receiving, an important issue is the influence of different kinds of obstacles located around the antennas and the environmental conditions.

12.1.2. Frequency Spectrum for Wireless Communications

The optimal *frequency band* for each propagation channel is determined and limited by the technical requirements of each communication system and by the conditions of radio propagation through each channel.

Extremely low and *very low frequencies* (ELF and VLF) are frequencies below 3 kHz and from 3 kHz up to 30 kHz, respectively. The VLF band corresponds to waves, which propagate through the waveguide formed by the earth's surface and the ionosphere at long distances with low attenuation [0.1–0.5 decibel (dB) per 1000 km]. Frequencies lower than 3 kHz (ELF band) are effective for underwater communication channels and for mines and subterranean communication.

Low frequencies (LF) and *medium frequencies* (MF) are frequencies from 30 kHz up to 300 kHz and 300 kHz to 3 MHz, respectively. They are useful for radio navigation of ships and aircrafts, and for broadcasting. Such radio waves propagate along the earth's surface by following the curvature of the earth, as shown in Fig. 12.2, and in the literature are called *surface* waves. Because of the long wavelengths of surface waves (for example, a 100-kHz signal has a wavelength of 3000 m), ground features, such as buildings, hills, trees, and built-up topography, do not affect the radiosignal propagation significantly.

High frequencies (HF) are those which are located in the band from 3 MHz up to 30 MHz. Signals in this spectrum propagate by means of reflections caused by the ionosphere and, therefore, are called the *sky* waves. This type of radio signals is used for long-distance land communications by use of broadcasting stations ("short-wave radio").

Very high frequencies (VHF) are located in the band from 30 MHz up to 300 MHz. They are used in a line-of sight (LOS) mode for TV communications, in long-range radar systems and in radio-navigation systems.

Ultra-high frequencies (UHF) are those that are located in the band from 300 MHz up to 3 GHz (in some literature its upper part from 0.5 GHz up to 3 GHz is also divided into P, L, S bands). This frequency band is very effective for wireless microwave links

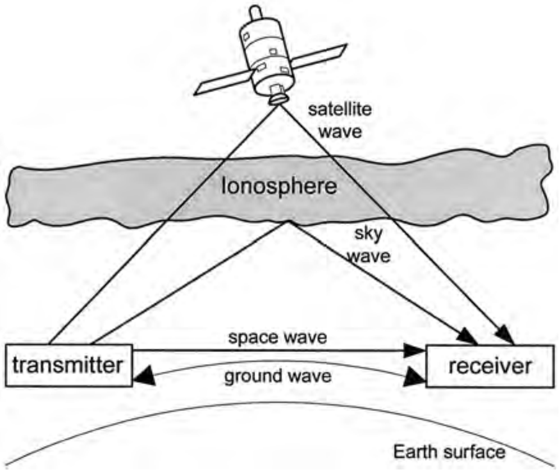


Figure 12.2 Different kinds of radio waves.

for cellular systems (fixed and mobile) and for satellite communication channels (since these frequencies penetrate the ionosphere). In the literature, these waves are sometimes called the *satellite* waves.

In recent decades radio waves with frequencies higher than 3 GHz (C, X, K bands, up to several hundred GHz, which are also loosely described as *microwaves*) have begun to be used for constructing new kinds of wireless communication channels.

12.1.3. Noise

The effectiveness of each wireless communication system depends on noise inside it, which in the literature is separated into the *additive* (or *white*) and the *multiplicative* noise [2–9]. Let us consider briefly the sources of such kinds of noise.

The *additive noise* arises from [1]

Noise in the receiver antenna
Noise within the electronic equipment that communicates with antenna
Background and ambient noise (galactic, atmospheric, man-made, etc.)

Now let us consider each type of noise, which exists in a complete communication system. Noise is generated within each element of electronic communication channel because of the random motion of electrons within the various components of the equipment. The noise power inside the transmitter–receiver electronic channel at a given system bandwidth B_w is given by [10,11]

$$N_F = k_B T_0 B_w \tag{12.1}$$

where $k_B = 1.38 \times 10^{-23}\,\mathrm{W\,s\,K^{-1}}$ is Boltzmann's constant, $T_0 = 290\,\mathrm{K}$ (17°C). Taking also into account the *noise figure* F of the receiver [2,4]

$$F = 1 + \frac{T_e}{T_0} \tag{12.2}$$

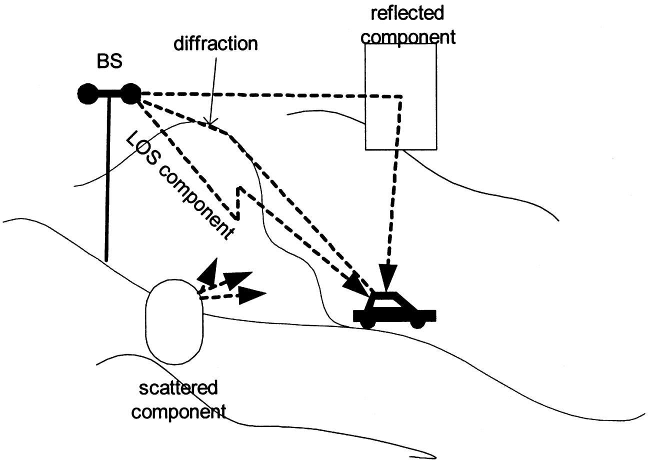


Figure 12.3 Presentation of the triple nature of fading phenomena.

where T_e is the effective noise temperature at the receiver, we can express the total effective noise power at the receiver input Eq. (12.1) as

$$N_F = k_B T_0 B_w F \tag{12.3}$$

The *multiplicative noise* arises from the processes encountered by transmitted radio waves during their travel from the transmitter to the receiver, such as (Fig. 12.3):

Multireflections from ground surface, walls, and hills
Multiscattering from rough surfaces such as the sea, rough terrain, buildings, and trees
Multidiffraction from the edges of walls, building rooftops, and hilltops

To gain a better understanding of the multiplicative noise, it is very important to define the *propagation characteristics* of the radio communication channel.

12.1.4. Main Propagation Characteristics

In real communication channels, the radio waves reach the receiver in a multipath situation in which the various waves arrive with different radiopaths and time delays. At the receiver, such waves are combined to give an oscillating resultant signal, the variations of which depend on the distribution of phases amongst the incoming component waves. The signal amplitude variations are known as *fading* [1–9]. Fading is basically a spatial phenomenon, but spatial signal variations are experienced as temporal variations by a receiver and/or transmitter moving through the multipath field or due to moving scatters, such as a truck passing the area between two terminal antennas. Thus we can talk here about space-domain and time-domain variations of EM field in land environments. Moreover, if one deals with mobile communication systems, one observes the effects of random fading in the frequency domain, i.e., the complicated interference picture of the received signal caused by receiver/transmitter movements, which is known in literature as the Doppler effect [2–9].

Numerous theoretical and experimental investigations of spatial and temporal variations of radio waves in conditions of built-up areas have shown that the urban propagation channel is approximately stationary in time, but the spatial variations of signal level have a *triple nature* (see Fig. 12.4 according to Ref. 5).

The *first* one is the *path loss*, which can be defined as an overall decrease in the signal strength with distance between two terminals, the transmitter and the receiver, when the signal is expressed in decibels. The physical processes, which cause this phenomenon are the spreading of electromagnetic wave radiated outward in space by the transmitter antenna and the obstructing effects of any natural and man-made object surrounding this antenna. The spatial and temporal variations of the signal path loss are large and slow.

Large-scale (in the space domain) and *long-term* (in the time domain) *fading* is the *second* one, which is usually called in the literature a *shadow* or *slow fading* [5,8], because it is caused by diffraction from the buildings' corners and their rooftops, or from the hills' tops located along the radio link surrounding the terminal antennas. The spatial scale of large-scale variations is of the order of the obstructions' dimensions, that is, from several to several tens of meters.

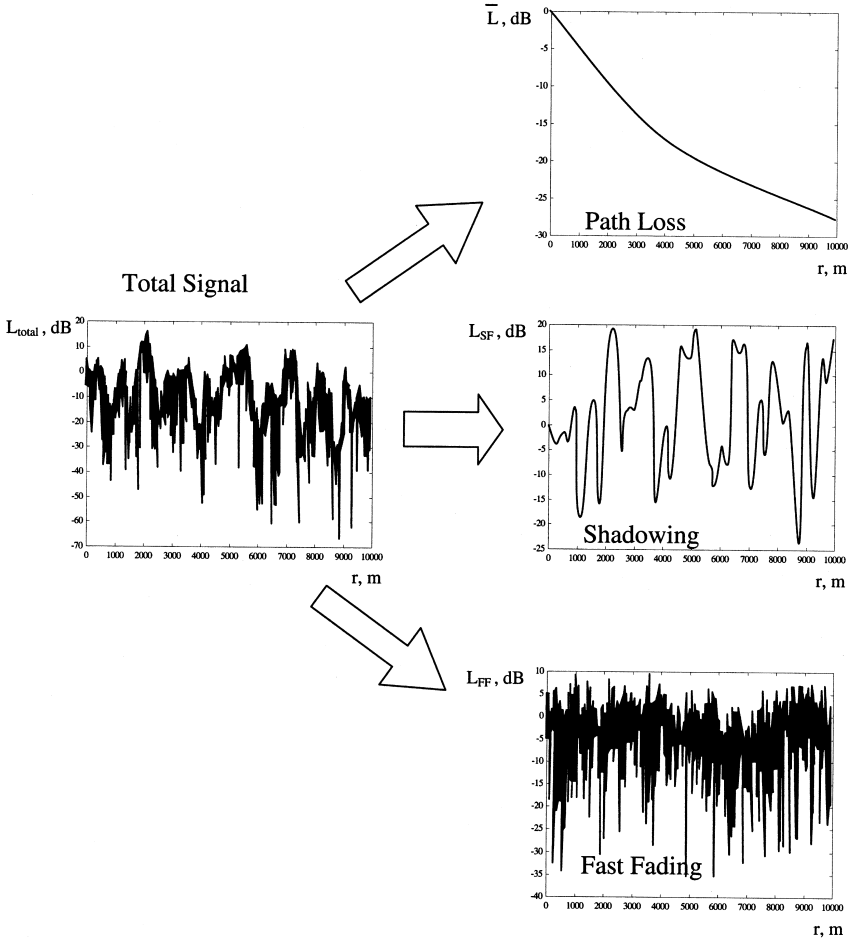


Figure 12.4 Illustration of Doppler effect.

The *third* one is the *small-scale fading* in the space domain and *short-term* or *fast* signal variations in the time domain, which are caused by the mutual interference of the wave components of the multiray field. The characteristic scale of such waves in the space domain is changed from half wavelength to three wavelength [3,8–12]. Therefore they are usually called *fast-fading* signals in the literature (see also bibliography in Refs. 5–9).

Path Loss

This is a principal characteristic that determines the effectiveness of the propagation channel in various kinds of environment. It defines variations of the signal amplitude or field intensity along the propagation trajectory (*path*) from point to point within the communication channel. For its quantitative evaluation we will assume that the signal–wave amplitude at the point $\mathbf{r}_1$ along the propagation path is $A_1(\mathbf{r}_1)$ or the signal–wave intensity is $J(\mathbf{r}_1) = A_1^2(\mathbf{r}_1)$. In the process of propagation along the path, at any next point $\mathbf{r}_2$ the signal–wave amplitude is $A_2(\mathbf{r}_2)$ or intensity $J(\mathbf{r}_2) = A_2^2(\mathbf{r}_2)$. In the literature the *path loss* is defined as a logarithmic difference between the amplitude or the intensity (sometime it is called *power*) at the points $\mathbf{r}_1$ and $\mathbf{r}_2$ along the propagation path in the medium.

In other words, *path loss*, which is denoted by L and measured in decibels (dB), can be evaluated as

For signal amplitude $A(\mathbf{r}_j)$ at two points $\mathbf{r}_1$ and $\mathbf{r}_2$ along the propagation path [1]

$$\begin{aligned} L &= 10\log\frac{A^2(r_2)}{A^2(r_1)} = 10\log A^2(r_2) - 10\log A^2(r_2) \\ &= 20\log A(\mathbf{r}_2) - 20\log A(\mathbf{r}_1)\,\mathrm{dB} \end{aligned} \tag{12.4}$$

For signal intensity $J(\mathbf{r}_j)$ at two points $\mathbf{r}_1$ and $\mathbf{r}_2$ along the propagation path

$$L = 10\log\frac{J(\mathbf{r}_2)}{J(\mathbf{r}_1)} = 10\log J(\mathbf{r}_2) - 10\log J(\mathbf{r}_1)\,\mathrm{dB} \tag{12.5}$$

If we take point $\mathbf{r}_1$ as the origin of the radiopath (the transmitter location) and assume $A(\mathbf{r}_1) = 1$, then the loss L at any arbitrary point $\mathbf{r}$ along the path is

$$L = 20\log A(\mathbf{r})\,\mathrm{dB} \tag{12.6a}$$

and

$$L = 10\log J(\mathbf{r})\,\mathrm{dB} \tag{12.6b}$$

The next question is: What are the units in which the received power is measured at the receiver? According to Refs. 1–8, the resulting output value is denoted in dB/(V/m), dB/(mV/m), and dB/(μV/m), if the reference signal–wave amplitude is specified as 1 V/m, mV/m, and μV/m, respectively. In the same way, the resulting output value is denoted in dB, dBm, and dBμ, if the reference signal/wave power is 1 W (watt), mW, and μW, respectively.

Taking into account relations between measured power units, that is, $1\,\mu\mathrm{W} = 10^{-3}\,\mathrm{mW} = 10^{-6}\,\mathrm{W}$, one can easily obtain

$$0\,\mathrm{dB}\mu = -30\,\mathrm{dBm} = -60\,\mathrm{dBW} \tag{12.7}$$

For example, if the received power level is $-15\,\mathrm{dBm}$, we have $10\log P_{\mathrm{mW}} = -15\,\mathrm{dBm}$, from which it immediately follows that $P_{\mathrm{mW}} = 10^{-1.5} = 0.0316\,\mathrm{mW}$.

The second main characteristic of communication channels is the *signal-to-noise ratio* (SNR or S/N). In decibels this characteristic can be presented as follows: for the receiver (output) channel where noise (artificial and natural) is significant

$$\mathrm{SNR} = P_R - N_F\,\mathrm{dB} \tag{12.8}$$

where P_R is the power at the receiver and N_F is described by Eq. (12.3). Both P_R and N_F are assumed to be in the same dB units, e.g., dBW.

Example. A receiver of the wireless communication system has a bandwidth of 250 kHz and requires that its input SNR should be not more than 10 dB when the input signal is $-105\,\mathrm{dBm}$. The background temperature at the input of the system is $T_0 = 300\,\mathrm{K}$. Find the maximum value of noise figure and the corresponding effective noise temperature at the input of such a receiver.

Solution. Using expression (12.3) in dB, that is,

$$N_F = 10\log(k_B T_0 B_w F)$$

we finally get

$$\mathrm{SNR} = P_R - N_F = P_R - F_{\mathrm{dB}} - 10\log(k_B T_0 B_w)$$

So, rewriting P_R in dBW according to Eq. (12.7), the unknown noise figure can be obtained as follows

$$\begin{aligned} F_{\mathrm{dB}} &= P_R - 10\log(k_B T_0 B_w) = (-105 - 30)\,\mathrm{dBW} - 10\,\mathrm{dB} \\ &\quad - 10\log\left(1.38 \times 10^{-23}\, W\,Hz^{-1}\,K^{-1} \times 300\mathrm{K} \times 250 \times 10^{3}\,\mathrm{Hz}\right) \\ &= -135 - 10 + 170 - 10\log 103.5 = 4.85\mathrm{dB} \end{aligned}$$

From Eq. (12.2) we finally get

$$T_e = T_0(F - 1) = 300(10^{4.85/10} - 1) = 616\mathrm{K}$$

12.1.5. Multipath Characteristics of the Multiplicative Noise

We will start, first of all, with a qualitative description of *slow* and *fast* fading following Refs. 2–9.

Long-Term or Slow Fading

As was shown [2–9], because the *slow* spatial signal variations (expressed in decibels, dB) tend to normal or gaussian distributions, the average signal power variations, as a result of their averaging within some individual small area, tend to the log-normal distribution (expressed in dB) with the standard deviation that depends on the relief of the terrain and on the type of built-up area [3–8,12].

Short-Term or Fast Fading

As follows from Fig. 12.4, the *fast* fading (expressed in dB) is observed over distances of about half or one wavelength. When talking about this phenomenon, we must contrast two main situations in the cellular propagation channel: the first one is when the subscribers' antennas are stationary with respect to the base station, which can be formally termed a *static multipath* situation [2–6]; the second one is when subscribers' antennas are in motion relative to the base station, which can be formally termed a *dynamic multipath* situation [2–6]. For the case of stationary receiver and transmitter (*static multipath channel*), due to multiple reflections and scattering from various obstructions around the arbitrary transmitter and receiver, the narrowband radio signals travel along different paths of varying lengths. In the case of a *dynamic multipath* situation, either the subscribers' antennas are in movement or the objects surrounding the stationary antennas move, the spatial variations of resultant signal at the receiver can be seen as temporal variations at the receiver as it moves through the multipath field. Moreover, in such a dynamic multipath situation a signal fading at the mobile receiver occurs in the time domain. This temporal fading relates to a shift of frequency radiated by the stationary transmitter. In fact, the time variations, or dynamic changes of the propagation path lengths are related to the Doppler effect, which is due to relative movements between a stationary transmitter and a moving receiver.

To illustrate the effects of phase change in the time domain due to Doppler frequency shift (called the *Doppler effect*), let us consider a mobile receiver moving at a constant velocity v, along the path X_1X_2, as it is shown in Fig. 12.5. As follows from the geometry presented in Fig. 12.5, the difference in path lengths traveled by a signal from source S to the mobile at points X_1 and X_2 is $\Delta\ell = \ell\cos\theta = v\Delta t\cos\theta$, where Δt is the time required for the moving receiver to travel from point X_1 to X_2 along the path, and θ is the angle between the mobile direction along X_1X_2 and direction to the source at the current point X_i, that is, X_iS, $i = 1, 2$. The phase change in the resultant received signal due to the

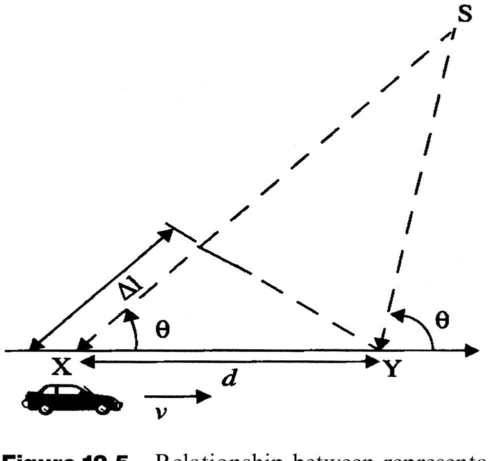


Figure 12.5 Relationship between representations of both signals' spectra according to [8].

difference in path lengths is therefore [8]

$$\Delta\Phi = k\Delta\ell = \frac{2\pi}{\lambda}\ell\cos\theta = \frac{2\pi v\Delta t}{\lambda}\cos\theta \tag{12.9}$$

Hence the apparent change in frequency radiated, or Doppler shift, is given by f_D, where

$$f_D = \frac{1}{2\pi}\frac{\Delta\Phi}{\Delta t} = \frac{v}{\lambda}\cos\theta \tag{12.10}$$

Because the Doppler shift relates, according to Eq. (12.10), to the mobile velocity and the spatial angle between the direction of mobile motion and the direction of arrival of the signal, it can be positive or negative depending on whether the mobile receiver is moving toward or away from the transmitter. In fact, as follows from Eq. (12.10), if the mobile moves *toward* the direction of arrival of the signal then $f_D > 0$, i.e., the apparent received frequency is increased, while if it moves away from the direction of arrival of the signal then $f_D < 0$, i.e., the apparent received frequency is decreased. Signals arriving from directly ahead of or directly behind the mobile correspond to the maximum rate of phase changes, giving $f_{D\max} = v/\lambda$.

12.1.6. Narrowband and Wideband Signal Representations

Now we will consider a question: what kinds of radio signals propagate in wireless communication channels. First of all we will consider a CW or *narrowband* signal representation. A voice modulated CW signal transmitted at the carrier frequency, f_c, which in the literature is called the transmitted *band-pass* or *RF* signal [2–8], can be expressed in the following form:

$$s(t) = A(t)\cos[2\pi f_c t + \varphi(t)] \tag{12.11}$$

where $A(t)$ is the signal envelope and $\varphi(t)$ is its phase.

For example, if a 3-kHz voice signal amplitude modulates a carrier at $f_c = 900$ MHz, its fractional bandwidth is very narrow, that is, $6\times10^3\,\text{Hz}/9\times10^8\,\text{Hz}\approx 7\times10^{-6}$ or $7\times10^{-4}\,\%$.

Since all information in the signal is contained within the phase and envelope time variations, usually in the literature one encounters an alternative form of band-pass signal $s(t)$

$$u(t) = A(t)\exp(j\varphi(t)) \tag{12.12}$$

which is called a *complex base-band* representation of $s(t)$. It is clear from Eqs. (12.11) and (12.12) that the relation between the *band-pass* (*RF*) and the *complex base-band* signal representations is

$$s(t) = \mathrm{Re}[u(t)\exp(j2\pi f_c t)] \tag{12.13}$$

The relationship between the representations of both signals, Eqs. (12.11) and (12.12), in the frequency domain is shown schematically in Fig. 12.6, from which follows that the complex base-band signal is a frequency-shifted version of the band-pass signal with

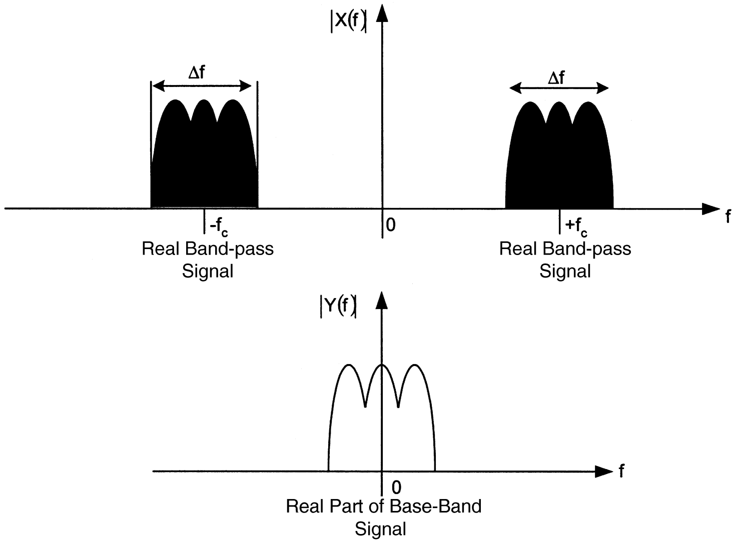


Figure 12.6 Schematic illustration of Doppler effect [according to Ref. 5].

the same spectral shape but centered in the close proximity of zero frequency despite the carrier f_c. Moreover, the mean power of the base-band signal is

$$\langle P_s(t)\rangle = \frac{\langle |u(t)|^2\rangle}{2} = \frac{\langle u(t)u^*(t)\rangle}{2} \tag{12.14}$$

which is the same result as the mean-square value of the real, band-pass signal $s(t)$.

The complex envelope of the received *CW* (*narrowband*) signal can be presented according to Eq. (12.12) within the multipath channel as the phasor sum of N baseband individual multiray components arriving at the receiver with the corresponding time delay, τ_i, $i = 0, 1, 2, \ldots$ [8],

$$r(t) = \sum_{i=0}^{N-1} u_i(t) = \sum_{i=0}^{N-1} A_i \exp\left(j\varphi_i(t,\tau_i)\right) \tag{12.15}$$

If we assume that during the vehicle movements over a local area the amplitude A_i variations are small enough, whereas phases φ_i vary greatly due to changes in propagation distance over the space, then as a result we obtain great random oscillations of the total signal $r(t)$ during the receiver movement over small distances. Finally, since $r(t)$ is the phasor sum Eq. (12.15) of the individual multipath components, the instantaneous phases of the multipath components cause the large fluctuations which typifies small-scale fast fading for CW signal $s(t)$. The average received power over a local area is then given by [8]

$$\langle P_{CW}\rangle \approx \sum_{i=0}^{N-1} \langle A_i^2\rangle + 2\sum_{i=0}^{N-1} \sum_{i,j\neq i} \langle A_i A_j\rangle \langle \cos(\varphi_i - \varphi_j)\rangle \tag{12.16}$$

Here the time averaging was done for *CW* measurements made by a mobile vehicle, as the receiver, over the local measured area [8].

For a *wideband* (or *pulse*) probing signal the total received power is simply related to a sum of the powers of the individual multipath components Eq. (12.15), where each component has a random amplitude and phase at any time t, and, then, the average small-scale received power can be presented as [8]

$$\langle P_{\text{pulse}} \rangle \approx \sum_{i=0}^{N-1} \langle A_i^2 \rangle \tag{12.17}$$

Hence, in the multipath wideband propagation channel the small-scale received power is simply the sum of the powers received in each multipath component. Because in practice, the amplitudes of individual multipath components do not fluctuate widely in a local area, then the received power of the wideband (pulse) signal does not fluctuate significantly when a vehicle moves over a local area. Comparison between the CW and pulse small-scale power presentation, Eqs. (12.16) and (12.17), shows that when $\langle A_i A_j \rangle = 0$ and/or $\langle \cos(\varphi_i - \varphi_j) \rangle = 0$, the average power for a CW signal is equivalent to the average received power for a pulse signal in a small-scale region. This can occur either when the path amplitudes are uncorrelated, that is, each multipath component is independent after reflection or scattering, or when multipath phases are independently and uniformly distributed over $[0, 2\pi]$. Thus we can conclude that in UHF–microwave bands, when the multipath components traverse differential path lengths, having hundreds of wavelengths: *The received local ensemble average powers of wideband signal and narrowband signal are equivalent.*

12.1.7. Characterization of Terrain Configurations

Now let us consider another principal question: terrain classification. The process of classification of terrain configurations is a very important stage in the construction of propagation models above the ground surface and finally, in predicting the signal attenuation (or "path loss") and fading characteristics within each concrete wireless propagation channel.

The simple classification of *terrain configuration* follows from practical research and experience of designers of such communication systems. It can be presented as [1–11]

Open area
Flat ground surface
Curved, but smooth terrain
Hilly terrain
Mountains

The *built-up areas* can also be simply classified as (1) rural areas, (2) suburban areas, and (3) urban areas. Many experiments that have been carried out in different built-up areas have shown that there are many specific factors, which must be taken into account to describe specific propagation phenomena in built-up areas, such as [1–11]

Buildings' density or terrain coverage by buildings (in percents)
Buildings' contours or their individual dimensions
Buildings' average height
Positions of buildings with respect to the base station and fixed or mobile receivers
Positions of both antennas, receiver and transmitter, with respect to the rooftops' level
Density of vegetation; presence of gardens, parks, lakes etc.
Degree of "roughness" or "hilliness" of a terrain surface

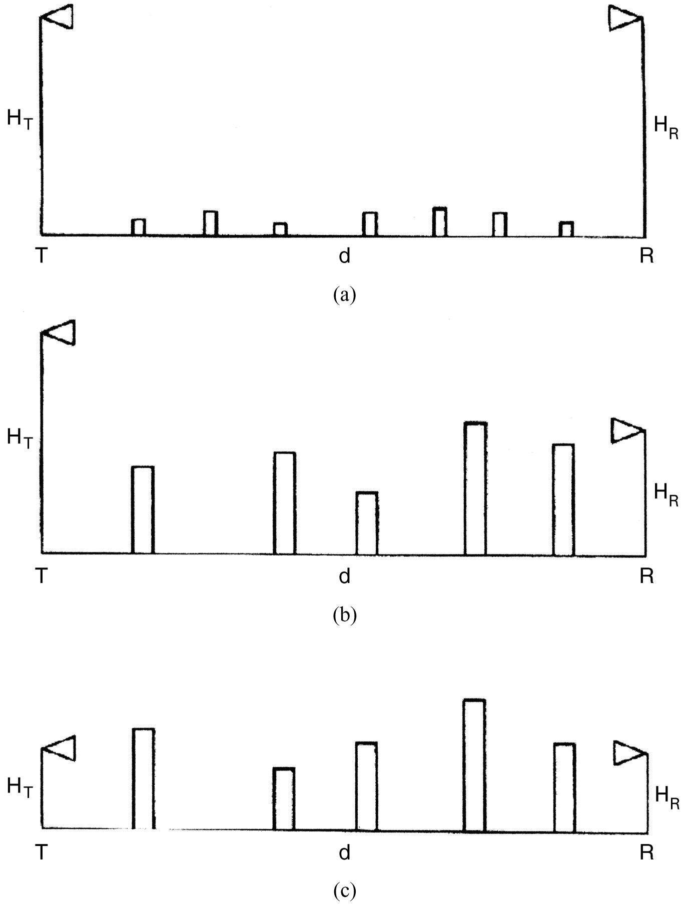
Using these specific characteristics and parameters, we can easily classify various kinds of terrain by examining topographic maps for each deployment of a wireless communication system.

12.1.8. Various Propagation Situations in Built-up Areas

As remarked earlier, a very important characteristic of the propagation channel is the location and position of both antennas with respect to the obstacles placed around them. Usually there are three possible situations shown in Fig. 12.7a–c, respectively [1]:

1. Both antennas, receiver and transmitter, are placed above the tops of obstacles (in a built-up area this means that they are above the rooftops' level).
2. One of the antennas is higher than the tops of the obstacles (namely, the roofs), but the second one is lower.
3. Both antennas are below the tops of the obstacles.

In the first situation they are in *direct visibility* or LOS conditions. In the last two situations, one or both antennas are in *clutter* or obstructive conditions, which call



(a)

(b)

(c)

Figure 12.7 Three possible situations with receiving and transmitting antennas.

non-line-of-sight (NLOS) conditions. In all these cases the profile of terrain surface is also very important and may vary from flat and smooth, with curvature, up to rough and hilly terrain.

12.2. ANTENNA BASICS

As was mentioned above and shown in Fig. 12.1 according to Ref. 13, a radio antenna, transmitting or receiving, is an independent element of the wireless communication system, which converts the current and/or voltage generated by the wire-based circuit, such as a transmission line, a waveguide or coaxial cable, into electromagnetic field energy propagating through space. In unbounded free space, the fields propagate in the form of spherical waves, whose amplitude, as will be shown below, is inversely proportional to the distance from the antenna. Each radio signal can be represented as a progressive electromagnetic wave [1], which propagates along a given direction.

12.2.1. Main Antenna Characteristics

The principal characteristics used to describe an antenna acting either as a transmitter or as a receiver are *radiation pattern*, *polarization*, *directivity*, *gain*, *efficiency*, and antenna *impedance*. We will define them briefly. More information about antennas and their characteristics can be found in Refs. 13–16.

Radiation Pattern

The radiation pattern of any antenna is defined usually as the relative distribution of electromagnetic power in space. Here we must differentiate a near-field and a far-field region of such radiation. As is shown in Fig. 12.8, the near-field region, called also the *Fresnel region*, is defined by a radius R [13]

$$R = \frac{2l^2}{\lambda} \tag{12.18}$$

beyond which lies the far field or the *Fraunhofer region*. Here l (m) is the diameter of the antenna or area of the smallest sphere where the antenna is embedded and λ(m) is the wavelength. The radiation pattern is a plot of the far-field radiation intensity from the antenna usually measured per unit solid angle (Fig. 12.8).

Example. Find the far-field distance for an antenna with dimension of 0.9 m and operating frequency of 1 GHz.

Solution. For the operating frequency of 1 GHz, the wavelength is $\lambda = c/f = (3 \times 10^8\,\text{m/s})/(10^9\,\text{Hz}) = 0.3\,\text{m}$. Then, according to Eq. (12.8), the minimum distance, from which the Fraunhofer zone has already begun, is

$$r_F \geq R = \frac{2l^2}{\lambda} = \frac{2 \times 0.81}{0.3} = 5.4\,\text{m}$$

Mathematically the radiation intensity can be presented as the product of the power density (or the time-averaged Poynting vector S in watts per square meter [1,13–16])

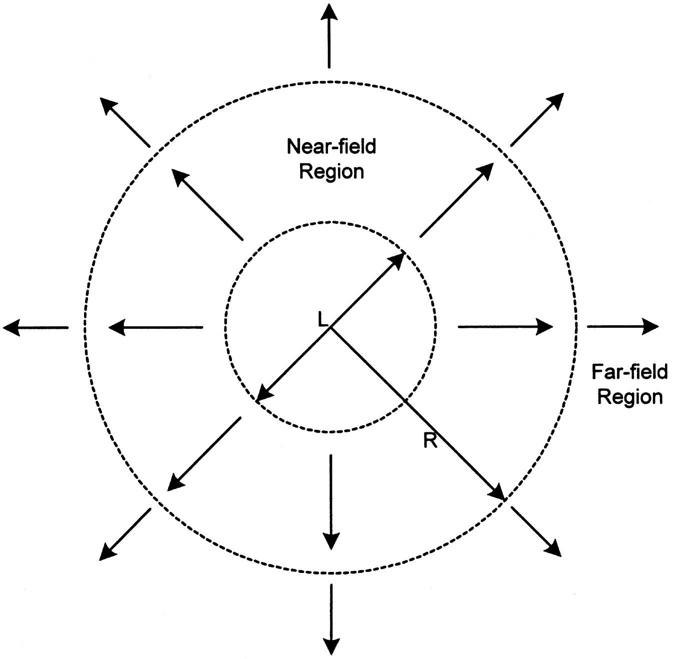


Figure 12.8 Definition of field regions surrounding the antenna [13].

and the square of the distance from the antenna r, that is, [13–16]

$$I = r^2 S = r^2 E_\theta H_\phi \tag{12.19}$$

where E_θ and H_ϕ are the components of the electrical and magnetic fields of antenna radiation in the spherical coordinate system shown in Fig. 12.9. The shape of the radiation pattern defines a type of the antenna: *isotropic*, *directional*, and *omnidirectional.*

An *isotropic antenna* refers to an antenna radiating equally in all directions, that is, its radiation power uniformly distributed in all directions. For such an antenna with a total power P, which spreads uniformly over a sphere of radius r, the power density at this distance in any direction equals [13–16]

$$S = \frac{P}{\text{area of a sphere}} = \frac{P}{4\pi r^2} \tag{12.20}$$

Then, according to Eq. (12.19), the radiation intensity of an isotropic antenna equals

$$I = \frac{P}{4\pi} \tag{12.21}$$

A *directional antenna* transmits (or receives) waves more efficiently in certain directions than in others. A radiation pattern plot for directional antenna is shown in Fig. 12.10 according to Ref. 6, illustrating the *main lobe*, which includes the direction of maximum radiation intensity, a *back lobe* with radiation in the opposite direction of the main

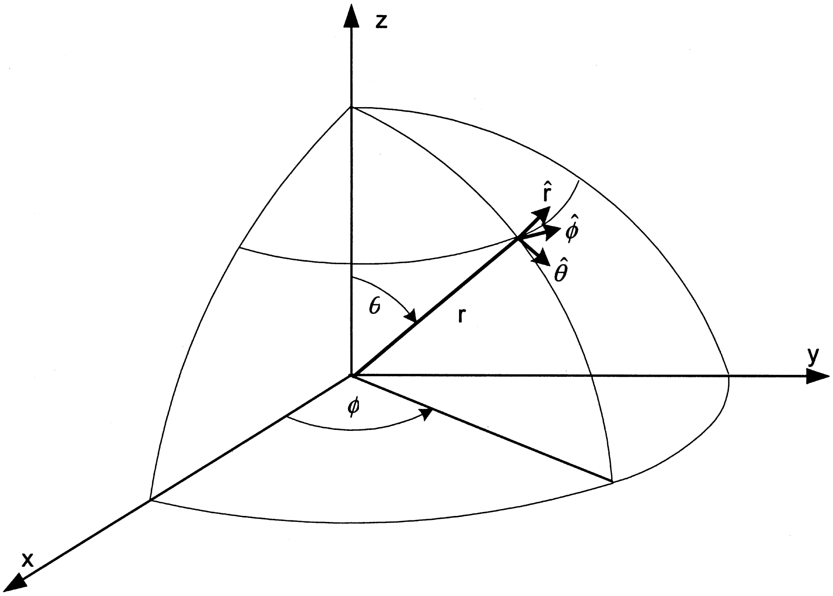


Figure 12.9 Spherical coordinate system for antenna parameters computation.

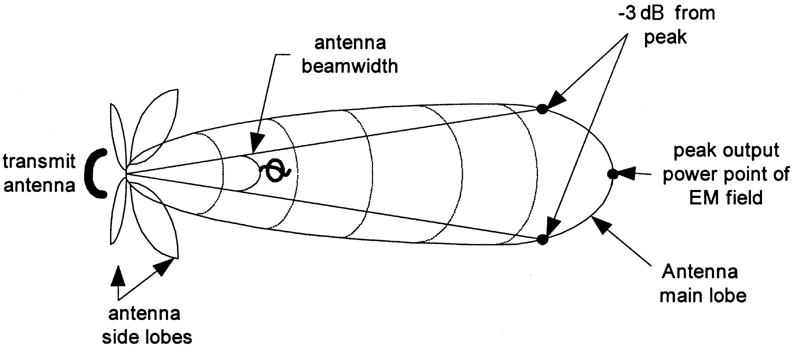


Figure 12.10 Radiation pattern of the directive antenna.

lobe, and several *side lobes* separated by nulls where no radiation occurs. For such kind of antennas, some unified parameters are usually used

The *half-power beam width* (see Fig. 12.10), or simply the beam width, is the solid angle that bounds the area of the main lobe where the half-power points are located.

The *front–back ratio* is the ratio between the peak amplitudes of the main and back lobes, usually expressed in decibels.

The *side-lobe level* is the amplitude of the biggest side lobe, usually expressed in decibels relative to the peak of the main lobe.

A more practical type of the antenna, as compared with the idealized isotropic one, is an *omnidirectional antenna*, whose radiation is constant in the plane of azimuth but may vary in the vertical plane.

The parameter "directivity" is used to describe non-isotropic antennas and the variation of its signal intensity in all directions. The concept of directivity indicates that an antenna concentrates the field energy in specific directions. If so, we can define the directivity D as the ratio between the power of the transmitting radiation in the specific direction determined by spherical coordinates (ϕ, θ), according to Fig. 12.9, to that of the equivalent isotropic antenna [13–16]:

$$D(\phi, \theta) = \frac{P(\phi, \theta)}{P_{\text{isotropy}}} \tag{12.22}$$

The use of an isotropic antenna as a reference in Eq. (12.22) allows one to measure the directivity in units dBi, i.e., relative to an isotropic antenna, $D\,\text{dBi} = 10\log D_{\text{isotropy}}$.

Because of power dissipation within the antenna circuit itself, a new parameter is introduced to describe the *radiation efficiency*, denoted by η and defined as the ratio between the power actually radiated by the antenna to the power accepted by the antenna.

The *power gain*, G, or *antenna gain*, expressed in dB, is defined as 4π times the ratio of the radiation intensity in a given direction to the total power accepted by the antenna. The gain is related to the radiation efficiency and the directivity by [13–16]

$$G(\phi, \theta) = \eta D(\phi, \theta) \tag{12.23}$$

Usually in practice the terms "directivity" and "gain" refer to the maximum value of $D(\phi, \theta) = D_{\max}$, denoted simply by D, and $G(\phi, \theta) = G_{\max}$, denoted simply by G. If so, we can rewrite Eq. (12.23) as a simple relation $G = \eta D$.

Using these notations we can rewrite now the formula Eq. (12.20) for radiated power density of the transmitting directive antenna as

$$S = \frac{P}{4\pi r^2}\, G = \frac{P}{4\pi r^2}\, \eta D \tag{12.24}$$

For a *receiving* antenna, a parameter called the *effective area* or *aperture* (also called *antenna cross-section*) is usually used, and denoted by A_e. The *effective aperture* of the receiving antenna is defined as the ratio of the power P_R, which is delivered to a matched receiver, to the power density S of electromagnetic radiation of the transmitter antenna, according to Eq. (12.20), which arrives at the receiver antenna, i.e.,

$$A_e = \frac{P_R}{S} \tag{12.25}$$

It should be noted that this value differs from the real geometrical area of the receiving antenna, which collects the arriving wave energy. The maximum antenna gain G is also related to the effective antenna aperture as follows [13–16]:

$$G = \frac{4\pi}{\lambda^2} A_e \tag{12.26}$$

As follows from Eq. (12.8), the power collected at the receiver antenna must be greater than the power of noise there. An improvement in the SNR can be obtained through use of antennas with high directivity or gain. In fact, for an isotropic antenna with $G = 4\pi\,(D = 1)$, we have $A_e = \lambda^2/4$ according to Eq. (12.26). By comparison, a short dipole antenna (about which we will talk later) has directivity $D = 1.5$, i.e., $A_e = 1.5\,(\lambda^2/4) = 3\lambda^2/8$. In this case the power collected at the receiving dipole antenna is 1.5 times greater than that for the isotropic antenna.

Polarization is also one of the important parameters of the antenna. It is defined by the orientation of the electric field component **E** of the radiated electromagnetic wave.

12.2.2. Antennas in Free Space

Formulas obtained above allow us to obtain the relation between power at the transmitter and the receiver antennas located in free space. This relation in called the *Friis transmission formula*. Let us present it below. According to formulas (12.24) and (12.25), for two antennas separated by a distance r, great enough to take into account only a far-field regions of both antennas, we get [13–16]

$$P_R = SA_{eR} = \frac{P_T}{4\pi r^2} G_T A_{eR} \tag{12.27}$$

Since the effective aperture of the receiver antenna equals [see Eq. (12.26)] $A_{eR} = \lambda^2 G_R/4\pi$, we finally get

$$P_R = SA_{eR} = P_T \left(\frac{\lambda}{4\pi r}\right)^2 G_T G_R \tag{12.28}$$

or introducing a new parameter, the path gain PG, for antennas in free space we get

$$PG = \frac{P_R}{P_T} = \left(\frac{\lambda}{4\pi r}\right)^2 G_T G_R \tag{12.29}$$

Here G_T and G_R are the maximum gains of the transmitting and the receiving antennas, respectively.

12.2.3. Types of Antennas

There is a wide variety of available antenna systems used in different areas of wireless communications. We describe briefly the dipole antenna, which is widely used in practical applications and will refer the reader to the literature [13–16] where all types of antennas are fully described.

Dipole Antennas

The basic structures and the current distributions of the hertzian and $\lambda/2$-dipole antennas are shown in Figs. 12.11a,b and 12.12a,b, respectively [6]. It is seen that for a hertzian dipole $l/\lambda \ll 1$, i.e., it can be considered as a "short" antenna with a uniform current distribution along its length and with maximum gain $G = 1.5$, which corresponds to angle $\theta = 90°$ (normal to the dipole axis). As for the $\lambda/2$-dipole antenna, presented in Fig. 12.12, the maximum of field radiation also occurs at $\theta = 90°$; the maximum gain is

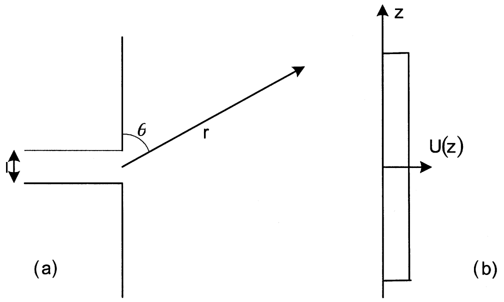


Figure 12.11 Hertzian dipole: (a) antenna and (b) current distribution according to [6].

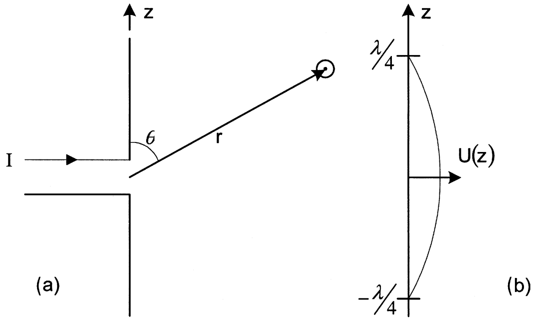


Figure 12.12 $\lambda/2$ dipole: (a) antenna and (b) current distribution according to [6].

now $G = 1.64$ [5,6]. So, the gains of the two kinds of dipole antennas are not very different. But, as was mentioned in [5,6,13–16], the radiation resistance of the $\lambda/2$-dipole antenna is much higher with respect to that of a Hertzian one (making the longer dipole easier to match to the feedline), and the loss resistance is small (higher efficiency).

Using now the Friis formula Eq. (12.29), we can compare the path gain for the isotropic and the $\lambda/2$-dipole antennas. In fact, according to Eq. (11.29), if two terminals, the transmitter and the receiver, are separated by a distance of 1 km, the ratio of their path gain in these two cases is [5,6]

$$\frac{(PG)_{\text{isotropic}}}{(PG)_{\lambda/2\text{-dipole}}} = \frac{0.57 \times 10^{-9}}{1.53 \times 10^{-9}} = 0.37 \qquad (12.30)$$

that is, the path gain of the $\lambda/2$-dipole antenna approximately in 3 times greater than that for the isotropic antenna.

12.3. PATH LOSS PREDICTION MODELS IN VARIOUS OUTDOOR COMMUNICATION LINKS

As was mentioned in Sec. 12.1, the path loss is a main characteristic of the radio propagation channel, which is investigated in more detail in the literature [1–12]. Below, we will present briefly the propagation models suitable for practical applications in wireless communications.

12.3.1. Free-Space Path Loss

Let us consider a nonisotropic antenna placed in free space as a transmitter of P_T watts and with a directivity gain G_T. At an arbitrary large distance r ($r > r_F$, where r_F is the Fraunhofer far-field range) from the source, the radiated power is uniformly distributed over the surface area of a sphere of radius r. If P_R is the power at the receiver antenna, which is located at distance r from the transmitter antenna and has a directivity gain G_R, then the *path loss* in decibels according to definition Eq. (12.5) and Friis formula Eq. (12.29) can be determined as [1–11]

$$L = 10\log\frac{P_T}{P_R} = 10\log\frac{(4\pi r/\lambda)^2}{G_T G_R} = L_0 + 10\log\frac{1}{G_T G_R} \tag{12.31}$$

Here L_0 is the path loss for an isotropic point source (with $G_R = G_T = 1$) in free space, which in decibels can be presented as

$$L_0 = 10\log\left(\frac{4\pi fr}{c}\right)^2 = 20\log\frac{4\pi fr}{c} = 32.44 + 20\log r + 20\log f \tag{12.32}$$

where the value 34.44 is obtained by the use of simple calculations, taking into account that the speed of light $c = 3 \times 10^8$ m/s:

$$32.44 = 20\log\frac{4\pi \times 10^3\text{m} \times 10^6(1/\text{s})}{3 \times 10^8\,\text{m/s}} = 20\log\frac{40\pi}{3} \tag{12.33}$$

In expression (12.32) the distance r is in kilometers (km), and frequency f is in megahertz (MHz). As the result, the path loss for both directive antennas, in free space, is

$$L_F = 32.44 + 20\,\log d_{\text{km}} + 20\,\log\,f_{\text{MHz}} - 10\,\log G_T - 10\,\log G_R \tag{12.34}$$

12.3.2. Path Loss over a Flat Terrain

The simplest case of radio wave propagation over terrain is that where the ground surface can be assumed as flat. The assumption of "flat terrain" is valid for radio links between subscribers up to 10–15 km apart [1–7]. The main process is a reflection from flat terrain, which is described by the reflection coefficients.

Reflection Coefficients

Following Refs. 1–6, we will present now the expressions for the complex coefficients of reflection (Γ) for waves with vertical (denoted by index V) and horizontal (denoted by index H) polarization, respectively.

For *horizontal* polarization,

$$\Gamma_H = |\Gamma_H|e^{-j\varphi_H} = \frac{\sin\psi - (\varepsilon_r - \cos^2\psi)^{1/2}}{\sin\psi + (\varepsilon_r - \cos^2\psi)^{1/2}} \tag{12.35a}$$

For *vertical* polarization,

$$\Gamma_V = |\Gamma_V| e^{-j\varphi_V} = \frac{\varepsilon_r \sin\psi - (\varepsilon_r - \cos^2\psi)^{1/2}}{\varepsilon_r \sin\psi + (\varepsilon_r - \cos^2\psi)^{1/2}} \tag{12.35b}$$

Here $|\Gamma_V|$, $|\Gamma_H|$ and φ_V, φ_H are the magnitude and the phase of the coefficients of reflection for vertical and horizontal polarization, respectively; $\psi = (\pi/2) - \theta_0$ is the grazing angle; θ_0 is the angle of wave incidence. The knowledge of reflection coefficient amplitude and phase variations is a very important factor in the prediction of path loss for different situations in the land propagation channels. In practice, for wave propagation over terrain, the ground properties are determined by the conductivity and the absolute dielectric permittivity (dielectric constant) of the subsoil medium, $\varepsilon = \varepsilon_0 \varepsilon_r$, where ε_0 is the permittivity of vacuum and ε_r is the complex relative permittivity of the ground surface.

Line-of-Sight (LOS) Two-Ray Model

The two-ray model was first proposed for describing the process of radio wave propagation over flat terrain [1,5–7], which is based on the superposition of a direct ray from the source and a ray reflected from the flat ground surface, as shown in Fig. 12.13. Following [1,5–7], we can present relation between the filed strength and power at the transmitter as

$$E = \sqrt{\frac{30 G_T G_R P_T}{r_1}} \tag{12.36}$$

where r_1 is the trajectory of the direct wave as presented in Fig. 12.13. Then the total field at the receiver is the sum of direct and received waves, that is,

$$E_R = E_T \left(1 + \frac{r}{r_1} |\Gamma| e^{-jk\Delta r} \right) \tag{12.37}$$

Here $\Gamma(\psi)$ is the reflection coefficient described by formulas (12.35a) and (12.35b) for horizontal and vertical polarization, respectively, $\Delta r = r_2 - r_1$ (see Fig. 12.13) is the

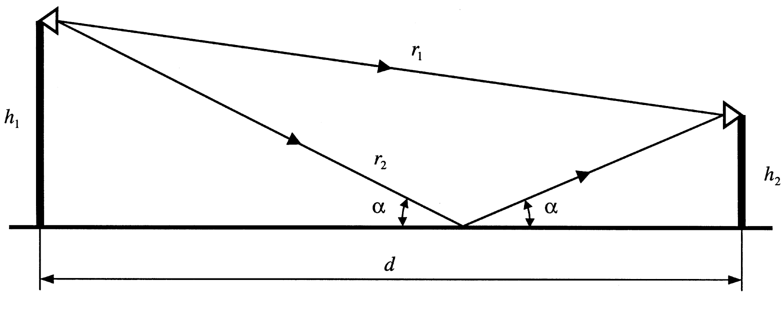


Figure 12.13 Geometry of two-ray model.

difference in the radio paths of the two waves, $\Delta\varphi = k\Delta r$ is the phase difference between the reflected and direct waves which can be presented as [1–7]

$$\Delta\varphi = k\,\Delta d = \frac{2\pi}{\lambda} r \left\{ \left[1 + \left(\frac{H_R + H_T}{r} \right)^2 \right] - \left[1 + \left(\frac{H_R - H_T}{r} \right)^2 \right] \right\} \tag{12.38}$$

where h_R and h_T are the receiver and transmitter antenna heights, respectively, and r is the distance between them. For $r_1 \gg (h_T \pm h_R)$ and $r \gg (h_T \pm h_R)$, with the assumption that $r_1 \approx r_2 \approx r$, Eq. (12.38) can be rewritten as

$$\Delta\varphi = \frac{4\pi h_R h_T}{\lambda r} \tag{12.39}$$

Furthermore, if we now assume that $G_R \approx G_T = 1$ (omnidirectional antennas) and that $\Gamma(\psi) \approx -1$ for the farthest ranges from transmitter (when the grazing angle is small), we will finally obtain the absolute value of the power at the receiver

$$\begin{aligned} |P_R| &= |P_T| \left(\frac{\lambda}{4\pi r} \right)^2 \left| 1 + \cos^2 k\,\Delta r - 2\cos k\,\Delta r + \sin^2 k\ \Delta r \right| \\ &= |P_T| \left(\frac{\lambda}{4\pi r} \right)^2 \sin^2 \frac{k\,\Delta r}{2} \end{aligned} \tag{12.40}$$

As follows from Eq. (12.40), the largest distance from transmitter, for which there is some maximum of received power, occurs when

$$\frac{k\,\Delta r}{2} \approx \frac{\pi}{2} \qquad \sin\frac{k\,\Delta r}{2} \approx 1 \tag{12.41}$$

This distance is called the *critical range*, denoted by r_b, and it is approximately determined according to Eq. (12.41) by the following formula [1,5–7,17–19]:

$$r_b \approx \frac{4 h_R h_T}{\lambda} \tag{12.42}$$

In other critical case of small incident angles, that is, when $\sin^2(k\,\Delta r/2) \approx (k\,\Delta r/2)^2$, $\Delta r = 2h_T h_R/r$, which is valid for large distances between antennas relative to antenna heights ($r \gg h_T$, h_R), from Eq. (12.40) an approximate formula of the path loss, called *path loss in the model of flat terrain*, can be obtained

$$L_{FT} = 10\log\frac{|P_T|}{|P_R|} = 10\log\frac{r^4}{h_T^2 h_R^2} = 40\log r_{\mathrm{m}} - 20\log(h_{T\mathrm{m}} h_{R\mathrm{m}}) \tag{12.43}$$

Using now the *critical range* definition (12.42), we can present the path loss over flat terrain in the following form [17]

$$L = \begin{cases} L_B + 20\log\dfrac{r}{r_B} & r \le r_B \\ L_B + 40\log\dfrac{r}{r_B} & r > r_B \end{cases} \tag{12.44}$$

where L_B is the path loss in free space at the distance that equals the critical range, i.e., $r = r_B$, which can be calculated from the following expression [17]:

$$L_B = 32.44 + 20\log r_{B\text{km}} + 20\log f_{\text{MHz}}$$

As follows from Eq. (12.44), there are two modes of field intensity decay at distances r less than the *break point* $r = r_B$, and beyond this point, that is, $\sim r^{-q}$, $q = 2$ for $r \leq r_B$, and $\sim r^{-q}$, $q = 4$ for $r > r_B$.

12.3.3. Path Loss in Clutter (NLOS) Conditions

We investigated above situations in communication links where the LOS conditions occur between two terminal antennas, the transmitter and receiver. Now we consider radio propagation above the terrain in the situation where both antennas are placed above the ground surface in non-line-of-sight (NLOS) conditions. Here a new effect of diffraction phenomena arises from various kinds of obstacles, such as trees or hills, placed on the terrain. The diffraction phenomenon is based on the Huygens' principle [1–12]. Let us briefly describe the diffraction from obstructions, using the Huygens' principle and replacing each obstruction by a *knife edge* [1–3,5–8].

Propagation over a single knife edge. If there is some obstacle that we may model as a simple knife edge (denoted as OO′, see Fig. 12.14) which lies between the receiver and the transmitter, the phase difference $\Delta\Phi$ between the direct ray from the source (at point O), denoted TOR, and that diffracted from the point O′, denoted TO′R, can be obtained in the standard manner by use of a simple presentation of the path difference, Δd, and the phase difference, $\Delta\Phi$, between these rays assuming that the height of the obstacle is much smaller than the characteristic ranges between the antennas and the obstacle ($h \ll d_1, d_2$) [1–3,8]:

$$\begin{aligned} \Delta d &\approx \frac{h^2}{2}\frac{d_1 + d_2}{d_1 d_2} \\ \Delta\Phi &= \frac{2\pi}{\lambda}\Delta d = \frac{2\pi}{\lambda}\frac{h^2}{2}\frac{d_1 + d_2}{d_1 d_2} \end{aligned} \tag{12.45}$$

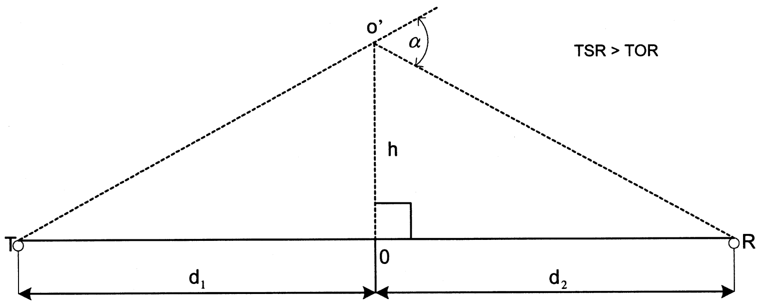


Figure 12.14 Schematical presentation of simple knife-edge model.

Fresnel–Kirchhoff diffraction parameter. If we now introduce the Fresnel–Kirchhoff diffraction parameter v according to [1–3,8]

$$v = h\sqrt{\frac{2(d_1 + d_2)}{\lambda d_1 d_2}} \tag{12.46}$$

the phase difference may be rewritten in terms of this parameter, i.e.,

$$\Delta\Phi = \frac{\pi}{2}v^2 \tag{12.47}$$

To estimate the effect of diffraction around obstructions we need a quantitative measure of the required clearance over any terrain obstruction, and as was shown in [1,6,8], this may be obtained analytically in terms of Fresnel-zone ellipsoids drawn around both ends of the radio link, receiver and transmitter (Fig. 12.15). The reader can find a full discussion of Fresnel ellipsoids in [1–3,6,8]. Here we will only repeat that the cross-sectional radius of any ellipsoid with number n from the family at a distance d_1 and $d_2 = d - d_1$ can be presented as a function of the parameters n, d_1, and d_2 as

$$r_n \equiv h_n = \left(\frac{n\lambda d_1 d_2}{d_1 + d_2}\right)^{1/2} \tag{12.48}$$

From Eq. (12.46) one can obtain the physical meaning of the Fresnel–Kirchhoff diffraction parameter:

$$v_n = h_n\left[\frac{2(d_1 + d_2)}{\lambda d_1 d_2}\right]^{1/2} = \left[\frac{2(d_1 + d_2)}{\lambda d_1 d_2}\frac{n\lambda d_1 d_2}{d_1 + d_2}\right]^{1/2} = (2n)^{1/2} \tag{12.49}$$

Thus the diffraction parameter v increases with the number n of ellipsoids. All the above formulas are correct for $h_n \ll d_1$, d_2, i.e., far from both antennas. The volume enclosed by the ellipsoid defined by $n = 1$ is known as a *first Fresnel zone*. The volume between this ellipsoid and that defined by $n = 2$ is the *second Fresnel zone*. The contributions to the total field at the receiving point from successive Fresnel zones tend to be in phase opposition and therefore interfere destructively rather than constructively. If an obstruction OO′ is placed at the middle of radio path TO′R (i.e., TO′ = O′R, see Fig. 12.14), then if the height of obstruction h increases from $h = r_1$ (corresponding to the *first* Fresnel zone) to $h = r_2$ (defining the limit of the *second* Fresnel zone), then to $h = r_3$ (i.e., to the *third*

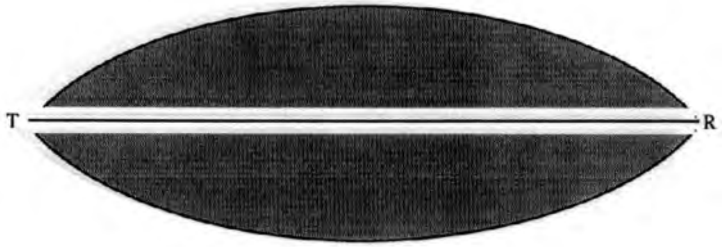


Figure 12.15 Geometrical presentation of the Fresnel zones.

Fresnel zone), etc., then the field at the receiver R would oscillate. The amplitude of oscillations would essentially decrease since a smaller amount of wave energy penetrates into the outer zone.

Diffraction losses. When there is a single obstacle between the transmitter and receiver, which can be modeled by a single "knife edge", losses of the wave energy take place. Such losses in the literature are called *diffraction losses.* They can be obtained analytically by use of so-called Fresnel complex integrals by the use of Huygens' principle [1–8]:

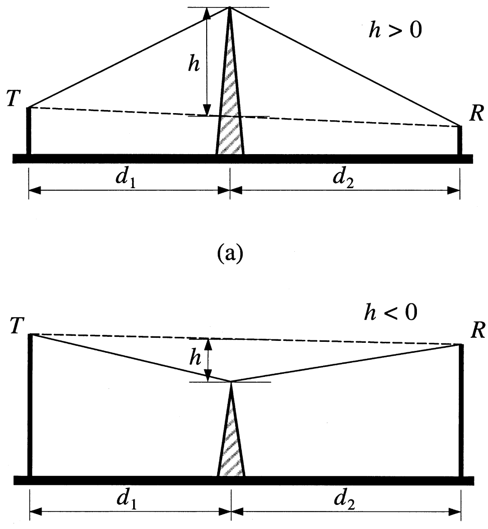
$$E = E_0 \frac{1+j}{2} \int_{\nu}^{\infty} \exp\left(-j\frac{\pi}{2}t^2\right) dt \tag{12.50}$$

The integral in the right side of Eq. (12.50) is the complex integral with parameter of integration ν defined by Eq. (12.46) for the height of the obstruction under consideration. We note that if the path TR between the transmitter and receiver (line-of-sight path) is actually obstructed by some obstacle modeled by a knife edge, as is shown in Fig. 12.16a, then the height h and the diffraction parameter ν are positive [it follows from Eq. (12.46)]. If the knife edge lies below the line-of-sight path (line TR in Fig. 12.16b), so that there is no interruption between T and R, then h and, hence, ν are negative [see again, Eq. (12.46)]. As is known, the Fresnel integral in Eq. (12.50) can be presented in the standard manner using the following integral presentations

$$\int_{\nu}^{\infty} \cos\left(-\frac{\pi}{2}t^2\right) dt = \frac{1}{2} - \int_{0}^{\nu} \cos\left(-\frac{\pi}{2}t^2\right) dt = \frac{1}{2} - C(\nu) \tag{12.51a}$$

and

$$\int_{\nu}^{\infty} \sin\left(-\frac{\pi}{2}t^2\right) dt = \frac{1}{2} - \int_{0}^{\nu} \sin\left(-\frac{\pi}{2}t^2\right) dt = \frac{1}{2} - S(\nu) \tag{12.51b}$$



(a)

Figure 12.16 Two variants of antenna locations relative to knife edge: (a) above and (b) below the direct visibility line TR [according to Refs. 1–3, 8].

At the same time, as follows from the classical theory of plane wave propagation [1,2], the total wave field E_{total} after diffraction at the edge of some arbitrary obstruction can be presented as

$$E_{\text{total}} = E_i \cdot \hat{D} \cdot \exp(j\Delta\Phi) \tag{12.52}$$

where E_i is the incident wave from the transmitter located in free space, $\hat{D}$ is the diffraction coefficient or matrix [1,2,5,6], and $\Delta\Phi$ is the phase difference between the diffracted and direct waves mentioned above. Taking into account Eqs. (12.50) and (12.51), the total field according to Eq. (12.52) can be rewritten as [1–3,8]

$$E = E_0 \frac{1+j}{2}\left[\left(\frac{1}{2} \pm C(v)\right) - j\left(\frac{1}{2} \pm S(v)\right)\right] \tag{12.53}$$

The main goal of strict diffraction theory is to obtain parameters D and $\Delta\Phi$ by use of Fresnel integrals. Comparing now Eqs. (12.52) and (12.53), it is easy to obtain the diffraction coefficient and the phase difference $\Delta\Phi$ through the Fresnel integrals [1–3,8]:

$$\widehat{D} = \frac{S + (1/2)}{\sqrt{2}\sin(\Delta\Phi + (\pi/2))} \tag{12.54a}$$

$$\Delta\Phi = \tan^{-1}\left[\frac{(S + (1/2))}{(C + (1/2))}\right] - \frac{\pi}{4} \tag{12.54b}$$

However, to obtain an exact solution by use of an integral equation such as Eq. (12.50), which is connected with the complex Fresnel integral, is a very complicated problem. Therefore, *empirical* and *semiempirical* models, which are based on numerous experimental data, are usually used to obtain the diffraction losses in NLOS communication links. A more effective empirical model to obtain the knife-edge diffraction losses is the Lee's approximate model [2,3] given by the following system of empirical equations:

$$L(v) = L_{\Gamma}^{(0)} = 0\,\text{dB} \qquad v \leq -1 \tag{12.55a}$$

$$L(v) = L_{\Gamma}^{(1)} = 20\log(0.5 - 0.62v)\,\text{dB} \qquad -0.8 < v < 0 \tag{12.55b}$$

$$L(v) = L_{\Gamma}^{(2)} = 20\log[0.5\exp(-0.95v)]\,\text{dB} \qquad 0 < v < 1 \tag{12.55c}$$

$$L(v) = L_{\Gamma}^{(3)} = 20\log\left\{0.4 - [0.1184 - (0.38 - 0.1v)^2]^{1/2}\right\}\text{dB} \qquad 1 < v < 2.4 \tag{12.55d}$$

$$L(v) = L_{\Gamma}^{(4)} = 20\log\frac{0.225}{v}\,\text{dB} \qquad v > 2.4 \tag{12.55e}$$

Results of calculations of such a model are shown in Fig. 12.17 according to Refs. 1–3, 8 with the corresponding notations according to Eqs. (12.55a)–(12.55e). Other empirical models, which give solution for diffraction losses after two, three and more obstructions are based on Lee's model (see detailed discussions in Refs. 1 and 2). All formulas above can be used mostly for open and rural communication links with LOS and NLOS conditions. More complicated situations with communication between terminal antennas are observed in mixed built-up environments with or without vegetation.

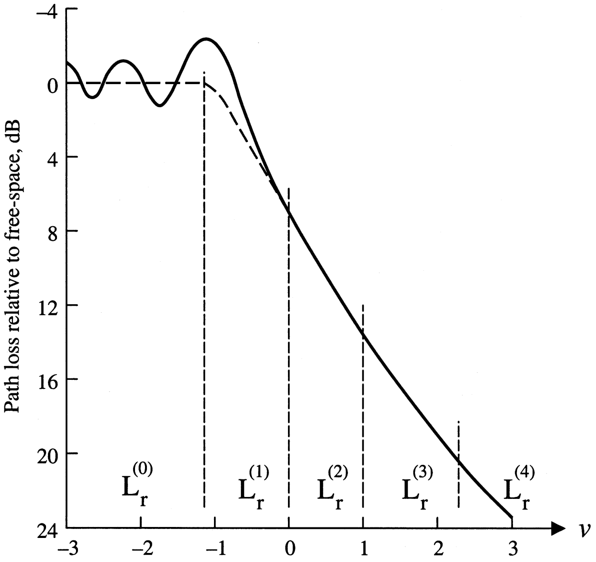


Figure 12.17 Computations of path loss according to Lee's knife-edge empirical model.

12.3.4. Path Loss Models in Land Communication Links with Regular Built-up Terrain

Here we consider several urban propagation environments and we start with the simplest case of wireless communication in the urban areas, when both antennas are placed above the flat ground surface in conditions of direct visibility (LOS conditions), but below the rooftops' level. Here we refer to the multislit street waveguide model [19–23], which was found to be in good agreement with experimental data of wave propagation in urban areas with a regular straight crossing streets. Then we will present the 2D diffraction model of propagation along the rows of buildings with different location of both terminal antennas relative to building rooftops.

Street-Multislit-Waveguide Model

The street is seen as a planar multislit waveguide. One waveguide plane is placed at the street side $z=0$, and the second one at $z=a$ (see Fig. 12.18), so a denotes the street width. The screen (building) L_n and slit (gap) l_n lengths are distributed according to the Poisson law with the average values of $\langle L\rangle = L$ and $\langle l\rangle = l$, respectively [20–23]:

$$f(L_n) = L^{-1}\exp\left(-\frac{L_n}{L}\right) \qquad f(l_n) = l^{-1}\exp\left(-\frac{l_n}{l}\right) \tag{12.56}$$

Following this model [1], we consider the resulting reflected and diffracted fields as a sum of the fields reaching the observer from the virtual image sources $\prod_n^+$ (for the reflections from plate $z=a$) and $\prod_n^-$ (for the reflections from plate $z=0$), (see Fig. 12.18), which finally gives us the approximate expression for the path loss at a large range

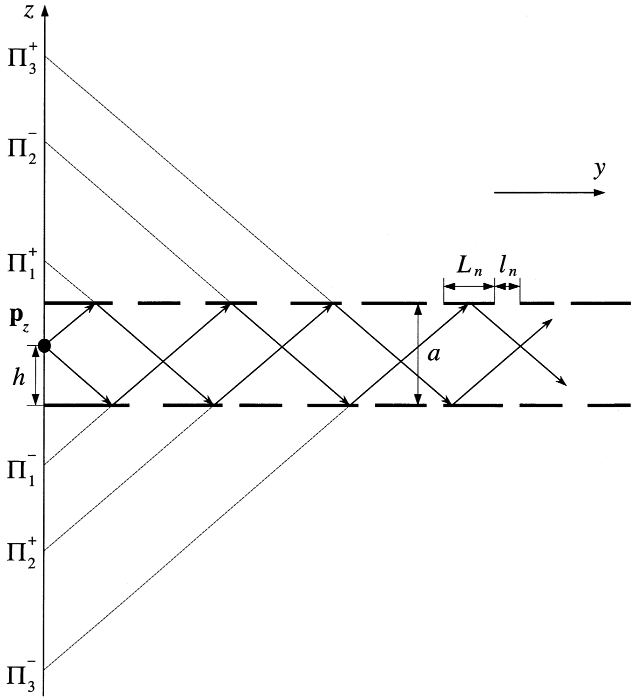


Figure 12.18 Two-dimensional geometry of street waveguide model according to [20–23].

from the source ($r \gg a$) [20–23]

$$L \approx 32.1 + 20 \log f_0 - 20 \log_{10} \frac{(1-\chi)^2}{(1+\chi)^2} + 17.8 \log r + 8.6 \left(|\ln \chi| \frac{\pi n}{a} \frac{r}{\rho_n^{(0)} a} \right) \qquad (12.57)$$

Here $\chi = L/(L + l)$ is the parameter of breakness, $\rho_n^{(0)} = \sqrt{k^2 - (n\pi/a)^2}$, n is a number of waveguide modes (number of reflections), which, as was shown in Refs. 1 and 20–23, must be less than $n = 2$ at the distances more than 50 m along the street (this range can be changed with changes of the street width a). Usually, a main mode with $n = 1$ propagates along the street waveguide.

Two-Dimensional Model of Straight Rows of Buildings

In obstructive ("clutter") conditions the receiver or transmitter antennas (or both) are placed in the shadow zones, when there are many nontransparent buildings surrounding them. In this case the diffraction from the roofs and corners of buildings plays a significant role and the total field depends not only on the reflected, but mostly on the diffracted waves [6,24,25]. Let us consider, according to [6,24,25], that an elevated antenna (base station) radiates a field that propagates in an environment with regularly distributed nontransparent buildings with various heights h_i and different separation distances d_i ($i = 1, 2, 3, \ldots$) between them. The height of the base station antenna, H, can be greater or smaller than the height of the first (near the antenna) building, h_1 (see Fig. 12.19a and b, respectively).

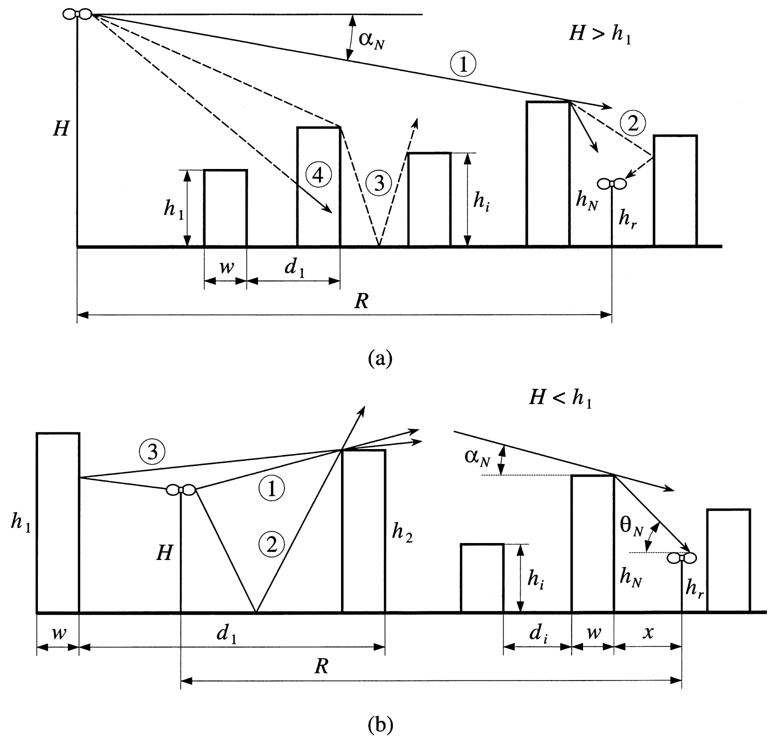


Figure 12.19 Two variants of the base station antenna locations: (a) above and (b) below the building rooftops in 2D-multiple-diffraction model according to [6].

The propagation over the rooftops involves diffraction past a series of buildings with dimensions larger than wavelength λ, i.e., h_i, $d_i \gg \lambda$. At each building a portion of the field will be diffracted toward the ground. The field reaching street level results from diffraction of the waves incident on the rooftops in the vicinity of the receiving antenna [6,24,25].

Treating the base station as a transmitter and assuming that the receiver is at street level, we can obtain the path loss in dB as the sum of the free space path loss

$$L_0 = -10 \log \frac{1}{(4\pi R)^2} \tag{12.58}$$

and excess loss L_{ex}. The last can be presented as the sum of two parts [6,24,25]:

1. The diffraction of the fields at the rooftops before the receiver down to the street level is

$$L_{e1} = -10 \log \left\{ \frac{G_1(\theta_N)}{\pi k r} \left[\frac{1}{\theta_N} - \frac{1}{(2\pi + \theta_N)} \right]^2 \right\} \tag{12.59}$$

   where $G_1(\theta_N)$ is the gain of the receiving antenna pattern in the direction θ_N as shown in Fig. 12.19a; simple geometrical constructions give

$\theta_N = \tan^{-1}\ [(h_N - h_r)/x]$ and $r = [(h_N - h_r)^2 + x^2]^{1/2}$, where x is the distance between the receiver and the building closest to the receiver; h_r is the receiver antenna height.

2. The reduction of the field at the rooftop before the receiver as a result of propagation past the previous rows of buildings

$$L_{e2} = -10\log\left(G_2 W^2\right) \tag{12.60}$$

where G_2 is the gain in the direction of the highest building edge visible from the base station antenna. To determine parameter W, let us consider two typical cases occur in the urban scene.

In the case when the base antenna is *higher* than the first building ($H > h_1$, see Fig. 12.19a) parameter W can be presented for small angle $\alpha_N = \tan^{-1}\{[H - h_N]/R\}$ and for $x \ll R$ [6,24,25] as

$$W = \frac{(d_N - w)/[R - (d_N - w)]}{\sqrt{2\pi k\left[(h_N - H)^2 + (d_N - w)^2\right]^{1/2}}} \times \left\{\frac{1}{\tan^{-1}[(h_N - H)/(d_N - w)]} + \frac{1}{2\pi + \tan^{-1}[(h_N - H)/(d_N - w)]}\right\} \tag{12.61}$$

In the case when the base antenna is *lower* than the first building ($H < h_1$, see Fig. 12.19b), according to [6,24,25] we have

$$W = \left(2.35\sqrt{\frac{d_N - w}{\lambda}}\ \tan^{-1}\left(\frac{H}{d_N - w}\right)\right)^{0.9} \tag{12.62}$$

Using above formulas, one can easily predict path loss effects in built-up areas with regularly distributed rows of straight crossing streets.

12.3.5. Path Loss Models in Land Communication Links with Irregular Built-up Terrain

Below we will describe propagation in built-up areas, when both terminals, the transmitter and receiver, are located in LOS and/or NLOS conditions at the street level, but with the assumption that buildings are randomly distributed over irregular terrain, as a main case of city topography, and will present some more realistic and more specific models that describe the propagation phenomena within the urban communication channel and predict the loss characteristics within it. We will start with empirical and semiempirical models, which are mostly used in urban communication link design, then we will present stochastic model on how to obtain path loss in built-up areas with array of buildings randomly distributed on the rough ground surface.

Okumura's Empirical Model

Based on numerous measurements carried out in and around Tokyo, Okumura proposed an empirical method of predicting the average power within the communication channel "mobile-base station" [26]. The method is based on a series of curves describing the average attenuation $A_{Ru}(f,d)$ relative to free space for quasismooth terrain in an urban environment. We present the average path loss, L_{50}, according to [26], as

$$L_{50} = L_{FS} + A_{Ru}(f,d) + H_{Tu}(h_T,d) + H_{Ru}(h_R,d) \tag{12.63}$$

Here as above, L_{FS} is the path loss in free space. The first correction factor in Eq. (12.63), $A_{Ru}(f,d)$, is expressed in Fig. 12.20 versus frequencies from 100 MHz to 1 GHz and distance from the transmitter (denoted by T) in the range 1–100 km. The reference transmitter antenna height is $h_T = 200$ m, and the reference moving vehicle antenna (denoted by R) height is $h_R = 3$ m. The second correction factor in (12.63), $H_{Tu}(h_T,d)$, is the base station antenna gain factor presented in Fig. 12.21 for the same reference heights of both antennas, $h_T = 200$ m and $h_R = 3$ m. The third correction factor in Eq. (12.63), $H_{Ru}(h_R,d)$, is the moving vehicle antenna height gain that is shown in Fig. 12.22. Here once more, the reference antenna heights are $h_T = 200$ m and $h_R = 3$ m. All corrections in Figs. 12.21 and 12.22 are changed in the positive or negative directions as the antenna height differ becoming greater or smaller than $h_T = 200$ m and $h_R = 3$ m.

As was mentioned in [1,2], the Okumura approach is probably the most widely quoted of the available models. It takes into account not only urban, suburban and rural environments, but also describes the effects of different kind of terrain. All phenomena and effects can be computed well in practice. However it is rather cumbersome to implement this model with all correction factors in a computer, because the data is available in graphical form. Thus, for computer implementation data has to be entered

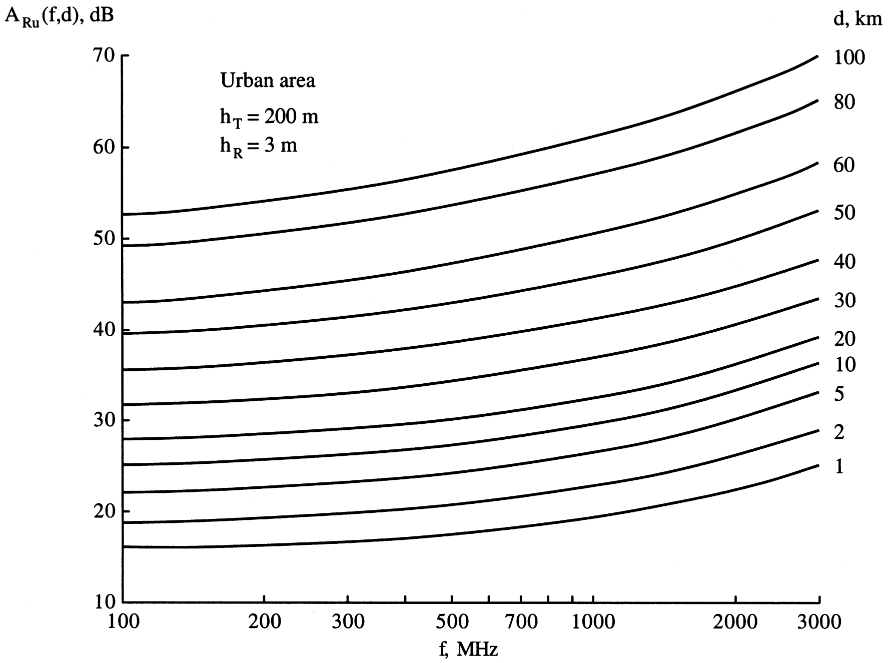


Figure 12.20 The correction factor $A_{Ru}(f,d)$ versus frequencies from 100 MHz to 1 GHz and ranges from 1 km to 100 km.

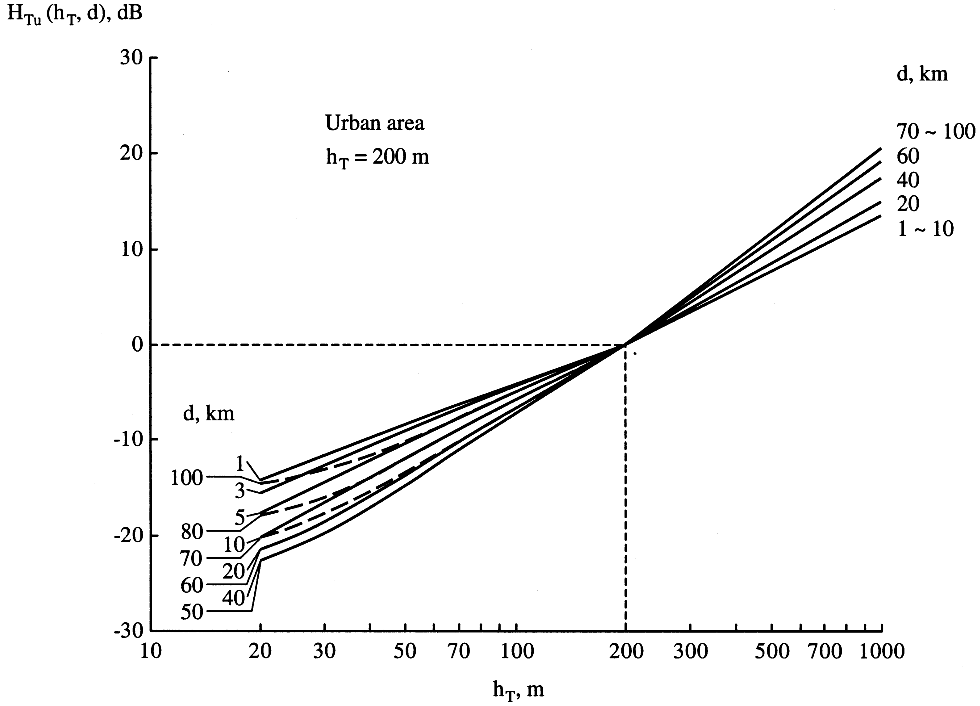


Figure 12.21 The base station antenna gain factor $H_{Tu}\ (h_T, d)$.

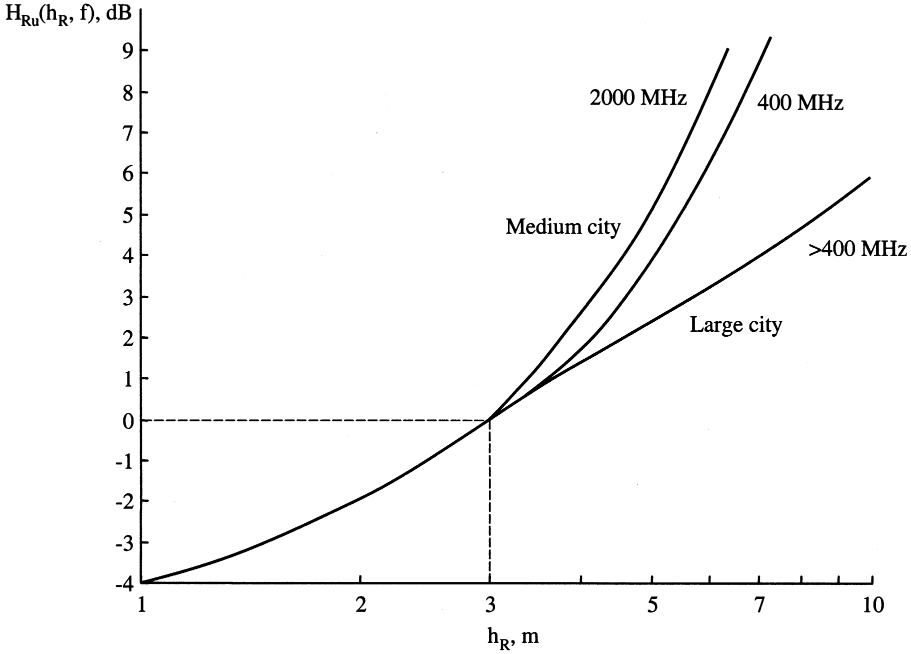


Figure 12.22 The vehicle antenna gain factor $H_{Ru}(h_R, d)$.

in the computer memory in *point-to-point* form and interpolation routines have to be written for intermediate computations.

Hata Model

In an attempt to make the Okumura technique suitable for computer implementation and easy to apply, Hata [27] developed an empirical model to describe the graphical information given by Okumura and presented in Figs. 12.20–12.22. His analytical expressions for average path loss, L_{50}, for urban, suburban and rural areas are applicable only over quasismooth terrain. The average path loss is given in dB as

$$L_{50} = 69.55 + 26.16 \log f_0 - 13.82 \log h_T - a(h_R) + (44.9 - 6.55 \log h_T) \log d \tag{12.64}$$

where $150 \leq f_0 \leq 1500\,\text{MHz}$, $30 \leq h_T \leq 200\,\text{m}$, $1 \leq h_R \leq 10\,\text{m}$, and $1 \leq d \leq 20\,\text{km}$. The function $a(h_R)$ is the correlation factor for mobile antenna height that is computed as follows [27]:

For medium-size cities,

$$a(h_R) = (1.1 \log f_0 - 0.7)h_R - (1.56f_0 - 0.8) \tag{12.65a}$$

For a large city,

$$a(h_R) = \begin{cases} 8.29(\log 1.54\, h_R)^2 - 1.1 & f_0 \leq 200\,\text{MHz} \\ 3.2(\log 11.75\, h_R)^2 - 4.97 & f_0 \geq 400\,\text{MHz} \end{cases} \tag{12.65b}$$

For suburban areas,

$$L_{50} = L_{50}(\text{urban}) - 2\left(\log \frac{f_0}{28}\right)^2 - 5.4\,\text{dB} \tag{12.66}$$

For open and rural areas,

$$L_{50} = L_{50}(\text{urban}) - 4.78(\log f_0)^2 + 18.33 \log f_0 - 40.94\,\text{dB} \tag{12.67}$$

The last formula also account for the difference in correction function for small, medium, and large cities. A comparison between results given by Hata's formulations and data obtained from Okumura's original curves for urban areas and for reference antenna heights $h_T = 200\,\text{m}$ and $h_R = 3\,\text{m}$ reveals negligible differences that, rarely exceed 1–2 dB [1,11].

Walfisch–Ikegami Model

This model gives a good path loss prediction for dense built-up areas such as medium and large cities [1,28]. It is based on important urban parameters such as building density, average building height, and street width. In this model antenna height is generally lower than the average buildings' height, so that the waves are guided along the street.

For *LOS conditions*, the path loss formula has the same form as the free-space formula changing only constants before log *d*, the distance between terminals *d*:

$$L_{50}(\mathrm{LOS}) = 42.6 + 20\log f_0 + 26\log d \tag{12.68}$$

As for *NLOS conditions*, the semiempirical path loss formula is [1,28]:

$$L_{50}(\mathrm{NLOS}) = 32.4 + 20\log f_0 + 20\log d + L_{\mathrm{RD}} + L_{\mathrm{MD}} \tag{12.69}$$

where L_{RD} represents rooftop diffraction loss, and L_{MD} represents multiple diffraction loss due to surrounding buildings. The rooftop diffraction loss is characterized as

$$L_{\mathrm{RD}} = -16.9 - 10\log\Delta a + 10\log f_0 + 20\log\Delta h_R + L(0) \tag{12.70}$$

where Δa is the distance between the vehicle and the building, h_R is the mobile vehicle antenna height, $L(0)$ is the loss due to elevation angle, and $\Delta h_R = h_{\mathrm{roof}} - h_R$.

The multiple-diffraction component is characterized by following equation [1,28]:

$$L_{\mathrm{MD}} = K_0 + K_a + K_d\log d + K_f\log f_0 - 9\log a \tag{12.71}$$

where

$$K_0 = -18\log(1 + \Delta h_T)$$

$$K_a = \begin{cases} 54 - 0.8\Delta h_T & d \geq 0.5\,\mathrm{km} \\ 54 - 1.3\Delta h_T & d < 0.5\,\mathrm{km} \end{cases}$$

$$K_d = 18 - 15\left(\frac{\Delta h_T}{h_{\mathrm{roof}}}\right)$$

$$K_f = \begin{cases} -4 + 0.7\left(\dfrac{f_0}{925} - 1\right) & \text{for suburban} \\ -4 + 0.7\left(\dfrac{f_0}{925} - 1\right) & \text{for urban} \end{cases}$$

a is the street width, h_{T} is the base station antenna height, h_{roof} is the average height of small buildings ($h_{\mathrm{roof}} < h_T$), $\Delta h_T = h_T - h_{\mathrm{roof}}$. In Walfisch–Ikegama model it was initially assumed that the base station antenna height is lower than a tall building but higher than the small buildings surrounding it.

Comparison between both empirical models, the Hata model and the Walfisch–Ikegama model for a dense urban area, shows that both of them have approximately the same polynomial signal power decay versus distance from both terminals with the parameter of attenuation $2.5 \leq \gamma \leq 4$.

Statistical Model of Path Loss in Outdoor Communication Links

We will present now a statistical approach, which is based on the stochastic models described in [1,29–31] for different kinds of the terrain, that is, on the knowledge of the terrain parameters and features.

To estimate the path loss, we, first of all, need information about the terrain features introduced and defined in Refs. 1 and 29–31:

Terrain elevation data, i.e., digital terrain map, consisting of ground heights as grid points $h_q(x, y)$.

A clutter map, that is, the ground cover of artificial and natural obstructions as a distribution of grid points, $h_0(x, y)$, for built-up areas this is the buildings' overlay profile; the average length or width of obstructions, $\langle L \rangle$ or $\langle d \rangle$; the average height of obstructions in the test area, $\bar{h}$; the obstructions density per km^2, ν.

The effective antenna height, that is, the antenna height plus a ground or obstruction height, if the antenna is assembled on a concrete obstruction: z_1 and z_2 for the transmitter and receiver, respectively.

As the result, there is a digital map (cover) with actual heights of obstructions can be performed according to buildings' overlay profile and topographic map of the built-up terrain. Using now all parameters of built-up terrain and both antennas, transmitter and receiver, the three-dimensional digital map can be analyzed. In the general case of rough terrain with randomly distributed obstacles (see Fig. 12.23), in obstructive conditions

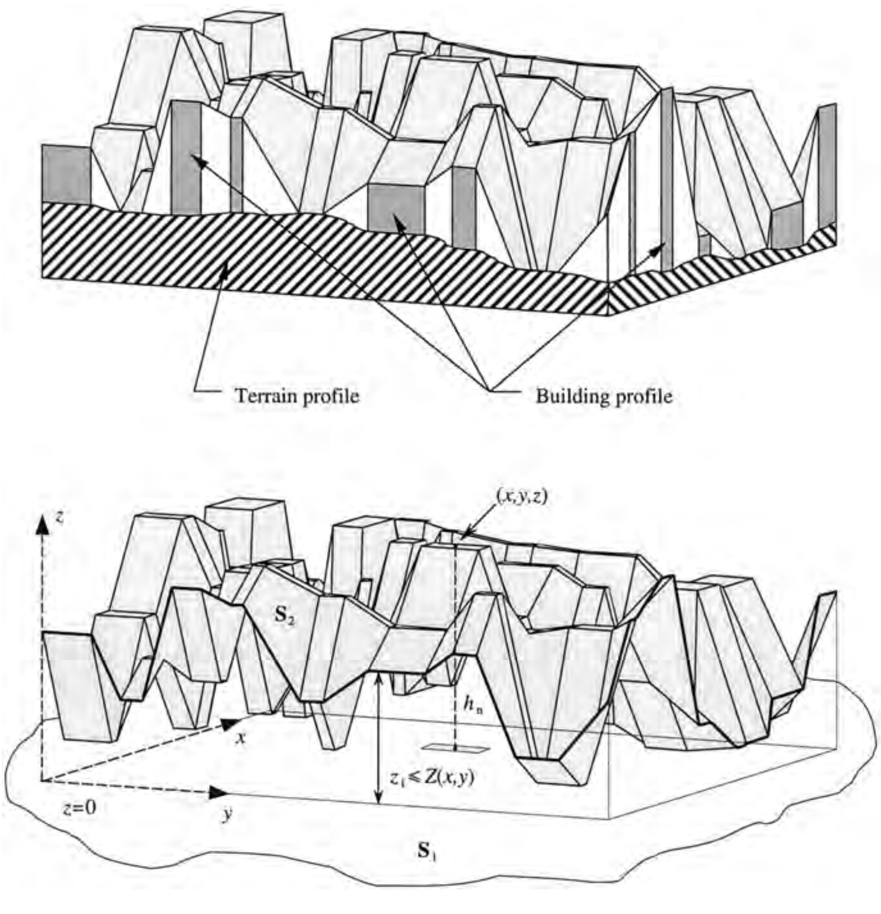


Figure 12.23 Outdoor multipath communication channel presentation according to [31].

between both antennas, the following parameters in addition to those presented above must be used: the typical correlation scales, ℓ_v and ℓ_h, of the complex reflection coefficient from the obstacles with absolute value Γ and the type of building material dominant in the tested area, defining the reflecting properties of obstacles. The geometrical parameters of the built-up terrain allow us to obtain the density of building contours at the ground level, $\gamma_0 = 2\langle L\rangle \nu/\pi$, and then the clearance conditions between receiver and transmitter, e.g., the average horizontal distance of the line of sight $\langle \rho\rangle$ as $\langle \rho\rangle = \gamma_0^{-1}$.

Path Loss ($\bar{L}$)

The various factors obtained above are then used for the computer program based on the three-dimensional parametric model for three types of irregular terrain.

In the case of *forested environments* the following formulas are used according to Ref. 31.

1. For description of the incoherent part of the total average field intensity created by multipath field components due to multiple scattering from trees:

$$\langle I_{\mathrm{inc}}\rangle \approx \frac{\gamma_0 \Gamma \exp(-\gamma_0 d)}{(4\pi)^2}\left[\frac{\Gamma^3}{4(8)^3}\frac{1}{d} + \frac{\Gamma}{32}\left(\frac{\pi}{2\gamma_0}\right)^{1/2}\frac{1}{d^{3/2}} + \frac{1}{2\gamma_0}\frac{1}{d^2}\right] \tag{12.72}$$

2. For the coherent part $\langle I_{\mathrm{co}}\rangle$ of the total field intensity created by the waves coming from the source and specularly reflected from the ground surface,

$$\langle I_{\mathrm{co}}\rangle = \frac{1}{(4\pi)^2}\frac{\exp(-\gamma_0 d)}{d^2}\left[2\sin\frac{kz_1z_2}{d}\right]^2 \tag{12.73}$$

Here d is the distance between the terminal antennas; all other parameters of the terrain and the terminal antennas are defined above.

In the case of *mixed residential areas* the following formulas are used according to Refs. 30 and 31.

1. For description of the incoherent part of the total average field intensity created by multipath components due to independent (single) scattering and diffraction from each obstacle:

$$\langle I_{\mathrm{inc}}\rangle = \frac{\Gamma\lambda\ell_h}{\lambda^2 + (2\pi\ell_h\gamma_0)^2}\frac{\lambda\ell_v}{\lambda^2 + \left[2\pi\ell_v\gamma_0(\bar{h} - z_1)\right]^2}\frac{\left[\left(\lambda d/4\pi^3\right)^2 + (z_2 - \bar{h})^2\right]^{1/2}}{8\pi d^3} \tag{12.74}$$

2. For the coherent part $\langle I_{\mathrm{co}}\rangle$ of the total field intensity created by the waves coming from the source and specularly reflected from the ground surface,

$$\langle I_{\mathrm{co}}\rangle = \exp\left(-\gamma_0 d\frac{\bar{h} - z_1}{z_2 - z_1}\right)\left[\frac{\sin(kz_1z_2/d)}{2\pi d}\right]^2 \tag{12.75}$$

In the case of *built-up* (*urban* and *suburban*) *areas* the following formulas are used according to Refs. 29–31:

1. The incoherent part of the total field intensity due to single scattering and diffraction from buildings' corners and rooftops:

$$\langle I_{\text{inc}1}\rangle = \frac{\Gamma\lambda l_v}{8\pi\left[\lambda^2 + \left[2\pi\ell_v\gamma_0(\bar{h}-z_1)\right]^2\right]d^3}\left[\frac{\lambda d}{4\pi^3} + \left(z_2-\bar{h}\right)^2\right]^{1/2} \tag{12.76a}$$

2. The incoherent part of the total field intensity due to double scattering and diffraction from buildings' corners and rooftops:

$$\langle I_{\text{inc}2}\rangle = \frac{\Gamma^2\lambda^2 l_v\left[(\lambda d/4\pi^3) + \left(z_2-\bar{h}\right)^2\right]}{24\pi^2\left[\lambda^2 + \left[2\pi l_v\gamma_0(\bar{h}-z_1)\right]^2\right]^2 d^3} \tag{12.76b}$$

The coherent part of the total field intensity is described by the same expression, as Eq. (11.75). The total average intensity of the receiving signal for all three types of the terrain is determined by the following formulas:

$$\langle I_{\text{total}}\rangle = \langle I_{\text{co}}\rangle + \langle I_{\text{inc}}\rangle \tag{12.77a}$$

for rural forested and residential areas and

$$\langle I_{\text{total}}\rangle = \langle I_{\text{co}}\rangle + \langle I_{\text{inc}1}\rangle + \langle I_{\text{inc}2}\rangle \tag{12.77b}$$

for urban and suburban areas.

The corresponding mean path loss in decibels (dB), taking into account the free space propagation, can be defined as [1,29–31]

$$\bar{L} = -10\log\left(\lambda^2\langle I_{\text{total}}\rangle\right) \tag{12.78}$$

All these formulas are used to design a link budget for different land wireless communication links, which also needs knowledge about fading phenomena characteristics, slow and fast, for different kinds of environment. Let us briefly describe fading characteristics in a multipath communication system.

12.4. FADING PHENOMENA IN WIRELESS OUTDOOR COMMUNICATION LINKS

As was mentioned above, most wireless communication systems operate in built-up areas where there is no direct line-of-sight (LOS) radio path between the terminals, the transmitter and the receiver, and where due to natural and artificially made obstructions (hills, trees, buildings, towers, etc.), there occur multidiffraction, multireflection, and multiscattering effects (see Fig. 12.3), which cause not only additional losses (with respect to those obtained in LOS above-the-terrain conditions) but also the multipath fading of the signal strength observed at the receiver, which can be separated into fully independent phenomena, the *slow* and the *fast* fading (see definitions in Sec. 12.1) Below we, first of all, will determine the main parameters of the multipath communication channel, the relations between channel parameters and those of the actual signal passing through it. Then, some general statistical descriptions of the multipath outdoor communication links, which are based on well-known stochastic laws, will be introduced.

12.4.1. Parameters of the Multipath Communication Links

In order to compare different multipath communication links and develop some general understanding of how to design such systems, some specific parameters, which grossly quantify the multipath channel, are used. First of all, we will describe the small-scale fast variations of mobile radio signal, which, as was shown in Refs. 4–8 and 31, directly relate to the impulse response of the channel. Next is a wideband (pulse) channel characterization and contains all information necessary to analyze and to simulate any type of radio transmission through the channel. Because of a time-varying impulse response, due to receiver/transmitter or obstructions motion in space, one can finally represent a total received signal as a sum of amplitudes and time delays of the multipath components arrived at the receiver at any instant of time. If through measurements one can obtain information about the signal power delay profile, one can finally determine the main parameters of the multipath communication channel.

Delay Spread Parameters

First important parameters for wideband channels, which can be determined from a signal power delay profile, are *mean excess delay*, *rms delay spread* and *excess delay spread* for the concrete threshold level X (in dB) of the channel (see Fig. 12.24). The *mean*

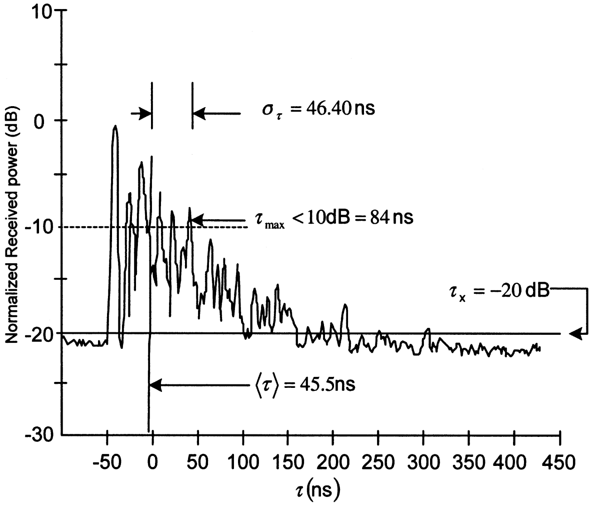


Figure 12.24 Example of the computation of the maximum excess delay for multipath components within 10 dB of the maximum according to [8].

excess delay is the first moment of the power delay profile of the pulse signal and is defined, using multipath signal presentation introduced in Sec. 12.1 by the formula Eq. (12.15), as

$$\langle\tau\rangle = \frac{\sum_{i=0}^{N-1} A_i^2 \tau_i}{\sum_{i=0}^{N-1} A_i^2} = \frac{\sum_{i=0}^{N-1} P(\tau_i)\tau_i}{\sum_{i=0}^{N-1} P(\tau_i)} \tag{12.79}$$

The *rms delay spread* is the square root of the second central moment of the power delay profile and is defined as

$$\sigma_\tau = \sqrt{\langle\tau^2\rangle - \langle\tau\rangle^2} \tag{12.80}$$

where

$$\langle\tau^2\rangle = \frac{\sum_{i=0}^{N-1} A_i^2 \tau_i^2}{\sum_{i=0}^{N-1} A_i^2} = \frac{\sum_{i=0}^{N-1} P(\tau_i)\tau_i^2}{\sum_{i=0}^{N-1} P(\tau_i)} \tag{12.81}$$

Usually [4–8,31], an additional parameter, the *maximum excess delay*, is introduced as the time delay during which multipath energy falls to the threshold level X (dB) below the maximum, that is,

$$\tau_{\max} = \tau_X - \tau_0 \tag{12.82}$$

Here, as above, τ_0 is the time of the first arriving signal at the receiver, τ_X is the maximum delay at which a multipath component is within X dB of the strongest arriving multipath signal (which does not necessarily arrive at τ_0). Figure 12.24 from Ref. 8 illustrates the computation of the maximum excess delay for multipath components within 10 dB of the maximum threshold. The value of τ_X is also called the *excess delay spread* of the power delay profile and in any case must be specified with a threshold of the ratio of the noise floor and the maximum received component.

Coherence Bandwidth

As shown in Refs. 4–8 and 31, the power delay profile in the time domain and the power spectral response in the frequency domain is related through the Fourier transform. Therefore, for the multipath channel full description, the *time delay* parameters in the time domain and the *coherence bandwidth* in the frequency domain are used simultaneously. The coherence bandwidth is the statistical measure of the frequency range over which the channel is considered to be “flat.” In other words, this is the frequency range over which two frequency signals are strongly amplitude correlated. Depending on the degree of amplitude correlation of two frequency separated signals, there are different definitions of this parameter.

The *first definition*: the *coherence bandwidth*, B_c, is a bandwidth over which the frequency correlation function is above 0.9 or 90%, and it equals [8]

$$B_c \approx 0.02\sigma_\tau^{-1} \tag{12.83}$$

The *second definition*: the *coherence bandwidth*, B_c, is a bandwidth over which the frequency correlation function is above 0.5 or 50%, and it equals [8]

$$B_c \approx 0.2\sigma_\tau^{-1} \tag{12.84}$$

Doppler Spread and Coherence Time

Above we considered two parameters, *delay spread* and *coherence bandwidth*, which describe the time dispersive nature of the multipath communication channel in a small-scale area. To obtain information about the time varying nature of the channel caused by movements of either transmitter or receiver or obstructions scatters located around them, new parameters, such as *Doppler spread* and *coherence time*, are usually introduced to describe time variation phenomena of the channel in a small-scale region.

Doppler spread B_D is a measure, which is defined as a range of frequencies over which the received Doppler spectrum is essentially nonzero. It shows the spectral spreading caused by the time rate of change of the mobile radio channel due to relative motions of vehicles (or scatters around them) with respect to the base station. According to Eq. (12.10), the Doppler spread B_D depends on Doppler shift f_D and on the angle θ between the direction of motion of any vehicle and direction of arrival of the reflected and/or scattered waves. If we deal with the complex base-band signal presentation Eq. (12.13), then we can introduce some criterion: if the baseband signal bandwidth is greater than the Doppler spread B_D, the effects of Doppler shift are negligible at the receiver.

Coherence time T_c is the time domain dual of *Doppler spread* and it is used to characterize the time varying nature of the frequency dispersive properties of the channel in time coordinates. There is a simple relationship between these two channel characteristics, that is [see also Eq. (12.10) and all notation there]:

$$T_c \approx \frac{1}{f_{D\max}} \approx \frac{\lambda}{v} \tag{12.85a}$$

We can also define the *coherence time* more strictly, according to Refs. 4–8 and 31, as "the time duration over which two multipath components of receiving signal have a strong potential for amplitude correlation." If so, one can, as above for coherence bandwidth, define the coherence time as the time over which the correlation function of two various signals in the time domain is above 0.5 (or 50%). Then according to Refs. 8 and 11 we get

$$T_c \approx \frac{9}{16\pi f_m} = \frac{9\lambda}{16\pi v} = 0.18\frac{\lambda}{v} \tag{12.85b}$$

As was shown in Ref. 12, this definition can be improved for modern digital communication channels by means of combination of Eqs. (12.85a) and (12.85b) as the geometric mean of them, that is,

$$T_c \approx \frac{0.423}{f_m} = 0.423\frac{\lambda}{v} \tag{12.85c}$$

The definition of coherence time implies that two signals, arriving at the receiver with a time separation greater than T_c, are affected differently by the channel.

12.4.2. Types of Fading

It is clear from channel parameters definitions that the type of signal fading within the mobile radio channel depends on the nature of the transmitting signal with respect to the characteristics of the channel. In other words, depending on the relation between the signal parameters, such as *bandwidth* B_S and *symbol period* T_S, and the corresponding channel parameters, such as *coherence bandwidth* B_c and *rms delay spread* σ_τ (or *Doppler spread* B_D and *coherence time* T_c), different transmitted signals will undergo different types of fading. As was shown by Rappaport [8], there are *four possible effects* due to the time and frequency dispersion mechanisms in a mobile radio channel, which are manifested depending on the balance of the above-mentioned parameters of the signal and of the channel. The multipath time delay spread leads to *time dispersion* and *frequency selective fading*, whereas Doppler frequency spread leads to *frequency dispersion* and *time selective fading*. Separation between these four types of small-scale fading for impulse response of multipath radio channel is explained in Table 12.1 according to Ref. 8:

A. Fading due to multipath time delay spread. Time dispersion due to multipath phenomena causes small-scale fading, either *flat* or *frequency selective*:

A.1. The small-scale fading is characterized as *flat* if the mobile channel has a constant-gain and linear-phase impulse response over a bandwidth, which is *greater* than the bandwidth of the transmitted signal. Moreover, the signal bandwidth in the time domain exceeds the signal delay spread, i.e., $T_S \gg \sigma_\tau$ (see the top rows of the last column of Table 12.1). As it can be seen, a flat fading channel can be defined as *narrowband* channel, since the bandwidth of the applied signal is *narrow* with respect to the channel flat fading bandwidth in the frequency domain, that is, $B_S \ll B_c$. At the same time, the flat fading channel is *amplitude-varying* channel, since there is a deep fading of the transmitted signal which occurs within such a channel.

A.2. The small-scale fading is characterized as a *frequency selective*, if the mobile channel has a constant-gain and linear-phase impulse response over a bandwidth which is *smaller* than the bandwidth of the transmitted signal in the

Table 12.1 Types of Fading according to [8]

General type of fading	Type of fading	Type of channel	
		Narrowband	
(A) *Small-scale fading*	(A.1): *Flat fading*	(A.1.1):	$B_S \ll B_c$
		(A.1.2):	$T_S \gg \sigma_\tau$
(Based on multipath time Delay Spread)		*Wideband*	
	(A.2): *Frequency selective fading*	(A.2.1):	$B_S > B_c$
		(A.2.2):	$T_S < \sigma_\tau$
		Narrowband	
(B) *Small-scale fading*	(B.1): *Fast fading*	(B.1.1):	$B_S < B_D$
		(B.1.2):	$T_S > T_c$
(Based on Doppler frequency spread)		*Wideband*	
	(B.2): *Slow fading*	(B.2.1):	$B_S \gg B_D$
		(B.2.2):	$T_S \ll T_c$

frequency domain, as well as its impulse response has a multiple delay spread greater than the bandwidth of the transmitted signal waveform, i.e., $T_S \ll \sigma_\tau$. These conditions are presented at the last column of Table 12.1. As can be seen, a frequency-selective fading channel can be defined as *wideband* channel, since the bandwidth of the spectrum $S(f)$ of the transmitted signal is *greater* than the channel frequency-selective fading bandwidth in the frequency domain (the coherence bandwidth), that is, $B_S > B_c$.

B. Fading due to Doppler spread. Depending on how rapidly the transmitted baseband signal changes with respect to the rate of change of the channel, a channel may be classified as a *fast fading* or *slow fading* channel.

B.1. The channel, in which the channel impulse response changes rapidly within the pulse (symbol) duration, is called a *fast fading* channel. In other words, in such a channel its coherence time is smaller than the symbol period of the transmitted signal. At the same time the Doppler spread bandwidth of the channel in the frequency domain is greater than the bandwidth of the transmitted signal (see the last column in Table 12.1). That is, $B_S < B_D$ and $T_c < T_S$. These effects cause frequency dispersion (also called *time-selective fading* [8]) due to Doppler spreading, which leads to signal distortion.

B.2. The channel, in which the channel impulse response changes at a rate *slower* than the transmitted baseband signal $u(t)$, is called a *slow fading* channel. In this case the channel may be assumed to be static over one or several bandwidth intervals. In the time domain, this implies that the reciprocal bandwidth of signal is much smaller than the coherence time of the channel and in the frequency domain the Doppler spread of the channel is less than the bandwidth of the baseband signal, that is, $B_S \gg B_D$ and $T_S \ll T_c$. Both these conditions are presented at the bottom rows of the last column in Table 12.1. It is important to note that velocity of the moving vehicle or moving obstructions within the channel, as well as the baseband signal determine whether a signal undergoes fast fading or slow fading. All situations within the wireless communication channel, described above, are summarized in Fig. 12.25 [8].

12.4.3. Mathematical Modeling of Fast Fading

Now we will discuss the question of the existence of a suitable statistical model for satisfactory description of multipath fast fading channels. Several multipath models have been proposed to describe the observed random signal envelope and phase in a mobile channel. The earliest 2D models were developed in Refs. 32–34 and were based on the random interference of direct (incident) waves and waves scattered from the flat sides of buildings, screens randomly distributed above the terrain. Because such a model is widely used for the description of wireless short-scale fading communication channels, we will present this model to the reader. The model assumes a fixed transmitter with a vertically polarized omnidirectional antenna and a moving receiver also with omnidirectional antenna. The signal at the receiver is assumed to comprise N horizontally traveling plane waves with each wave with number i having equal average amplitude A_i and with statistically independent angles of arrival (α_i) (azimuth angles) and phase angle (ϕ_i) distributions. The assumption of equal average amplitude of each ith wave is based on the absence of an LOS component with respect to scattered components arriving at the receiver. Moreover, phase angles distribution is assumed to be uniform in the interval $[0, 2\pi]$, that is, the angle distribution function is equal $P(\phi_i) = (2\pi)^{-1}$. A typical ith wave

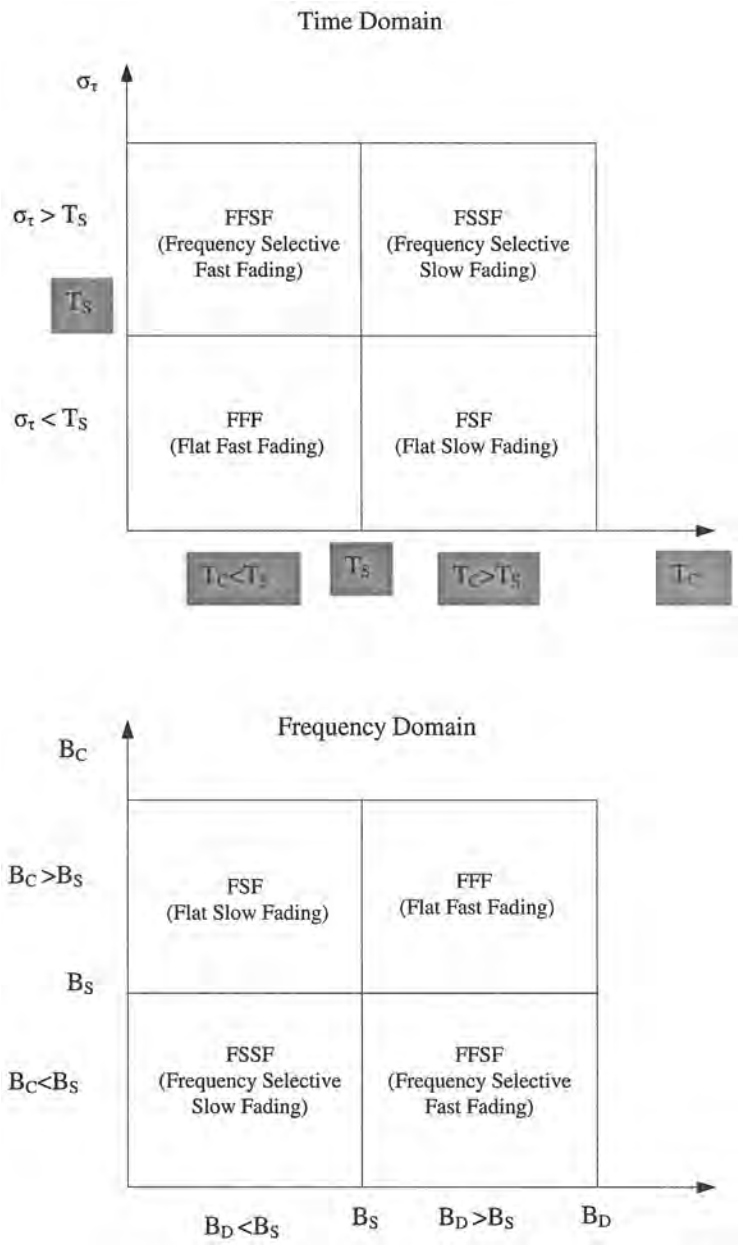


Figure 12.25 Types of fading experienced in the multipath outdoor communication link [8].

arriving at an angle α_i to the x axis is shown in Fig. 12.26. The receiver moves with a velocity v in the x direction, so the Doppler shift in z axis, according to Eq. (12.10), can be now rewritten as

$$f_D = \frac{v}{\lambda}\cos\alpha_i \tag{12.86}$$

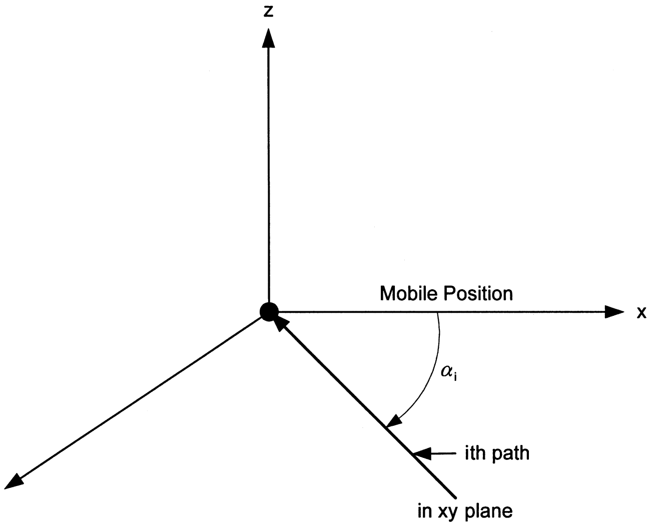


Figure 12.26 Graphical presentation of the multipath phenomena according to Clarke's model [32].

The mean square value of the amplitude A_i of such uniformly distributed individual waves is constant

$$E\{A_i^2\} \equiv \langle A_i^2 \rangle = \frac{E_0}{N} \tag{12.87}$$

because $N =$ constant and the real amplitude of local average field E_0 is also assumed to be a constant.

Let us consider the vertically polarized plane electromagnetic waves arriving at the moving receiver, which usually have one E-field component (E_z) and two H-field components (H_x and H_y, see [4–8]). Without any loss of generality of the problem, because for each field component the same technique is used, let us consider only the E-field component and present it at the receiving point as [32–34]

$$E_z = E_0 \sum_{i=1}^{N} A_i \cos(\omega_c t + \theta_i) \tag{12.88}$$

where $\omega_c = 2\pi f_c$, f_c is the carrier frequency, $\theta_i = \omega_i t + \phi_i$ is the random phase of the ith arriving component of total signal, and $\omega_i = 2\pi f_i$ represents the Doppler shift experienced by the ith individual wave. The amplitudes of all three electromagnetic field components are normalized such that the ensemble average of the amplitude A_i is given by $\sum_{i=1}^{N} \langle A_i^2 \rangle = 1$.

Since the Doppler shift is small with respect to the carrier frequency, all field components may be modeled as narrow band random processes and approximated as gaussian random variables, if $N \to \infty$, with a uniform phase distribution in the interval $[0,2\pi]$. If so, the E-field component can be expressed in the following form:

$$E_z = C(t)\cos(\omega_c t) - S(t)\sin(\omega_c t) \tag{12.89}$$

where $C(t)$ and $S(t)$ are the in phase and quadrature components that would be detected by a suitable receiver [32–34]:

$$\begin{aligned} C(t) &= \sum_{i=1}^{N} A_i \cos(\omega_i t + \theta_i) \\ S(t) &= \sum_{i=1}^{N} A_i \sin(\omega_i t + \theta_i) \end{aligned} \tag{12.90}$$

According to the assumptions above, both components $C(t)$ and $S(t)$ are independent Gaussian random processes. They are uncorrelated zero-mean gaussian random variables, that is,

$$\langle S \rangle = \langle C \rangle = \langle E_z \rangle = 0 \tag{12.91}$$

with an equal variance σ^2 (the mean signal power) given by

$$\sigma^2 \equiv \langle |E_z|^2 \rangle = \langle S^2 \rangle = \langle C^2 \rangle = \frac{E_0^2}{2} \tag{12.92}$$

The envelope of the received E-field component can be presented as

$$|E(t)| = \sqrt{S^2(t) + C^2(t)} = r(t) \tag{12.93}$$

Since components $C(t)$ and $S(t)$ are independent gaussian random variables that satisfy Eqs. (12.91)–(12.93), the random received signal envelope r has a Rayleigh distribution [4–8,12] (below, we will talk about the probability density (PDF) and cumulative distribution (CDF) functions of the signal envelope $r(t)$). Using such a definition of the signal envelope, we can now describe two mainly used descriptions of wireless multipath communication links, the Rayleigh and the Rician, according to Refs. 4–8 and 12.

Rayleigh Multipath Fast Fading Statistics

In wireless communication channels, stationary or mobile, the Rayleigh distribution is commonly used to describe the signal's spatial or temporal (i.e., small-scale or fast) fading. As was shown above, a Rayleigh distribution can be obtained mathematically as the limit envelope of the sum of two quadrature gaussian signals, $C(t)$ and $S(t)$. Again, if the phase of multipath components is uniformly distributed over the range of $[0, 2\pi]$, then we deal with a zero-mean Rayleigh distribution of random variable r, the PDF of which can be presented in the following form [2–5,8,9]:

$$\text{PDF}(r) = \frac{r}{\sigma_r^2} \exp\left(-\frac{r^2}{2\sigma_r^2}\right) \qquad \text{for } r \geq 0 \tag{12.94}$$

Here the variance σ_r^2 or average power of the received signal envelope for the Rayleigh distribution can be determined as $\sigma_r^2 \equiv E[r^2] - E^2[r]$, where $E[r]$ is an expected value usually used in statistics [2]. The PDF distribution Eq. (12.94) completely describes the random received signal envelope $r(t)$ defined in Clarke's model by the Eq. (12.93).

In the formula Eq. (12.94) the maximum value of $\mathrm{PDF}(r) = \exp(-0.5)/\sigma_r = 0.6065/\sigma_r$ corresponds to random variable $r = \sigma_r$.

One can also operate with the so-called *mean value*, the *rms value* and the *median value* of random variable x. The definition of these parameters follows from the Rayleigh CDF presentation, which describes the probability of the event that the envelope of received signal strength (voltage) does not exceed a specified value R [2–5,8,9]:

$$\mathrm{CDF}(R) = \Pr(r \le R) = \int_0^R \mathrm{PDF}(r)\,dr = 1 - \exp\left(-\frac{R^2}{2\sigma_r^2}\right) \tag{12.95}$$

The mean value of the Rayleigh distributed signal strength (voltage), r_{mean} (in the literature it is also denoted as an expected value $E[x]$ [2]), can be obtained from the following conditions:

$$r_{\mathrm{mean}} \equiv E[r] = \int_0^\infty r\,\mathrm{PDF}(r)\,dr = \sigma\sqrt{\frac{\pi}{2}} \approx 1.253\sigma_r \tag{12.96a}$$

If so, the *rms value* of the signal envelope is defined as the square root of the mean square, that is,

$$\mathrm{rms} = \sqrt{2}\,\sigma_r \approx 1.414\sigma_r \tag{12.96b}$$

The *median value* of Rayleigh distributed signal strength envelope is defined from the following conditions [2]:

$$\frac{1}{2} = \int_0^{r_{\mathrm{median}}} \mathrm{PDF}(r)\,dr$$

from which follows that

$$r_{\mathrm{median}} = 1.177\sigma_r \tag{12.96c}$$

As follows from Eqs. (12.96a)–(12.96c), the difference between the *mean* and the *median* values is $\sim 0.076\sigma_r$ and their PDF for a Rayleigh fading signal envelope differ by only 0.55 dB. The differences between the rms value and two other values are higher.

Rician Multipath Fading Statistics

As was mentioned in Sec. 12.1, in a wireless communication link, multipath components arrive at the receiver due to multiple reflection, diffraction and scattering from various obstruction around the two terminals, the transmitter and the receiver. Also, a line-of-sight (LOS) component, which describes signal loss along the path of direct visibility (called the *dominant path* [4–9,12]) between both antennas, is often found at the receiver. The PDF of such a received signal is usually said to be *Rician*. To estimate the contribution of each component, dominant (or LOS) and multipath, for the resulting signal at the receiver, the Rician parameter K is usually introduced, as a ratio between these components, i.e.,

$$K = \frac{\mathrm{LOS} - \mathrm{component\ power}}{\mathrm{multipath} - \mathrm{component\ power}} \tag{12.97}$$

The Rician PDF distribution of the signal strength or voltage envelope r can be defined as [4–9,12]:

$$\mathrm{PDF}(r) = \frac{r}{\sigma_r^2}\exp\left(-\frac{r^2 + A^2}{2\sigma_r^2}\right)I_0\left(\frac{Ar}{\sigma_r^2}\right) \qquad \text{for } A > 0, \quad r \geq 0 \tag{12.98}$$

where A denotes the peak strength or voltage of the dominant component envelope and $I_0(\cdot)$ is the modified Bessel function of the first kind and zero order. According to the definition Eq. (12.98), we can now rewrite the parameter K, which was defined above as the ratio between the *dominant* and the *multipath* component power. It is given by

$$K = \frac{A^2}{2\sigma_r^2} \tag{12.99a}$$

or in terms of dB

$$K = 10\log\frac{A^2}{2\sigma_r^2}\ \mathrm{dB} \tag{12.99b}$$

Using Eqs. (12.99a) in (12.98), we can rewrite Eq. (12.98) as a function only of K:

$$\mathrm{PDF}(x) = \frac{r}{\sigma_r^2}\exp\left(-\frac{r^2}{2\sigma_r^2}\right)\exp(-K)I_0\left(\frac{r}{\sigma_r}\sqrt{2K}\right) \tag{12.100}$$

from which for $K=0$ and $\exp(-K)=1$ follows the worst-case Rayleigh PDF Eq. (12.94) when there is no dominant signal component. Conversely, in a situation of good clearance between two terminals with no multipath components, that is $K \to \infty$, the Rician fading approaches a gaussian one yielding a "Dirac-delta shaped" PDF described by the formula Eq. (12.101) (see below). Hence, the Rician distribution's PDF approaches the Rayleigh PDF and the gaussian PDF, if the Rician K factor approaches zero and infinity, respectively. These features of the Rician PDF and CDF (in dB) can be seen from illustrations presented in Figs. 12.27 and 12.28, respectively.

12.4.4. Mathematical Modeling of Slow Fading

Gaussian Fading Statistics

A very interesting situation within the wireless communication link is that, when *good clearance* between two terminals or *slow* fading at the receiver occurs, we find a tendency to gaussian (also called *normal*) distribution of the random received signal strength or voltage r [4–8,12,31] with the following PDF:

$$\mathrm{PDF}(r) = \frac{1}{\sigma_L\sqrt{2\pi}}\exp\left[-\frac{(r-\bar{r})^2}{2\sigma_L}\right] \tag{12.101}$$

Here $\bar{r} \equiv \langle r\rangle$ is the mean value of the random signal level, σ_L is the value of the received signal strength or voltage envelope and $\sigma_L^2 = \langle r^2 - \bar{r}^2\rangle$ is the variance or

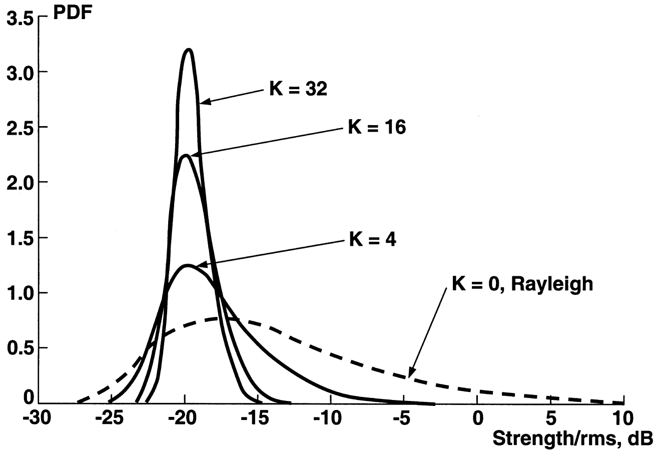


Figure 12.27 The Rician PDF distribution in the logarithmic scale for different parameters $K=0$, 4, 16, and 32 [11].

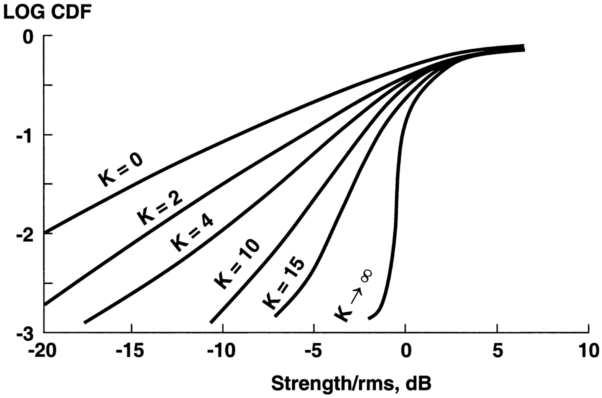


Figure 12.28 The Rician LOG CDF distribution for different parameters K: $K=0$ corresponds to the Rayleigh distribution; $K \to \infty$ corresponds to the gaussian distribution [11].

time-average power ($\langle w \rangle$ is a sign of averaging of variable w) of the received signal envelope. This PDF can be obtained only as a result of the random interference of a large number of signals with randomly distributed amplitudes (strength or voltage) and phase. If the phase of the interfering signals is uniformly distributed over the range of $[0, 2\pi]$, then one can talk about a zero-mean gaussian distribution of random variable r. In this case we define the PDF of such a process by Eq. (12.101) with $\bar{r}=0$ and $\sigma_L^2 = \langle r^2 \rangle$, and CDF as [10–12]

$$\begin{aligned} \mathrm{CDF}(R) &= \Pr(r \le R) = \int_0^R \mathrm{PDF}(r)\,dr \\ &= \frac{1}{\sigma_L\sqrt{2\pi}} \int_{-\infty}^{X} \exp\left[-\frac{(r-\bar{r})^2}{2\sigma_L^2}\right] = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{R-\bar{r}}{\sqrt{2}\sigma_L}\right) \end{aligned} \tag{12.102}$$

where the error function is defined by

$$\mathrm{erf}(w) = \frac{2}{\sqrt{\pi}} \int_0^w \exp\left(-y^2\right) dy \tag{12.103}$$

From Eq. (12.101) it is clearly seen that the value $\bar{r} = 0$ corresponds to the maximum of the PDF, which equals $\mathrm{PDF}(0) = 1/\sigma_L\sqrt{2\pi}$. For $r = \sigma_L$, it follows from Eq. (12.101) that $\mathrm{PDF}(\sigma_L) = 1/\sigma_L\sqrt{2\pi e}$, where $e \approx 2.71\ldots$ Using probability distribution functions Eqs. (11.101) and (12.102) and obtaining during measurements some information about the variance $\sigma_L^2 = \langle r^2 \rangle$ of received signals, we can easily predict the slow fading in corresponding communication links.

We must note here that *in decibels* slow fading, described in voltage by normal or gaussian distributions Eqs. (12.101) and (12.102), is usually described by the log-normal distribution [4–9,12]. In the propagation channels with log-normal *slow fading* or *shadowing*, the effects of fading can be represented by the local path loss $L(r)$ at an arbitrary local point x of radio path randomly distributed inside the propagation channel [8,12]:

$$L(x) = \bar{L}(x) + R_\sigma \,\mathrm{dB} \tag{12.104}$$

where R_σ is a gaussian distributed random variable (in dB) with mean value $\bar{r} = 0$ and with standard deviation $\sigma_L = \sqrt{\overline{(r - \bar{r})^2}}$ (also in dB) and $\bar{L}(x)$ is the average large-scale path loss for arbitrary point r between both terminal antennas, which can be presented as follows [8]:

$$\bar{L}(x) = \bar{L}(x_0) + 10n \log \frac{x}{x_0} \,\mathrm{dB} \tag{12.105}$$

Here x_0 is the reference distance, close to the transmitter, which is determined from concrete measurements in the urban scene, $\bar{L}(x_0)$ is the average path loss at the distance x_0 from the transmitter, and n is the path loss exponent, which indicates the rate of signal power attenuation with distance.

As follows from Eqs. (12.104) and (12.105), the *log-normal shadowing*, as a slow fading phenomenon, implies that measured signal levels at a concrete distance between the transmitter and the receiver have a normal, or gaussian, distribution about the distance-dependent mean path loss from Eq. (12.105), where the measured signal levels have values in dB units. The standard deviation σ_L of the gaussian distribution, which describes the shadowing effect also has units in dB. The probability density function of the shadowing component R_σ in Eq. (12.104), as a zero-mean gaussian variable with standard deviation σ_L, can be presented at the same manner than Eq. (12.101), that is, [8]

$$\mathrm{PDF}(R_\sigma) = \frac{1}{\sigma_L\sqrt{2\pi}} \exp\left(-\frac{R_\sigma^2}{2\sigma_L^2}\right) \tag{12.106}$$

The probability that the shadowing increases in Eq. (12.105) the average path loss $\bar{L}(r)$ by at least Z dB can be presented using the *complimentary cumulative distribution* function $\mathrm{CCDF}(R) = 1 - \mathrm{CDF}(R)$, which we denote as $Q(Z/\sigma_L)$,

$$Q\left(\frac{Z}{\sigma_L}\right) \equiv \mathrm{CCDF}\left(\frac{Z}{\sigma_L}\right) = 1 - \mathrm{CDF}\left(\frac{Z}{\sigma_L}\right) \equiv \Pr(R_\sigma > Z) \tag{12.107}$$

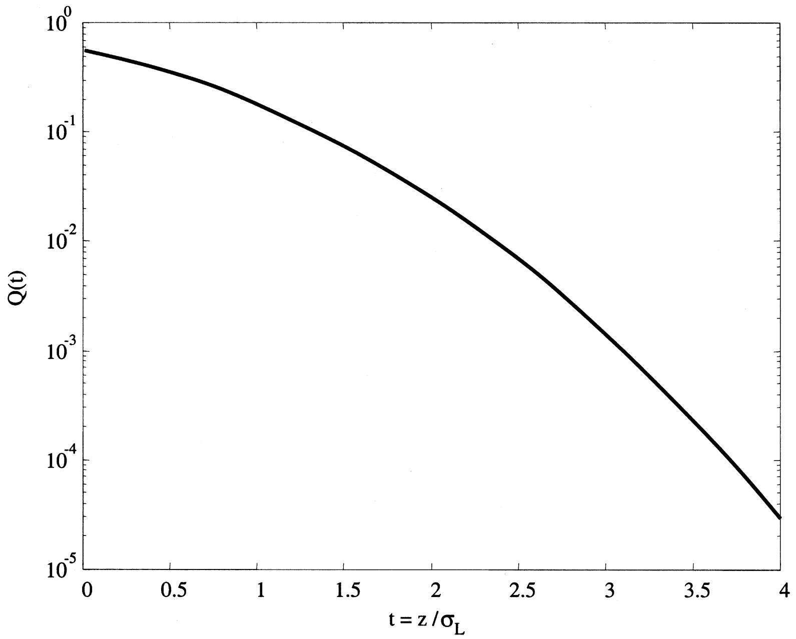


Figure 12.29 The complementary cumulative normal distribution function $Q(Z/\sigma_L)$ versus normalized parameter $w = Z/\sigma_L$ [8].

where $\mathrm{CDF}(Z/\sigma_L)$ is described in the same manner, as in Eq. (12.102), through the error function (erf) defined by Eq. (12.103), that is,

$$\mathrm{CDF}\left(\frac{Z}{\sigma_L}\right) \equiv \Pr(R_\sigma < Z) = \int_0^Z \mathrm{PDF}(R_\sigma)\, dR_\sigma \tag{12.108}$$

If so, we finally have for the complementary cumulative normal distribution function $Q(Z/\sigma_L)$ the following expression by putting in Eq. (12.107) the normalized variable $w = Z/\sigma_L$ [8]:

$$Q(w) = \frac{1}{\sqrt{2\pi}} \int_{r=w}^{\infty} \exp\left(-\frac{r^2}{2}\right) dx = \frac{1}{2}\mathrm{erf}\left(\frac{w}{\sqrt{2}}\right) \tag{12.109}$$

This function is plotted in Fig. 12.29 and can be used to evaluate the shadowing margin needed for any location variability in accordance with Eq. (12.109).

12.5. LINK BUDGET DESIGN IN WIRELESS OUTDOOR COMMUNICATION SYSTEMS

Link budget is the main parameter of wireless communication systems, both indoor and outdoor. Because the subject of our chapter is the outdoor communication, below we will deal with the land communication links, taking into account all essential parts of the communication system, as is shown in Fig. 12.1, that is, the total path loss within the communication channel, including fading propagation phenomena, the losses inside the terminal antennas, the transmitter and receiver, and also the thermal noise (adaptive) inside the electronic channels. Some characteristics, such as thermal noise, gains of antennas and antenna losses, and link average losses, (called above the *average path loss*) are simple to evaluate using knowledge obtained in Secs. 12.1–12.3. A more complicated question is how to obtain information about the multiplicative noise, that is, about the long-term (or slow) and short-term (or fast) fading. Let us briefly discuss this subject and give some simple examples on how to estimate such propagation characteristics within the outdoor communication link.

12.5.1. Link Budget Accounting Shadowing Effects (Slow Fading)

To take into account the slow fading or the shadow effects within the communication link, the corresponding graph is plotted in Fig. 12.30 for link budget design. The later takes into account both LOS and NLOS (clutter) conditions, and by the slow fading component caused by shadowing, that is,

$$L_{\text{total}} = L_{\text{LOS}} + L_{\text{NLOS}} + R_{\sigma} \tag{12.110}$$

An example of such a link budget design is shown in Fig. 12.30 (according to Ref. 5) for a wireless communication system providing 90% successful communications at the fringe of radio coverage. In other words, the communication system was examined for the case where 90% of locations at the boundaries of the tested area had acceptable radio coverage. Within the tested area, a greater percentage of vehicle antenna location has acceptable coverage, so here the total path loss will be less. We can illustrate this effect

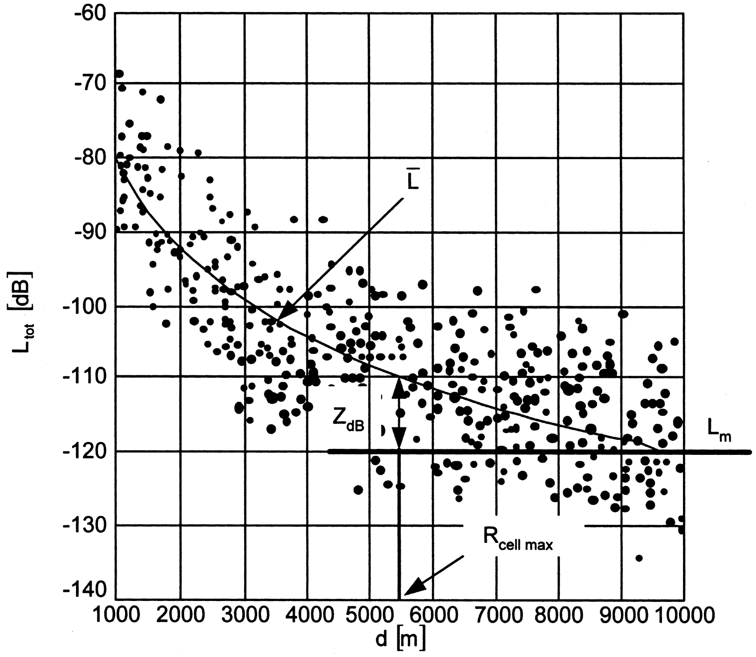


Figure 12.30 Effect of shadowing margin on tested site range.

by using the results presented in Fig. 12.30. In fact, if the maximum acceptable pass loss is 120 dB, the probability that $L_{\text{total}} > 120\,\text{dB}$ is [5]

$$\Pr(L_{\text{total}} > 120) = \Pr(L_{\text{LOS}} + L_{\text{NLOS}} + R_\sigma > 120)$$
$$= \Pr(R_\sigma > 120 - L_{\text{LOS}} - L_{\text{NLOS}}) \equiv Q\left(\frac{120 - L_{\text{LOS}} - L_{\text{NLOS}}}{\sigma_L}\right) \tag{12.111}$$

This probability was denoted as an outage probability P_{out}. According to Eq. (12.107) the fraction of locations covered by the transmitter at a range r is simply [5]

$$\text{Coverage fraction} = P_{cf}(d) = 1 - P_{\text{out}} \tag{12.112}$$

We must note here, that the outage probability Eq. (12.112) does not take into account the effects of signal-to-interference ratio, which we consider below, and here is purely caused by shadow effects. In general terms Eq. (12.112) can be expressed as [5]

$$P_{cf}(d) = 1 - Q\left(\frac{L_m - \bar{L}(d)}{\sigma_L}\right) = 1 - Q\left(\frac{Z}{\sigma_L}\right) \tag{12.113}$$

where L_m is the maximum acceptable path loss and $\bar{L}(d)$ is the median path loss within the actual communication system, evaluated at a distance d; $Z = L_m - \bar{L}(d) \equiv L_{SF}$ is the fade margin chosen for such a system (see all notations in Fig. 12.30).

Example. Find a distance d between the terminal antennas within a wireless land communication link, which operates at frequency of 1 GHz and provides 80% of successful communications at the fringe of coverage. Let us assume that propagation LOS effects are described by a free space model Eq. (12.32) for isotropic antennas and the clutter factor of the rough terrain (hills and trees) is $L_{\text{NLOS}} = 38.5\,\text{dB}$, with shadowing of location variability $\sigma_L = 8\,\text{dB}$. The maximum acceptable path loss is $L_m = 150\,\text{dB}$.

Solution. According to link budget Eq. (12.110) and free-space propagation model Eq. (12.32) for isotropic antennas:

$$L_{\text{total}} = 32.44 + 20\log d_{\text{km}} + 20\log f_{\text{MHz}} + L_{\text{NLOS}} + Z$$

To find Z we take value $t = Z/\sigma_L$ from Fig. 12.26, taking into account that the probability of shadow is $Q(t) = 100\% - 80\% = 20\%$ or $Q = 0.2$. From Fig. 12.26 this occurs when $t = 0.75$, that is, $Z = t\sigma_L = 0.75(8) = 6\,\text{dB}$, from which we easily obtain that

$$20\log d_{\text{km}} = L_{\text{total}} - 32.44 - 20\log f_{\text{MHz}} - L_{\text{NLOS}} - Z$$

or

$$\log d_{\text{km}} = \frac{L_{\text{total}} - 32.44 - 20\log f_{\text{MHz}} - L_{\text{NLOS}} - Z}{20} \approx 0.65$$

So the distance between antennas in such a wireless communication system is $d = 10^{0.65} = 4.5\,\text{km}$.

12.5.2. Link Budget Design of the Channels with Fast Fading

As was mentioned above, fast fading is a stochastic phenomenon that occurs within the multipath communication links and can be described commonly by Rician statistics, that is, by Rician PDF and CDF, from which main parameters of fading can be obtained. In fact, if we now rewrite Eq. (12.110) through parameter rms and Rician K parameter [8],

$$\mathrm{PDF}(r) = 2\frac{r}{(\mathrm{rms})^2}\exp\left(-\frac{r^2}{(\mathrm{rms})^2}\right)\exp(-K)I_0\left(2\frac{r}{\mathrm{rms}}\sqrt{K}\right) \tag{12.114}$$

we can present the fast fade margin, L_{FF}, as [9]

$$L_{\mathrm{FF}} = 10\log\sigma_{\mathrm{FF}} = 10\log\left[\int_0^{\infty} r^2\mathrm{PDF}(r)\,dr - \left(\int_0^{\infty} r\,\mathrm{PDF}(r)\,dr\right)^2\right]\mathrm{dB}$$

Using derivations carried out in Ref. 9, we finally get the expression

$$\sigma_{\mathrm{FF}} = \left[2(\mathrm{rms})^2 e^{-K}\right]\left[\frac{1}{2}e^{K}\int_0^{\infty} y^3 e^{-y^2} I_0(2y\sqrt{K})\,dy - \left(\int_0^{\infty} y^2 e^{-y^2} I_0(2y\sqrt{K})\,dy\right)^2\right]^{1/2} \tag{12.115}$$

Then, the total path loss in the multipath channel with the fast fading only, can be easily obtained by the knowledge of average path loss, consisting the LOS effects and NLOS effects, according to the commonly used models described in Sec. 12.3, and the fade margin, i.e.,

$$L_{\mathrm{total}} = \bar{L} + L_{\mathrm{FF}} \tag{12.116a}$$

In the *common case* of a communication link with multipath and shadowing effects of radio propagation, the corresponding link budget equation is [9]

$$L_{\mathrm{total}} = \bar{L} + L_{\mathrm{FF}} + L_{\mathrm{SF}} \tag{12.116b}$$

Through the corresponding knowledge of the signal-to-noise ratio, the antenna gains and the sensitivity of wireless system, i.e., the maximum acceptable path loss (see above all definitions), we can finally design a full budget of a wireless outdoor communication system.

Example. Let us consider a communication system designed in the built-up area with, as discussed above, a probability of shadowing $Q(t) = 0.2$ with shadowing of location variability $\sigma_L = 8\,\mathrm{dB}$, and with a probability of fast fading $\mathrm{CDF}(\sigma_{\mathrm{FF}}/\mathrm{rms}) = 0.1$ described by the Rician statistics with the standard deviation about the mean $\sigma_r = 5\,\mathrm{dB}$ and with Rician parameter $K = 15$.

Find the link budget of such a wireless outdoor communication system, if the average path loss inside the system is 105 dB, the antenna gains are $G_T = G_R = -5\,\mathrm{dB}$, and signal-to-noise ratio is 10 dB. Estimate the efficiency of its performance if the maximum acceptable path loss of the system $L_m = 120\,\mathrm{dB}$.

Solution. Taking into account Eq. (12.32) for mean path loss in free space for isotropic antennas and the effects of slow and fast fading described by Eq. (11.116b), we will rewrite the link budget of the system as

$$L_{\text{total}} = \bar{L} + L_{\text{FF}} + L_{\text{SF}} + \frac{S}{N} + G_T + G_R$$

1. For $Q(t) = 0.2$ and $\sigma_L = 8\,\text{dB}$, we get $t = 0.75$ (see example above) and then the slow fade margin is $L_{\text{SF}} \equiv Z = 8 \times 0.75 = 6\,\text{dB}$.
2. Taking into account that $\text{CDF}(\sigma_{\text{FF}}/\text{rms}) = 0.1$ and that $\text{rms} \approx 1.4\sigma_r = 7\,\text{dB}$, we have from Fig. 12.28 (check curve for $K = 15$), that $\sigma_{\text{FF}\,\text{dB}} - (rms)_{\text{dB}} \equiv L_{\text{FF}} - 7 = -3\,\text{dB}$, from which we get $L_{\text{FF}} = 4\,\text{dB}$.
3. The total path loss of the system is

$$L_{\text{total}} = 105 + 6 + 4 + 10 - 5 - 5 = 115\,\text{dB}$$

Because the maximum acceptable path loss is 120 dB, than is, higher than the real loss within the system, the sensitivity of the system is enough to obtain information from any vehicle or subscriber located within the area of service.

12.6. CELLULAR CONCEPT FOR WIRELESS SYSTEMS

Usually the actual design of modern wireless systems relates to the so-called *cellular concept* of wireless communications in built-up areas [1–8,28,35,36], which allows the designers to decrease natural background noises within the propagation channels and to exclude deep fading affecting the signal at the input of the receiver.

Let us ask a question: What is the "cellular principle" and how may we construct each "cell" in a completed cellular system? The simplest "radio cell" one can construct uses a base station at the center of such a cell, which determines the coverage area from its antenna. This coverage area is defined by the range where a stable signal from this station can be received. Figure 12.31 illustrates the distribution of such cells. It is seen that there exist regions of overlap with neighboring "radio cells," where stable reception from neighboring base stations can be obtained. From this scheme it also follows that different frequencies should be used in these cells which surround the tested central "cell." On the other hand, the same frequencies can be used for the cells farthest from the central one. This is the so-called *cells repeating* or *reuse of operating frequencies* principle. At the same time, the reuse of the same radio channels and frequencies within the neighboring cells is limited by preplanned *cochannel interference*. Moreover, in the process of cellular systems design in various built-up areas, it is very important to predict the influence of propagation phenomena within the corresponding communication channels on variations of the main parameters of the cellular system, on the construction and splitting of cellular maps. All these questions will be discussed below.

12.6.1. Main Characteristics of a Cell

The main question is why is it useful to use such a cell structure with a lot of base stations, as shown in Fig. 12.31, instead of using a more powerful antenna which will cover a large

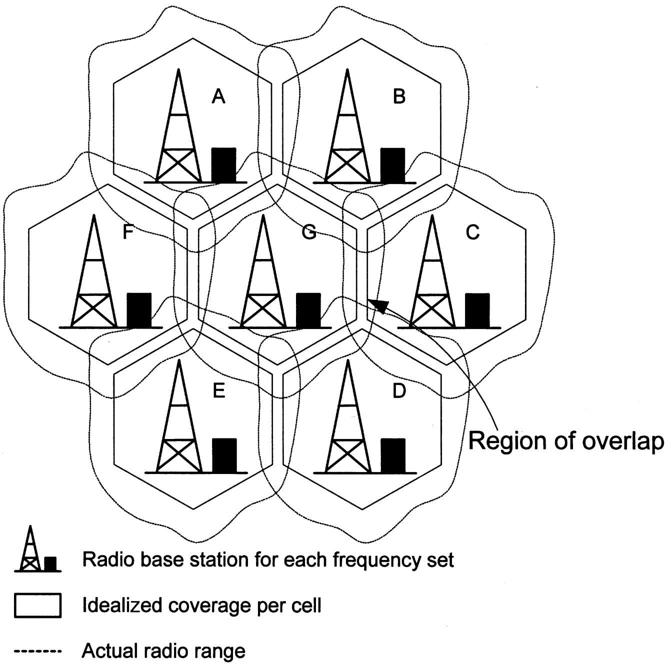

A
B
F
G
C
E
D
Region of overlap
Radio base station for each frequency set
Idealized coverage per cell
------- Actual radio range

Figure 12.31 The concept of cell distribution and cellular map pattern according to [35,36].

area and will service enough subscribers? Let us present a simple example. For a flat terrain, R can be defined as the radius of a circle surrounding the base station on the topographical map of selected area. Then the area that is covered by the base station antenna is approximately πR^2(in km^2). If, for example, radius $R = 2$ km, the single cell coverage area is $S = 6.28$ km^2, which provides service to about 200 subscribers of a wireless personal communication channel [5]. For $R = 20$ km, $S = 628$ km^2 and the number of subscribers grows to 20,000. But to service 120,000 subscribers, the cell should be designed with a radius $R = 25$ km, with an area $S = 1960$ km^2 [5]. As follows from the above estimations, to use only a single antenna (or single cell) for stable wireless communication in urban conditions with the complicated multipath propagation phenomena, caused by the multireflections, multidiffraction, multiscattering, etc., is in practice quite unrealistic.

This is why, the concept of cellular wireless communication has been introduced with numerous cells of a small radius, which provide a sufficient signal-to-noise ratio and a low level of fading, slow and fast, of the received signals within the communication channel. As an example, a characteristic cell layout plan for London, U.K. is presented in Fig. 12.32 according to Ref. 35, at an early stage of its implementation. As follows from this figure, the early strategy of cell communications design is based on the following principles:

With an increase of the number of subscribers, the dimensions of the cells become smaller (usually this was done for centers of cities, where the number of users is bigger and building density is higher).

Cells are arranged in *clusters*. Only clusters with a hexagonal shape are possible; the designed cluster sizes of 4, 7, and 12 cells are shown in Fig. 12.32;

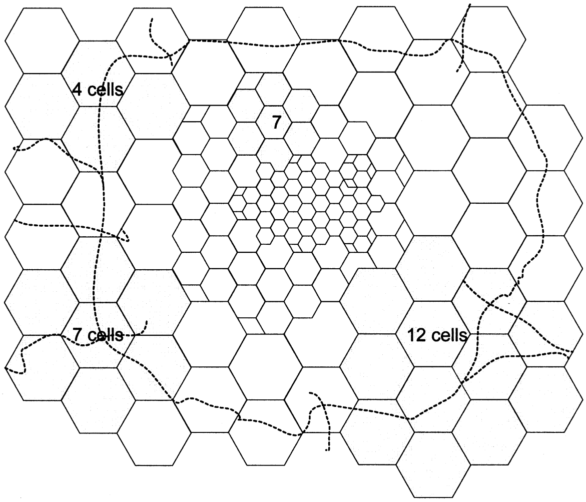


Figure 12.32 A typical city cellular map, where cluster sizes of 4, 7, and 12 are also indicated [35].

Cells are split. The installation of additional base stations within each cell depends on the degree of cell density in each cluster and on the coverage effect of each base station antenna.

Moreover, this proposed strategy of cells design has shown [5] that each base station antenna in such cells, with an effective power of 100 mW, covers a cell with radius $R = 1$ mil (~1.6 km). At the same time, to cover the area of one cell with radius $R = 10$ mil (~16 km), the transmitting antenna requires a power of 100 W, i.e., 1000 times higher. Thus the antenna power problem has been successfully solved by the use of a cell splitting strategy.

However the question regarding the regions of overlap of coverage between neighboring cells is not solved yet (as can be clearly seen from Fig. 12.31). The circle-shaped cell was therefore replaced by a regular hexagon-shaped cell. It is clearly seen from Fig. 12.33, where both circle-shaped and regular hexagon-shaped cells are presented, that the hexagon-shaped cell is more geometrically attractive than the circle-shaped cell. Moreover, in the hexagon-shaped multiple cells structure (plan), the hexagonal cells are closely covered by each other. Thus, each hexagonal cell can be packed into clusters "side to side" with neighboring cells. The size of such a hexagonal cell can be defined by use of its radius R and the angle of 120° (see Fig. 12.33).

12.6.2. Cell Design Strategy

Now we will describe the main characteristics of a cell and will show how to create the cell structure. The real distance from the center of a cell, where the base station is located ("based cell"), to the center of the "repeat cell," which is denoted in Fig. 12.34 by the same letter, is called the *reuse distance*, D; the cell size is determined by its *radius* R.

The cluster size is designated by the letter N and is determined by the equation [1–5,35,36]

$$N = i^2 + ij + j^2 \tag{12.117}$$

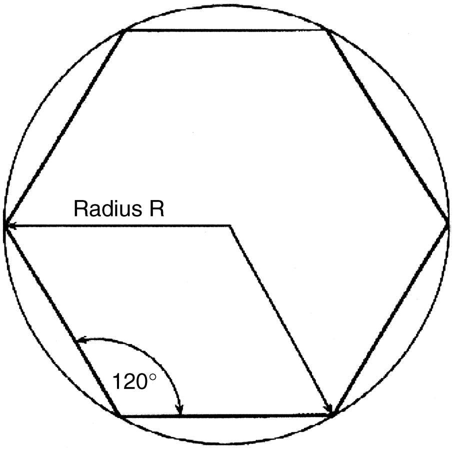


Figure 12.33 Circle-shaped and regular hexagon-shaped cells presentation according to [35].

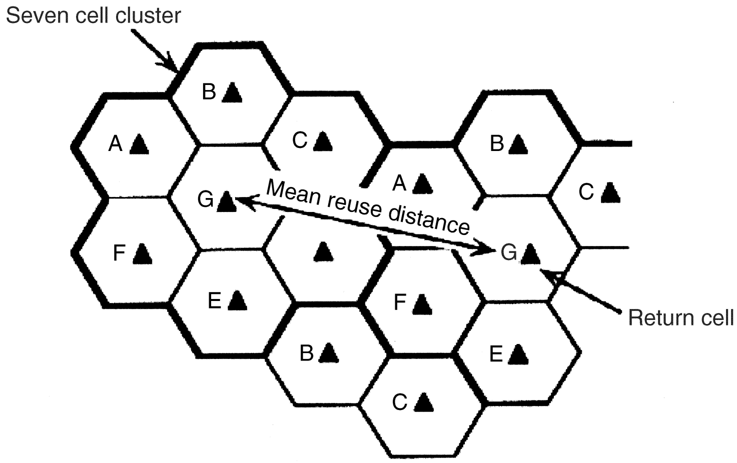
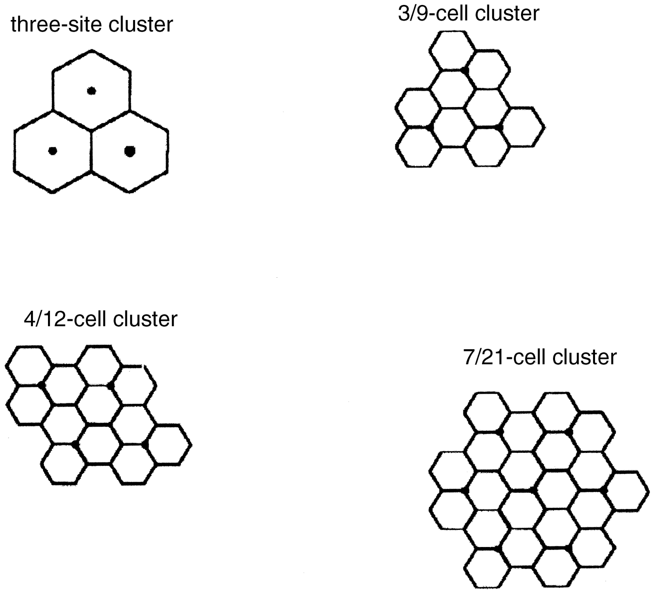


Figure 12.34 The popular 7-cell cluster arrangement. *D* is the reuse distance, and *R* is the cell's radius according to [35].

where $i,j = 0, 1, 2, \ldots$, etc. As follows from Eq. (12.117), only the cluster sizes 3, 4, 7, 9, 12, etc., are possible. However each cluster can be divided into 3 clusters each consisting of 3 cells. It is called a 3/9-cell cluster (see Fig. 12.35). Other variants of sectored clusters are presented in Fig. 12.35. As can be seen, each sector has one base station antenna (or radio port).

We donot enter into this subject deeply, but will only remark that it is necessary to divide clusters into sub-clusters because of the necessity to use the same repeating frequencies in different cells. In fact, if we focus on the popular 7-cell cluster arrangement, which is depicted in Fig. 12.34, we first notice that the allocation of frequencies into seven sets is required. In Fig. 12.34, the mean reuse distance, is illustrated, in which the cells (say, denoted by $G \leftrightarrow G$) use the *same frequency set*. This is a simple way to use the repeat frequency set in the other clusters.


three-site cluster

3/9-cell cluster

4/12-cell cluster

7/21-cell cluster

Figure 12.35 Different variants of sectored clusters according to [35].

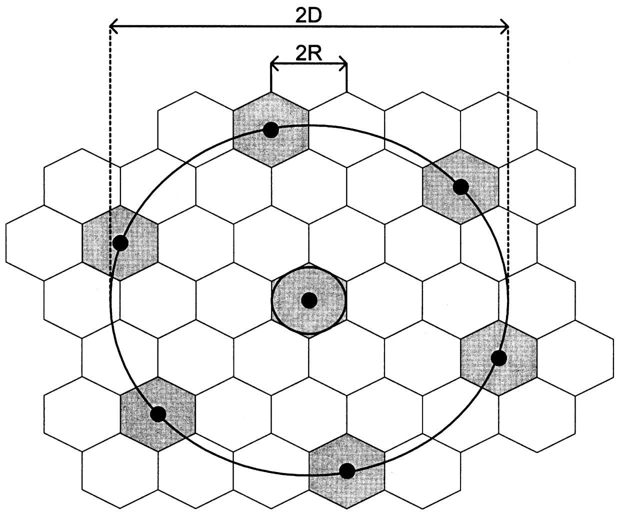


Figure 12.36 Directional frequency reuse plan according to [35].

Between D and the cell radius R (see Fig. 12.36) there exists a relationship, which is called the *reuse ratio*. This parameter for a hexagonal cell is a function of cluster size, [1–5,35,36], i.e.,

$$\frac{D}{R} = \sqrt{3N} \tag{12.118}$$

Example. For a 7-cell cluster of 2-mile-radius cells, the repeat cell centers, which operate with the same frequency set would be separated by

$$D = R\sqrt{3N} = 2\sqrt{21} \approx 9.2\,\text{mil}$$

Within other cells in a cluster, interference inside the communication channel can be expected at the same frequencies. Hence, for a 7-cell cluster there could be up to six immediate interferers, as is shown in Fig. 12.36.

12.6.3. Cochannel Interference Concept

Now we will discuss the question of how to predict the optimal cell size and the cluster splitting using the law of signal decay, described by many independent radio propagation models constructed to predict the propagation effects within various wireless communication channels. This question is related closely to another major problem of cochannel interference caused by frequent reuse of channels within the cellular communication system. To illustrate the cochannel interference concept, let us consider a pair of cells with radius R, separated by a reuse distance D, as shown in Fig. 12.36. Since the cochannel site is located far from the transmitter ($D \gg R$), which is located within the initial cell, its signal at the servicing site will suffer multipath attenuation. We consider here the situation in the built-up environments where both antennas are lower than the surrounding buildings' rooftops. To predict the degree of cochannel interference in such a situation with moving subscribers within the cellular system, a new parameter, "carrier-to-interference ratio," C/I, is introduced in the literature [2–9]. This parameter in turn depends on frequency planning and antenna engineering. As pointed out in Ref. 9, a cochannel interferer has the same nominal frequency as the desired frequency. It arises from the multiple use of the same frequency. Thus, referring to the part of cellular map depicted in Fig. 12.34, we find that cochannel sites are located in the second cluster. For omnidirectional antennas located inside each site, the theoretical cochannel interference in dB is given by [1–4,8]:

$$\frac{C}{I} = 10\log\left[\frac{1}{j}\left(\frac{D}{R}\right)^{\gamma}\right] \tag{12.119}$$

where j is the number of cochannel interferers ($j = 1, 2, \ldots, 6$), γ is the path loss slope constant, which determines the signal decay in various propagation environments. For a typical seven-cell cluster ($N = 7$) with one cell as basic (with the transmitter inside it) and with six other interferers ($j = 6$) as the cochannel sites in first tier (see Fig. 12.36), this parameter depends on conditions of wave propagation within the urban communication channel. To understand this fact, let us present, according to Ref. 1, a simple propagation model for the regular urban environment with $\gamma = 4$. In this case one can rewrite Eq. (12.119) as

$$\frac{C}{I} = 10\log\left[\frac{1}{6}\left(\frac{D}{R}\right)^{4}\right] \tag{12.120}$$

In this case, according to Eq. (12.120), $D/R = \sqrt{3N} = 4.58$, and $C/I = 18.6\,\text{dB}$.

In the general case, by introducing Eq. (12.118) in Eq. (12.120) we have that [1–4,8]

$$\frac{C}{I} = 10\log\left[\frac{1}{6}(3N)^2\right] = 10\log\left(1.5N^2\right) \tag{12.121}$$

i.e., C/I is also a function of cluster size N and is increased with increase of cells' number in each cluster or with decrease of cell radius R.

According to the propagation situation in urban scene the servicing and cochannel sites can lie both inside and outside the break point range r_B (see Fig. 12.37). If both of them are within this range, as follows from Fig. 12.38a, the cochannel interference parameter can be described instead [Eq. (12.119)] by the C/I-ratio prediction equation (in dB) as [1–4,8]

$$\frac{C}{I} = 10\log\left[\frac{1}{6}\left(\frac{D}{R}\right)^2\right] \tag{12.122}$$

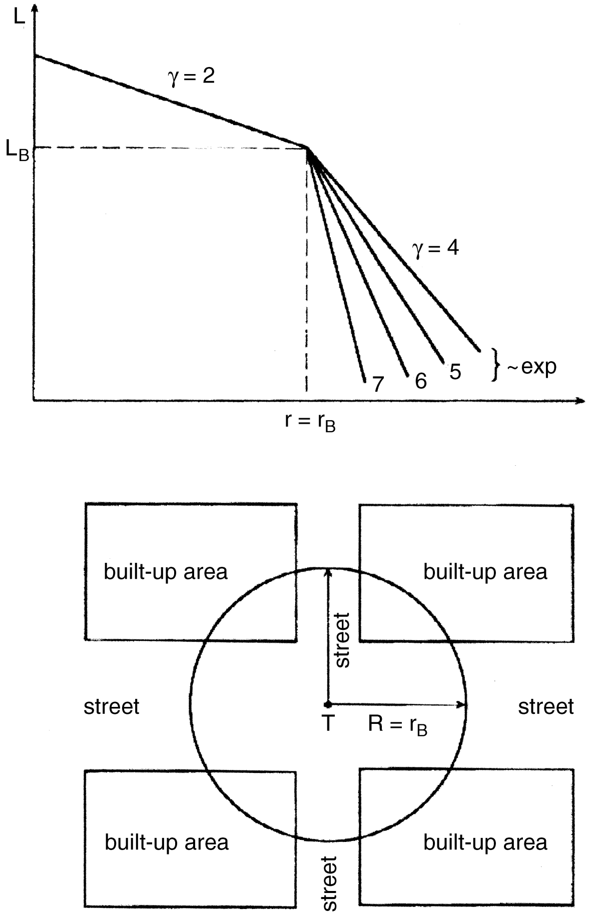
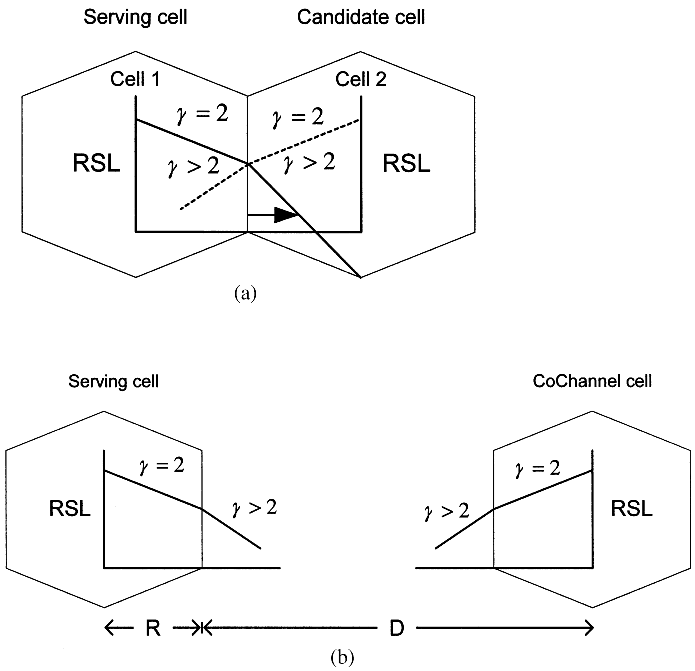


Figure 12.37 Cochannel interference evaluation scheme.



(a)

(b)

Figure 12.38 A cell in an urban area with grid-plan streets (a) and (b) according to [1,35,36].

For cell sites located beyond the break point range (see Fig. 12.38b) this equation can be modified taking into account the multipath phenomenon and obstructions which change the signal decay law from D^{-2} to $D^{-\gamma}$, $\gamma = 2 + \Delta\gamma, \Delta\gamma \geq 1$. Hence, we finally have instead of Eq. (12.118) [1]:

$$\frac{C}{I} = 10\log\left[\frac{1}{6}\left(\frac{D^{(2+\Delta\gamma)}}{R^2}\right)\right] \tag{12.123}$$

We can now rewrite Eq. (12.123) versus number of cells in cluster, N, and of radius of the individual cell, R, by use of Eq. (12.118):

$$\frac{C}{I} = 10\log\left[\frac{N}{2}(3N)^{\Delta\gamma/2}R^{\Delta\gamma}\right] \tag{12.124}$$

Let us examine this equation for two typical cases in the urban scene described above in Sec. 12.3.

City with Regularly Planned Streets

In this case for a typical straight wide avenue, for which according to street-multislit-waveguide model $\Delta\gamma = 2(\gamma = 4)$ (see Sec. 12.3) and

$$\frac{C}{I} = 10\log\left[\frac{3}{2}N^2R^2\right] \tag{12.125}$$

For the case of narrow streets (more realistic case in urban scene) one can put in Eq. (12.125) $\Delta\gamma = 3 - 7$ ($\gamma = 5 - 9$), which is close to the exponential signal decay that follows from the street waveguide model. In this case, the cell size R can be approximately described, using the multislit street waveguide model (see Sec. 12.3), as [1]

$$R_{\text{cell}} \equiv r_B = \frac{4h_T h_R}{\lambda}\frac{(1+\chi)}{(1-\chi)}\left(1+\frac{h_b}{a}+\frac{h_T h_R}{a^2}\right) \tag{12.126}$$

where all parameters in Eq. (12.126) are described in Sec. 12.3.

City with Nonregularly Planned Streets

For the case of propagation over irregular built-up terrain, as follows from the probabilistic approach, described in Sec. 12.3, $\Delta\gamma = 1$ and the C/I-ratio prediction equation is as follows [1]:

$$\frac{C}{I} = 10\log\left[\frac{N}{2}(3N)^{1/2}R\right] \tag{12.127}$$

As follows from this approach, the average distance of the direct visibility $\bar{\rho}$ between two arbitrary points, the source and the observer, is described by the following formula:

$$\bar{\rho} = (\gamma_0)^{-1}\,\text{km} \tag{12.128}$$

where all parameters are presented in Sec. 12.3.

As follows from Eqs. (12.125)–(12.128), the C/I ratio strongly depends on conditions of wave propagation within the urban communication channels (on path loss slope parameter $\gamma = 2 + \Delta\gamma,\ \Delta\gamma \geq 1$) and on the cellular map splitting strategy (on parameters N and R).

Let us now introduce the important celluar parameters and present them in Table 12.2. Here in column (A) the reuse ratio D/R is presented; number of channels per cell is presented in column (B). The data presented in column (C) is obtained by use of the standard presentation of formula Eq. (12.121), that is, $C_{\text{i}} = C/I = 1.5N^2$. To obtain

Table 12.2 Important Cellular Parameters in Column (a) the Reuse Ratio D/R; in Column (b) Number of Channels Per Cell; in Column (c) $C_i = C/I = 1.5 \cdot N^2$; in Column (d) the Number of Subscribers Per Cell [35]

Cluster size (N)	(A) Reuse ratio (D/R)	(B) Number of channels per cell (279/N)	(C) Cochannel ($C_i = C/I = 1.5N^2$) (dB)	(D) Number of subscribers per cell ($\tilde{n}$)
3	3	93	11	2583
4	3.5	69	14	1840
7	4.6	39	18	937
9	5.2	31	21	707
12	6	23	23	483
21	7.9	14	28	245

the number of subscribers per cell, described by column (D) in Table 12.2, we need additional information about the urban area and additional formulations, such as [35]

The urban area of operation and servicing: A, km^2
The number of citizens in the operating urban area: P (per thousands)
The mean radius of the cell: R, km
The number of channels in one cell: n_c

Thus, in Fig. 12.39 according to Ref. 35 the dependence of n_c versus cells' number N for various C/I ratio is shown. If, for example, 30 subscribers use the same channel in the considered cell, then the number of subscribers in this cell equals

$$\tilde{n} = 30\, n_c \approx 10\, \pi n_c \tag{12.129}$$

For regularly distributed cells over the built-up terrain, the number of cells in the urban area concerned equals

$$K = \frac{A}{\pi R^2} \tag{12.130}$$

Then the total number of subscribers in the urban area considered equals

$$\tilde{N} = \tilde{n} K = \frac{10\, n_c A}{R^2} \tag{12.131}$$

The parameter $\tilde{N}$ calculated by use of 1this formula is presented in column (D) in Table 12.2. From Eq. (12.131) we can estimate as a percentage the number of subscribers from the population located in the urban area. In fact,

$$\tilde{N}(\%) = \frac{10\, n_c A}{R^2 P(1000)} 100\% = \frac{A n_c}{P R^2} (\%) \tag{12.132}$$

Example. The number of citizens is 600,000 located within an area with radius $R_a = 8$ km. The cell size is $R = 2$ km, the number of channels in each cell is $n_c = 40$. Find the number of subscribers (in %) for effective servicing by wireless communication system.

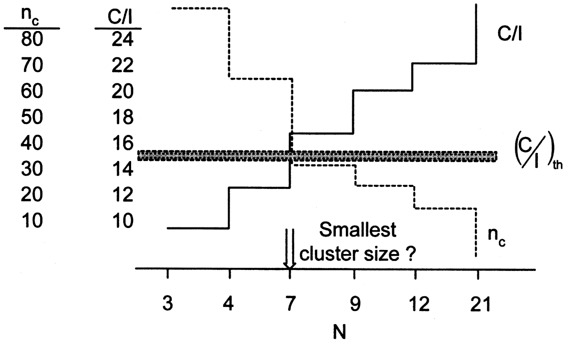


Figure 12.39 Number of channels per cell n_c versus cells number N for various C/I ratio.

Solution.

First step. We calculate a city area: $A = \pi R_a^2 \approx 200\,\text{km}^2$.

Second step. We calculate the number of citizen per thousands: $P = 600{,}000/1000 = 600$.

Third step. We calculate, using Eq. (12.132) in first tier, $\tilde{N}(\%) = 200(40)/600(4) \approx 3.3\%$.

The result of this example shows that for cities with a high density of population (A/P is small), it is very hard to plan the wireless service by use of a simple propagation model. If we try to increase (up to the maximum) the number of channels by splitting the operating radio frequency band, the cell size (radius R) remains critically limited by the conditions of radio wave propagation in the urban area.

We note that the formula Eq. (12.132) can be rewritten by introducing into it new parameters as the frequency band of the total cellular service system, ΔF, and the frequency band of each channel, Δf_c. In this case the number of radio channels in each cell equals [35,36]

$$n_c = \frac{\Delta F}{\Delta f_c N} \tag{12.133}$$

Then the number of subscribers, which can effectively communicate by using existing cellular cervicing system equals, as a percentage,

$$\tilde{N}(\%) = \frac{A \Delta F}{\Delta f_c P N R^2} \tag{12.134}$$

Equations (12.132) and (12.134) show that to increase the efficiency of the cellular communication system in various urban environments, an effective frequency splitting strategy over the channels within each cell is required. Moreover, by a decrease of the cell size and the cluster size (or number N) one can also increase the efficiency of the cellular system. The latter depends on the strategy of cellular map construction and splitting. However, the experience of cellular systems designers shows that it is very difficult to decrease the number N of cells in each cluster (see Fig. 12.39 according to Ref. 35). Apparently, as follows from this picture, number $N = 7$ is the smallest size of cluster constructed, because for $N < 7$ the acceptable C/I level of 16 dB cannot be reached. Initially the parameter N was selected as $N = 12$ by the TACS cellular system constructed in England (more detailed information is presented in Refs. 2–9). However, while analyzing the C/I ratio and its optimization, the optimal number $N = 7$ was found. In fact, as follows from Fig. 12.39, for 300 working radio channels with 21 channels required to control the total cellular system, we obtain for $N = 12$ and $N = 7$, respectively, $n_c = 39$ and $n_c = 23$ communication channels in each cell. This result follows from Table 12.2 and Fig. 12.39, where value 23 from first column in table lies between 20 and 30 (the level corresponding to $N = 7$) and value 39 from this column lies between 30 and 40 (the level corresponding to $N = 12$).

REFERENCES

1. Blaunstein, N. *Radio Propagation in Cellular Networks*; Artech Houses: Boston–London, 1999, p. 386.
2. Parsons, L.D. *The Mobile Radio Propagation Channels*; Pentech Press: New York–Toronto, 1992, p. 313.

3. Lee, W.Y.C. *Mobile Communication Design Fundamentals*; McGraw Hill: New York, 1993, p. 365.
4. Yacoub, M.D. *Foundations of Mobile Radio Engineering*; CRC Press: NY, 1993, p. 290.
5. Saunders, S.R. *Antennas and Propagation for Wireless Communication Systems*; Wiley: New York, 1999, p. 409.
6. Bertoni, H.L. *Radio Propagation for Modern Wireless Systems*; Prentice Hall PTR: New Jersey, 2000, p. 258.
7. Feuerstein, M.L.; Rappaport, T.S. *Wireless Personal Communication*; Artech House: Boston–London, 1992, p. 315.
8. Rappaport, T.S. *Wireless Communications*; Prentice Hall PTR: New York, 1996, p. 641.
9. Blaunstein, N.; Jorgen Bach Andersen, *Multipath Phenomena in Cellular Networks*; Artech Houses: Boston–London, 2002, p. 296.
10. Jakes, W.C. *Microwave Mobile Communications*; Wiley: New York, 1974, p. 642.
11. Steele, R. *Mobile Radio Communication*; IEEE Press: New York, 1992, p. 779.
12. Proakis, J.G. *Digital Communications*; McGraw Hill: New York, 1995.
13. Balanis, C.A. *Advanced Engineering Electromagnetics*; Wiley: New York, 1997, p. 941.
14. Kraus, J.D. *Antennas*, 2nd Ed.; McGraw-Hill: New York, 1988.
15. Siwiak, K. Radiowave *Propagation and Antennas for Personal Communications*, 2nd Ed.; Artech House: Boston–London, 1998.
16. Vaughan, R.; Bach Andersen, J. *Channels, Propagation, and Antennas for Mobile Communications*; IEEE: London, 2002.
17. Milstein, L.B.; Schilling, D.L.; Pickholtz, R.L. On the feasibility of a CDMA overlay for personal communications networks. IEEE Select. Areas Commun. **May 1992**, *10*(4), 665–668.
18. Rustako, A.J., Jr.; Amitay, N.; Owens, M.J.; Roman, R.S. Radio propagation at microwave frequencies for line-of-sight microcellular mobile and personal communications. IEEE Trans. Veh. Technol. **Feb. 1991**, *40*(2), 203–210.
19. Tan, S.Y.; Tan, H.S. UTD propagation model in an urban street scene for microcellular communications. IEEE Trans. Electromag. Compat. **1993**, *35*(4), 423–428.
20. Blaunstein, N., and M. Levin, VHF/UHF wave attenuation in a city with regularly spaced buildings. Radio Sci. **1996**, *31*(2), 313–323.
21. Blaunstein, N.; Levin, M. Propagation loss prediction in the urban environment with rectangular grid-plan streets. Radio Sci. **1997**, *32*(2), 453–467.
22. Blaunstein, N.; Giladi, R.; Levin, M. Los characteristics' prediction in urban and suburban environments. IEEE Trans. on Vehic. Tech. **1998**, *47*(1), 11–21.
23. Blaunstein, N. Average field attenuation in the nonregular impedance street waveguide. IEEE Trans. Anten. Propagat. **1998**, *46*(12), 1782–1789.
24. Xia, H.H.; Bertoni, H.L.; Maciel, L.R.; Honcharenko, W. Radio propagation characteristics for line-of-sight microcellular and personal communications. IEEE Trans. Anten. Propag. **Oct. 1993**, *41*(10), 1439–1447.
25. Bertoni, H.L.; Honcharenko, W.; Maciel, L.R.; Xia, H.H. UHF propagation prediction for wireless personal communications. Proc. IEEE. **Sept. 1994**, *82*(9), 1333–1359.
26. Okumura, Y.; Ohmori, E.; Kawano, T.; Fukuda, K. Field strength and its variability in the VHF and UHF land mobile radio service. Review Elec. Commun. Lab. **1968**, *16*, 825–843.
27. Hata, M. Empirical formula for propagation loss in land mobile radio services. IEEE Trans. Veh. Technol. **1980**, *VT-29*, 317–325.
28. Saleh Faruque, *Cellular Mobile Systems Engineering*; Artech House: Boston–London, 1994.
29. Ponomarev, G.A.; Kulikov, A.N.; Telpukhovsky, E.D. *Propagation of Ultra-Short Waves in Urban Environments*; Tomsk: Rasko, Russia, 1991.
30. Blaunstein, N. Prediction of cellular characteristics for various urban environments. IEEE Anten. Propagat. Magazine. **1999**, *41*(6), 135–145.
31. Blaunstein, N.; Katz, D.; Censor, D.; Freedman, A.; Matityahu, I.; Gur-Arie, I. Prediction of loss characteristics in built-up areas with various buildings' overlay profiles. IEEE Anten. Propagat. Magazine. **2001**, *43*(6) 181–191.

32. Clarke, R.H. A statistical theory of mobile-radio reception. Bell Systems Tech. J. **1968**, *47*, 957–1000.
33. Aulin, T. A modified model for the fading signal at a mobile radio channel. IEEE Trans. Veh. Technol. **1979**, *28*(3), 182–203.
34. Suzuki, H. A statistical model for urban propagation. IEEE Trans. Communication **1977**, *25*, 673–680.
35. Mehrotra, A. *Cellular Radio Performance Engineering*; Artech House: Boston–London, 1994, p. 249.
36. Linnartz, J.P. *Narrowband Land-Mobile Radio Networks*; Artech House: Boston–London, 1993, p. 335.

13
Satellite Communication Systems

Matthew N. O. Sadiku
Prairie View A&M University
Prairie View, Texas, U.S.A.

Satellite-based communication has become a major facet of the telecommunication industry for two major reasons. First, it provides a means of broadcasting information to a large number of people simultaneously. Thus, satellite communication systems are an important ingredient in the implementation of a global communication infrastructure. Second, satellite communication provides a means of reaching isolated places on earth, where terrestrial telecommunications infrastructure does not exist or teledensity is low.

Satellite communication was first deployed in the 1960s and has its roots in military applications. Since the launch of the Early Bird satellite (first commercial communication satellite also known as Intelsat I) by NASA in 1965 proved the effectiveness of satellite communication, satellites have played an important role in both domestic and international communications networks. They have brought voice, video, and data communications to areas of the world that are not accessible with terrestrial lines. By extending communications to the remotest parts of the world, virtually everyone can be part of the global economy.

Satellite communications is not a replacement of the existing terrestrial systems but rather an extension of the wireless system. However, satellite communication has the following merits over terrestrial communications:

Coverage: Satellites can cover a much large geographical area than the traditional ground-based systems. They have the unique ability to cover the globe.

High bandwidth: A Ka-band (27–40 GHz) system can deliver a throughput of gigabits per second rate.

Low cost: A satellite communications system is relatively inexpensive because there are no cable-laying costs and one satellite covers a large area.

Wireless communication: Users can enjoy untethered mobile communication anywhere within the satellite coverage area.

Simple topology: Satellite networks have simpler topology, which results in more manageable network performance.

Broadcast/multicast: Satellite are naturally attractive for broadcast/multicast applications.

Maintenance: A typical satellite is designed to be unattended, requiring only minimal attention by customer personnel.

Immunity: A satellite system will not suffer from disasters such as floods, fire, and earthquakes and will, therefore, be available as an emergency service should terrestrial services be knocked out.

Table 13.1 Advantages and Disadvantages of Satellite Communication [1]

Advantages	Disadvantages
Wide-area coverage	Propagation delay.
Easy access to remote sites	Dependency on a remote facility.
Costs independent of distance	Less control over transmission.
Low error rates	Attenuation due to atmospheric particles (e.g., rain) can be severe at high frequencies.
Adaptable to changing network patterns	Continual time-of-use charges.
No right-of-way necessary, earth stations located at premises	Reduced transmission during solar equinox.

Of course, satellites systems do have some disadvantages. These are weighed with their advantages in Table 13.1. Some of the services provided by satellites include fixed satellite service (FSS), mobile satellite service (MSS), broadcasting satellite service (BSS), navigational satellite service, and meteorological satellite service.

This chapter explores the integration of satellites with terrestrial networks to meet the demands of highly mobile communities. After looking at the fundamentals of satellite communications, we will discuss the orbital and propagation characteristics and the various applications of satellite-based communications systems.

13.1. FUNDAMENTALS

A satellite communication system may be viewed as consisting of two parts: the space and ground segments. The space segment consists of the satellites and all their on-board tracking and control systems. The earth segment comprises the earth terminals, their associated equipment, and the links to terrestrial networks [2].

13.1.1. Types of Satellites

There were only 150 satellites in orbit by September 1997. The number was expected to be roughly 1700 by the year 2002. With this increasing trend in the number of satellites, there is a need to categorize them. According to the height of their orbit and "footprint" or coverage on the earth's surface, they are classified as follows [3].

Geostationary Earth Orbit (GEO) Satellites

They are launched into a geostationary or geosynchronous orbit, which is 35,786 km above the equator. (Raising a satellite to such an altitude, however, required a rocket, so that the achievement of a GEO satellite did not take place until 1963.) A satellite is said to be in geostationary orbit when the space satellite is matched to the rotation of the earth at the equator. A GEO satellite can cover nearly one-third of the earth's surface, i.e., it takes three GEO satellites to provide global coverage. Due to their large coverage, GEO satellites are ideal for broadcasting and international communications. [GEO is sometimes referred to as *high earth orbit* (HEO).] Examples of GEO satellite constellations are Spaceway designed by Boeing Satellite Systems and Astrolink by Lockheed Martin. Another example is Thuraya, designed by Boeing

Satellite Systems to provide mobile satellite services to the Middle East and surrounding areas.

There are at least three major objections to GEO satellites [4]. First, there is a relatively long propagation delay (or latency) between the instant a signal is transmitted and when it returns to earth (about 240 ms). This is caused by speed-of-light transmission delay and signal processing delay. This may not be a problem if the signal is going only one way. However, for signals such as data and voice, which go in both directions, the delay can cause problems. GEO satellites, therefore, are less attractive for voice communication. Second, there is lack of coverage at far northern and southern latitudes. This is unavoidable because a GEO satellite is below the horizon and may not provide coverage at latitudes as close to the equator as 45°. Unfortunately, many of the European capitals, including London, Paris, Berlin, Warsaw, and Moscow, are north of this latitude. Third, both the mobile unit and the satellite of a GEO system require a high transmit power. In spite of these objections, the majority of satellites in operation today are GEO satellites but that may change in the near future.

Middle Earth Orbit (MEO) Satellites

They orbit the earth at 5,000 to 12,000 km. GEO satellites do not provide good coverage for places far north and satellites in inclined elliptical orbits are an alternative. Although the lower orbit reduces propagation delay to only 60 to 140 ms round trip, it takes 12 MEO satellites to cover most of the planet. MEO systems represent a compromise between LEO (see below) and GEO systems, balancing the advantages and disadvantages of each. [MEO is sometimes referred to as *intermediate circular orbit* (ICO).]

Low Earth Orbit (LEO) Satellites

They circle the earth at 500 to 3000 km. For example, the Echo satellite circled the earth every 90 min. To provide global coverage may require as many as 200 LEO satellites. Latency in a LEO system is comparable with terrestrial fiber optics, usually less than 30 ms round trip. LEO satellites are suitable for personal communication systems (PCS). However, LEO systems have a shorter life space of 5–8 years (compared with 12–15 years for GEO systems) due to the increased amount of radiation in low earth orbit. The LEO systems have been grouped as Little LEO and Big LEO. The Little LEOs have less capacity and are limited to nonvoice services such as data and message transmission. An example is OrbComm designed by Orbital Corporation, which consists of 36 satellites, each weighing 85 lb. The Big LEOs have larger capacity and voice transmission capability. An example is Loral and Qualcomm's Globalstar which will operate in the L-band frequencies and employ 48 satellites organized in eight planes of six satellites each.

The arrangement of the three basic types of satellites is shown in Fig. 13.1. The evolution from GEO to MEO and LEO satellites has resulted in a variety of global satellite systems. The convenience of GEO was weighed against the practical difficulty involved with it and the inherent technical advantages of LEO, such as lower delay and higher angles of elevation. While it has been conceded that GEO is in many respects theoretically preferable, LEO or MEO systems would be preferred for many applications. Although a constellation (a group of satellites) is required instead of only one for hemispheric coverage, the loss of individual satellites would cause only gradual degradation of the system rather than a catastrophic failure. A comparison of the three satellite types is given in Table 13.2.

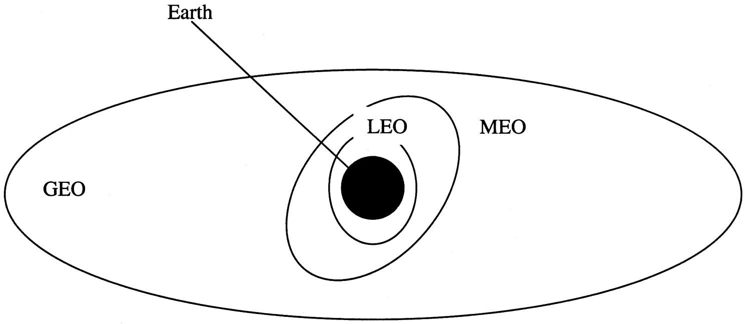


Figure 13.1 The three common types of satellites: GEO, MEO, and LEO.

Table 13.2 Comparison of the GEO, MEO, and LEO [5]

Type	Altitude	Coverage	Advantages	Disadvantages
LEO	300–1000 km	Spot	Low path loss High data rate Low delay Low launch cost Less fuel	Coverage is less. Need many satellites. Short orbital life. High Doppler. Highly complex.
MEO	1000–10,000 km	Region	Moderate path loss Moderate launch cost Less fuel	Multiple satellites. Moderate coverage. Highly complex.
GEO	36,000 km	Earth	Global coverage Need few satellites Long orbital life Low Doppler Less complex	High path loss. Long delay. Low data rate. High launch cost. Fuel for station keeping.

13.1.2. Frequency Bands

Every nation has the right to access the satellite orbit and no nation has a permanent right or priority to use any particular orbit location. Without a means for the nations to coordinate the use of satellite frequency bands, the satellite services of one nation could interfere with those of another, thereby creating a chaotic situation in which neither country's signals could be received clearly.

To facilitate satellite communications and eliminate interference between different systems, international organizations govern the use of satellite frequency. The International Telecommunication Union (ITU) is responsible for allocating frequencies to satellite services. Since the spectrum is a limited resource, the ITU has reassigned the same parts of the spectrum to many nations and for many purposes throughout the world.

The frequency spectrum allocations for satellite services are given in Table 13.3. Notice that the assigned segment is the 1–40 GHz frequency range, which is the microwave portion of the spectrum. As microwaves, the signals between the satellite and the earth stations travel along line-of-sight paths and experience free-space loss that increases as the square of the distance.

Satellite services are classified into 17 categories [6]: fixed, intersatellite, mobile, land mobile, maritime mobile, aeronautical mobile, broadcasting, earth exploration, space research, meteorological, space operation, amateur, radiodetermination, radionavigation, maritime radionavigation, and standard frequency and time signal. The Ku band is presently used for broadcasting services, and also for certain fixed satellite services. The C band is exclusively used for fixed satellite services, and no broadcasting is allowed. The L band is employed by mobile satellite services and navigation systems.

A satellite band is divided into separation portions: one for earth-to-space links (the uplink) and one for space-to-earth links (the downlink). Like a terrestrial microwave relay, a satellite must use separate frequencies for sending to the satellite (the uplink) and receiving from the satellite (the downlink); otherwise, the powerful signal transmitted by the satellite would interfere with the weak incoming signal. Table 13.4 provides the general frequency assignments for uplink and downlink satellite frequencies. We notice from the table that the uplink frequency bands are slightly higher than the corresponding downlink frequency band. This is to take advantage of the fact that it is easier to generate higher frequency RF power within a ground station than it is onboard a satellite. In order to direct the uplink transmission to a specific satellite, the uplink radio beams are highly focused. In the same way, the downlink transmission is focused on a particular *footprint* or area of coverage.

All satellite systems are constrained to operate in designed frequency bands depending on the kind of earth station used and service provided. The satellite industry, particularly in the United States, is subject to several regulatory requirements,

Table 13.3 Satellite Frequency Allocations

Frequency band	Range (GHz)
L	1–2
S	2–4
C	4–8
X	8–12
Ku	12–18
K	18–27
Ka	27–40

Table 13.4 Typical Uplink and Downlink Satellite Frequencies

Uplink frequencies (GHz)	Downlink frequencies (GHz)
5.925–6.426	3.700–4.200
7.900–8.401	7.250–7.750
14.00–14.51	11.70–12.20
27.50–31.0	17.70–20.20

domestically and internationally, depending upon which radio services and frequency bands are proposed to be used on the satellite. In the United States, the Federal Communications Commission (FCC) is the independent regulatory agency that ensures that the limited orbital or spectrum resource allocated to space radiocommunications services is used efficiently. After receiving an application for a U.S. domestic satellite, FCC initiates the advance publication process for a U.S. satellite. This is to ensure the availability of an orbit position when the satellite is authorized. FCC does not guarantee international recognition and protection of satellite systems unless the authorized satellite operator complies with all coordination requirements and completes the necessary coordination of its satellites with all other administrations whose satellites are affected [7].

13.1.3. Basic Satellite Components

Every satellite communication involves the transmission of information from a ground station to the satellite (the uplink), followed by a retransmission of the information from the satellite back to the earth (the downlink). Hence the satellite system must typically have a receiver antenna, a receiver, a transmitter antenna, a transmitter, some mechanism for connecting the uplink with the downlink, and a power source to run the electronic system. These components are illustrated in Fig. 13.2 and explained as follows [8,9]:

Transmitters: The amount of power required by a satellite transmitter to send out depends on whether it is GEO or LEO satellite. The GEO satellite is about 100 times farther away than the LEO satellite. Thus, a GEO would need 10,000 times as much power as a LEO satellite. Fortunately, other parameters can be adjusted to reduce this amount of power.

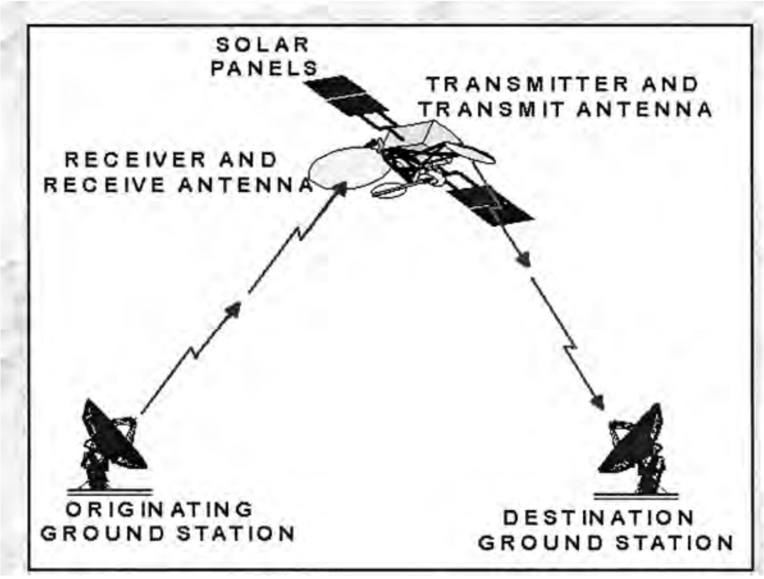

SOLAR PANELS
TRANSMITTER AND TRANSMIT ANTENNA
RECEIVER AND RECEIVE ANTENNA
ORIGINATING GROUND STATION
DESTINATION GROUND STATION

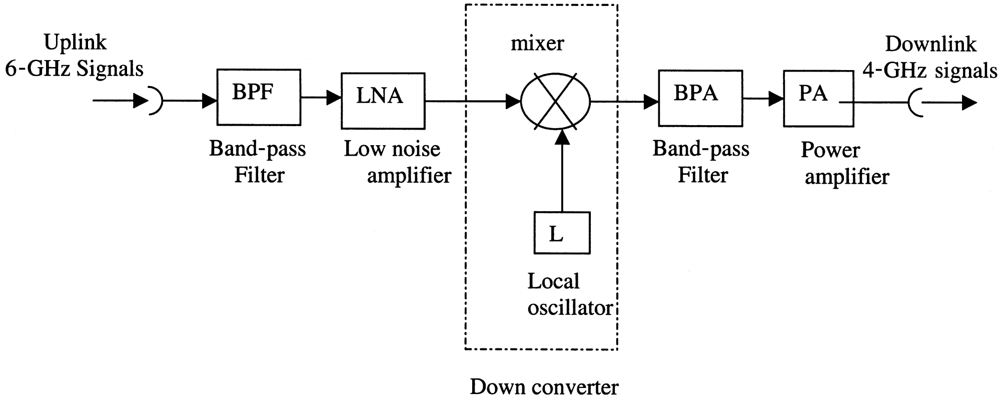
Figure 13.2 Basic components of a communication satellite link (with permission of Regis Leonard, NASA Lewis Research Center).

Antennas: The antennas dominate the appearance of a communication satellite. Antenna design is one of the more difficult and challenging parts of a communication satellite project. The antenna geometry is constrained physically by the design and the satellite topology. A major difference between GEO and LEO satellites is their antennas. Since all the receivers are located in the coverage area, which is relatively small, a properly designed antenna can focus most of the transmitter power within that area. The easiest way to achieve this is to simply make the antenna larger. This is one of the ways the GEO satellite makes up for the apparently larger transmitter power it requires.

Power generation: The satellite must generate all of its own power. The power is often generated by large solar cells, which convert sunlight into electricity. Since there is a limit to how large the solar panel can be, there is also a practical limit to the amount of power that can be generated. Satellites must also be prepared for periods of eclipse, when the earth is between the sun and the satellite. This necessitates having batteries on board that can supply power during eclipse and recharge later.

Transponders: These are the communication devices each satellite must carry. A transponder is a piece of equipment that receives a weak signal at one frequency, amplifies it, and changes its frequency to another for transmission to another earth station. The block diagram of a typical transponder is shown in Fig. 13.3. For example, a GEO satellite may have 24 transponders with each assigned a pair of frequencies (uplink and downlink frequencies).

Ground Stations: The ground (or earth) stations form the ground segment of the satellite communication system. The ground station is responsible for interacting and communicating with the satellites. Most ground or earth stations simply transmit and receive signals with a fixed antenna. At least one ground station must perform the task of controlling and monitoring the satellite. In a transmitting ground station, the information signal (voice, video, or data) is processed, amplified, and transmitted. In a receiving ground station, the reverse process takes place. In the past, ground stations were massive, expensive, and owned by common carriers and the military. Now, earth stations are small, less expensive, and owned or leased by



Figuer. 13.3 A simplified block diagram of a typical transponder.

private organizations. The antenna is a vital component of the ground station, and its size varies considerably from a 1-m diameter parabolic reflector used for TV programs at home to a 64-m diameter reflector used in the deep space network.

Telemetry and Control: This is partly implemented by the ground stations and partly by the satellite. Telemetry, tracking, and command (TTC) are used to monitor and control the satellite while in orbit. Telemetry is the means by which a measurement made is transmitted to an observer at a distance. Tracking is collecting data to monitor the movement of an object. Command is the process of establishing and maintaining control. Tracking and commands are used by the terrestrial control station to determine the position of the satellite and predict its future location. The satellite contains telemetry instrumentation that continuously gathers information and transmits it to the ground station.

13.2. ORBITAL CHARACTERISTICS

Since a satellite is a spacecraft that orbits the earth, an intuitive question to ask is "What keeps objects in orbit?" The answer to the question is found in the orbital mechanical laws governing satellite motion. Satellite orbits are essentially elliptical and obey the same laws of Johannes Kepler that govern the motion of planets around the sun. The three Kepler's laws are stated as follows [10]:

First law: The orbit of each planet follows an elliptical path in space with the sun serving as the focus.

Second law: The line linking a planet with the sun sweeps out equal areas in equal time.

Third law: The square of the period of a planet is proportional to the cube of its mean distance from the sun.

Besides these laws, Newton's law of gravitation states that any two bodies attract each other with a force proportional to the product of their masses and inversely proportional to the square of the distance between them, i.e.,

$$\mathbf{F} = -\frac{GMm}{r^2}\mathbf{a}_r \tag{13.1}$$

where M is the mass of one body (earth), m is the mass of the other body (satellite), $\mathbf{F}$ is the force on m due to M, r is the distance between the two bodies, $\mathbf{a}_r = \mathbf{r}/r$ is a unit vector along the displacement vector $\mathbf{r}$, and $G = 6.672 \times 10^{-11}\,\mathrm{Nm/kg^2}$ is the universal gravitational constant. If M is the mass of the earth, the product $GM = \mu = 3.99 \times 10^{14}\,\mathrm{m^3/s^2}$ is known as Kepler's constant.

Kepler's laws in conjunction with Newton's laws can be used to completely describe the motion of the planets around the sun or that of the satellite around the earth. Newton's second law can be written as

$$\mathbf{F} = m\frac{d^2r}{dt^2}\mathbf{a}_r \tag{13.2}$$

Equating this with the force between the earth and the satellite in Eq. (1) gives

$$\frac{d^2r}{dt^2}\mathbf{a}_r = -\frac{\mu}{r^2}\mathbf{a}_r \tag{13.3}$$

or

$$\ddot{\mathbf{r}} + \frac{\mu}{r^3}\mathbf{r} = 0 \tag{13.4}$$

where $\ddot{\mathbf{r}}$ is the vector acceleration. The solution to the vector second-order differential Eq. (13.4) is not simple but it can be shown that the resulting trajectory is in the form of an ellipse given by [11,12]

$$r = \frac{p}{1 + e\cos\theta} \tag{13.5}$$

where r is the distance between the geocenter and any point on the trajectory, p is a geometric constant, $e\,(0 \leq e < 1)$ is the eccentricity of the ellipse, and θ (known as the *true anomaly*) is the polar angle between r and the point on the ellipse nearest to the focus. These orbital parameters are illustrated in Fig. 13.4. The eccentricity e is given by

$$e = \sqrt{1 - \left(\frac{a}{b}\right)^2} \tag{13.6}$$

The point on the orbit where the satellite is closest to the earth is known as the *perigee*, while the point where the satellite is farthest from the earth is known as the *apogee*. The fact that the orbit is an ellipse confirms Kepler's first law. If a and b are the semimajor and semiminor axes (see Fig. 13.4), then

$$b = a\sqrt{(1 - e^2)} \tag{13.7a}$$

$$p = a(1 - e^2) \tag{13.7b}$$

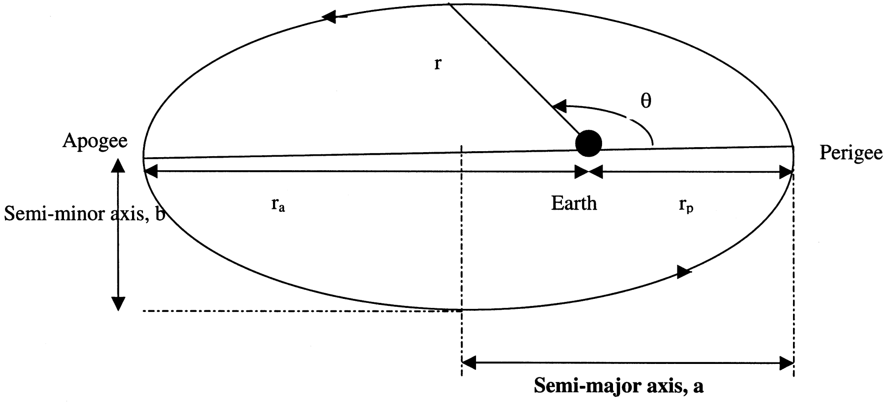


Figure 13.4 Orbital parameters.

Thus, the distance between a satellite and the geocenter is given by

$$r = \frac{a(1 - e^2)}{1 + e\cos\theta} \tag{13.8}$$

Note that the orbit becomes circular orbit when $e = 0$.

The apogee height and perigee height are often required. From the geometry of the ellipse, the magnitudes of the radius vectors at apogee and perigee can be obtained as

$$r_a = a(1 + e) \tag{13.9}$$

$$r_p = a(1 - e) \tag{13.10}$$

To find the apogee and perigee heights, the radius of the earth must be subtracted from the radii lengths.

The period T of a satellite is related to its semimajor axis a using Kepler's third law as

$$T = 2\pi\sqrt{\frac{a^3}{\mu}} \tag{13.11}$$

For a circular orbit to have a period equal to that of the earth's rotation (a sidereal day 23 h, 56 min, 4.09 s), an altitude of 35,803 km is required. In this equatorial plane, the satellite is "geostationary."

The velocity of a satellite in an elliptic orbit is obtained as

$$v^2 = \mu\left(\frac{2}{r} - \frac{1}{a}\right) \tag{13.12}$$

For a synchronous orbit ($T = 24$ h), $r = a = 42{,}230$ km, and $v = 3074$ m/s or 11,070 km/h. The closer the satellite is to the earth, the stronger is the effect of gravity, which constantly pulls it toward the earth, and so the greater must be the speed of the satellite to avoid falling to the earth.

A constellation is a group of satellites. The total number N of satellites in a constellation depends on the earth central angle γ and is given by [13]

$$N \approx \frac{4\sqrt{3}}{9}\left(\frac{\pi}{\gamma}\right)^2 \tag{13.13}$$

13.3. PROPAGATION CHARACTERISTICS

There are two major effects space has on satellite communications. First, the space environment, with radiation, rain, and space debris, is harsh on satellites. The satellite payload, which is responsible for the satellite communication functions, is expected to be simple and robust. Traditional satellites, specially GEOs, serve as "bent pipes" and act as repeaters between communication points on the ground, as shown in Fig. 13.5a.

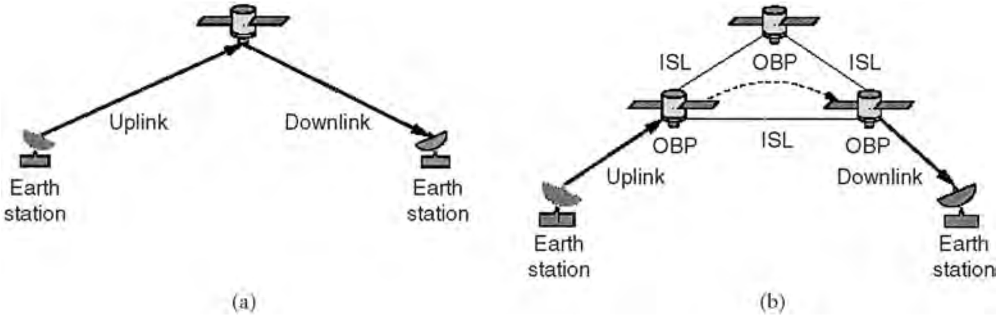


Figure 13.5 Satellite configuration types: (a) bent pipe and (b) onboard processing (OBP) switching and routing.

There is no onboard processing (OBP). However, new satellites allow OBP, including decoding/recoding, demodulation/remodulation, transponder, beam switching, and routing [14], as in Fig. 13.5b where a network of satellites is connected by intersatellite links (ISL).

The second effect is that of wave propagation. Attenuation due to atmospheric particles (rain, ice, dust, snow, fog, etc.) is not significant at L, S, and C bands. Above 10 GHz, the main propagation effects are [15,16]

Tropospheric propagation effects: attenuation by rain and clouds, scintillation, and depolarization

Effects of the environment on mobile terminals: shadowing, blockage, and multipath caused by objects in the surrounding of the terminal antenna

The troposphere can produce significant signal degradation at the Ku, Ka, and V bands, particularly at lower elevation angles. Most satellite systems are expected to operate at an elevation angle above roughly 20°. Rain constitutes the most fundamental obstacle encountered in the design of satellite communication systems at frequencies above 10 GHz. The resultant loss of signal power makes for unreliable transmission. Based on empirical data, the specific attenuation per unit α (dB/km) is related to rain intensity or rain rate R (in mm/h) as

$$\alpha = aR^b \tag{13.14}$$

where a and b are frequency dependent coefficients. The approximate expressions for $a(f)$ and $b(f)$ given in Table 13.5 are suitable for engineering purposes. The total attenuation loss in dB is given by

$$A = \alpha L_{\text{eq}} \tag{13.15}$$

where L_{eq} is the equivalent length, which is determined by the height of the freezing level and it depends on the rain rate R and the elevation angle θ. It is given by

$$L_{\text{eq}} = \left[7.413 \times 10^{-3} R^{0.766} + (0.232 - 1.803 \times 10^{-4} R) \sin\theta\right]^{-1} \tag{13.16}$$

Table 13.5 Attenuation Coefficients [17]

Frequency f (GHz)	a	b
8.5–25	$4.21 \times 10^{-5}(f)^{2.42}$	$1.41(f)^{-0.0779}$
25–54	$4.21 \times 10^{-5}(f)^{2.42}$	$2.63(f)^{-0.272}$
54–100	$4.09 \times 10^{-2}(f)^{0.699}$	$2.63(f)^{-0.272}$

The rain rate R is given by Rice-Holmberg model. The percent of an average year for which the rain rate exceeds R at a medium location is given by the Rice-Holmberg distribution

$$P(R) = ae^{-0.03R} + be^{-0.258R} + ce^{-1.63R} \tag{13.17}$$

where

$$a = \frac{M\beta}{2922} \qquad b = M(1-\beta)/438.3 \qquad c = 1.86\beta \tag{13.18}$$

M is the total mean yearly rainfall in millimeters and β is the ratio of thunderstorm rain accumulation to total accumulation. Attenuation due to other hydrometeors such as oxygen, water vapor, and fog is discussed in Refs. 18 and 19.

To determine the amount of power received on the ground due to satellite transmission, we consider the power density

$$\Psi = \frac{P_t}{S} \tag{13.19}$$

where P_t is the power transmitted and S is the terrestrial area covered by the satellite. The value of P_t is a major requirement of the spacecraft. The coverage area is given by

$$S = 2\pi R^2(1-\cos\gamma) \tag{13.20}$$

where $R = 6378$ km is the radius of the earth. S is usually divided into a cellular pattern of spot beams, thereby enabling frequency reuse. The effective area of the receiving antenna is a measure of the ability of the antenna to extract energy from the passing electromagnetic wave and is given by

$$A_e = G_r \frac{\lambda^2}{4\pi} \tag{13.21}$$

where G_r is the gain of the receiving antenna and λ is the wavelength. The power received is the product of the power density and the effective area. Thus,

$$P_r = \Psi A_e = \frac{G_r \lambda^2}{4\pi S} P_t \tag{13.22}$$

This is known as the *Friis equation* relating the power received by one antenna to the power transmitted by the other. We first notice from this equation that for a given transmitted power P_t, the received power P_r is maximized by minimizing the coverage area S. Second, mobile terminals prefer having nondirectional antennas, thereby making their gain G_r fixed. Therefore, to maximize P_r encourages using as long a wavelength as possible, i.e., as low a frequency as practicable within regulatory and technical constraints.

The path loss accounts for the phenomenon, which occurs when the received signal becomes weaker as the distance between the satellite and the earth increases. In free space, the strength of the radiated signal diminishes as the square of the distance it travels, so the received power density is inversely proportional to the square of the distance. The path "free-space" loss (in dB) is given by

$$L_p = 92.45 + 20\log_{10} f + 20\log_{10} r \tag{13.23}$$

where f is the frequency (in GHz) and r is the distance (in km).

The noise density N_o is given by

$$N_o = kT_o \tag{13.24}$$

where $k = 1.38 \times 10^{-23}$ Ws/K is Boltzmann's constant and T_o is the equivalent system temperature, which is defined to include antenna noise and thermal noise generated at the receiver. Shannon's classical capacity theorem for the maximum error-free transmission rate in bits per second (bps) over a noisy power-limited and bandwidth-limited channel is

$$C = B\log_2\left(1 + \frac{P_r}{N_o}\right) \tag{13.25}$$

where B is the bandwidth and C is the channel capacity.

13.4. APPLICATIONS

Satellite communication services are uniquely suited for many applications involving wide area coverage. Satellites provide the key ingredient in the development of broadband communications and information processing infrastructure. Here, we consider five major applications of satellite communications: the use of very small aperture terminals (VSATs) for business applications; fixed satellite service (FSS), which interconnects fixed points, and mobile satellite (MSAT) service (MSS), which employs satellite to extend cellular network to mobile vehicles; satellite radio, which continuously provide entertainment to listeners; and satellite-based Internet, which enables IP-over-satellite connectivity.

13.4.1. VSAT Networks

A very small aperture terminal (VSAT) is a dish antenna that receives signals from a satellite. (The dish antenna has a diameter that is typically in the range of 1.2 m to 2.8 m

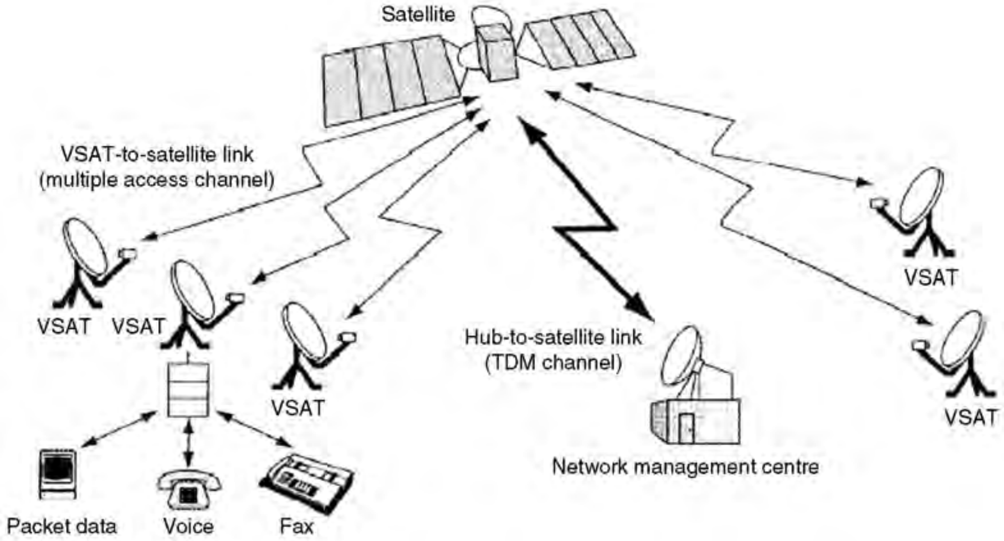


Figure 13.6 A typical VSAT network.

but the trend is toward smaller dishes, not more than 1.5 m in diameter.) A VSAT may also be regarded as a complete earth station that can be installed on the user's premises and provide communication services in conjunction with a larger (typically 6–9 m) earth station acting as a network management center (NMC), as illustrated in Fig. 13.6.

VSAT technology brings features and benefits of satellite communications down to an economical and usable form. VSAT networks have become mainstream networking solutions for long-distance, low-density voice and data communications because they are affordable to both small and large companies. Other benefits and advantages of VSAT technology include lower operating costs, ease of installation and maintenance, ability to manage multiple protocols, and ability to bring locations where the cost of leased lines is very high into the communication loop.

Satellite links can support interactive data applications through two types of architectures [3,20]: mesh topology (also called *point-to-point connectivity*) and star topology (also known as *point-to-multipoint connectivity*). Single-hop communications between remote VSATs can be achieved by full-mesh connectivity. Although the mesh and star configurations have different technical requirements, it is possible to integrate the two if necessary.

The star network employs a hub station. The hub consists of an RF terminal, a set of baseband equipment and network equipment. A VSAT network can provide transmission rates of up to 64 kbps. As common with star networks, all communication must past through the hub. That is, all communication is between a remote node and the hub; no direct node-to-node information transfer is allowed in this topology. This type of network is highly coordinated and can be very efficient. The point-to-multipoint architecture is very common in modern satellite data networks and is responsible for the success of the current VSAT.

A mesh network is more versatile than star network because it allows any-to-any communications. Also, the star network can provide transmission rates up to 64 kbps per remote terminal, whereas the mesh network can have its data rates increased to 2 Mbps or more. Mesh topology was used by the first satellite networks to be implemented. With time, there was a decline in the use of this topology but it remains an effective means of transferring information with least delay. Mesh topology applies to either temporary connections or dedicated links to connect two earth stations. All full-duplex point-to-point connectivities are possible and provided, as typical of a mesh configuration. If there are N nodes, the number of connections is equal to the permutation of $N(N-1)/2$. Mesh networks are implemented at C and Ku bands. The transmission rate ranges from about 64 kbps to 2.048 Mbps (E1 speed). Users have implemented 45 Mbps.

Several types of VSAT networks are now in operation, both domestically and internationally. There were over 1000 VSATs in operation at the beginning of 1992. Today, there are over 100,000 two-way Ku band VSATs installed in the United States and over 300,000 worldwide. Almost all of these VSATs are designed primarily to provide data for private corporate networks, and almost all two-way data networks with more than 20 earth stations are based on some variation of an ALOHA protocol for access [21,22]. The price of a VSAT started around $20,000 and dropped to around $6,000 in 1996.

13.4.2. Fixed Satellite Service

Several commercial satellite applications are through earth stations at fixed locations on the ground. The international designation for such an arrangement is *fixed satellite service* (FSS). The FSS is to provide communication service between two or more fixed points on earth, as opposed to mobile satellite services (MSS) (to be discussed later), which provides communication for two moving terminals. Although ITU defined FSS as a space radiocommunication service covering all types of satellite transmissions between given fixed points, the borderline between FSS and Broadcasting Satellite Service (BSS) for satellite television is becoming more and more blurred [23]. FSS applies to systems which interconnect fixed points such as international telephone exchanges. It involves GEO satellites providing 24 hour per day service.

Table 13.6 shows the WARC (World Administrative Radio Conference) frequency allocations for FSS. The table only gives a general idea and is by no way comprehensive. The FSS shares frequency bands with terrestrial networks in the 6/4 GHz and 14/12 GHz bands. Thus, it is possible that a terrestrial network could affect a satellite on the uplink or that a terrestrial network may be affected by the downlink from a satellite.

As exemplified by Intelsat, FSS has been the most successful part of commercial satellite communications. Early applications were point-to-point telephony and major trunking uses. Current applications of the FSS can be classified according to frequency (from about 3 MHz to above 30 GHz), the lowest frequency being the HF band. They

Table 13.6 Frequency Allocations for FSS (Below ~30 GHz)

Downlinks (in GHz)	Uplinks (in GHz)
3.4–4.2 and 4.5–4.8	5.725–7.075
7.25–7.75	7.9–8.4
10.8–11.7	
11.7–12.2 (Region 2 only)	12.75–13.25 and 14.0–14.5
12.6–12.75 (Region 1 only)	
17.7–21.2	27.5–31.0

include high-frequency (HF) service, private fixed services, auxiliary broadcasting (AUXBC) services, cable relay service (CARS), and federal government fixed services.

Although the telecommunications industry as a whole is growing rapidly, the FSS industry is not. The market trend is toward the replacement of long-haul microwave system with fiber. Fiber provides much greater capacity than microwaves.

13.4.3. Mobile Satellite Service

There is the need for global cellular service in all geographical regions of the world. The terrestrial cellular systems serve urban areas well; they are not economical for rural or remote areas where the population or teledensity is low. Mobile satellite (MSAT) systems can complement the existing terrestrial cellular network by extending communication coverage from urban to rural areas. Mobile satellite services (MSS) are not limited to land coverage but include marine and aeronautical services [6,24]. Thus, the coverage of mobile satellite is based on geographical and not on population coverage as in terrestrial cellular system and could be global.

MSAT or satellite-based PCS/PCN is being developed in the light of the terrestrial constraints. The low cost of installation makes satellite-based PCS simple and practical. The American Mobile Satellite Corporation (AMSC) along with Telesat Mobile of Canada are designing a geosynchronous MSAT to provide PCS to North America. The concept of MSAT is illustrated in Fig. 13.7.

Satellite communication among mobile earth stations is different from the cellular communication. First, the cells move very rapidly over the earth, and the mobile units, for all practical purposes, appear stationary—a kind of inverted cellular telephone system. Second, due to different designs, use of a handheld is limited to the geographical coverage

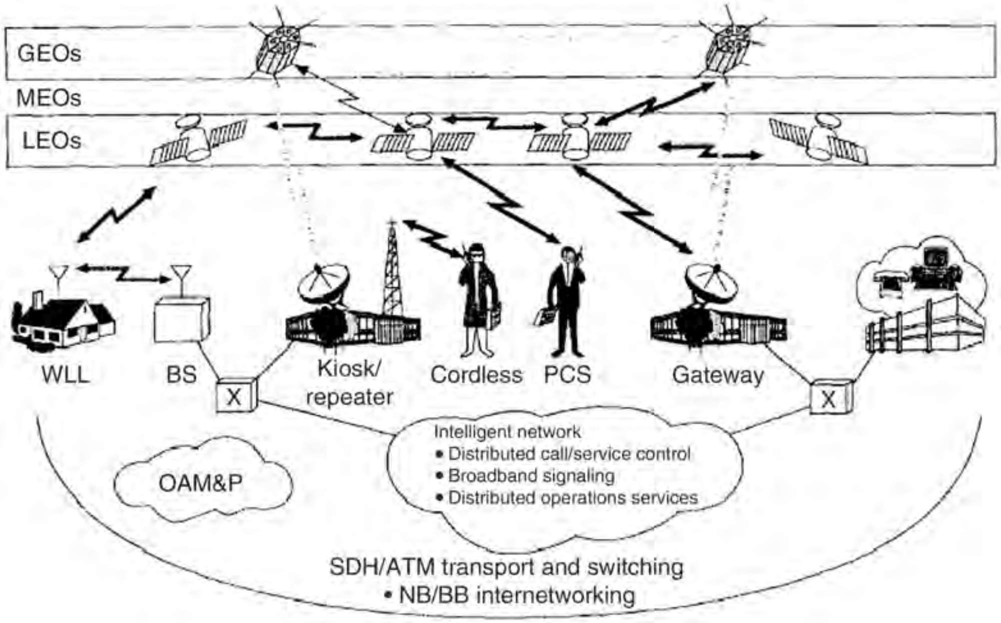


Figure 13.7 MSAT concept.

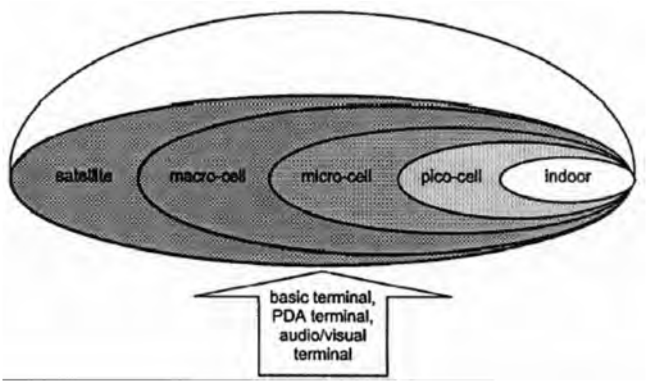


Figure 13.8 Various cell sizes.

of a specific satellite constellation and roaming of handheld equipment between different satellite systems will not be allowed. With personal communication systems (PCS), there will be a mix of broad types of cell sizes: the picocell for low-power indoor applications, the microcell for lower-power outdoor pedestrian application; macrocell for high-power vehicular applications; and supermacro cell with satellites, as shown in Fig. 13.8. For example, a mirocell of a PCS has a radius of 1 to 300 m.

There are two types of constellation design approaches to satellite-based PCS. One approach is to provide coverage using three GEO satellites at approximately 36,000 km above the equator. The other approach involves using the LEO and MEO satellites at approximately 500 to 1500 km above the earth's surface. Thus, MSS are identified as either GEO or nongeostationary orbit (NGSO) satellites [25].

The main purpose of MSAT or MSS is to provide data and/or voice services into a fixed or portable personal terminal, close to the size of today's terrestrial cellular phones, by means of interconnection via satellite. LEO and MEO satellites have been proposed as an efficient way to communicate with these handheld devices. The signals from the handheld devices are retransmitted via a satellite to a gateway (a fixed earth station) which routes the signals through the public switched telephone network (PSTN) to its final destination or to another handheld device.

Satellite systems designed for personal communications include the Iridium, Globalstar, and ICO systems [26–29]. All are global system covering everywhere on earth. Each of these is characterized by two key elements: a constellation of non-geosynchronous satellites (LEO or MEO) arranged in multiple planes and a handheld terminal (handset) for accessing PCS.

Iridium (www.iridium.com), which began in 1990, is the first mobile satellite telephone network to offer voice and data services to and from handheld telephones anywhere in the world. It uses a network of inter-satellite switches for global coverage and GSM-type technology to link mobile units to the satellite network. Several modifications have been made to the original idea, including reducing the number of satellites from 77 to 66 by eliminating one orbital plane. (The name Iridium was based on the fact that Iridium is the element in the periodic table whose atom has 77 electrons.) Some of the key features of the current Iridium satellite constellation are [30–33]:

Number of (LEO) satellites: 66 (each weighing 700 kg or 1500 lb)
Number of orbital planes: 6 (separated by 31.6° around the equator)

Number of active satellites per plane: 11 (uniformly spaced, with one spare satellite per plane at 130 km lower in the orbital plane)
Altitude of orbits: 780 km (or 421.5 nmi)
Inclination: 86.4°
Period of revolution: 100 min
Design life: 8 y

In spite of some problems expected of a complex system, Iridium is already at work. Its 66 LEO satellites were fully commercial as of November 1, 1998. But on August 13, 1999, Iridium filed for bankruptcy and was later bought by Iridium Satellite LCC. Vendors competing with Iridium include Aries, Ellipso, Globalstar, and ICO.

The second system is Globastar (www.globalstar.com), which is a satellite-based cellular telephone system that allows users to talk from anyplace in the world. It serves as an extension of terrestrial systems world-wide except for polar regions. The constellation is capable of serving up to 30 million subscribers. Globalstar is being developed by the limited partnership of Loral Aerospace Corporation and Qualcomm with ten strategic partners. A functional overview of Globalstar is presented in Fig. 13.9. The key elements are [34–36]:

Space segment: It comprises a constellation of 48 active LEO satellites located at an altitude of 1414 km and equally divided in 8 planes (6 satellites per plane). The satellite orbits are circular and are inclined at 52° with respect to the equator. Each satellite illuminates the earth at 1.6 GHz L band and 2.5 GHz S band with 16 fixed beams with service links, assignable over 13 FDM channels.
User segment: This includes mobile and fixed users.

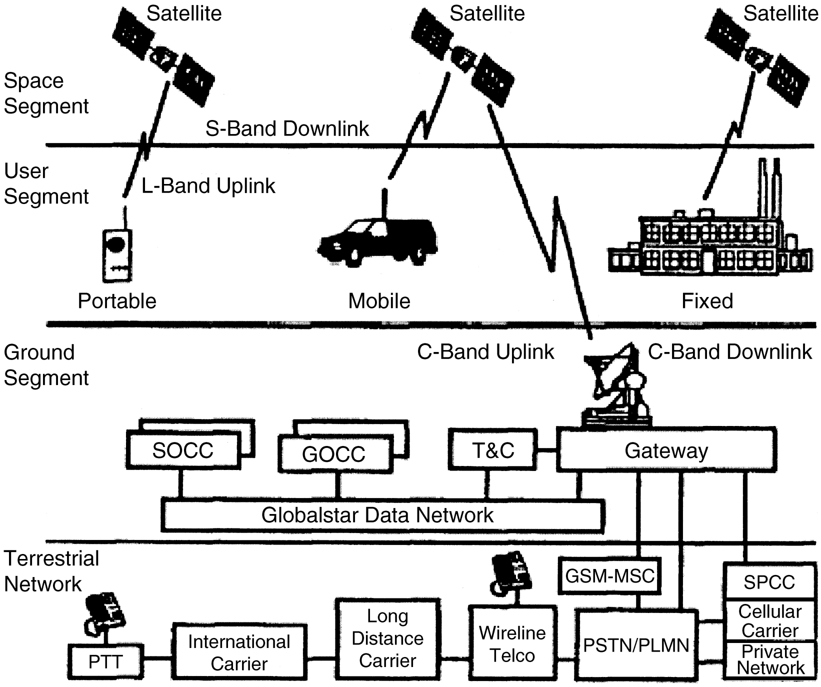


Figure 13.9 Globalstar system architecture.

Ground segment: This consists of gateways (large ground station), ground operations control center (GOCC), satellite operations control center (SOCO), and Globalstar data network (GDN). The gateway enables communications to and from handheld user terminals (UTs), relayed via satellite, with Public Switched Telephone Network (PSTN). A gateway with a single radio channel transmits on a single frequency.

The Globastar satellites employ "bent pipe" transponders with the feeder link at C band. Each satellite weighs about 704 lb and has a capacity of 2800 full-duplex circuits. It covers the earth with only 16 spots beams.

Since Globalstar plans to serve the military with commercial subscriptions, it employs signal encryption for protection from unauthorized calling party. Unlike Iridium, which offers a global service, Globalstar's business plan calls for franchising its use to partners in different countries.

The third system is the ICO system (originally called Inmarsat-P), which was built by Hughes Space and Communications (now Boeing Satellite Systems). ICO constellation is made of [37–39]:

Ten operational MEO satellites with five in each of the two inclined circular orbits at an altitude of 10,355 km.

One spare satellite in each plane, making 12 total launched.

Each satellite employs 163 spot beams.

Each satellite will carry an integrated C- and S-band payload.

Twelve satellite access nodes (SANs) located globally.

The inclination of the orbits is 45°—making it the lowest of the systems described. Although this reduces the coverage at high latitudes, it allows for the smallest number of satellites. The ICO system (www.ico.com) is designed to provide the following services:

Global paging

Personal navigation

Personal voice, data, and fax

The three constellations are compared in Table 13.7.

Table 13.7 Characteristics of Satellite PCS Systems

Parameter	Iridium	Globastar	ICO
Company	Motorola	Loral/Qualcomm	ICO-Global
No. of satellites	66	48	10
No. of orbit planes	6	8	2
Altitude (km)	780	1414	10,355
Weight (lb)	1100	704	6050
Bandwidth (MHz)	5.15	11.35	30
Frequency up/down (GHz)	30/20	5.1/6.9	14/12
Spot beams/satellite	48	16	163
Carrier bit rate (kps)	50	2.4	36
Multiple access	TDMA/FDMA	CDMA/FDMA	TDMA/FDMA
Cost to build ($ billion)	4.7	2.5	4.6
Service start date	1998	1999	2003

13.4.4. Satellite Radio

Satellite radio is broadcasting from satellite. With satellite radio, one can drive from Washington DC to Los Angeles, CA without changing the radio station and without static interference. Satellite eliminates localization, which is the major weakness of conventional radio. It transforms radio from a local medium into a national one. Satellite radio will permanently change radio just as cable changed television. It is regarded as radio beyond AM, beyond FM, or radio to the power of *X*.

Figure 13.10 displays a typical architecture of satellite radio. Satellite radio is based on digital radio, which produces a better sound from radio than analog radio. Digital radio systems are used extensively in communication networks. Digital radio offers CD quality sound, efficient use of the spectrum, more programming choice, new services, and robust reception even under the most challenging condition.

Satellite radio is both a new product and a service. As a product, it is a new electronic device that receives the satellite signal. As a service, it will provide consumers with 100 national radio stations, most of which will be brand-new, comprising various music, news, sports, and comedy stations.

Satellite radio service is being provided by DC two companies: XM Satellite Radio (also known as XM Radio), based in Washington, DC, and Sirius Satellite Radio, based in New York. The two companies obtained FCC licenses to operate digital audio radio service (DARC) system coast-to-coast throughout continental U.S.A. To avoid competition with terrestrial radio broadcasters, both satellite broadcasters will carry advertisement of nationally branded products.

XM Satellite Radio is made possible by two satellites, officially named "Rock and Roll," placed in geostationary orbit, one at 85 degrees West longitude and the other at

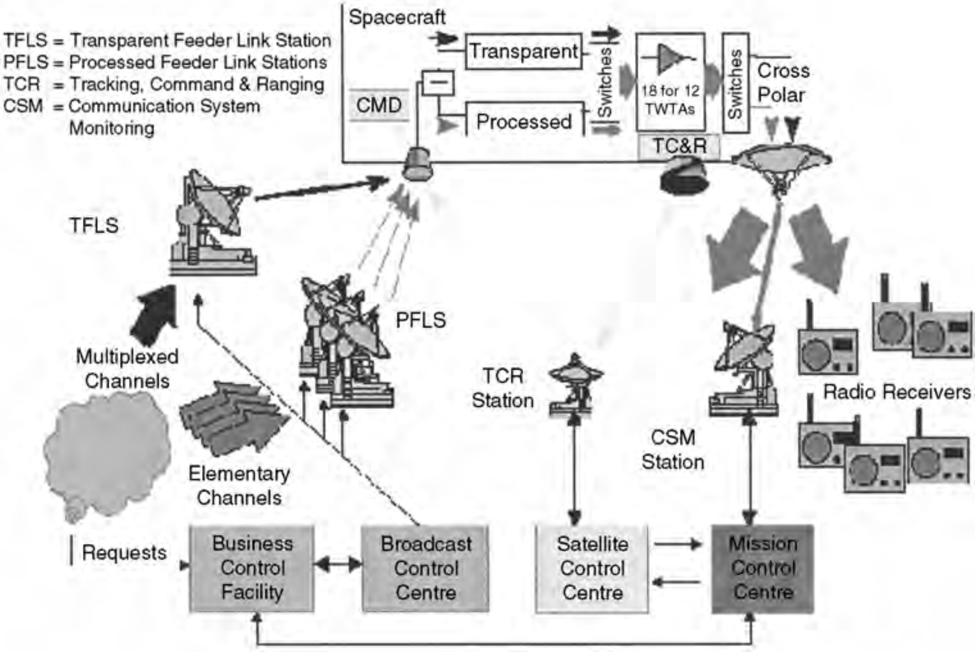


Figure 13.10 A typical architecture of satellite radio. (Source: EBU Technical Review.)

Table 13.8 Comparison of XM and Sirius systems

	XM	Sirius
Constellation	2 Satellites	3 Satellites
Satellite type	Boeing 702	SS/Loral FS-3000
Terrestrial repeaters	1500 in 70 areas	105 in 46 areas
Satellite costs (million)	\$439	\$120
Transmission rate	4 Mbps	4.4 Mbps
Uplink frequencies	7.05–7.075 GHz	7.06–7.0725 GHz
Downlink frequencies	2.3325–2.345 GHz	2.32–2.3325 GHz

115 degrees West longitude. Rock and Roll are Boeing 702 satellites, built by Boeing Satellite Systems. The satellites will be positioned above the United States. In September, 2001, XM Satellite Radio started to broadcast. Subscribers pay as little as \$9.95 per month after they purchase an AM/FM/XM radio.

Sirius Satellite Radio, on the other hand, does not use GEO satellites. Rather, it is flying three satellites which are equally spaced in an elliptical 47,000 × 24,500-km orbit that takes 24 h to complete. This ensures that each satellite spends about 16 h a day over the continental U.S.A., with at least one satellite over the country at any time. It also means that Sirius will be higher in the sky than XM, which is at the zenith only at the equator. Sirius charges \$12.95 a month for its service. The systems of both satellite radio companies are compared in Table 13.8.

Besides XM Satellite Radio and Sirius Satellite Radio that operate in the United States, WorldSpace is another radio satellite broadcasting company already broadcasting in Africa and Asia. With a constellation of three satellites (AfriStar to cover Africa and Middle East, AmeriStar to serve Latin America and the Caribbean, and AsiaStar to serve nearly all Asia), WorldSpace intends to touch all or parts of the four continents, especially those areas of the world that most conventional radio stations cannot reach.

As a new technology, satellite radio is not without its own peculiar problems. First, people are not yet used to paying for radio programming. If the programming of the satellite broadcasters is not better than what people are getting free from regular, terrestrial radio, they will be reluctant to pay. So the real question is: How many people are going to subscribe? Second, satellite broadcasting requires a near-omnidirectional receive antennas for cars, which in turn requires a powerful signal from the satellite. Third, some believe that the two companies will face a big hurdle in transforming radio from a local medium into a national one. The many-pie-in-the-sky companies are faced with great risks ahead of them.

Satellite radio may will transform radio industry, which has seen little technological change since the discovery of FM, some 40 y ago. Receiving digital-quality music from radio satellite is a major technical milestone. It is as revolutionary to the entertainment industry as was the invention of radio itself. The future of radio by satellite is exciting but uncertain [40–42].

13.4.5. Satellite-Based Internet

The Internet is becoming an indispensable source of information for an evergrowing community of users. The thirst for Internet connectivity and high performance remains unquenched. This has led to several proposals for integrating satellite networks with terrestrial ISDN and the Internet [43–46].

Several factors are responsible for this great interest in IP-over-satellite connectivity. First, satellites cover areas where land lines do not exist or cannot be installed. Satellites can serve as an access link between locations separated by great distances. Second, developments in satellite technology allow home users to receive data directly from a geostationary satellite channel at a rate 20 times faster than of an average telephone modem. With more power transponders utilizing wider frequencies, commercial satellite links can now deliver up to 155 Mbps. Third, the unique positioning of satellites between sender and receivers lends itself to new applications such as IP multicast, streaming data, and distributed web caching. Fourth, satellite connectivity can be rapidly deployed because trenches and cable installation are unnecessary. Moreover, satellite communication is highly efficient for delivering multimedia content to businesses and homes [47].

As an inherently broadcast system, a satellite is attractive to point-to-multipoint and multipoint-to-multipoint communications especially in broadband multimedia applications. The asymmetrical nature of Web traffic suggests a good match to VSAT systems since the VSAT return link capacity would be much smaller than the forward link capacity.

A typical network architecture for a satellite-based Internet service provider (ISP) is shown in Fig. 13.11, which has been simplified to focus on the basic functionality. It includes its own satellite network and a network of ground gateway stations. The ground gateway stations interface with the public network through which access to the Internet is gained. The number of satellites may vary from dozens to hundreds, and they may be GEO, MEO, or LEO. Thus, the satellite-based Internet has several architectural options due to the diverse designs of satellite systems, orbit types, payload choice, and intersatellite links designs [48].

There is ongoing research into various aspects of implementation and performance of TCP/IP over satellite links. Related issues include the slow start algorithm, the ability to accommodate large bandwidth-delay products, congestion control, acknowledgment, and error recovery mechanisms.

More information about satellite communications systems can be obtained from Refs. 49–53.

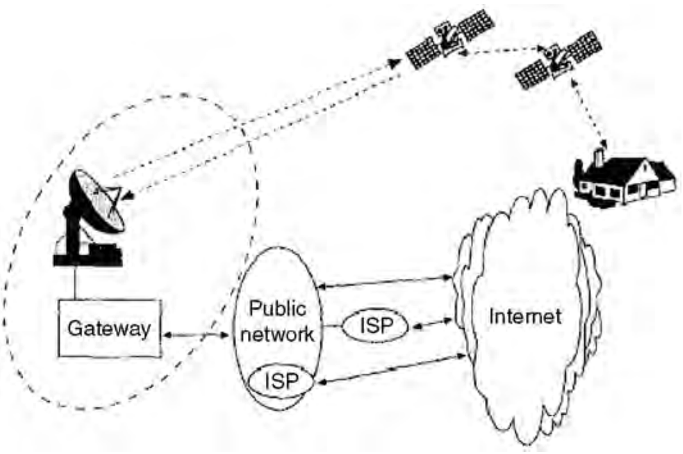


Figure 13.11 A typical configuration for satellite-based Internet.

REFERENCES

1. Marihart, D.J. Communications technology guidelines for EMS/SCADA systems. IEEE Trans. Power Delivery **April 2001**, *16* (2), 181–188.
2. Calcutt, D.; Tetley, L. *Satellite Communications: Principles and Applications*; Edward Arnold: London, 1994; 3–15, 321–387.
3. Elbert, B.R. *The Satellite Communication Applications Handbook*; Artech House: Norwood, MA, 1997; 3–27, 257–320.
4. Pritchard, W. Geostationary versus nongeostationary orbits. Space Comm. **1993**, *11*, 205–215.
5. Farserotu, J.; Prasad, R. *IP/ATM Mobile Satellite Networks*; Artech House: Boston, MA, 2002; 14.
6. Ha, T.T. *Digital Satellite Communications*; McGraw-Hill: New York, 1990; 1–30, 615–633.
7. Tycz, T.S. Fixed satellite service frequency allocations and orbit assignment procedures for commercial satellite systems. Proc IEEE **July 1990**, *78* (7), 1283–1288.
8. Morgan W.L.; Gordon, G.D. *Communications Satellite Handbook*; Wiley: New York, 1989; 573–589.
9. Martin, J.S. *Communications Satellite Systems*; Prentice Hall: Englewood Cliffs, NJ, 1978; 17–28.
10. Richharia, M. *Satellite Communication Systems*; McGraw-Hill: New York, 1995; 16–49.
11. Fthenakis, E. *Manual of Satellite Communications*; McGraw-Hill: New York, 1984; 31–44.
12. Pratt, T. *Satellite Communications*; Wiley: New York, 1986; 11–51.
13. Wu, W.W. Mobile satellite communications. Proc. IEEE **Sept. 1994**, *82* (9), 1431–1448.
14. Hu, Y.; Li, V.O.K. Satellite-based Internet: a tutorial. IEEE Comm. Mag. **March 2001**, 154–162.
15. Propagation special issue. Int. J. Satellite Comm. **May/June 2001**, *19* (3).
16. Hogg D.C.; Chu, T.S. The role of rain in satellite communication. Proc IEEE **1975**, *63*, 1308–1331.
17. Bargellini P.L.; Hyde, G. Satellite and space communications. In *Reference Data for Engineers*; 8th Ed.; SAMS: Carmel, IN, 1993; Chapter 27.
18. Ippolito, L.J. *Radiowave Propagation in Satellite Communications*; Van Nostrand Reinhold: New York, 1986.
19. Gordon G.D.; Morgan, W.L. *Principles of Communications Satellites*; Wiley: New York, 1993.
20. Elbert, B.R. *Introduction to Satellite Communication*; Artech House: Norwood, MA, 1999; 390–395.
21. Hadjitheodosiou, M.H. Next generation multiservice VSAT networks. Electron. Comm. Eng. J. **June 1997**, 117–126.
22. Abramson, N. VSAT data networks. Proc. IEEE **July 1990**, *78* (7), 1267–1274.
23. Raison, J.C. Television via satellite: convergence of the broadcasting-satellite and fixed-satellite service—the European experience. Space Comm. **1972**, *9*, 129–141.
24. Wood, P. Mobile satellite services for travelers. IEEE Comm. Mag. **Nov. 1991**, 32–35.
25. Abrishamkar, F. PCS global mobile services. IEEE Comm. Mag. **Sept. 1996**, 132–136.
26. Comparetto G.; Ramirez, R. Trends in mobile satellite technology. Computer **Feb. 1997**, 44–52.
27. Evans, J.V. Satellite systems for personal communications. IEEE Ant. Prop. Mag. **June 1997**, *39* (3), 7–20.
28. Satellite systems for personal communications. Proceedings of the IEEE **June 1997**, *39* (3), 7–20.
29. Satellite communications—a continuing revolution. IEEE Aerospace Electron. Sys. Mag. **Oct. 2000**, 95–107.
30. Pattan, B. *Satellite-Based Cellular Commuications*; McGraw-Hill: New York, 1998; 45–88.
31. Lemme, P. Iridium: Aeronautical satellite communications. IEEE AES Sys. Mag. **Nov. 1999**, 11–16.
32. Hubbel, Y.C. A comparison of the iridium and AMPS systems. IEEE Network **March/April 1997**, 52–59.

33. Leopold R.J.; Miller, A. The iridium communications system. IEE Potentials **April 1993**, 6–9.
34. Hirshfield, E. The Globalstar system: breakthroughs in efficiency in microwave and signal processing technology. Space Comm. **1996**, *14*, 69–82.
35. Dietrich, F.J. The Globalstar cellular satellite system. IEEE Trans. Ant. Prop. **June 1998**, *46* (6), 935–942.
36. Hendrickson, R. Globalstar for the military. Proc. MILCOM **1998**, *3*, 808–813.
37. Poskett, P. The ICO system for personal communications by satellite. Proc. IEE Colloquim (Digest), Part 1 **1998**, 211–216.
38. Ghedia, L. Satellite PCN—the ICO system. Int. J. Satellite Comm. **1999**, *17*, 273–289.
39. Werner, M. Analysis of system parameters for LEO/ICO-satellite communication networks. IEEE J. Selected Areas Comm. **Feb. 1995**, *13* (2), 371–381.
40. Sadiku, M.N.O. XM radio. IEEE Potentials **April/May 2002**.
41. Wood, D. Digital radio by satellite. EBU Tech. Rev. **Summer 1998**, 1–9.
42. Layer, D.H. Digital radio takes to the road. IEEE Spectrum **July 2001**, 40–46.
43. Otsu, T. Satellite communication system integrated into terrestrial ISDN. IEEE Trans. Aerospace Electron. Sys. **Oct. 2000**, *36*(4), 1047–1057.
44. Metz, C. TCP over satellite ... the final frontier. IEEE Internet Computer **Jan./Feb. 1999**, *3* (1), 76–80.
45. Choi, H.K. Interactive web service via satellite to the home. IEEE Comm. Mag. **March 2001**, 182–190.
46. Hu Y.; Li, V.O.K. Satellite-based internet: a turorial. IEEE Comm. Mag. **March 2001**, 154–162.
47. Metz, C. IP over satellite: Internet connectivity blasts off. IEEE Internet Computer **July/August 2000**, 84–89.
48. Cooper P.W.; Bradley, J.F. A space-borne satellite-dedicated gateway to the Internet. IEEE Comm. Mag. **Oct. 1999**, 122–126.
49. Special issue, Satellite communications. Proc. IEEE **March 1977**, *65* (3).
50. Special issue, Satellite communication networks. Proc. IEEE, **Nov. 1984**, *72* (11).
51. Special issue, Global satellite communications technology and system. Space Comm. **2000**, *16*.
52. Maral G.; Bousquet, M. *Satellite Communications Systems*, 3rd Ed.; Wiley: New York, 1998.
53. Sadiku, M.N.O. *Optical and Wireless Communications: Next Generation Networks*; CRC Press: Boca Raton, FL, 2002.

14
Optical Communications

Joseph C. Palais
Arizona State University
Tempe, Arizona, U.S.A.

14.1. INTRODUCTION

Electromagnetics plays a key role in modern optical telecommunications systems. This chapter emphasizes the electromagnetic phenomena peculiar to fiber optic communications.

Optical communications was first seriously considered just after the invention of the laser in 1960 [1,2]. Atmospheric propagation was proposed and much research was done in the following decade. Problems with weather, line-of-site clearance, beam spreading, and safety (and maybe others) removed free-space optical communication as a major player in the communications area.

In the mid-1960s, guided propagation in a glass fiber was proposed as a strategy for overcoming the many problems of optical atmospheric propagation for telecommunications. In 1970, the first highly transparent glass fiber was produced, making fiber-optic communications practical.

14.1.1. Fiber-Optic System

To help understand the electromagnetic features of fiber-optic communications, we will look briefly at the major components of an optical link. The basic fiber system is pictured in Fig. 14.1. The message is assumed to be available in electronic form, usually as a current.

The transmitter contains a light source that is modulated so that the optical beam carries the message. As an example, for a digital signal, the light beam is electronically turned on (for binary ones) and off (for binary zeros). The optical power in a digitally modulated signal is pictured in Fig. 14.2. The optical beam is the carrier of the digital message. The most likely choices for fiber-optic light sources are the light-emitting diode and the laser diode. Several characteristics of the light source determine the behavior of the propagating optical wave. Because of that we will briefly describe some of the properties of common sources in a later section of this chapter.

The modulated light beam is coupled into the transmission fiber. At the receiver, the signal is collected by a photodetector, which converts the information back into electrical form. The photodetectors do not affect the propagating properties of the wave but certainly must be compatible with the rest of the system. For completeness then, we will briefly describe properties of the photodetectors in a later section of this chapter.

14.1.2. Optical Spectrum

Fiber communications utilize carrier wavelengths in the optical region of the electromagnetic spectrum. A partial view of the optical spectrum, Fig. 14.3, shows the ultraviolet, visible, and near-infrared portions. Most fiber communications use carriers in the infrared, because that is where glass fiber attenuation losses are lowest. There is some activity in the visible using plastic fibers (which have higher losses than glass) for short paths.

The wavelength regions where fiber systems have been constructed appear in Table 14.1.

The relationship between wavelength (λ) and frequency (f) is

$$\lambda = \frac{c}{f} \tag{14.1}$$

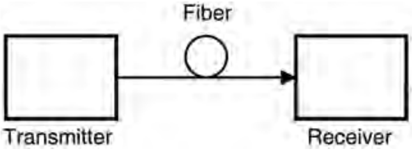


Figure 14.1 Basic fiber communications system.

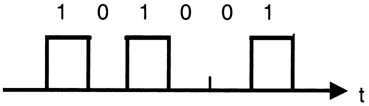


Figure 14.2 Digitally modulated optical signal.

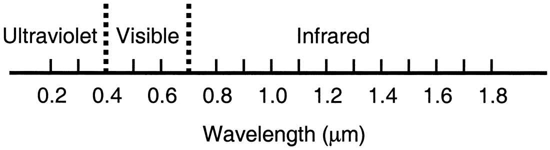


Figure 14.3 Optical spectrum.

Table 14.1 Major Wavelength Regions for Fiber Systems

Wavelength (nm)	Fiber material	Loss (dB/km)
650–670	Plastic	120–160
800–900	Glass	3
1250–1350	Glass	0.5
1500–1600	Glass	0.25

The velocity of light in empty space is $c = 3 \times 10^8$ m/s. As an example, a wavelength of 1.5 μm corresponds to a frequency of 2×10^{14} Hz (and a period of oscillation of 0.5×10^{-14} s). Because the amount of information that can be transmitted is proportional to the carrier frequency, the amount of information that can be transmitted on an optical carrier is enormous. In addition, multiple optical carriers (different wavelengths) can travel the fiber, further enhancing their information carrying capacity.

14.1.3. Light Sources

The most commonly used light sources in optical communication are the light-emitting diode (LED) and the laser diode (LD) [3–6].

Light-Emitting Diodes

The LED is a pn junction semiconductor. The LED emits light in the visible or infrared regions when forward biased. Free charges (electrons and holes) injected into the junction region spontaneously recombine with the subsequent emission of radiation.

The electronic driving circuit and output characteristic appear in Fig. 14.4. Ideally the output optical power increases linearly with input current. Thus, the optical power (P) waveform is a replica of the driving current (i). In the forward-biased region, the output power is given by

$$P = a_1 i \tag{14.2}$$

Both analog and digital modulation is possible with the LED. Bandwidth limitations restrict modulation to a few hundred megahertz and a few hundred megabits per second.

Operating voltages are on the order of a volt or 2. Operating currents are on the order of a few tens of milliamps. Output powers are on the order of a few milliwatts.

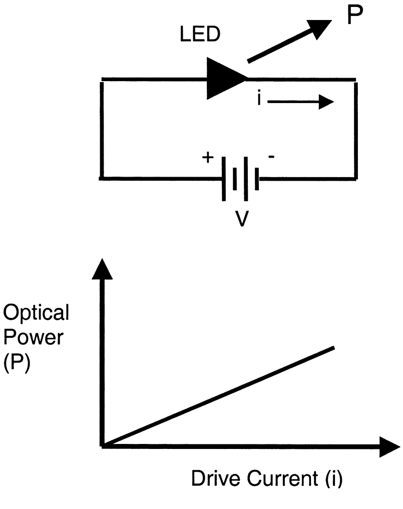


Figure 14.4 LED driving circuit and output characteristic.

Table 14.2 Materials for Semiconductor Light Sources

Material	Band-gap energy (eV)	Wavelength (μm)
GaInP	1.82–1.94	0.64–0.68
AlGaAs	1.4–1.55	0.8–0.89
InGaAsP	0.73–1.34	0.93–1.7
GaAs	1.4	0.89
InGaAs	0.95–1.24	1.0–1.31

The output wavelength is determined by the band-gap energy (W_g) of the semiconductor material. In particular, the output wavelength is given by

$$\lambda = \frac{1.24}{W_g} \tag{14.3}$$

In this equation, the wavelength is in micrometers and the band-gap energy is in electron volts. Table 14.2 indicates the materials commonly used for LEDs.

An important property of light sources used in fiber-optic communications is its *spectral width*. This is the range of wavelengths (or frequencies) over which it emits significant amounts of power. Ideally a light source would be monochromatic, emitting a single wavelength. In practice, no such light source exists. All radiate over a range. For LEDs, the range is typically on the order of 20 to 100 nm. Coherence refers to how close the source radiation is to the ideal single wavelength. The smaller the spectral width, the more coherent the source.

Laser Diodes

The laser diode shares a number of characteristics with the LED. It is also a pn junction semiconductor that emits light when forward biased. Light amplification occurs when photons stimulate free charges in the junction region to recombine and emit. The light beam is reflected back and forth through the amplifying medium by reflectors at each end of the junction. The amplification together with the feedback produce an oscillator emitting at optical frequencies.

The stimulated recombination leads to radiation that is more coherent than that produced by spontaneous recombination. Spectral widths for LDs are typically in the range of 1 to 5 nm. Specially constructed LDs can be designed to have even smaller spectral widths.

The laser diode is much faster than the LED, allowing for much higher modulation rates. Bandwidths of several gigahertz and several gigabits per second are achievable. For higher rates, external modulation is required.

The driving circuit and the output characteristic appear in Fig. 14.5. Note that the output power does not increase until the input current is beyond a threshold value (I_{TH}). Thresholds are on the order of a few to a few tens of milliamperes. Voltages are on the order of a few volts and output powers of a few milliwatts.

The materials used for LDs are the same as those used in constructing LEDs. Equation (14.3) applies to laser diodes, as does Table 14.2.

14.1.4. Photodetectors

The most common photodetector for fiber communications is the semiconductor junction photodiode. The photodetector converts optical power (P) to an electric current. The received current (i) is [7]

$$i = \rho P \tag{14.4}$$

The term ρ is the photodetector responsivity. A table of photodetector materials, their operating wavelengths, and their peak responsivities appears in Table 14.3. The cutoff wavelength is determined by the band-gap energy and is given by

$$\lambda_c = \frac{1.24}{W_g} \tag{14.5}$$

Just as in Eq. (14.3) for the emission of an LED, the wavelength is in micrometers and the band-gap energy is in electron volts. Only wavelengths equal to or smaller than the cutoff wavelength can be detected.

Because the optical power waveform is a replica of the message current, the receiver current is then a replica of the original signal current. This is ideally the case. Distortions

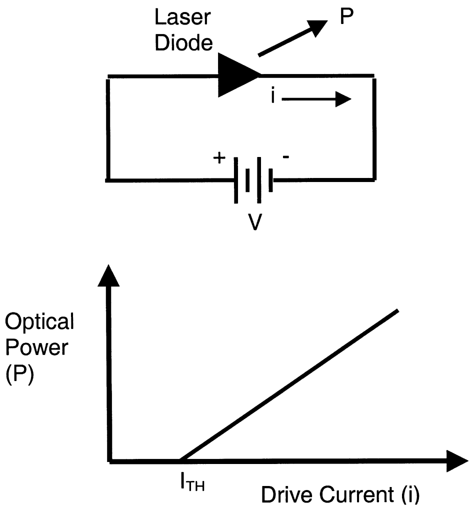


Figure 14.5 Laser diode driving circuit and output characteristic.

Table 14.3 Materials for Semiconductor Photodetectors

Material	Wavelength (μm)	Peak response (μm)	Peak responsivity (A/W)
Si	0.3–1.1	0.8	0.5
Ge	0.5–1.8	1.55	0.7
InGaAs	1.0–1.7	1.7	1.1

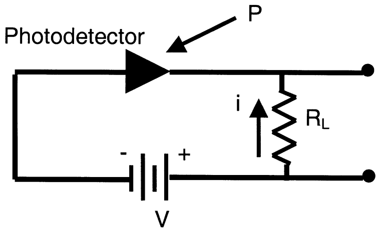


Figure 14.6 Receiver circuit.

in the waveform caused by transmission and caused by transmitter and receiver irregularities will degrade the signal. We will be describing the degradations due to transmission along the fiber in detail later in this chapter.

The circuit for the simplest type of receiver is shown in Fig. 14.6. According to Eq. (14.4) the photodetector acts like a constant current source. Therefore, the output voltage $v = iR_L$ can be increased by increasing the load resistance R_L. However, the receiver bandwidth is no larger than $B = 1/2\pi R_L C_d$, so that doing so decreases the receiver bandwidth. The photodiode's shunt capacitance is C_d.

14.1.5. Electromagnetic Problems

There is a set of electromagnetic problems that bears upon fiber-optic communications. These problems can be described as follows:

Propagation of a plane wave in unbounded media
Reflection at a plane boundary
Waves in an electromagnetic cavity
Guided propagation in a rectangular dielectric waveguide
Propagation in an optical fiber

Analyses of these electromagnetic phenomena in the next few sections help explain the design, operation, capabilities, and limitations of fiber transmission lines, fiber-optic components, and fiber-optic systems.

14.2. PROPAGATION

This section describes properties of traveling waves, emphasizing traveling pulses as in an optical digital communication system.

14.2.1. Wave Properties

All electromagnetic fields must satisfy the wave equation [8,9]. It evolves from Maxwell's equations. For dielectric materials it is given by

$$\nabla^2 E = \mu\varepsilon \frac{\partial^2 E}{\partial t^2} \tag{14.6}$$

In rectangular coordinates, the Laplacian is

$$\nabla^2 E = \frac{\partial^2 E}{\partial x^2} + \frac{\partial^2 E}{\partial y^2} + \frac{\partial^2 E}{\partial z^2} \tag{14.7}$$

It is possible to understand many complex electromagnetic wave phenomena by studying the simple case of a plane wave traveling in an unbounded medium. The electric field for such a wave traveling in the z direction can be written (in complex form) as

$$E = E_0 e^{-\alpha z} e^{j(\omega t - kz)} \tag{14.8}$$

The instantaneous form of this field is generated by taking the real part of the complex field. In this case, the result is

$$E = E_0 e^{-\alpha z} \cos(\omega t - kz) \tag{14.9}$$

This is a solution to the electromagnetic wave equation that is appropriate for a wave propagating in an unbounded medium. The amplitude of the field is $E_0 e^{-\alpha z}$, its radian frequency is $\omega = 2\pi f$, and k is the *propagation factor*. The term $\omega t - kz$ is the *phase* of the wave, while kz is the phase shift over a distance z. In the case being considered, at any instant of time the phase is constant over any plane given by a fixed value of z. Since a fixed value of z defines a plane (parallel to the xy plane), the field described above is a plane wave.

The factor α is the *attenuation coefficient*. It is determined by the losses in the medium. For an ideal (lossless) medium, α would be zero. If α is given in units of km^{-1}, the loss in dB/km is related to it by

$$\text{dB/km} = -8.685\alpha \tag{14.10}$$

The propagation factor is related to frequency and the phase velocity (v) of the wave by

$$k = \frac{\omega}{v} \tag{14.11}$$

The wave velocity in a medium is determined by its *refractive index,* as given by

$$n = \frac{c}{v} \tag{14.12}$$

That is, the *index of refraction* is the ratio of the velocity of light in empty space to that in the medium. Because most media slow light beams, their indices of refraction are greater than unity. Glasses used for fibers have an index close to 1.5. This tells us that a light beam travels in a glass fiber at a speed of approximately

$$v = \frac{c}{n} = \frac{3 \times 10^8}{1.5} = 2 \times 10^8 \text{ m/s} \tag{14.13}$$

Fortunately, the travel delay time at this speed is short compared to the response time of human beings for any terrestrial distances. For example, a fiber telephone link including a path under the Pacific ocean may be as long as 10,000 km. The propagation time along this path would be 0.05 s. The two way delay would be 0.1 s, a bit shorter than most humans can detect. We conclude that telephone transmission over fibers appears to be instantaneous to human participants. The wavelength is given in terms of the propagation coefficient by

$$\lambda = \frac{2\pi}{k} \tag{14.14}$$

The electric field of a traveling wave is sketched in Fig. 14.7. The wave is shown at two instants of time, indicating movement to the right. The dotted lines form the envelope of the wave. The peak wave amplitude diminishes with distance traveled because of attenuation. The wavelength, which is the distance between adjacent points of equal phase, is also indicated on the figure.

We know that plane waves travel in unbounded media with the electric field vector lying in a plane perpendicular to the direction of travel. For a plane wave traveling in the z direction, this would be the xy plane. The field given in Eq. (14.8) is the scalar component of the vector field. For simplicity, it can be thought of as the x component of the electric field. It could represent the y component just as well. Any other vector in the xy plane can be resolved into its x and y components. Thus, we say there are two independent ways in which a plane wave can propagate in the unbounded medium. The different ways in which a field can propagate in a given environment are called its *modes.*

Polarization refers to the direction of the electric field. A field that points in just one direction (say the x direction) is *linearly polarized.* A plane wave in an unbounded medium can propagate in two linearly polarized modes.

A *circularly polarized* wave is produced by the simultaneous transmission of two orthoganally polarized waves that are 90 degrees out of phase with each other. For example, the x-directed field

$$E = E_0 e^{-\alpha z} e^{j(\omega t - kz)} \tag{14.15}$$

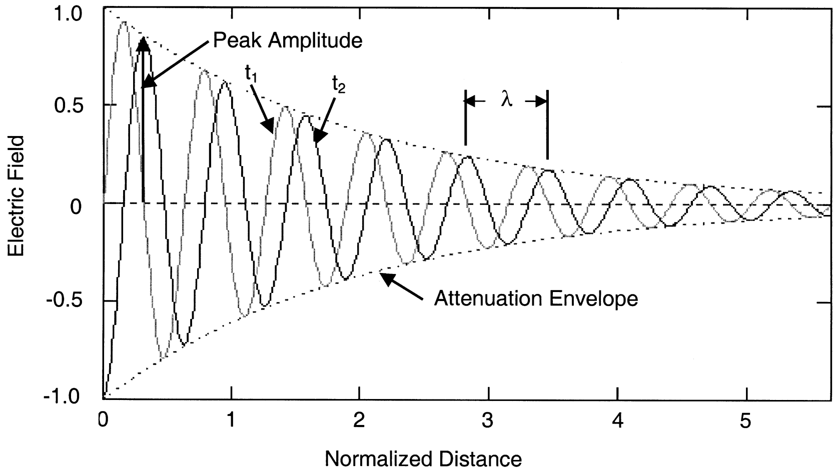


Figure 14.7 Traveling wave at two instances of time, showing peak amplitude, wavelength, and attenuation envelope. Time t_2 is later than time t_1 indicating wave movement to the right.

combined with the y-directed field

$$E = E_0 e^{-\alpha z} e^{j(\omega t - kz + \pi/2)} \tag{14.16}$$

produces a total electric field vector that rotates in a circle at radian frequency ω as time progresses.

If these last two orthogonal fields had a random phase difference (replacing $\pi/2$ in the last equation with a randomly varying function of time), the resulting electric vector would trace out a random pattern with time. This represents a *nonpolarized* (or *unpolarized*) wave. Equations for the fields such as those above imply that the waves are perfectly monochromatic. This is a simplification that produces reliable results in many cases, but is not always applicable. As we mentioned earlier in this chapter, the radiation emitted by practical sources exists over a range of wavelengths. This can be modeled mathematically by including a randomly time varying term in the phase of the wave. For example,

$$E = E_0 e^{-\alpha z} e^{j[\omega t - kz + \phi(t)]} \tag{14.17}$$

where $\phi(t)$ is a random function. In most fiber applications, the light is unpolarized, either because the light source produces unpolarized light or the wave becomes unpolarized during transmission. Bends and twists in the fiber along with other discontinuities in the path (e.g., at connectors and splices) together with random motion of the fiber (e.g., caused by vibrations) are the cause of the depolarization.

14.2.2. Pulse Transmission

A problem in transmission of pulses occurs because of two factors. One is that the source light is not emitted at a single wavelength but exists over a range of wavelengths (the source spectral width). The other factor is that the index of refraction is not the same for all wavelengths. In fact, for glasses used in fiber, the refractive index varies with wavelength. *Dispersion* is the name given to this property of velocity dependence on wavelength. *Material Dispersion* is the appropriate name when the dispersion is due to a property of the material [10].

Dispersion causes a pulse of light to lengthen as it traverses the fiber. This is because each of the component wavelengths travels at a different speed, each arriving with a slight delay with respect to the others. The amount of pulse spreading ($\Delta\tau$) per unit of length of fiber (ΔL) is given by

$$\Delta\left(\frac{\tau}{L}\right) = -M\Delta\lambda \tag{14.18}$$

The factor M is the *material dispersion factor* (or just the *material dispersion*) and is plotted in Fig. 14.8 for pure silica. Note that it is high in the region around 800 nm, goes to zero near 1300 nm, and is small and negative around 1550 nm. Clearly the 1300 nm region is favored to minimize pulse spreading. Spectral widths of common fiber light sources were given earlier in this chapter.

In the range 1200 to 1600 nm, the material dispersion factor can be approximated by

$$M = \frac{M_0}{4}\left(\lambda - \frac{\lambda_0^4}{\lambda^3}\right) \tag{14.19}$$

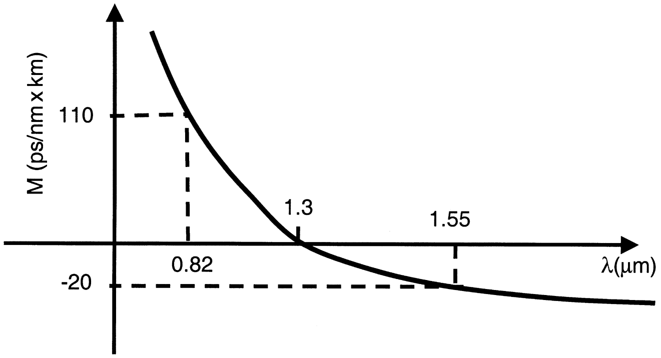


Figure 14.8 Material dispersion for silica glass.

The constant M_0 is approximately $-0.095\,\text{ps}/(\text{nm}^2 \times \text{km})$. The zero dispersion wavelength is λ_0. It is close to 1300 nm for silica fibers. As an example, at 1550 nm the material dispersion factor is close to $-20\,\text{ps}/(\text{nm} \times \text{km})$. Using an LED with spectral width of 20 nm, yields a pulse spread per unit length of the transmission path

$$\Delta\left(\frac{\tau}{L}\right) = -M\Delta\lambda = -(-20)20 = 400\,\text{ps/km} \tag{14.20}$$

The problem with pulse spreading is that it limits the information carrying capacity of the fiber. Pulses that spread eventually overlap with neighboring pulses, creating intersymbol interference. This leads to transmission errors and must be avoided. The direct way to avoid this is to place pulses further apart at the transmitter. This means lowering the data rate. The limits on data capacity caused by pulse spreading for non-return-to-zero and return-to-zero pulse codes [11] is

$$R_{\text{NRZ}} \times L = \frac{0.7}{\Delta(\tau/L)} \tag{14.21}$$

$$R_{\text{RZ}} \times L = \frac{0.35}{\Delta(\tau/L)} \tag{14.22}$$

Using the numerical values in the preceding example yields

$$R_{\text{NRZ}} \times L = 1.75\,\text{Mb/s} \times \text{km} \tag{14.23}$$

and

$$R_{\text{RZ}} \times L = 0.875\,\text{Mb/s} \times \text{km} \tag{14.24}$$

Similarly, pulse spreading reduces the bandwidth of an analog system. The 3-dB bandwidth limit is

$$f_{3\text{-dB}} \times L = \frac{0.35}{\Delta(\tau/L)} \tag{14.25}$$

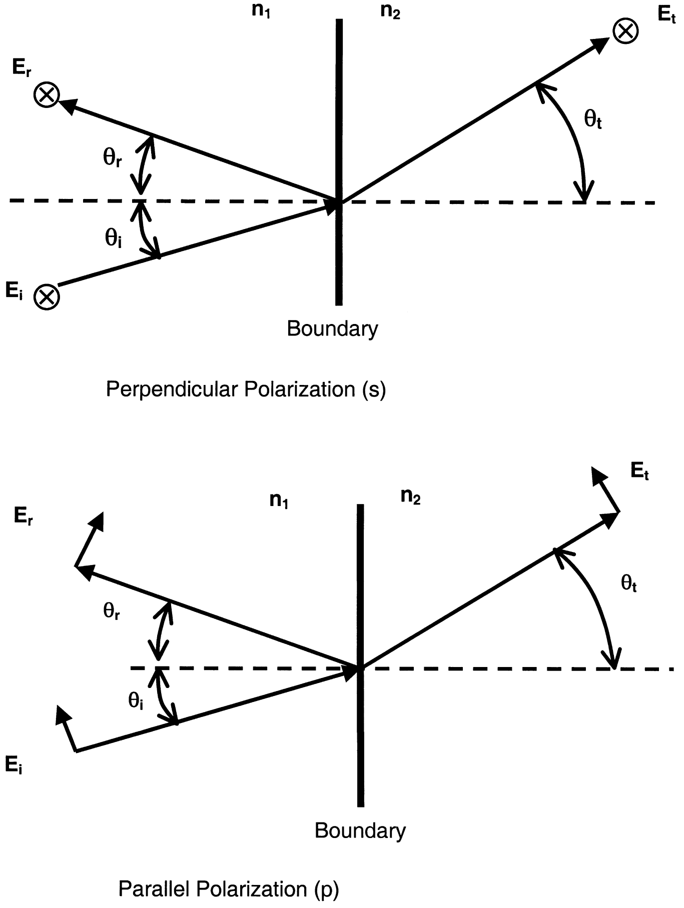
Using the same numerical values as above yields the limit as

$$f_{3\text{-dB}} \times L = 0.875\,\text{MHz} \times \text{km} \tag{14.26}$$

At modulation frequencies much lower than that calculated above, the analog signal propagates without distortion. At the 3-dB frequency the amplitude of the signal diminishes to 50% of what it was at lower frequencies. At modulation frequencies well above the 3-dB value, the signals are attenuated greatly. Pulse spreading causes the fiber to act as a low-pass filter, allowing only the lower modulating frequencies to pass.

14.2.3. Snell's Law and Total Reflection

Reflection of light at a plane boundary is one of the classical problems in optics [12,13]. Results from the analysis of this problem explain the operation of numerous optical components. The physical situation is drawn in Fig. 14.9. A plane boundary exists between two dielectric (insulators) media having refractive indices n_1 and n_2, respectively. Because the materials are insulators, no electric currents will flow. Glass and plastic materials used



Perpendicular Polarization (s)

Parallel Polarization (p)

Figure 14.9 Reflection at a plane boundary. Perpendicular and parallel polarizations are illustrated.

in fiber-optic waveguides satisfy the assumptions of this model. The incident ray in medium 1 strikes the boundary at an angle θ_i as measured with respect to the boundary normal. There is a reflected wave at angle θ_r and a transmitted wave at angle θ_t. The angle of reflection equals the angle of incidence

$$\theta_r = \theta_i \tag{14.27}$$

and the transmitted angle is determined by Snell's law

$$\frac{\sin\theta_t}{\sin\theta_i} = \frac{n_1}{n_2} \tag{14.28}$$

Figure 14.10 illustrates what happens when a light ray travels from a lower index to a higher index material. The ray bends toward the boundary normal. Figure 14.11 illustrates what happens when a light ray travels from a higher index to a lower index material. The ray bends away from the boundary normal. A particularly interesting result occurs when

$$\sin\theta_i = \frac{n_2}{n_1} \tag{14.29}$$

At this angle, we find that *sin* $\theta_t = 1$, so that θ_t itself becomes 90°. This means that the energy from the incident beam does not penetrate into the second medium. The incident angle at which this occurs is called the *critical angle*, θ_c. Clearly, a critical angle exists only if $n_1 > n_2$, because the sine function must be unity or less for a solution to exist.

As can be seen from Snell's law, for all incident angles greater than the critical angle, there is no solution for a transmitted angle. This means that for all angles of incidence equal to or greater than the critical angle there will be no transmitted wave. This condition

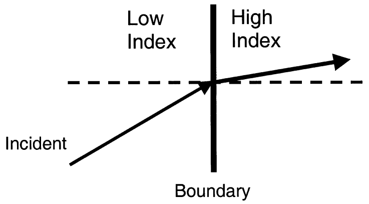


Figure 14.10 Ray bending when a wave travels from a low-index to a high-index material.

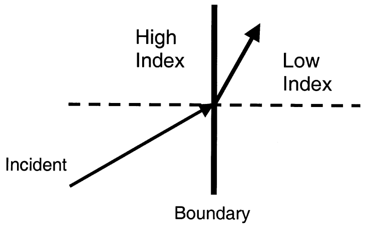


Figure 14.11 Ray bending when a wave travels from a high-index to a low-index material.

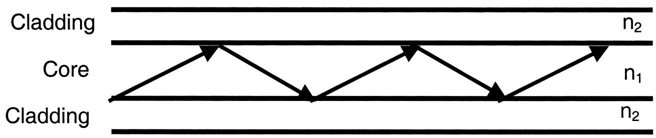


Figure 14.12 Wave guiding by total internal reflection.

is called *total internal reflection*. It is the basis of wave guiding by an optical fiber, as illustrated in Fig. 14.12. In the figure, the central region is the *core* (having refractive index n_1) and the outer material is the *cladding* (having refractive index n_2). The light stays inside the glass fiber structure by continual reflection at the boundaries. Note that this can only occur (total reflection) if the inner material has a higher refractive index than the outer material so that a critical angle exists.

14.2.4. Reflection at a Boundary

The preceding section considered the problem of reflection at a plane boundary from a ray approach. The only results generated by this type of analysis are the directions of the reflected and transmitted beams. A more complete solution of the problem is based upon a wave analysis and predicts the fraction of light reflected and transmitted as well as the angles of reflection and transmission.

In the wave analysis, the electric and magnetic fields in the two regions are written and the electromagnetic conditions of continuity of the fields at the boundary are applied. The results of this type of analysis are presented in the following paragraphs.

The *plane of incidence* is the plane defined by the normal to the boundary and the direction of travel of the incident wave. That would be the plane of the page in Fig. 14.9. For the problem under consideration, the amount of reflection depends upon the polarization of the incident field. Recall that the two orthogonal linearly polarized modes of a plane wave lie in a plane that is perpendicular to the direction of travel. The corresponding electric fields are either polarized perpendicular to the plane of incidence (this is called s polarization) or parallel to the plane of incidence (this is called p polarization).

The *reflection coefficient* ρ is defined as the ratio of the reflected electric field to the incident electric field when they are written in complex form. The results are known as *Fresnel's laws of reflection.*

For parallel polarization, the reflection coefficient is

$$\rho_p = \frac{-n_2^2 \cos\theta_i + n_1\sqrt{(n_2^2 - n_1^2 \sin^2\theta_i)}}{n_2^2 \cos\theta_i + n_1\sqrt{(n_2^2 - n_1^2 \sin^2\theta_i)}} \tag{14.30}$$

For perpendicular polarization, the reflection coefficient is

$$\rho_s = \frac{n_1 \cos\theta_i - n_1\sqrt{(n_2^2 - n_1^2 \sin^2\theta_i)}}{n_1 \cos\theta_i + n_1\sqrt{(n_2^2 - n_1^2 \sin^2\theta_i)}} \tag{14.31}$$

The reflection coefficients give us the relationship between the incident and reflected electric fields. Because the power is proportional to the square of the field, the fractional reflected power (called the *reflectance*) is determined by taking the magnitude of the square of the reflection coefficient. Thus, the reflectance R is

$$R = |\rho|^2 \tag{14.32}$$

Clearly, the fraction of transmitted power is

$$1 - R = 1 - |\rho|^2 \tag{14.33}$$

For the case of normal incidence ($\theta_i = 0$), both cases of polarization reduce to

$$R = \left(\frac{n_1 - n_2}{n_1 - n_2}\right)^2 \tag{14.34}$$

For an air-to-glass interface ($n_1 = 1$ and $n_2 = 1.5$), the result is $R = 0.04$. In this case, 4% of the light is reflected and 96% of the light is transmitted. Reflectance plots are shown in Figs. 14.13 and 14.14 for air-to-glass and glass-to-air interfaces, respectively. In the first case, the wave goes from a region of lower refractive index to one that is higher. In the second case, the wave goes from a region of higher refractive index to one that is lower. These figures illustrate a few interesting points upon which we will elaborate.

For small angles of incidence, the reflectance does not vary much with a change in the incident angle.

At a certain incident angle, the reflectance is zero for the case of parallel incidence. The angle at which zero reflection occurs is called the *Brewster angle* and is found from Eq. (14.30) to be

$$\tan\theta_B = \frac{n_2}{n_1} \tag{14.35}$$

At the Brewster angle, all the light is transmitted.

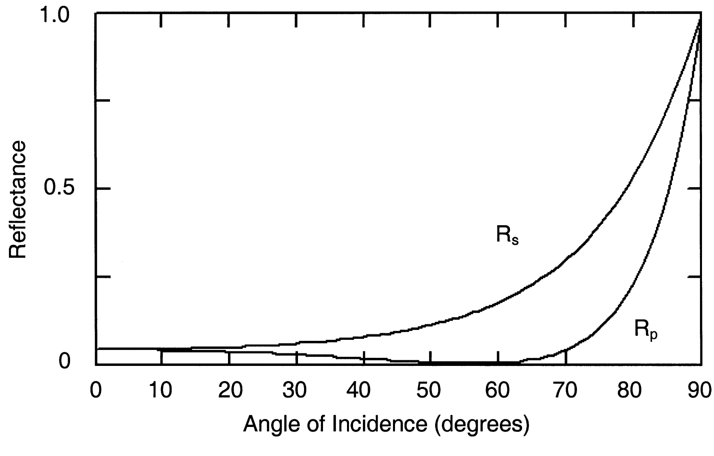


Figure 14.13 Reflectance for an air-to-glass interface.

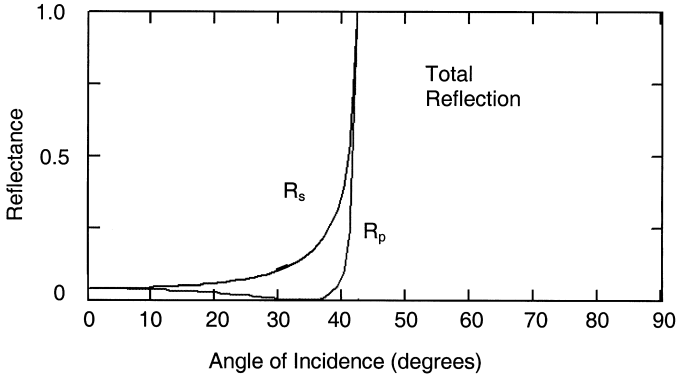


Figure 14.14 Reflectance for a glass-to-air interface.

As we determined earlier, all the light is reflected when incident at, or beyond, the critical angle. This shows up on Fig. 14.14 where the wave is going from a high-index to a low-index region. The figure shows that for incident angles equal to or greater than the critical angle all the light is reflected. The critical angle is calculated from either of Eqs. (14.30) and (14.31) by setting the term under the square root sign to zero. When that is done, $\rho_p = -1$ and $\rho_s = 1$. The reflectance is unity in either case. The critical angle condition is thus

$$n_2^2 - n_1^2 \sin^2 \theta_i = 0 \tag{14.36}$$

The incident angle that satisfies this equation is the critical angle. The solution is

$$\sin \theta_c = \frac{n_2}{n_1} \tag{14.37}$$

just as we found before in Eq. (14.29).

For angles greater than θ_c, we find that $n_1 \sin \theta_i > n_2$, making the term under the square root sign negative and making the second term in the numerator and denominator of both reflection coefficient equations imaginary. Both equations are then of the form

$$|\rho| = \left| \frac{A - jB}{A + jB} \right| \tag{14.38}$$

where A and B are real numbers and j is the imaginary term $j = \sqrt{-1}$. The magnitude of this complex number is unity. We conclude that the reflectance is unity for all angles of incidence equal to or greater than the critical angle. Again, this is the principle of the fiber waveguide. All rays that strike the core-cladding boundary at angles that are equal to or greater than the critical angle are bound to the core.

When we have total internal reflection, you might think there would be no electric field in the second medium. This is not the case, however. The boundary conditions require that the electric field be continuous at the boundary (that is, the field in region 1 and that in region 2, as measured at the boundary, must be equal). The exact solution shows a finite field in medium 2 that decays exponentially away from the boundary and carries no power into the second medium. This is called an *evanescent* field. It is not unlike the field that surrounds an inductor carrying a sinusoidal time-varying current. The stored energy is $Li^2/2$, where L is the inductance and i is the peak current. A magnetic field exists in the region around the inductor (having this same energy), but the energy does not flow away from the inductor. The energy can be captured by the circuit by discharging the inductor.

The field in medium 2 has a decay away from the boundary given by

$$E \propto e^{-\alpha z} \tag{14.39}$$

where the attenuation factor is

$$\alpha = k_0\sqrt{n_1^2 \sin^2\theta_i - n_2^2} \tag{14.40}$$

and k_0 is the free-space propagation factor. As can be seen, α is zero at the critical angle and increases as the incident angle increases beyond the critical angle. Because α is so small near the critical angle, the evanescent fields penetrate deeply beyond the boundary but do so less and less as that angle increases.

In medium 1, the reflected field interferes with the incident field to produce a standing wave. The envelope of this standing wave and the evanescent wave are pictured in Fig. 14.15.

This chapter is introducing, in a step-wise manner, the necessary electromagnetic fundamentals upon which fiber-optic communications is built. At this point, we can understand how a wave can be trapped in the fiber core due to total internal reflection at the core-cladding boundary, that a standing wave will exist in the fiber core, and that an evanescent field will exist in the fiber cladding.

14.2.5. Gaussian Beams

The electric field of a plane wave, such as that given by Eq. (14.8), has an amplitude that is the same over any plane given by $z = \text{constant}$. We call this a *uniform* plane wave. While this type of wave does not exactly exist in nature, it is a good approximation in many cases. For example, the radiation from a finite-sized light source (such as a LED or laser diode) spreads out as the observer moves away from the source. At distances

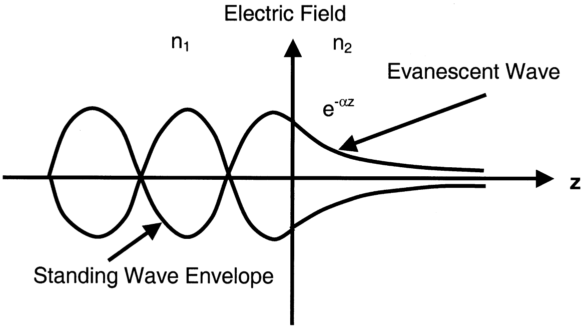


Figure 14.15 Electric field amplitudes near a reflecting boundary.

far from a source the beam pattern is nearly uniform over a small region near the axis of propagation, approximating a uniform plane wave.

A more common beam distribution is the *gaussian* beam. The gaussian beam is often generated by a laser and is the beam pattern in some fibers. The intensity of a gaussian distribution is given by

$$I = I_0 e^{-2(r/w)^2} \tag{14.41}$$

Intensity is proportional to the power in the wave. Technically, the intensity is the magnitude of the square of the electric field. This beam distribution is plotted in Fig. 14.16. It is cylindrically symmetric. The radial distance from the origin is r, I_0 is the peak intensity (it occurs in the center at $r = 0$), and the term w is called the *spot size*. It is the radial distance at which the intensity decreases to $1/e^2 = 0.135$ of its peak value I_0. That is, when $r = w$, then $I/I_0 = 1/e^2$.

A picture of the gaussian field pattern appears in Fig. 14.17. Since the field is circularly symmetric, the beam appears to be a circular spot of light. The electric field of a gaussian plane wave traveling in the z direction could be written as

$$E = E_0 e^{-r^2/w^2} e^{-\alpha z} e^{j(\omega t - kz)} \tag{14.42}$$

The intensity is

$$I = EE^* = E_0^2 e^{-2\alpha z} e^{-2r^2/w^2} \tag{14.43}$$

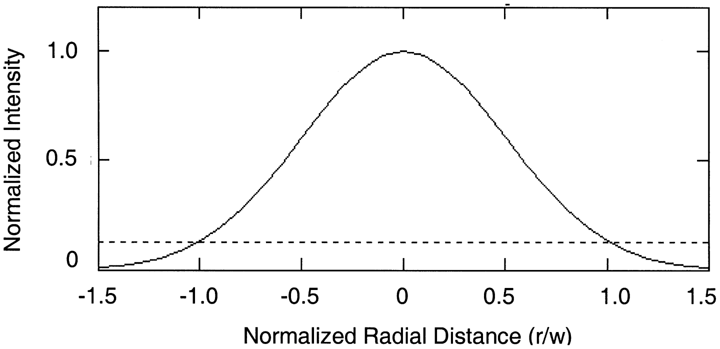


Figure 14.16 Gaussian intensity distribution.

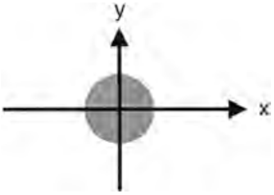


Figure 14.17 Gaussian transverse light pattern.

where E^* is the complex conjugate of E. The result is as expected, where we recognize $E_0^2e^{-2\alpha z}$as the peak intensity at the center of the beam for any position z and E_0^2 as the peak intensity at the origin ($r = 0$, $z = 0$).

14.2.6. Electromagnetic Cavity

A simple electromagnetic cavity is drawn in Fig. 14.18. Its analysis reveals a number of facets of operation of optical devices including the fiber. For simplicity, we consider the propagation of plane waves between infinitely extended perfect mirrors. The waves move along the cavity axis (the z axis) back and forth between the two reflecting mirrors. Assuming the electric fields are linearly polarized in the y direction as indicated on the figure, the forward- and backward-traveling waves are given (respectively) by

$$E_+ = E_1 e^{-jkz} \tag{14.44a}$$

$$E_- = E_2 e^{jkz} \tag{14.44b}$$

where the time variation $e^{j\omega t}$ has been suppressed. The total field at any point in the cavity is then

$$E = E_1 e^{-jkz} + E_2 e^{jkz} \tag{14.45}$$

To be perfectly reflecting, the mirrors must have infinite conductivity. The electromagnetic boundary conditions in such a case require that the total field be zero at the mirrors. That is,

$$E(z = 0) = 0 \tag{14.46a}$$

$$E(z = L) = 0 \tag{14.46b}$$

Applying the first of these conditions leaves

$$0 = E_1 + E_2 \tag{14.47a}$$

or

$$E_2 = -E_1 \tag{14.47b}$$

We see that the two opposing fields must be equal in magnitude. The minus sign indicates a 180° phase shift at the first mirror needed to make the fields cancel at that mirror.

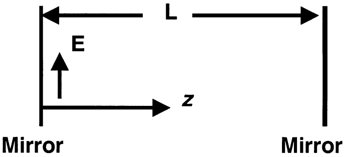


Figure 14.18 Electromagnetic cavity.

The total field can now be written as

$$E = E_1 e^{-jkz} - E_1 e^{jkz} \tag{14.48}$$

or

$$E = 2jE_1 \sin kz \tag{14.49}$$

Applying the boundary condition at the second boundary yields

$$0 = 2jE_1 \sin kL \tag{14.50}$$

The conclusion is that

$$0 = \sin kL \tag{14.51}$$

requiring that

$$kL = m\pi \tag{14.52}$$

where m is a positive integer. Since we know that $k = 2\pi/\lambda$, this becomes

$$L = \frac{m\lambda}{2} \tag{14.53}$$

This is a classic result called the *resonance condition*. It states that the only waves that can exist in the steady state within the cavity are those for which the cavity is an integral number of half wavelengths long. The wavelengths satisfying this result are said to be the *resonant wavelengths* of the structure. In fact, they are the wavelengths for which the interference between the forward and backward waves is constructive. At any point in the cavity, the two waves (satisfying the resonance condition) always have the same phase relationship with respect to each other. This is *constructive interference*. Fields at wavelengths not satisfying this condition interfere *destructively*. Their relative phase difference changes for each pass across the cavity. The result at any point in the cavity is the summation of a large number of randomly phased waves. The sum of such a sequence of fields is zero.

The *resonant frequencies* corresponding to the resonant wavelengths in Eq. (14.53) are given by

$$f = \frac{mc}{2nL} \tag{14.54}$$

where n is the refractive index of the material filling the cavity. A picture of the cavity resonant frequencies appears in Fig. 14.19. The different frequencies that can exist within the cavity are the allowed modes of the cavity.

The total field in the cavity can be written in the simplified form

$$E = E_0 \sin kz \tag{14.55}$$

This represents a standing wave pattern within the cavity. The envelope of this wave is drawn in Fig. 14.20.

The cavity problem is significant for a number of reasons. One is that it is an easily solvable electromagnetic boundary value problem. The solution strategy is to write electric fields that are solutions to the wave equation and that can ultimately be made to satisfy the boundary conditions. In fact, it can be proven that if a field satisifies the wave equation and the boundary conditions of a structure, it is a valid solution. Many electromagnetic boundary value problems are more complex, but they are solved with the same basic strategy. The problems of interest for optical communication are the dielectric slab waveguide and the fiber waveguide. These structures will be considered in later sections of this chapter.

Another reason for studying the resonant cavity is that it is the structure of the laser diode. The amplifying semiconductor fills the cavity. The cavity provides the feedback necessary to produce oscillations. The laser output will be at wavelengths where there is an amplification resonance as in Fig. 14.21. The amplification is indicated on the figure by the dashed curve. The distinct output wavelengths are the *longitudinal modes* of the device. The spectral width of the laser diode is $\Delta\lambda$, as also shown on the figure.

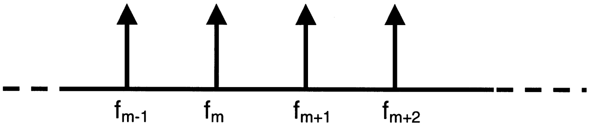


Figure 14.19 Cavity-resonant frequencies.

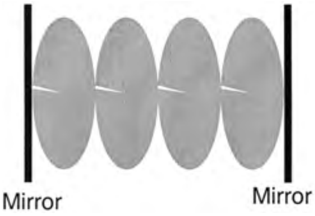


Figure 14.20 Envelope of the cavity standing-wave pattern.

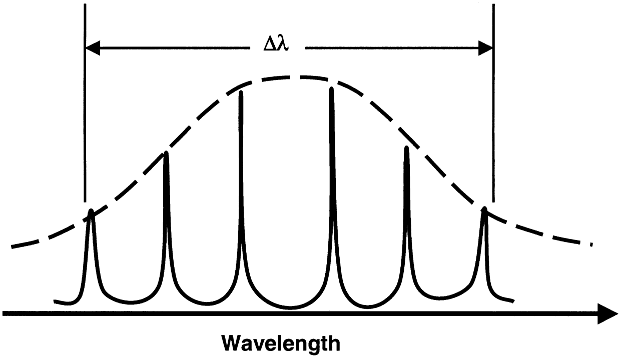


Figure 14.21 Laser diode output spectrum.

14.3. INTEGRATED OPTICS

In this section we describe the fundamentals of integrated optics. Integrated optics is the technology of constructing optical components on substrates [14–18]. Components that have successfully utilized integrated optics include directional couplers, beam dividers, modulators, phase shifters, and switches.

The study of electromagnetic propagation in the integrated optic structure parallels that of propagation in the fiber. The analysis is simpler to do for integrated optics because of its rectangular geometry as compared to the circular geometry of the fiber structure. Despite the different geometries, the integrated optic analysis to follow tells us a great deal about propagation in the fiber.

14.3.1. Slab Waveguide

The slab waveguide is drawn in Fig. 14.22. It consists of three layers of dielectric materials, having refractive indices n_1, n_2, and n_3. The middle layer is the guiding region and has the largest index of refraction. From our earlier description of total internal reflection, it is apparent how this structure guides optical waves. Rays at, or beyond, the critical angle are reflected at the upper and lower boundaries and cannot escape the structure. The wave zigzags down the waveguide as indicated on the figure. As we expect from our earlier discussion of total internal reflection, evanescent fields in the upper and lower regions exist and will travel along with the wave propagating in the middle guiding layer.

When (as is often the case) the central layer thickness (d) is very small, this layer is referred to as a *film* (or a *thin film*). The critical angles at the lower and upper boundaries, respectively, are given by

$$\sin\theta_{c12} = \frac{n_2}{n_1} \tag{14.56}$$

$$\sin\theta_{c13} = \frac{n_3}{n_1} \tag{14.57}$$

For complete guiding, the ray angles must be equal to, or greater than, the largest of these two critical angles calculated above. Otherwise the wave would not undergo total reflection at one of the two boundaries, and optical energy would escape from the structure.

For integrated optic devices, the upper and lower materials are usually different. In fact, the upper region is commonly air ($n_3 = 1$). This type of structure is *asymmetrical.*

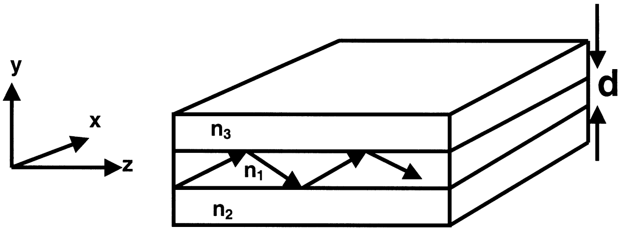


Figure 14.22 The dielectric slab waveguide.

If the upper and lower materials are the same ($n_3 = n_2$), the structure is *symmetrical.* The study of the symmetric waveguide most nearly parallels that of the circularly symmetric fiber. The electromagnetic solution of this structure follows the same strategy as that for the electromagnetic cavity: i.e., write equations for fields in the three regions that satisfy the wave equation and apply the boundary conditions. The details are more complicated however. We will sketch the solution, but without going through all the specifics.

As we did in the case of reflection at a plane boundary, we divide the problem into two possible linear polarizations. In Fig. 14.23, the yz plane is the plane of incidence. The perpendicular polarization (s) has the electric field pointing in the x direction. The electric field points in a direction perpendicular to the plane of incidence. We call this *transverse electric* (TE) polarization because the electric field always points transverse to the direction of net travel (the z direction).

The other possible polarization has the electric field parallel to the plane of incidence (p polarization) as indicated in Fig. 14.24. In this case, the magnetic field (which lies perpendicular to both the local direction of travel and the electric field) is polarized in the x direction. It is the magnetic field that now points transverse to the z direction. This is transverse magnetic (TM) polarization.

14.3.2. TE Mode Chart

Consider a TE wave in a symmetrical waveguide. The electric field points in the x direction. The field in the central region is made up of the superposition of two plane waves, one moving upward at angle θ and one moving downward at that angle. The equation for the upward-traveling plane wave in region 1 is given by

$$E_+ = 0.5E_0 e^{j(\omega t - k_1 y \cos\theta - k_1 z \sin\theta)} \tag{14.58}$$

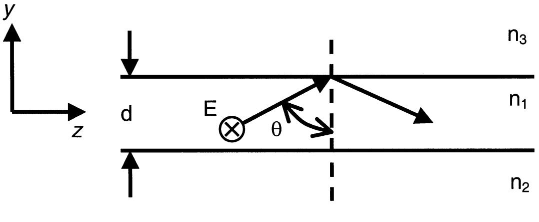


Figure 14.23 TE polarization in the slab waveguide.

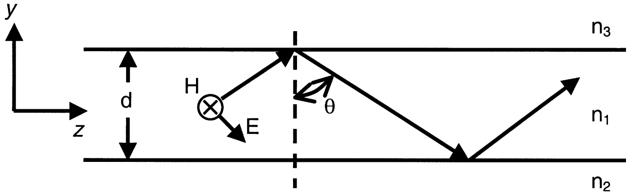


Figure 14.24 TM polarization in the slab waveguide.

while the downward wave is given by

$$E_{-} = 0.5E_0 e^{j(\omega t + k_1 y \cos\theta - k_1 z \sin\theta)} \tag{14.59}$$

where k_1 is the propagation coefficient in the middle layer. The amplitudes of the two waves are the same due to total reflection at the boundaries, just as we found for the fields at the mirrors in the cavity problem.

The total field in the guiding region is the sum of the preceding two waves. Adding the two waves and simplifying yields

$$E_1 = E_0 e^{j\omega t} e^{-jk_1 z \sin\theta} \cos(k_1 y \cos\theta) \tag{14.60}$$

We further simplify by defining

$$h = k_1 \cos\theta \tag{14.61}$$

and

$$\beta = k_1 \sin\theta \tag{14.62}$$

yielding

$$E_1 = E_0 e^{j(\omega t - \beta z)} \cos hy \tag{14.63a}$$

This field is symmetrical in the transverse (xy) plane, as indicated by the cosine function. That is, the field pattern is an even function of y. A field with odd symmetry can also exist. It is given by

$$E_1 = E_0 e^{j(\omega t - \beta z)} \sin hy \tag{14.63b}$$

These equations represents a nonuniform plane wave traveling in the z direction. It is nonuniform because it varies in the transverse plane in the manner indicated by the cosine or sine terms. By comparing this result with the field for the uniform plane wave, we note that β is the effective propagation factor. The pattern of the wave in the transverse plane is a standing wave (as we might have expected) because it is made up of two interfering plane waves.

The field in the region above the middle layer must travel in the z direction at the same speed as the central field. It must also have a term indicating the decaying evanescent wave. A field satisfying this condition is

$$E_2 = A_2 e^{-\alpha(y - d/2)} e^{j(\omega t - \beta z)} \tag{14.64}$$

The last term indicates that the wave travels with respect to the z axis at the same speed (same propagation factor) as does the field in the central region. This field decays away from the central region with attenuation factor α. The field amplitude is A_2 at the boundary ($y = d/2$).

If we substitute this last equation into the wave equation, we find that α must be

$$\alpha = k_0 \sqrt{n_1^2 \sin^2\theta - n_2^2} \tag{14.65}$$

just as we found earlier when considering the evanescent wave in the transmitted region for the problem of plane wave reflection at a plane boundary.

The electromagnetic problem now reduces to finding the relative amplitudes of the waves and the allowed values of β, the propagation factor. These are found by applying the boundary electromagnetic conditions, continuity of the tangential electric field. Because of the somewhat complicated structure, the z component of the magnetic fields must also be found and matched at the boundaries. The magnetic fields can be found from the electric fields since they are related by Faraday's law. The results of applying the boundary conditions are

$$A_2 = E_0 \cos\frac{hd}{2} \tag{14.66}$$

so that the evanescent field is now

$$E_2 = E_0 \cos\frac{hd}{2} e^{-\alpha(y-d/2)} e^{j(\omega t-\beta z)} \tag{14.67}$$

Note that E_1 and E_2 are now equal at the boundary. In addition, the boundary conditions yield

$$\tan\frac{hd}{2} = \frac{1}{n_1 \cos\theta}\sqrt{(n_1 \sin\theta)^2 - n_2^2} \tag{14.68}$$

This is a transcendental equation that must be solved graphically or numerically. It is called the *characteristic equation* or the *mode equation*. Solving it reveals the allowed propagation angles θ for a given central film thickness d.

An example plot of the mode equation appears in Fig. 14.25 for the case where $n_1 = 3.6$ and $n_2 = 3.55$. Note that there are multiple solutions because the tangent function repeats itself. That is, for a given value of the right-hand side of the mode equation, there is an infinite set of solutions for $hd/2$.

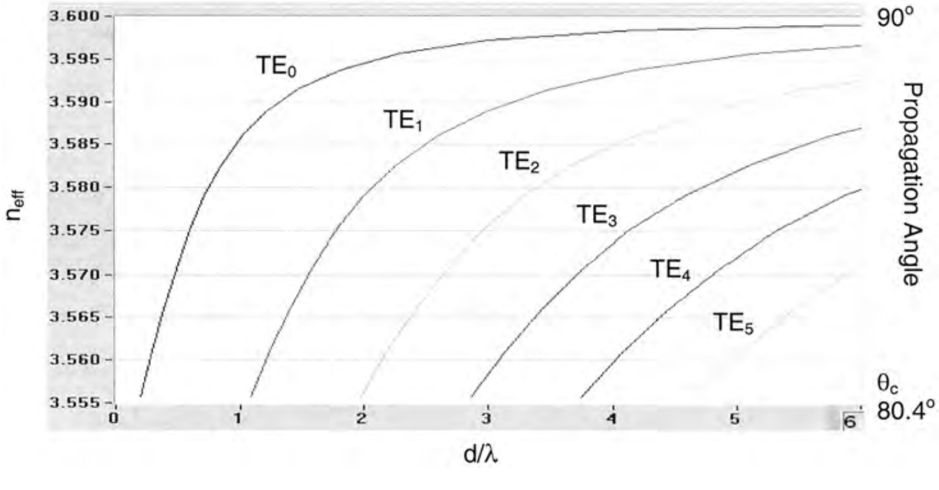


Figure 14.25 Mode chart for a symmetrical slab waveguide with $n_1 = 3.6$ and $n_2 = 3.55$.

The figure itself is called a *mode chart*. The horizontal axis represents the normalized film thickness d/λ. The vertical axis on the right refers to the allowed propagation angle θ. Because the critical angle for this structure is

$$\theta_c = \sin^{-1}\frac{n_2}{n_1} = \sin^{-1}\frac{3.55}{3.6} = 80.4^\circ$$

the range of angles allowing propagation is between the 80.4° and 90° as indicated on the mode chart. The critical angle is also called the *cutoff angle*, because rays that strike the interface at lesser angles cannot propagate (they are cut off).

The vertical axis on the left is the *effective index of refraction*, defined as

$$n_{\mathrm{eff}} = n_1 \sin\theta \tag{14.69}$$

When the ray angle is 90° (an axial ray), $n_{\mathrm{eff}} = n_1$ and when the angle is equal to the critical angle, $n_{\mathrm{eff}} = n_2$. We see that the range of the effective index of refraction lies between the indices of the two waveguide materials. This is also indicated on the mode chart.

The modes refer to the different ray angles allowed for a fixed film thickness. For example, when $d/\lambda = 2$ the mode chart (Fig. 14.25) shows that three different propagation angles are allowed. The corresponding modes, effective indices of refraction, and ray angles are given in Table 14.4. As indicated, TE_0, TE_1, and TE_2 modes can all travel simultaneously. This represents a multimode waveguide. If the operating wavelength were 1.55 μm in this example, the film thickness would be 3.1 μm. The subscript m in the mode designation (TE_m) is called the *mode order*. The order of the lowest ordered mode is zero in the slab waveguide.

A single-mode waveguide will exist if the film is thin enough. For the structure in this example, if $d/\lambda < 0.836$, the only TE mode that can propagate is the TE_0 mode. The *cutoff condition* is

$$\frac{d}{\lambda} \le \frac{1}{2\sqrt{n_1^2 - n_2^2}} \tag{14.70}$$

For values of thickness that satisfy this inequality, the integrated waveguide is single mode, allowing propagation of only the $m = 0$ mode.

The reason that only specific ray angles are allowed has to do with the interference between the upward- and downward-traveling waves. Only for certain angles will the interference be constructive. These are the allowed angles, or modes, of the waveguide. This is the same phenomenon discussed in the description of the electromagnetic cavity. They appear automatically when the electromagnetic problem is solved explicitly by finding fields that satisfy the wave equation and the boundary conditions. The number of

Table 14.4 Modes in the Slab Waveguide

Mode	n_{eff}	Ray angle
TE_0	3.595	87
TE_1	3.58	84
TE_2	3.557	81

propagating modes is the integer part of

$$N = 1 + \frac{2d\sqrt{n_1^2 - n_2^2}}{\lambda} \tag{14.71}$$

The number of modes that can propagate decreases as the guiding layer thickness gets smaller and as the two indices of refraction get closer to each other.

As illustrated, a mode chart displays many of the propagation characteristics of a wave-guiding structure. A similar mode chart will be shown for the fiber-optic waveguide to be discussed later in this chapter.

14.3.3. TM Mode Chart

Next consider the mode chart for TM waves. The mode equation for this case is

$$\tan\frac{hd}{2} = \frac{n_1}{n_2^2 \cos\theta}\sqrt{(n_1 \sin\theta)^2 - n_2^2} \tag{14.72}$$

A mode chart that includes both TE and TM modes appears in Fig. 14.26 using the same values of refractive indices as in the previous example. Because the two indices are so close, both mode equations yield nearly identical results. That is why the TE and TM modes appear to be the same. If two or more modes share the same propagation characteristics, they are *degenerate*. If the two refractive indices were quite different, the TE and TM mode curves would separate from each other and the modes would no longer be degenerate.

In the degenerate case, a single-mode waveguide satisfying the cutoff condition in Eq. (14.70) actually sustains two modes. Both the TE_0 and TM_0 modes propagate, but with the same effective propagation factors. The total number of allowed modes (including both TE and TM) is twice that calculated for the TE case alone.

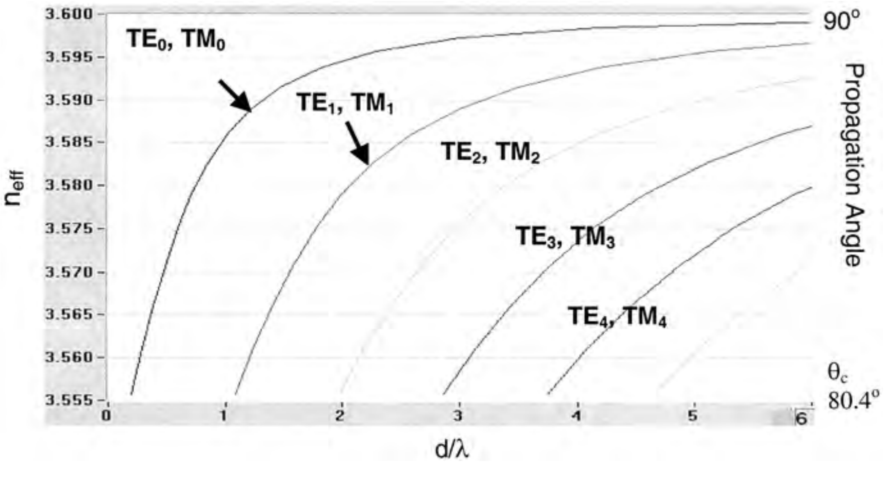


Figure 14.26 Mode chart showing both TE and TM modes.

14.3.4. Mode Field Patterns

The light distribution in the transverse plane is the *transverse mode pattern.* These patterns are particularly important when designing components that connect to, or are built within, the integrated optic structure. The field distributions of all components must match closely to avoid losses.

For the TE mode in the symmetrical waveguide the transverse pattern is given within the film by the term

$$E_1 \propto E_0 \cos hy \tag{14.73}$$

and outside the film by

$$E_2 \propto E_0 \cos\frac{hd}{2} e^{-\alpha(y-d/2)} \tag{14.74}$$

The standing wave patterns are plotted in Fig. 14.27 for the four lowest-ordered modes. At any point in the transverse plane the actual electric field amplitude is oscillating at a frequency on the order of 10^{14} Hz. The standing wave pattern is the envelope of the field amplitude variation. For the slab waveguide, the mode order is the number of zero crossings in the standing wave pattern. If we were to view projections of the various mode patterns using visible light we would see a single central spot for the TE_0 mode, two spots for the TE_1 mode, three spots for the TE_2 mode and so on.

The evanescent fields outside the central guiding layer are also indicated on the figure. As the mode order *m* increases, the attenuation factor α decreases and the wave penetrates further into the outer layers. Higher ordered modes travel with ray angles closer to the critical angle than do lower ordered modes. As pointed out in the discussion on reflection from a plane boundary, the attenuation factor decreases as the critical angle is approached accounting for the increased wave penetration.

We can now expand our mode definition. We have been saying that modes refer to the different propagation paths allowed in a waveguide. Now we can also say that modes refer to the different transverse field patterns that are possible in a waveguide.

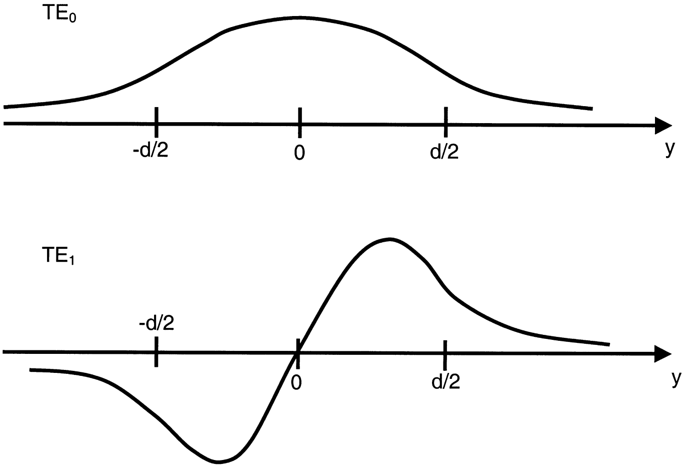


Figure 14.27 Transverse mode patterns in the symmetric slab waveguide.

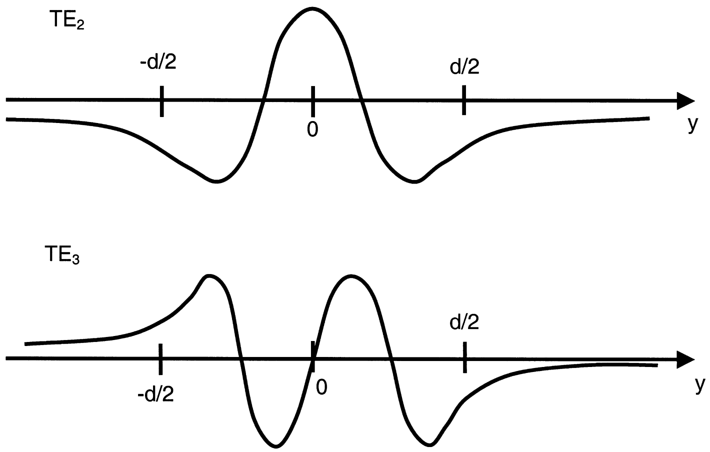


Figure 14.27 Continued.

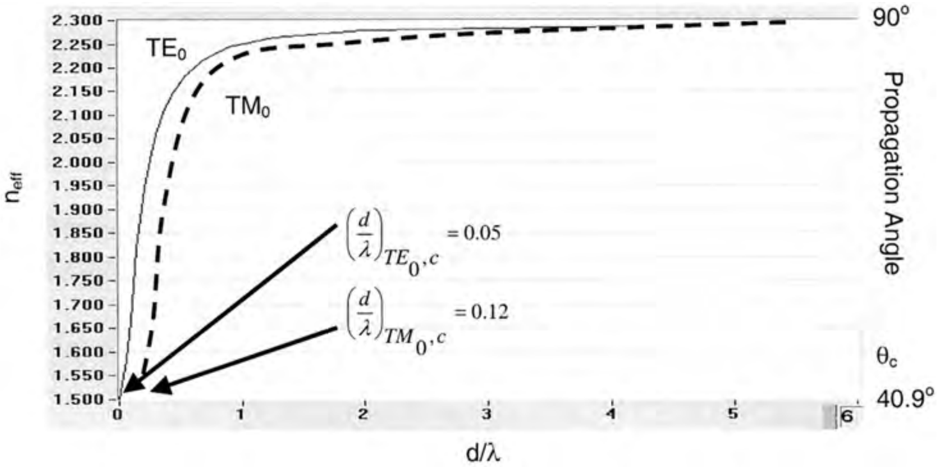


Figure 14.28 Mode chart for an asymmetric slab waveguide.

14.3.5. Asymmetric Waveguide

The equations and the solution for the asymmetric waveguide are more complicated than that for the symmetric waveguide. Here we will simply show a mode chart indicate some of the features of propagation. A mode chart appears in Fig. 14.28 for the case of a zinc sulfide (ZnS) film deposited onto a glass substrate. Air covers the space above the film. Thus, $n_1 = 2.29$, $n_2 = 1.5$, and $n_3 = 0$. The critical angle at the air–ZnS interface is 25.9° and at the glass–ZnS interface it is 41°. Propagation only occurs when the ray angles are greater than the largest of these two values, 41°; otherwise, light will leak from the ZnS film into the glass substrate. Thus, the range of propagating angles is from 41° to 90°. The corresponding range of effective refractive indices (recall that $n_{\text{eff}} = n_1 \sin\theta$) is from n_1 to n_2 (that is, 1.5 to 2.29).

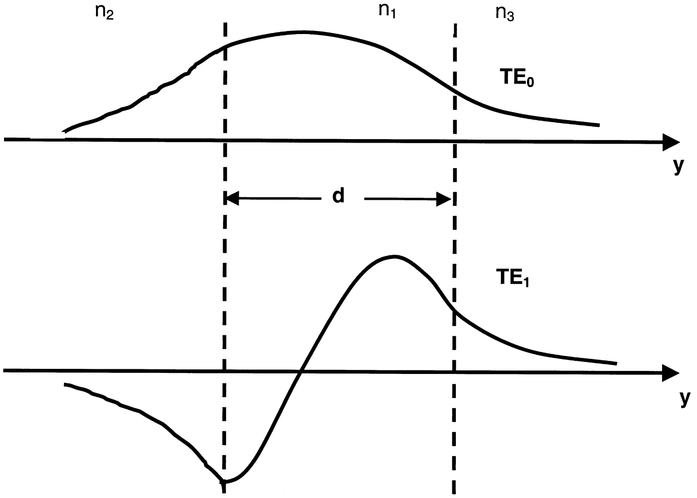


Figure 14.29 Transverse mode patterns in the asymmetric slab waveguide.

The mode chart shows both TE and TM modes. Because the three indices of refraction are not close, the modes are not degenerate. The TE and TM modes are clearly separate. As found from the mode chart, truly single-mode propagation exists if $d/\lambda < 0.12$ for this structure. This represents the cutoff condition for the TM_0 mode. Only the TE_0 mode can propagate in this case.

The mode patterns are similar to those of the symmetric waveguide except for the lack of symmetry. This is indicated in Fig. 14.29 for a few lower order modes.

14.3.6. Modal Distortion

Earlier in this chapter material dispersion was described as the cause of pulse spreading, ultimately limiting the information capacity of the transmission line. Another cause of pulse spreading can now be described. As illustrated by the mode chart, the different modes travel with different effective indices of refraction and thus with different velocities with respect to the waveguide axis. An input pulse will distribute its energy among all the allowed modes. Because of the different mode velocities, parts of the wave arrive ahead of (or behind) other parts. The result is that the pulse at the receiver is wider than that originally transmitted. Once again we have pulse spreading. This is called *modal distortion* or *modal dispersion*. The limitations on bandwidth and data rate previously given apply regardless of the cause of the spreading.

A simple analysis allows calculation of the amount of spreading. The earliest arriving pulse will be that of the lowest order axial ray. Energy in this mode travels straight down the transmission line, a distance L. The last arriving pulse will be that of the highest order mode, traveling at the critical angle. A little geometry shows that this pulse travels a distance Ln_1/n_2. The difference in time of arrival between the fastest and slowest modes will be the pulse spread.

The axial ray travel time will be L/v, where $v = c/n_1$. Thus, for the axial ray

$$t = \frac{Ln_1}{c} \tag{14.75}$$

For the critical angle ray, length L is replaced by the zigzag path length Ln_1/n_2. The critical angle travel time is then

$$t = \frac{Ln_1^2}{cn_2} \tag{14.76}$$

The difference between the two arrival times is the pulse spread. Subtracting and simplifying yields

$$\Delta\left(\frac{\tau}{L}\right) = \frac{n_1(n_1 - n_2)}{cn_2} \tag{14.77}$$

If we define the *fractional refractive index* change as

$$\Delta = \frac{n_1 - n_2}{n_1} \tag{14.78}$$

then the modal pulse spread can be expressed as

$$\Delta\left(\frac{\tau}{L}\right) = \frac{n_1\Delta}{c} \tag{14.79}$$

To generalize, modal distortion can be minimized by designing a waveguide with materials having refractive indices which are close to each other. Notice that modal distortion is independent of the operating wavelength and it is independent of the spectral width of the light source. This is unlike material dispersion, which is highly dependent on wavelength and spectral width.

14.4. FIBER OPTICS

In this section we describe several types of optical fibers and their properties [19–25]. Coverage includes step-index and graded-index fibers and single-mode and multimode fibers. Properties of interest are the modes, attenuation, pulse distortion, and bandwidth limitations.

14.4.1. Step-Index Fiber

The *step-index* (SI) fiber (Fig. 14.30) consists of a central core having radius a and refractive index n_1, surrounded by a cladding having refractive index n_2. In order to have total internal reflection, the core index must be greater than that of the cladding. For analytical purposes, it is convenient to assume that the cladding is infinitely thick. This removes any problems associated with the outer boundary of the cladding. As we already know, there is a decaying evanescent field associated with total internal reflection. This field decays rapidly so that we might expect that effects of a finite cladding thickness are negligible. That is, the field at the outer edges of the cladding are so small there is no chance of interaction with any material placed around the cladding itself. Therefore, the infinite cladding assumption is reasonable.

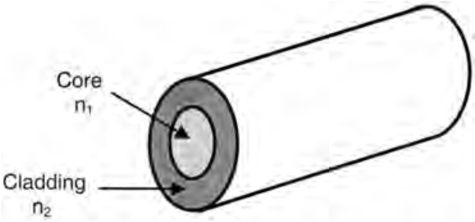


Figure 14.30 Step-index fiber.

Propagation in the step-index fiber is very much like propagation within the slab waveguide. Rays zigzag down the core, contained because of total internal reflection. Only discrete modes are allowed because of the requirement for constructive interference between the waves bouncing back and forth off the core-cladding interface. The fiber can transmit many modes if the core is large enough or can restrict transmission to a single mode if the core is small enough. A mode chart describes many of the propagation properties of the fiber. Material dispersion and modal distortion cause pulse spreading, affecting the fiber's ability to transmit unlimited bandwidths and data rates.

These general attributes are known (or at least expected) from the close analogy with the symmetrical slab waveguide. They also become evident from the electromagnetic solution of this boundary value problem. Unfortunately, the analysis is complicated by the circular symmetry of the fiber requiring the use of cylindrical coordinates. While the solution to the wave equation in rectangular coordinates consists of relatively simple trigonometric functions (sines, cosines, and exponentials), in cylindrical coordinates the solutions are Bessel functions.

The solution strategy is the same as described for the electromagnetic cavity and the slab waveguide. Functions for the electric field in the core and in the cladding that satisfy the wave equation in cylindrical coordinates are found. The boundary conditions are then applied. This leads to a characteristic equation from which the mode chart can be constructed.

A linearly polarized (LP) electric field pointing in the y direction can be written as

$$E_y = E_1 J_\ell\left(\frac{ur}{a}\right)(\cos \ell\phi)e^{j(\omega t - \beta z)} \tag{14.80}$$

in the core, and as

$$E_y = E_2 K_\ell\left(\frac{wr}{a}\right)(\cos \ell\phi)e^{j(\omega t - \beta z)} \tag{14.81}$$

in the cladding. In these equations J_ℓ is the Bessel function of the first kind of order ℓ, while K_ℓ is the modified Bessel function of the second kind of order ℓ. For simplicity, we have assumed a lossless transmission fiber. Otherwise, an exponential decay term (of the form $e^{-\alpha z}$) would need to be added to the electric field equations.

The terms u and w are given by

$$u = a\sqrt{n_1^2 k_0^2 - \beta^2} \tag{14.82a}$$

$$w = a\sqrt{\beta^2 - n_2^2 k_0^2} \tag{14.82b}$$

In these equations, k_0 is the propagation factor in free space.

We will be using a term called the *normalized frequency*, derived from u and w, which is given by

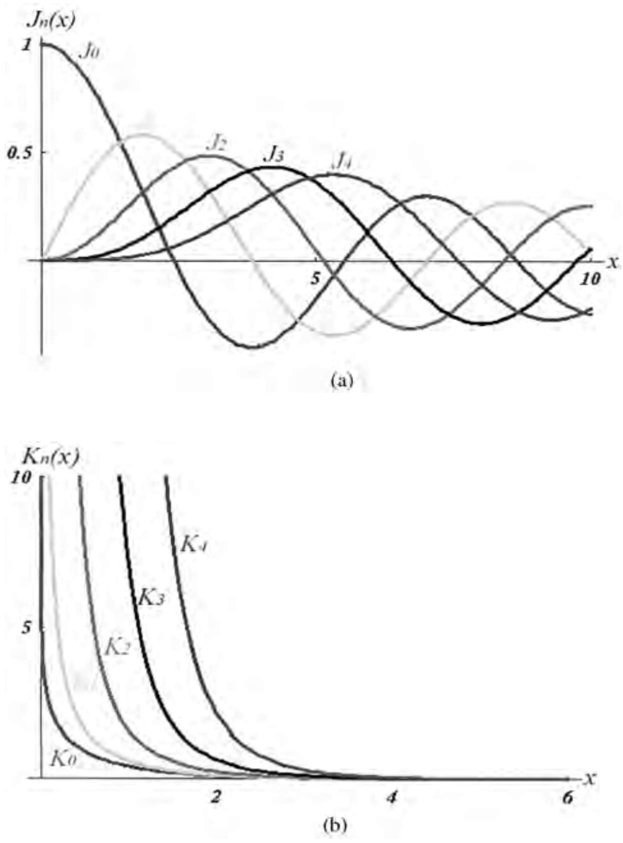
$$V = \sqrt{u^2 + w^2} \tag{14.83}$$

This can be rewritten as

$$V = \frac{2\pi a}{\lambda}\sqrt{n_1^2 - n_2^2} \tag{14.84}$$

It is sometimes simply called the *V parameter*.

The Bessel functions are solutions to the wave equation. The Bessel function J_ℓ resembles a sinusoid (appropriate for the field within the core). The modified Bessel function K_ℓ resembles an exponential decay (appropriate for the field in the cladding). Plots of these functions in Fig. 14.31 illustrate this behavior.



(a)

(b)

Figure 14.31 Bessel functions.

The following steps are taken in solving this problem. Faraday's law determines the magnetic fields from the electric fields, and Ampere's law determines the z component of the electric field from the magnetic field. Applying the boundary conditions to these fields yields the characteristic equation for the linearly polarized (LP) modes

$$u\frac{J_{\ell-1}(u)}{J_\ell(u)} = w\frac{K_{\ell-1}(w)}{K_\ell(w)} \tag{14.85}$$

Solving this equation yields the mode chart for the step-index waveguide. A few of the lowest-ordered modes are plotted in Fig. 14.32. If the V parameter of the fiber is known, the mode chart reveals the parameter $b = w^2/V^2$. From this, the propagation factor β can be determined from Eq. (14.82), the corresponding propagation angle from Eq. (14.62), and the corresponding effective refractive index from Eq. (14.69).

Several approximations were made in deriving the characteristic equation for the $\mathrm{LP}_{\ell m}$ modes. A more exact (and more complicated) approach yields the exact mode chart plotted in Fig. 14.33. Comparison shows that the LP approximation is reasonable and yields good results. The equivalent LP modes are indicated on the exact mode chart. We will continue the discussion with reference to the exact solution. Notice that if $V < 2.405$, only the lowest-ordered mode (HE_{11}) will propagate. This is the condition for design of a single-mode fiber. For larger values of V more than one mode propagates and we have a multimode fiber.

The single-mode fiber has the great advantage of eliminating modal distortion and, consequently, increasing the information capacity (rate length and frequency length). Almost all long-distance optical links use single-mode fibers because of the high capacities and lengths required. Multimode fibers are sufficient for shorter paths, such as used in LANs. The transverse pattern of the HE_{11} mode is nearly gaussian. It can be written as

$$E_y = E_0 e^{-(r/w)^2} \tag{14.86}$$

where the spot size w is given by

$$\frac{w}{a} = 0.65 + 1.619V^{-3/2} + 2.879V^{-6} \tag{14.87}$$

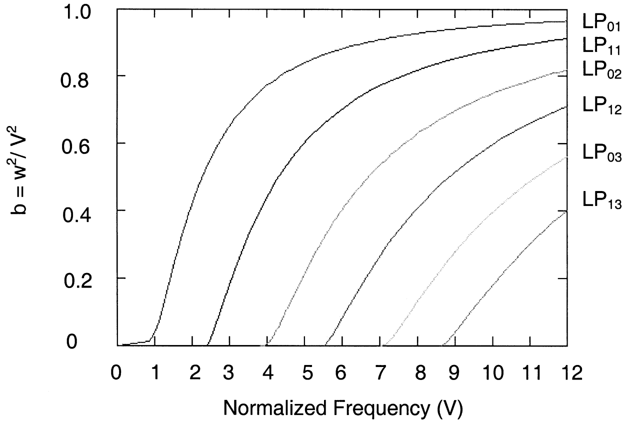


Figure 14.32 $\mathrm{LP}_{\ell m}$ mode chart for the step-index fiber.

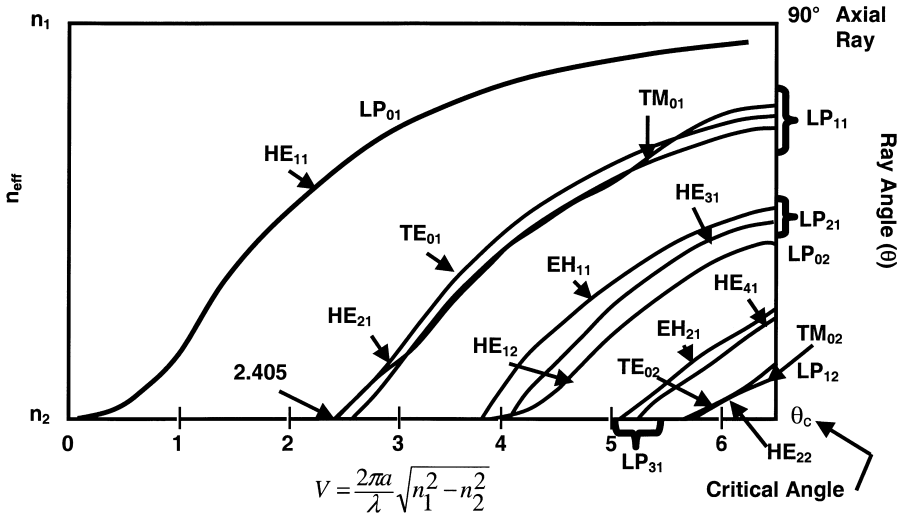


Figure 14.33 Exact mode chart for the step-index fiber.

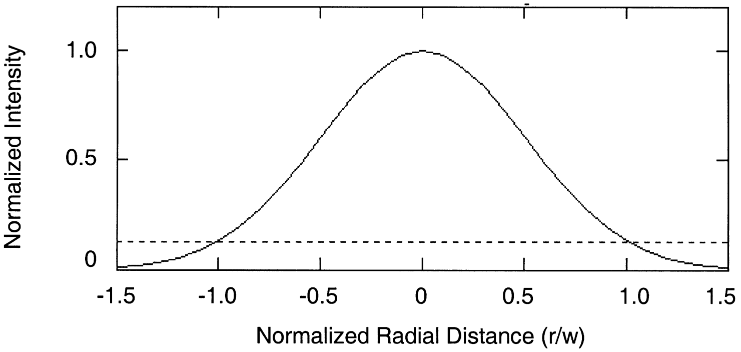


Figure 14.34 HE_{11} mode gaussian intensity distribution.

Because the intensity is the square of the electric field, it can be written as

$$I = I_0 e^{-2(r/w)^2} \tag{14.88}$$

A plot of the gaussian beam pattern appears in Fig. 14.34. The spot size variation with normalized frequency appears in Fig. 14.35. When V is close to 2.405, the spot size is only about 10% larger than the core radius. This implies that the energy in the beam is tightly bound to the core of the fiber. For smaller values of V, the spot size increases. This is undesirable as the energy is no longer tightly bound to the core. In this situation, energy can penetrate deeply into the cladding at bends in the fiber, eventually radiating out the sides. Best operation of a single-mode fiber has V in the range from 2.0 to 2.2.

Modal distortion in the multimode step-index fiber can be treated in exactly the same way as was done for the slab waveguide. The total pulse spread in a highly

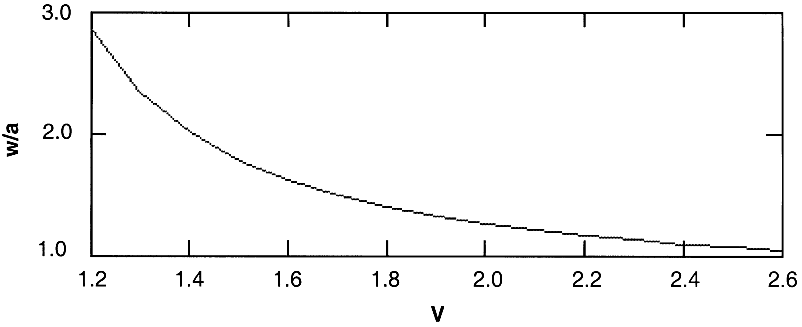


Figure 14.35 Spot size variation with the normalized frequency.

multimode fiber is determined by calculating the difference in arrival times between an axial ray and one traveling at the critical angle. As before, the result is

$$\Delta\left(\frac{\tau}{L}\right) = \frac{n_1 \Delta}{c} \tag{14.89}$$

As an example, if $n_1 = 1.48$ and $n_2 = 1.465$, then $\Delta = 0.01$ so that

$$\Delta\frac{\tau}{L} = 5 \times 10^{-11}\,\text{s/m}$$

or

$$\Delta\frac{\tau}{L} = 50\,\text{ns/km}$$

This may be compared to the much smaller material dispersion calculated in an earlier example of 400 ps/km = 0.4 ns/km. We conclude that modal distortion is the major source of pulse spreading in a step-index multimode fiber. That is, modal distortion is much greater than material dispersion.

14.4.2. Graded-Index Fiber

The graded-index (GRIN) fiber (Fig. 14.36) was developed to overcome the large modal distortion in a multimode step-index fiber. It has a refractive index variation across the core and cladding given in the core by

$$n(r) = n_1\sqrt{1 - 2\left(\frac{r}{a}\right)^{\alpha}\Delta} \tag{14.90a}$$

and in the cladding by

$$n(r) = n_1\sqrt{1 - 2\Delta} = n_2 \tag{14.90b}$$

If $\alpha = 2$, the refractive index can be reduced to

$$n(r) = n_1\left[1 - \left(\frac{r}{a}\right)^2\Delta\right] \tag{14.91a}$$

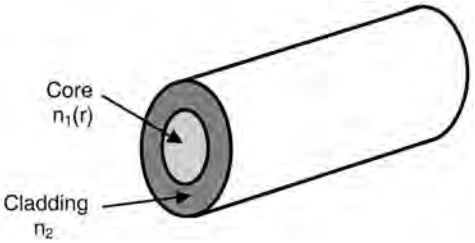


Figure 14.36 Graded-index fiber.

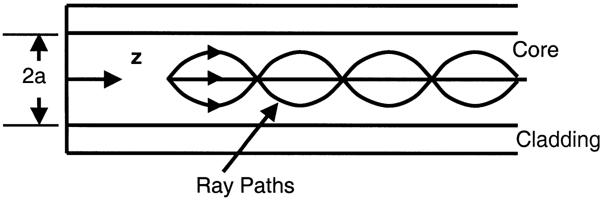


Figure 14.37 Ray paths in a graded-index fiber.

in the core and

$$n(r) = n_2 = n_1(1 - \Delta) \tag{14.91b}$$

in the cladding. This refractive index distribution is called the *parabolic profile*.

For the parabolic profile, the ray paths in the core are given by

$$r(z) = r_0 \cos(\sqrt{A}z) + \frac{1}{\sqrt{A}} r_0' \sin(\sqrt{A}z) \tag{14.92}$$

where r_0 is the initial ray position (at $z = 0$), r_0' is the initial slope and $A = 2\Delta/a^2$, a property of the GRIN fiber. The ray slopes are given by

$$r'(z) = -\sqrt{A} r_0 \sin(\sqrt{A}z) + r_0' \cos(\sqrt{A}z) \tag{14.93}$$

Several ray paths are illustrated in Fig. 14.37.

Modal distortion still exists in the multimode GRIN fiber because of the different path lengths traversed by the various rays. As with the multimode SI fiber, we can calculate the difference in arrival times between pulses traveling axially (the shortest route) and pulses whose trajectories approach the core-cladding boundary (the longest route). This represents the amount of pulse spreading. Note that in the GRIN fiber case, the index of refraction decreases as the ray moves away from the fiber's axis. Therefore (because $v = c/n$), rays speed up as they move away from the fiber's axis. In doing so they tend to catch up with the axial rays, diminishing the amount of pulse spreading. This is the great advantage of the multimode GRIN fiber. An approximate expression for the pulse spread in a GRIN fiber is

$$\Delta\left(\frac{\tau}{L}\right) = \frac{n_1 \Delta^2}{2c} \tag{14.94}$$

Comparison with the results for the SI fiber shows a reduction in the pulse spread by a factor of $2/\Delta$. As an example, if $n_1 = 1.48$ and $n_2 = 1.465$, then $\Delta = 0.01$ so that

$$\Delta\left(\frac{\tau}{L}\right) = 2.53 \times 10^{-13}\text{s/m}$$

or

$$\Delta\left(\frac{\tau}{L}\right) = 0.253\,\text{ns/km}$$

The reduction in pulse spread (and resultant increase in fiber capacity) is close to a factor of 200.

The multimode GRIN fiber is used when path lengths are moderate (such as in LAN applications). Path lengths up to a few kilometers and information rates of a few Gb/s can be accommodated. Longer paths and higher rates require single-mode fibers, where modal distortion is no longer a factor.

For the parabolic GRIN fiber the wave equation can be solved explicitly, without the need for numerical solutions of a characteristic equation. Some of the results follow.

The effective index of refraction for a mode described by positive integers p and q is

$$n_{\text{eff}} = \frac{\beta_{pq}}{k_0} = n_1 - (p + q + 1)\frac{\sqrt{2\Delta}}{k_0 a} \tag{14.95}$$

The factors k_0 and β have the same meaning as before.

The lowest ordered mode has $p = q = 0$. Its electric field is

$$E_{00} = E_0 e^{-\alpha^2 r^2/2} e^{j(\omega t - \beta z)} \tag{14.96}$$

where $\alpha = (k_0 n_1/a)^{1/2}(2\Delta)^{1/4}$. The corresponding transverse field plot appears in Fig. 14.38. It is circularly symmetric and gaussian shaped. For simplicity, we have assumed a lossless fiber in writing the filed equations.

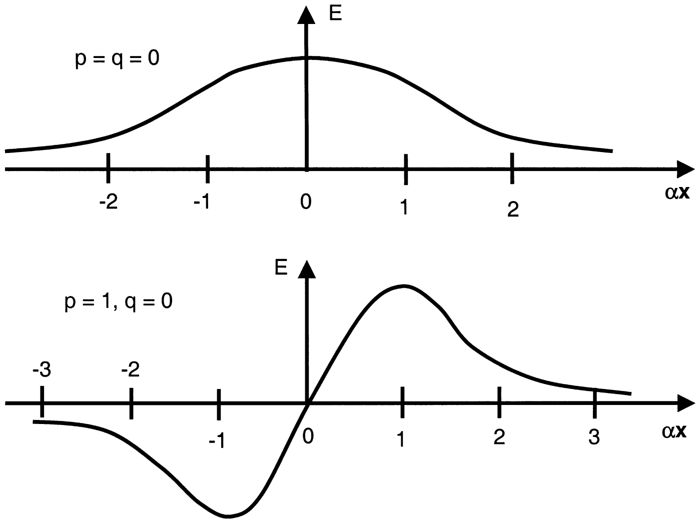


Figure 14.38 Graded-index fiber transverse field patterns for the lowest-ordered ($p = 0$, $q = 0$) mode, the $p = 1$, $q = 0$ mode, and the $p = 2$, $q = 0$ mode.

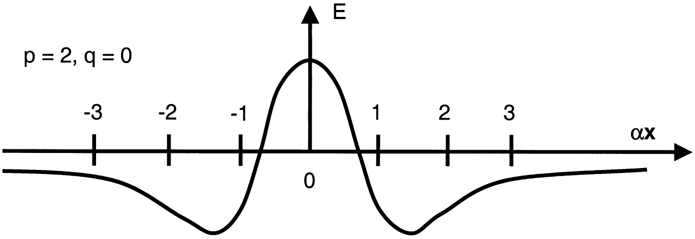


Figure 14.38 Continued.

The $p=1$, $q=0$ and $p=2$, $q=0$ are given, respectively, by

$$E_{10} = E_1 \alpha x e^{-\alpha^2 r^2/2} e^{j(\omega t - \beta z)} \tag{14.97}$$

and

$$E_{20} = E_2[2(\alpha x)^2 - 1] e^{-\alpha^2 r^2/2} e^{j(\omega t - \beta z)} \tag{14.98}$$

These modes are plotted in Fig. 14.38. They are not circularly symmetric nor are they gaussian, although they have a gaussian envelope as indicated by the term $e^{-\alpha^2 r^2/2}$. The mode index p gives the number of zero crossings of the field pattern along the x direction. The mode index q does the same with respect to the y direction.

14.4.3. Attenuation

When we include loss in the equations for the electric fields in fibers, we do so by adding a term of the form $e^{-\alpha z}$, where α is the attenuation coefficient. Typically, fiber attenuation is given in dB/km rather than in terms of the attenuation coefficient. As mentioned earlier, they are related by

$$\text{dB/km} = -8.685\alpha \tag{14.99}$$

where the units of the attenuation coefficient are km^{-1}. For convenience, the minus sign is often omitted when writing the fiber loss in decibels.

A plot of the fiber loss for a silica glass appears in Fig. 14.39. The wavelength regions where fiber systems have been constructed are in the regions around 800–900 nm and from 1250 to 1650 nm. The region around 1400 nm where there is a local increase in attenuation is usually avoided. The nomenclature for the longer wavelength region is presented in Table 14.5.

The local peak in the loss curves near 1380 nm is caused by absorption in the hydroxyl ions (OH) present. This is an impurity whose concentration is minimized during the manufacturing process. The lowest loss region is near 1550 nm. This is where the longest fiber systems are designed to operate. Shorter paths (on the order of a few hundred meters) can be served in the 800-nm region. Moderately long systems, having path lengths up to a few kilometers, can be served by 1300-nm systems. Recall that this wavelength is advantageous because of the low material dispersion.

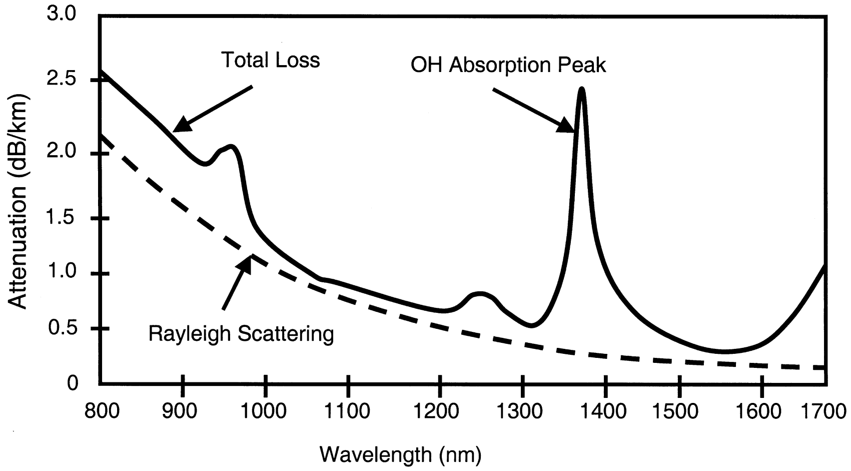


Figure 14.39 Silica glass fiber attenuation.

Table 14.5 Transmission Bands in the Long-wavelength Region

Nomenclature	Descriptor	Range (nm)
0 band	Original	1260–1360
E band	Extended	1360–1460
S band	Short wavelength	1460–1530
C band	Conventional	1530–1565
L band	Long wavelength	1565–1625
U band	Ultra-long wavelength	1625–1675

14.4.4. Waveguide Dispersion and Polarization-Mode Dispersion

Earlier we presented the concepts behind pulse spreading caused by modal distortion and material dispersion. There are two other major pulse spreading mechanisms, *waveguide dispersion* and *polarization-mode dispersion* (PMD).

Waveguide dispersion arises because different wavelengths travel at different speeds, even if traveling in the same mode. To illustrate this statement refer to the exact mode chart in Fig. 14.33 and consider only the HE_{11} mode. Because the source emits light over a range of wavelengths, the V parameter has a range of values associated with it resulting in a corresponding range of effective refractive indices (and related range of velocities). Just as occurs with material dispersion, component wavelengths travel at a different speeds, each arriving with a slight delay with respect to the others. The amount of pulse spreading is given by an equation very similar to that for material dispersion

$$\Delta\left(\frac{\tau}{L}\right) = -M_g \Delta\lambda \tag{14.100}$$

where $\Delta\lambda$ is the source width and M_g is the *waveguide dispersion*. For the SI fiber the waveguide dispersion looks as in Fig. 14.40.

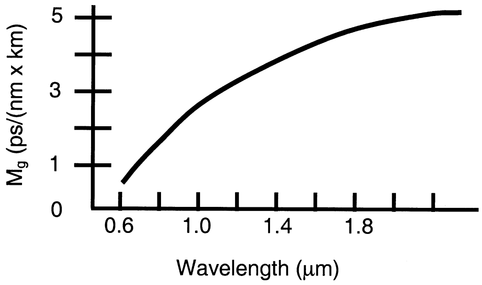


Figure 14.40 Waveguide dispersion in a step-index fiber.

Because material and waveguide dispersion act upon the various wavelengths present in the same way, the two pulse-spreading phenomena combine as

$$\Delta\left(\frac{\tau}{L}\right) = -(M + M_g)\,\Delta\lambda \tag{14.101}$$

By comparing the material and waveguide dispersion values in Figs. 14.8 and 14.40, we see that in the 800-nm region material dispersion dominates. In the 1550-nm region, waveguide and material dispersion are of the same order of magnitude but opposite sign. They tend to cancel each other, but not entirely. Just above 1300 nm the material dispersion is about −4 ps/(nm × km) and waveguide dispersion is about −4 ps/(nm × km). They do cancel each other. In single-mode fibers there is a zero dispersion wavelength, typically near 1310 nm.

By changing the structure of the waveguide (for example, by designing a fiber with a core index profile that is triangular), the waveguide dispersion can be increased to about 20 ps/(nm × km). This just cancels the −20 ps/(nm × km) material dispersion. We call such a fiber, a *dispersion-shifted fiber*. The result is a fiber with minimum loss and minimum pulse spreading at the same wavelength, a very desirable fiber characteristic for high-rate long-path links.

Other index profiles are available that result in other desirable characteristics. A particularly useful one property is a uniform low dispersion over a range of wavelengths. For example, a dispersion of 5 ps/(nm × km) over wavelengths from 1500 to 1600 nm. This is needed when the fiber supports a number of independent carriers in a scheme called *wavelength-division multiplexing*. Tens, and even hundreds, of independent channels can be transmitted simultaneously in this manner. The analytical solution for the fields in waveguides having unusual refractive index profiles can be quite complicated, well beyond the range of what is covered in this chapter.

We have indicated how material and waveguide dispersion add together. If we include modal distortion as well, the total pulse spread is

$$\Delta\left(\frac{\tau}{L}\right) = \sqrt{(\Delta\tau)^2_{\text{modal}} + (\Delta\tau)^2_{\text{dispersive}}} \tag{14.102}$$

The modal spread $(\Delta\tau)_{\text{modal}}$ disappears for a single mode fiber. The dispersive spread $(\Delta\tau)_{\text{dispersive}}$ includes both waveguide and material dispersion.

A final phenomenon causing pulse spreading is polarization-mode dispersion (PMD). PMD occurs in a single-mode fiber because two orthogonally polarized fields can exist simultaneously. For example, we wrote the fields in the step-index fiber for a y-polarized electric field. An x-polarized field can also propagate. Therefore, even in what is called a single-mode fiber ($V < 2.405$), two fields can propagate. In most fibers these two fields will travel at slightly different velocities due to birefringence. *Birefringence* refers to having the index of refraction depend upon the field polarization. Most fibers are birefringent. The birefringence could be caused by an elliptical core (rather than a perfectly circular core). It could also be caused by unequal stresses in the two orthogonal transverse directions occurring during manufacture.

14.5. FURTHER STUDY

This chapter detailed several electromagnetic problems relating to fiber-optic communications. Several books are suggested [26–33] for further study relating to fiber networks and components.

REFERENCES

1. Chaffee, C.D. The rewiring of America. *The Fiber Optics Revolution*; Academic Press: New York, 1987.
2. Hecht, J. *Understanding Fiber Optics*; Prentice-Hall: Upper Saddle River, New Jersey, 1999.
3. Agrawal, G.P.; Dutta, N.K. *Long-wavelength Semiconductor Lasers*; Van Nostrand Reinhold Company: New York, 1986.
4. Kressel, H.; Butler, J.K. (Eds.) *Semiconductor Lasers and Heterojunction LEDs*; Academic Press: New York, 1977.
5. Kressel, H. (Ed.) *Semiconductor Devices for Optical Communications*; Springer-Verlag: New York, 1980.
6. Morthier, G.; Vankwikelberge, P. *Handbook of Distributed Feedback Laser Diodes*; Artech House: Norwood, MA, 1997.
7. Palais, J.C. *Fiber Optic Communications*, 4th Ed.; Prentice-Hall: Upper Saddle River, New Jersey, 1998.
8. Chang, D.K. *Field and Wave Electromagnetics*; Addison Wesley: Reading, Massachusetts, 1983.
9. Born, M.; Wolf, E. *Principles of Optics*, 3rd Ed.; Pergamon Press: New York, 1965.
10. Palais, J.C. *Fiber Optic Communications*, 4th Ed.; Prentice-Hall: Upper Saddle River, New Jersey, 1998.
11. Morris, D.J. *Pulse Code Formats for Fiber Optical Data Communications*; Marcel Dekker: New York, 1983.
12. Chang, D.K. *Field and Wave Electromagnetics*; Addison Wesley: Reading, MA, 1983.
13. Born, M.; Wolf, E. *Principles of Optics*, 3rd Ed.; Pergamon Press: New York, 1965.
14. Coldren, L.A.; Corzine, S.W. *Diode Lasers and Photonic Integrated Circuits*; Wiley: New York, 1995.
15. Ebeling, K.J. *Integrated Optoelectronics*; Springer-Verlag Publishing Co.: New York, 1993.
16. Hunsberger, R.G. *Integrated Optics: Theory and Technology*, 4th Ed.; Springer-Verlag: New York, 1995.
17. März, R. *Integrated Optics: Design and Modeling*; Artech House: Norwood, MA, 1995.
18. Murphy, E.J. *Integrated Optical Circuits and Components, Design and Applications*; Marcel Dekker: New York, 1999.
19. Cherin, A.H. *An Introduction to Optical Fibers*; McGraw-Hill, Inc.: New York, 1983.
20. Ghatak, A.K.; Thyagarajan, K. *Introduction to Fiber Optics*; Cambridge University Press: New York, 1998.

21. Jeunhomme, L.B. *Single-Mode Fiber Optics*, 2nd Ed.; Marcel Dekker: New York, 1990.
22. Marcuse, D. *Theory of Dielectric Optical Waveguides*, 2nd Ed.; Academic Press: New York, 1991.
23. Okamoto, K. *Fundamentals of Optical Waveguides*; Academic Press: New York, 2000.
24. Sodha, M.S.; Ghatak, A.K. *Inhomogeneous Optical Waveguides*; Plenum Press: New York, 1977.
25. Yariv, A. *Introduction to Optical Electronics*; 4th Ed.; Holt, Rinehart and Winston: New York, 1991.
26. Agrawal, G.P. *Fiber-Optic Communication Systems*, 2nd Ed.; John Wiley and Sons: New York, 1997.
27. DeCusatis, C.; Clement, D.; Maass Eric; Lasky, R. *Handbook of Fiber Optic Data Communication*; Academic Press Inc.: New York, 1997.
28. Goff, D.R. *Fiber Optic Reference Guide*; Focal Press: Boston, 1996.
29. Green, P.E. *Fiber Optic Networks*; Prentice-Hall: Englewood Cliffs, New Jersey, 1993.
30. Keiser, G.E. *Optical Fiber Communications*, 3rd Ed.; McGraw-Hill: New York, 2000.
31. Ramaswami, R.; Sivarajan, K.N. *Optical Networks: A Practical Perspective*, 2nd Ed.; Morgan Kaufmann: New York, 2002.
32. Weik, M.H. *Fiber Optics Standard Dictionary*, 3rd Ed.; Chapman & Hall: New York, 1997.
33. Bass, M. (Ed.) *Fiber Optics Handbook*; McGraw Hill: New York, 2002.

15
Numerical Techniques

Randy L. Haupt
Utah State University
Logan, Utah, U.S.A.

15.1. INTRODUCTION

Numerical techniques for calculating electromagnetic fields surpassed analytical techniques many years ago. Analytical methods work for only a few basic geometries that do not apply to most practical problems. Most undergraduate electromagnetics texts now contain sections on numerical methods for calculating fields. The IEEE Transactions on Antennas and Propagations have more articles on numerical calculation of fields than on analytical calculation. Professors must skip teaching some of the traditional analytical methods in favor of the newer numerical methods. Classroom and technical presentations make use of electromagnetic movies in which the viewer watches a very colorful display of a gaussian pulse striking an object and scattering. The computer has become a critical part of electromagnetics.

Computational electromagnetics is the simulation of Maxwell's equations and their variations on a computer. Numerical approaches to solving Maxwell's equations find the fields in either the time domain or frequency domain. Time-domain models contain many frequencies and can model transient behavior. On the other hand, frequency-domain methods calculate solutions for one frequency at a time and are appropriate for steady-state behavior. Fourier transforms allow transitioning between the two domains.

Maxwell's equations are a function of space and time. For instance, Faraday's law and Ampere's law in vector form are

$$\nabla \times \boldsymbol{E} = -\mu \frac{\partial \boldsymbol{H}}{\partial t} \tag{15.1}$$

$$\nabla \times \boldsymbol{H} = \sigma \boldsymbol{E} + \varepsilon \frac{\partial \boldsymbol{E}}{\partial t} \tag{15.2}$$

where the time-dependent electric ($\boldsymbol{E}$) and magnetic ($\boldsymbol{H}$) fields are given by

$$\boldsymbol{E}(x,y,z,t) = \mathbf{a}_x E_x(x,y,z,t) + \mathbf{a}_y E_y(x,y,z,t) + \mathbf{a}_z E_z(x,y,z,t) \tag{15.3}$$

$$\boldsymbol{H}(x,y,z,t) = \mathbf{a}_x H_x(x,y,z,t) + \mathbf{a}_y H_y(x,y,z,t) + \mathbf{a}_z H_z(x,y,z,t) \tag{15.4}$$

and μ = permeability, ε = permittivity, and σ = conductivity. Spatial dependence of the material properties and random fluctuations add to the complexity of representing electromagnetic parameters. The time-dependent forms of Maxwell's equations require boundary values and initial conditions in order to find the fields.

Before finding the behavior of a field at a single frequency, ω, Maxwell's equations must be converted to a form that has a single frequency. Assuming that the time portion of the field is the fundamental harmonic in a Fourier series, the x component of the electric field is

$$E_x(x,y,z,t) = E_x(x,y,z)\cos\omega t = E_x(x,y,z)\mathrm{Re}\left\{e^{j\omega t}\right\} \rightarrow E_x(x,y,z)e^{j\omega t} \tag{15.5}$$

Since all the components have the same time factor, it divides out of Maxwell's equations leaving

$$\nabla \times \mathbf{E} = -j\omega\mathbf{B} \tag{15.6}$$

$$\nabla \times \mathbf{H} = \sigma\mathbf{E} + j\omega\varepsilon\mathbf{E} \tag{15.7}$$

The time-harmonic form of Maxwell's equations is a function of space and frequency (frequency terms result from time derivatives in Maxwell's equations) but not time. Consequently, this formulation requires the specification of boundary values. Time behavior of the field comes from calculating the fields at many frequencies and Fourier transforming the results to the time domain.

Unlike analytical methods, computer solutions are not in closed form but are just numbers assigned to grid points. The computer calculates fields and currents at discrete points or grid points specified in the region of interest. Most numerical methods set up a grid of points equally separated in space and time. More grid points increase accuracy but increase computation time as well. The Nyquist rate requires sampling the waveform at twice the highest frequency. Generally, numerical methods limit the maximum spacing between points to be less than $\lambda/10$. For time-domain methods, λ is the wavelength at the center frequency. Grid spacing inside penetrable materials depends on the wavelength inside the material.

Not all numerical methods use uniform grids. A good example is the popular gaussian quadrature formulas for numerical integration. This powerful approach places sample points at the zeros of polynomials and weights and adds the function at those points to find the answer. They have a higher order of accuracy than equally sampled formulas. Some solution domains have small regions where the fields or currents change rapidly. One strategy that maintains accuracy while keeping the number of grid points reasonable is to transition from a coarse grid where fields slowly change to a finer grid where fields rapidly change. Undersampling violates the Nyquist rate resulting in aliasing and the corruption of results.

Realistic radiating objects are very difficult to model with current electromagnetics codes. For instance, accurately modeling the scattered field due to a radar pulse incident on an airplane cannot be done without many reasonable approximations that cut down on the computational load. Some of the more common approximations include

1. Modeling in 1D or 2D instead of 3D. Looking at cuts through a 3D object are often sufficient for many applications.
2. Assuming a surface or wire is infinitely thin. This approach for integral equations and high-frequency methods simplifies the calculations.

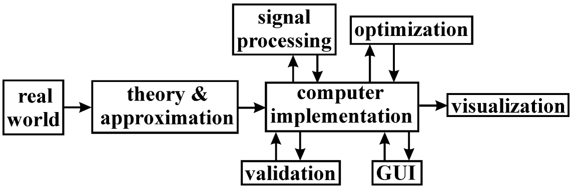


Figure 15.1 Block model of the computational electromagnetics process.

3. Simple sources, e.g., plane wave, point source, constant current, constant voltage, and gaussian pulse.
4. Replacing curved lines with straight lines. This assumption can result in stair-step boundaries, straight line instead of a curve, and less complicated math.
5. Ignoring mutual coupling. Only consider mutual coupling between adjacent array elements, using point sources in place of dipoles, discarding small elements in the MOM impedance matrix. Coupling increases storage and calculations.
6. Applying far field approximations. The far-field assumption ignores small amplitude terms.

Knowing the approximations used in a given numerical calculation is essential to the proper interpretation of the output.

Figure 15.1 is a flowchart of computer modeling and of this chapter. People develop theory to explain the real world. For electromagnetics, that theory is Maxwell's equations. Approximations save computational effort. Section 15.2 presents the computer implementation of Maxwell's equations in the form of numerical algorithms. The output from the computer models usually goes to an optimization or signal processing algorithm for practical use in system design. These functions are discussed in Sec. 15.3. Results of the computer model must be verified and validated in order to be accepted and used by the electromagnetics community. Graphical user interfaces (GUI) and proper visualization of the output are necessary for a good software package. Programming issues appear in Sec. 15.4.

15.2. SOLVING MAXWELL'S EQUATIONS

The four most commonly used methods for numerically finding electromagnetic fields appear in this chapter. All these methods had their beginnings solving static or frequency-domain problems. Adding time domain capability had to await the computational resources of the 1960s. Finite differencing and integral equation methods proved easiest to convert from frequency to time domain.

15.2.1. High-Frequency Methods

This presentation on high-frequency methods is based on material found in Refs. 1 and 2 with original work coming from Ref. 3. High-frequency methods assume the wavelength approaches zero. This assumption works well when applied to very large objects (at least a few wavelengths across). Geometrical optics, also known as *ray tracing*, forms the basis for these techniques. The electromagnetic rays are orthogonal trajectories to the phase fronts of a wave described by the Eikonal equation:

$$\left(\frac{\partial \zeta}{\partial x}\right)^2+\left(\frac{\partial \zeta}{\partial y}\right)^2+\left(\frac{\partial \zeta}{\partial z}\right)^2=n^2(x,y,z) \tag{15.8}$$

where ζ is known as the *eikonal* or surface of constant phase and n is the index of refraction. Rays are lines perpendicular to the constant phase fronts. As the phase fronts curve due to changes in the index of refraction, the rays correspondingly bend.

When electromagnetic rays impinge on an object, they reflect from and/or transmit through the surface. Geometrical optics (GO) or ray tracing ignores diffraction effects and assumes that an electromagnetic wave is a series of rays traveling in straight lines. The reflected electric field (E^r) at a distance s from the reflection point p is calculated from the incident field (E^i) by

$$\begin{bmatrix} E^r_{\|}(s) \\ E^r_{\perp}(s) \end{bmatrix} = \begin{bmatrix} E^i_{\|}(p)R_{\|} \\ E^i_{\perp}(p)R_{\perp} \end{bmatrix} \sqrt{\frac{\rho_1\rho_2}{(\rho_1+s)(\rho_2+s)}}e^{-jks} \tag{15.9}$$

where the subscripts correspond to parallel and perpendicular polarizations. GO treats the reflected wave as a local phenomenon at point p. In other words, the reflected wave is a function of the incident wave, shape of the object (ρ_1 and ρ_2 are the orthogonal radii of curvature of the reflected wavefront), and material makeup of the object ($R_{\|}$ and $R_{\perp}$ are the parallel and perpendicular Fresnel reflection coefficients) of the object it strikes. Parallel and perpendicular quantities are referenced to the plane of incidence, which is the plane containing the incident ray and the edge of the object. The total field is the sum of the incident and reflected fields.

$$\mathbf{E}=\mathbf{E}^i+\mathbf{E}^r \tag{15.10}$$

The incident field takes one of the following forms

$$E^i=\begin{cases} e^{-jk\rho_i} & \text{plane waves} \\ \dfrac{e^{-jk\rho_i}}{\sqrt{\rho_i}} & \text{cylindrical waves} \\ \dfrac{e^{-jk\rho_i}}{\rho_i} & \text{spherical waves} \end{cases} \tag{15.11}$$

where ρ_i is the distance from the source to the point of reflection.

The shooting and bouncing ray (SBR) technique [4] represents a plane wave by a bundle of rays that "shoot" into the cavity and bounce around. Each ray is traced using GO as it reflects from conductors and passes through dielectrics. Integrating the exiting rays over the aperture yields the scattered field. This technique works well for very complex shaped cavities that have various materials inside.

Sometimes the surface currents are a more important quantity than the fields. Reflected or scattered fields are then calculated from the induced surface currents. This approach is known as *physical optics* (PO). The PO current only exists where the incident field directly illuminates a surface (lit region) and is found from the tangential component of the incident magnetic field.

$$\mathbf{J}_s = \begin{cases} 2\hat{\mathbf{n}} \times \mathbf{H}^i & \text{lit region} \\ 0 & \text{shadow region} \end{cases} \tag{15.12}$$

where $\hat{\mathbf{n}}$ is the unit normal to the surface. Once the current is found, the reradiated fields are calculated using the appropriate radiation integral.

PO fields are most accurate in the specular direction, since the shadow regions have no surface current. GO and PO work reasonably well for large objects and at angles near the specular direction. PO ignores the edge effects and results in uniform induced currents on the surface. PO and GO are used to derive simple radiation formulas like the radar cross section (RCS) of simple shapes.

Techniques have been developed to supplement GO and PO in order to take into account edge effects. The geometrical theory of diffraction (GTD) adds a diffracted field to the GO approximation. The physical theory of diffraction (PTD) [5], developed independently from GTD, adds a nonuniform or fringe current to the PO approximation. These similar approaches result in the same fields. Only GTD and its extensions are presented here.

Figure 15.2 shows three regions that arise from a wave incident on a finite curved conductor: (1) direct, reflected, and diffracted fields, (2) direct and diffracted fields, and (3) diffracted fields (shadow region). Diffraction results when an incident wave impinges on an edge, corner, or tip and scatters. It also occurs as a creeping ray around a smooth object when the incident field is at grazing incidence. Like the reflected field from a large object, the diffracted field is a localized phenomenon.

The GTD diffracted electric field (superscript *d*) is given by

$$\begin{bmatrix} E_{\|}^d(s) \\ E_{\perp}^d(s) \end{bmatrix} = \begin{bmatrix} E_{\|}^i(p)D_{\|} \\ E_{\perp}^i(p)D_{\perp} \end{bmatrix} A(s)e^{-jks} \tag{15.13}$$

where $s =$ distance from diffraction point to observation point and $D_{\|}$ is the parallel and $D_{\perp}$ is the perpendicular polarization diffraction coefficient. The spatial attenuation

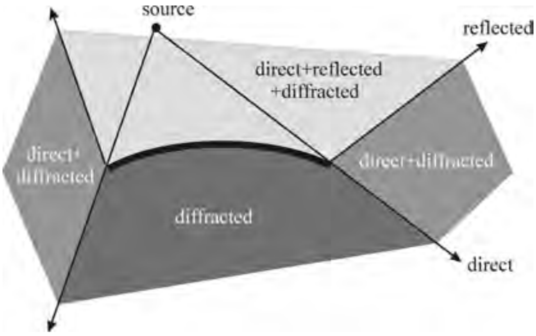


Figure 15.2 A curved object illuminated by a source has three field regions containing the direct, reflected, and/or diffracted fields.

factor, $A(s)$, is a function of the incident wave (plane, cylindrical, or spherical). Diffraction coefficients exist for various geometries, including

1. Reflection at a plane or curved surface
2. Diffraction at a straight or curved wedge
3. Diffraction at a corner in a plane or doubly curved surface
4. Creeping waves around curved objects like cones, cylinders, and ellipsoids

Adding Eq. (15.13) to the GO field in Eq. (15.9) produces the total electric field. The diffraction coefficient depends on the polarization of the incident wave as well as the geometry and material composition of the object.

Intuition dictates that the discontinuous boundaries inherent in GTD should be smooth transitions. The uniform theory of diffraction (UTD) [6] multiplies the singularities in the diffraction coefficients by transition functions that go to zero at the boundaries resulting in a smooth transition between regions. In addition, UTD takes into account creeping waves that arise from a ray incident tangential to a curved surface by including another term for the scattered field that has a launching coefficient associated with the creeping wave. The creeping wave travels around a geodesic (Fermat's principle) and radiates as it travels around the surface.

UTD and PTD work well for a two-dimensional configuration like an infinite wedge. A finite-length wedge requires the use of incremental diffraction coefficients (ILDCs) [7]. ILDCs come from the closed form diffraction coefficients that correspond to two-dimensional geometries, such as the wedge, strip, polygonal cylinder, and slit. The ILDCs are integrated over the length of an edge. If the edge is infinite, then the result corresponds to the two-dimensional diffraction coefficients. See Ref. 8 for practical applications.

Time-domain UTD models result from Fourier transforming the frequency-domain UTD solutions [9]. The TD-UTD solution is accurate during the time it takes for a ray to propagate from the source to the observation point. The TD-UTD is most useful when the pulse width of the incident field is small compared to the geometric dimensions of the radiating object.

Are high-frequency techniques numerical or analytical? They are presented here as numerical methods because the calculation of the field points for large scattering objects requires a computer. Also, adding time domain to UTD and using UTD with other numerical methods results in complicated numerical algorithms. Finally, the SBR method is computationally intensive.

15.2.2. Integral Equations

Integral equations work well for finding the radiating fields from perfectly conducting objects. Integral equations are derived from the tangential boundary conditions for the electric and magnetic fields

$$\hat{\boldsymbol{n}} \times \boldsymbol{H}^t(r,t) = \hat{\boldsymbol{n}} \times \left[\boldsymbol{H}^i(r,t) + \boldsymbol{H}^s(r,t)\right] = \boldsymbol{J}(r,t) \tag{15.14}$$

$$\hat{\boldsymbol{n}} \times \boldsymbol{E}^t(r,t) = \hat{\boldsymbol{n}} \times \left[\boldsymbol{E}^i(r,t) + \boldsymbol{E}^s(r,t)\right] = 0 \tag{15.15}$$

where the superscripts t, i, and s stand for total, incident, and scattered, respectively, and $\hat{\boldsymbol{n}}$ is the unit normal. One of the two most commonly used integral equations in electromagnetics is the electric field integral equation (EFIE).

$$\frac{Z}{k}\left[k^2\iint_S \boldsymbol{J}_s(r')G(\boldsymbol{r}_s,\boldsymbol{r}')ds' + \nabla\iint_S \nabla'\cdot\boldsymbol{J}_s(r')G(\boldsymbol{r}_s,\boldsymbol{r}')ds'\right] = E_t^i(r=r_s) \tag{15.16}$$

where

Z = impedance.
k = wave number.
S = surface of scatterer.
$\boldsymbol{J}_s$ = surface current.
Primed quantities = source points.
r_s = observation point on the surface.
$G(\boldsymbol{r},\boldsymbol{r}')$ = Green's or transfer function.

It enforces the boundary conditions on the tangential electric field and can be used on open or closed surfaces. The second integral equation is the magnetic field integral equation (MFIE).

$$\boldsymbol{J}_s(r') - \lim_{r\to S}\left\{\hat{\boldsymbol{n}}\times\iint_S \boldsymbol{J}_s(r')\times[\nabla' G(\boldsymbol{r},\boldsymbol{r}')]\,ds'\right\} = \boldsymbol{H}_t^i(r=r') \tag{15.17}$$

It enforces the boundary conditions on the tangential magnetic field but can only be used on closed surfaces.

The integral operators take into account all surfaces of the computational domain. Calculating the current at one point on an object includes interactions from all other points on all surfaces. Singularities associated with the Green function inside the integrals must be carefully dealt with to get accurate results. Fields are found from the currents, so only the desired field points need to be calculated. Fields between the current and desired field points are not calculated.

The method of moments (MOM) finds the currents on an object due to an incident wave or an induced voltage or current [10]. It is a way of converting integral equations to matrix equations. This technique works best for wires and flat plates. More complex systems are assembled from wires and/or metal plates. Each wire or metal plate is further subdivided into wire segments or patches that are small compared to the frequency's wavelength. Figure 15.3 shows an example of a solid sphere modeled using wires. Currents

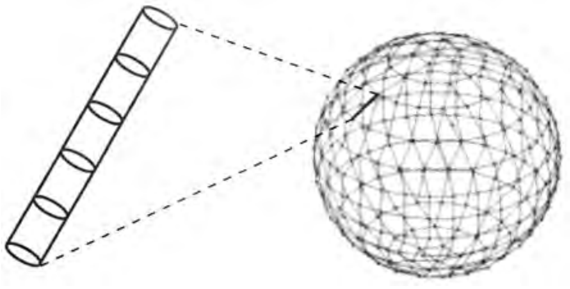


Figure 15.3 A solid perfectly conducting sphere is modeled with a wire grid. Each wire is divided into subsegments. MOM calculates the current induced on each subsegment.

on a thin wire are easier to calculate than currents on a thick wire, since the current only flows in one dimension on a thin wire. The wire radius must be large enough that the total surface area of the wires equals the total surface area of the true structure. The wires should be less than a quarter wavelength, with lengths of less than 0.1 wavelength very common. Certain regions, particularly near an induced source, need the wires broken into even shorter segments. The MOM technique determines the current on every wire segment and surface patch due to the sources and all the other currents on the other wire segments and surface patches. Once these currents are known, then the electric field at any point in space is found by integrating the contributions from all the wire segments and surface patches.

The MOM formulation begins by representing the current as the sum of weighted (a_n) simple functions known as basis functions (F_n)

$$J(r') = \sum_{n=1}^{N} a_n F_n(r') \tag{15.18}$$

The basis functions may be full wave functions like sine and cosine or piecewise functions like pulses or triangles. This expansion is substituted into the EFIE or MFIE, then the inner product is taken with N weighting functions. When the weighting functions are the same as the expansion functions the approach is known as *Galerkin's method.* If the weighting functions are delta functions, then the technique is known as *point matching* or *collocation*.

The MOM results in a $N \times N$ matrix known as the impedance matrix (Z). The unknowns in the vector I are the coefficients, a_n, in Eq. (15.18), and the right-hand side of the matrix equation is the source vector, V.

$$ZI = V \tag{15.19}$$

The induced voltage is either an applied voltage or incident electric field. This equation is solved for the a_n in the I vector. Then, the current on the object is found from Eq. (15.18).

The impedance matrix is usually a full matrix (almost all matrix elements are nonzero). A standard routine for solving this equation is LU decomposition and back substitution. The decomposition part takes about N^2 operations, while the back substitution takes N operations. Multiple right-hand sides use the same decomposition; so finding I takes N steps. Thus, each of multiple incidence angles takes only N operations to find the current. The impedance matrix localization (IML) method uses basis and testing functions that localize strong interactions to only a small number of elements within the impedance matrix [11]. The other matrix elements are small (typically 10^{-4} to 10^{-6} in relative magnitude) and set equal to zero.

The EFIE and MFIE formulations produce spurious currents for cavities near resonance. These result when the eigenvalues of the integral equation go to zero and an ill-conditioned matrix in the MOM formulation results [12]. The combined field integral equation (CFIE) formulation linearly combines the EFIE and MFIE formulations in order to reduce the spurious currents. The EFIE equation times a constant α where $0 \leq \alpha \leq 1$, plus the MFIE times $1 - \alpha$ produce the CFIE. More recently, Shore [13] advocates dual surface MFIE and EFIE to eliminate the resonant problems still associated with some formulations of the CFIE.

The most well-known electromagnetic modeling code is NEC (Numerical Electromagnetics Code) and its commercial variations [14]. This code combines an integral equation for smooth surfaces with one for wires in order to model a variety of antenna structures. The antenna model can have nonradiating networks and transmission lines connecting parts of the structure, perfect or imperfect conductors, and lumped-element loading. The NEC program uses both EFIE and MFIE. The EFIE works best for thin-wire structures of small conductor volume, while the MFIE (does not work for the thin-wire case) works best for voluminous structures with large smooth surfaces. The EFIE models thin surfaces very well. Although the EFIE is specialized to thin wires in NEC, it is frequently used to represent surfaces that may be modeled by wire grids with reasonable success for far-field quantities but with variable accuracy for surface fields.

15.2.3. Finite Difference Methods

Maxwell's equations may be directly solved by replacing the derivatives with finite difference formulas. For instance, replacing a spatial derivative with a central difference approximation of a function, F, on an equally spaced grid along the x axis yields

$$\frac{df(x_i)}{dx} = \frac{F(x_{i+1}) - F(x_{i-1})}{2h} + O\left(h^2\right) \tag{15.20}$$

where x_n is between x_{n-1} and x_{n+1} and the grid spacing is h. The central difference approximation is a second order approximation because the error is on the order of h^2. Higher order derivatives use more grid points in the finite difference approximation. Assuming $h \ll 1$, a fourth-order finite difference formula is approximately two orders of magnitude better than a second-order formula.

For example, consider approximating the Poisson equation with second order differencing of the scalar function, v, and source, s

$$s = \nabla^2 v = \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} \simeq \frac{v_{i+1,j} - 2v_{i,j} + v_{i-1,j}}{h^2} + \frac{v_{i,j+1} - 2v_{i,j} + v_{i,j-1}}{h^2} \tag{15.21}$$

Solving for v at the grid point (i,j) yields the Jacobi iterative formula

$$v_{i,j}^{n+1} = \frac{1}{4}\left(v_{i-1,j}^{n} + v_{i+1,j}^{n} + v_{i,j-1}^{n} + v_{i,j+1}^{n}\right) - \frac{sh^2}{4} \tag{15.22}$$

New values of v (represented by $n+1$) are found by averaging the previous (represented by n) four adjacent values of v. Equation (15.22) is a local operator, since it only uses nearby grid points. Local operators allow more detail in the model, such as changing material properties, detailed shapes, and nonconducting objects. Gauss Seidel iteration uses updated values of v on the right-hand side of Eq. (15.22) when available; so it is preferred over the Jacobi formulation.

An example of finite differencing without time dependence is the Laplace equation ($s = 0$) over a square grid with the top of the square at 5 V and the other three sides at 0 V.

If the grid of unknown voltages is $N \times N$, then there are N^2 equations and N^2 unknowns. The matrix equation takes the form

$$\begin{bmatrix} -4 & 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 1 & -4 & 1 & 0 & \cdots & 0 & 1 & \ddots & \vdots \\ 0 & 1 & -4 & 1 & 0 & \cdots & 0 & \ddots & 0 \\ \vdots & 0 & 1 & \ddots & \ddots & \ddots & \vdots & \ddots & 1 \\ 0 & \vdots & \ddots & \ddots & -4 & 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 & 1 & -4 & 1 & \ddots & \vdots \\ 0 & 1 & 0 & \cdots & 0 & 1 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \ddots & \ddots & -4 & 1 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 1 & -4 \end{bmatrix} \begin{bmatrix} v_{2,2} \\ \vdots \\ \\ \\ \\ \\ \\ \end{bmatrix} = \begin{bmatrix} c_{1,1} \\ \vdots \\ \\ \\ \\ \\ \\ \end{bmatrix} \tag{15.23}$$

A sparse matrix, like this one, has most of its elements equal to zero. If there is some well-defined pattern to the elements in the matrix, then storage becomes simpler and solution of (15.23) is by an optimized direct method or by iteration using an algorithm like conjugate gradient. See Press et al. [15] for more details.

Multigrid is an important breakthrough in quickly solving boundary value problems like the Poisson equation model. Multigrid iteratively solves the problem on a coarse grid (spacing between grid points is $4h$)

$$\nabla_{4h}^2 V_{4h} = s_{4h} \tag{15.24}$$

then interpolates this solution to a finer grid (spacing between grid points is $2h$). Iteration finds the solution on the finer grid. This process continues until reaching the finest grid (h). Next, the residual (r) of the equation on the fine grid (difference between the right and left hand sides) is restricted or converted back to the coarse grid where the error (e) on the coarse grid is found through iteration.

$$\nabla_{2h}^2 e_{2h} = r_{2h} \tag{15.25}$$

Interpolating the error to the fine grid and added to the fine grid solution. This completes a "V" cycle (Fig. 15.4). Multigrid works by reducing the low-frequency components of the error on the coarse grids while reducing the high-frequency error on the fine grids. Normal iterative techniques work with only a fine grid and take a long time to reduce the low frequencies in the error. Reference 16 provides a tutorial on applying multigrid to some electrostatic problems.

Maxwell's equations have a time dependence in addition to the spatial dependence. Consequently, a temporal grid exists in conjunction with the spatial grid. Time implies initial conditions, and space implies boundary conditions. The most common approach to the finite difference solution of Maxwell's equations is the finite difference time-domain (FDTD) method. Excellent references for FDTD include Refs. 1, 17, 18, and the original work by Yee in Ref. 19.

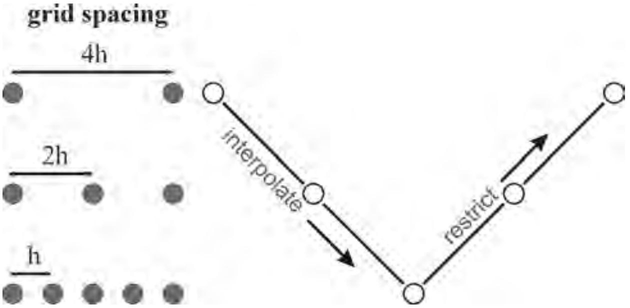


Figure 15.4 A V cycle for multigrid when solving the Poisson equation.

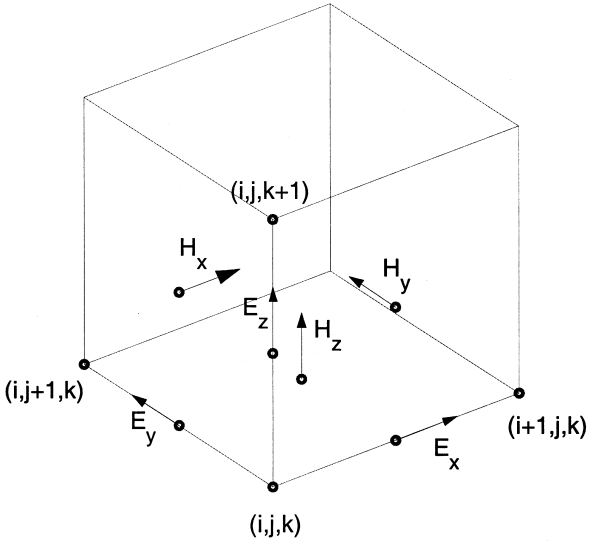


Figure 15.5 The positions of the six field components for the FDTD cube.

FDTD replaces the curl on the left side of Ampere's and Faraday's laws and the partial derivative with respect to time on the right side with second-order finite difference approximation. The resulting grid is complicated and difficult to picture. It is useful to visualize the field components and locations using the smallest complete unit called the *Yee cube* as shown in Fig. 15.5. Note that each of the six field components has a different location in space. Stacking these cubes in the three orthogonal directions covers the computational domain with a three-dimensional grid. The spatial location of grid points is given by $i\Delta x$, $j\Delta y$, and $k\Delta z$ where i, j, and k are integers. Yee's three-dimensional, lossless, source-free ($\mathbf{J} = 0$) equations are written as

$$H^{n+1/2}_{x,i,j+1/2,k+1/2} = H^{n-1/2}_{x,i,j+1/2,k+1/2} + \frac{\Delta t}{\mu_{i,j,k}} \times \left(\frac{E^n_{y,i,j+1/2,k+1/2} - E^n_{y,i,j+1/2,k-1/2}}{\Delta z} - \frac{E^n_{z,i,j,k+1/2} - E^n_{z,i,j+1,k+1/2}}{\Delta y} \right) \tag{15.26}$$

$$H^{n+1/2}_{y,i+1/2,j,k+1/2} = H^{n-1/2}_{y,i+1/2,j,k+1/2} + \frac{\Delta t}{\mu_{i,j,k}}\left(\frac{E^n_{z,i,j,k+1/2} - E^n_{z,i+1,j,k+1/2}}{\Delta x} - \frac{E^n_{x,i+1/2,j,k} - E^n_{x,i+1/2,j,k+1}}{\Delta z}\right) \tag{15.27}$$

$$H^{n+1/2}_{z,i+1/2,j+1/2,k} = H^{n-1/2}_{z,i+1/2,j+1/2,k} + \frac{\Delta t}{\mu_{i,j,k}}\left(\frac{E^n_{x,i+1/2,j,k} - E^n_{x,i+1/2,j+1,k}}{\Delta y} - \frac{E^n_{y,i,j+1/2,k} - E^n_{y,i+1,j+1/2,k}}{\Delta x}\right) \tag{15.28}$$

$$E^{n+1}_{x,i+1/2,j,k} = E^n_{x+1/2,i,j,k} + \frac{\Delta t}{\varepsilon_{i,j,k}}\left(\frac{H^{n+1/2}_{z,i+1/2,j+1/2,k} - H^{n+1/2}_{z,i+1/2,j-1/2,k}}{\Delta y} - \frac{H^{n+1/2}_{y,i+1/2,j,k+1/2} - H^{n+1/2}_{y,i+1/2,j,k-1/2}}{\Delta z}\right) \tag{15.29}$$

$$E^{n+1}_{y,i,j+1/2,k} = E^n_{y,i,j+1/2,k} + \frac{\Delta t}{\varepsilon_{i,j,k}}\left(\frac{H^{n+1/2}_{x,i,j+1/2,k+1/2} - H^{n+1/2}_{x,i,j+1/2,k-1/2}}{\Delta z} - \frac{H^{n+1/2}_{z,i+1/2,j+1/2,k} - H^{n+1/2}_{z,i-1/2,j+1/2,k}}{\Delta x}\right) \tag{15.30}$$

$$E^{n+1}_{z,i,j,k+1/2} = E^n_{y,i,j,k+1/2} + \frac{\Delta t}{\varepsilon_{i,j,k}}\left(\frac{H^{n+1/2}_{y,i+1/2,j,k+1/2} - H^{n+1/2}_{z,i-1/2,j,k+1/2}}{\Delta x} - \frac{H^{n+1/2}_{x,i,j+1/2,k+1/2} - H^{n+1/2}_{x,i,j-1/2,k+1/2}}{\Delta y}\right) \tag{15.31}$$

The spatial samples are often at half increments on the grid. Subscripts indicate the field component (x, y, or z) and the rest of the subscript denotes the location of the field component on the grid. The notation takes some time to learn. Try writing Eqs. (15.24) to (15.29) from Maxwell's equations and drawing your own version of the Yee cube to gain a sufficient understanding of the spatial grid.

Superscripts on the field components indicate the time increment, where n is an integer. All magnetic field components are spaced on the half grid and the electric field components are on the whole integer grid for time. All electric field components are Δt apart. Similarly, all magnetic field components are Δt apart.

An example of one-dimensional space and time grid appears in Fig. 15.6. A wave, such as a gaussian pulse, propagates via Eqs. (15.25) and (15.29). A leapfrog scheme that first calculates the electric field from the magnetic field grid then the magnetic field from the electric field grid, updates the fields. The electric field at the current time comes from the electric field at one previous time step and at the same spatial location and the magnetic field at a previous one-half time step and one-half a space step on either side of the electric field. A complementary logic is used to find the magnetic field.

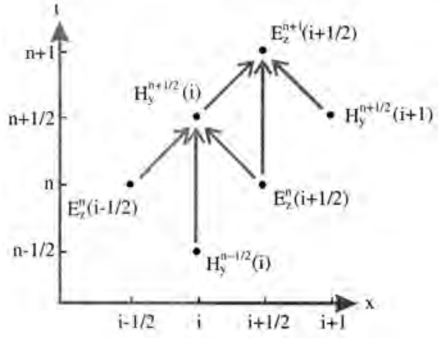


Figure 15.6 A one-dimensional FDTD grid in space and time.

The original Yee FDTD algorithm is second-order accurate in space and time. Numerical dispersion occurs because the phase velocity in the grid is not the same as the phase velocity in the physical problem. Small grid spacing in time and space minimize numerical dispersion. A finer grid implies more unknowns, creating a trade-off between accuracy and computational load. The spatial sampling (in this case Δx) is less than or equal to $\lambda/10$. A full time step for a three-dimensional spatial problem is calculated from the Courant condition given by

$$\Delta t \leq \frac{1}{c\sqrt{1/(\Delta x)^2 + 1/(\Delta y)^2 + 1/(\Delta z)^2}} \tag{15.32}$$

where c is the speed of light. Computationally large problems require many grid points per wavelength to reduce the dispersion error to an acceptable level. One way around this problem is to use higher order derivatives like second-order accuracy in time and fourth-order accuracy in space [20]. Modeling boundary conditions and discontinuities are still a topic of research for the higher order approaches.

Many FDTD applications involve modeling waves traveling in free space, such as with an antenna. In order to model these open region problems accurately, a relatively large free-space area around the objects of interest must be gridded. The end of the gridded space actually forms a numerical boundary that reflects incident waves. Absorbing boundary conditions (ABC) significantly reduce or eliminate these artificial reflections. The perfectly matched layer (PML) technique has proven to be a standard for FDTD [21]. It absorbs the electromagnetic waves from any angle of incidence and of any frequency. Figure 15.7 shows the magnitude of the scattered field from a perfectly conducting metal cube with a sinusoidal incident field. The grid is visible on the interior of the cube because the electric field does not penetrate the cube.

Most of the time, the space between the radiating object and the far field is too large to grid and solve for field values at all the grid points. Instead, near-field data must be transformed into the far field [22]. A transformation boundary surrounds the radiating object while lying within the FDTD grid boundaries. Tangential field components calculated using FDTD are converted into equivalent electric and magnetic surface currents on the transformation boundary. Far-field quantities are then calculated from these surface currents.

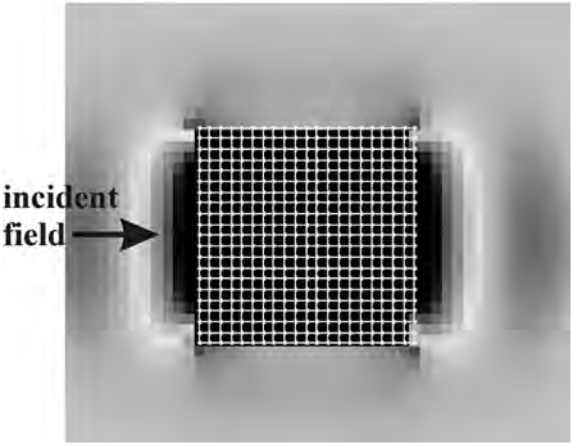


Figure 15.7 Scattered field magnitude due to a sinusoidal incident field on a perfectly conducting cube.

15.2.4. Finite Element Method (FEM)

R. Courant developed FEM in 1943 [23,24]. He used the Rayleigh-Ritz method for finding approximate solutions to variational problems by replacing the functions with appropriate combinations of basic elements then finding the minimum solution. The first step converts a boundary value problem into an equivalent variational problem. The equation for the variational form is given by

$$\begin{aligned} F_v(E) = {} & \frac{1}{2}\iiint_V \left[\frac{1}{\mu_r}(\nabla \times \boldsymbol{E})\cdot(\nabla \times \boldsymbol{E}) - k_0^2\varepsilon_r \boldsymbol{E}\cdot\boldsymbol{E}\right]dV \\ & + \frac{1}{2}\iint_S [\boldsymbol{E}\cdot(\hat{\boldsymbol{n}} \times \nabla \times \boldsymbol{E})]dS + \iiint_V \boldsymbol{E}\cdot\left[jk_0 Z_0 \boldsymbol{J}^i + \nabla \times \left(\frac{1}{\mu_r}\right)\boldsymbol{M}^i\right]dV \end{aligned} \tag{15.33}$$

and the equation for the weighted residual method is

$$\begin{aligned} F_w(E) = {} & \iiint_V \left[\frac{1}{\mu_r}(\nabla \times \boldsymbol{E})\cdot(\nabla \times \boldsymbol{W}) - k_0^2\varepsilon_r \boldsymbol{E}\cdot\boldsymbol{W}\right]dV \\ & + \frac{1}{2}\iint_S [\boldsymbol{W}\cdot(\hat{\boldsymbol{n}} \times \nabla \times \boldsymbol{E})]dS + \iiint_V \boldsymbol{W}\cdot\left[jk_0 Z_0 \boldsymbol{J}^i + \nabla \times \left(\frac{1}{\mu_r}\right)\boldsymbol{M}^i\right]dV \end{aligned} \tag{15.34}$$

where

V = volume containing the unknowns.
S = boundary enclosing the volume.
$\boldsymbol{W}$ = weighting function.
$\boldsymbol{J}^i$ = induced electric current source.
$\boldsymbol{M}^i$ = induced magnetic current source.

Equation (15.32) is also known as the *weak form* because the order of differentiation of the electric field in Eq. (15.32) is less than that of Eq. (15.31), or the strong form.

The FEM then divides an electromagnetic domain into many discrete, easy to analyze, polygon-shaped elements that conform to irregularly shaped subdivisions of the object. FEM begins with a model drawn in 1D, 2D, or 3D space using a preprocessor or a CAD drafting package. Next, automatic mesh generators create triangular/tetrahedral meshes throughout the model. Meshing is the process of breaking up a physical domain into smaller subdomains (elements). Surface domains may be subdivided into triangle or quadrilateral shapes, while volumes may be subdivided primarily into tetrahedra or hexahedra shapes. There are some requirements on the shape of elements. In general, the elements should be as equiangular as possible in equilateral triangles and regular tetrahedra. Highly distorted elements (long, thin triangles, squashed tetrahedra) lead to numerical instability. Manual meshing becomes necessary in regions where an automatic generator fails to create regular meshes. Connecting elements should have the same number of nodes along the common side. Areas with high gradients require a mesh with small elements—the finer the mesh, the better the results. Picking a good mesh density is an art. If the mesh is too coarse, then errors are too large. Alternatively, if the mesh is too fine, then computing time becomes unacceptably long. A fine mesh is necessary in regions having high parameter gradients, whereas a coarse mesh is sufficient elsewhere. The mesh must not have holes, self-intersections, or faces joined at two or more edges, and must conform to the boundary of the domain. Figure 15.8 shows a triangular mesh overlaying a model of a waveguide T junction. Note that the mesh is much finer around the rectangular inset at the top of the T, because the field variations are greater there.

The next step in the FEM is to select the interpolation function that approximates the unknown over an element. Polynomials are the most common because they are simple and have a limited extent. Nodal-based elements come from interpolating function values at the nodes. These elements are generally not used for vector electromagnetic fields, because they produce spurious modes and it is difficult to impose tangential boundary conditions. Edge-based elements overcome these limitations by assigning degrees of freedom to the edges instead of the nodes. The most common two-dimensional elements are rectangles and triangles, while the most common three-dimensional elements are bricks and tetrahedrals.

Assembly is the process of taking all the equations and developing a matrix equation to solve for the weights. All the element equations surrounding a given node are added

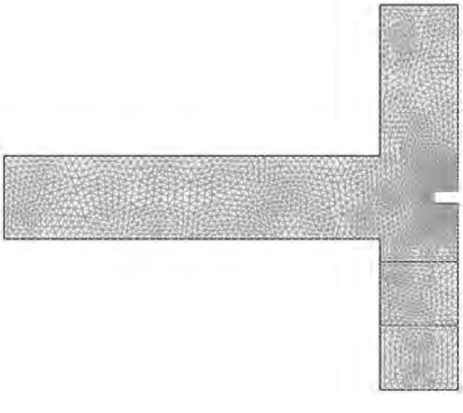


Figure 15.8 A T-junction waveguide is gridded for FEM.

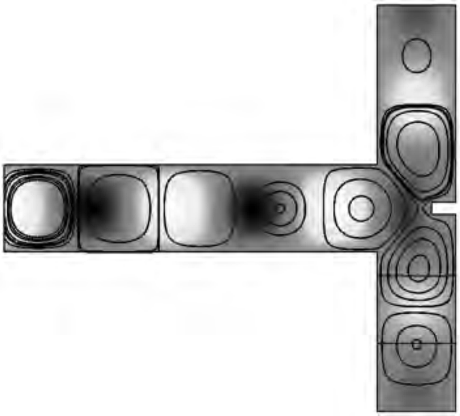


Figure 15.9 The electric field for the T-junction waveguide calculated using FEM.

together to get a single equation. The coefficients of this equation form a row in the matrix. An important step in the assembly process is establishing a global numbering scheme to keep track of all the nodes in the mesh. The assembled matrix is quite sparse. Finally, the boundary conditions are incorporated and the matrix equation solved using sparse solution methods. Figure 15.9 is a plot of the magnitude of the electric field for the T-waveguide problem as computed by FEM.

15.2.5. Other Techniques for Finding the Fields

In the generalized multipole technique (GMT), the boundaries of the problem are discretized then a number of radiating sources are placed off the boundaries. These sources act as basis functions that are analytical solutions to the field equations in the medium. The sources are weighted such that the boundary conditions are met in a least squares sense. Arranging the sources and boundary points are key to having a well conditioned matrix and good results. The number of boundary points should be much greater than the number of sources.

The transmission line method (TLM) makes use of the fact that Maxwell's equations are analogous to transmission line equations through the following equivalences:

$$E \leftrightarrow V \qquad H \leftrightarrow I \qquad \varepsilon \leftrightarrow 2C \qquad \mu \leftrightarrow L \tag{15.35}$$

where V is voltage, I is current, C is capacitance per unit length, and L is inductance per unit length. TLM models space and objects using a rectangular mesh of transmission lines. Each node has an associated scattering matrix. The reflected voltages are found by multiplying the scattering matrix (S) for the node by the incident voltages at the input ports.

$$\begin{bmatrix} V_1^r \\ V_2^r \\ V_3^r \\ V_4^r \end{bmatrix} = S \begin{bmatrix} V_1^i \\ V_2^i \\ V_3^i \\ V_4^i \end{bmatrix} \tag{15.36}$$

The three-dimensional version of TLM has six ports with two orthogonal polarizations per port. Thus, there are 12 incident and reflected voltages and the scattering matrix is 12 × 12. Although TLM is a time-domain method, frequency-domain information is obtained using a Fourier transform as was done in FDTD.

15.3. RECENT NUMERICAL TOOLS FOR ELECTROMAGNETICS

Section 15.2 presents the most common methods of finding fields. This section presents methods that make use of the field points generated by the numerical models in Sec. 15.2. Electromagnetics makes use of the many numerical methods developed in signal processing.

15.3.1. Model-Based Parameter Estimation (MBPE)

A few years ago, most calculations were done over a narrow bandwidth. The introduction of time-domain methods, wideband antennas, and high Q resonant circuits stimulated the need for very detailed computations to achieve the desired accuracies and not miss important features in the computed output. Many times the output is sampled and connected by straight lines (linear interpolation) to generate the output plots. A closer sampling distinguishes fine features but increases the computation cost. Uniform sampling usually means that some regions with slow varying details are oversampled, while regions with high variations are undersampled. More sophisticated interpolation like splines make use of derivative information to produce a smoother curve through the calculated data points.

MBPE is an interpolation/extrapolation technique for measured or computed data [27]. Unlike splines, polynomials, or Fourier series, MBPE uses interpolating functions derived from physical parameters of the problem. It is smart curve fitting. Complex exponentials are typical solutions to time-domain electromagnetic differential equations, while complex poles are typical solutions to frequency-domain electromagnetic differential equations. Consequently, exponentials and poles seem to be appropriate physically based curve fitting functions for electromagnetic problems:

$$q(t) = \sum_{m=1}^{M} A_m e^{s_m t} + q_{np}(t) \tag{15.37}$$

$$Q(f) = \sum_{m=1}^{M} \frac{A_m}{f - s_m} + Q_{np}(f) \tag{15.38}$$

where

M = number of terms.
q = waveform domain.
Q = transform domain.
A_m = residues.
s_m = poles.

q_{np} = nonpole component of the waveform function.
Q_{np} = nonpole component of the transform function.

The nonpole parts of (15.37) and (15.38) represent the nonresonant response. Equations (15.37) and (15.38) indicate that the transform pairs are the time and frequency domains. In electromagnetics, the frequency-space and space-angle transform pairs are also of great importance.

Prony's method was the original approach to MBPE [27]. This algorithm finds an infinite impulse response (IIR) filter with a prescribed time-domain impulse response. The classical method of Count de Prony models a sequence of $2p$ observations made at equally spaced times by a linear combination of p exponential functions. Prony's ingenious method converts the problem to a system of linear equations. Advances in signal processing have resulted in many other approaches to MBPE.

A closely related technique is the singularity expansion method (SEM) [28]. SEM characterizes an object's response in the time and frequency domains in terms of poles, branch cuts, and entire functions (singularities) in the complex frequency plane. Since most scattering objects have a transient response dominated by a small number of damped sinusoids, the damped sinusoids are poles of the Laplace transformed response. Natural frequencies or resonances are the basic starting ideas for SEM.

15.3.2. Optimization

The numerical techniques discussed so far find a single solution for specific problem parameters. Optimization finds the best set of problem parameters that yield an optimized or desired solution. Bounds or constraints may be placed on the parameters due to physical limitations, prior knowledge, or computational limits. Optimizing implies either finding a minimum or maximum of the output $(y_1, \ldots, y_N)$ from an objective function, F, given the input $(x_1, \ldots, x_N)$.

$$F(x_1, \ldots, x_N) = \{y_1, \ldots, y_N\} \tag{15.39}$$

Objective functions can have multiple inputs and multiple outputs. In computational electromagnetics, the objective function is a numerical model of an antenna, scattering object, microwave circuit, etc. Inputs to an antenna objective function may include parameters such as size, spacing, material properties, etc. Common output variables include gain, null depth, and sidelobe level. More than one solution must be generated in order to find the best set of parameters. Thus, optimization tends to be very time consuming.

Numerical optimization traditionally took two approaches: downhill methods or random methods. The downhill methods primarily rely upon derivative information to find a local minimum and are based on Newton's formula

$$v_{n+1} = v_n - \alpha_n Q_n^{-1} \nabla F(v_n) \tag{15.40}$$

where

v = vector containing the coordinates.
n = iteration number.
α_n = step size.

$Q_n = n$th approximation to the Hessian matrix $= H$.

$$H = \begin{bmatrix} \frac{\partial^2 F}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 F}{\partial x_1 \partial x_N} \\ \cdot & \cdots & \cdots \\ \frac{\partial^2 F}{\partial x_N \partial x_1} & \cdots & \frac{\partial^2 F}{\partial x_N \partial x_N} \end{bmatrix}$$

$\nabla F(v_n) =$ gradient of the objective function.

A myriad of techniques sprouted around solving Eq. (15.37). Some of the more popular include [29]

Steepest descent (in use for over one hundred years): $Q_n =$ identity matrix.
Newton's method: $Q_n = H =$ Hessian matrix.
Conjugate gradient: indirectly constructs Q_n.
Davidon-Fletcher-Powell (DFP): $Q_{n+1} = Q_N + C_{\mathrm{DFP}}$, where C_{DFP} is a correction term.
Broyden-Fletcher-Coldfarb-Shanno (BFGS): $Q_{n+1} = Q_N + C_{\mathrm{DFP}} + C_{\mathrm{BFGS}}$, where C_{BFGS} is a correction term.

The Nelder Mead downhill simplex algorithm [30] is commonly used by software packages like MATLAB, Mathematica, etc. The algorithm iteratively attempts to surround the optimum point with a simplex. A simplex is the most elementary geometrical figure that can be formed in dimension n and has $n+1$ sides (e.g., a triangle in two-dimensional space). Each iteration creates a new vertex for the simplex. The vertex corresponding to the highest function value is discarded. In this way, the simplex creeps towards the minimum. The simplex shrinks its diameter when it surrounds the minimum. The creeping and shrinking stop when the diameter reaches a specified tolerance. Since the algorithm does not use derivatives, it has a certain robustness that makes it attractive.

Optimization methods are classified as local or global. The downhill minimization algorithms are local because they start at a single point and move downhill to the local minimum. Practical problems often have many local minima. The local minimum found depends upon the initial starting point. Global optimization techniques incorporate random components that allow them to jump out of local minima and explore vast regions of the objective function space. A pure random search is just a guessing game and is rarely used.

Simulated annealing (SA) is random search based on the principles of thermodynamics [31]. The physical process of annealing occurs when a solid melts and its particles try to organize into a low-energy state during the cooling process. The probability that a particle is at a certain energy level is calculated by use of the Boltzmann distribution. As the temperature of the material decreases, the Boltzmann distribution tends toward the lowest energy particle configuration.

SA guesses at the optimum solution and then perturbs that solution. If the new cost (C) is less than the old cost, then it is accepted. If the new cost is greater than the old cost, then it is accepted if $P > p$ and rejected if $p < P$, where P is a uniform random number. The threshold probability, p, is given by

$$p = e^{-C/T} \tag{15.41}$$

The variable, T, corresponds to the temperature in the annealing process. T is slowly reduced so that the probability of accepting a higher cost decreases with time. The formula for reducing T is called the cooling schedule and is critical to the success of SA. Both the step size and T determine the convergence properties of the SA algorithm. Suggested step sizes and T values are approximately 80% of the higher costs accepted.

Another naturally based random search algorithm is the genetic algorithm (GA). The GA is a type of evolutionary algorithm that models the biological processes of genetics and natural selection to optimize highly complex objective functions. A GA helps a population composed of many individuals or potential solutions to evolve under specified selection rules to a state that contains the "most fit" individuals (i.e., minimizes the objective function, assuming it has been written so that the minimum value is the desired solution). The method was developed by John Holland [33] over the course of the 1960s and 1970s and popularized by his student, David Goldberg [34]. Michielssen first applied GAs to the design of radar absorbers in [35], and Haupt first used GAs for antenna design in [36]. An introductory article with the code for a very simple GA helped popularize GAs in electromagnetics [37].

The following explanation follows the flow chart in Fig. 15.10. The first step is defining an objective function with inputs and outputs. A binary GA encodes the value of each input parameter (e.g., a, b, c, d) as a binary number. The parameter values are then placed side-by-side in an array known as a *chromosome*. A population is a matrix with each row representing a chromosome. The algorithm begins with a population consisting of random ones and zeros (see Fig. 15.11). These random binary digits translate into guesses to values of the input parameters. Next, the binary chromosomes are converted to continuous values, which are evaluated by the objective function. Mating takes place between selected chromosomes. Mates are randomly selected with a probability of selection greater for those chromosomes yielding desirable output from the objective function (tournament or roulette wheel selection). Offspring (new chromosomes) produced from mating inherit binary codes from both parents. A simple crossover scheme randomly picks

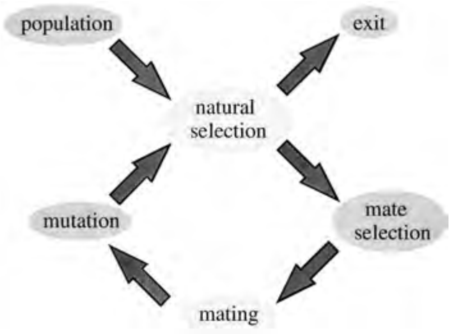
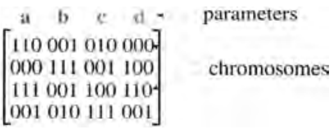


Figure 15.10 Flow chart of a genetic algorithm.



$$\begin{array}{cccc} a & b & c & d \\ \hline 110 & 001 & 010 & 000 \\ 000 & 111 & 001 & 100 \\ 111 & 001 & 100 & 110 \\ 001 & 010 & 111 & 001 \end{array}$$

Figure 15.11 The values of the parameters are encoded in a binary representation. All the parameters are placed in a chromosome, and the chromosomes are rows in the population matrix.

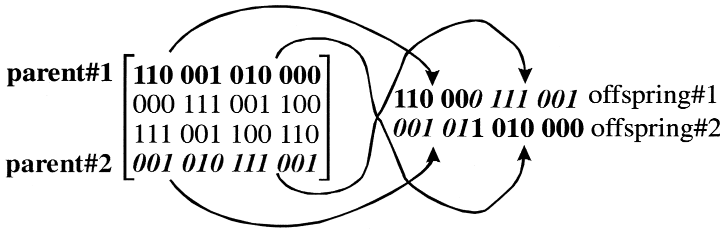


Figure 15.12 Two parents are randomly selected from the population matrix. A random crossover point splits the parents. Two new offspring are formed from parts of the parents.

a crossover point in the chromosome. Two offspring result by keeping the binary strings to the left of the crossover point for each parent and swapping the binary strings to the right of the crossover point, as shown in Fig. 15.12. Crossover mimics the process of meiosis in biology. Mutations randomly convert some of the bits in the population from "1" to "0" or visa versa. The objective function outputs associated with the new population are calculated and the process repeated. The algorithm stops after finding an acceptable solution or after completing a set number of iterations.

Selecting the best population size, mating scheme, and mutation rate is still an area of controversy. References 38 and 39 address this issue for electromagnetics problems. Since the GA is a random search, a certain population size and mutation rate can give considerably different answers for different independent runs. A GA run will give you a good answer found from a wide exploration of the search space but not necessarily the best answer.

Most real world optimization problems have multiple objectives, such as maximizing gain and maximizing bandwidth for the same antenna. Multiple objectives can be handled by weighting and adding the fitness from each objective. Multiobjective optimization does not have a single optimum solution relative to all objectives. Instead, there is a set of optimal solutions, known as Pareto-optimal or noninferior solutions. A Pareto GA attempts to find as many Pareto-optimal solutions as possible, since all these solutions have the same cost.

Some of the advantages of a GA include that it

Optimizes with continuous or discrete parameters.
Doesn't require derivative information.
Simultaneously searches from a wide sampling of the objective function surface.
Deals with a large number of parameters.
Is well suited for parallel computers.
Optimizes parameters with extremely complex objective function surfaces.
Provides a list of semioptimum parameters, not just a single solution.
May encode the parameters so that the optimization is done with the encoded parameters.
Works with numerically generated data, experimental data, or analytical functions.

15.3.3. Wavelets

Traditionally, we have thought of time domain signals as lasting forever. A function starts at $t = -\infty$ and continues until $t = +\infty$. Consequently, the Fourier transform is an efficient method of finding the spectrum or amplitudes of the frequency components

$$\Im(\omega) = \int_{-\infty}^{\infty} F(t)e^{j\omega t}\,dt \tag{15.42}$$

where $\omega = 2\pi f$. In reality, signals last for a finite duration. Even if they continue for a long period, our patience and computer limits stipulate that only a portion of the signal can be examined at any one time. Thus, most engineering problems do not use Eq. (15.42).

The more practical alternative to the Fourier transform is the short-time Fourier transform (STFT) or the windowed Fourier transform. The STFT windows or works with finite segments of the data.

$$\mathrm{STFT}(t,\omega) = \int F(\tau)w^*(\tau - t)e^{-j\omega t}\,d\tau \tag{15.43}$$

Many different windows (w) have been developed for various purposes. As an example, consider the linear chirp signal plus impulse at $t = 0.1$ s.

$$F(t) = \cos\left(100\pi t^2\right) + 2\delta(t - 0.1) \tag{15.44}$$

as shown in Fig. 15.13. Note that the frequency increases with time and there is an impulse function at $t = 0.1$ s. Figure 15.14 shows the STFT of (15.44). This plot nicely shows a linear increase in frequency with time but cannot accurately show the location of the impulse.

The STFT plot demonstrates the need for multiresolution analysis. A new approach was needed that could specify low-frequency signals accurately in frequency and high-frequency signals accurately in time. Precisely locating low-frequency signals in

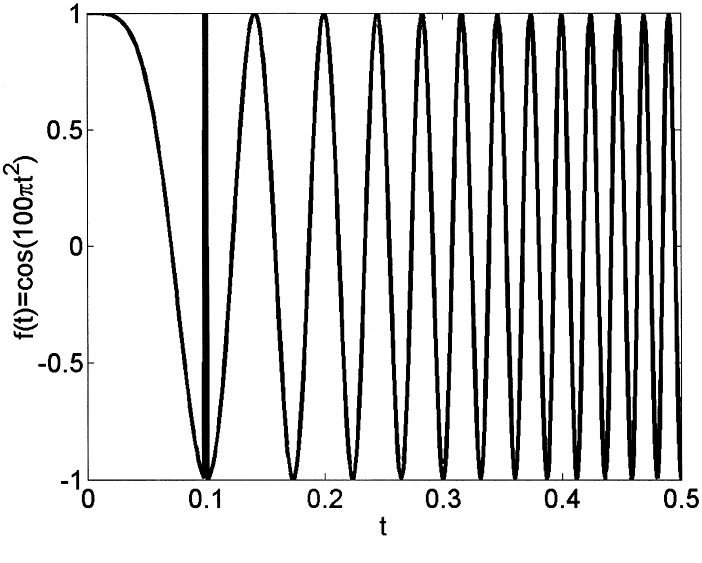


Figure 15.13 Linear chirp signal with an impulse at $t = 0.1$ s.

time is not critical, because they are slowly changing. On the other hand, fast changes require higher sampling rates. Wavelets provide the variable sampling capability that sinusoids cannot. References 40 and 41 provide an excellent introduction to time-frequency analysis.

A *wavelet* is defined to be any function that satisfies the following constraints:

1. The function has compact support (i.e., it has a definite start and end).
2. The area under the curve equals zero (i.e., the functions average value is zero).
3. The area under the wavelet is zero (i.e., no dc component).

A single cycle of a square wave satisfies that requirement. In fact, the first wavelet was the Haar wavelet and is one cycle of a square wave. Figure 15.15 shows a graph of the Mexican Hat wavelet given by the equation

$$\Psi(x) = \frac{2}{\sqrt[4]{\pi}\sqrt{3}}\left(1 - x^2\right)e^{-x^2/2} \tag{15.45}$$

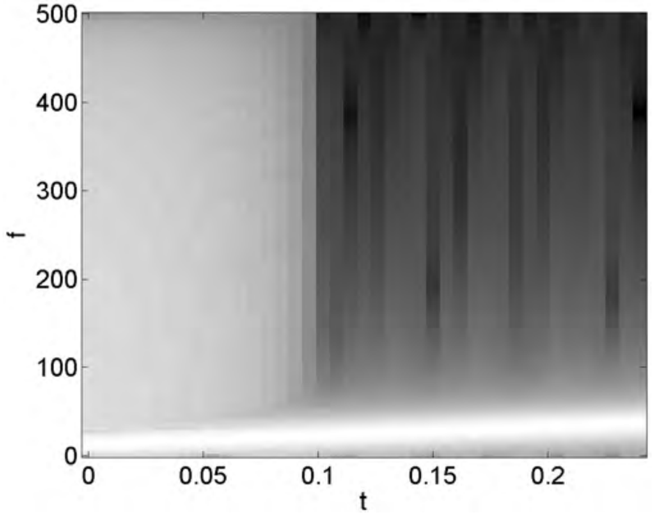


Figure 15.14 STFT of the linear chirp plus impulse signal.

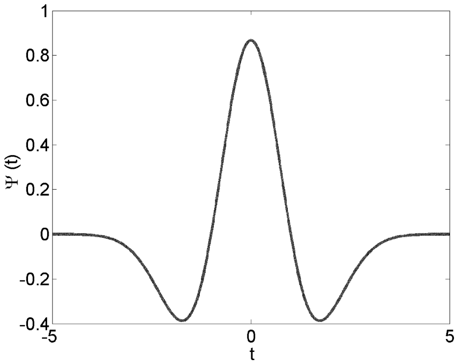


Figure 15.15 Mexican hat wavelet.

The continuous wavelet transform (CWT) is given by

$$\mathrm{CWT}_{a,b} = \frac{1}{\sqrt{|a|}} \int f(t)\Psi^{*}\left(\frac{t-b}{a}\right)dt \qquad a \neq 0 \tag{15.46}$$

where a is the scale index (inverse of frequency) and b is the time shift (translation). The CWT of the chirp signal in Eq. (15.44) is shown in Fig. 15.16. This plot shows the increase of frequency over time (remember that $a \propto 1/f$) and shows the precise location of the impulse at $t = 0.1$s.

The wavelet is an appropriate basis function when the scattering object is large and contains features with scales ranging from fractions of a wavelength to many wavelengths. The basis functions are shifted and dilated forms of a mother wavelet. Wavelets with a short expanse are used near edges or small features, while wavelets with a long expanse are used over large smooth features of the object. The resulting MOM matrix has many small elements. By applying a threshold level to the matrix elements, many can be set equal to zero. The resulting sparse matrix is of a form that can be quickly solved [42]. Figure 15.17 is a plot of a 200 × 200 impedance matrix with white indicating elements having a magnitude of at least 10% of the maximum magnitude in the matrix. This matrix has 11,944 elements with a magnitude of at least 10% of the maximum. A wavelet transform of that matrix has 4416 elements with a magnitude of at least 10% of the maximum (see Fig. 15.18). This 63% savings increases as the matrix gets bigger.

15.3.4. Hybrid Methods

Hybrid methods combine two or more of the solution methods described earlier. Building on the strengths of two techniques allows the modeling of more complex structures. Hybrid methods can also include the weaknesses inherent in both techniques. A combination of MOM and GTD is described in Ref. 1. Combining the frequency-domain MOM

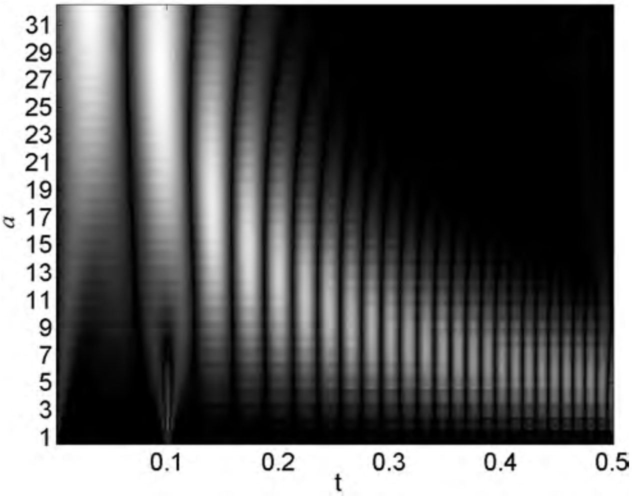


Figure 15.16 Continuous wavelet transform (with Mexican hat wavelet) of the linear chirp plus impulse signal.

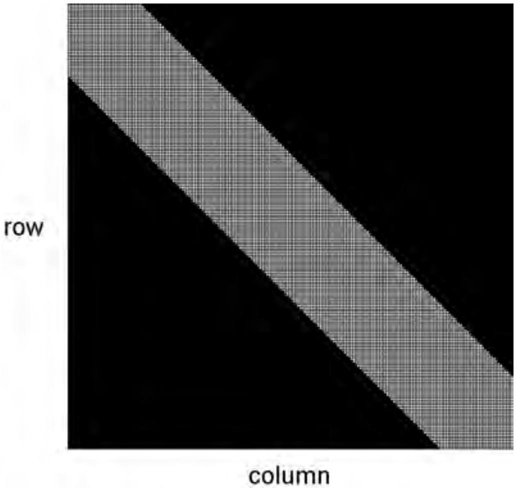


Figure 15.17 Magnitude of a typical MOM impedance matrix. The white elements have a magnitude that is at least 10% of the maximum magnitude element.

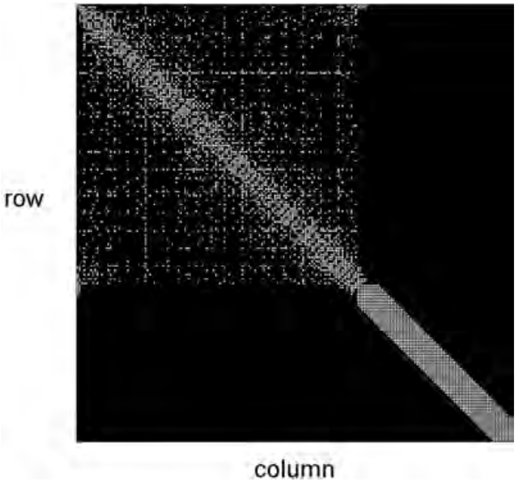


Figure 15.18 Magnitude of a wavelet transform of a typical MOM impedance matrix. The white elements have a magnitude that is at least 10% of the maximum magnitude element.

and the FDTD methods takes advantage of MOM's ability to solve exterior problems using patch models and of the ability of FDTD to model localized regions containing metal structures, dielectrics, permeable media, anisotropic or nonlinear media, as well as wires [43]. Another approach is to combine ray tracing and FDTD methods for site-specific modeling of indoor radio-wave propagation [44]. FDTD is only used to study areas close to complex discontinuities where ray-based solutions are not accurate. Since MOM and PO are current based methods, combining these approaches to solve for currents on large complex objects results in suitable accuracy in a reasonable computation time [45].

15.4. SOFTWARE CONCERNS

Developing the numerical method to solve an electromagnetic problem is the first step toward creating a useful computer program. Next, the programming language must be selected from the myriad available. Proper visualization of the results is essential to proper interpretation of the results. Finally, the code must be verified and validated in order to be accepted by users.

15.4.1. Programming Languages

The numerical solution of an electromagnetics problem may involve a programming language, general-purpose software, or specialized software. A programming language has the advantages of portability, fast execution, wide usage, and cheap or free software. The more popular languages include Java, C, C++, Fortran, and BASIC. Fortran and Basic are the primary languages used for various versions of the Numerical Electromagnetics Code (NEC). For the most part, programming languages must be compiled and take a long time to write and debug. Java is the newest of these languages and has gained popularity. Java is a good object-oriented language for quickly writing programs that run on multiple platforms and has found extensive use on the internet. Unfortunately, it is slower at mathematical operations than the other languages and lacks the extensive library functions for various numerical analysis routines.

General-purpose software has the advantages of fast program design, extensive prewritten routines, fast debugging, and excellent graphics. This type of software is designed to do basic mathematical operations and graphics. The most popular versions include MATLAB, Mathematica, MathCad, and Maple. These programs are interpreted, so they do not have to be compiled. Their advantages include very fast development time; run times as fast as programming languages; extensive mathematical, science, and engineering functions; excellent graphics; and symbolic mathematical manipulations. They tend to be very expensive except for steep academic discounts for student use. Portability can be an issue between different general-purpose software or even between old and new versions of the same software package.

Specialized software has excellent graphics, limited applications, few commands to learn, and usually operates with a GUI (graphical user interface). Specialized software is difficult to link with other software. For instance, combining a UTD code with a MOM code from two different vendors is at best a difficult endeavor. In addition, using a programming language or general-purpose software package to optimize the output of a specialized software package is difficult.

Developing your own software package today requires a GUI for easy interaction with users. An outstanding GUI encourages use by people that did not develop the code. Some pitfalls with GUI design include

Assuming the user knows too much
Limiting user access to the application
Placing too many features at the top level
Terms that are unclear and inconsistent
Being too verbose

Good GUIs are intuitive, consistent, and fast. Users also appreciate knowing how much longer a given operation will take before they can enjoy the fruits of their patience. Easy to use online help is necessary.

15.4.2. Visualization

Not long ago, visualization of electromagnetic fields required good abstract thinking. Today visualization means displaying the physical characteristics of the model as well as the electromagnetic fields associated with the model. Computer graphics quickly convey verbal and numerical information through imagery. A well-designed graphic should [46]

Show the data.
Induce thinking about the substance rather than about methodology, graphic design, or technology of graphic production.
Avoid distorting what the data have to say.
Make large data sets coherent.
Encourage the eye to compare different pieces of data.
Examine the data at several levels of detail.
Serve a reasonably clear purpose.
Be closely integrated with the statistical and verbal descriptions of a data set.

Common pitfalls in visualization are data distortion and putting too much data on a single plot. Distortion is easy to fall prey to with computer graphics that autoscale data. A sphere can look like an ellipsoid if not all axes are to the same scale. Graphics software does not warn you that there are too many lines on the plot or that fine detail of interest is obscured by the rest of the data.

Electromagnetics models are particularly difficult to visualize. A single graph cannot show

Three spatial dimensions
Time
All polarization components
Material properties of the objects
Currents, fields, and charge

Consequently, the software designer and user must decide how to represent the data. Some tradeoffs include still shots vs. movies, 2D vs. 3D, dB vs. magnitude, and color vs. symbol.

15.4.3. Verification and Validation

Any computer model must be validated and verified. Validation compares the computer output with known physical results. Equations and approximations along with other aspects of modeling the physical problem with a computer algorithm must be checked. Validation is the engineering and science part of the computer model. Accepted validation standards include mathematical expressions, experimental results, and other computer models. Validation is a continuous process that compares the computer output to new information as it becomes available.

Verification is the process of correctly solving the equations developed for the model. Unlike validation, it neglects errors caused by the choice of equation and parameters of the equation. It is the numerical analysis part of the computer modeling. Changes to the coefficients of an ill-conditioned numerical model result in large changes in the solution. A large condition number for a matrix implies that solving for the unknown vector in an equation that contains that matrix may result in significant errors.

Iterative solutions are prone to chaotic behavior due to the highly nonlinear formulation of the equation.

Experimentalists are used to plots that use error bars and solutions that have a degree of accuracy assigned. Numerical models rarely contain similar information. Many a novice falls prey to giving a numerical model some input and believing the output. Can you justify your results? Experience plays a major role. Verifying and validating numerical results should be a standard requirement for numerical models.

15.5. CONCLUSIONS

This chapter presented the four major numerical approaches to electromagnetics. The advantages and disadvantages of these methods are summarized in Table 15.1. No one technique is perfect. Hybrid methods are becoming more popular for very complicated problems due because they have the advantages of two or more of the numerical approaches.

Many challenges remain in computational electromagnetics. Computation speed is still too slow for most practical problems. Making use of specialized codes, parallel computers, and increasing clock speeds gradually move us toward modeling complex objects. Three-dimensional problems are still tricky and slow for most codes. Some progress has been made in using signal processing techniques to compress data and speed calculations. Cross-fertilization between these two fields needs to continue. Genetic algorithms have opened the possibility of not only modeling electromagnetic behavior but optimizing the design of electromagnetic systems as well. Visualization of results will continue to improve with even the incorporation of artificial intelligence and virtual reality to help. As codes become more complex, verifying and validating their results will become more challenging. Improved experimental measurements and the general availability of the measurement data will be extremely important.

As a closing note, there are many books available for the budding computational electromagneticist besides the ones already referenced.

Table 15.1 Characteristics of the Four Main Numerical Methods Used in Electromagnetics. E = excellent, G = good, and F = fair

	High frequency	MOM	FD	FEM
Single frequency	E	E	G	G
Transient	F	G	E	E
Materials	F	F	E	E
Thin objects	E	E	F	F
Far field	E	E	G	G
Large objects	E	F	F	F
Small objects	F	E	E	E
Ease of formulation	G	G	E	F
3D objects	F	F	E	E
Bandwidth	G	G	E	E

REFERENCES

1. Stutzman, W.L.; Thiele, G.A. *Antenna Theory and Design*; Wiley: New York, 1998.
2. Balanis, C.A. *Advanced Engineering Electromagnetics*; Wiley: New York, 1989.
3. Keller, J.B. Geometrical theory of diffraction. J. Opt. Soc. Am. **1962**, *52*, 116–130.
4. Ling, H.; Chou, R.; Lee, S.-W. Shooting and bouncing rays: calculating the RCS of an arbitrarily shaped cavity. IEEE Trans. Antennas Propagat. **1989**, *37*, 194–205.
5. Ufimtsev, P.Ia. Approximate computation of the diffraction of plane electromagnetic waves at certain metal bodies: PT. I. Diffraction patterns at a wedge and a ribbon. Zh. Tekhn. Fiz. (USSR) **1957**, *27*, 1708–1718.
6. Kouyoumjian, R.G.; Pathak, P.H. A uniform theory of diffraction for an edge in a perfectly conducting surface. Proc. IEEE **1974**, *62*, 1448–1461.
7. Mitzner, K.M. Incremental length diffraction coefficients. Technical Rep. No. AFAL-TR-73-296, Northrop Corp., Aircraft Division, Apr 1974.
8. Shore, R.A.; Yaghjian, A.D. Incremental diffraction coefficients for planar surfaces. IEEE AP-S Trans. **1988**, *36*, 55–70.
9. Veruttipong, T.W. Time domain version of the uniform GTD. IEEE AP-S Trans. **1990**, *38*, 1757–1764.
10. Harrington, R.F. *Field Computation by Moment Methods*. Robert E. Krieger Publishing Co.: Malabar, FL, 1968.
11. Canning, F.X. The Impedance Matrix Localization (IML) method for moment-method calculations. IEEE Antennas Propagat. Mag. **1990**, *32*, 17–30.
12. Peterson, A.F.; Ray, S.L.; Mittra, R. *Computational Methods for Electromagnetics*; IEEE Press: New York, 1998.
13. Shore, R.A.; Yaghjian, A.D. Dual surface integral equations in electromagnetics. International Union of Radio Science XXVIIth General Assembly, Maastricht, Netherlands, 2002.
14. Burke, J.G.; Poggio, A.J. Numerical Electromagnetic Code (NEC)—Method of Moments Parts I, II, and III. Technical Document No. 116, Lawrence Livermore National Laboratory, U.S.A., 1981.
15. Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P. *Numerical Recipes in C.*; Cambridge University Press: Cambridge, U.K., 1997, 71–89.
16. Haupt, R.L.; Haupt, S.E. An introduction to multigrid using matlab. Computer Appl. Engg. J. **1994**, *2*, 421–431.
17. Kunz, K.S.; Luebbers, R.J. *The Finite Difference Time-Domain Method for Electromagnetics*; CRC Press: Boca Raton, FL, 1993.
18. Taflove, A. *Computational Electrodynamics: The Finite Difference Time-Domain Method*; Artech House: Boston, 1995.
19. Yee, K.S. Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. IEEE AP-S Trans. **1966**, *14*, 302–307.
20. Georgakopoulos, S.V.; Birtcher, C.R.; Balanis, C.A.; Renaut, R.A. Higher order finite difference schemes for electromagnetic radiation, scattering, and penetration, part 1: theory. IEEE AP-S Magazine **2002**, *44*, 134–142.
21. Berenger, J.P. A perfectly matched layer for the absorption of electromagnetic waves. Journal of Computational Physics **1994**, *114*, 185–200.
22. Furse, C.M. Faster than Fourier: ultra-efficient time-to-frequency-domain conversions for FDTD simulations. IEEE AP-S Magazine **2000**, *42*, 24–33.
23. Jin, J. *The Finite Element Method in Electromagnetics*; Wiley: New York, 1993.
24. Volakis, J.L.; Chatterjee, A.; Kempel, L.C. *Finite Element Method for Electromagnetics*; IEEE Press: New York, 1998.
25. Ludwig, A.C. A comparison of spherical wave boundary value matching versus integral equation scattering solutions for a perfectly conducting body. IEEE Trans. On Antennas Propagat. **1986**, *34*, 857–865.
26. Christopoulos, C. *The Transmission-Line Modeling Method*; IEEE Press: Piscataway, NJ, 1995.

27. Miller Edmund, K.; Sarkar Tapan, K. Model-order reduction in electromagnetics using model-based parameter estimation. In *Frontiers in Electromagnetics*; Werner, D.H., Mittra, R. Eds.; IEEE Press: NY, 1999, 371–436.
28. Baum, C.E. The singularity expansion method: background and developments. IEEE AP-S Mag. **1986**, *28*, 15–23.
29. Luenberger, D.G. *Linear and Nonlinear Programming*; Addison-Wesley: Reading, MA, 1984.
30. Nelder, J.A.; Mead, R. Computer J. **1965**, *7*, 308–313.
31. Metropolis, N.; Rosenbluth, A.; Rosenbluth, M.; Teller, A.; Teller, E. Equation of state calculations by fast computing machines. J. Chem. Phys. **1953**, *21*, 1087–1092.
32. Kirkpatrick, S.; Gelatt Jr. C.D.; Vecchi, M.P. Optimization by simulated annealing. Science **1983**, *220*, 671–680.
33. Holland, J.H. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Ann Arbor, 1975.
34. Goldberg, D.E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: Reading, MA, 1989.
35. Michielssen, E.; Sajer, J.M.; Ranjithan, S.; Mittra, R. Design of lightweight broad-band microwave absorbers using genetic algorithms. IEEE Trans. Microwave Theory Tech. **1993**, *41*, 1024–1031.
36. Haupt, R.L. Thinned arrays using genetic algorithms. IEEE AP-S Transactions, 1994, *42*, 993–999.
37. Haupt, R.L.; An introduction to genetic algorithms for electromagnetics. IEEE Antennas Propagat. Mag. **1995**, *37*, 7–15.
38. Haupt, R.L.; Haupt Sue Ellen. *Practical Genetic Algorithms*; Wiley: New York, 1998.
39. Haupt, R.L.; Haupt, S.E.; Optimum population size and mutation rate for a simple real genetic algorithm that optimizes array factors. Appl. Computat. Electromagnet. Soc. J. **2000**, *15*, 94–102.
40. Rioul, O.; Vetterli, M. Wavelets and signal processing. IEEE Signal Proc. Mag. **Oct. 1991**, *11*, 14–38.
41. Strang, G. Wavelets. American Scientist **1994**, *82*, 250–255.
42. Steinberd, B.Z.; Leviatan, Y.; On the use of wavelet expansions in the method of moments. IEEE AP-S Trans. **1993**, *41*, 610–619.
43. Taflove, A.; Umashankar, K. A hybrid moment method/finite difference time domain approach to electromagnetic coupling and aperture penetration into complex geometries. IEEE Trans. Antennas Propagat. **1982**, *30*, 617–627.
44. Wang, Y.; Safavi-Naeini, S.; Chaudhuri, S.K. A hybrid technique based on combining ray tracing and FDTD methods for site-specific modeling of indoor radio wave propagation. IEEE Trans. Antennas Propagat. **2000**, *48*, 743–754.
45. Jakobus, U.; Landstorfer, F.M. Improved PO-MM hybrid formulation for scattering from three-dimensional perfectly conducting bodies of arbitrary shape. IEEE Trans. Antennas Propagat. **1995**, *43*, 162–169.
46. Miller, E.K.; Shaeffer, J. Theory, techniques and applications of electromagnetic visualization. Short Course Notes from the IEEE AP-S Symposium, Salt Lake City, UT, July 2000.

16

Biological Effects of Electromagnetic Fields

Riadh Habash
University of Ottawa
Ottawa, Ontario, Canada

16.1. INTRODUCTION

Electromagnetic (EM) fields have become a driving force of our civilization through their numerous applications. However, there are concerns about the hazards that might exist due to exposure to such fields. Actually, such concerns began as early as the eighteenth century, which saw rapid developments in medical applications and physiological effects of electricity and magnetism.

EM field is classified as either nonionizing or ionizing. There is a fundamental distinction made between ionizing field, which has enough energy to physically break chemical bonds at the molecular level, and nonionizing field, which does not. Nonionizing fields (frequencies below the ultraviolet range), which are the subject of this chapter, have photon energy less than 10 eV, a level not enough to produce ions by ejection of orbital electrons from atoms, but still have a strong effect, which is heating.

Investigations started after World War II, and much of the concern was directed toward possible health hazards of radio-frequency radiation (RFR). In the following years, with the help of the media, public concern diverted from RFR to electric and magnetic fields (EMFs). Also, attention shifted from the strong electric fields near high-voltage power lines to those relatively weak magnetic fields produced by distribution lines and electrical appliances. In recent years, concerns regarding RF exposure from mobile phones have grown considerably. These concerns are generated because of the wide use of such equipment and they are largely inflamed by the fact that the mobile phone is placed very close to the user's head.

This chapter traces the various components of the entire subject including interaction mechanisms, safety standards and protection guidelines, sources and exposure scenarios, description of large-scale epidemiological studies involving humans as well as research on exposure of cells and animals relevant to adverse health effects.

16.2. ELECTRIC AND MAGNETIC FIELDS

There are two types of EMFs classified according to the frequency range: extremely low frequency (ELF) fields and very low frequency (VLF) fields. ELF fields are defined as those having frequencies up to 3 kHz. VLF fields cover the frequency range 3–30 kHz. Because of the quasistatic nature of EM fields at these frequencies, electric and magnetic fields act independently of one another and are measured separately. Electric fields created by voltage and measured in volts per meter (V/m), are present whenever an electric appliance is plugged in. The appliance need not be turned on for electric fields to be detected. Magnetic fields, induced by alternating current (AC) and measured using the derived quantity magnetic flux density (B) in tesla (T) or gauss (G), are present when the appliance is turned on. The strength of EMFs decreases as we move away from their sources.

Any residential or occupational site is subject to coincident exposure from many EMF sources external and internal to the site itself. External sources include high-voltage power lines, distribution lines, underground cables, substations, transformers, and transportation systems. In the workplace, sources of EMFs include computers, fax machines, copy machines, fluorescent lights, printers, scanners, telephone switching systems (PBX), motors, induction heaters, electronic article surveillance (EAS), demagnetizers, security systems, and metal detectors. In homes, there are two immediate sources of EMFs. The first type includes internal wiring, meters, service panels, subpanels, and grounding systems. The second type includes electrical appliances such as electric blankets, electric waterbed heaters, hairdryers, electric shavers, television (TV) sets, video display terminals (VDTs), stereo systems, air conditioners, fluorescent lights, refrigerators, blenders, portable heaters, clothes washers and dryers, coffee makers, vacuum cleaners, toasters, and other household appliances.

16.2.1. Interaction Mechanisms

There are several proposed mechanisms for the interaction of EM fields with living systems. These include induced electric currents, direct effect on magnetic biological materials, effects on free radicals, and excitation of cell membranes.

Before discussing these mechanisms, one must understand the relationship between electric and magnetic fields outside and inside biological systems (coupling), which varies greatly with frequency. Electric fields are greatly diminished by many orders of magnitude inside biological tissues from their values in air external to the tissues. Biological tissues are nonmagnetic materials, which mean the magnetic field inside the human body is same outside it.

The first mechanism involves the ability, through magnetic induction, to stimulate eddy currents at cell membranes and in tissue fluids, which circulate in a closed loop that lies in a plane normal to the direction of the magnetic field. The above current can be calculated using only Faraday's law and Laplace's equations, without simultaneously solving Maxwell's equations. Both current and electric fields are induced inside living systems by external time-varying magnetic fields [2].

All living organisms are basically made of diamagnetic organic compounds, but some paramagnetic molecules (e.g., O_2), and ferromagnetic microstructures (hemoglobin core, magnetite) are also present. Biological magnetites are usually found in single domain units, covered with thin membranes called *magnetosomes* (Fe_3O_4). These microstructures behave like small magnets and are influenced by external fields changing their energy content. They are found in bacteria and other small biological elements. Such bacteria and biological elements orient along the applied magnetic fields.

According to Foster [3],

> Low-frequency electric fields can excite membranes, causing shock or other effects. At power line frequencies, the threshold current density required to produce shock is

around 10 A/m^2, which corresponds to electric field of 100 V/m in the tissue. Electric fields can create pores in cell membranes by inducing electric breakdown. This requires potential differences across the membranes at levels between 0.1 and 1 V, which, in turn, requires electric field in the medium surrounding the cell of at least 10^5 V/m.

Many life scientists through series of studies [4–6] believe that the cell membrane plays a principal role in the interaction of EM fields with biological systems. Indications point to cell membrane receptors as the probable site of initial tissue interactions with EM fields for many neurotransmitters, growth-regulating enzyme expressions, and cancer-promoting chemicals.

Scientists theorizing this mechanism conclude that biological cells are bioelectrochemical structures, which interact with their environment in various ways, including physically, chemically, biochemically, and electrically. According to Dr. William Ross Adey at the University of California, Riverside [7], "The ions, especially calcium ions could play the role of a chemical link between EM fields and life processes. The electrical properties and ion distribution around cells are perfect for establishing effects with external steady oscillating EM fields."

The impact of EM fields may also be understood in terms of amplification and/or the cooperative sensing associated with simultaneous stimulation of all membrane receptors. Litovitz et al. [8] hypothesized that oscillating EM fields need to be steady for certain period of time (approximately 1 s) for a biological response to occur. This allows cells to discriminate external fields from thermal noise fields, even though they might be smaller than the noise fields.

16.2.2. Laboratory Studies

Scientists look to laboratory studies as a source of information that will address concerns regarding likely health effects. Laboratory studies on cells or whole organisms play a key role in evaluating the response of different systems of the body. Laboratory studies are easier to control and provide the opportunity to check whether EMFs cause cancer or other illnesses, something that is not possible with human volunteers. However, laboratory studies entail complications especially those related to extrapolation to humans. Numerous health effects from EMFs have been discussed in the literature, but most of the attention has focused on possible relationship with DNA and cancer.

Numerous cellular studies referred to as in vitro have been carried out to find out if EMFs can damage DNA or induce mutations. In general, it is believed that the energy associated with EMFs is not enough to cause direct damage to DNA; however, it is understood that indirect effects might be possible by EMF changing processes within cells that could lead to DNA breakage. Meanwhile, EMFs well above environmental field intensities might enhance DNA synthesis, change the molecular weight distribution during protein synthesis, delay the mitotic cell cycle, and induce chromosome aberrations [9–11]. In contrast, EMFs, according to a number of studies [12–15], are unable to induce chromosomal aberrations even under relatively strong magnetic field exposure.

Studies of animals referred to as in vivo aim to determine the biological effects of EMFs on whole animals. Animal studies are very important because they supplement epidemiological studies and can provide a reliable model in which to look at exactly how EMFs characteristics cause the risk. There has been no absolute evidence in any study that low-level EMFs alone can cause cancer in animals. This is supported by the findings of many studies conducted during the last few years [16–18]. A study of animals treated with a known chemical initiator have shown greater numbers of tumors in those animals subsequently or concurrently exposed to magnetic fields at moderate to high exposure levels [19].

It is clear from the literature that the energy associated with EMF environmental exposures is not enough to cause direct damage to DNA or cause cancer in animals.

16.2.3. Melatonin Hypothesis

One possible interaction hypothesis under investigation is that exposure to EMFs suppresses the production of melatonin, which is a hormone produced by the pineal gland, a small pinecone-shaped gland located deep near the center of the brain. Melatonin is produced mainly at night and released into the blood stream to be dispersed throughout the body. It surges into almost every cell in the human body, destroying the free radicals and helping cell division to take place with undamaged DNA. Melatonin reduces secretion of tumor-promoting hormones. It has the ability to increase cytotoxicity of the immune system's killer lymphocytes; therefore, its production is essential for the immune system, which protects the body from infection and cancer cells. Various cancers might proliferate if melatonin is lowered such as breast cancer, prostate cancer, and ovarian malignancies. Figure 16.1 displays consequences of melatonin reduction.

Several studies [20–22] have found melatonin reduction in cells, animals, and humans exposed to EMFs. The effect varies according to the period of exposure and strength of EMFs.

In contrast, Rogers et al. [23,24] exposed baboons to 60-Hz fields at 6 kV/m plus 50 μT or at 30 kV/m and 100 μT (12 h/d for 6 weeks). They noticed no evidence of any effect on melatonin levels. Graham et al. [25] found also no effects on melatonin levels among young men volunteers exposed on four continuous nights to 60-Hz fields at 28.3 μT.

16.2.4. Human Studies

Effects of EMFs might be studied safely and effectively in the laboratory with human volunteers in spite of limitations to the duration of exposure and types of tests that are performed. Laboratory studies on humans have certain advantages. They focus directly on

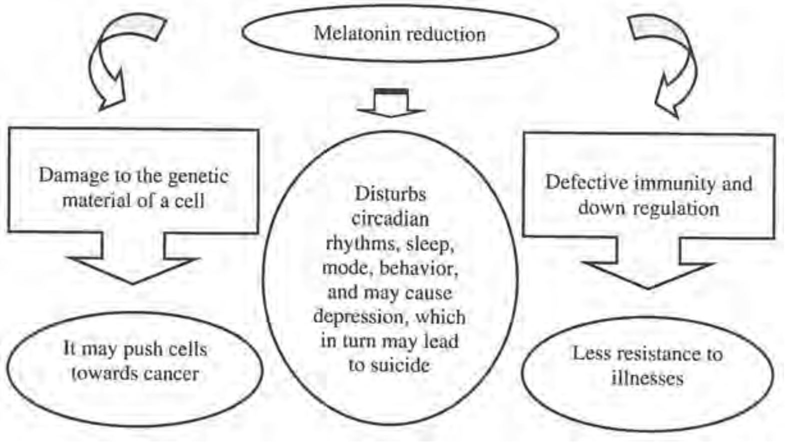

Melatonin reduction
Damage to the genetic material of a cell
Disturbs circadian rhythms, sleep, mode, behavior, and may cause depression, which in turn may lead to suicide
Defective immunity and down regulation
It may push cells towards cancer
Less resistance to illnesses

Figure 16.1 Biological consequences of melatonin reduction.

the "right" species, therefore avoiding the problem of extrapolation from data obtained in other species. Even negative results can be of immediate use in addressing public concerns. Such studies may also be used to directly evaluate the effects of exposure on "real-life" functions. The main sources of information in this field are surveys of people and workers living close to potential sources of EMFs, laboratory tests, and epidemiological data.

EMFs may affect the brain and nervous system and may cause effects on normal behavior or cognitive abilities of humans have been a persistent concern. In the early studies of occupational exposure to EMFs [26], switch yard workers in the former Soviet Union who differed in the duration and intensity of their exposure to 50-Hz fields suffered from an abnormally high incidence of neurophysiological complaints. A recent review on behavioral effects of EMFs was conducted by Zenon [27].

Heart rate and blood pressure may assess cardiovascular functions. Current densities of about $0.1\,\mathrm{A/m^2}$ can stimulate excitable tissues, while current densities above about $1\,\mathrm{A/m^2}$ interfere with the action of the heart by causing ventricular fibrillation, as well as producing heat. For example, Sazonova [26] observed that the pulse rates of people among workers with an average exposure of 12–16 kV/m for more than 5 h/d were lower by 2–5 beats/min at the end of the day, although they had been equivalent at the start of the day. According to a review by Stuchly [28], exposure of healthy male volunteers to 20-μT electric and magnetic fields at 60 Hz has been linked to a statistically significant slowing of the heart rate and to changes in a small fraction of the tested behavioral indicators. Korpinen et al. [29] used ambulatory recording techniques to carry out an extensive study on the effects of EM occupational exposure on heart rate. No field-related changes in mean heart rate were found as a result of exposure to 50-Hz fields directly under power lines ranging in intensity from 110 to 400 kV.

16.2.5. Epidemiological Studies

Epidemiological studies address the observed effects of possibly harmful EMF exposure on human health and whether the level of exposure is related quantitatively to the severity of health effects. These studies are limited in the sense that they are indirect experiments where the exposure can only be assessed through different substitute measures. An epidemiological association, if found, might not be related directly to exposure, it may be due to chance, confounding factors, or some unrecognized factors related to the way the data have been collected.

Childhood Leukemia

Childhood is a critical period of rapid cell growth and the cancer development cycle is correspondingly much quicker than adults. In addition, a child's immune system is underdeveloped and melatonin production is lower. Childhood exposure to EMFs has been studied intensively for many decades. However, research into this area gained momentum in 1979, when one of the first epidemiological studies [30] showed an association between exposure to EMFs and cancer among children living near power lines. Many studies have since been conducted but they do not support the notion that EMF exposure increases the risk of childhood cancer [31–36].

The association between EMF exposure and childhood cancer is inadequate and inconclusive. Some studies showed a link but their findings have small risk magnitudes by epidemiological standards with odds ratio $(\mathrm{OR}) < 5$ and were unable to exclude other environmental influences.

Adult Cancers

Occupational exposure was studied considering various health problems as well as adult cancers, including brain tumors and leukemia [37–42]. Sahl et al. [37] studied utility workers at Southern California Edison. Comparisons in the cohort study focused on electrical versus nonelectrical workers, and exposure was characterized on the basis of job history. The authors noticed no difference in risk for brain cancer among electrical workers compared to the reference group. However, small but significant increases in brain cancer risk were observed for electricians with risk ratio (RR) = 1.6 and plant operators (RR = 1.6).

Researchers from Canada and France [38] conducted a study of 223,292 workers at three large utilities, two in Canada (Hydro Quebec and Ontario Hydro) and a national utility in France (Electricite de France). The result shows that workers with acute myeloid leukemia (AML) were about three times more likely to be in the half of the workforce with higher cumulative exposure to magnetic fields. In the analysis of median cumulative magnetic field exposure, no significant elevated risks were found for most types of cancer studied.

The elevated risks of leukemia were also seen among senior workers who spent the most time in electric fields above certain thresholds, in the range of 10 to 40 V/m [41]. In a recent Canadian population-based control study, Villeneuve et al. [42] conducted a study among men in eight Canadian provinces, for 543 cases of brain cancer confirmed histologically (no benign tumors included). Astrocytoma and glioblastomas accounted for over 400 of these. Population based controls (543) were selected to be of similar age. They reported a nonsignificant increased risk of brain cancer among men who had ever held a job with an average magnetic field exposure $> 0.6\,\mu T$ relative to those with exposures $< 0.3\,\mu T$. A more pronounced risk was observed among men diagnosed with glioblastoma multiforme (the most malignant of neuroepithelial neoplasms) (OR = 5.36). There are rather notable differences in adult cancer studies with two kinds of results: (1) null association [37,40] and (2) mixed but in general strongly positive results from Canada-France study [38] and Canadian senior workers Villeneuve et al. [41,42]. RRs in the upper exposure categories were above 2.0 and for the more highly exposed groups between 1.1 and 1.3. RRs of this magnitude are below the level at which a casual association between EMFs and cancer can be assessed.

16.2.6. Safety Standards and Protection Guidelines

Safety standard is a standard specifying measurable field values that limit human exposure to levels below those deemed hazardous to human health. The standard consists of regulations, recommendations, and guidelines that would not endanger human health.

There are many worldwide institutions and organizations that have recommended safety limits for EM exposure. These include the Institute of Electrical and Electronic Engineers (IEEE) [43], the National Radiological Protection Board (NRPB) of the United Kingdom [44], the International Commission on Non-Ionizing Radiation Protection (ICNIRP) [45], the Swedish Radiation Protection Institute [46], Safety Code 6 of Canada [47], and Australian Radiation Protection and Nuclear Safety Agency (ARPANSA) [48].

Most of the protection guidelines use a two-tier standard, indicating a basic restriction (current density) and corresponding investigation levels or reference levels (external field strengths). The exposure limits range from few microtesla (μT) up to

1300 μT. The levels for those occupationally involved in various electrical industries are set higher than those for the general public.

The IEEE has a standard covering exposures above 3 kHz but, at present, does not have a standard covering the lower frequencies relevant to the electricity power system. However, a new standard is being prepared by Subcommittee 28 that will be based on known interactions of internal electric fields with the different parts of the nervous system.

The recommended NRPB guidelines are same for occupational and public environments. The basic restriction specified by the NRPB is an induced current density of 10 mA/m^2 in the head and trunk, while the investigation levels for electric and magnetic fields at 50 Hz are 12 kV/m and 1600 μT, respectively.

Sweden has been a leader in developing recommended visual ergonomic and EM emission standards for computer displays. The Swedish Confederation of Professional Employees, or TCO, which represents over a million workers, published its own series of guidelines [46], which include guidelines for energy consumption, screen flicker, luminance, and keyboard use.

16.3. RADIO-FREQUENCY RADIATION

As defined by the Institute of Electrical and Electronics Engineers (IEEE), RFR is a band in the electromagnetic spectrum that lies in the frequency range of 3 kHz to 300 GHz. Microwave (MW) radiation is usually considered a subset of RFR, although an alternative convention treats RF and MW as two separate spectral regions. Microwaves occupy the spectral region between 300 MHz to 300 GHz, while RF includes 3 kHz to 300 MHz. Since they have similar characteristics, RF and MW are recognized together, and referred to as RFR throughout this chapter.

Many frequencies of RFR are used in various applications. For example, the frequency range of 5 to 16 kHz is used in AM radio transmission, while 76 to 108 MHz is used for FM radio. Cellular and personal communication uses frequencies between 800 MHz and 3 GHz. The 2.45 GHz is reserved for industrial, scientific, and medical (ISM) applications, mainly microwave cooking.

The interaction of RF fields with living systems, and consequently their related bioeffects, can be considered at various levels including the molecular, subcellular, organ, system level, or the entire body. Biological effects due to RF exposure are classified as high-level (thermal) effects, intermediate-level (athermal) effects, and low-level (nonthermal) effects.

16.3.1. Thermal Effects

An obvious outcome of RFR absorption by the human body is heating (thermal effect), where the core temperature of the body rises despite the process of thermoregulation by the body. Many of the biological effects of RFR that have significant implications for human health are related to induced heating or induced current. Heating is the primary interaction of RF fields at high frequencies especially above about 1 MHz. Below about 1 MHz, the induction of currents in the body is the dominant action of RFR. Heating from RFR best relates to specific absorption rate (SAR) rather than to incident power density to account for differences in coupling.

Biological systems alter their functions as a result of a change in temperature. It is worth mentioning that most adverse health effects due to RF exposure between 1 MHz and 10 GHz are consistent with responses to induced heating, resulting in raising tissue temperatures higher than 1°C. Elevated temperatures have obvious effects on humans such as increased blood pressure, dizziness, weakness, disorientation, and nausea.

16.3.2. Athermal and Nonthermal Effects

Controversy surrounds two issues regarding biological effects of intermediate- and low-level RFR. First, whether RFR at such levels can even cause harmful biological changes in the absence of demonstrable thermal effects. Second, whether effects can occur from RFR when thermoregulation maintains the body temperature at the normal level despite the EM energy deposition or when thermoregulation is not challenged and there is no significant temperature change. In response to the first issue, investigations on the extremely low-level RFR have been established and some results confirmed but knowledge is yet inconclusive.

Regarding the second issue, a biological effect may have two meanings. It may mean an effect that occurs under circumstance of no evident change in temperature or the exposure level is low enough not to trigger thermoregulation in the biological body under irradiation, suggesting that physiological mechanisms maintain the exposed body at a constant temperature. Such case is related to nonthermal effect where the effect occurs through mechanisms other than those due to macroscopic heating. The second meaning is that RFR causes a biological effect, without the involvement of heat. This is sometimes referred to as *athermal effect*.

16.3.3. Toxicological Studies

Health effects are often the result of biological effects that accumulate over time and depend on exposure dose. For example, if an effect of EM exposure has been noticed on cultured cells, this does not essentially mean that the exposure will lead to adverse effect for the health of the organism as a whole. In general, the number of cellular and animal studies in the literature is large due to the large number of cellular processes and systems that may probably be affected by RFR.

A relationship between RFR and cancer would indicate that RFR somehow induces mutations in the DNA which, in turn, can disrupt cell growth and developing, leading to cancer. A number of laboratory experiments have been conducted to assess possible effects of RFR on genetic material. Investigations on different cell systems found no evidence for any direct genotoxic or mutagenic effects of continuous and pulsed RFR at different power densities. Tice et al. [49], as a part of comprehensive investigation of the potential genotoxicity of RF signals emitted by mobile phones, demonstrated that under extended exposure conditions, RFR from mobile phones at an average SAR of at least 5 W/kg are capable of inducing chromosomal damage in human lymphocytes. Similar findings were reported by d'Ambrosio et al. [50] while radiating human cells to 1748 MHz at 5 W/kg and by Mashevich et al. [51] when radiating human lymphocytes to continuous 830-MHz RF energy at SAR in the range 1.6–8.8 W/kg for 72 h. These results show that RFR has a genotoxic effect. Since the positive findings in the literature were consistently associated with hyperthermia, it will be concluded that RFR at low-intensity levels do not induce any genetic damage under nonthermal conditions.

Disturbance of normal cell cycle is a possible sign of uncontrolled cell growth, or cancer. Czerska et al. [52] reported an increased proliferation of cells exposed to 2.45-GHz RFR at SAR of 1 W/kg when the radiation was pulsed. Continuous wave (CW) RFR increased proliferation only when absorbed energy was high enough to induce heating. Other investigators reported increased and decreased cell proliferation rates after applying

RFR of various SARs [53–55]. In contrast, d'Ambrosio et al. [50] found no significant changes in cell distribution or cell proliferation in cells exposed to 1748 MHz, either CW or phase only modulated wave (GMSK) for 15 min.

16.3.4. Noncancerous Effects on Animals

While most of experimental studies focus on carcinogenesis, tumor promotion, and mutagenic effects, other noncancer effects also need to be considered. RFR can induce morphological and physiological changes. According to Adey et al. [56], RF carriers sinusoidally modulated at ELF fields can induce changes to the CNS. However, Tsurita et al. [57] found no significant changes in the groups of rats exposed for 2–4 weeks to a 1439-MHz (2 W/kg) TDMA signal on the morphological changes of the brain. The exposure period was 2 or 4 weeks.

RFR can induce cataracts if the exposure intensity and the duration are sufficient. Many studies on the ocular effect of RFR on animals have reported no effects, despite the fact that most studies employed exposure levels greatly in excess of that seen with mobile phones [58,59].

Some changes in learning behavior occurred after RF exposure. Lai et al. [60] observed retarded learning of a task in rats exposed to 2.45 GHz. However, Bornhausen and Scheingraber [61] found that exposure in utero to the GSM (900 MHz, 217-Hz pulse-modulated RFR; 17.5 and 75 mW/kg) field did not induce any measurable cognitive deficits in exposed Wistar rats during pregnancy. Dubreuil et al. [62] noted that head-only exposure of rats to 900 MHz pulsed RFR (SAR of 1 or 3.5 W/kg) for 45 min had no effect on learning.

RFR-induced breakdown of the blood–brain barrier (BBB) have been studied either alone or in combination with magnetic fields. Many authors agree that exposure to RFR affects BBB in vivo [63–65]. However, other studies have not found RFR-induced disruption of the BBB [66,67].

16.3.5. Human Studies

Ocular Effects

The cornea and lens are the parts of the eye most exposed to RFR at high levels by their surface location and because heat produced by the RFR is more effectively removed from other eye regions by blood circulation.

One related modeling study of the human eye by Hirata et al. [68] showed that 5 mW/cm^2 caused a temperature change in the lens less than 0.3°C at frequencies from 0.6 to 6 GHz. This small temperature change is overestimated because the eye model was thermally isolated from the head and did not consider blood flow. Therefore, RF exposures much in excess of currently allowable exposure limits would be required to produce cataracts in human beings, and exposures below the cataractogenic level would be expected to cause other effects in other parts of the eye and face.

Reviews of the literature of RFR-induced cataracts [69,70] concluded that clinically significant ocular effects, including cataracts, have not been confirmed in human populations exposed for long periods of time to low-level RFR.

Brain Functions

The close placement of RFR sources such as mobile phones to the user's head has elevated possibilities of interference with brain activities. While many studies have addressed this issue, they have only investigated the short-term effects of RFR. The controversial findings in the literature suggest that some aspects of cognitive functions and measures of brain physiology may be affected without offering a uniform view. These include changes in memory tasks, response patterns, normal sleeping EEG patterns, and other brain functional changes. Subjective symptoms such as dizziness, disorientation, nausea, headache, and other unpleasing feelings such as a burning sentient or a faint pain might be a direct result of RFR although such symptoms are very general and may have many causes.

The actual outcomes from the majority of studies have no serious implications for human health since the effects were seen for just a few of many tests and they were far too small to have any serious functional significance.

In a review, Hossmann and Hermann [71] concluded that "Most of the reported effects are small as long as the radiation intensity remains in the nonthermal range. However, health risks may evolve from indirect consequences of mobile telephony, such as the sharply increased incidence rate of traffic accidents caused by telephony during driving, and possibly also by stress reactions which annoyed bystanders may experience when mobile phones are used in public places."

Cardiovascular System

Jauchem [72] reviewed cardiovascular changes in humans exposed to RFR. Both acute and long-term effects were investigated. The author reported that most studies showed no acute effect on blood pressure, heart rate, or electrocardiogram (ECG) waveform; others reported subtle effects on the heart rate.

16.3.6. Epidemiological Studies

Navy Personnel

Robinette et al. [73] conducted a study of mortality results on males who had served in the U.S. Navy during the Korean War. They selected 19,965 equipment-repair men who had occupational exposure to RFR. They also chose 20,726 naval equipment-operation men who, by their titles, had lower occupational exposure to RFR as a control group. The researchers studied mortality records for 1955–1974, in-service morbidity for 1950–1959, and morbidity for 1963–1976 in veterans administration hospitals. No difference on cancer mortality or morbidity was seen among the high-exposure and low-exposure groups.

Military Workers

Szmigielski [74] showed strong association between RF exposure and several types of cancer (including brain cancer and cancer of the alimentary canal) was reported in a cohort of about 120,000 Polish military personnel, of whom 3% had worked with RF heat sealers. Exposure was determined from assessments of field levels at various locations. The study did not consider the length of time at the location, the nature of the job, or the number of cases observed.

Traffic Radar Devices

Davis and Mostofi [75], in a brief communication, reported six cases of testicular cancer in police who used handheld radars between 1979 and 1991 among a cohort of 340 police officers employed at two police departments within contiguous counties in the north-central United States. The six cases had been employed as police officers as their primary lifetime occupation, and all had been exposed to traffic radar on a routine basis. The mean length of service prior to testicular-cancer diagnosis was 14.7 y, the mean age at diagnosis was 39 y, and all had used radar at least 4½ y before the diagnosis.

Finkelstein [76] presented the results of a retrospective cohort cancer study among 22,197 officers employed by 83 Ontario police departments. The standardized incidence ratio (SIR) for all tumor sites was 0.90. There was an increased incidence of testicular cancer (SIR = 1.3) and melanoma skin cancer (SIR = 1.45). No information about individual exposures to radar devices was provided.

16.3.7. RF Heat Sealers

Lagorio et al. [77] reported higher cancer mortality among Italian plastic ware workers exposed to RFR generated by dielectric heat sealers for the period 1962–1992. Six types of cancers were found in the exposed group. The standardized mortality ratio (SMR) analysis was applied to a small cohort of 481 women workers, representing 78% of the total person-years at risk. Mortality from malignant neoplasms was slightly elevated, and increased risks of leukemia and accidents were detected. The all-cancer SMR was higher among women employed in the sealing. Exposure assessment was based on the time assigned on jobs. Exposure to RFR was based on a previous survey, which showed that the radiation exceeded 1 mW/cm^2. The work area also included exposure to chemicals associated with cancer (solvents and vinyl chloride), which may have impact on the result.

Telecom Operators

In Norway, Tynes et al. [78] studied breast cancer incidence in female radio and telegraph operators with potential exposure to light at night, RFR (405 kHz–25 MHz), and ELF fields (50 Hz). The researchers linked the Norwegian Telecom cohort of female radio and telegraph operators working at sea to the Cancer Registry of Norway to conduct their study. The cohort consisted of 2619 women who were certified to work as radio and telegraph operators. The incidences of all cancers were not significant, but an excess risk was seen for breast cancer. They noted that these women were exposed to light at night, which is known to decrease melatonin levels, an expected risk factor for breast cancer.

Radio and Television Transmitters

An association between proximity of residences to TV towers and an increased incidence of childhood leukemia was found in an Australian study conducted by Hocking et al. [79]. The researchers studied the leukemia incidence among people living close to television towers (exposed group) and compared this to the incidence among those living further out from the towers (unexposed or control group). People were assigned to one of the two groups based on data from the New South Wales Cancer Registry and their accompanying address. The Hocking study concluded that there was a 95% increase in childhood leukemia associated with proximity to TV towers. No such association was found between RFR emitted by the TV towers and adult leukemia. McKenzie et al. [80] repeated the Hocking study, using more accurate estimates of the exposure to RFR. The researchers looked at the same area and at the same time period, but with more accurate estimates of the RF exposure that people received in various areas. They found increased childhood leukemia in one area near the TV antennas but not in other similar areas near the same TV antennas. They found no significant correlation between RF exposure and the rate of childhood leukemia. They also found that much of the "excess childhood leukemia" reported by the Hocking study occurred before high-power 24-h TV broadcasting had started.

In Italy, Michelozzi et al. [81] conducted a small area study to investigate a cluster of leukemia near a high-power radio transmitter in a peripheral area of Rome. The leukemia mortality within 3.5 km (5863 inhabitants) was higher than expected. The excess was due to a significant higher mortality among men (seven cases were observed). Also, the results showed a significant decline in risk with distance from the transmitter, only among men.

Mobile Phones

Most of the mobile phone studies (Table 16.1 [82–88]) reported no increased incidence of brain tumors among mobile phone users (analog or digital phones). Furthermore, there was no relationship between brain tumor incidence and duration of mobile phone use. Only one group of researchers in Sweden [82] has reported associations between analog phone use and brain tumors. Their results have found no support in the investigation of other researchers. It is also doubtful whether results for analog phone users can be extrapolated to digital phone users.

16.3.8. Safety Standards and Exposure Guidelines

How much RF energy is safe? This is a complex problem, comprising public health, life sciences, engineering, and social (including economic and legal) considerations. Currently, there are various safety standards established for RF exposure in most of the industrial world (Table 16.2 [43–48,89]).

SAR is the rate at which RF energy is absorbed by the tissue and thus is a good predictor of thermal effects. SAR is defined as

$$\mathrm{SAR} = \frac{\sigma |E|^2}{\rho} = c\frac{dT}{dt}$$

where E is the effective value of the electric field intensity (V/m), dT/dt is the time derivative of the temperature (K/s), σ is the electrical conductivity (S/m), ρ is the mass density (kg/m^3), and c is the specific heat (J/kg K). The unit of SAR is W/kg. The SAR is the dosimetric measure that is used for extrapolating across species.

SAR calculations and estimates usually use many EM properties of biological tissues (e.g., complex dielectric constants and conductivity of different tissues) whose accuracy depends on their acquisition techniques, which are mostly in vivo.

There are two major types of SAR: (1) a whole-body average SAR and (2) a local (spatial) peak SAR when the power absorption takes place in a confined body region, as in the case of the head exposed to mobile phone. Whole-body SAR measurements are significant to estimate elevations of the core body temperature. As SAR increases, the possibility for heating and, therefore, tissue damage also rises. The whole-body SAR for a given organism will be highest within a certain resonant frequency range, which is dependent on the size of the organism and its orientation relative to the electric and magnetic field vectors and the direction of wave propagation. For the average human the

Table 16.1 Summary of Epidemiological Studies of Cellular Phones and Cancer Risk

Investigator	Description[a]	Risk measure	Outcome
		Brain tumors	

Hardell et al., 1999 [82]	CC: Sweden (1994–1996);
	(GSM/NMT phones);
	209 brain tumor cases;
	425 controls.
		OR = 0.98 (0.69–1.41);
		Same side of the head: OR = 2.42 (0.97–6.05)
			Right brain tumors for users who used the phone at their right ear.
			Stronger for temporal or occipital localization of the tumor on right side (only for analog phones).
			Temporal or occipital localization of the tumor on the same side as phone use for the left side use.

Muscat et al., 2000, 2002 [83,84]	CC: USA (1994–1998);
	469 brain cancer;
	422 controls.
		OR = 0.85 (0.6–1.2)	No significant association between primary brain cancer and years of mobile phone use, number of hours of use per month, or the cumulative number of hours of use.

Inskip et al., 2001 [85]	CC: USA (1994–1998);
	489 Glioma;
	197 Meningiomad;
	96 Acoustic neuroma;
	799 controls.
		OR = 1.0 (0.6–1.5);
		Glioma: 0.9 (0.5–1.6);
		Meningioma: 0.2 (0.3–1.7);
		Acoustic neuroma: 1.4 (0.6–3.5).
			The results do not support the existence of an association between mobile phone use and certain cancers (glioma, meningioma, or acoustic neuroma). There was no difference for side of head.

Johansen et al., 2001 [86]	CE: Denmark (1982–1995);
	420,095 users from two operators; 3391 cancers;
	3825 expected.
		SIR = 0.89 (0.86–0.92);
		Brain: SIR = 0.95 (0.81–1.12);
		Salivary gland:
		SIR = 0.72 (0.29–1.49);
		Leukemia: SIR = 0.97 (0.78–1.21).
			No relationship between brain tumor risk and RF dose compared by duration of phone use, date since first subscription, age at first subscription, or type of phone used.

Auvinen et al., 2002 [87]	CC: Finland (1996);
	398 brain tumors;
	198 gliomas;
	34 salivary gland;
	5 controls per case.
		Brain tumor: OR = 1.3 (0.9–1.8);
		Salivary gland: OR = 1.3 (0.4–4.7);
		Gliomas: OR = 2.1 (1.3–3.4) (Analog); Gliomas:
		OR = 1.0(0.5–2.0) (Digital).
			No clear association between use of mobile phones and risk of cancer has been provided. Gliomas were associated with the use of analog but not digital phones.

		Melanoma of the eye	
Stang et al., 2001 [88]	HBPBCC: Germany; (1994–1997); 118 case; 475 control.	OR = 3.0 (1.4–6.3)	Association between RF exposure from mobile phones and uveal melanoma.

[a] OR: Odds Ratio; CC: Case Control; CE: Case ecological; HBPBCC: Hospital-based population-based case control.

Table 16.2 Maximum Permissible Exposures to RFR

		Whole body SAR (W/kg)		Local SAR in head (W/kg)		Local SAR in limbs (W/kg)	
Standard	Frequency range	Public	Occupational	Public	Occupational	Public	Occupational
ARPANSA [48]	100 kHz–6 GHz	0.08 (6)[a]	0.4 (6)	2 [10][b] (6)	10 [10] (6)	4 [10] (6)	20 [10] (6)
Safety Code 6 [47]	100 kHz–10 GHz	0.08 (6)	0.4 (6)	1.6 [1] (6)	8 [1] (6)	4 [10] (6)	20 [10] (6)
ICNIRP [45]	100 kHz–6 GHz	0.08 (6)	0.4 (6)	2 [10] (6)	10 [10] (6)	4 [10] (6)	20 [10] (6)
FCC [89]	100 kHz–6 GHz	0.08 (30)	0.4 (6)	1.6 [1]	8 [1] (6)	4 [10]+[c]	20 [10] (6)+
NRPB [44]	100 kHz–6 GHz	0.4 (15)		10 [10] (6)		20 [100] (6)	
ANSI/IEEE[43]	100 kHz–6 GHz	0.08 (30)	0.4 (6)	1.6 [1] (30)	8 [1] (6)	4 [10] (30)+	20[10] (6)+

[a] () Averaging time in minutes.

[b] [] Averaging mass in grams.

[c] + in hands, wrists, feet and ankles.

peak whole-body SAR occurs in a frequency range of 60–80 MHz, while the resonant frequency for a laboratory rat is about 600 MHz.

Both SARs are averaged over a specific period of time and tissue masses of 1 or 10 g (defined as a tissue volume in the shape of a cube). Averaging the absorption over a larger amount of body tissue gives a less reliable result. The 1-g SAR is a more precise representation of localized RF energy absorption and a better measure of SAR distribution. Local SAR is generally based on estimates from the whole-body average SAR. It incorporates substantial safety factors (for example, 20).

There are two local SAR safety limits applicable to mobile phones: 1.6 W/kg averaged over 1 g ($SAR_{1\,g}$) in North America, and 2 W/kg averaged over 10 g ($SAR_{10\,g}$) developed by the ICNIRP of the European Union and accepted for use in Australia, Japan, and other parts of the world. Whether 1.6 W/kg or 2 W/kg is a right limit for RF exposure remains controversial.

Exposure to RFR from mobile phones is in the region close to antenna, the near field. However, exposure from other sources is in the far field, which is often quantified in terms of power density, expressed in units of watts per square meter (W/m^2). At the lower frequencies, about 0.1 to 10 MHz, the energy absorbed is less important than current density and total current, which can affect the nervous system. There is an overlap region at the upper part of this range where either current density or energy absorption rate is the limiting quantity. The standards at the lower frequencies are concerned with preventing adverse effects on the central nervous system (CNS) and electric shock. Exposure limits at these lower frequencies also involve numerous technical issues as well, but are not the focus of this review.

16.4. DOSIMETRY

Dosimetry in this manner considers the measurement or determination by calculation of the internal fields, induced current density, specific absorption (SA), or SAR distributions in objects like models (phantoms), animals, humans, or even parts of human body exposed to RFR.

Internal dosimetry can be divided into two categories [90]: macroscopic and microscopic dosimetry. In macroscopic dosimetry, the EM fields are determined as an average over some volume of space, such as in mathematical cells that are small in size. For example, the electric field in a given mathematical cell of 1 mm is assumed to have the same value everywhere within 1 mm^3 volume of the cell. The same is applied for magnetic fields. While in microscopic dosimetry, the fields are determined at a microscopic (cellular) level. Or the other way, the mathematical cells over which the EM fields are determined are microscopic in size. Microscopic dosimetry is useful for studies at the cellular level, which may throw light on EM interaction mechanisms.

16.4.1. Theoretical Dosimetry

The internal field in any biological material irradiated by RFR is calculated by solving Maxwell's equations. Practically, this is a difficult task and may be done for only a few special cases. Because of the mathematical difficulties encountered in the process of calculation a combination of techniques is used to find SAR in any biological object. Each technique gives information over a limited range of parameters in such a way that suits the chosen model.

In general, computational methods for analyzing EM problems fall into three categories: analytical techniques, numerical techniques, and expert systems. Analytical techniques apply assumptions to simplify the geometry of the problem in order to apply a closed-form solution. Numerical techniques attempt to solve basic field equations directly, due to boundary conditions posed by the geometry. However, expert systems estimate, but do not calculate, the values of fields for the parameters of interest based on a rules database.

Finite difference time-domain (FDTD) method is one of the most popular modeling techniques currently being used for EM interactions and SAR analysis. Starting from the evaluated SAR distribution, the thermal responses as a function of time, until the steady-state reaches may also be calculated through the application of finite difference method (FDM) or other numerical techniques.

16.4.2. Dosimetry of Induced Electric Fields

Induced electric field and current density values in biological systems are used as dosimetric measures in quantifying interactions with EMFs. It is known that an electric field that is initially uniform becomes distorted in the immediate vicinity of any biological body (for example, human). Whether the human is electrically grounded or is standing on an insulating platform also will considerably affect the field distribution.

Professor Maria A. Stuchly and her team at the University of Victoria focus on development of efficient numerical EM modeling techniques and their applications to solving complex problems at low frequencies. A practical example of the electric fields and current densities induced in a human body in close proximity to a 60-Hz transmission line was evaluated by the group [91]. The total-scattered field formulation was employed, along with a quasistatic formulation of the finite difference time-domain (FDTD) method. The demonstrated induced fields and current densities were significantly higher than originally predicted for the uniform electric field exposure on a ground plane.

Dawson et al. [92], within the same group, employed the scalar potential finite difference (SPFD) method to estimate tissue- and site-specific electric fields and current densities due to contact currents in anatomically realistic models of an adult and a child. Three pathways of contact current were modeled: hand to opposite hand and both feet, hand to hand only, and hand to both feet. For a contact current of 1 mA, as the occupational reference level set by the ICNIRP, the current density in brain does not exceed the basic restriction of 10 mA/m^2. The restriction is exceeded slightly in the spine by a factor of more than 2 in the heart. For a contact current of 0.5 mA, as the general public reference level, the basic restriction of 2 mA/m^2 is exceeded several folds in the spine and heart. Several microamperes of contact current produced tens of mV/m within the child's lower arm bone marrow. The differences in induced electric field and current density values between child body and those of adult body are due to larger size of adult relative to the child. The above findings are supported by Akimasa et al. [93] of the same group.

16.4.3. Instrumentation

Since measuring actual SAR in the human body is difficult, SAR is estimated by measurements using phantom models. Phantoms are tissue-equivalent synthetic materials simulating biological bodies. They may be simple or complex depending on the tissue composition as well as the shape.

SAR distribution in the phantom is derived from the measurement of electric field strength inside the body with implantable isotropic electric field probes (small antennas). An electric field probe often consists of electrically short dipole with a diode sensor across its terminals and highly resistive lines to carry the detected signal for measurement. Sensors of the probe are designed to function as true square-law detectors where the output voltage is proportional to the square of the electric field.

A multiple-axis probe positioning system and additional instrumentation regarding data processing and calibration are needed while measuring SAR. Probe placement is conducted either manually or by a robot. A probe supported by nonmetallic robotic arm moves from one point to another in a homogeneous liquid simulating tissue. The liquid is contained in a manikin (a RF transparent shell for the phantom) simulating a human head or another part of the human body. The head model, for example, is usually placed on its side (left or right ear) that allows a handset to be placed underneath the head to facilitate field measurement. A SAR measurement system is illustrated in Fig. 16.2 [1].

A few major factors influence the results of SAR measurements including probe calibration in phantom, tissue properties, and data acquisition system.

16.4.4. In-Head Dosimetry of Mobile Phones

Dosimetry of mobile phones targets SAR generated in the human head due to RFR. The energy absorbed in the head is mainly due to electric fields induced by the magnetic fields generated by currents flowing through the feed point, along the antenna and the body of the phone. The RF energy is scattered and attenuated as it propagates through the tissues of the head, and maximum energy absorption is expected in the more absorptive high-water-content tissues near the surface of the head.

In-head dosimetry can be achieved by evaluating mobile devices with a dummy head model called *phantom*. A phantom is a device that simulates the size, contours, and electrical characteristics of human tissue at normal body temperature. It is composed of a mannequin (solid shell) cut in half and filled with tissue-equivalent synthetic material solution, which has electrical properties of tissues. The phantom is typically set up in relation to other SAR measurement equipment. Measured pieces of equipment for this setup include a robot arm and miniature isotropic electric field probe. A phone is positioned against the mannequin operating at full power while the computer-controlled probe inserted into the tissue maps the electric fields inside. Computer algorithms determine

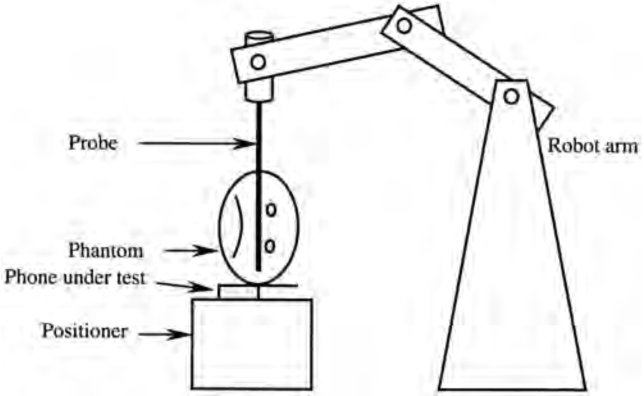


Figure 16.2 A schematic of a SAR measurement system.

Table 16.3 Summary of SAR Levels and Temperature Rise in Human Head

Investigator	Description of source	SAR (W/kg)	Temperature rise

Dimbylow, 1994 [94]	900 MHz: λ/4; 600 mW;
	1.8 GHz; λ/4; 125 mW;
	Calculated.
		For 900 MHz
		$SAR_{1\,g} = 2.17$; $SAR_{10\,g} = 1.82$
		For 1.8 GHz
		$SAR_{1\,g} = 0.7$; $SAR_{10\,g} = 0.48$	

Balzano et al., 1995 [95]	Motorola: 800–900 MHz; 600 mW and 2 W; Measured.	For analog (600 mW) Classic antenna: $SAR_{1\,g} = 0.2$–0.4; Flip antenna: $SAR_{1\,g} = 0.9$–1.6; Extended antenna: $SAR_{1\,g} = 0.6$–0.8. For GSM (2 W) Classic antenna: $SAR_{1\,g} = 0.09$–0.2; Flip antenna: $SAR_{1\,g} = 0.2$–0.3; Extended antenna: $SAR_{1\,g} = 0.1$–0.2.	
Anderson and Joyner, 1995 [96]	AMPS phones; 600 mW; 800/900 MHz.	SAR in the eye: 0.007–0.21; Metal-framed spectacles enhanced SARs in the eye by 9–29%; SAR in brain: 0.12–0.83.	Eye: 0.022°C due to SAR of 0.21 W/kg. Brain: 0.034°C due to SAR of 0.83 W/kg.
Okoniewski and Stuchly, 1996 [97]	Handset; 1 W; 915 MHz; λ/4; Calculated.	$SAR_{1\,g} = 1.9$; $SAR_{10\,g} = 1.4$	
Lazzi and Gandhi, 1998 [98]	Handset; Helical antenna 600 mW; 835 MHz. 125 mW; 1900 MHz; Calculated and measured.	$SAR_{1\,g} = 3.90$ (calculated); $SAR_{1\,g} = 4.02$ (measured). $SAR_{1\,g} = 0.15$ (calculated); $SAR_{1\,g} = 0.13$ (measured).	

Gandhi et al., 1999 [99]	AMPS phones; 600 mW; 800/900 MHz; Calculated and measured.	$SAR_{1\,g} > 1.6$ unless antennas are carefully designed and placed further away from the head.	
Van Leeuwen et al., 1999 [100]	Mobile phones; 250 mW; Calculated.	$SAR_{10\,g} = 1.6$	0.11°C
Wang and Fujiwara, 2000 [101]	Portable phone: 900 MHz; 600 mW; Helical antenna; Calculated.	$SAR_{1\,g} = 2.10$; $SAR_{10\,g} = 1.21$	
Bernardi et al., 2000 [102]	AMPS phones; 600 mW; 900 MHz; Calculated.	$SAR_{1\,g} = 2.2$–3.7	Ear: 0.22–0.43°C. Brain: 0.08°C to 0.19°C.
Van de Kamer and Lagendijk, 2002 [103]	Dipole antenna; 250 mW; 900 MHz; Calculated.	Cubic $SAR_{1\,g} = 1.72$; Arbitrary $SAR_{1\,g} = 2.55$; Cubic $SAR_{10\,g} = 0.98$; Arbitrary $SAR_{10\,g} = 1.73$.	

the maximum electric field and then calculate a 1-g or 10-g average over a body to give a SAR value.

The local peak SARs differ depending on many factors such as the antenna type, antenna radiation efficiency, antenna inclination with the head, distance of antenna from head, effect of the hand holding the handset, and the structural accuracy and resolution of the head model. Therefore, values of SARs are a function of various conditions set by each investigator. In other words, SAR is a result of a complex physical phenomenon of reactive coupling of the whole radiating structure with the human tissue. A significant contributor to the uncertainty in estimating SAR is the absence of a standard tissue averaging technique of the local SAR values over 1 or 10 g.

In recent years, many dosimetrical studies have been performed for calculating or measuring power absorbed in phantoms simulating human heads exposed to RFR (Table 16.3 [94–103]). It is evident that many SAR values exceeded the safety limits. However, the temperature rise is far too small to have any lasting effects. Temperature measurements are significant only in case of high SARs. Increases in temperature (0.03–0.19°C) are much lower than the threshold temperature for neuron damage (4.5°C for more than 30 min), cataract induction (3–5°C), and physiological effects (1–2°C). Therefore, the temperature increases caused by mobile phone exposure have no effect on the temperature-controlling functions of the human brain. In fact, the thermostabilizing effect of brain perfusion often prevents temperature increase.

16.5. CONCLUSION AND RESEARCH NEEDS

In evaluating the significant amount of information and wide range of cases studied, the conclusion seems to be that the current studies indicate no evident pattern of increased health risk associated with EM fields. Many of the early studies are methodologically weak and the results not reliable.

In conclusion, effects of EM fields are only a threat if the dosage of exposure is very high. In the case of most EM sources, especially those from mobile phones, the dose is not very high but still detectable. The detection of biological responses to low-level EM exposure requires the design of sophisticated sensitive research procedure. The sensitivity creates a greater possibility of producing contradictory results. Such research depends critically on the skill and experience of the researcher and it is necessary that results be compared with prudent investigation in properly structured and independent research laboratories. Further research is required to narrow a gap of knowledge.

REFERENCES

1. Habash, R.W.Y. *Electromagnetic Fields and Radiation: Human Bioeffects and Safety*; Marcel Dekker: New York, NY, 2001.
2. Moulder, J.E. Biological studies of power-frequency fields and carcinogenesis. IEEE Engg. Med. Biol. **1996**, *15*, 31–40.
3. Foster, K.R. Electromagnetic field effects and mechanisms. IEEE Engg. Med. Biol. **1996**, *15*, 50–56.
4. Adair, P.K. Constraints on biological effects of weak extremely low-frequency electromagnetic fields. Phy. Rev. Lett. **1991**, *A43*, 1039–1048.

5. Eichwald, C.; Walleczek, J. Magnetic field perturbations as a tool for controlling enzyme-regulated and oscillatory biochemical reactions. Biophy. Chem. **1998**, *74*, 209–224.
6. Magnussen, T. *Electromagnetic Fields*; EMX Corporation: San Jose, CA, 1999.
7. Adey, W.R. Cell membranes: The electromagnetic environment and cancer promotion; Neurochem. Res. **1988**, *13*, 671–677.
8. Litovitz, T.A.; Krause, D.; Mullins, J.M. Effect of coherence time of the applied magnetic field on ornithine decarboxylase activity. Biochem. Biophys. Res. Comm. **1991**, *178*, 862–865.
9. Lai, H.; Singh, N.P. Acute exposure to a 60-Hz magnetic field increases DNA strand breaks in rat brain cells. *Bioelectromagnetics* **1997**, *18*, 156–65.
10. Wu, R.W.; Yang, H.; Chiang, H.; Shao, B.J.; Bao, J.L. The effects of low-frequency magnetic fields on DNA unscheduled synthesis induced by methylnitro-nitrosoguanidine in vitro. Electro Magnetobiol. **1998**, *17*, 57–65.
11. Tofani, S.; Barone, D.; Cintorino, M.; de Santi, M.M.; Ferrara, A.; Orlassino, R.; Ossola, P.; Peroglio, F.; Rolfo, K.; Ronchetto, F. Static and ELF magnetic fields induce tumor growth inhibition and apoptosis. *Bioelectromagnetics* **2001**, *22*, 419–428.
12. Cohen, M.M.; Kunska, A.; Astemborski, J.A.; McCulloch, D.; Paskewitz, D.A. The effect of low-level 60-Hz electromagnetic fields on human lymphoblastoid cells. II. Sister-chromatid exchanges in peripheral blood lymphocytes and lymphoblastod cell lines. Mutation Res. **1985**, *172*, 177–184.
13. Rosenthal, M.; Obe, G. Effects of 50-Hz Electromagnetic fields on proliferation and on chromosomal alterations in human peripheral lymphocytes untreated or pretreated with chemical mutagens. Mutation Res. **1989**, *210*, 329–335.
14. Scarfi, M.R.; Lioi, M.B.; Zeni, O.; Franceschetti, G.; Franceschi, C.; Bersani, F. Lack of chromosomal aberration and micronucleus induction in human lymphocytes exposed to pulsed magnetic fields. Mutation Res. **1994**, *306*, 129–133.
15. Paile, W.; Jokela, K.; Koivistoinen, A.; Salomaa, S. Effects of 50-Hz sinusoidal magnetic fields and spark discharges on human lymphocytes in vitro. Bioelectrochem. Bioenerg. **1995**, *36*, 15–22.
16. Sasser, L.B.; Morris, J.E.; Miller, D.L.; Rafferty, C.N.; Ebi, K.L.; Anderson, L.E. Lack of a co-promoting effect of a 60-Hz magnetic field on skin tumorigenesis in SENCAR mice. *Carcinogenesis* **1998**, *19*, 1617–1621.
17. Babbitt, J.T.; Kharazi, A.I.; Taylor, J.M.G.; Rafferty, C.N.; Kovatch, R.; Bonds, C.B.; Mirell, S.G.; Frumkin, E.; Dietrich, F.; Zhuang, D.; Hahn, T.J.M. Leukemia/lymphoma in mice exposed to 60-Hz magnetic fields. *Results of the Chronic Exposure Study* TR-110338, EPRI, Los Angeles, 1998.
18. Boorman, G.A.; McCormick, D.L.; Findlay, J.C.; Hailey, J.R.; Gauger, J.R.; Johnson, T.R.; Kovatch, R.M.; Sills, R.C.; Haseman, J.K. Chronic toxicity/oncogenicity evaluation of 60-Hz (power frequency) magnetic fields in F344/N Rats. Toxicologic Pathol. **1999**, *27*, 267–278.
19. Stuchly, M.A.; McLean, J.R.N.; Burnett, R.; Goddard, M.; Lecuyer, D.W.; Mitchel, R.E.J. Modification of tumor promotion in the mouse skin by exposure to an alternating magnetic field. Cancer Lett. **1992**, *65*, 1–7.
20. Liburdy, R.P.; Sloma, T.R.; Sokolic, R.; Yaswen, P. ELF magnetic fields, breast cancer and melatonin: 60-Hz fields block melatonin's oncostatic action on ER + breast cancer cell proliferation. J. Pineal Res. **1993**, *14*, 89–97.
21. Selmaoui, B.; Touitou, Y. Sinusoidal 50-Hz magnetic fields depress rat pineal NAT activity and serum melatonin role of duration and intensity of exposure. Life Sciences **1995**, *57*, 1351–1358.
22. Harland, J.D.; Liburdy, R.P. ELF inhibition of melatonin and tamoxifen action on MCF-7 cell proliferation: field parameters. *BEMS Meeting*, Victoria, British Columbia, Canada, 1996.
23. Rogers, W.R.; Reiter, R.J.; Barlow-Walden, L.; Smith, H.D.; Orr, J.L. Regularly scheduled, daytime, slow-onset 60-Hz electric and magnetic field exposure does not depress serum melatonin concentration in nonhuman primates. Bioelectromagnetics **1995**, Suppl. 3, 111–118.
24. Rogers, W.R.; Reiter, R.J.; Smith, H.D.; Barlow-Walden, L. Rapid-onset/offset, variably scheduled 60-Hz electric and magnetic field exposure reduces nocturnal serum melatonin concentration in nonhuman primates. Bioelectromagnetics **1995**, Suppl. 3, 119–122.

25. Graham, C.; Cook, M.R.; Sastre, A.; Riffle, D.W.; Gerkovich, M. Multi-night exposure to 60-Hz magnetic fields: Effects on melatonin and its enzymatic metabolite. J. Pineal Res. **2000**, *28*, 1–8.
26. Sazonova, T. A Physiological Assessment of the Work Conditions in 400 kV and 500 kV Open Switch Yards. In *Scientific Publications of the Institute of Labor Protection of the All-Union Central Council of Trade Unions* 46, Profizdat, USSR, 1967. (Available from IEEE, Piscataway, NJ, Special Issue Number 10.)
27. Zenon, S. Behavioural effects of EMFs mechanisms and consequences of power frequency electromagnetic field exposures, *Electromagnetics Meeting*, Bristol, UK, 24–25 September 1998.
28. Stuchly, M.A. Human exposure to static and time-varying magnetic fields. Health Phy. **1986**, *51*, 215–225.
29. Korpinen, L.; Partanen, J.; Uusitalo, A. Influence of 50-Hz electric and magnetic fields on the human heart. Bioelectromagnetics **1993**, *14*, 329–340.
30. Wertheimer, N.; Leeper, E. Electrical wiring configurations and childhood cancer. Am. J. Epidemiol. **1979**, *109*, 273–284.
31. Savitz, D.A.; Wachtel, H.; Barnes, F.A.; John, E.M.; Tvrdik, J.G. Case-control study of childhood cancer and exposure to 60-Hz magnetic fields, Am. J. Epidemiol. **1988**, *128*, 21–38.
32. London, S.J.; Thomas, D.C.; Bowman, J.D.; Sobel, E.; Chen, T.S.; Peters, J.M. Exposure to residential electric and magnetic fields and risk of childhood leukemia. Am. J. Epidemiol. **1991**, *134*, 923–937.
33. Feychting, M.; Ahlbom, A. Magnetic fields and cancer in children residing near swedish high-voltage power lines. Am. J. Epidemiol. **1993**, *138*, 467–481.
34. Linet, M.S.; Hatch, E.E.; Kleinerman, R.A.; Robison, L.L.; Kaune, W.T.; Friedman, D.R.; Severson, R.K.; Haines, C.M.; Hartsock, C.T.; Niwa, S.; Wacholder, S.; Tarone, R.E. Residential exposure to magnetic fields and acute lymphoblastic leukemia in children. New England J. Med. **1997**, *337*, 1–7.
35. McBride, M.L.; Gallagher, R.P.; Theriault, G.; Armstrong, B.G.; Tamaro, S.; Spinelli, J.J.; Deadman, J.E.; Finchman, S.; Robson, D.; Choi, W. Power-frequency electric and magnetic fields and risk of childhood of leukemia in Canada. *Am. J. Epidemiol.* **1999**, *149*, 831–842.
36. Skinner, J.; Mee, T.J.; Blackwell, R.P.; Maslanyj, M.P.; Simpson, J.; Allen, S.G. Exposure to power frequency electric fields and the risk of childhood cancer in the UK. Br. J. Cancer **2002**, *87*, 1257–1266.
37. Sahl, J.D.; Kelsh, M.A.; Greenland, S. Cohort and nested case-control studies of hematopoietic cancers and brain cancer among electric utility workers. *Epidemiology* **1993**, *4*, 104–114.
38. Theriault, G.; Goldberg, M.; Miller, A.B.; Armstrong, B.; Guenel, P.; Deadman, J.; Imbernon, E.; To, T.; Chevalier, A.; Cyr, D.; Wall, C. Cancer risks associated with occupational exposure to magnetic fields among electric utility workers in Ontario and Quebec, Canada, and France: 1970–1989. Am. J. Epidemiol. **1994**, *139*, 550–572.
39. London, S.J.; Bowman, J.D.; Sobel, E.; Thomas, D.C.; Garabrant, D.H.; Pearce, N.; Bernstein, L.; Peters, J.M. Exposure to magnetic fields among electrical workers in relation to leukemia risk in Los Angeles County. Am. J. Indust. Med. **1994**, *26*, 47–60
40. Johansen, C.; Olsen, J. Risk of Cancer among danish utility workers—A nationwide cohort study, Am. J. Epidemiology **1998**, *147*, 548–555.
41. Villeneuve, P.J.; Agnew, D.A.; Miller, A.B.; Corey, P.N.; Purdham, J.T. Leukemia in electric utility workers: The evaluation of alternative indices of exposure to 60-Hz electric and magnetic fields. Am. J. Indust. Med. **2000**, *37*, 607–617.
42. Villeneuve, P.J.; Agnew, D.A.; Johonson, K.C.; Mao, Y. Brain cancer and occupational exposure to magnetic fields among men: Results from a canadian population-based case-control study. Int. J. Epidemiol. *31*, 210–217.
43. IEEE C95.1–1991, Safety levels with respect to human exposure to radio-frequency electromagnetic fields, 3 kHz to 300 GHz, IEEE, Piscataway, NJ, 1992.

44. NRPB, Board Statement on Restrictions on Human Exposure to Static and Time-Varying Electromagnetic Fields. Documents of the PRPB, Vol. 4, No. 5, National Radiological Protection Board, Chilton, Didcot, Oxon, UK, 1993.
45. ICNIRP, Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz). Health Phy. **1998**, *74*, 494–522.
46. TCO'99. Certification, Display (CRT), TCO Report No. 1, Stockholm, Sweden, 1999.
47. Safety Code 6, Limits of Human Exposure to Radiofrequency Electromagnetic Fields in the Frequency Range from 3 kHz to 300 GHz, Environmental Health Directorate, Health Protection Branch, Health Canada, Canada, 1999.
48. ARPANSA, Maximum exposure levels to Radio-frequency Fields, 3 kHz–300 GHz, Radiation Protection Series No. 3, Australian Radiation Protection and Nuclear Safety Agency, Australia, 2002.
49. Tice, R.R.; Hook, G.G.; Donner, M.; McRee, D.; Guy, A.W. Genotoxicity of radio-frequency signals. I. Investigation of DNA damage and micronuclei induction in cultured human blood cells. Bioelectromagnetics **2002**, *23*, 113–126.
50. D'Ambrosio, G.; Massa, R.; Rosaria, M.; Zeni, S.O. Cytogenetic damage in human lymphocytes following GMSK phase-modulated microwave exposure. Bioelectromagnetics **2002**, *23*, 7–13.
51. Mashevich, M.; Folkman, D.; Kesar, A.; Barbul, A.; Korenstein, R.; Jerby, E.; Avivi, E. Exposure of human peripheral blood lymphocytes to electromagnetic fields associated with cellular phones leads to chromosomal instability, Bioelectromagnetics **2003**, *24*, 82–90.
52. Czerska, E.M.; Elson, E.C.; Davis, C.C.; Swicord, M.L.; Czerski, P. Effects of continuos and pulsed 2450-MHz radiation on spontaneous lymphoblastoid tranformation of human lymphocytes in vitro. Bioelectromagnetics **1992**, *13*, 247–259.
53. Cleary, S.F.; Du, Z.; Cao, G.; Liu, L.M.; McCrady, C. Effect of radio-frequency radiation on cytolytic T lymphocytes. Fed. Am. Soc. Experimental Biol. J. **1996**, *10*, 913–919.
54. Kwee, S.; Raskmark, P. Changes in cell proliferation due to environmental non-ionizing radiation: 2. Microwave radiation. Bioelectrochem. Bioenerget. **1998**, *44*, 251–255.
55. Velizarov, S.; Raskmark, P.; Kwee, S. The effects of radio-frequency fields on cell proliferation are nonthermal. Bioelectrochem. Bioenerget. **1999**, *48*, 177–180.
56. Adey, W.R., Bawin, S.M.; Lawrence, A.F. Effects of weak amplitude modulated microwave fields on calcium efflux from awake cat cerebral cortex. *Bioelectromagnetics* **1982**, *3*, 295–307.
57. Tsurita, G.; Nagawa, H.; Ueno, S.; Watanabe, S.; Taki, M. Biological and morphological effects on the brain after exposure of rats to a 1439-MHz TDMA field. Bioelectromagnetics **2000**, *21*, 364–371.
58. Kamimura, Y.; Saito, K.–I.; Saiga, T.; Amenyima, Y. Effect of 2.45 GHz microwave irradiation on monkey eyes. IEICE Trans. Comm. **1994**, *E77-B*, 762–765.
59. Lu S.–T.; Mathur, S.P.; Stuck, B.; Zwick, H.; D'Andrea, H.; Zeriax, J.M.; Merritt, J.H.; Lutty, G.; McLeod, D.S.; Johnson, M. Effects of high-peak-power microwaves on the retina of the rhesus monkey. Bioelectromagnetics **2000**, *21*, 439–454.
60. Lai, H.; Horita, A.; Guy, A.W. Microwave irradiation affects radial-arm maze performance in the rat. Bioelectromagnetics **1994**, *15*, 95–104.
61. Bornhausen, M.; Scheingraber, H. Prenatal exposure to 900-MHz, cell-phone electromagnetic fields had no effect on operant-behavior performances of adult rats. Bioelectromagnetics **2000**, *21*, 566–574.
62. Dubreuil, D.; Jay, T.; Edeline, J.M. Does head-only exposure to GSM-900 electromagnetic fields affect the performance of rats in spatial learning tasks? Behavioural Brain Res. **2002**, *129*, 203–210.
63. Lin, J.C.; Lin, M.F.; Microwave hyperthermia-induced blood–brain barrier alterations, Radiat. Res. **1982**, *89*, 77–87.
64. Neubauer, C.; Phelan, A.M.; Kues, H.; Lange, D.G. Microwave irradiation of rats at 2.45 GHz activates pinocytotic-like uptake of tracer by capillary endothelial cells of cerebral cortex. Bioelectromagnetics **1990**, *11*, 261–268.

65. Persson, B.R.R.; Salford, R.L.G.; Brun, A. Blood–brain barrier permeability in rats exposed to electromagnetic fields used in wireless communication. Wireless Network **1997**, *3*, 455–461.

66. Fritze, K.; Wiessner, C.; Kuster, N.; Sommer, C.; Gass, P.; Hermann, D.P.; Kiessling, M.; Hossmann, K.-A. Effect of GSM microwave exposure on the genomic response of the rat brain. Neuroscience **1997**, *81*, 627–639.

67. Finnie, J.W.; Blumbergs, P.C.; Manavis, J.; Utteridge, T.D.; Gebski, V.; Davies, R.A.; Vernon-Roberts, B.; Kuchel, T.R. Effect of long-term mobile communication microwave exposure on vascular permeability in mouse brain. Pathology **2002**, *34*, 344–347.

68. Hirata, A.; Matsuyama, S.; Shiozawa, T. Temperature rises in the human eye exposed to em waves in the frequency range 0.6–6 GHz. IEEE Trans. Electromagnet. Compatibility **2000**, *42*, 386–393.

69. Elder, J.A. Special senses: Cataractogenic effects. In *Biological Effects of Radio-frequency Radiation*; Elder, J.A., Cahill, D.F., Eds.; Washington, DC: Environmental Protection Agency, 1984; Environmental Protection Agency Report, EPA-600/8-83-026F: 5-64-5-68.

70. Elder, J.A. Occular Effects of Radio-Frequency Radiation. IEEE Subcommittee 28.4 White Paper, 2001.

71. Hossmann, K.-A.; Hermann, D.M. Effects of electromagnetic radiation of mobile phones on the central nervous system. Bioelectromagnetics **2003**, *24*, 49–62.

72. Jauchem, J.R.; Ryan, K.L.; Frei, M.R.; Dusch, S.J.; Lehnert, H.M.; Kovatch, R.M. Repeated exposure of C3H/HeJ mice to ultra-wideband electromagnetic pulses: Lack of effects on mammary tumors. Radiation Res. **2001**, *155*, 369–377.

73. Robinette, C.D.; Silverman, C. Causes of health following occupational exposure to microwave radiation (Radar) 1950–1974. In *Symposium on Biological Effects and Measurement of Radiofrequency/Microwaves*; Hazzard, D.G. Ed.; Dept. of Health, Education, and Welfare, HEW Publication No. (FDA) 77-8026, Washington, DC, 1977.

74. Szmigielski, S. Cancer morbidity in subjects occupationally exposed to high-frequency (Radio frequency and microwave) Electromagnetic radiation. Science Total Environment **1996**, *180*, 9–17.

75. Davis, R.L.; Mostofi, F.K. Cluster of testicular cancer in police officers exposed to handheld radar. Am. J. Indust. Med. **1993**, *24*, 231–233.

76. Finkelstein, M.M. Cancer incidence among ontario police officers. Am. J. Indust. Med. **1998**, *34*, 157–162.

77. Lagorio, S.; Rossi, S.; Vecchia, P.; De Santis, M.; Bastianini, L.; Fusilli, M.; Ferrucci, A.; Desideri, E.; Comba, P. Mortality of plastic-ware workers exposed to radio frequencies. Bioelectromagnetics **1997**, *18*, 418–421.

78. Tynes, T.; Hannevik, M.; Andersen, A.; Vistnes, A.I.; Haldorsen, T. Incidence of breast cancer in Norwegian female radio and telegraph operators. Cancer Causes Control **1996**, *7*, 197–204.

79. Hocking, B.; Gordon, I.; Grain, H.; Hatfield, G. Cancer incidence and mortality and proximity to TV towers. Med. J. Australia **1996**, *65*, 601–605.

80. McKenzie, D.R.; Yin, Y.; Morrell, S. Childhood incidence of acute lymphoblastic leukemia and exposure to broadcast radiation in sydney—a second look. Australia and New Zealand J. Public Health **1998**, *22*, 360–367.

81. Michelozzi, P.; Ancona, C.; Fusco, D.; Forastiere, F.; Perucci, C.A. Risk of leukemia and residence near a radio transmitter in Italy. Epidemiology **1998**, *9* (Suppl), 354.

82. Hardell, L.; Nasman, A.; Pahlson, A.; Hallquist, A.; Mild, K.H. Use of cellular telephones and the risk for brain tumors: a case-control study. Int. J. Oncol. **1999**, *15*, 113–116.

83. Muscat, J.E.; Malkin, M.G.; Thompson, S.; Shore, R.E.; Stellman, S.D.; McRee, D.; Neugut, A.I.; Wynder, E.L. Handheld cellular telephone use and risk of brain cancer. J. AMA **2000**, *284*, 3001–3007.

84. Muscat, J.E.; Malikin, M.G.; Shore, R.E.; Thompson, S.; Neugut, A.I.; Stellman, S.D.; Bruce, J. Handheld cellular telephones and risk of acoustic neuroma. *Neurology* **2002**, *58*, 1304–1306.

85. Inskip, P.D.; Tarone, R.E.; Hatch, E.E.; Wilcosky, T.C.; Shapiro, W.R.; Selker, R.G. Cellular telephone use and brain tumors. New England J. Med. **2001**, *344*, 79–86.

86. Johansen C.; Boice, Jr. J.D.; McLaughlin, J.K.; Olsen, J.H. Cellular telephones and cancer—A nationwide cohort study in denmark. J. Nat. Cancer Inst. **2001**, *93*, 203–207.

87. Auvinen, A.; Hietanen, M.; Luukkonen, R.; Koskela, R.S. Brain tumors and salivary and cancers among cellular telephone users. Epidemiology **2002**, *13*, 356–359.

88. Stang A.; Anastassiou, G.; Wolfgang, A.; Bromen, K.; Bornfeld, N.; Jöckel, K.-H. The possible role of radio-frequency radiation in the development of uveal melanoma. Epidemiology **2001**, *12*, 7–12.

89. FCC, Guidelines for Evaluating the Environmental Effects of Radio Frequency Radiation, Federal Communications Commission, 96–326, Washington, DC, 1996.

90. Durney, C.H.; Christensen, D.A. *Basic Introduction to Bioelectromagnetics*; CRC Press: Boca Raton, FL, 1999.

91. Potter, M. E.; Okoniewski, M.; Stuchly, M.A. low-frequency finite difference time-domain (FDTD) for modeling of induced fields in humans close to line sources. J. Computational Phy. **2000**, *162*, 82–103.

92. Dawson, T.W.; Caputa, K.; Stuchly, M.A.; Kavet, R. Induced electric fields in the human body associated with 60-Hz contact currents. IEEE Trans. Biomed. Engg. **2001**, *48*, 1020–1026.

93. Akimasa, H.; Caputa, K.; Dawson, T.W.; Stuchly, M.A. Dosimetry in models of child and adult for low-frequency electric fields. IEEE Trans. Biomed. Engg. **2001**, *48*, 1007–1011.

94. Dimbylow, P.J.; Mann, S.M. SAR calculations in an anatomically realistic model of the head for mobile communications transceivers at 835 MHz and 1.8 GHz. Phys. Med. Biol. **1994**, *39*, 1537–1553.

95. Balzano, Q.; Garay, O.; Manning, T.J.; Electromagnetic energy exposure of simulated users of portable cellular telephones. IEEE Trans. Vehicular Technol. **1995**, *44*, 390–403.

96. Anderson, V.; Joyner, K.H. Specific absorption rate levels measured in a phantom head exposed to radio-frequency transmissions from analog hand-held mobile phones. Bioelectromagnetics **1995**, *16*, 60–69.

97. Okoniewski, M.; Stuchly. M.A. A study of the handset antenna and human body interaction. IEEE Trans. Microwave Theory Techni. **1996**, *44*, 1855–1864.

98. Lazzi, G.; Gandhi, O.P. On modeling and personal dosimetry of cellular telephone helical antennas with the FDTD code. IEEE Trans. Antennas Propagat. **1998**, *46*, 525–529.

99. Gandhi, Om, P.; Gianluca, L.; Tinniswood, A.; Yu, Q.-S. Comparison of numerical and experimental methods for determination of SAR and radiation patterns of handheld wireless telephones. Bioelectromagnetics **1999**, *20*, 93–101.

100. Van Leeuwen, G.M.; Lagendijk, J.J.; Van Leersum, B.J.; Zwamborn, A.P.; Hornsleth, S.N.; Kotte, A.N. Calculation of change in brain temperatures due to exposure to a mobile phone. Phy. Med. Biol. **1999**, *44*, 2367–2379.

101. Wang, J.; Fujiwara, O. FDTD analysis of dosimetry in human model for a helical antenna portable telephone. Electronics Communications in Japan **2000**, *E83-B*, 549–554.

102. Bernardi, P.; Cavagnaro, M.; Pisa, S.; Piuzzi, E. Specific absorption rate and temperature increases in the head of a cellular-phone user. IEEE Trans. Microwave Techniq **2000**, *48*, 1118–1125.

103. Van de Kamer, J.B.; Lagendijk, J.J.W. Computation of high-resolution SAR distributions in a head due to a radiating dipole antenna representing a hand held mobile phone. Phy. Med. Biol. **2002,** *47*, 1827–1835.

17

Biomedical Applications of Electromagnetic Engineering

James C. Lin
University of Illinois at Chicago
Chicago, Illinois, U.S.A.

17.1. INTRODUCTION

Electromagnetic energy in the frequency region below 300 GHz is nonionizing and has wavelengths in air longer than 1 mm. Energy with wavelengths longer than 10 m (frequencies lower than 30 MHz) has conduction and propagation properties that differ greatly from those of wavelengths that approximate the human body's physical dimensions. Since its interaction with biological media differs according to the specific spectral band, these properties can give rise to a wide range of applications. Moreover, physiological responses can often vary at different frequencies. For example, thermal therapies avoid frequencies lower than 10 kHz to prevent stimulation of excitable muscular and cardiac tissues. Thus, advances in the use of electromagnetic technology for biomedical research and practice would be enhanced by a thorough understanding of biophysical interactions.

Short-wave and microwave diathermy have been used to heat muscle masses to relieve stress and strain, to stimulate blood circulation, and to reduce inflammation for more than 50 y [1,2]. Although nuclear magnetic resonance imaging or MRI has been developed only during the past 20 y, it has become one of the most extensively used diagnostic radiological imaging procedures. This chapter reviews recent biomedical applications that involve electromagnetic technology. Specifically, it describes applications in neuromagnetic imaging and stimulation, physiological monitoring, elimination of hypothermia, thermal ablation therapy, and hyperthermia treatment of cancer. Some of the applications have already found their niche in clinical practice.

17.2. THERMAL ABLATION THERAPIES

Percutaneous catheter ablation of arrhythmogenic foci has become an important therapy for selected patients with drug refractory, symptomatic tachyarrhythmias [3]. The increased interest stems, in part, from the nonpharmacological approach and minimally invasive nature of the procedure [3–5]. Microwave catheter antennas and radio-frequency

(RF) electrodes are used to deliver electromagnetic (EM) fields and waves into the surrounding heart tissue. The frequencies used for the RF band are 500 to 750 kHz [6,7] and that for the microwave band are 915 and 2450 MHz [5,8]. Endocardiac conducting tissues responsible for causing arrhythmia or abnormal heart rhythm are destroyed by thermal energy applied through a catheter to the tissue.

17.2.1. EM Energy Propagation in Tissue

When RF energy is used, the applied voltage induces a current to flow between a small electrode inside or on the surface of the body to a large grounded, dispersive electrode on the surface (Fig. 17.1). In cauterizing tissue, a train of short RF pulses at high voltage is delivered through a pair of electrodes to create cutting and coagulation. For catheter ablation, RF energy is applied as a sinusoidal current through a small endocardial electrode to provide effective tissue heating [4].

A characteristic of RF frequency is that the associated wavelength is at least an order of magnitude longer than the dimensions of the human body. Its propagation behavior is therefore quasistatic and can be approximated using Laplace's formulation in electromagnetic field theory [7].

The absorption of EM energy in tissues is governed by the dielectric permittivity and conductivity. At the RF frequencies used for ablation, the conductive energy dissipation is considerably higher than dielectric energy dissipation. Accordingly, a reasonable approximation is obtained by neglecting dielectric permittivity and considering only the tissue conductivity such that Laplace's equation becomes

$$\nabla \cdot \sigma \nabla V = 0 \tag{17.1}$$

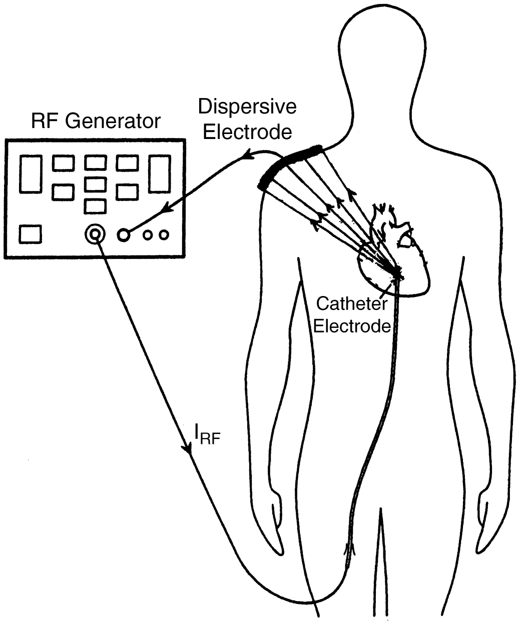


Figure 17.1 Schematic diagram of a RF ablation system. The applied RF voltage induces a current to flow between a small electrode inside or on the surface of the body to a large grounded, dispersive electrode on the surface.

where σ is the tissue electrical conductivity and V is the electrical potential. The density of current (**J**) flowing at any point in the tissue is given from the Ohm's law,

$$\mathbf{J} = -\sigma \nabla V \tag{17.2}$$

The current flow is impeded by tissue resistance (which is inversely proportional to conductivity), and RF energy is extracted or transferred to the tissue. The transferred or absorbed energy is converted to heat in accordance with Joule's law, which states

$$\mathbf{W} = \frac{\mathbf{J}^2}{\sigma} = \sigma(\nabla V)^2 \tag{17.3}$$

where **W** denotes the rate of energy absorption or the heating potential generated by RF energy as applied through the catheter electrode in a unit volume of tissue. The SI unit of **W** is watts per cubic meter (W/m^3).

A solution to Eq. (17.1) requires that V be specified at all points of the boundary throughout the region of interest. A pertinent set of boundary conditions is the voltage on the surface of the electrodes. In the case of unipolar ablation, RF energy is delivered from the electrode at the catheter tip inside the heart to a large, flat dispersive electrode on the skin surface. The voltage distribution between the active and dispersive electrodes in a homogeneous tissue is approximately given by

$$V(r) \sim \frac{V_\mathrm{o}}{r} \tag{17.4}$$

where V_o is the voltage on the surface of the spherical unipolar electrode and r the radial distance from the active electrode. The inverse proportion of voltage distribution to radial distance indicates a lower risk of cardiac stimulation or muscle contraction associated RF energy from cardiac ablation at 100 V or less, which is the usual voltage used in RF ablation. The current density and time rate of heat generation are given, respectively, by

$$\mathbf{J} \sim \sigma \frac{V_\mathrm{o}}{r^2} \tag{17.5}$$

$$\mathbf{W} \sim \sigma \frac{V_\mathrm{o}^2}{r^4} \tag{17.6}$$

Clearly, RF energy absorption decreases as the fourth power of distance from the active electrode. The rapid decrease suggests that applied RF energy diverges from the small electrode. Consequently, active tissue heating is localized to a very short distance from the electrode–tissue interface. For effective cardiac ablation, it is essential to maintain direct contact between the RF electrode and cardiac tissue. Slight pressure exerted on the myocardium by the catheter is useful. If the density of tissue (ρ) is known, **W** can also be expressed as a specific absorption rate (SAR) quantity in units of W/kg, such that

$$\mathrm{SAR} = \frac{\mathrm{W}}{\rho} \tag{17.7}$$

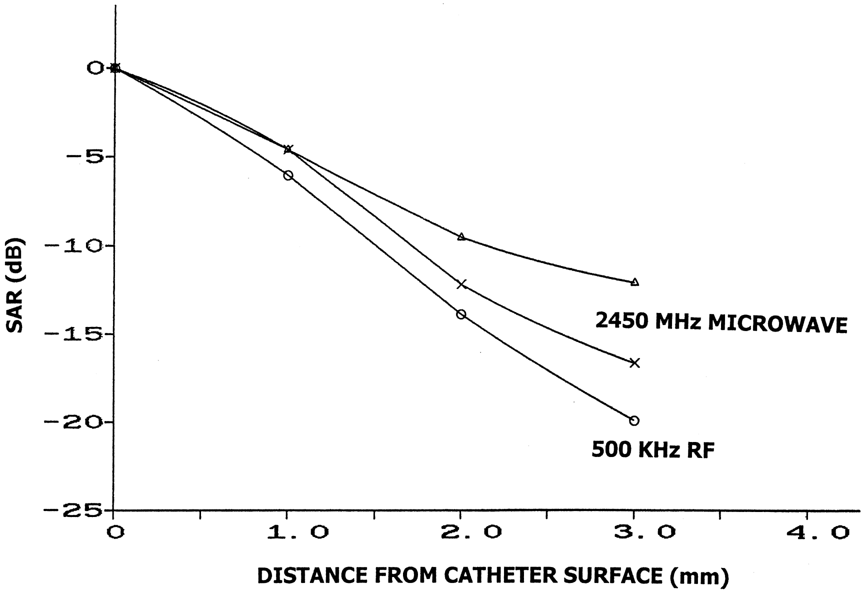


Figure 17.2 A comparison of SAR distributions produced by microwave catheter antennas and RF electrode as a function of radial distance from the catheter surface.

Note that SAR is a measure of the rate of RF energy deposition at points surrounding the catheter electrode. This distribution of SAR serves as the source of lesion formation in cardiac ablation following thermalization of absorbed RF energy. Figure 17.2 illustrates the SAR measured in a tissue-equivalent model for a catheter electrode operating at 500 kHz. It can be seen that the drop in SAR is about 7.5 dB/mm away from the electrode. At 3 mm the decrease is about 20 dB or 100 times [9]. The size of lesions produced would be the combined result of SAR, duration of RF application, and heat conduction in tissue.

RF heating would quickly become insignificant beyond a few millimeters from the active electrode. Heating would fall well below the thermal noise floor at the dispersive electrode on the body surface, provided that the dispersive electrode is large and in good contact everywhere. Increasing RF power delivery influences the magnitude of active heating (SAR) and its subsequent passive spread in lesion formation by RF ablation, but it has less influence on the region of active heat generation.

Desiccation and coagulation of tissue close to the electrode would decrease the tissue's electrical conductivity and raise the resistance to current flow. This, in turn, would further impede effective tissue heating and limit the size of RF-induced lesions. Lesions beyond the immediate vicinity of the electrode–tissue interface occur as a result of passive heat transfer from the shallow high temperature region. Indeed, studies have shown that RF-induced lesions increase rapidly in size during the initial period of power application. Subsequently, the rate of increases diminishes rapidly as the resistance at the electrode–tissue interface rises, and the current flow falls [10–13]. This inevitable phenomenon of thermal lesion production may be assessed through changes in the impedance of the catheter electrode. It is noteworthy that measures to maintain good electrode–tissue contact such as increasing the contact pressure can enhance RF coupling to the tissue [12].

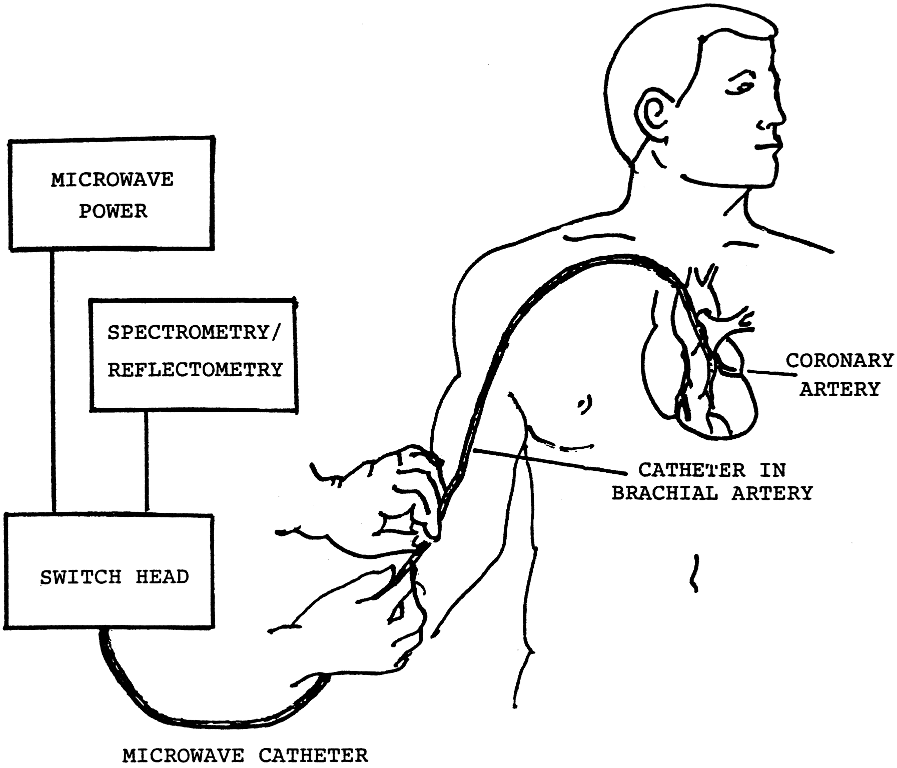


Figure 17.3 A microwave cardiac ablation system microwave energy is delivered through a radiating antenna mounted to the tip of a catheter.

Thermal microwave energy has been investigated for its potential in producing larger and deeper lesions. Unlike RF ablation, microwave energy is delivered through a radiating antenna mounted to the tip of a catheter (Fig. 17.3). A dispersive electrode at the body surface is not needed. Tissue heating is produced exclusively by absorption of radiated microwave energy in the biological dielectric [13]. Endocardiac microwave antennas should increase the volume of direct heating as compared to RF ablation since the lesion size is determined by the antenna radiation pattern, microwave power, and duration of power delivery. Comparison of phantom and in vivo results from RF and microwave ablation catheters showed that the volume of direct heating is indeed larger (Fig. 17.2) and that microwave energy is suitable for transcatheter ablation procedures [9,14–19]. Several microwave catheter antennas have been developed with efficient energy transfer into the myocardium [20–23].

The propagation and radiation of microwaves in biological tissue are governed by frequency, power, and the antenna radiation pattern, as well as by tissue composition and dielectric permittivity. The biological tissues of interest in microwave cardiac ablation are blood, muscle, and tissues with low water content, such as fat, bone, or desiccated tissue. Some typical values of microwave dielectric constant and conductivity at 37°C are given in Table 17.1 for 915 and 2450 MHz—two frequencies of most interest [13,24]. There is a modest change in dielectric constant and conductivity as a function of frequency for all three types of tissues. However, differences among the three types of tissues are quite large.

Table 17.1 Dielectric Constant and Conductivity of Biological Tissues at 37°C.

	Dielectric constant			Conductivity (S/m)		
Frequency (MHz)	Blood	Muscle	Fat[a]	Blood	Muscle	Fat[a]
915	60	51	5.6	1.4	1.6	0.10
2450	58	47	5.5	2.1	2.2	0.16

[a] Fat, bone, or desiccated tissue.

As microwave fields propagate in the tissue medium, energy is extracted from the field and absorbed by the medium. This absorption results in a progressive reduction of the microwave power intensity as it advances into the tissue. The portion of energy extracted from the propagating microwave field is converted into heat production. The reduction is quantified by the depth of penetration. At 2.45 GHz, the depths of plane-wave penetration for blood, muscle, and fat are 19, 17, and 79 mm, respectively. For microwave-catheter antennas that do not have plane wavefronts, the penetration depth is reduced according to the specific antenna design. Nevertheless, these values clearly suggest that microwave energy can deposit energy directly into tissue at a distance, through radiative interaction of microwaves with cardiac tissues. Furthermore, the differences in the dielectric permittivity yield a depth of penetration for tissues with low water content that is about four times deeper than for muscle (or higher water content tissue) at 2.45 GHz. This means that a microwave field can propagate more readily through (is absorbed less by) low-water-content tissues than in tissues of high water content. It also implies that microwaves can propagate through intervening desiccated tissue to deposit energy directly in more deeply lying tissue. At 2.45 GHz, the dielectric constant for muscle is 20% lower than that for blood, but is about 800% higher than that for fat. While conductivities for blood and muscle are approximately the same, they are about 300% higher than those for tissues with low water content. Indeed, these inherent features have been demonstrated in phantom, animal, and human subjects for microwave energy. Specifically, larger and deeper lesions have been produced by microwave radiation [9,14].

17.2.2. Temperature Elevation in Tissue

In thermal therapeutic applications, the final temperature may be affected by tissue blood flow and thermal conduction. Specifically, the dynamic temperature variation in tissue is a function of tissue composition, blood perfusion, thermal conductivity of tissue, and heat generation due to metabolic processes in addition to RF energy absorption. The Pennes approximation to biological heat transfer via diffused conduction and blood convection in tissue can provide useful insights into RF and microwave ablation under the condition of uniform tissue perfusion by blood [25]. Specifically, the Pennes bioheat transfer equation states that

$$\frac{d(\rho c T)}{dt} = \mathbf{W} + \eta \nabla^2 T - V_s(T - T_o) \tag{17.8}$$

where T is the temperature in tissue (°C), T_o is the temperature in blood and tissue (°C), ρ is the density of tissue (kg/m^3), c is the specific heat of tissue (J/kg °C), $\mathbf{W}$ is the heating potential generated by RF power deposition (W/m^3), η is the coefficient of heat conduction of tissue (W/(°C m^3)), and V_s is the product of flow rate and heat capacity of blood

(W/(°C m^3)). The numerical values of these parameters for muscle tissue are $\rho = 1000$ (kg/m^3), $c = 3500$ (J/kg °C), and $\eta = 0.508$ (W/°C m^3). However, heat transfer by blood convection is considerably faster in well perfused tissue, as is seen by a high $V_s = 7.780 \times 10^6$ (W/°C m^3). Note that this equation neglects metabolic heat production since the usual period of RF application is 30 to 60 s, a period insufficient for significant metabolic heat contribution to tissue ablation. Note that for applications of short duration, the time rate of heating and the spatial distribution of radiated microwave energy are function of power deposition (the rate of energy absorption) and the antenna radiation patterns, respectively.

17.2.3. Cardiac Ablation

Radiofrequency ablation has become the preferred treatment modality for a variety of cardiac arrhythmias [26,27]. When RF energy is used, the applied voltage induces a current to flow between a small electrode inside or on the surface of the body to a large grounded, dispersive electrode on the surface. An important precaution is that the dispersive electrode must be large and in uniform contact with the skin to safely guard against high current density occurring at the edges. Considerations of the biophysical aspects of RF cardiac ablation indicate that the dissipation or absorption of RF energy is the source of tissue temperature elevation and the cause for subsequent lesion formation in RF ablation [7]. The divergent nature of RF current flow from the catheter electrode and the rapid dissipation of RF energy by tissue resistance combine to limit the depth and size of lesions produced. Several electronic [10–12,28,29] and mechanical techniques [30–37] can be invoked to enhance catheter ablation, produce larger and deeper lesions, and to overcome some of the aforementioned limitations. However, all of these techniques are baffled fundamentally by the effective current density required for ablation through resistive heating. Treatment of certain subepicardial arrhythmogenic substrates remains a challenge for RF ablation.

Several microwave catheter antennas have been developed to deliver ablating energy to the target tissue substrate. A drawback of some catheter antennas is that a considerable amount of microwave energy is reflected by the antennas to the skin surface and is deposited at the point of antenna insertion into the blood vessel [38–41]. The problem was addressed by several catheter antennas with efficient energy transfer into the myocardium [20–23]. A particular design feature of these catheter antennas is that there is good dielectric and impedance matching and a minimal amount of power is reflected from the antenna or flowing up the transmission line. Thus, they minimize heating of the coaxial cable or at the insertion point of the catheter into the body. These catheter antennas also serve as bipolar electrodes for endocardiac electrogram recording. Studies in dogs both during cardiopulmonary bypass and closed-chest operations have shown microwave energy greater than 200 J (joules) delivered to the heart through a split-tip dipole catheter antenna can produce an irreversible block of the heart rhythms [22,42,45]. This energy was achieved either by increasing the delivered power from 20 to 40 W or by increasing the treatment duration from 7 to 11 s (210 to 330 J per application). It produced a myocardial temperature of 65°C. These results show that microwave catheter ablation is a safe and suitable procedure for treatment of cardiac arrhythmias, including arrhythmias due to conduction pathways located deep in the myocardium.

17.2.4. Angioplasty

The primary goal of angioplasties is in dilation of atherosclerotic arteries to achieve maximum function for a prolonged period of time [45–49]. Under sterile condition, a percutaneous microwave catheter antenna or RF catheter electrode is guided through the arterial system and positioned in the appropriate coronary artery. A moderate dose (10–20 W) of 2450-MHz microwave or 500-kHz RF energy is delivered to the atherosclerotic plaque to reduce arterial narrowing by microwave-induced heating, softening and spreading of the plaque. In the microwave case, the catheter antenna consists of either a simple junction or sleeved slot radiator terminating a flexible coaxial cable 1 to 2 mm in diameter. Alternatively, the microwave thermal angioplasty is used in combination with conventional balloon angioplasties. A microwave catheter antenna is used to apply energy to heat the plaque through a water-filled balloon in remodeling the atherosclerotic lesion. Peak spatial temperatures in dog and rabbit myocardium have reached 60 to 80°C in 10 to 60-s, respectively, for a net power of 10 to 40 W to the catheter antenna. These techniques have been used successfully in animals and are presently undergoing clinical trials.

17.2.5. Benign Prostate Hyperplasia

Benign prostatic hyperplasia or hypertrophy is a major cause of morbidity in the adult male. At present, open surgery and transurethral resections of the prostate are the gold standards for treatment of benign prostatic hypertrophy. They can provide immediate relief of obstructive symptoms that remain fairly to extremely durable [50]. A new, less invasive procedure uses thermal energy delivered by microwaves (Fig. 17.4) [51]. In particular, an early report of thermal microwave technique from 1985 employed a transurethral microwave applicator [52]. It showed coagulation of the prostate in mongrel dogs and some salutary effects in an initial six patients treated with this device. An ensuing study used 2450 MHz microwave energy to treat 35 patients, and it compared a transurethral resection alone to preliminary microwave coagulation followed by transurethral resection of the gland [53]. Significant reduction in blood loss by initial treatment with the microwave thermal therapy was observed.

Numerous reports have appeared since that time on various aspects of both transrectal and transurethral microwave therapy of the prostate using 915 and 2450 MHz energy [54–59]. Most of the research in human subjects to date has focused on methods of delivery. Initial attempts to deliver the energy transrectally have not been effective and injury to the rectal mucosa has occurred due to the difficulty of interface cooling of this organ. Recent investigations have focused on transurethral delivery of the energy with cooling systems within the catheter to ensure urethral preservation [56–61]. Sensors placed

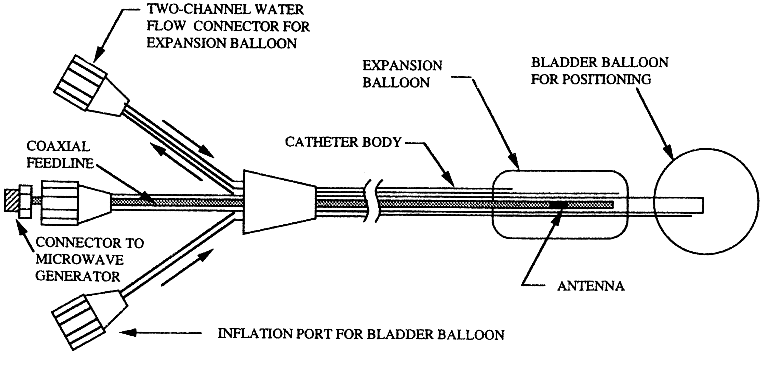

TWO-CHANNEL WATER FLOW CONNECTOR FOR EXPANSION BALLOON
EXPANSION BALLOON
BLADDER BALLOON FOR POSITIONING
COAXIAL FEEDLINE
CATHETER BODY
CONNECTOR TO MICROWAVE GENERATOR
ANTENNA
INFLATION PORT FOR BLADDER BALLOON

Figure 17.4 Schematic diagram of transurethral microwave balloon catheter used in hyperthermia treatment for benign prostatic hyperplasia (from Ref. 51).

in the microwave antenna maintain temperature on the urethral surface between 43 and 45°C. It is noted that while the number of treatment sessions and the temperature attained are extremely important predictors of response, sufficient hyperthermia volume is crucial for enhanced efficacy.

Virtually no data clearly demonstrating reduction in prostate volume in human subjects has been reported, although most investigators have shown improvement in measured urinary flow rates compared to preoperative studies. Randomized studies comparing microwave thermotherapy to transurethral resections conclude that microwave hyperthermia treatment had a definite therapeutic effect on symptomatic prostatic hypertrophy [62–69]. Thus, microwave thermal ablation of prostatic tissue and enlargement of the urethral with minimal clinical complications offers a therapeutic alternative to surgery in select patients with benign prostatic hyperplasia. Indeed, transurethral microwave thermotherapy is on its way to become a standard in minimally invasive treatment.

17.3. HYPERTHERMIA CANCER TREATMENT

Hyperthermia cancer therapy is a treatment procedure in which tumor temperatures are elevated to the range of 42–45°C. An important aspect of this development is the production of adequate temperature distribution in superficial and deep-seated tumors. A large number of antennas and applicators have been designed to produce therapeutic heating of a tumor of different volumes in a variety of anatomic sites [70–73]. For superficial tumors, single-contact applicators operating between 433 and 2450 MHz have been used. Because of the limited depth of energy penetration, these antennas have been applied to the heating of well-localized tumors extending to depths of up to a few centimeters. Techniques devised to provide noninvasive heating of deep-seated tumors include capacitive plates, helical coils and multiapplicator arrays. Capacitive plates operate at low frequencies so that the wavelengths are long compared to typical body dimensions. With appropriate design, this simple applicator can provide fairly uniform heating of tissue between the plates of two or three-electrode systems. The helical coil applicator gives rise to a power deposition pattern that varies slowly with radial distance and gives good penetration. The multiapplicator array concept is rapidly gaining utility in the clinic. Commercially available annular phased array systems operating at 60–90 MHz are designed to heat large anatomical regions such as the thorax, abdomen, and pelvic areas. The most promising region for this system appears to be the pelvis, since heating of the upper body is frequently limited by systemic hyperthermia and hemodynamic compensation, and by excessive heating of adjacent normal tissue structures. A major advantage of multiapplicator array systems is the ability to steer the hot spot electronically by varying the phase of each element, thereby allowing phased arrays operating at microwave frequencies to be used to effect selective heating of deep-seated tumors in a variety of anatomic sites.

Under certain conditions invasive method of power deposition such as thermoseed implants and interstitial antennas may be preferable for local hyperthermia of deep-seated tumors. Magnetic induction heating of arrays of thermoseed implants holds great promise for localized hyperthermia in deep-seated tumors. Quantitative study of power deposition at the implant and elsewhere inside the head has shown at least two orders of magnitude greater energy absorption by the implant than by the rest of the head. For tumors of large volume, interstitial techniques have been employed to generate the desired hyperthermic field (Fig. 17.2). RF electrodes operating in the frequency range of 0.5 to 10 MHz and microwave antennas operating between 300 and 2450 MHz have been employed for treatment of breast, head, and neck tumors. In most cases, it was necessary to implant an array of microwave antennas in order to produce the desired volume heating (Fig. 17.5). Interstitial techniques are attractive because, in combination with radiotherapy, interstitial hyperthermia renders a new modality of treatment for these malignancies with little additional risk to the patient (Fig. 17.6). Moreover, hyperthermia has been shown to be especially effective against hypoxic tumor cells that are low in pH and are resistant to ionizing radiation. Thus, the synergistic effect of this combined mode of treatment may portend reduced therapeutic dosages of drugs and/or ionizing radiation. This may not only lessen the extent of unacceptable radiation necrosis of normal tissue, but also enhance the mean survival rate.

While clinical and laboratory results have indicated a promising future for hyperthermia [74–80], its efficacy critically depends on the induction of a sufficient temperature rise throughout the tumor volume. Since each modality has its own liabilities there is no universally preferred modality. It is often necessary to use a wide range of external and implanted antennas and applicators to produce therapeutic heating of localized and regional tumors of different volumes at a variety of anatomical sites. The type of applicator must therefore be selected, ad hoc, on the basis of specific site and type of the tumor which greatly complicates logistics and compromises quality control of the treatment.

Moreover, monitoring and control of tumor temperature in real time during hyperthermia treatment is essential for effective therapy. While progress in temperature sensing in vivo has been dramatic, considerable advance is needed prior to widespread clinical application of hyperthermia for cancer. Among approaches that may impact this outcome include invasive multipoint sensing [81,82], and noninvasive magnetic resonance imaging temperature mapping [83–86].

17.4. HYPOTHERMIA ELIMINATION

Severe hypothermia occurring at internal body temperature below 32°C is a clinical condition with a high rate of mortality. Current rewarming methods rely on either slow

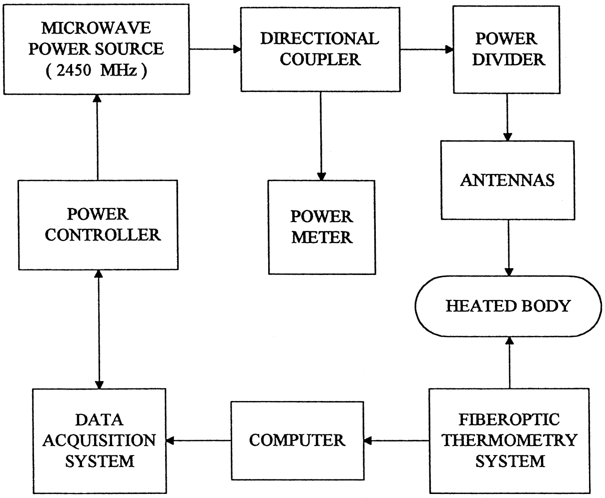
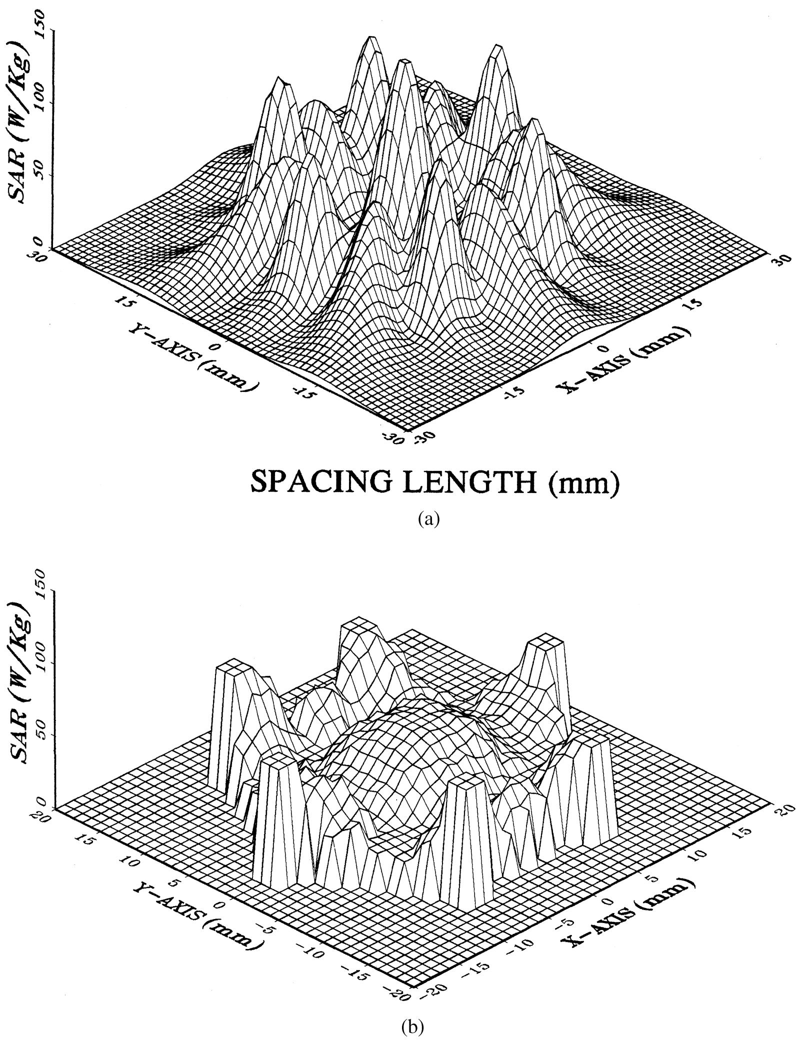


Figure 17.5 A system block diagram for interstitial hyperthermia.



(a)

(b)

Figure 17.6 SAR distributions produced by an hexagonal array of six 2450-MHz interstitial miniature microwave catheter antennas: (a) computer simulation and (b) measurements made in phantom tissue model (additional details are given in Ref. 161).

superficially conducted heat or invasive core perfusion. RF energy can potentially provide a noninvasive rapid whole-body rewarming that is safe and effective under most conditions. Likewise, extreme cold can seriously impair the performance and threaten the well-being of persons engaged in certain occupations. Conventional passive cold protection is inadequate because it necessitates the use of bulky gloves and precludes full use of hands and fingers, for example. Gloves with resistive wires through which a direct electrical current is passed to apply heat are limited by surface burns and excessive heat loss to the environment when used in water. In contrast, RF energy is known to be capable of bulk heating at depth in tissues and therefore has been studied to assess its feasibility both in elimination of deep hypothermia and in warming of the extremities exposed to severe environmental cold. In one study, a 13-turn helical coil RF rewarming system measured 34 cm long and 22 cm in diameter produced an average energy deposition rate of

5.5 W/kg in 10-kg rhesus monkeys with 60 W of 13.56-MHz power. The rectal temperature of the primate was raised from 28.5 to 34.5°C in 1 h compared to more than 2 h for conventional surface rewarming [87–90]. A study in human volunteers compared the rewarming effectiveness of a 13.56-MHz RF coil at a specific absorption rate of 2.5 W/kg with warm water immersion (40°C) and a mummy-type insulating sack under simulated mildly hypothermic conditions (35°C). It found that for mildly hypothermic individuals, active rewarming with RF at an SAR of 2.5 W/kg is about equivalent to passive rewarming with insulating sack, but less effective than warm water immersion [91].

Similar designs of helical coils operating at resonating frequency of 27.12 MHz have been used to efficiently warm hands and feet of subjects in cold air and water. The hand warming system consisted of a series of smaller insulated copper coils for each finger. Net powers of 15 and 25 W were delivered to the hands and feet, respectively. Subjects exposed to cold air at 10°C promptly reported warming sensations and gradual decrease of cold discomfort in tests where recorded finger tip temperatures were as low as 15°C [92]. These preliminary results from human volunteers demonstrate that RF technology can be applied to maintain hand and foot temperatures to within comfort ranges, and higher levels of the same energy can be used to rewarm whole-body hypothermia.

17.5. TISSUE IMAGING

A majority of imaging devices used by physicians to facilitate the diagnosis of diseases are based on sources that emit ionizing electromagnetic energy. Magnetic resonance imaging relies on the nonionizing static and RF magnetic fields and has been shown to offer a distinct advantage in a multitude of disease processes when compared to ionizing modalities. The current generation of MRI machines produces magnetic fields of 1.5 T (tesla) or less and therefore for imaging of hydrogen protons in tissue the frequency of applied RF field must be 63.9 MHz or lower according to Larmor resonance. MRI is the preferred imaging technique in the diagnosis of soft tissue disorders of the head, neck and

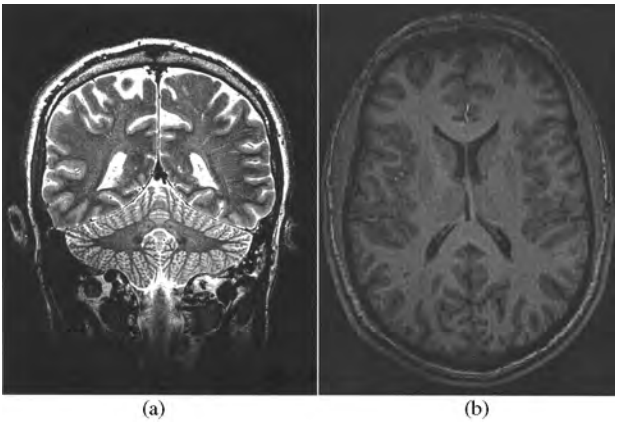


Figure 17.7 High-resolution MRI images of brain anatomy with different MR techniques: (a) coronal T2-weighted image obtained with a spin echo technique; (b) axial T1-weighted image obtained with a gradient echo technique. (Courtesy of Noam Alperin, Department of Radiology, University of Illinois, Chicago.)

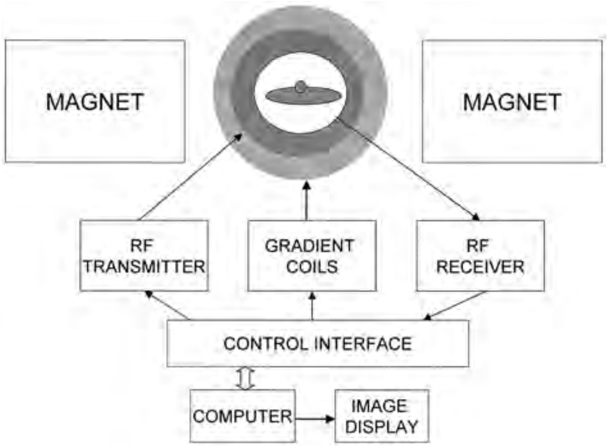


Figure 17.8 Instrumentation diagram for a typical magnetic resonance imaging system.

spinal regions (see Fig. 17.7 for MRI images of brain anatomy). Also, it has been shown to be important in the evaluation of disease progression [93]. The use of MRI is increasing for diagnostic evaluation of cardiovascular diseases and abdominal disorders in the method of MR angiography. While cardiovascular evaluation may soon become the next major domain of application for MRI after neurological studies [94,95], the problem of motion artifact remains as a significant challenge for MRI of the abdomen. An instrumentation diagram for magnetic resonance imaging is shown in Fig. 17.8.

MRI is based on the nuclear magnetic resonance of water molecules. An applied static magnetic field induces the magnetic dipole moments associated with hydrogen protons in tissue to be aligned in the direction of the applied field. Since it takes time to reach complete alignment, the magnetization of the dipole moments, $\boldsymbol{M}$ is described by,

$$\boldsymbol{M} = \boldsymbol{M}_{\mathrm{o}}[1 - e^{(-t/T_1)}] \tag{17.9}$$

where $\boldsymbol{M}_{\mathrm{o}}$ is the magnitude of the instantaneous magnetization and T_1 is the longitudinal relaxation time. In the absence of an applied magnetic field, the spatial orientations of the magnetic dipole moments are random, $\boldsymbol{M} = 0$. Each protons or dipole moment precesses about the static field, B_{o}, in the transverse plane, at the Larmor frequency,

$$\omega_0 = \gamma B_{\mathrm{o}} \tag{17.10}$$

where γ is the gyromagnetic ratio. For protons, $\gamma/2\pi = 42.6\,\mathrm{MHz/T}$. A time-varying magnetic field, $\boldsymbol{B}_1$ that oscillates at ω_0, is applied in the direction transverse to $\boldsymbol{B}_{\mathrm{o}}$ and then by choosing the duration of the excitation, the magnetic dipole moments can be tipped into the transverse direction and precess about $\boldsymbol{B}_1$. The decaying signal following the excitation is detected using a radio-frequency (RF) receiver.

To form an image, the signals originating from dipole moments in the body must be spatially resolved. This is accomplished through spatial encoding by using a gradient magnetic field. Specifically, a linear magnetic field gradient, G_x, in the x direction, is introduced so that,

$$\boldsymbol{B} = \boldsymbol{B}_{\mathrm{o}} + G_x \boldsymbol{x} \tag{17.11}$$

The magnetic resonance signal coming from each anatomic region is spatially encoded such that each region contributes a signal whose frequency is proportional to,

$$\omega = \gamma \mathbf{B} = \gamma[\mathbf{B}_{\mathrm{o}} + G_x x] \tag{17.12}$$

Thus the spatial location of the dipole moment giving rise to the MRI signal is obtained by a simple inversion of the Fourier transform of the received signal. The formation and display of the MR image is done efficiently, with the aid of a computer.

There are several other tissue imaging techniques currently under development that utilize electromagnetic technology [96–113]. The wide range of dielectric property variations offers a potential for higher contrast and better tissue characterization. Microwave tomography has been explored to reconstruct images associated with dielectric property variations in body cross section [96–101]. Microwave thermoelastic imaging uses microwave pulse-induced thermoelastic pressure waves to form planar or tomographic images [102–113]. Since the generation of thermoelastic or thermoacoustic pressure wave depends on permittivity, specific heat, and acoustic properties of tissue, microwave thermoelastic imaging possesses some unique features that are being explored as an imaging modality for noninvasive characterization of tissues.

17.6. NONINVASIVE AND REMOTE SENSING

Microwaves provide a convenient approach to detect and monitor physiological movements without compromising the integrity of these physiological events. In this case, microwave energy is directed to the target and the reflected signal is processed to yield information on the organ of interest or the physiological event under interrogation (Fig. 17.9). This noninvasive technique provides a capability for continuous monitoring as well as quantifying time-dependent changes in the cardiovascular and respiratory systems [114].

Currently, there are several areas in which noninvasive microwave contact or noncontact, close range, or remote approaches hold promise. These include heart rate, respiration, ventricular movement, pressure pulse sensing, and monitoring of superficial arterial circulation [115–126]. Low-frequency displacement of the precordium overlying the apex of the heart is related to movement in the left ventricle, and it echoes the hemodynamic events within the left ventricle. Microwave apexcardiograms obtained using 2.45 GHz showed close correlation to the hemodynamic events occurring within the left ventricle [118]. They involve detecting the reflected Doppler signal using an antenna

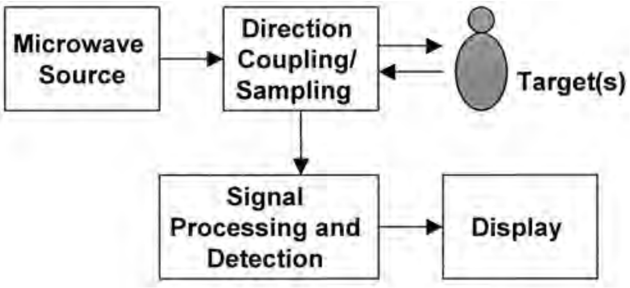


Figure 17.9 A microwave system for noninvasive sensing of physiological movements.

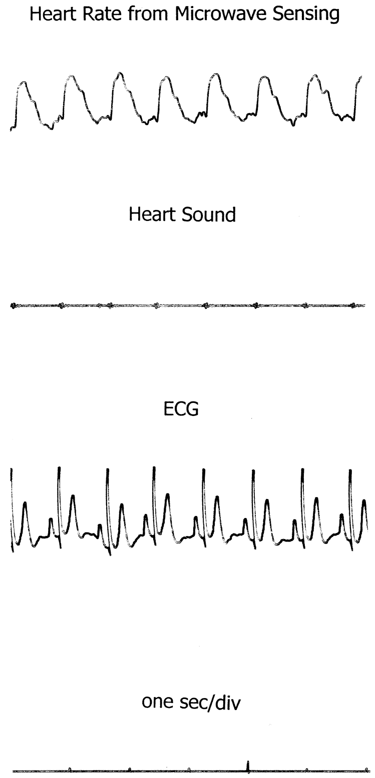


Figure 17.10 Human heart signal from remote, noncontract sensing using microwave radiation, showing simultaneous heart sound and electrocardiogram (ECG) recordings.

located a few centimeters over the apex of the heart (Fig. 17.10). This approach has several advantages over more conventional techniques because it does not require any physical contact with the subject. Problems such as skin irritation, restriction of breathing and electrode connections are easily eliminated.

Doppler microwaves have been employed to interrogate the wall properties and pressure pulse characteristics at a variety of arterial sites, including the carotid, brachial, radial, and femoral arteries [119–121]. Microwave-sensed carotid pulse waveforms have been obtained in patients using contact application of 25-GHz energy along with simultaneously recorded intra-aortic pressure waves. The resemblance of the microwave-sensed arterial pulse and the invasively recorded pressure wave is remarkable. These results confirm that a noninvasive Doppler microwave sensor can successfully and reproducibly detect pressure pulse waveforms of diagnostic quality. Because of its basis on motion detection, this continuous wave device will detect other movements as well. Other interesting applications are the use of microwaves for sensing cerebral edemas [122,123] and for speech articulator measurement [127].

The ability to detect remotely such vital signs as heart beat and respiration rates are particularly useful in situations where direct contact with the subject is either impossible or undesirable. Indeed, heart beat and respiration have been detected at distances of a few to tens of meters, with or without intervening nonmetallic barriers [124–126]. This approach has several advantages over more conventional techniques because it does not require any physical contact with the subject. Similar approaches may conceivably find uses in a variety of rescue related operations where direct physical contact with the subject is either impossible or undesirable.

17.7. MAGNETIC IMAGING AND STIMULATION

Magnetic fields are emerging as functional imaging modalities in the form of magnetocardiography (MCG) and magnetoencephalography (MEG) [128,129]. Magnetic stimulation for the diagnosis of neurological disorders represents yet another noncontact application of electromagnetic energy that is gaining in popularity [130,131]. MCG is a promising noninvasive modality for obtaining functional information on the electrical activity of the heart. The small biomagnetic signals (1 pT or less) are recorded using multichannel superconducting quantum interference devices (SQUID) with subjects in the supine position and the SQUID sensors placed directly over the thoracic surface [132]. It has been explored as a mapping tool to localize noninvasively cardiac electrical sources responsible for fetal atrial flutter and ventricular fibrillation [133–135]. In a like manner, high-resolution MEG measures the minute magnetic fields generated by the ionic currents in the brain [136]. MEG correlates well with abnormal electrical activity measured by electroencephalography (EEG), and sometimes it shows abnormalities not seen on an EEG. MEG is more accurate in spatial localization, as the magnetic field passes through the scalp and skull unimpeded, while the electrical signal of an EEG is attenuated and dispersed. It has been successfully used to localize cerebral sources of electrical activity [137–140].

In magnetic stimulation of the nervous system, a time-varying magnetic field, produced by passing a current through a wire coil, gives rise to an induced electric field in proximity to excitable tissues of the central and peripheral nervous system. Currents are delivered as 50–200 μs pulses with peak values of several thousand amperes. The technique is used to study the somatosensory or neuromuscular systems in humans [141–149]. The principal advantages of magnetic stimulation are that it is noninvasive and less painful than applying electrical currents through surface electrodes, and it has the ability to reach nerves lying well below the skin surface. The main disadvantages are that magnetic stimulation is not selective enough to restrict the region of excitation and it has relatively poor controllability and reproducibility.

17.8. BONE AND SOFT TISSUE HEALING

The therapeutic effects of low-frequency electric and magnetic fields have been studied extensively for their promotion of connective tissue repair. These studies concern bone repair and deal with acceleration of the healing of fresh fractures, delayed and nonunion, incorporation of bone grafts, osteoporosis, and osteonecrosis [150–152]. The most commonly used techniques included inductive and capacitive coupling, and implanted electrodes that induce voltages and currents similar to those produced, normally, during dynamic mechanical deformation of connective tissues. Indeed, there are three FDA-approved technologies:

1. Pulsed electromagnetic field (PEMF) technology using noncontact coils
2. Sinusoidal electric field (SEF) technology using skin contact electrodes
3. Direct-current technology using surgically implanted electrodes

Most studies indicate a dose-response relationship with current densities of 5 to 100 μA/cm^2, peak or 0.3 to 10 μA/cm^2, average in the bone and surrounding soft tissue. Sinusoidal and specialized waveforms with frequency contents extending from 10 Hz to 60 kHz have been employed. Although the majority of patients were treated with PEMF technology, more than 250,000 of them have been treated with these three technologies [150,152]. Success of treatment ranges from 65 to 90% depending upon such variables as infection, prior operation, and patient compliance. The total elapse times since initial trauma also influence rates of success; higher success rates are associated with short times since initial trauma. Moreover, several experimental and clinical investigations [153–160] involving tibial, scaphoid and other fractures have concluded that pulsed electromagnetic field is a safe and effective treatment for nonunion of bone fractures without discomfort, or the high costs of surgical repair.

As the specific requirements for field parameters are being defined, the range of treatable ills has been broadened to include nerve regeneration and wound healing, Specifically, the noninvasive treatment techniques of PEMF and SEF are being adapted for treatment of soft tissue injuries to reduce swelling and to accelerate wound healing [119–121]. While an acceleration of extracellular matrix synthesis and tissue healing has been observed in all these experimental systems and clinical applications, the underlying cellular mechanism of interaction of these fields upon the repair of cartilage and soft fibrous tissues is presently obscure.

REFERENCES

1. Lehmann, J.F. Diathermy. In *Handbook of Physical Medicine and Rehabilitation*; Krusen, F.H., Kottke, F.J., Elwood, P.M., Eds.; Saunders: Philadelphia, 1971, 273–345.
2. Lehmann, J.F. (Ed.). *Therapeutic Heat and Cold*; Williams & Wilkins: Baltimore, 1990.
3. Huang, S.K.S. (Ed.). *Radio-frequency Catheter Ablation of Cardiac Arrhythmias*; Futura: Armond, NY, 1995.
4. Wagshal, A.B.; Huang, S.K.S. Application of radiofrequency energy as an energy source for ablation of cardiac arrhythmias. In *Advances in Electromagnetic Fields in Living Systems*; Lin, J.C., Ed.; Plenum: New York, 1997; Vol. 2, 205–254.
5. Lin, J.C. Catheter microwave ablation therapy for cardiac arrhythmias. Bioelectromagnetics. **1999**, *20*, 120–132, Supplement 4.
6. Huang, S.K.; Bharati, S.; Graham, A.R.; Lev, M.; Marcus, F.I.; Odall, R.S. Closed chest catheter desiccation of the atrioventricular junction using radio-frequency energy—a new method of catheter ablation. J. Am. Coll. Cardiol. **1987**, *9*, 349–358.
7. Lin, J.C. Biophysics of radio-frequency ablation. In *Radio-frequency Catheter Ablation of Cardiac Arrhythmias: Basic Concepts and Clinical Applications*, 2nd Ed.; Huang, S.K.S., Wilber, D.J. Eds.; Futura: Armonk, New York, 2000; 13–24.
8. Beckman, K.J.; Lin, J.C.; Wang, Y.; Illes, R.W.; Papp, M.A.; Hariman, R.J. Production of reversible and irreversible atrio-ventricular block by microwave energy. Circulation **1987**, *76*, 1612.
9. Lin, J.C.; Wang, Y.J.; Hariman, R.J. Comparison of power deposition patterns produced by microwave and radio-frequency cardiac ablation catheters. Electronics Lett. **1994**, *30*, 922–923.
10. Blouin, L.T.; Marcus, F.I. The effect of electrode design on the efficiency of delivery of RF energy to cardiac tissue in vitro. PACE **1989**, *12*, 136–143.
11. Wittkampf, F.H.M.; Hauer, R.N.W.; EO Robles de Medina. Control of RF lesions size by power regulation. Circulation **1989**, *80*, 962–968.
12. Hoyt, R.H.; Huang, S.K.S.; Marcus, A.I.; Odell, R.S. Factors influencing transcatheter radiofrequency ablation of the myocardium. J. Appl. Cardiol. **1986**, *1*, 469–486.

13. Lin, J.C. Engineering and biophysical aspects of microwave and radio-frequency radiation. In *Hyperthermia*; Watmough, D.J., Ross, W.M., Eds.; Blackie: Glasgow, 1986; 42–75.
14. Wonnell, T.L.; Stauffer, P.R.; Langberg, J.J. Evaluation of microwave and radio-frequency catheter ablation in a myocardium-equivalent phantom model. IEEE Trans. Biomed. Engg. **1992**, *39*, 1086–1095.
15. Langberg, J.J.; Wonnell, T.; Chin, M.C.; Finkbeiner, W.; Scheinman, M.; Stauffer, P. Catheter ablation of the atrioventricular junction using a helical microwave antenna: a novel means of coupling energy to the endocardium. PACE **1991**, *14*, 2105–2113.
16. Liem, L.B.; Mead, R.H.; Shenasa, M.; Chun, S.; Hayase, M.; Kernoff, R. Microwave catheter ablation using a clinical prototype system with a lateral firing antenna design. Pacing Clin. Electrophysiol. **1998**, *21*, 714–721.
17. Shetty, S.; Ishii, T.K.; Krum, D.P.; Hare, J.; Mughal, K.; Akhtar, M.; Jazeyeri, M.R. Microwave applicator design for cardiac tissue ablations. J. Microwave Power and Electromagnetic Energy **1996**, *31*, 59–66.
18. Haugh, C.; Davidson, E.S.; Estes 3rd, N.A.; Wang, P.J. Pulsing microwave energy: a method to create more uniform myocardial temperature gradients. J. Interv. Card. Electrophysiol. **1997**, *1*, 57–65.
19. Nevels, R.D.; Arndt, G.D.; Raffoul, G.W.; Carl, J.R.; Pacifico, A. Microwave catheter design. IEEE Trans. Biomed. Engg. **1998**, *45*, 885–890.
20. Lin, J.C.; Wang, Y.J. A catheter antenna for percutaneous microwave therapy. Microwave Optical Technol. Lett. **1995**, *8*, 70–72.
21. Lin, J.C.; Wang, Y.J. The cap-choke catheter antenna for microwave ablation treatment. IEEE Trans. Biomed. Engg. **1996**, *43*, 657–660.
22. Lin, J.C.; Hariman, R.J.; Wang, Y.J.; Wang, Y.G. Microwave catheter ablation of the atrioventricular junction in closed-chest dogs. Med. Biolog. Engg. Comput. **1996**, *34*, 295–298.
23. Pisa, S.; Cavagnaro, M.; Bernardi, P.; Lin, J.C. A 915-MHz Antenna for microwave thermal ablation treatment: physical design, computer modeling, and experimental measurement. IEEE Trans. Biomed. Engg. **2001**, *48*, 599–601.
24. Michaelson, S.M.; Lin, J.C. *Biological Effects and Health Implications of Radio-Frequency Radiation*; Plenum: New York, 1987.
25. Pennes, H.H. Analysis of tissue and arterial blood temperatures in the resting human forearm, J. Appl. Physiol. **1948**, *1*, 93–122.
26. Huang, S.K.S.; Wilber, D.J. (Eds.). *Radio-Frequency Catheter Ablation of Cardiac Arrhythmias: Basic Concepts and Clinical Applications*, 2nd Ed.; Futura: Armonk, New York, 2000.
27. Wagshal, A.B.; Huang, S.K.S. Application of radiofrequency energy as an energy source for ablation of cardiac arrhythmias. In *Advances in Electromagnetic Fields in Living Systems*; Lin, J.C., Ed.; Plenum: New York, 1997; Vol. 2, 205–254.
28. Wagshal, A.B.; Pires, L.A.; Bonavita, G.J.; Mittleman, R.S.; Huang, S.K.S. Does the baseline impedance measurement during radio-frequency catheter ablation influence the likelihood of an impedance rise. Cardiology **1996**, *87*, 42–45.
29. Strickberger, S.A.; Weiss, R.; Knight, B.P.; Bahu, M.; Bogun, F.; Brinkman, K.; Harvey, M.; Goyal, R. Randomized comparison of two techniques for titrating power during radio-frequency ablation of accessory pathways. J. Cardiovascular Electrophysiology **1996**, *7*, 795–801.
30. Langberg, J.J.; Gallagher, M.; Strickberegere, A.S.; Amirana, O. Temperature-guided radio-frequency catheter ablation with very large distal electrode. Circulation **1993**, *88*, 245–249.
31. Dinerman, J.L.; Berger, R.D.; Calkins, H. Temperature monitoring during radiofrequency ablation. J. Cardiovascular Electrophysiol **1996**, *7*, 163–173.
32. Pires, L.A.; Huang, S.K.S.; Wagshal, A.B.; Mittleman, R.S.; Rittman, W.J. Temperature-guided radio-frequency catheter ablation of closed-chest ventricular myocardium with a noval thermister-tipped catheter. Am. Heart J. **1994**, *127*, 1614–1618.
33. Wen, Z.C.; Chen, S.A.; Chiang, C.E.; Tai, C.T.; Lee, S.H.; Chen, Y.Z.; Yu, W.C.; Huang, J.L.; Chang, M.S. Temperature and impedance monitoring during radio-frequency catheter ablation

of slow av node pathway in patients with atrioventricular node reentrant tachycardia. International J. Cardiol. **1996**, *57*, 257–263.

34. Mackey, S.; Thornton, L.; He, S.; Marcus, F.I.; Lampe, L.F. Simultaneous multipolar radio-frequency ablation in the monopolar mode increases lesion size. PACE **1996**, *19*, 1042–1048.
35. Langberg, J.J.; Lee, N.A.; Chin, M.C.; Rosenqvist, M. Radio-frequency catheter ablation: The effect of electrode size on lesion volume in vivo. PACE **1990**, *13*, 1242–1248.
36. Mittleman, R.S.; Huang, S.K.S.; Deguzman, W.T.; Cuenoud, H.; Wagshal, A.B.; Pires, L.A. Use of the saline infusion electrode catheter for improved energy delivery and increased lesion size in radio-frequency catheter ablation. PACE **1995**, *18*, 1022–1027.
37. Simmons, W.N.; Mackey, S.; He, D.S.; Marcus, F.I. Comparison of gold versus platinum electrodes on myocardial lesion size using radio-frequency energy. PACE **1996**, *19*, 398–402.
38. Langberg, J.J.; Wonnell, T.; Chin, M.C.; Finkbeiner, W.; Scheinman, M.; Stauffer, P. Catheter ablation of the atrioventricular junction using a helical microwave antenna: a novel means of coupling energy to the endocardium. PACE **1991**, *14*, 2105–2113.
39. Wonnell, T.L.; Stauffer, P.R.; Langberg, J.J. Evaluation of microwave and radio-frequency catheter ablation in a myocardium-equivalent phantom model. IEEE Trans. Biomed. Engg. **1992**, *39*, 1086–1095.
40. Liem, L.B.; Mead, R.H.; Shenasa, M.; Kernoff, R. In vitro and in vivo results of transcatheter microwave ablation using forward-firing tip antenna design. Pacing Clin. Electrophysiol. **1996**, *19*: 2004–2008.
41. Liem, L.B.; Mead, R.H.; Shenasa, M.; Chun, S.; Hayase, M.; Kernoff, R. Microwave catheter ablation using a clinical prototype system with a lateral firing antenna design. Pacing Clin. Electrophysiol. **1998**, *21*, 714–721.
42. Lin, J.C.; Beckman, K.J.; Hariman, R.J. Microwave ablation for tachycardia. Proc. IEEE/EMBS International Conf. 1989, pp. 1141–1142.
43. Lin, J.C.; Beckman, K.J.; Hariman, R.J.; Bharati, S.; Lev, M.; Wang, Y.J. Microwave ablation of the atrioventricular junction in open heart dogs. Bioelectromagnetics **1995**, *16*, 97–105.
44. Lin, J.C. Catheter microwave ablation therapy for cardiac arrhythmias. Bioelectromagnetics. **1999**, *20*, Supplement 4, 120–132.
45. Lin, J.C. Transcatheter microwave technology for treatment of cardiovascular diseases. In *Emerging Electromagnetic Medicine*; O'Connor, M.E., Bentall, R.H.C., Monahan, J.C., Eds.; Springer-Verlag: New York, 1990; 125–132.
46. Rosen, A.; Wallinsky, P.; Smith, D.; Shi, Y.; Kosman, Z.; Martinez-Hernandez, A.; Rosen, H.; Sterzer, F.; Mawhinney, D.; Presser, A.; Chou, J.S.; Goth, P.; Lowery, G. Percutaneous transluminal microwave balloon angioplasty. IEEE. Trans. Microwave Theory Tech. **1990**, *38*, 90–93.
47. Lin, J.C. Microwave technology for minimally invasive interventional procedures. Chinese J. Med. Biolog. Engg. **1993**, *13*, 293–304.
48. Nardone, D.T.; Smith, D.L.; Martinez-Hernandez, A.; Consigny, P.M.; Kosman, Z.; Rosen, A.; Walinsky, P. Microwave thermal balloon angioplasty in the atherosclerotic rabbit. Am. Heart J. **1994**, *27*, 198–203.
49. Landau, C.; Currier, J.W.; Haudenschild, C.C.; Minihan, A.C.; Heymann, D.; Faxon, D.P. Microwave balloon angioplasty effectively seals arterial dissections in an atherosclerotic rabbit model. J. Am. Coll. Cardiol. **1994**, *23*, 1700–1707.
50. Aagaard, J.; Jonler, M.; Fuglsig, S.; Christensen, L.L.; Jorgensen, H.S.; Noorgaard, J.P. Total transurethral resection vs minimal transurethral resection of the prostate—A 10-year followup study of urinary symptoms, uroflowmetry, and residual volume. Bri. J. Urol. **1994**, *74*, 333–336.
51. Sterzer, F.; Mendecki, J.; Mawhinney, D.D.; Friedenthal, E.; Melman, A. Microwave treatments for prostate disease. IEEE Trans. Microwave Theory Tech. **2000**, *48*, 1885–1891.
52. Harada, T.; Etori, K.; Nishizawa, O.; Noto, H.; Tsuchida, S. Microwave surgical treatment of diseases of the prostate. Urology **1985**, *26*, 572–76.

53. Harada, T.; Tsuchida, S.; Nishizawa, O.; Kigure, T.; Noto, H.; Etori, K.; Kumazaki, T.; Koh, D.; Shimoda, J. Microwave surgical treatment of diseases of the prostate: Clinical application of microwave surgery as a tool for improved prostatic electroresection. Urologia Internationalis **1987**, *42*, 127–31.
54. Montorsi, F.; Galli, L.; Guazzoni, G.; Colombo, R.; Bulfamante, G.; Barbieri, L.; Grazioli, V.; Rogatti, P. Transrectal microwave hyperthermia for benign prostatic hyperplasia—long-term clinical, pathological and ultrastructural patterns. J. Urology **1992**, *148*, 321–325.
55. Debicki, P.; Astrahan, M.A.; Ameye, F.; Oyen, R.; Baert, L.; Haczewski, A.; Petrovich, Z. Temperature steering in prostate by simultaneous transurethral and transrectal hyperthermia. Urology **1992**, *40*, 300–307.
56. Astrahan, M.A.; Sapozink, M.D.; Cohen, D.; Luxton, G.; Kampp, T.D.; Boyd, S.; Petrovich, Z. Microwave applicator of transurethral hyperthermia of benign prostatic hyperplasia. Intern. J. Hyperthermia **1989**, *5*, 383–396.
57. Strohmaier, W.L.; Bichler, K.H.; Fruchter, S.H.; Wilbert, D.M. Local microwave hyperthermia of benign prostatic hyperplasia. J. Urol. **1990**, *144*, 913–917.
58. Lindner, A.; Braf, Z.; Lev, A.; Golomb, J.; Lieb, Z.; Seigel, Y.; Servadio, C. Local hyperthermia of the prostatic gland for the treatment of benign prostate hypertrophy and urinary retention. Br. J. Urol. **1990**, *65*, 201–203.
59. Carter, SStC.; Patel, A.; Reddy, P.; Royer, P.; Ramsay, J.W.A. Single-session transurethral microwave thermotherapy for the treatment of benign prostate obstruction. J. Endourol. **1991**, *5*, 137–143.
60. Baert, L.; Willemen, P.; Ameye, F.; Astrahan, M.A.; Lanholz, B.; Petrovich, Z. Transurethral microwave hyperthermia: An alternative treatment for prostdynia. Prostate **1991**, *19*, 113–119.
61. Bostwick, D.G.; Larson, T.R. Transurethral microwave thermal therapy—pathologic findings in the canine prostate. Prostate **1995**, *26*, 116–122.
62. Zerbib, M.; Steg, A.; Conquy, S.; Martinache, P.R.; Flam, T.A.; Debre, B. Localized hyperthermia vs the sham procedure in obstructive benign hyperplasia of the prostate—a prospective randomized study. J. Urology **1992**, *147*, 1048–1052.
63. Dahlstrand, C.; Walden, M.; Geirsson, G.; Pettersson, S. Transurethral microwave thermotherapy versus transurethral resection for symptomatic benign prostatic obstruction: a prospective randomized study with a 2-year follow-up. Br. J. Urol. **1995**, *76*, 614–618.
64. Larson, T.R.; Collins, J.M.; Corica, A. Detailed interstitial temperature mapping during treatment with a novel transurethral microwave thermoablation system in patients with benign prostatic hyperplasia. J. Urology **1998**, *159*, 258–264.
65. Jepsen, J.V.; Bruskewitz, R.C. Recent developments in the surgical management of benign prostatic hyperplasia. Urology **1998**, *51*, 23–31.
66. Ramsey, E.W.; Miller, P.D.; Parsons, K. A novel transurethral microwave thermal ablation system to treat benign prostatic hyperplasia: results of a prospective multicenter clinical trial. J. Urol. **1997**, *158*, 112–119.
67. D'Ancona, F.C.; Francisca, E.A.; Hendriks, J.C.; Debruyne, F.M.; De La Rosette, J.J. High-energy transurethral thermotherapy in the treatment of benign prostatic hyperplasia: criteria to predict treatment outcome. Prostate Cancer Prostatic Dis. **1999**, *2*, 98–105.
68. Wagrell, L.; Schelin, S.; Nordling, J.; Richthoff, J.; Magnusson, B.; Schain, M.; Larson, T.; Boyle, E.; Duelund, J.; Kroyer, K.; Ageheim, H.; Mattiasson, H.A. Feedback microwave thermotherapy versus TURP for clinical BPH—a randomized controlled multicenter study. Urology **2002**, *60*, 292–299.
69. Norby, B.; Nielsen, H.V.; Frimodt-Moller, P.C. Transurethral interstitial laser coagulation of the prostate and transurethral microwave thermotherapy vs. transurethral resection or incision of the prostate: results of a randomized, controlled study in patients with symptomatic benign prostatic hyperplasia. Br. J. Urol. Int. **2002**, *90*, 853–862.
70. Lin, J.C. (Ed.). Special issue on phased arrays for hyperthermia treatment of cancer. Trans. IEEE Microwave Theory Tech. **1986**, *34*, 481–482.

71. Fessenden, P.; Hand, J.W. Hyperthermia therapy physics. In *Medical Radiology—Radiation Therapy Physics*; Smith, A.R., Ed.; Springer-Verlag: Berlin, 1995; 315–363.
72. Lin, J.C. Hyperthermia therapy. In *Encyclopedia of Electrical and Electronics Engineering*; Webster, J.G., Ed.; Wiley: New York, 1999; Vol. 9. 450–460.
73. Rosen, A.; Vender Vosrt, A.; Kotsuka, Y. (Eds.). Special issue on medical applications in medicine. IEEE Trans. Microwave Theory Tech. **2000**, *48*, 1885–1891.
74. Vernon, C.C.; Hand, J.W.; Field, S.B.; Machin, D.; Whaley, J.B.; Vanderzee, J.; Vanputten, W.L.J.; Vanrhoon, G.C.; Vandijk, J.D.P.; Gonzalez, D.G.; Liu, F.F.; Goodman, P.; Sherar, M. Radiotherapy with or without hyperthermia in the treatment of superficial localized breast cancer—results from five randomized controlled trials. Int. J. Radiation Oncol. Biol. Phy. **1996**, *35*, 731–744.
75. Kuwano, H.; Sumiyoshi, K.; Watanabe, M.; Sadanaga, N.; Nozoe, T.; Yasuda, M.; Sugimachi, K. Preoperative hyperthermia combined with chemotherapy and irradiation for the treatment of patients with esophageal carcinoma. Tumori. **1995**, *81*, 18–22.
76. Emami, B.; Scott, C.; Perez, C.A.; Asbell, S.; Swift, P.; Grigsby, P.; Montesano, A.; Rubin, P.; Curran, W.; Delrowe, J.; Arastu, H.; Fu, K.; Moros, E. Phase III study of interstitial thermoradiotherapy compared with interstitial radiotherapy alone in the treatment of recurrent or persistent human tumors—a prospectively controlled randomized study by the Radiation Therapy Oncology Group. Int. J. Radiation Oncol. Biol. Phy. **1996**, *34*, 1097–1104.
77. Matsuda, T. The present status of hyperthermia in Japan. Ann. Acad. Med. Singapore **1996**, *25*, 420–424.
78. Overgaard, J.; Gonzalez, D.G.; Hulshof, M.C.C.M.; Arcangeli, G.; Dahl, O.; Mella, O.; Bentzen, S.M. Hyperthermia as an adjuvant to radiation therapy of recurrent or metastatic malignant melanoma—a multicenter randomized trial by the European Society for Hyperthermic Oncology. Int. J. Hyperthermia **1996**, *12*, 3–20.
79. Falk, M.H.; Issels, R.D. Hyperthermia in oncology. Int. J. Hyperthermia, **2001**, *17*, 1–18.
80. Moroz, P.; Jones, S.K.; Gray, B.N. Status of hyperthermia in the treatment of advanced liver cancer. J. Surg. Oncol. **2001**, *77*, 259–269.
81. Vanbaren, P.; Ebbini, E.S. Multipoint temperature control during hyperthermia treatments—theory and simulation. IEEE Trans. Biomed. Engg. **1995**, *42*, 818–827.
82. Qi, C.; Li, D.J. Thermometric analysis of intra-cavitary hyperthermia for esophageal cancer. Int. J. Hyperthermia **1999**, *15*, 399–407.
83. Lewa, C.J.; de Certaines, J.D. Body temperature mapping by magnetic resonance imaging. Spectroscopy Lett. **1994**, *27*, 1369–1419.
84. Young, I.R.; Hand, J.W.; Oatridge, A.; Prior, M.V. Modeling and observation of temperature changes in vivo using MRI. Magnetic Resonance Med. **1994**, *32*, 358–369.
85. Macfall, J.R.; Prescott, D.M.; Charles, H.C.; Samulski, T.V. H-1 MRI phase thermometry in vivo in canine brain, muscle, and tumor tissue. Med. Phys. **1996**, *23*, 1775–1782.
86. Kowalski, M.E.; Behnia, B.; Webb, A.G.; Jin, J.M. Optimization of electromagnetic phased-arrays for hyperthermia via magnetic resonance temperature estimation. IEEE Trans. Biomed. Engg. **2002**, *49*, 1229–1241.
87. Olsen, R.G.; David, T.D. Hypothermia and electromagnetic rewarming in the rhesus monkey. Aviation, Space, and Environmental Medicine **1984**, *55*, 1111–1117.
88. Olsen, R.G.; Ballinger, M.B.; David, T.D.; Lotz, W.G. Rewarming of the hypothermic rhesus monkey with electromagnetic radiation. Bioelectromagnetics **1987**, *8*:183–93.
89. Olsen, R.G. Reduced temperature afterdrop in rhesus monkeys with RF rewarming. Aviat Space Environ Medicine **1988**, *59*, 78–80.
90. Hesslink, Jr. R.L.; Pepper, S.; Olsen, R.G.; Lewis, S.B.; Homer, L.D. Radio frequency (13.56 MHz) energy enhances recovery from mild hypothermia. J. Appl. Physiol. **1989**, *67*, 1208–1212.
91. Kaufman, J.W.; Hamilton, R.; Dejneka, K.Y.; Askew, G.K. Comparative effectiveness of hypothermia rewarming techniques: radio-frequency energy vs. warm water. Resuscitation. **1995**, *29*, 203–214.

92. Lloyd, J.R.; Olsen, R.G. Radiofrequency energy for rewarming of cold extremities. Undersea Biomed. Res. **1992**, *19*, 199–207.

93. Stark, D.D.; Bradley, Jr. W.G. *Magnetic Resonance Imaging*, 3rd Ed.; Mosby-Year Book: St. Louis, 1999.

94. Nield, L.E.; Qi, X.; Yoo, S.J.; Valsangiacomo, E.R.; Hornberger, L.K.; Wright, G.A. MRI-based blood oxygen saturation measurements in infants and children with congenital heart disease. Pediatr. Radiol. **2002**, *32*, 518–522.

95. Plein, S.; Ridgway, J.P.; Jones, T.R.; Bloomer, T.N.; Sivananthan, M.U. Coronary artery disease: assessment with a comprehensive MR imaging protocol—initial results. Radiology **2002**, *225*, 300–307.

96. Larsen, L.E.; Jacobi, J.H. Microwave scattering parameter imagery of an isolated canine kidney. Med. Phys. **1979**, *6*, 394–403.

97. Guerquin-Kern, J.L.; Gautherie, M.; Peronnet, G.; Jofre, L.; Bolomey, J.C. Active microwave tomographic imaging of isolated, perfused animal organs. Bioelectromagnetics **1985**, *6*, 145–156.

98. Meaney, P.M.; Paulsen, K.D.; Hartov, A.; Crane, R.K. Microwave imaging for tissue assessment: initial evaluation in multitarget tissue-equivalent phantoms. IEEE Trans. Biomed. Engg. **1996**, *43*, 878–890.

99. Semenov, S.Y.; Svenson, R.H.; Boulyshev, A.E.; Souvorov, A.E.; Borisov, V.Y.; Sizov, Y.; Starostin, A.N.; Dezern, K.R.; Tatsis, G.P.; Baranov, V.Y. Microwave tomography—two-dimensional system for biological imaging. IEEE Transactions on Biomedical Engineering **1996**, *43*, 869–877.

100. Franchois, A.; Joisel, A.; Pichot, C.; Bolomey, J.C. Quantitative microwave imaging with a 2.45-GHz planar microwave camera. IEEE Trans. Med. Imag. **1998**, *17*, 550–561.

101. Semenov, S.Y.; Svenson, R.H.; Bulyshev, A.E.; Souvorov, A.E.; Nazarov, A.G.; Sizov, Y.E.; Posukh, V.G.; Pavlovsky, A.; Repin, P.N.; Starostin, A.N.; Voinov, B.A.; Taran, M.; Tatsis, G.P.; Baranov, V.Y. Three-dimensional microwave tomography: initial experimental imaging of animals. IEEE Trans. Biomed. Eng. **2002**, *49*, 55–63.

102. Olsen, R.G.; Lin, J.C. Acoustical imaging of a model of a human hand using pulsed microwave irradiation. Bioelectromagnetics **1983**, *4*, 397–400.

103. Lin, J.C.; Chan, K.H. Microwave thermoelastic tissue imaging—system design. IEEE Trans. Microwave Theory Tech. **1984**, *32*, 854–860.

104. Chan, K.H. Microwave-induced thermoelastic tissue imaging. PhD dissertation, University of Illinois, Chicago, 1988.

105. Chan, K.H.; Lin, J.C. Microwave-induced thermoelastic tissue imaging. Proc. IEEE/EMBS Annual International Conference, New Orleans, 1988, pp. 445–446.

106. Su, J.L. Computer-assisted tomography using microwave-induced thermoelastic waves. Thesis PhD dissertation, University of Illinois, Chicago, 1988.

107. Su, J.L.; Lin, J.C. Computerized Thermoelastic Wave Tomography, World Cong. Med Phys Biomed Engg, Kyoto, Japan, July, 1991.

108. Lin, J.C. Auditory perception of pulsed microwave radiation. In *Biological Effects and Medical Applications of Electromagnetic Fields*; Gandhi, O.P. Ed.; Prentice-Hall: New York, 1990, Chapter 12, 277–318.

109. Dajani, N.F. 3D finite difference time-domain scattering and computed tomography in microwave medical imaging. PhD dissertation, University of Illinois, Chicago, 2001.

110. Kruger, R.A.; Reinecke, D.R.; Kruger, G.A. Thermoacoustic computed tomography—technical considerations. Med. Phys. **1999**, *26*, 1832–1837.

111. Kruger, R.A.; Miller, K.D.; Reynolds, H.E.; Kiser, W.L.; Reinecke, D.R.; Kruger, G.A. Breast cancer in vivo: Contrast enhancement with thermoacoustic CT at 434 MHz—Feasibility study. Radiology **2000**, *216*, 279–283.

112. Ku, G.; Wang L.V. Scanning thermoacoustic tomography in biological tissue. Med. Phys. **2000**, *27*, 1195–1202.

113. Xu, M.; Wang, L.V. Pulsed-microwave-induced thermoacoustic tomography: filtered back-projection in a circular measurement configuration. Med. Phys. **2002**, *29*, 1661–1669.

114. Lin, J.C. Microwave sensing of physiological movement and volume change. Bioelectromagnetics **1992**, *13*, 557–565.

115. Lohman, B.; Boric-Lubecke, O.; Lubecke, V.M.; Ong, P.W.; MM Sondhi. A digital signal processor for Doppler radar sensing of vital signs. IEEE Eng. Med. Biol. Mag. **2002**, *21*, 161–164.

116. Lin, J.C. Noninvasive Microwave measurement of respiration, *Proc. IEEE* **1975**, *63*, 1530.

117. Lin, J.C.; Dawe, E.; Majcherek, J. A noninvasive microwave apnea detector, Proceedings 1977 San Diego Biomedical Symposium, Academic Press, 1977, pp. 441–443.

118. Lin, J.C.; Kiernicki, J.; Kiernicki, M.; Wollschlaeger, P.B. Microwave apexcardiography. IEEE Trans. Microwave Theory Tech. *27*, 618–620, 1979.

119. Lee, J.Y.; Lin, J.C. A microprocessor based noninvasive pulse wave analyzer. IEEE Trans. Biomed. Engg. **1985**, *32*, 451–455.

120. Papp, M.A.; Hughes, C.; Lin, J.C.; Pouget, J.M. Doppler microwave: a clinical assessment of its efficacy as an arterial pulse sensing technique. Invest. Radiol. **1987**, *22*, 569–573.

121. Thansandote, A.; Stuchly, S.S.; Smith, A.M. Monitoring variations of biological impedances using microwave Doppler radar. Phys. Med. Biol. **1983**, *28*, 983–990.

122. Lin, J.C.; Clarke, M.J. Microwave imaging of cerebral edema, Proc. IEEE, **1982**, *70*, 523–524.

123. Clarke, M.J.; Lin, J.C. Microwave sensing of increased intracranial water content. Invest. Radiol. **1983**, *18*, 245–248.

124. Lin, JC. Microwave propagation in biological dielectrics with application to cardiopulmonary interrogation. In *Medical Applications of Microwave Imaging*; Larsen, L.E., Jacobi, J.H. Eds.; IEEE Press: New York, 1986, 47–58.

125. Chen, K.M.; Misra, D.; Wang, H.; Chuang, H.R.; Postow, E. An X-band microwave life-detection system. IEEE Trans. Biomed. Eng. **1986**, *33*, 697–701.

126. Chan, K.H.; Lin, J.C. Microprocessor based cardiopulmonary rate monitor, Med. Biol. Engg. and Comput. **1987**, *25*, 41–44.

127. Holzrichter, J.F.; Burnett, G.C.; Ng, L.C.; Lea, W.A. Speech articulator measurements using low power EM-wave sensors. J. Acoust. Soc. Am. **1998**, *103*, 622–625.

128. Hukkinen, K.; Kariniemi, V.; Katila, T.E.; Laine, H.; Lukander, R.; Makipaa, P. Instantaneous fetal heart rate monitoring by electromagnetic methods. Am. J. Obstet. Gynecol. **1976**, *125*, 1115–1120.

129. Tesche, C.D. Noninvasive detection of ongoing neuronal population activity in normal human hippocampus. Brain Res. **1997**, *749*, 53–60.

130. Ueno, S.; Tashiro, T.; Harada, K. Localized stimulation of neural tissues the brain by means of a paired configuration of time-varying magnetic fields. J. Appl. Phys. **1988**, *64*, 5862–5864.

131. Ueno, S.; Matsuda, T.; Hiwaki, O. Localized stimulation of the human brain and spinal cord by a pair of opposing pulsed magnetic fields. J. Appl. Phys. **1990**, *67*, 5838–5840.

132. Stroink, G. Principles of cardiomagnetism. In *Advances in Biomagnetism*; Williamson, S.J. Ed.; Plenum Press: New York, 1989; 47–56.

133. Peters, M.J.; Stinstra, J.G.; van den Broek, S.P.; Huirne, J.A.F.; Quartero, H.W.F.; ter Brake, H.J.M.; Rogalla, H. On the fetal magnetocardiogram. Bioelectrochem. Bioenerget. **1998**, *47*, 273–281.

134. Wakai, R.T.; Leuthold, A.C.; Martin, C.B. Atrial and ventricular fetal heart rate patterns in isolated congenital complete heart block detected by magnetocardiography. Am. J. Obstet. Gynecol. **1998**, *179*, 258–260.

135. Stinstra, J.; Golbach, E.; van Leeuwen, P.; Lange, S.; Menendez, T.; Moshage, W.; Schleussner, E.; Kaehler, C.; Horigome, H.; Shigemitsu, S.; Peters, M.J. Multicenter study of fetal cardiac time intervals using magnetocardiography. BJOG (an international journal of obstetrics and gynaecology). **2002**, *109*, 1235–1243.

136. Barth, D.S. Magnetoencephalography. In *The Treatment of Epilepsy: Principles and Practice*; Wyllie, E., Ed.; Lea & Febiger: Philadelphia, 1993, 285–297.
137. Sekihara, K.; Abraham-Fuchs, K.; Stefan, H.; Hellstrandt, E. Multichannel biomagnetic system for study of electrical activity in the brain and heart. Radiology **1990**, *176*, 825–830.
138. Pantev, C.; Gallen, C.; Hampson, S.; Buchanan, S.; Sobel, D. Reproducibility and validity of neuromagnetic source localization using a large array biomagnetometer, Am. J. EEG Technol. **1991**, *31*, 83–101.
139. Babb, C.W.; Coon, D.R.; Rechnitz, G.A. Biomagnetic neurosensors. 3. Noninvasive sensors using magnetic stimulation and biomagnetic detection. Anal. Chem. **1995**, *67*, 763–769.
140. Roberts, T.P.; Rowley, H.A. Magnetic source imaging as a tool for presurgical functional brain mapping. Neurosurg. Clin. N. Am. **1997**, *8*, 421–438.
141. Amassian, V.E.; Cracco, R.Q.; Maccabe, J.P. Focal stimulation of human cerebral cortex with the magnetic coil: A comparison with electrical stimulation, Electroencephal Clin. Neurophysiol. **1989**, *74*, 401–416.
142. Chokroverty, S. *Magnetic Stimulation in Clinical Neurophysiology*; Butterworth: Boston, 1990.
143. Ueno, S.; Tashiro, T.; Harada, K. Localized stimulation of neural tissues the brain by means of a paired configuration of time varying magnetic fields. J. Appl. Phys. **1988**, *64*, 5862–5864.
144. Ueno, S.; Masuda, T.; Fujiki, M. Functional mapping of the human motor cortex obtained by focal and vectorial magnetic stimulation of the brain. IEEE Trans. Magn. **1990**, *26*, 1539–1544.
145. Kobayashi, M.; Ueno, S.; Kurokawa, T. Importance of soft tissue inhomogeneity in magnetic peripheral nerve stimulation. Electroencephalog. Clin. Neurophysiol.: Electromyog. Motor Control **1997**, *105*, 406–413.
146. Evans, B.A. Magnetic stimulation of the peripheral nervous system, J. Clin. Neurophysiol. **1991**, *8*, 77–84.
147. McMillan, A.S.; Watson, C.; Walshaw, D. Transcranial magnetic-stimulation mapping of the cortical topography of the human masseter muscle. Arch. Oral Biol. **1998**, *43*, 925–931.
148. Ishii, R.; Schulz, M.; Xiang, J.; Takeda, M.; Shinosaki, K.; Stuss, D.T.; Pantev, C. MEG study of long-term cortical reorganization of sensorimotor areas with respect to using chopsticks. NeuroReport **2002**, *13*, 2155–2159.
149. Kanno, A.; Nakasato, N.; Hatanaka, K.; Yoshimoto, T. Ipsilateral area 3b responses to median nerve somatosensory stimulation. Neuroimage **2003**, *18*, 169–177.
150. Bassett, C.A. Beneficial effects of electromagnetic fields. J. Cell Biochem. **1993**, *51*, 387–393.
151. Aaron, R.K.; Ciombor, D.M. Therapeutic effects of electromagnetic fields in the stimulation of connective tissue repair. J. Cell Biochem. **1993**, *51*, 42–46.
152. Polk, C. Therapeutic applications of low frequency electric and magnetic fields. In *Advances in Electromagnetic Fields in Living Systems*; Lin, J.C., Ed.; Plenum Press: New York, 1994; Vol. 1, 129–153.
153. Pienkowski, D.; Pollack, S.R.; Brighton, C.T.; Griffith, N.J. Low-power electromagnetic stimulation of osteotomized rabbit fibulae. A randomized, blinded study. J. Bone Joint Surg. Am. **1994**, *76*, 489–501.
154. Grace, K.; Revell, W.; Brookes, M. The effects of pulsed electromagnetism on fresh fracture healing: osteochondral repair in the rat femoral groove. Orthopedics **1998**, *21*, 297–302.
155. Godley, D. Nonunited carpal scaphoid fracture in a child: treatment with pulsed electromagnetic field stimulation. Orthopedics **1997**, *20*, 718–719.
156. Kenkre, J.E.; Hobbs, F.D.; Carter, Y.H.; Holder, R.L.; Holmes, E.P. A randomized controlled trial of electromagnetic therapy in the primary care management of venous leg ulceration. Fam. Pract. **1996**, *13*, 236–241.
157. Patino, O.; Grana, D.; Bolgiani, A.; Prezzavento, G.; Mino, J.; Merlo, A.; Benaim, F. Pulsed electromagnetic fields in experimental cutaneous wound healing in rats. J. Burn Care Rehabil. **1996**, *17*, 528–531.

158. Scardino, M.S.; Swaim, S.F.; Sartin, E.A.; Steiss, J.E.; Spano, J.S.; Hoffman, C.E.; Coolman, S.L.; Peppin. B.L. Evaluation of treatment with a pulsed electromagnetic field on wound healing, clinicopathologic variables, and central nervous system activity of dogs. Am. J. Vet. Res. **1998**, *59*, 1177–1181.
159. Ito, H.; Shirai, Y. The efficacy of ununited tibial fracture treatment using pulsing electromagnetic fields: relation to biological activity on nonunion bone ends. J. Nippon Med. Sch. **2001**, *68*, 149–153.
160. Inoue, N.; Ohnishi, I.; Chen, D.; Deitz, L.W.; Schwardt, J.D.; Chao, E.Y. Effect of pulsed electromagnetic fields (PEMF) on late-phase osteotomy gap healing in a canine tibial model. J. Orthop. Res. **2002**, *20*, 1106–1114.
161. Lin, J.C.; Hirai, S.; Chiang, C.L.; Hsu, W.L.; Su, J.L.; Wang, Y.J. Computer simulation and experimental studies of SAR distributions of interstitial arrays of sleeved-slot microwave antennas for hyperthermia treatment of brain tumors. IEEE Trans. Microwave Theory Techniq. **2000**, *48*, 2191–2197.

18

Measurement Techniques for the Electromagnetic Characterization of Biological Materials

Mohammad-Reza Tofighi and Afshin Daryoush
Drexel University
Philadelphia, Pennsylvania, U.S.A.

18.1. INTRODUCTION

The knowledge of material parameters at RF frequencies and above is important in various industrial [1,2], medical, and regulatory applications [3,4]. Data are available for a variety of materials ranging from PCB substrates and laminates [1,2] to biological samples [5].

Complex permittivity measurements of biological materials are of particular importance since interest exists to understand the interaction of electromagnetic energy with biological tissues up to millimeter wave frequencies. Some of this interest stems from the potential health hazards due to the advent of wireless communications [4], therapeutic applications of microwaves [3], and microwave imaging [6].

The therapeutic [3], dosimetric [4], or other applications [7–10] require a knowledge of dielectric properties of the tissues. From the electromagnetic engineering point of view, studying the bulk dielectric properties remains the most direct way of characterizing any substance. With the knowledge of these properties as they appear in Maxwell's equation, the absorption of energy and the field distribution which are the results of the solution to a boundary value problem can be obtained. Today, varieties of techniques exist to numerically solve Maxwell's equation. On the other hand, at the microscopic level, the underlying physics is much more complicated compared with the existing formulations for the bulk effects and not well understood yet. However, observations made at the macroscopic level can greatly contribute to an understanding of the microscopic phenomena.

There are many good references dedicated to the issue of the interaction of RF/microwaves with biological systems and medical applications of microwaves [3,4,6–10]. Most popular applied publications are in regards to EM energy absorption in human body as a result of cell phone radiation [4,11–16], applications where heating is needed [2,17,18], hyperthermia treatment of cancer tumors [19–21], microwave catheters for ablation [3,22], and microwave imaging [6,23–25].

This chapter is primarily focused on biological materials. Section 18.2 presents the theory of relaxation and the relaxation phenomena in biological substances. Various techniques for permittivity measurement are addressed in Sec. 18.3. Section 18.4 highlights the technical issues related to the characterization of biological materials using the commonly used open-ended coaxial probe. Moreover, a two-port measurement test fixture is presented that provides the accurate complex permittivity of brain tissue up to 50 GHz.

18.2. COMPLEX PERMITTIVITY OF MATERIALS

The dielectric theory literature, mathematical models (i.e., Debye and Cole-Cole), and physical mechanisms involved are briefly reviewed in this section. Techniques for permittivity measurement are reviewed in the rest of the chapter.

18.2.1. Applications and Significance

From a macroscopic view, in electromagnetic problems dielectric properties of materials are quantified by their bulk complex permittivity. This complex permittivity is represented by:

$$\varepsilon = \varepsilon_0(\varepsilon' - j\varepsilon'') = \varepsilon_0\left(\varepsilon' - \frac{j\sigma}{\omega\varepsilon_0}\right) = \varepsilon_0\varepsilon'(1 - j\tan\delta) \tag{18.1}$$

where $\varepsilon_0 = 8.854 \times 10^{-12}$ (F/m) is the free space permittivity, σ is the conductivity (S/m), and $\tan\delta = \varepsilon''/\varepsilon'$ is the loss tangent. The imaginary part is the term associated with the absorption of electrical energy. Since a field in linear systems can be represented by a summation of plane waves, the plane wave propagation parameters bear physically meaningful information. Three important wave parameters are the phase constant (β), the attenuation constant (α), and the wavelength (λ). For a plane wave, these parameters are related to the medium constitutive parameters and at frequency of (ω are given as [24]:

$$\alpha = \frac{\omega}{c}\sqrt{\frac{\varepsilon'}{2}}\sqrt{\sqrt{1+\tan\delta^2} - 1}\,(\text{rad/m}) \tag{18.2}$$

$$\beta = \frac{\omega}{c}\sqrt{\frac{\varepsilon'}{2}}\sqrt{\sqrt{1+\tan\delta^2} + 1}\,(\text{Np/m}) \tag{18.3}$$

$$\lambda = \frac{\lambda_0}{\sqrt{\varepsilon'/2}\sqrt{\sqrt{1+\tan\delta^2} + 1}}\ (\text{m}) \tag{18.4}$$

where c is the speed of the light in vacuum, and λ_0 is the wavelength in free space. The inverse of attenuation, i.e., the penetration depth, is also an important parameter describing the wave penetration in lossy materials.

18.2.2. Relaxation Theory

The extent to which the fields interact with the materials depends on the dielectric parameters of those materials. The classical theory of dielectrics can be found in books by Frohlich [26] and Daniel [27]. Frohlich's book [26] covers the basic macroscopic theory, static and dynamic properties. That book reviews important topics in dielectric theory such as dipolar interaction, dipolar molecules in gases and dilute solutions, Debye theory, and resonance absorption. Classical relaxation theory is presented comprehensively in Daniel's book [27].

Due to the complexity of biological substances, a complete understanding of bioelectrical interactions requires at least a sufficient knowledge of biochemistry, biology, electrical engineering, and physics [28–32]. The book by Pethig [28] presents the electronic and dielectric properties of biological material by bringing together the relevant issues from all these disciplines. He reviews not only the basic dielectric theory but also the dielectric properties of biopolymers (which are amino acids, polypeptides, and side chains), the role of water in biological systems, heterogeneous material (Maxwell-Wagner and counterion theory), and quantum mechanical aspects.

Works by Schwan and coworkers, started as early as 1950s [30–32], have provided significant enhancement in relating the macroscopic theory to the microscopic phenomena and in the interpretation of different relaxation mechanisms, from dc up to the microwave region.

The electrical properties of a material exposed to electromagnetic fields are in general frequency dependent. A material, which demonstrates significant permittivity changes in the frequency range of interest is referred to as a dispersive one in that range.

In order to understand the physics of dielectrics, the first step is developing a theory for the static properties of dielectric materials, i.e., when the electric field has no time variation. A good review of these properties in sufficient details relevant to biological tissues can be found in Ref. 28.

A phenomenological approach for the mathematical modeling of dispersion is the Debye theory [27,28]. The theory suggests a first-order differential equation system, similar to charge of a linear RC circuit. The complex permittivity in the frequency domain reduces to the well-known Debye equation

$$\varepsilon = \varepsilon' - j\varepsilon'' = \varepsilon_\infty + \frac{\varepsilon_s - \varepsilon_\infty}{1 + j\omega\tau} \tag{18.5}$$

where ε_s and ε_∞ are the static and optical dielectric constants and τ (i.e., the relaxation time) is a time constant of this first-order system. To study the Debye relaxation phenomena at microwave frequency range, ε_∞ is considered as the permittivity value at sufficiently high frequencies, where the orientational effects have disappeared. For small and relatively simple molecular structures (e.g., water), there is often only a single relaxation process (cf. Fig. 18.1). In contrast, for polymers and biological tissues, the dielectric dispersion can consist of several components associated with small side chain movements and the whole macromolecular movement (cf. Fig. 18.2). From Eq. (18.5), the real and imaginary parts of complex permittivity, ε, are

$$\varepsilon' = \varepsilon_\infty + \frac{\varepsilon_s - \varepsilon_\infty}{1 + (\omega\tau)^2} \tag{18.6}$$

$$\varepsilon'' = \frac{(\varepsilon_s - \varepsilon_\infty)\omega\tau}{1 + (\omega\tau)^2} \tag{18.7}$$

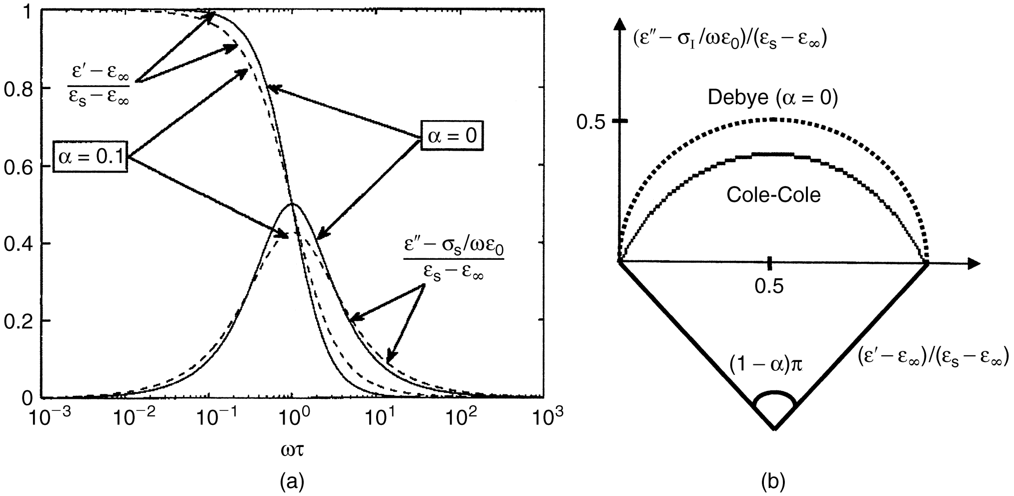


Figure 18.1 Cole-Cole plot for ε' and ε'' as (a) a function of ω and (b) arc plot in complex plane. $\alpha = 0$ corresponds to Debye equation.

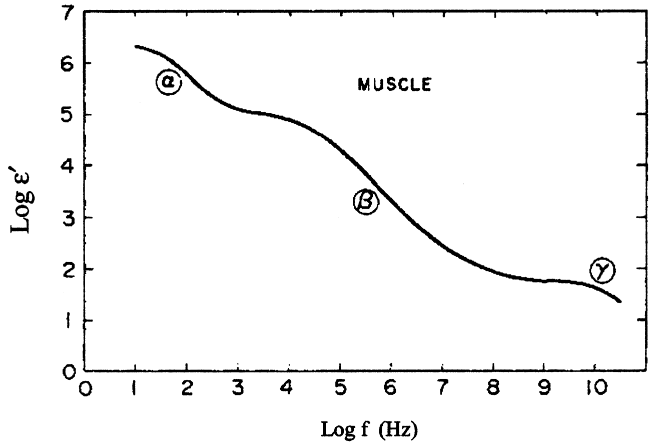


Figure 18.2 The real part of permittivity of muscle tissues. α, β, and γ dispersions are identified [31] (with permission from IEEE).

For Debye relaxation the characteristic frequency is defined as $f_c = 1/2\pi\tau$, which is half way between its low and high frequency values. Dielectric relaxation behavior can be best represented by Argand diagrams [27] on the complex plane

$$\left(\varepsilon' - \frac{\varepsilon_s + \varepsilon_\infty}{2}\right)^2 + \left(\varepsilon'' - \frac{\sigma_I}{\omega\varepsilon_0}\right)^2 = \left(\frac{\varepsilon_s - \varepsilon_\infty}{2}\right)^2 \tag{18.8}$$

where σ_I is the ionic conductivity of the medium. The plot of $(\varepsilon'' - \sigma_I/\omega\varepsilon_0)$ vs. ε' will lead to a semicircle centered at $(\varepsilon_s + \varepsilon_\infty)/2$. Debye theory is the basis of relaxation models proposed for interpretation of the observed dispersion of real materials.

18.2.3. Models for Relaxation

Most real materials do not exhibit single time constant relaxation behavior. In concentrated systems, the electrical interaction between the relaxing species will usually lead to a distribution of relaxation time, $p(\tau)$, and with the help of this distribution, the following relation is used [27,28].

$$\varepsilon = \varepsilon_\infty + (\varepsilon_s - \varepsilon_\infty)\int_0^\infty \frac{p(\tau)}{1 + j\omega\tau} d\tau - \frac{j\sigma_I}{\omega\varepsilon_0} \tag{18.9}$$

Gaussian, Cole-Cole, Fuoss-Kirkwood, and Davidson-Cole are some of the distribution function introduced in the literature [27]. The most useful distribution was first introduced by Cole and Cole [33], which leads to

$$\varepsilon = \varepsilon_\infty + \frac{(\varepsilon_s - \varepsilon_\infty)}{1 + (j\omega\tau)^{1-\alpha}} \tag{18.10}$$

In this expression, the ionic conductivity σ_I is ignored. From Eq. (18.10) the real and imaginary parts of complex permittivity are obtained from the following relations $(0 < \alpha < 1)$:

$$\varepsilon' - \varepsilon_\infty = \frac{(\varepsilon_s - \varepsilon_\infty)\left[1 + (\omega\tau)^{1-\alpha}\sin(\alpha\pi/2)\right]}{1 + 2(\omega\tau)^{1-\alpha}\sin(\alpha\pi/2) + (\omega\tau)^{2(1-\alpha)}} \tag{18.11}$$

$$\varepsilon'' = \frac{(\varepsilon_s - \varepsilon_\infty)(\omega\tau)^{1-\alpha}\cos(\alpha\pi/2)}{1 + 2(\omega\tau)^{1-\alpha}\sin(\alpha\pi/2) + (\omega\tau)^{2(1-\alpha)}} \tag{18.12}$$

Figure 18.1 compares a Cole-Cole plot for $\alpha = 0.1$ with the Debye plot (i.e., Cole-Cole plot with $\alpha = 0$). Cole-Cole plot for ε' and ε'' as a function of ω is illustrated in Fig. 18.1a. It is seen that the Cole-Cole distribution yields a broader spectrum. A plot of real versus imaginary parts of permittivity reveals that the locus is a semicircle in the case of the Debye equation and is an arc with a subtended angle of $(1 - \alpha)\pi$ in the case of the Cole-Cole model (Figure 18.1b).

Both Debye and Cole-Cole models are examples of physically realizable system. The complex permittivity of a physically realizable system follows the Kramers-Kronig relations [27–29].

$$\varepsilon'(f) - \varepsilon_\infty = \frac{2}{\pi}\int_0^\infty \frac{x\varepsilon''(x)}{x^2 - f^2} dx \tag{18.13}$$

$$\varepsilon''(f) = \frac{-2f}{\pi}\int_0^\infty \frac{\varepsilon'(f) - \varepsilon_\infty}{x^2 - f^2} dx \tag{18.14}$$

18.2.4. Relaxation Mechanisms in Biological Substances

Three different relaxation mechanisms are defined by Schwan and Foster [31]. These mechanisms are interfacial polarization (Maxwell-Wagner effect), dipolar orientation, and ionic diffusion.

Biological tissues are electrically heterogeneous and hence composed of different entities. In a heterogeneous medium, a charge accumulation exists between the interfaces. This charge accumulation is a consequence of the boundary conditions that the internal electric fields must satisfy at interfaces between different media. Consequently, a dielectric relaxation is observed in bulk properties. A useful example of this theory, known as Maxwell-Wagner theory, is a simple model for the cell suspension as analyzed by Schwan [30–32], where spheres covered by shells with different permittivity values are suspended in a third medium.

On the other hand, the partial orientation of permanent dipoles is responsible for dipolar relaxation. The time constant for dipolar relaxation ranges from microseconds for large globular proteins, to picoseconds for smaller polar molecules such as water. Consequently, the center frequency of the dispersion will be in the MHz to GHz region [31].

Dipolar relaxation effects are major contributors to the permittivity of tissues in the MHz to GHz region. In contrast to the tissues, water as a pure liquid exhibits a nearly single relaxation time with a characteristic frequency close to 20 GHz at room temperature and 25 GHz at 37°C [28].

The third major class of polarization mechanisms arises from ionic diffusion in the electrical double layers adjacent to charged surfaces [28,29]. This phenomenon is called the counterion effect. In contrast to the Maxwell-Wagner effect, which is a macroscopic phenomenon, this effect is a surface phenomenon. Counterion effect is responsible for the dispersion in the KHz frequency range of the solutions of biological particles and long chain macromolecules such as DNA, which show dielectric constants in the order of 10^4 below 1 KHz [28,29]. The reason for the observation of this dispersion is assumed to be the formation of a double layer of charge around a particle with surface charges.

Figure 18.2 illustrates a typical dielectric relaxation behavior exhibited by all tissues. Two significant features of the plot are highlighted by Schwan [32]. These features are the very large dielectric constant at low frequencies and three distinct relaxation regions at low, medium, and very high frequencies. These regions are called α, β, and γ, respectively [32]. Each of these relaxation regions is in its simplest form characterized by a Cole-Cole relation [cf. Eq. (18.10)].

The Maxwell-Wagner effect is responsible for β dispersion, around 50 KHz. Among the three dispersion regions (i.e., α, β, γ), the α dispersion (about 80 Hz) is the least understood one [31]. One possibility for α dispersion is the counterion effect. Frequency dependent impedance of intracellular structures, such as tubular apparatus in muscle cells, is assumed as the other possibility [31].

γ dispersion has a relaxation frequency near 20 GHz. The temperature dependence of the relaxation frequency for γ dispersion in tissues is equal to that of water and is about 2%/°C [31]. This dispersion is primarily due to the presence of water. A minor additional relaxation between β and γ dispersion was first observed by Schwan [33] (for Hemoglobin) and then reaffirmed by Grant et al. [34]. It is called δ dispersion and is observed in a broad frequency range from some 200 to 3000 MHz and is mainly due to proteins bound water [31,32]. The variability of the characteristic frequencies for the various mechanisms i.e., α, β, γ, and δ from one biological tissue to another is given at Table 18.1 [32].

It is believed that the tissue water sets the γ-dispersion relaxation frequency in a similar manner to the pure water [35]. Therefore, the dielectric property of pure water which is well established [36–38], from dc up to the infrared, becomes important behavior of these tissues.

Table 18.1 Range of Characteristic Frequencies Observed for Biological Materials for α, β, γ, and δ Dispersions [32]

Dispersion	Frequency range (Hz)
α	$1–10^4$
β	$10^4–10^8$
δ	$10^8–10^9$
γ	2×10^{10}

Source: Reprinted with permission from [32]. Copyright (2003) American Chemical Society.

18.3. TECHNIQUES FOR PERMITTIVITY MEASUREMENT

Since the 1940s, many techniques have been introduced for measuring the complex permittivity of materials at RF and microwave frequencies. Reviews of some of early techniques are available in book chapters by Westphal [39], Fox and Sucher [40], and a survey by Bussey [41]. State of the art techniques of the 1970s, especially for biological tissues and liquids, are illustrated in a book by Grant et al. [42]. A review of the coaxial line reflection method, the most widely used technique for biological materials today, is provided by Stuchly and Stuchly [43]. Afsar et al. [44] review a variety of methods available by 1986, for complex permittivity measurement of both lossy and low-loss materials in a broad range of frequencies from 1 MHz to 1500 GHz.

Recent advances in processing speeds and improvement in functionality and accuracy of test equipment, have been major factors for the development of automated complex permittivity measurement systems in recent years. Today, dielectric probe measurement systems, such as the Agilent 85070 family [45], exist that can measure the complex permittivity of materials conveniently and quickly and are compatible with a variety of network analyzers. Moreover, with the emergence of numerical techniques capable of solving Maxwell's equations in reasonable time for complex structures, we are witnessing the arrival of new methodologies that can increase the measurement accuracy, specifically at millimeter wave range where traditional methods face some limitations.

In what follows we review some of the technique identified in above mentioned references and elsewhere. We are particularly interested in techniques historically used in characterization of dielectrics from a few hundred MHz up to millimeter wave frequencies. We would also like to emphasize methods that are widely used for biological and lossy materials. In this regard, this section will address those works, which are extensively referenced or are unique in their nature.

18.3.1. Basic Methods

Westphal [39] provides the formulation of some transmission line techniques. These methods are the basis of measurement instrumentation later developed and used by many researchers [30–32,34–35,46–47]. Figure 18.3 illustrates some of these techniques, in which an air-filled coaxial line or a hollow waveguide, short terminated at the end, is partially filled with a disk shape sample (a–d) or is filled with a liquid (e, f).

Measurement of the standing pattern using a slotted line and a detector yield the complex permittivity of the sample. The process is as follows: If $\beta_a (= 2\pi/\lambda_{\mathrm{ga}}$, where λ_{ga}

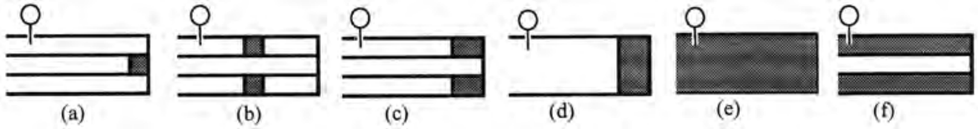


Figure 18.3 Various transmission-line techniques for complex permittivity measurement.

is the air-filled transmission line wavelength) and Z_a are the propagation constant and characteristic impedance of the air field transmission line, $S(=E_{\max}/E_{\min})$ is the standing wave ratio, and x_a is the location of the first minimum with respect to the sample surface, the input impedance for configurations (b), (c), and (d), at the sample surface, is given by [39]

$$Z_{\text{in}} = Z_a \frac{1/S - j\tan\beta_a x_a}{1 - j(1/S)\tan\beta_a x_a} \tag{18.15}$$

For (d) and (c),

$$Z_{\text{in}} = Z_s \tanh\gamma_s d \tag{18.16}$$

where index s refers to the sample region and d is the sample thickness. The characteristic impedance of the TEM (coaxial case) or dominant TE_{10} mode (waveguide case) for a nonmagnetic medium is

$$Z_i = \frac{j\omega\mu_0}{\gamma_i} \qquad (i = a,s) \tag*{(18,17)}$$

The unknown value of γ_s is obtained by solving the following equation

$$\frac{\tanh\gamma_s d}{\gamma_s d} = -\frac{jZ_{\text{in}}}{\beta_a d Z_a} \tag{18.18}$$

The desired equation for configuration (b), which is a preferred method for a thin sample by placing it at the region of high electric field and setting the distance between the sample and end wall as quarter wavelength, is

$$\frac{\coth\gamma_s d}{\gamma_s d} = -\frac{jZ_{\text{in}}}{\beta_a d Z_a} \tag{18.19}$$

The above equations are solved for γ_s. Ambiguity in the solution can be resolved if an approximate knowledge of its value is available, or two consecutive measurements for (b) and (c) are performed, and Eqs. (18.18) and (18.19) are multiplied.

Knowing the value of γ_s, the complex permittivity of sample is obtained using

$$\varepsilon' - j\varepsilon'' = \lambda_0^2\left[\left(\frac{1}{\lambda_c}\right)^2 - \left(\frac{\gamma_s}{2\pi}\right)^2\right] \tag{18.20}$$

where λ_c is the cutoff wavelength (infinity for coaxial line). For low-loss sample, the loss of waveguide cannot be ignored. Westphal provides correction for the waveguide loss [39].

A small sample at the end of inner conductor in configuration (a) is a capacitively terminated coaxial line. This structure resembles a reentrant structure where a capacitance $C = \varepsilon C_0$ at the end is introduced, where C_0 is the capacitance with no sample (i.e., the sample is air). From a knowledge of C_0, by neglecting the fringing field, the relations for obtaining the complex permittivity is straightforward:

$$\varepsilon' - j\varepsilon'' = \frac{1}{(j\omega\varepsilon_0 C_0 Z_{\mathrm{in}})} \tag{18.21}$$

Configurations (e) and (f) are ideal for liquid samples [42]. Direct measurement of γ_s can be made from the standing wave pattern quite easily if the liquid is low loss. For high-loss liquids, the standing wave pattern diminishes rapidly and a traveling wave technique using a microwave bridge is suggested. The bridge has two arms, one containing the liquid cell and one containing an attenuator and a phase shifter. By balancing the bridge, an estimate of $\gamma_s = \alpha_s + j\beta_s$ is obtained [42].

For liquids, Steel and Sheppard [46,47] and Grant et al. [38] change the sample thickness by a movable short. The amplitude of the signal as a function of sample thickness is recorded and is used for parameter extraction. In some cases, a combination of the above-mentioned techniques is more appropriate [37].

18.3.2. Resonant Cavity Measurement

Measurement of the complex permittivity using a resonant cavity coupled to the sample provides more accuracy than the previous methods. However, this method suffers from the fact that it is applicable only at resonance frequencies of the cavity in the lowest order mode (or a specified mode in the case of waveguide), which are distinctly away from each other or from other modes. The basic idea in using the resonant method is that the resonance frequency (f_0) and quality factor (Q) of the cavity will change as a result of coupling by the sample. The characteristic parameters of the sample, i.e., permittivity and permeability, can be measured by monitoring the changes in these two parameters.

Perturbation Technique

Perturbing the fields inside the cavity by a small sample has been a popular method for decades, where the perturbation theory [48] can be applied to find the electric or magnetic properties of the sample [49–55].

The basic formula for this method relates the change in the resonance frequency of the cavity to the field distribution inside, before and after placing the sample, which is [49,56]

$$\frac{\Delta\omega}{\omega_0} = \frac{\iiint_{V_1} [(E_1 D_0 - E_0 D_1) - (H_1 B_0 - H_0 B_1)]\, dv}{\iiint_V (E_0 D_0 - H_0 B_0)\, dv} \tag{18.22}$$

where, V_1 and V are the corresponding volumes of the sample and the cavity respectively, $\Delta\omega (= \omega - \omega_0)$ is the change in the resonance frequency of the cavity, and field indices

0 and 1 refer to the cavity fields before and after placing the sample. Assuming that the perturbation is small and an isotropic and homogeneous sample [56] is placed in an air filled cavity,

$$\frac{\Delta\omega}{\omega_0} = \frac{\iiint_{V_1} [\varepsilon_0(1-\varepsilon_r)E_0 \cdot E_1 - \mu_0(1-\mu_r)H_0 \cdot H_1]\,dv}{\iiint_{V} \left(\varepsilon_0 E_0^2 - \mu H_0^2\right)dv} \tag{18.23}$$

The quantity in the denominator is proportional to the stored energy in the cavity. The above relation can be further simplified if it is assumed that the fields outside the sample (volume V_1) are unchanged, and the field inside the sample can be related to the field outside by a quasistatic approximation. Some of these approximated relations between E and E_1 are given by Harrington [48] as shown in Fig. 18.4.

The dielectric samples are usually placed at the location of the maximum electric field intensity, where as magnetic samples are measured when inserted in a high magnetic field region. In any case, applying the perturbation theory by further simplification of Eq. (18.23) requires no appreciable field variation within the sample. For a nonmagnetic material, by satisfying this requirement, the numerator of Eq. (18.23) reduces to $\varepsilon_0(1-\varepsilon_r)E_0E_1V_1$, in which E_0 and E_1 are the fields at the location of sample placement. On the other hand, the denominator depends on the cavity shape and modal distribution and is four times the energy stored in the cavity.

In the case that sample is lossy, ε_r is a complex quantity and $\Delta\omega/\omega_0$ in the above relations is replaced by [48,49,51]

$$\frac{\Delta\omega}{\omega_0} = \frac{\omega-\omega_0}{\omega_0} + \frac{j}{2}\left(\frac{1}{Q} - \frac{1}{Q_0}\right) \tag{18.24}$$

where Q_0 and Q are the unloaded quality factors of the cavity before and after placing the sample. Still by replacing ε_r by ε' in Eqs. (18.22) and (18.23), these equations can be used for obtaining ε' from the change in real frequency, and ε'' is obtained by the following relation [48].

$$\frac{\varepsilon'-1}{\varepsilon''} = 2\left(\frac{Q_0Q}{Q_0-Q}\right)\left(\frac{\omega_0-\omega}{\omega_0}\right) \tag{18.25}$$

As an example, by placing a thin sample (Figure 18.4b) in the middle of a TE_{101} rectangular resonant cavity (Figure 18.4e), along the direction of the E field, complex

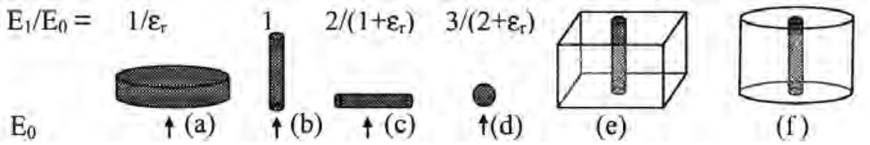


Figure 18.4 (a–d) Useful relations for applying perturbation theory. (e, f) Samples embedded in rectangular and circular resonators.

permittivity is obtained by [41,54]

$$\begin{aligned} \varepsilon' &= 1 + \frac{f_0 - f}{f}\frac{V}{2V_1} \\ \varepsilon'' &= \frac{Q_0 - Q}{Q_0 Q}\frac{V}{4V_1} \end{aligned} \tag{18.26}$$

A very popular technique is using a cylindrical cavity resonating at TM_{010} mode (Figure 18.5). In this case as used by Land and Campbell [55],

$$\varepsilon' = 1 + 2\frac{f_0 - f}{f}C \tag{18.27}$$

$$\varepsilon'' = \frac{Q_0 - Q}{Q_0 Q}C \tag{18.28}$$

where, for a sample of radius s, inside a dielectric field cavity with the radius a and relative permittivity ε'_c, C is given by

$$C = \varepsilon'_c\left(\frac{a}{s}\right)^2 J_1^2(ka) \tag{18.29}$$

In this relation k is the wavenumber in the medium filling the cavity and $J_1()$ is the first order Bessel function of the first kind. Land and Campbell [55] propose a cavity method, where the sample is placed in three sample container holes made inside a PTFE filled cylindrical resonant cavity (cf. Fig. 18.5). The cavity resonates at 3.2 GHz

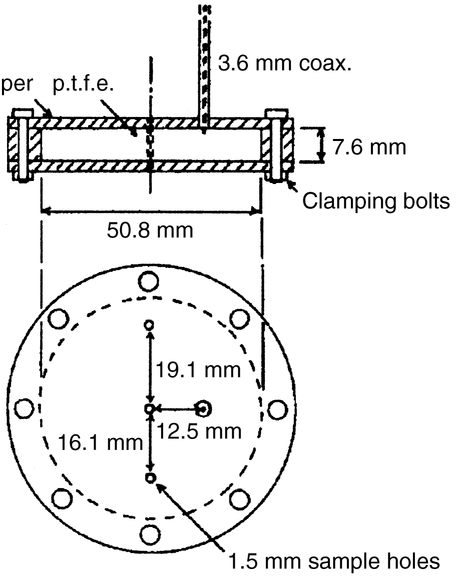


Figure 18.5 The cylindrical cavity used by Land and Campbell [55] for measurement of biological samples at 3.2 GHz (with permission from IOP Publishing Ltd.).

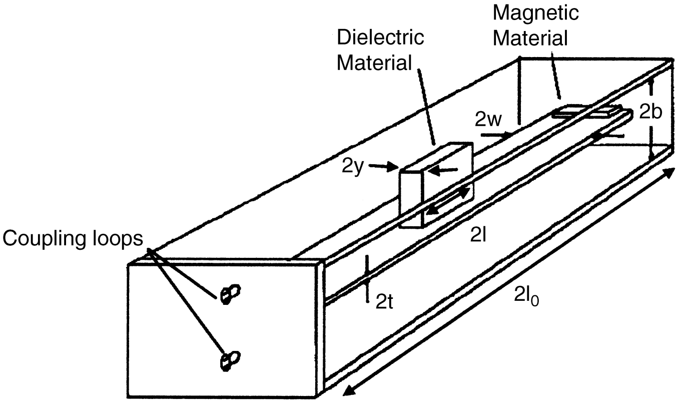


Figure 18.6 NIST strip-line cavity [53] (with permission from IEEE).

for the TM_{010} mode. It has a length of 7.6 mm and a diameter of 50.8 mm. The sample is so small that the perturbation technique can be used to find the complex permittivity from the measurement of Q and f_0. They use this technique to measure variety of liquids and tissues such as water, saline, fat, and breast. The accuracy is reported to be ±2.5 % for the dielectric constant and ±3.5% for the material loss factor.

Another practical system is the strip line cavity suggested by Waldron [49] and developed by NIST [52,53] (Fig. 18.6).

The formulas for a cavity with height $2b$, length $2l_0$, with a strip with thickness $2t$ and width $2w$, holding a sample with the height $b - t$, length $2l$, and width $2y$ are as follows [49,52].

$$\frac{\omega - \omega_0}{\omega} + \frac{j}{2}\left(\frac{1}{Q} - \frac{1}{Q_0}\right) = -A(\varepsilon - 1)\frac{2yl}{l_0} \tag{18.30}$$

with ε being the relative complex permittivity and

$$A = \frac{\pi\alpha}{2(\alpha + \beta)K(1/\alpha)}\sqrt{\frac{\alpha^2 - \beta^2}{\alpha - 1}} \tag{18.31}$$

α and β are parameters depending on the cavity dimensions (w, t, b) [49,52] and K is the complete elliptic integral of the first kind.

Sample-Terminated Coaxial Cavity

In practice, the resonance method can be used in all the cases shown in Fig. 18.3 by terminating the left side of transmission lines to a short wall as explained by Westphal [39]. A more recent approach is the one used by Tanabe and Joines [57] and Xu et al. [58], in which a coaxial line is terminated to a sample (cf. Fig. 18.7). Resonator is constructed by placing a capacitive gap at the input. C_r, L_r, and G_r are the equivalent components associated with the coaxial line length constituting the resonator. C_f is the fringing field inside due to the higher order mode, and C and G are the capacitance and conductance outside the interface due to the radiation and capacitive effects.

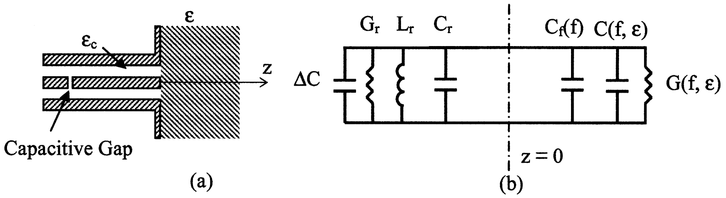


Figure 18.7 (a) Sample-terminated coaxial resonator and (b) its equivalent circuit [58] (© 2003 IEEE).

In the method discussed by Tanabe and Joines [57], the fringing and radiation effects are ignored and the admittance at the interface is considered to be a result of the open end static capacitor $C(\varepsilon')$ given by

$$C(\varepsilon') = \varepsilon'\varepsilon_0 h(\varepsilon') \tag{18.32}$$

h is a function of inner and outer radii of the coaxial lines ($a = 0.14364\,\text{cm}$ and $b = 0.47250\,\text{cm}$) and the dielectric of the coaxial line ($\varepsilon_c = 2.05$). It is obtained by static methods and is represented by an experimental relation as a function of ε'. For a time varying case, a conductance G can be recognized that is

$$G = \omega C(\varepsilon') \tan\delta \tag{18.33}$$

Note that

$$\frac{G}{C} = \omega \tan\delta \tag{18.34}$$

Tanabe and Joines [57] use the equivalent transmission line methods where the capacitive gap and the open end are treated as the extension of the coaxial cavity with the same characteristic impedance as the coaxial line. The overall cavity will have an effective resonance length of Le $(= nc/2f)$, where n is the mode order of TEM$_n$ mode and $c\,(= 2.998 \times 10^8\,\text{m/s})$ is the speed of light in free space.

The mathematical derivation of the end results are somewhat tedious and are not presented. The end formulas are

$$\varepsilon' = \frac{Y_0}{\omega\varepsilon_0 h(\varepsilon')} \frac{A_1(1 - A_2^2)}{1 + A_1^2 A_2^2} \tag{18.35}$$

$$\tan\delta = \frac{C_e}{CQ}\left(1 - \frac{Q f_0 C_{e0}}{Q_0 f C_e}\right) \tag{18.36}$$

where

$$A_1 = \tan\left[\frac{f}{f_0}\tan^{-1}\left(\frac{\omega_0 C_0}{Y_0}\right) + n\pi\frac{\Delta f}{f_0}\right] \tag{18.37}$$

$$A_2 = \tanh\left[\left(\frac{\sqrt{1+\tan^2\delta} - 1}{\sqrt{1+\tan^2\delta} + 1}\right)^{1/2} \tan^{-1} A_1\right] \tag{18.38}$$

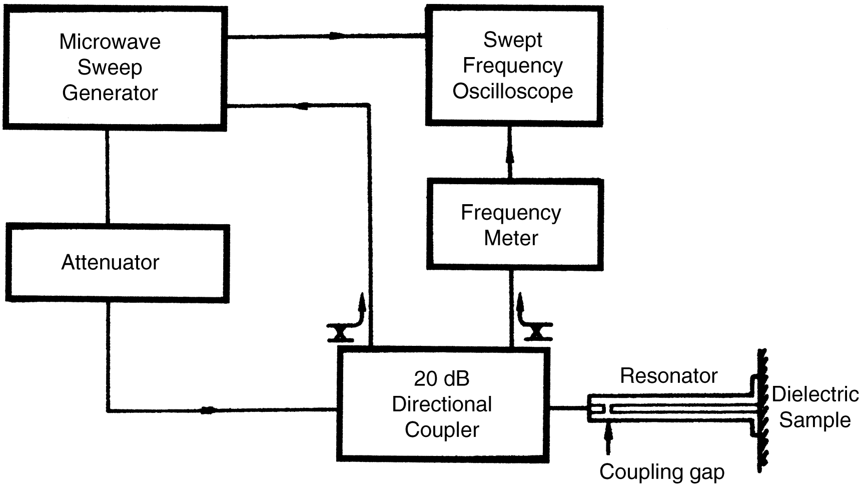

Microwave Sweep Generator
Swept Frequency Oscilloscope
Frequency Meter
Attenuator
20 dB Directional Coupler
Resonator
Dielectric Sample
Coupling gap

Figure 18.8 Schematic block diagram of the dielectric measurement technique reported by Tanabe and Joines [57] (with permission from IEEE).

$$C_e = \frac{1}{2}\left[C + \frac{Y_0 l}{v_1}\left(1 + \frac{\omega^2 C^2}{Y_0}\right)\right] \tag{18.39}$$

$$C_{e0} = \frac{1}{2}\left[C_0 + \frac{Y_0 l}{v_1}\left(1 + \frac{\omega_0^2 C_0^2}{Y_0}\right)\right] \tag{18.40}$$

In the above relations C_e and C_{e0} are the total equivalent capacitance of the resonator, $Y_0 (= 1/Z_0)$ is the characteristic admittance of the line, v_1 is the wave velocity in the coaxial line, $C_0 = C(\varepsilon' = 1)$, and l is the equivalent length of the coaxial cavity plus the gap capacitor. This latter quantity can be obtained from the air terminated measurement and C_0. Measurement of Q and f, and Q_0 and f_0 (cf. Fig. 18.8) provide enough information to calculate the complex permittivity through Eqs. (18.35) and (18.37) by successive iterations.

Tanabe and Joines [57] apply their technique for measuring the permittivity of water, methanol, and skin, as well as loss less dielectric such as polyethylene. They claim an accuracy of 5% for real part of permittivity and 25% on loss tangent within the frequency range of 1 to 4 GHz range.

The above method is further modified by Xu et al. [58]. They introduce the effect of fringing capacitance and the radiation at the sample side. They employ the method of Marcuvitz [59] for radiation admittance of the probe and end up with a series presentation for both G and C as functions of ε and f. The method is similar to the previous one and a pair of nonlinear equations for ε' and $\tan\delta$ is obtained. The coefficients of the capacitance series, truncated to the first two terms, and the fringing capacitance are obtained by resonance measurement of open, short, and a known dielectric-terminated cavity. The coefficient for the conductance series are explicitly given as a function of coaxial line parameters. They employ this method for a variety of tissues from 0.1 to 11 GHz.

Open Resonators

At frequencies above 30 GHz, the size and the quality factor of closed cavities decrease. Yet resonance methods are still applicable by employing open and Fabry-Perot resonators [60–65]. These structures can be used for measurements up to 300 GHz [61].

Developing a theory for his method, Jones [60] uses an open resonator that is constructed by a flat and a concave mirror and operating at 35 GHz. The sample sheet is placed on top of the flat mirror. In his procedure, the resonance for the empty cavity is obtained by moving the mirror using a micrometer. Then the quality factor of the empty cavity is measured by changing the frequency. After inserting the sample, the resonance for the same mode is restored by reducing the resonator length. Finally, *Q* of the loaded resonator is measured.

Jones' technique has been used by researchers over time by modifying it to be automated and applied with a fixed-frequency source (Afsar et al. [64–65]) or used in conjunction with a network analyzer [62].

Afsar and others [63–65] introduce an automated system based on a fixed frequency of 60 GHz and changing the cavity length (cf. Fig. 18.9).

The transcendental equation for the refractive index, obtained by applying a wave impedance matching condition of gaussian beam boundary condition to the fields at the air sample interface, is

$$\frac{1}{n}\tan(nkt - \Phi_{\mathrm{AM}}) = -\tan(kd - \Phi_{\mathrm{AC}}) \tag{18.41}$$

where n is the refractive index of the specimen, k is the free-space wave number. Furthermore, Φ_{AM} and Φ_{AC} are expressed as

$$\Phi_{\mathrm{AM}} = \tan^{-1}\frac{t}{nZ_0} - \tan^{-1}\frac{1}{nkR_1(t)} \tag{18.42}$$

$$\Phi_{\mathrm{AC}} = \tan^{-1}\frac{d'}{Z_0} - \tan^{-1}\frac{1}{kR} - \tan^{-1}\frac{t}{nZ_0} + \tan^{-1}\frac{1}{kR_2(t)} \tag{18.43}$$

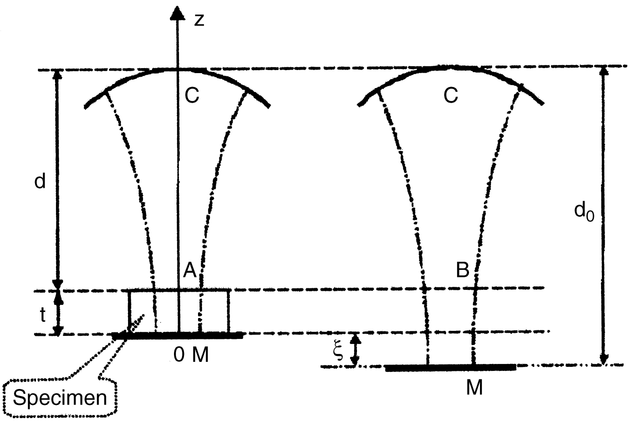


Figure 18.9 Diagram of a hemispherical open resonator with and without the specimen (from Afsar et al. [65], with permission from IEEE).

R is the radius of the curvature of the mirror and also

$$R_1(t) = t + \frac{n^2 Z_0^2}{t} \qquad R_2(t) = \frac{R_1(t)}{n} \tag{18.44}$$

Finally, $Z_0 = \sqrt{d'(R - d')}$, with $d' = d + t/n$. For an empty cavity ($t = 0$, $n = 1$), the above relations reduce to

$$kd_0 = q\pi + \tan^{-1}\sqrt{\frac{d_0}{R - d_0}} - \tan^{-1}\frac{1}{kR} \tag{18.45}$$

where q is the mode number of $\mathrm{TEM}_{0,0,q}$ mode that yields d_0, the resonance length of the empty cavity. A correction for mismatching between the wave front of the gaussian beam and the surface of the spherical mirror, as well as the upper surface of the sample, is

$$d = d_0 - t - \xi + \frac{t(n - \Delta)}{n^2 k^2 w_t^2} + \frac{3}{4k^2 R} \tag{18.46}$$

$$w_t^2 = \frac{2Z_0(1 + t^2/n^2 Z_0^2)}{k} \tag{18.47}$$

$$\Delta = \frac{n^2}{n^2 \cos^2(nkt - \Phi_{\mathrm{AM}}) + \sin^2(nkt - \Phi_{\mathrm{AM}})} \tag{18.48}$$

where ξ is the shift length to restore the resonance with and without specimen. Equations (18.41) and (18.46) can be solved to find n and d. The loss tangent is found through the following relation:

$$\tan\delta = \frac{1}{Q_\varepsilon} \frac{2nk(d + t\Delta)}{2nkt\Delta - \Delta \sin 2(nkt - \Phi_{\mathrm{AM}})} \tag{18.49}$$

where

$$\frac{1}{Q_\varepsilon} = \frac{1}{Q_L} - \frac{1}{Q'_L} \qquad \text{and} \qquad Q'_L = Q_0 \frac{2(t\Delta + d)}{d_0(\Delta + 1)} \tag{18.50}$$

in which Q_L and Q_0 are the quality factor of the cavity with and without the sample.

Using this technique, Afsar et al. [64,65] have measured the dielectric properties of various dielectrics at 60 GHz. For instance, $\varepsilon' = 2.063 \pm 0.004$ and $\tan\delta = 0.00029 \pm 0.00003$ is calculated for Teflon.

18.3.3. Open-Ended Transmission Line

The advent of accurate Automatic Network Analyzers (ANA) and new calibration techniques has revolutionized the measurement of microwave device and networks, since the late 1970s. In the last two decades, the extraction of complex permittivity from the

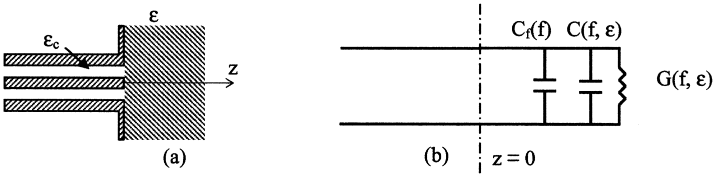


Figure 18.10 A coaxial probe terminated to a lossy medium with complex permittivity ε: (a) the schematic of the probe and (b) the equivalent circuit model.

measured reflection coefficient (input admittance) of a coaxial line terminated to a specimen (cf. Fig. 18.10) has been the most widely used method of the tissue permittivity extraction [43,45,66–89].

Stuchly and Stuchly [43] review a variety of coaxial line terminated structure including those already shown in Fig. 18.3. The basic idea involves the reflection coefficient (S_{11}) measurement at the input of the coaxial line. First, the values for the equivalent circuit models (cf. Fig. 18.10) are analytically or numerically calculated as a function of frequency and permittivity. Then, the unknown real and imaginary parts of the complex permittivity are obtained. A variety of approaches exist in the literature to formulate the terminating admittance of the coaxial line. Some of them are briefly reviewed here.

Lumped-Capacitance Method

This approach yields the same complex permittivity extraction relations for structures in Figs. 18.3a (i.e., reentrant coaxial line) and 18.10 (i.e., open-ended coaxial line) [43,66,67]. The former has the advantage of requiring a small sample and the latter is useful for a sufficiently big sample having a flat surface to contact with the probe. The fringing capacitance C_f is due to higher order, nonpropagating coaxial modes on the left side of the coaxial–sample interface and is assumed not to change by changing the sample permittivity. G and C are the conductance and capacitance on the right side of the interface. The lumped-element approach assumes that these two quantities are directly proportional to their corresponding complex permittivity components (real or imaginary) as if they are created as a result of filling the static fringing capacitance in air (C_0), in the right side of the interface, with a medium with the permittivity of ε.

$$C = \varepsilon' \varepsilon_0 C_0 \qquad G = \varepsilon'' \varepsilon_0 \omega C_0 \tag{18.51}$$

Note that their relation is the same as Eq. (18.34). If $S_{11} = |S_{11}| \exp j\phi$ is obtained by a measurement, ε' and ε'' can be obtained by

$$\varepsilon' = \frac{2|S_{11}|\sin(-\phi)}{\omega Z_0 C_0 \left(1 + 2|S_{11}|\cos\phi + |S_{11}|^2\right)} - \frac{C_f}{C_0} \tag{18.52}$$

$$\varepsilon'' = \frac{1 - |S_{11}|^2}{\omega Z_0 C_0 \left(1 + 2|S_{11}|\cos\phi + |S_{11}|^2\right)} \tag{18.53}$$

Note that C_f and C_0 should be known separately by analytical or numerical methods or from the measurements of two known samples. The latter is usually a more practical method. However, a numerical solution for these quantities using the method of moment (MOM) [90] or finite element technique (FEM) [91] has also been suggested.

Note that this method neglects the dependence of C_f, C, and G on frequency and ignores the fact that aperture field distribution at the interface creates radiation on the outside. Radiation effect and the presence of significant higher mode effects will reduce the merit of this method at frequencies higher than a few GHz.

Stationary Solution of Aperture Admittance

To remedy some of the shortcomings of the previous method, a stationary solution of the aperture admittance as found in Marcuvitz [59] is suggested [58,78–80]. In this approach the admittance of the probe is represented by a stationary integral. The integral is solved by a series expansion technique and the final solutions are in the form

$$G = G_1\varepsilon^{5/2}f^4 + G_2\varepsilon^{7/2}f^6 + B_5\varepsilon^{9/2}f^8 + \cdots \tag{18.54}$$

$$B = B_1\varepsilon f + B_3\varepsilon^2 f^3 + B_5\varepsilon^3 f^5 + \cdots \tag{18.55}$$

where G_i and B_i are constants related to the coaxial line parameters. A good approximation is to consider only the first term of Eq. (18.54) and the first two terms of Eq. (18.55) [79]. Note that in this formulation, the impact of higher order modes inside the waveguide on the aperture admittance is ignored [78–80].

Full Wave Modeling

The full wave solution is possible by using a well-known numerical technique such as mode matching and MOM. Mosig et al. [73] present a formulation for the aperture using this method. They use a modal expansion (only TEM and $\mathrm{TM_{0n}}$ modes) of the field inside the coaxial probe. They even provided some charts for this purpose at different frequencies. The charts are for a SR7 coaxial cable (i.e., $a = 1.05\,\mathrm{mm}$, $b = 3.675\,\mathrm{mm}$, $\varepsilon_r = 2.3$) and at frequencies 1, 3, and 10 GHz. However, due to the computational cost and the limited computational power at the time, their method did not seem to be a practical one.

Stuchly et al. [82] also use MOM to find the aperture admittance for a wide range of inner conductor diameter and permittivity of a 50-Ω coaxial cable. Recognizing the need for introducing such data in a suitable form to be used for complex permittivity extraction, they fit the numerical results to a rational function that is

$$Y(j\omega a,\, \varepsilon) = \frac{\sum_{n=1}^{N}\sum_{p=1}^{P}\alpha_{np}\zeta^p(j\omega a)^n}{1 + \sum_{m=1}^{M}\sum_{q=0}^{Q}\beta_{mq}\zeta^q(j\omega a)^m} \tag{18.56}$$

where $\zeta = \sqrt{\varepsilon}$ and α_{np} and β_{mq} are parameters of the function. For fitting they employ 56 dielectric constants in the range of $1 \le \varepsilon' \le 80$ and 20 normalized frequency in the range of $0.01 \le ka \le 0.19$, resulting in a total of 1120 data points. For fitting the simulated data to the above function they use a modified Levenberg-Marquardt algorithm [92] to minimize the total sum of absolute error between the admittance from Eq. (18.56) and admittance obtained by MOM at the 1120 points. This procedure yields fitting parameters α_{np} and β_{mq}. Truncation of the series is made by $M = N = 4$ and $P = Q = 8$. Therefore 68 parameters are given to yield a function that can represent the aperture admittance explicitly. If measured admittance is available, Eq. (18.56) can be solved for the complex permittivity of the medium.

Virtual Line Method

This method models the sample medium by a virtual transmission line with length L filled with the unknown dielectric of the specimen. If a coaxial line of length D is terminated to the sample, the following relation is valid [87]:

$$\varepsilon = \frac{-jc\sqrt{\varepsilon_c}}{2\pi fL}\,\frac{1-\Gamma_m e^{2jkD}}{1+\Gamma_m e^{2jkD}}\cot\frac{2\pi fL\sqrt{\varepsilon}}{c} \tag{18.57}$$

where ε_c is the dielectric constant of the probe, k is the wave number of the probe, c is the speed of light in free space, and Γ_m is the reflection coefficient measured at the input of the probe. In this technique D and L are measured from measurement of two known media, i.e., air and distilled water. Knowing these quantities, complex permittivity of an unknown medium can be evaluated by measuring the reflection coefficient through Eq. (18.57).

18.3.4. Free-Space (Quasioptical) Measurement Techniques

This method is based on placing a flat sample of thickness d between transmitting and reflecting horn antennas. This method is particularly suitable for W-band (70–120 GHz) as reported by Friedsam and Biebl [93] and Afsar et al. [94], at which quasioptical method based on a gaussian beam can be applied. However, it may also be applied at lower frequencies as has been the case for a system reported by Ghodgaonkar et al. [95] at 8.6 to 13.4 GHz.

Figure 18.11 illustrates the set up for these three methods. In Figure 18.11a, from simple plane wave theory, knowing the transmission and reflection coefficient (S_{11} and S_{21}), we obtain the complex permittivity and permeability as [95]

$$\varepsilon = \frac{\gamma}{\gamma_0}\left(\frac{1-\Gamma}{1+\Gamma}\right) \tag{18.58}$$

$$\mu = \frac{\gamma}{\gamma_0}\left(\frac{1+\Gamma}{1-\Gamma}\right) \tag{18.59}$$

In the above relations, $\gamma_0\,(=j2\pi/\lambda_0)$ is the propagation constant of free space and

$$\gamma = \frac{\ln(1/T)}{d} \tag{18.60}$$

$$\Gamma = K \pm \sqrt{K^2-1} \tag{18.61}$$

where

$$K = \frac{S_{11}^2 - S_{21}^2 + 1}{2S_{11}} \tag{18.62}$$

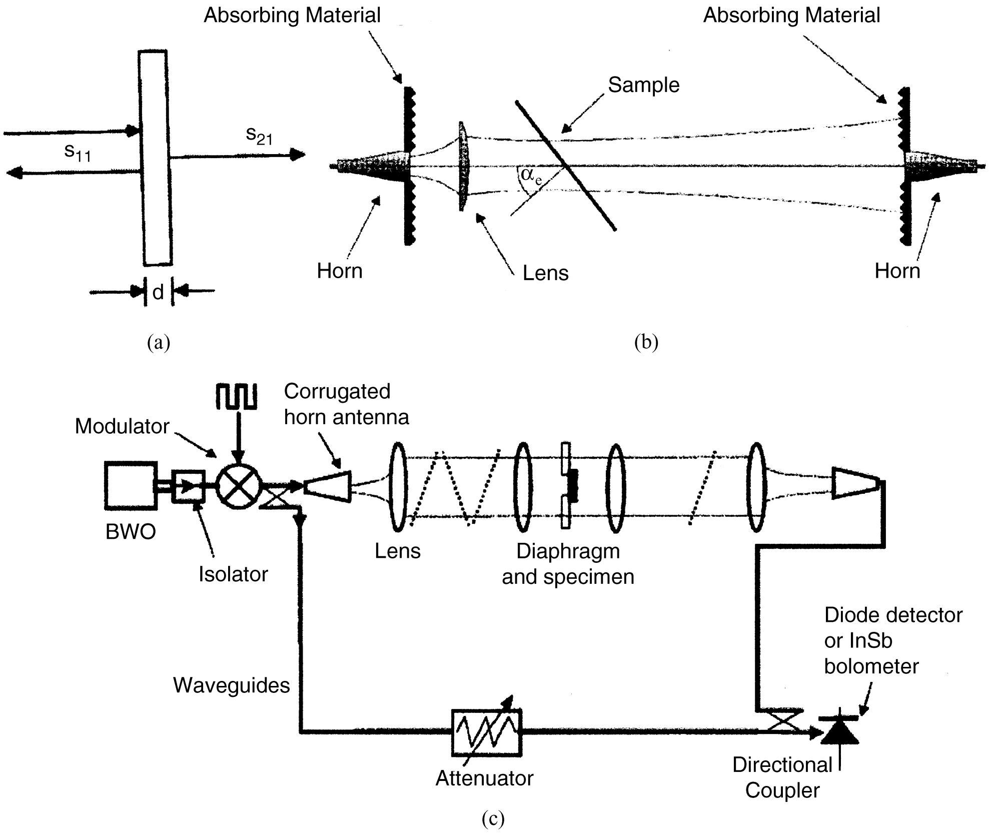


Figure 18.11 (a) Reflection and transmission of a dielectric sample, (b) Friedsam and Biebl [93] setup, and (c) Afsar et al. method [94] (with permission from IEEE).

$$T = \frac{S_{11} + S_{21} - \Gamma}{1 - (S_{11} + S_{21})\Gamma} \tag{18.63}$$

In their method they use a TRL calibration method [96–98] for the accurate measurement of S_{11} and S_{21} using a HP8510B network analyzer.

If the material is nonmagnetic, the transmission measurement is adequate. In the Friedsam and Biebl [93] set up (Fig. 18.11b), for 75 to 95 GHz, a sheet of sample is located between corrugated horn antennas. By using a collimating lens before the sample, a quasioptical gaussian beam illuminates the sample. Calibration procedure involved performing a reference measurement. This is performed by normalizing the transmission coefficient with sample in place to the case without sample. The mean value of the complex permittivity is obtained for various angle α_e and a nonlinear least square method to minimize an error function of difference between the theoretical and measured transmission coefficients. The reported uncertainties are 0.1 % for dielectric constant and 2×10^{-4} for loss tangent.

In Afsar et al. method [94] (cf. Fig. 18.11c), an unbalanced bridge, with a waveguide as the reference arm, is used. The refractive and absorption indices are obtained from the distance between the maxima and the amplitude of transmittance spectra.

18.3.5. Two-Port Network Analyzer Measurement/Extraction

One-port measurement and extraction of dielectric properties of materials using a coaxial probe is generally adopted for measurement of biological tissues up to 20 GHz. There are some advantages offered by using two-port calibration and measurements. A powerful calibration technique known as TRL [96–98] is available for modern network analyzers. The TRL method potentially provides much better accuracy than the traditional calibration methods (which use imperfect match and open standards), specially at higher frequencies, and is inherently two port. In addition, a two-port measurement [99–102] gives more information (four S parameters) than a one-port measurement with only the reflection coefficient being measured.

Belhadj-Tahar et al. [99] presented a technique for the simultaneous measurement of the complex permittivity and permeability of a given material using ANA. A gap built in a coaxial line was filled with the material under test. Complex permittivity and permeability were computed from the two-port S parameters for several materials from 45 MHz to 18 GHz. They employed a mode matching method for the numerical solution of the structure, and the gradient method for the inverse problem solution. HP8510A ANA and APC7 coaxial standards were used. For avoiding the contact resistance and capacitance, the sample was metalized on the surfaces that were in contact with the coaxial connectors.

Measurement of teflon and alumina yielded accurate results for the dielectric constant; however, they concluded that an accurate measurement is not possible for the imaginary part of permittivity in the range of less than 0.1. They also commented that by using a 2.4-mm connector the method could be extended to 50 GHz.

Abdulnour et al. [100] developed a generic approach for permittivity measurement in microwave and millimeter wave frequencies. They first determined the scattering parameters of a discontinuity containing a material having a wide range of complex permittivity that was known a priori. The discontinuity was a tube containing a sample under test (i.e., a cylindrical shape dielectric). For the direct problem, they used a boundary integral equation method combined with a modal expansion approach. They developed some simple generic formulas from graphs for constant ε' and constant ε'' on S_{21} plane, which directly provide ε' and ε'' from measured S_{21}. An accuracy of better than 1% was claimed.

To cover a broad measurement range, they used a microstrip structure for 1 to 7 GHz (covering L-, S-, C-bands), WR90, WR62, and WR42 waveguides for 8 to 26 GHz (covering X-, Ku-, and K-bands). The frequency bandwidth was between 1.2 to 1.8 times the cutoff frequency of the dominant TE_{10} mode of the rectangular waveguide and up to half of the cutoff frequency for the microstrip structure. They verified the complex permittivity of teflon, plexiglas, and some polymers. It should be mentioned that they used the HP8510B along with the TRL calibration method.

Queffelec and Gelin [101] also applied the microstrip line and two-port measurement by placing a dielectric sample with specified dimensions over a microstrip line, where the sample covers the whole width of the enclosure (cf. Fig. 18.12a). A 25 mils thick alumina substrate and an enclosure width of 25.4 mm were considered. They analyzed the structure based on the mode matching technique, over the frequency range of 45 MHz to 14 GHz. TRL method was used for calibration. The accuracy of the method was claimed to be better than 5% over this frequency range and for the dielectric constant less than 10. However, the measurement accuracy deteriorated for higher permittivity values.

It is worth mentioning that none of the above mentioned techniques was developed for the measurement of biological samples. A very simple technique is introduced by

Vander Vorst and colleagues [103,104] for the measurement of blood, dioxane, and methanol up to 110 GHz. This method applies an HP network analyzer system with waveguide standards. They introduced LL calibration that is a reduced form of TRL calibration, and is designed for evaluating complex propagation constant γ. In their waveguide measurement system, a liquid sample holder is used, where its cross section exactly matches with the reference air-filled waveguide cross section.

A newer technique is the one introduced by Tofighi and Daryoush [25,105–109], based on a two-port microstrip test fixture. Fig. 18.13 depicts this two-port test fixture.

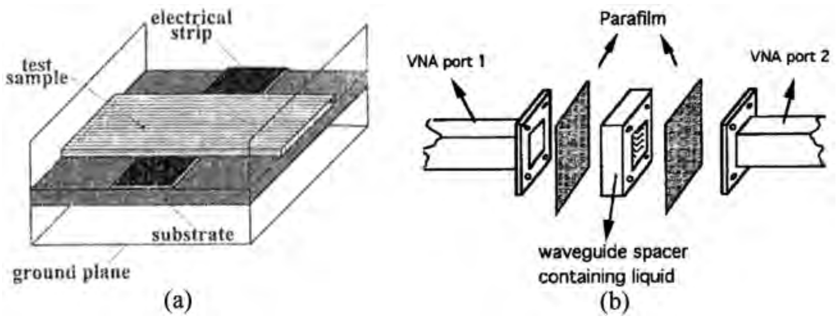


Figure 18.12 (a) Microstrip test structure used by Queffelec and Gelin [101], and (b) waveguide technique introduced by Vander Vorst and colleagues [103] (with permission from IEEE).

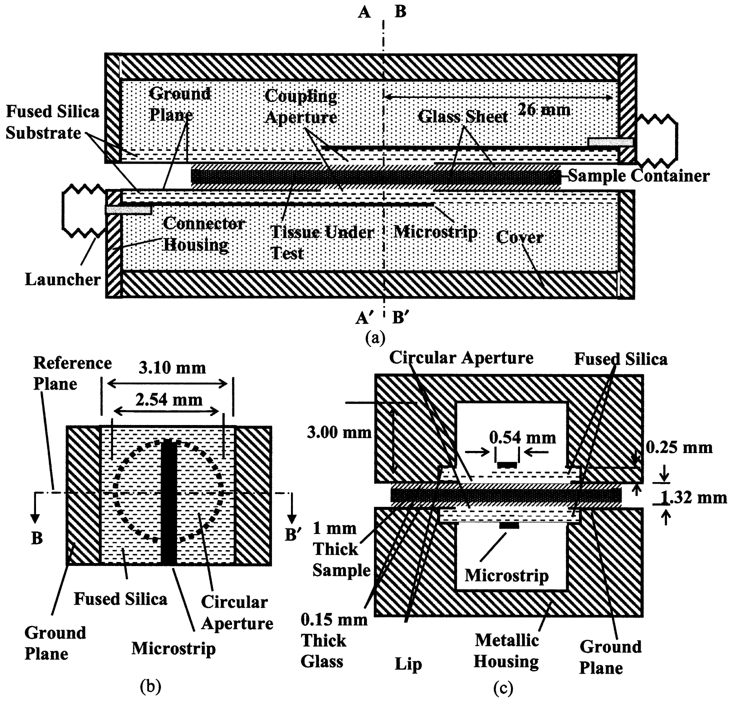


Figure 18.13 A two-port microstrip test fixture for complex permittivity measurement of biological materials; schematic of (a) side, (b) top, and (c) front views, where the two microstrip lines are coupled through two apertures, glass sheets, and TUT [108] (with permission from IEEE).

Open-circuited microstrip transmission lines are coupled to tissue under test (TUT) through two small apertures. The sample is sandwiched between glass plates and then inserted between the microstrip ground planes. Planes 1–1′ and 2–2′ are the reference planes where the effects of embedding networks are removed. Microstrip lines and apertures are etched over the two sides of a fused silica substrate ($\varepsilon_r = 4.1$). Two coaxial to microstrip line launchers provide transitions from the network analyzer test set cables to the fixture.

A 100 mil (i.e., 2.54 mm) diameter circular aperture is considered (cf. Fig. 18.13b), as a compromise between the coupling factor and the spatial resolution. To have a realizable microstrip line test fixture with the aperture on the ground plane facing the sample, the width of the fused silica substrate should be slightly more than the width of air-filled space above. The substrate sits inside lips provided in the enclosure. Therefore, the microstrip cross section is not a complete rectangle (cf. Fig. 18.13c). This technique has been used for the extraction of the complex permittivity of brain grey and white matters, neurological cell solutions, and dielectric imaging of brain slices up to 50 GHz and will be explained in detail in later sections.

18.3.6. Characteristic Impedance Determination Method

If the propagation constant of a transmission line with the specimen as part of its structure is known, then the complex permittivity can be obtained [105,110,111]. The requirements for the applicability of this method are that the specimen should be uniform across its length and the transmission line loaded with it operates only at its dominant mode.

Although it is conceptually a simple method, its general theory was formalized quite recently. The formulation for this technique is similar to the propagation constant determination as part of TRL [96–98,105] or multiline [112] calibrations and is omitted here. The propagation constant, γ, can be found from at least two successive two-port measurements of the transmission line with samples having two different lengths (cf. Fig. 18.14). The length difference of a quarter wavelength (wavelength of the transmission line loaded with the sample) at the middle of frequency band yields the most reliable results.

This method is used by Janezic and Jargon [110] for the measurement of polystyrene from 8 to 12 GHz using a coaxial line, by Wan et al. [111] for PVC in a rectangular waveguide at 8 to 18 GHz, and, by Tofighi [105] for fused silica as the substrate of a microstrip line of Fig. 18.12 from 5 to 50 GHz.

18.3.7. Other Methods

Methods for the measurement of the permittivity of materials are not restricted to those explained in the previous sections. Time-domain techniques are also alternatives to the

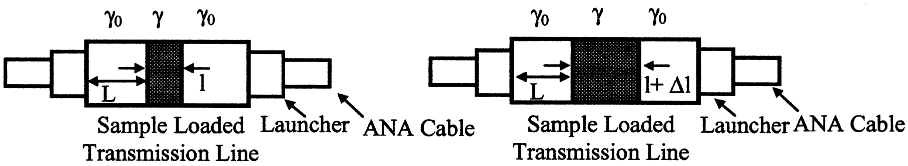


Figure 18.14 Conceptual diagram of method for propagation constant measurement with the transmission lines are loaded with two uniform specimen with difference in length of Δl.

frequency-domain methods [113–115]. One method, particularly popular in 1960s is described by Nicolson and Ross [113] and involves measurements in the time domain. They insert a sample with a certain width in an air-filled coaxial cable. The sample forms the unknown dielectric of a coaxial line. The coaxial line is excited by a pulse with subnanosecond rise time (e.g., ≤ 100 ps). Time-domain reflected and transmitted waveforms are taken by using a computer controlled sampling oscilloscope. The Fourier transforms of these signals provide necessary information to obtain the frequency dependent S_{11} and S_{21} of the dielectric field coaxial line section. The complex permittivity and permeability of the material can be retrieved using transmission line equations. The method was able to measure the material parameters from 0.4 to 10 GHz with a spectral resolution of 400 MHz.

One of the well-known techniques for dielectric measurement at millimeter and submillimeter range is dispersive fourier transform spectroscopy (DFTS) first developed by Afsar [116]. This method has been claimed to be very accurate for wavelengths from 3 to 0.25 mm (100 to 1200 GHz) and is extended at best up to 5 mm long wavelength (60 GHz). This method includes a two-beam interferometer with the sample located in one of the active arms (mirror arm) of the interferometer.

The recent advances in the field of picosecond optoelectronics pulses have made them interesting tools for being used in a reliable measurement technique in microwave and millimeter wave frequencies. Arjavalingam et al. [117] use a technique called coherent microwave transient spectroscopy (COMITS) to measure the permittivity from 10 to 125 GHz. No microwave source or detector is used in their experiment setup. The setup consists of a transmitting broadband antenna and an identical receiving antenna. The antennas are made from coplanar strip lines, exponentially tapered at one end (e.g., Vivaldi antenna).

18.4. CHARACTERIZATION OF BIOLOGICAL MATERIALS

Certain technical issues for complex permittivity extraction are highlighted in this section. Addressing these issues for one-port as well as two-port measurement techniques, this section provides the extracted results of the complex permittivity of brain tissues up to 50 GHz, obtained by using the two-port approach.

18.4.1. Open-Ended Coaxial Probe

As explained before, biological materials experience a large amount of dispersion at microwave and millimeter wave frequencies due to the water like dipolar absorption around 20 GHz. To appropriately characterize them, complex permittivity in a broad frequency range has to be obtained and fit to the Cole-Cole representation. Fortunately, such dispersion, manifested by power loss, has propelled the medical and industrial applications of microwave heating. Furthermore, as the signal attenuates into the medium, a finite *sensing volume* exists around the region of the source contact to the medium, which gives the justification to using measurement techniques such as open-ended coaxial line on finite samples.

As stated before, a lot of studies have been reported that provide various techniques for permittivity measurement at microwave frequencies, and attempts have been made to extend this knowledge beyond 20 GHz. As a result of these efforts the complex permittivity of tissues and liquids are generally well-known below 20 GHz [118–130].

Regarding to both measurement and characterization, there are several technical issues that are covered in this section. It is almost impossible to cover these issues for all the measurement techniques mentioned in the previous section. However, in what follows, we try to address issues related to new one-port and two-port approaches that employ network analyzers for reflection and/or transmission measurements.

One-Port Measurement System

One of the well-known studies with identifying practical details has been reported by Burdette et al. [71,72]. They were able to perform in vivo measurement, obtain continuous data from 0.1 GHz to 11 GHz, and process data in real time. They used probes with 0.085 in (2.16 mm) diameter semirigid coaxial cable as shown in Fig. 18.15.

The set up of their measurement system was based on an HP 8410B network analyzer. Short circuit, open circuit, and matched loads were employed for network analyzer calibration. Data collection was accomplished by using a semiautomated data acquisition and processing system, whose key component was an A/D converter. The system was utilized for the determination of the in vitro and in vivo dielectric properties of various material, which included saline, distilled water, methanol, ethylene glycol, canine and rat muscle, canine kidney, canine fat, rat brain, and rat blood. The standard error of mean of their results for these measurements was at most ± 3.25 for real part and ± 2.25 for imaginary part of ε. However, they found it hard to comment on the absolute accuracy of their measurement since the variability of data from the reference literature was greater.

Burdette's work is one of the earliest ones in terms of applying the automatic network analyzer for measurements. It also stands out because they did in vivo measurement and identified the sources of inaccuracy. Those sources are tissue dehydration, accumulation of dried tissue at the probe tip, variation of probe contact pressure, improper probe positioning, temperature change, and tissue inhomogeneity. However, the method used was based on the lumped element model.

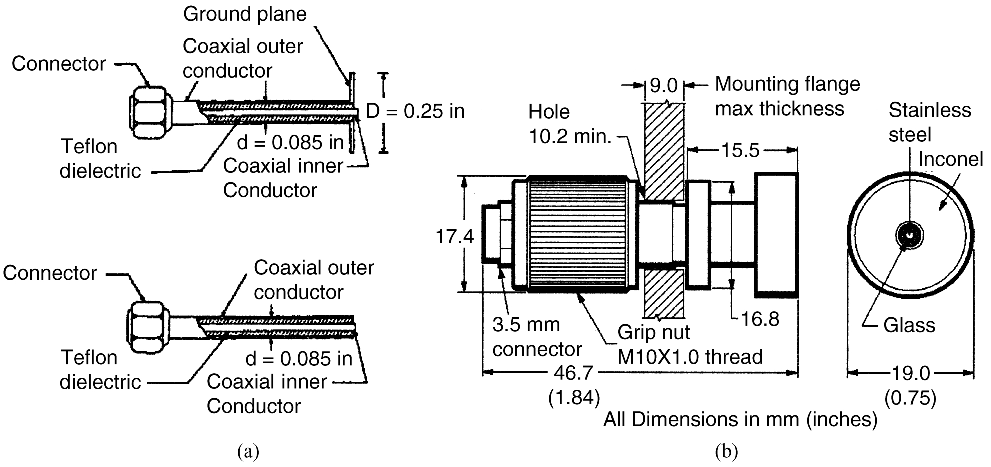


Figure 18.15 Schematics of (a) a coaxial probe used by Burdette et al. to measure the complex permittivity of biological materials [71] (with permission from IEEE), and (b) Agilent 85070D probe [45] (© 2002 Agilent Technologies, Inc. Reproduced with permission, Courtesy of Agilent Technologies, Inc.).

Athey and Stuchly [74] used a similar system up to 1 GHz and introduced the uncertainty analysis of the resulting complex permittivity. The reported uncertainties were due to the error in the probe termination capacitance, line characteristic impedance, and measured reflection coefficient. They reported the measurement uncertainty for distilled water, NaCl solutions, and low and high water content tissues.

As another example, Fig. 18.15b illustrates an Agilent 85070D probe [45]. The specified operating ranges are −40 to 200°C for temperature, 200 MHz to 20 GHz for frequency, which requires greater than 20 mm for the sample diameter and $20/\sqrt{\varepsilon}$ for its thickness. The accuracy is claimed to be 5% for ε' and ± 0.05 for $\tan\delta$. The typical repeatability is specified as four time better than the accuracy. The recommended range of materials to be measured is $\varepsilon' < 100$ and $\tan\delta > 0.05$. Material should make a flat and air-gap free contact with the probe.

One-Port Calibration

An imperfect reflectometer can be modeled by taking all the linear errors of the system and combining them into a two-port error adapter between the reflectometer and the unknown one-port (Fig. 18.16) [97]. This linear error adaptor is associated with the network analyzer internal test set and transfer switches as well as the external cables and adaptors embedded up to measurement reference plane. In the case of noncoaxial media such as microstrip or waveguide, this adaptor includes coaxial to microstrip or waveguide launchers and the length of microstrip or waveguide from the launcher to the measurement reference plane.

The relationship between the actual and measured reflection coefficients (Γ_L and Γ_M) can easily be obtained using flowchart reduction techniques [97].

$$\Gamma_L = \frac{\Gamma_M - e_{00}}{e_{11}(\Gamma_M - e_{00}) + e_{01}e_{10}} \tag{18.64}$$

The error parameters are obtained from the measurement with three calibration standards usually open, short, and match termination [71,72,132,133]. Note that e_{01} and e_{10} cannot be separated that does not matter in practice. A direct formulation for an unknown load, knowing Y_i and Γ_i $(i = 1, 2, 3)$, the aperture admittance and reflection coefficient of the standard, is [80]

$$\frac{(Y_L - Y_1)(Y_2 - Y_3)}{(Y_L - Y_2)(Y_3 - Y_1)} = \frac{(\Gamma_M - \Gamma_1)(\Gamma_2 - \Gamma_3)}{(\Gamma_M - \Gamma_2)(\Gamma_3 - \Gamma_1)} \tag{18.65}$$

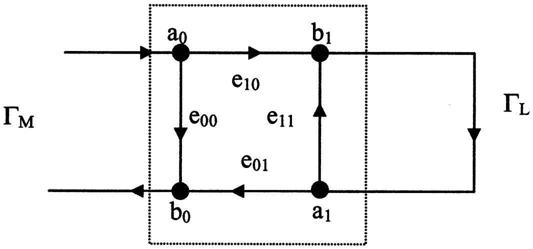


Figure 18.16 One-port signal flowchart.

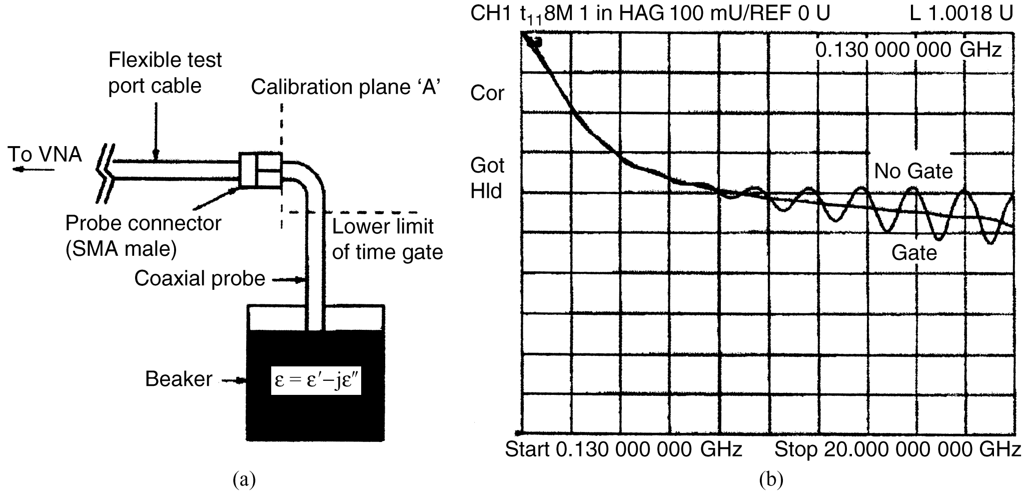


Figure 18.17 (a) Open-ended coaxial line probe system used by Anderson et al. [83], and (b) improvement by gating out the effect of connector in time domain (with permission from IEEE).

There is no limitation for the choice of standards and in fact the use of known (reference) liquids for standards is highly recommended [80,81]. A thorough study of this concept can be found in the report by Misra et al. [80]. They demonstrate that using liquids such as water, which have similar electromagnetic properties as biological materials, significantly reduces the error in measurement that otherwise will be high.

It is worth mentioning that the uncertainty in the Cole-Cole parameter of reference liquid might significantly contribute to the uncertainty in the permittivity of tested sample. This issue has been studied by Nyshadham et al. [81] up to 18 GHz. For instance, for a 3.6-mm semirigid probe, and by employing short, open, and methanol as the reference, 10% and −4% uncertainties are observed for real and imaginary part of permittivity of saline at 18 GHz. They observed that generally the resulting uncertainty in the permittivity of the test material is smaller than the uncertainty in the reference liquid itself.

Besides reference calibration, ANA time gating feature [80,83], can also be helpful, in some applications, to remove the ripple artifact due to embedded adaptors between the ANA and the probe. Anderson et al. [83] employ this technique (cf. Fig. 18.17) in their study. They use this method in the absence of reference liquid calibration. In this method, the probe should be sufficiently long (> 7 cm) to allow separation of the desired time domain components from the connector reflection.

Modeling Issues

Gabriel et al. [85] compared the calculated results using the static capacitance model, Marcuvitz model, and the MOM solution. It was concluded that the Marcuvitz formulation is accurate enough to predict the complex permittivity up to 20 GHz. They used this model for their later comprehensive measurement of various biological tissues up to 20 GHz [5].

A similar study was done by Misra et al. [80]. They used water, methanol, and dioxane as the standard liquid and measured various water–dioxane mixtures from 1 to 18 GHz. The study was performed for both lumped element and the Marcuvitz model for probe admittance. For a 80% water and 20% dioxane mixture, the deviation of measured permittivity from the literature values at 18 GHz was negligible for the Marcuvitz method but was 15% and 28% for real and imaginary parts for the lumped element model.

At higher frequencies and for better accuracy, resorting to numerical techniques seems inevitable [105,106]. In this scenario, it is useful to fit the modeling data to a function [82,107] by Eq. (18.56), a well-known example for open-ended probe. Numerical simulations such as FDTD [134,135] are particularly useful when the impact of deviation from ideal model such as probe flange [86] or sample container effects [89] are studied.

Sensitivity and Uncertainty

Qualitatively speaking, the higher is the change in the reflection coefficient of a probe (a more sensitive probe) the lower is the measurement uncertainty. For lumped-capacitance model, Stuchly and coworkers [43,66,67] quantifies this in terms of the optimum fringing capacitance in air (C_0) of a probe that yields the lowest uncertainty:

$$C_0 = \frac{1}{2\pi f\, Z_0} \frac{1}{\sqrt{\varepsilon'^2 + \varepsilon''^2}} \tag{18.66}$$

where Z_0 is the characteristic impedance of the probe. This quantity sets the probe dimension and apparently depends on the range of material to be measured. For instance, for water at 10 GHz the optimum capacitance is 0.005 pF [43].

A formal definition of the sensitivity is given by Stuchly et al. [82]. The sensitivity of a parameter Γ with respect to the change of another parameter ε is defined as

$$S_{\varepsilon}^{\Gamma} = \frac{\partial\Gamma/|\Gamma|}{\partial\varepsilon/|\varepsilon|} = \frac{|\varepsilon|}{|\Gamma|}\frac{\partial\Gamma}{\partial\varepsilon} \tag{18.67}$$

The magnitude of the sensitivity of a response to the change in the unknown parameter (which is to be extracted from the response) has some practical implications in error/uncertainty evaluation. For instance, a sensitivity of 2 dB implies that with a 10% error in reflection coefficient response, the error in the extracted parameter is 8%. Clearly, a negative sensitivity (in dB scale) is not very demanding, whereas a highly positive one can significantly reduce the impact of measurement error (systematic or random) on the extracted parameters. Knowing the relation between Γ and ε in closed form [see Eq. (18.56)], one can calculate this quantity. Figure 18.18 represents the sensitivity for water and methanol of 2.2 mm and 3.6 mm probes up to 18 GHz [83].

Higher Order Modes

For a fixed probe size, higher order modes' effect starts to appear at higher frequencies. This is the limiting factor of using the standard probe sizes in frequencies above 20 GHz. For a coaxial cable the cutoff frequency of first higher order (i.e., TE_{11}) mode is given by

$$f_c \approx \frac{v}{\pi(a+b)} \tag{18.68}$$

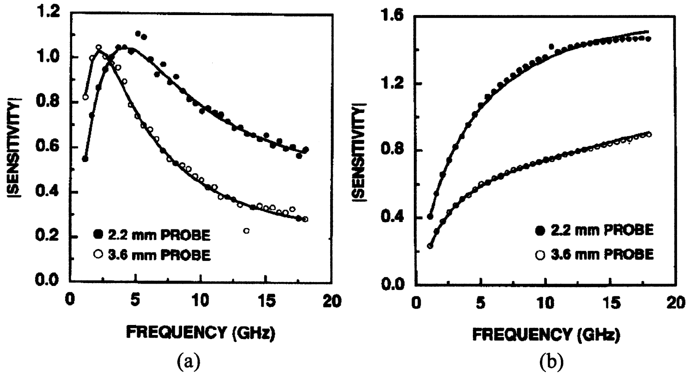


Figure 18.18 Sensitivity for (a) water and (b) methanol of 2.2-mm and 3.6-mm probes up to 18 GHz [83] (with permission from IEEE).

where v is the phase velocity of the dominant TEM mode in the coaxial probe. Even a nonpropagating higher order mode, if it reaches the adaptor connecting the ANA cable to the probe, after reflecting at the probe aperture, can affect the measurement results.

Cole-Cole Fitting

In a recent survey, Gabriel et al. [123] have represented a comprehensive collection of complex permittivity of various tissues. They also measured the tissue dielectric parameter [5] from 10 Hz to 20 GHz at 37°C, and provide parametric models for 17 types of tissues [123]. The parametric model they used is a four-term Cole-Cole relation:

$$\varepsilon(\omega) = \varepsilon' - j\varepsilon'' = \varepsilon_\infty + \sum_{n=1}^{4} \frac{\Delta\varepsilon_n}{1 + (j\omega\tau_n)^{1-\alpha_n}} - \frac{j\sigma_I}{\omega\varepsilon_0} \tag{18.69}$$

This equation corresponds to four different dispersion regions and the corresponding parameters are obtained by fitting the measurement results to the above equation. These parameters are also tabulated by Gabriel et al. [123]. Furthermore, they provide graphs for complex permittivity of the 17 tissue types from dc up to 100 GHz obtained from Eq. (18.69). The complex permittivity is given in terms of the real part (i.e., ε') and the total conductivity of $\sigma = \omega\varepsilon_0\varepsilon''$.

In another study, Bao et al. [88] measure the complex permittivity of the rat brain grey and white matters up to 26.5 GHz at 25°C (24°C for white matter) and 37°C. Using a nonlinear least square algorithm, they fit the measured results to a two-term Cole-Cole relation covering 45 MHz to 26.5 GHz.

18.4.2. A Two-Port Microstrip Measurement System

This section provides an overview of a state-of-the art technique to extract complex permittivity of biological tissues up to 50 GHz, based on the microstrip test fixture shown in Fig. 18.13. This test fixture is employed as part of the two-port measurement system, where complex permittivity of dielectric and biological samples under test are extracted by comparing the measured scattering parameters with the simulated ones.

Test Fixture Setup

Figure 18.19 presents a close look photo of the setup developed by Tofighi and Daryoush [25,105–109]. The test fixture microstrip cavity structure, ANA test port cables and connectors, input launchers, and TUT embedded in the test fixture are illustrated in the figure. TUT is placed between two microstrip cavities as shown in Fig. 18.19. A special sample container is also designed for containing biological tissues and liquids (cf. Fig. 18.19b). The sample container consists of a rectangular frame of an acrylic material made from a 0.04-in (1.02 mm) thick polycarbonate sheet. Two glass coverslips (Corning No.1 cover glass, $\varepsilon_r = 6.6$) with dimensions of $22 \times 50\,\text{mm}^2$ and

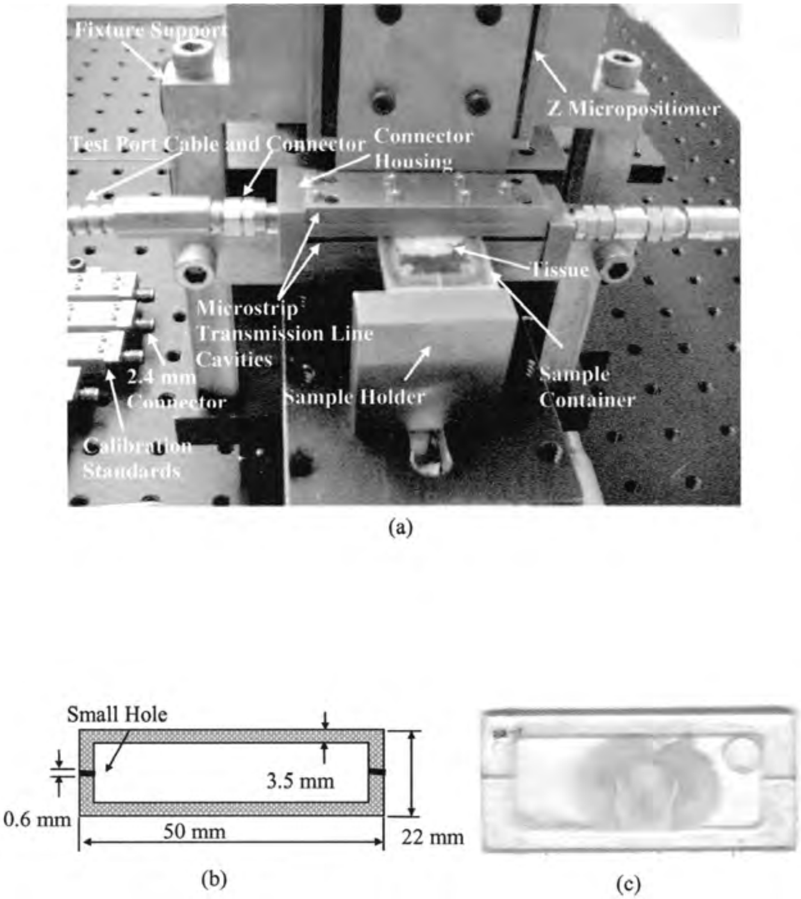


Figure 18.19 (a) Close-up photos of the experimental setup of the test fixture. This structure is designed for characterizing the complex permittivity of TUT with different thicknesses. (b) Sample container structure made from an acrylic sheet (1.02 mm thick). (c) Photos of a typical slice of rat brain inside a sample container filled with saline. (With permission from IEEE.)

0.15 mm thick, are glued to the two sides of the frame using rubber cement glue. Liquids (e.g., water and saline) are injected inside the container. This test fixture has been used for the extraction of complex permittivity of white and grey brain matters of rat brain slices (cf. Fig. 18.19c) at 27°C at frequency range of 15 to 50 GHz as explained later in this section.

Two-Port Calibration

A 10-term model (cf. Fig. 18.20) is the simplified version of a well-known 12-term model [97], assuming that ANA transfer switches do not introduce different impedances in forward and reverse regimes. Using measurements of known standards (e.g., short, open, matched load, transmission), a efficient number of independent equations are obtained, where their simultaneous solutions will lead to the error terms [97,134] necessary to deembed the error networks. Error terms e_{30} and e_{03} are due to isolation and are usually negligible in low to moderate insertion loss measurements.

Thru reflect line (TRL) calibration is one of the most accurate two-port calibration techniques (cf. Fig. 18.21). Most of the today's ANA are capable of performing this type of calibration [131,132]. This technique requires the availability of two transmission lines, which are 90°$(=\beta\Delta l)$ different in their electrical length at the center frequency of the calibration (cf. Fig. 18.21c). A third standard is also needed, which should be a reflective one-port network either open or short (not necessarily perfect ones). Besides the input and output error networks, the propagation constant of the microstrip line and consequently the substrate permittivity can be obtained by TRL calibration routine. The theory behind TRL calibration is somewhat involved and can be found everywhere [96].

Higher Order Modes

To analyze the test fixture, a commercial finite element method (FEM) package from Agilent (HFSS 5.2) is used for full-wave characterization of the test fixture

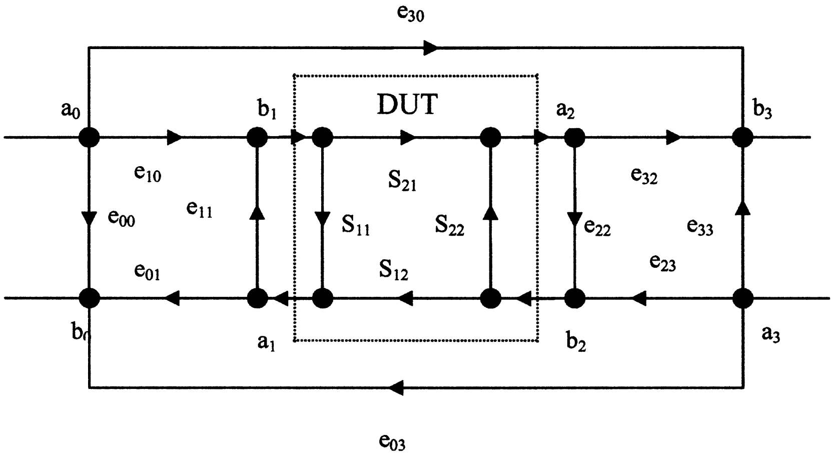


Figure 18.20 Ten-term error model of a two-port network used to identify the embedded network from the DUT.

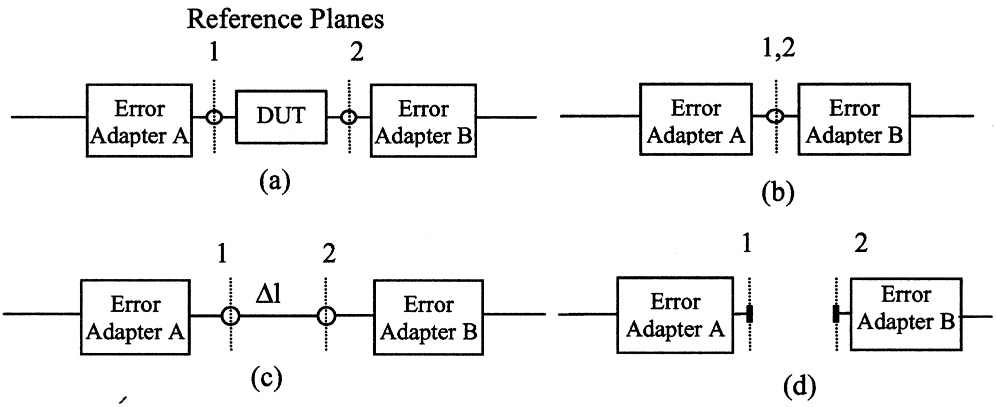


Figure 18.21 TRL calibration procedure to calibrate ANA at the reference planes 1 and 2: (a) DUT and error adapters, (b) through connection, (c) delay line with a length of L, and (d) reflect are placed between planes 1 and 2.

[105–107]. Based on the modal analysis, a microstrip line with the dimensions as shown in Fig. 18.13 on a fused silica substrate supports single mode operation up to 53.5 GHz. The second order mode has a 37 dB/cm attenuation at 50 GHz. In other words, the second order mode attenuates about 96.7 dB at 50 GHz before reaching the launchers for a launcher to launcher distance (see Fig. 18.13a) of 52 mm used in the design. The characteristic impedance of the line should be close to 50 Ω. Although having a 50-Ω characteristic impedance is not a requirement, it provides a lower return loss for coaxial to microstrip launcher than any other choice of the line characteristic impedance. As a result, the dynamic range of the system is expected to be maximized for this choice.

Modeling and Extraction Procedures

Performance of the test fixture including the TUT is evaluated using the FEM model in terms of its S parameters. The model includes the microstrip line, apertures, glass plates, TUT, and metallic housing. The absorbing boundary condition is defined to delimit the tissue boundary. A homogeneous distribution of tissue is considered in the modeling [105,106]. The modeling is performed for various frequencies from 5 to 50 GHz and for values of $\tan\delta$ and ε', over the range of $0.15 \leq \tan\delta \leq 2.5$ and $2.5 \leq \varepsilon' \leq 75$. These values are within the expected range of biological tissues reported in the literature [122]. The simulation is performed once and the results are fitted to a complex function. Tofighi and Daryoush [107] employ a rational function relation similar to the one presented in Eq. (18.56) for their two-port test fixture. The S parameter at each frequency is fitted to a rational function of complex permittivity, as described in the following equation, as opposed to fitting the modeling S parameter for all frequencies to a single function:

$$S_{ij} = \frac{\sum_{p=0}^{P} A_p \varepsilon^P}{1 + \sum_{q=1}^{Q} B_q \varepsilon^q} \tag{18.70}$$

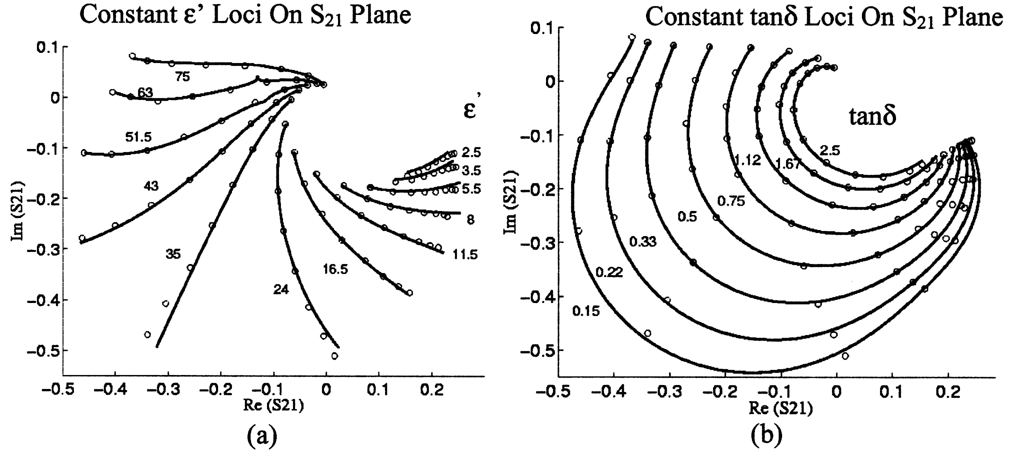


Figure 18.22 The loci of fixed (a) ε' and (b) $\tan\delta$ plotted in S_{21} plane at 15 GHz, obtained by the fitting method (—) to the data from the modeling (○) (with permission from IEEE).

A plot representing loci of S_{21} for fixed values of ε' and $\tan\delta$ is very helpful to visualize the success of this fitting procedure. Figure 18.22 represents a comparison between the simulated S_{21} and S_{21} fitted to Eq. (18.70) at 15 GHz. These loci are, in fact, perpendicular to one another at each modeling point (o).

Measurement and Extraction of the Complex Permittivity

Measurements are performed for distilled water and biological tissues (white and grey brain matters) at 27 ± 0.5°C. The extracted ε' and ε'' for distilled water are given at Fig. 18.23 for frequencies of 15 to 50 GHz. A single-term Debye relaxation is generally accepted for water dispersion at microwave frequencies [129]. The ripple-like behavior is a result of the lack of repeatability in performance of the coaxial to microstrip launchers, which were employed in TRL calibration standards and the test fixture. They can be removed by a correction technique, which employs S_{21} of water as a known reference material [107,108].

Biological tissue characterization can also be performed using this test fixture. S parameter measurements were performed for the cerebral cortex in the front brain and pons in the back of the brain as grey and white matters respectively. A relatively high proportion of nerve cell nuclei exist in grey matter, whereas white matter consists mainly of axons. Small ripple-like behaviors are observed for the extracted ε' and ε''. The results are presented in Fig. 18.24. The results are compared to the results of a four-term Cole-Cole dispersion relation provided by Gabriel et al. (the relation is obtained based on the measured data below 20 GHz) [123], and a newer two-term Cole-Cole dispersion relation for rat brain at 24°C for grey matter and 25°C for white matter given by Bao et al. (the relation is obtained from measured data below 26.5 GHz) [88]. These results are also compared to the measurement results for rabbit by Steel and Sheppard [47] at 35 GHz, by applying a linear interpolation for 27°C from their tabulated results given at two different temperatures (i.e., 20°C and 37°C). The results suggest that the extracted complex permittivity values for grey matter and white matter match better to the model provided by Bao et al. [88] and

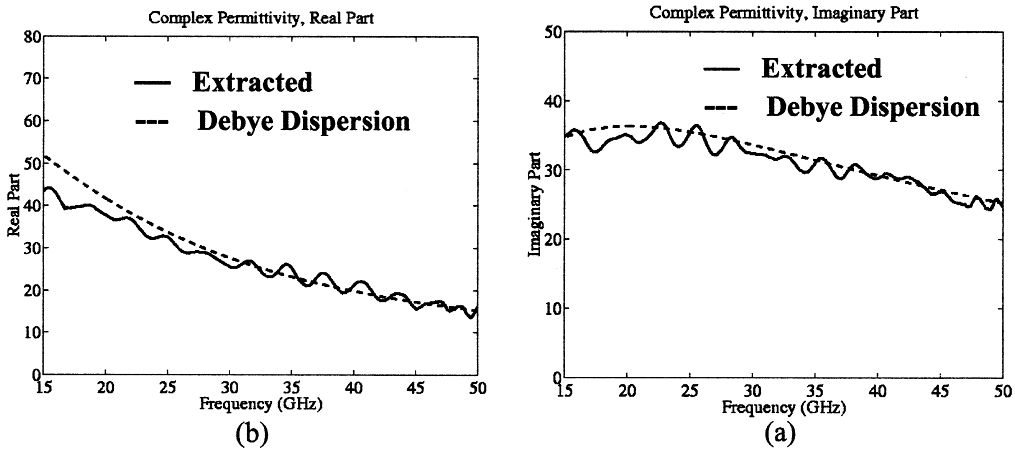


Figure 18.23 Extracted (—) and Debye model [129] (- - -) results of the complex permittivity ε ($\varepsilon = \varepsilon' - j\varepsilon''$) for distilled water at 27°C as a function of frequency: (a) real part (ε') and (b) imaginary part (ε'') [107] (with permission from IEEE).

Gabriel et al. [123], respectively. Nonetheless, a further refinement of the published Cole-Cole models seems necessary as the extracted results are matched with the published ones at 35 GHz [47].

Fitting the Complex Permittivity to Cole-Cole Relation

The extracted results show Cole-Cole like dispersion characteristics, with a characteristic frequency (peak of absorption) around 22 GHz for both grey and white matters. Using a nonlinear least square fitting method [92], the results of Fig. 18.24 are fitted to a single-term Cole-Cole relation for γ dispersion:

$$\varepsilon = \varepsilon_\infty + \frac{(\varepsilon_s - \varepsilon_\infty)}{1 + (j\omega\tau)^{1-\alpha}} + \frac{\sigma}{j\omega\varepsilon_0} \tag{18.71}$$

To account for the existing data at lower frequencies, published permittivity results above 1 GHz for white and grey matters [120,121] are also included in the fitting procedure. Table 18.2 provides the Cole-Cole parameters obtained by this fitting, where a further refinement to the published model parameters in references 88 and 123 is made.

Measurement Sensitivity and Uncertainty

Figure 18.25 illustrates the sensitivity of S parameters, defined in the previous sections

$$S_{\varepsilon}^{S_{ij}} = \frac{|\varepsilon|}{|S_{ij}|}\frac{\partial S_{ij}}{\partial \varepsilon}, \tag{18.72}$$

for water and brain grey and white matters (i.e., high water content tissues) and methanol (i.e., a low-loss liquid). Figure 18.25 shows that for water and high water

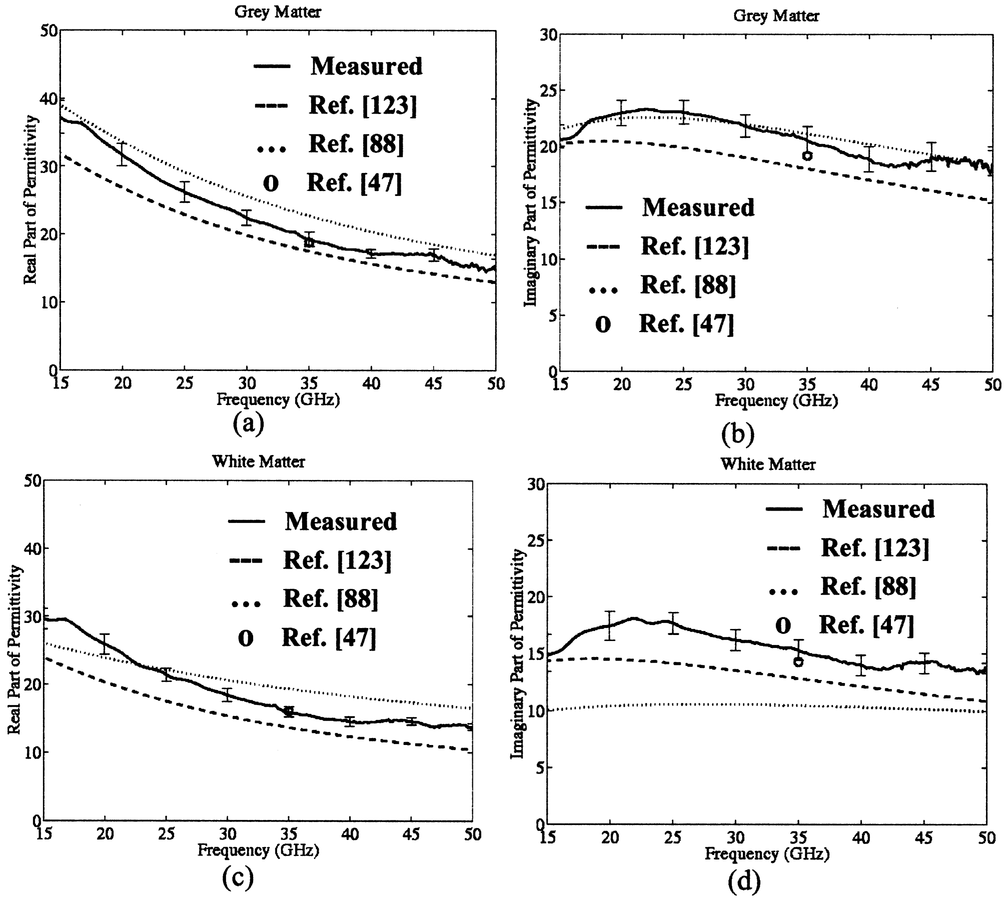


Figure 18.24 Extracted complex permittivity ε ($\varepsilon = \varepsilon' - j\varepsilon''$) for grey and white matters at 27°C as a function of frequency compared with the literature: (a) grey, real part (ε'), (b) grey, imaginary part (ε''), (c) white, real part (ε'), and (d) white, imaginary part (ε'') [107] (with permission from IEEE).

Table 18.2 The Cole-Cole Parameters of Brain White and Grey Matters for γ-Dispersion above 1 GHz

	τ (ps)	ε_s	ε_∞	σ (S/m)	α
Grey matter[a]	6.75	49.5	5.8	0.96	0
White matter	6.34	37.3	5.3	0.79	0

[a] Measurement results of this study above 15 GHz at 27°C and the published results of the literature [120,121] above 1 GHz were included to obtain these parameters [107] (with permission from IEEE).

content tissues (i.e., waterlike media) the sensitivity of S_{21} is better than S_{11} sensitivity by 15 dB. In this case S_{21} sensitivity is 3–7 dB. Recalling Fig. 18.18, we note that the sensitivity of a 2.2 mm is at best 1 (= 0 dB) at 5 GHz that reduces to −4 dB at 18 GHz and expected to decrease even further at higher frequencies. This argument highlights the advantage of using transmission parameter (S_{21}) measurement as opposed

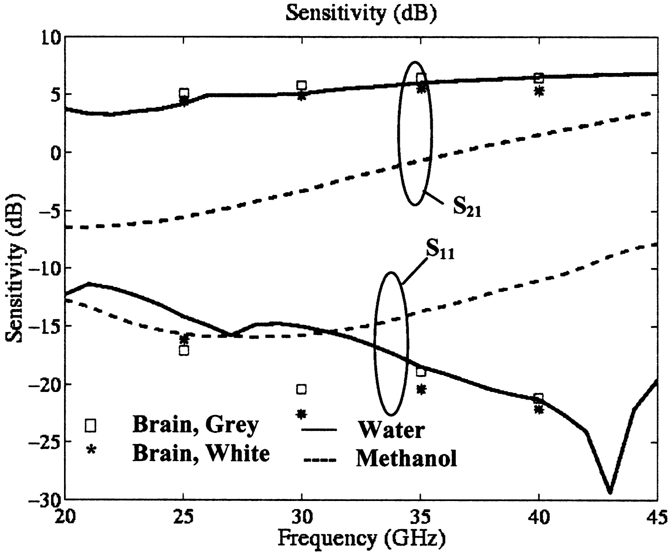


Figure 18.25 The sensitivity of S_{21} and S_{11} for water, brain (grey and white) matters, and methanol.

to the reflection parameter (S_{11}) measurement as far as the extraction accuracy is concerned [108].

The higher sensitivity implies lower uncertainty in the extracted results [107]. The uncertainty in the extracted values of ε' and ε'' are due to S_{21} error and are associated with a number of error sources: (1) sample container thickness tolerance, (2) fitting error, and (3) error in the reference water complex permittivity due to its temperature uncertainty. The measurement's systematic errors are already reduced to a level sufficiently below the other errors by TRL calibration and water based correction methods that were explained before. Numerical modeling error is also reduced by careful study of the fixture response for known two-port networks formed by removing TUT (i.e., replacing it with air gap) and changing the separation between the apertures.

The uncertainties are referred to as δ'_i and δ''_i (for real and imaginary parts respectively), where i is the index for identifying various independent contributors. Then, the total uncertainties are

$$\delta\varepsilon'_{\text{tot}} = \sqrt{\sum_i (\delta\varepsilon'_i)^2} \qquad \delta\varepsilon''_{\text{tot}} = \sqrt{\sum_i (\delta\varepsilon''_i)^2} \tag{18.73}$$

The uncertainties in S_{21}, i.e., $\delta S_{21,i}$ are estimated through modeling and simulation [107]. Once they are known the corresponding uncertainty in the complex permittivity can be estimated through relation

$$\delta\varepsilon_i = \delta\varepsilon'_i - j\delta\varepsilon''_i = \frac{\delta S_{21,i}}{(\partial S_{21}/\partial\varepsilon)} \tag{18.74}$$

where the denominator can be easily obtained using Eq. (18.70).

Table 18.3 Measurement Uncertainty, δ'_i and δ''_i, from Various Sources, the Total Uncertainties, $\delta\varepsilon'_{\text{tot}}$ and $\delta\varepsilon''_{\text{tot}}$, and the Total Relative Uncertainties, $\delta'_{\text{tot}}/\varepsilon'$ and $\delta''_{\text{tot}}/\varepsilon''$

	$\delta\varepsilon'_1$	$\delta\varepsilon''_1$	$\delta\varepsilon'_2$	$\Delta\varepsilon''_2$	$\delta\varepsilon'_3$	$\delta\varepsilon''_3$	$\delta\varepsilon'_4$	$\delta\varepsilon''_4$	$\delta\varepsilon'_{\text{tot}}$	$\delta\varepsilon''_{\text{tot}}$	$\delta\varepsilon'_{\text{tot}}\varepsilon'$	$\delta\varepsilon''_{\text{tot}}\varepsilon''$
Grey	0.7	0.1	0.1	0.2	0.2	0.2	1.2	1.0	1.4	1.1	0.06	0.05
White	0.6	0.1	0.1	0.2	0.1	0.2	1.0	0.9	1.1	1.0	0.06	0.06

[a] $i=1$; fitting error, 2; tissue thickness, 3; reference temperature, and 4; aperture placement on tissue.
[b] Obtained for white and grey matters at 30 GHz [108].
Source: With permission from IEEE.

On the other hand, there is some uncertainty due to the placement of aperture on top of a white or grey tissue region, which might not be exactly of the same texture, composition, or placement for all measurements (i.e., placement uncertainty, $\delta S_{21,4}$). Table 18.3 lists the result of various uncertainties and the total uncertainties obtained from Eq. (18.73) for measurement at 30 GHz [108]. The placement uncertainty ($\delta\varepsilon_4$) is obtained by repetitive tissue measurements and taking the standard deviation of the extracted results. This source of uncertainty is clearly the dominant factor. The same analysis is repeated for selected frequencies from 15 to 50 GHz. The results of this analysis have been already shown as error bars in Fig. 18.24. The uncertainty for both real and imaginary parts of the complex permittivity varies from 4% to 8% for the entire range, except the real part above 45 GHz, where error exceeds 10%.

18.4.3. Tissue Cole-Cole Parameters

A large body of knowledge has been accumulated on the dielectric properties of biological materials since the pioneering works by Cook [118] and Schwan [30] in the 1950s. These data extended to in vivo for some tissues, following the work by Burdette et al. and others [71,72,119].

In an effort to collect tissue behavior at electromagnetic frequencies, Stuchly and Stuchly [120] tabulated the available data in the form of ε' and ε'' over the frequency range of 10 KHz to 10 GHz. The listed materials in their tabulation include a large body of biological tissues and phantoms for some of these tissues [120]. A complete set of information on the biological tissues including their electrical properties is also summarized in a book by Duck [121]. The previously mentioned parametric model of various tissues' complex permittivity using a four-term Cole-Cole relation of Eq. (18.69), given by Gabriel et al. [123], complements these studies. Table 18.4 lists parameters of Eq. (18.69) used to predict the dielectric properties of tissues.

To appreciate the impact of Cole-Cole parameters on the complex permittivity of biological tissues, two examples of high and low water contents such as muscle and fat are depicted over RF frequencies of interest (viz. 100 MHz to 100 GHz). Figure 18.26 depicts the real (a) and imaginary part (b) of complex permittivity. The overall range of variation from dc to 100 GHz is over seven orders of magnitude, however over frequencies of 100 MHz to 100 GHz this variation is as high as one order of magnitude. Note that above 400 MHz, where the dipolar relaxation of water is the dominant polarization mechanism, the single term relaxation relation of Eq. (18.71) can be used. Particularly, σ in that equation represents the lower frequency polarization mechanisms, for which $\omega\tau \gg 1$ at RF frequencies. In fact, the conductivity due to the ionic drift is negligible at these frequencies as well.

Table 18.4 The Cole-Cole Parameters for a Variety of Biological Tissues [123]

Tissue type	ε_∞	$\Delta\varepsilon_1$	τ_1 (ps)	α_1	$\Delta\varepsilon_2$	τ_2 (ns)	α_2	$\Delta\varepsilon_3$	τ_3 (μs)	α_3	$\Delta\varepsilon_4$	τ_4 (ms)	α_4	σ_I
Blood	4.0	56.0	8.38	0.10	5200	132.63	0.10	0.0			0.0			0.7000
Bone (cancellous)	2.5	18.0	13.26	0.22	300	79.58	0.25	2.0×10^4	159.15	0.20	2.0×10^7	15.915	0.00	0.0700
Bone (cortical)	2.5	10.0	13.26	0.20	180	79.58	0.20	5.0×10^3	159.15	0.20	1.0×10^5	15.915	0.00	0.0200
Brain (grey matter)	4.0	45.0	7.96	0.10	400	15.92	0.15	2.0×10^5	106.1	0.22	4.5×10^7	5.305	0.00	0.0200
Brain (white matter)	4.0	32.0	7.96	0.10	100	7.96	0.10	4.0×10^4	53.05	0.30	3.5×10^7	7.958	0.02	0.0200
Fat (infiltrated)	2.5	9.0	7.96	0.20	35	15.92	0.10	3.3×10^4	159.15	0.05	1.0×10^7	15.915	0.01	0.0350
Fat (not infiltrated)	2.5	3.0	7.96	0.20	15	15.92	0.10	3.3×10^4	159.15	0.05	1.0×10^7	7.958	0.01	0.0100
Heart	4.0	50.0	7.96	0.10	1200	159.15	0.05	4.5×10^5	72.34	0.22	2.5×10^7	4.547	0.00	0.0500
Kidney	4.0	47.0	7.96	0.10	3500	198.94	0.22	2.5×10^5	79.58	0.22	3.0×10^7	4.547	0.00	0.0500
Lens Cortex	4.0	42.0	7.96	0.10	1500	79.58	0.10	2.0×10^5	159.15	0.10	4.0×10^7	15.915	0.00	0.3000
Liver	4.0	39.0	8.84	0.10	6000	530.52	0.20	5.0×10^4	22.74	0.20	3.0×10^7	15.915	0.05	0.0200
Lung (inflated)	2.5	18.0	7.96	0.10	500	63.66	0.10	2.5×10^5	159.15	0.20	4.0×10^7	7.958	0.00	0.0300
Muscle	4.0	50.0	7.23	0.10	7000	353.68	0.10	1.2×10^6	318.31	0.10	2.5×10^7	2.247	0.00	0.2000
Skin (dry)	4.0	32.0	7.23	0.00	1100	32.48	0.20	0.0			0.0			0.0002
Skin (wet)	4.0	39.0	7.96	0.10	280	79.58	0.00	3.0×10^4	1.59	0.16	3.0×10^4	1.592	0.20	0.0004
Spleen	4.0	48.0	7.96	0.10	2500	63.66	0.15	2.0×10^5	265.26	0.25	5.0×10^7	6.366	0.00	0.0300
Tendon	4.0	42.0	12.24	0.10	60	6.37	0.10	6.0×10^4	318.31	0.22	2.0×10^7	1.326	0.00	0.2500

Source: With permission from IOP Publishing Ltd.

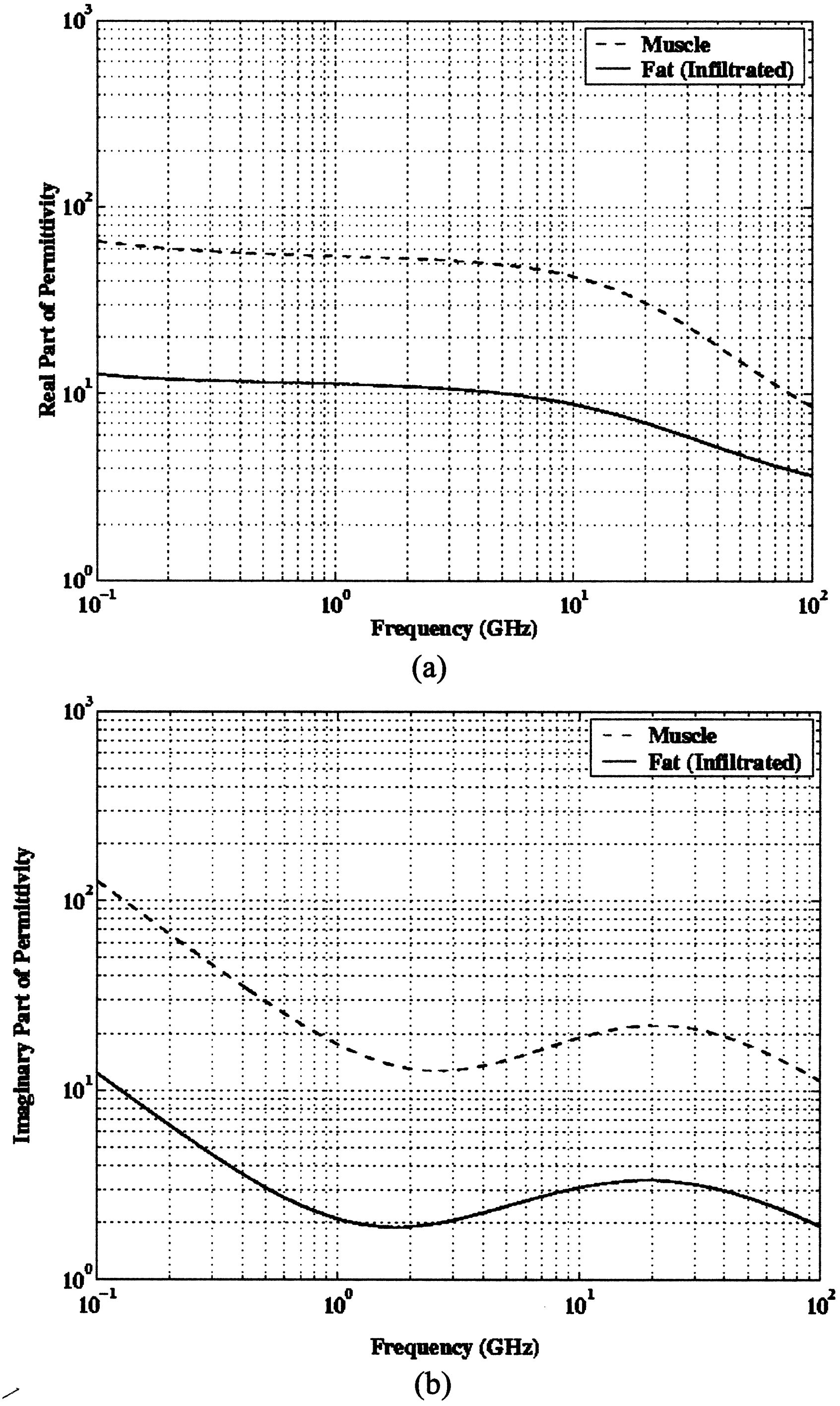


Figure 18.26 The simulated complex permittivity of low (fat) and high (muscle) water content biological tissues as a function of RF frequencies: (a) real part of permittivity (ε') and (b) imaginary part of permittivity (ε''). The graphs are generated based on Cole-Cole parameters from Table 18.4.

18.5. CONCLUSION

A number of techniques dealing with measurements of complex permittivity of dielectrics are reviewed, and strengths and weaknesses of each technique are presented. These techniques are extended for modeling and measurements of biological tissues. Even though complex permittivity data on biological tissues are primarily limited to 20 GHz today, narrowband measurements of biological tissues in millimeter wave frequencies of even close to 100 GHz are reported. Moreover, a new methodology based on two-port measurement approach is used to measure and extract complex permittivity of white and gray brain matters up to 50 GHz. Because of the high dynamic range of the variation for S_{21} compared to S_{11}, this technique provides great accuracy and repeatability and is employed to correct the Cole-Cole parameters for brain matter up to millimeter wave frequencies. Nonetheless, more studies are needed to characterize the electromagnetic field interactions with biological systems in millimeter wave frequencies, by characterization of various tissues at these frequencies.

REFERENCES

1. Javadi, H.S. Microwave Material, In: *Handbook of Microwave Technology*; Ishii, T.K. Ed.; Academic Press: San Diego, 1995; Vol. 2, 605–643, Chapter 19.
2. Baker-Jarvis, J.; Bill Riddle.; Janezic, M.D. *Dielectric and Magnetic Properties of Printing Wiring Boards and Other Substrate Material*; NIST Technical Note 1512, **March 1999**.
3. Rosen A.; Rosen, H. (Eds.). *New Frontiers in Medical Device Technologies*; John Wiley: New York, 1995.
4. Gandhi, O.P.; Lazzi, G.; Furse, C.M. Electromagnetic absorption in the human head and neck for mobile telephones at 835 and 1900 MHz. IEEE Trans. Microwave Theory Techn. **Oct. 1996**, *MTT-44*(10), 1884–1897.
5. Gabriel, S.; Lau, R.W.; Gabriel, C. The dielectric properties of biological tissues: II. Measurements in the frequency range 10 Hz to 20 GHz. Phys. Med. Biol. **1996**, *41*, 2251–2269.
6. Larsen, L.E.; Jacobi, J.H. (Eds.). *Medical Application of Microwave Imaging*; IEEE Press: New York, 1986.
7. Special Issue on Medical Application and Biological Effects of RF/Microwaves, Rosen, A., Vander Vorst, A. Eds.; IEEE Trans. Microwave Theory Tech. **Oct. 1996**, *MTT-44*(10), Part II, 1753–1973.
8. Special Issue on Medical Application and Biological Effects of RF/Microwaves, Rosen, A., Vander Vorst, A., Kotsuka, Y. Eds.; IEEE Trans. Microwave Theory Tech. **Nov. 2000**, *MTT-48*(11), Parts I & II, 1781–2198.
9. Polk C.; Postow, E. (Eds.). *CRC Handbook of Biological Effects of Electromagnetic Fields*; CRC Press: Boca Raton, Florida, 1986.
10. Battocletti, J.H. Biomedical applications of microwave engineering. In *Handbook of Microwave Technology*; Ishii, T.K. Ed.; Academic Press: San Diego, 1995; 309–345, Chapter 11.
11. Gandhi, O.P.; Lazzi, G.; Tinniswood, A.; Yu, Q. Comparison of numerical and experimental methods for determination of SAR and radiation patterns of handheld wireless telephones. Bioelectromagnetics **1999**, *20*, 93–101.
12. Okoniewski, M.; Stuchly, M. A study of the handset antenna and human body interaction. IEEE Trans. Microwave Theory Techn. **Oct. 1996**, *44*(10), 1855–1865.
13. Jensen M.; Rahmat-Samii, Y. EM interaction of handset antennas and a human in personal communications. Proc. IEEE **Jan. 1995**, *83*(1), 7–17.

14. Dimblow P.J.; Mann, S.M. SAR calculations in an anatomically realistic model of the head for mobile communication transceivers at 900 MHz and 1.8 GHz. Phys. Med. Biol. **1994**, *39*, 1537–1533.
15. Bernardi, P.; Cavagnaro, M.; Pisa, S.; Piuzzi, E. Human exposure to radio base-station antennas in urban environment. IEEE Trans. Microwave Theory Tech. **Nov. 2000**, *MTT-48*(11), Part II, 1996–2002.
16. Schiavoni, A.; Bertotto, P.; Richiardi, G.; Bielli, P. SAR generated by commercial cellular phone modeling, head modeling, and measurements. IEEE Trans. Microwave Theory Tech. **Nov. 2000**, *MTT-48*(11), Part II, 2064–2071.
17. Rosen, A.; Rosen, D.; Tuma, G.A.; Bucky, L.P. RF/Microwave-aided tumescent liposuction. IEEE Trans. Microwave Theory Tech. **Nov. 2000**, *MTT-48*(11), Part I, 1879–1884.
18. Sterzer, F.; Mendecki, J.; Mawhinney, D.D.; Friedenthal, E.; Melman, A. Microwave treatments for prostate disease. IEEE Trans. Microwave Theory Tech. **Nov. 2000**, *MTT-48*(11), Part I, 1885–1891.
19. Hiraoka, M.; Mitsumori, M.; Hiroi, N.; Ohno, S.; Tanaka, Y.; Kotsuka, Y.; Sugimachi, K. Development of RF and microwave heating equipment and clinical applications to cancer treatment in Japan. IEEE Trans. Microwave Theory Tech. **Nov. 2000**, *MTT-48*(11), Part I, 1789–1799.
20. Dunn, D.; Rappaport, M.; Terzuoli, A.J. Verification of deep-set brain tumor hyperthermia using a spherical microwave source distribution. IEEE Trans. Microwave Theory Tech. **Oct. 1996**, *MTT-44*(10), 1769–1776.
21. Camart, J.; Despretz, D.; Chive, M.; Pribetich, J. Modeling of various kinds of applicators used for microwave hyperthermia based on the FDTD method. IEEE Trans. Microwave Theory Tech. **Oct. 1996**, *MTT-44*(10), 1811–1818.
22. Labonte, S.; Blais, A.; Legault, S.R.; Ali, H.O.; Roy, L. Monopole antennas for microwave catheter ablation. IEEE Trans. Microwave Theory Techn. **Oct. 1996**, *MTT-44*(10), 1832–1839.
23. Jofre, L.; Hawley, M.S.; Broquetas, A.; Reyes, E.; Ferrando, M.; Elias-Fuste, A.R. Medical imaging with a microwave tomographic scanner. IEEE Trans. Biomed. Eng. **Mar, 1990**, *BME-37*(3), 303–310.
24. Hagness, S.C.; Taflove, A.; Bridges, J.E. Two-dimensional FDTD analysis of a pulsed microwave confocal system for breast cancer detection: fixed-focus and antenna-array sensors. IEEE Trans. Biomed. Eng. **Dec. 1998**, *BME-45*(12), 1470–1479.
25. Tofighi, M.R.; Daryoush, A.S. Near field microwave imaging of brain. Electron. Lett. **June 2001**, *37*(13), 807–808.
26. Frohlich, H. *Theory of Dielectrics, Dielectric Constants, and Dielectric Loss*; Oxford University Press: Amen House, London, 1958.
27. Daniel, V.V. *Dielectric Relaxation*; Academic Press: London, 1967.
28. Pethig, R. *Dielectric and Electronic Properties of Biological Materials*; Wiley: New York, 1979.
29. Pethig, R.; Kell, D.B. The passive electrical properties of biological systems: their significance in physiology, biophysics and biotechnology. Phys. Med. Biol. **1987**, *32*(8), 933–970.
30. Schwan, H.P. Electrical properties of tissues and cell suspensions. Advanced Phys. Med. Biol. **1957**, *5*, 147–209.
31. Schwan, H.P.; Foster, K.R. RF-field interactions with biological systems: electrical properties and biophysical mechanism. Proc. IEEE **1980**, *68*, 104–113.
32. Schwan, H.P. Dielectric properties of biological tissues and biophysical mechanisms of electromagnetic field interaction. In *Biological Effects of Nonionizing Radiation*; Illinger, K.H. Ed.; ACS Symposium Series: Washington D.C., 1981, 109–131.
33. Cole, K.S.; Cole, R.H. Dispersion in dielectrics; I. Alternating current characteristics. J. Chem. Phys. **Apr. 1941**, *9*, 341–351.
34. Grant, E.H.; Keefe, S.E.; Takashima, S. The dielectric behavior of aqueous solutions of bovine serum albumin from radiowave to microwave frequencies. J. Phys. Chem. **1968**, *72*, 4373–4380.

35. Foster, K.R.; Schepps, J.L.; Schwan, H.P. Microwave dielectric relaxation in muscle: A second look. Biophys. J. **1980**, *29*, 271–281.
36. Afsar, M.N.; Hasted, J.B. Measurements of optical constants of liquid H_2O and D_2O. J. Opt. Soc. Amer. **July 1977**, *67*, 902–904.
37. Grant, E.H.; Nightingale, R.V.; Sheppard, R.J. Dielectric properties of water in myoglobin solution. In *Biological Effects of Nonionizing Radiation*; Illinger, K.H. Ed.; ACS Symposium Series: Washington D.C., 1981, 57–62.
38. Grant, E.H.; Szwarnowski, S.; Sheppard, R.J. Dielectric properties of water in microwave and far-infrared regions. In: *Biological Effects of Nonionizing Radiation*; Illinger, K.H. Ed.; ACS Symposium Series: Washington D.C., 1981, 47–56.
39. Westphal, W.B. Dielectric measuring techniques. In *Dielectric Material and Applications*; Von Hippel, A.R. Ed.; Wiley: New York, 1954, 63–122.
40. Fox J.; Sucher, M. In: *Handbook of Microwave Measurements*; Polytechnique Institute of Brooklyn: New York, 1954.
41. Bussey, H.E. Measurement of RF Properties of Materials, A Survey. Proc. IEEE **June 1967**, *55*(6), 1046–1053.
42. Grant, E.H.; Sheppard, R.J.; South, G.P. *Dielectric Behaviour of Biological Molecules in Solution*; Oxford University Press: Oxford, 1978.
43. Stuchly, M.M.; Stuchly, S.S. Coaxial line reflection methods for measuring dielectric properties of biological substances at radio and microwave frequencies— A review. IEEE Trans. Instrum. Meas. **Sept**. **1980**, *IM-29*(3), 176–183.
44. Afsar, M.N.; Birch, J.R.; Clarke, R.N. The measurement of the properties of materials. Proc. IEEE **Jan**. **1986**, *74*(1), 183–199.
45. Agilent 85070D Dielectric Probe Kit, Product Overview, Agilent Technology, www.agilent.com.
46. Steel, M.C.; Sheppard, R.J. The dielectric properties of rabbit tissue, pure water and various liquids suitable for tissue phantoms at 35 GHz. Phys. Med. Biol. **1988**, *33*, 467–472.
47. Steel, M.C.; Sheppard, R.J.; Collin, R. Precision waveguide cells for the measurement of complex permittivity of lossy liquids and biological tissue at 35 GHz. J. Phys. E. **1987**, *20*, 872–877.
48. Harrington, R.F. *Time-Harmonic Electromagnetic Fields*; McGraw-Hill: New York, 1961.
49. Waldron, R.A. Theory of strip-line cavity for measurement of dielectric constants and gyromagnetic-resonance line-width. IEEE Trans. Microwave Theory Tech. **Jan. 1964**, *MTT-12*(1), 123–131.
50. Lakshminarayana, M.R.; Partain, L.D.; Cook, W.A. Simple microwave technique for independent measurement of sample size and dielectric constant with results for a gunn oscillator system. IEEE Trans. Microwave Theory Tech. **July 1979**, *MTT-27*(7), 661–665.
51. Parkash, A.; Vaid, J.K.; Mansingh, A. Measurement of dielectric parameters at microwave frequencies by cavity perturbation technique. IEEE Trans. Microwave Theory Tech. **Sept. 1979**, *MTT-27*(9), 791–795.
52. Jones, C.A.; Kantor, Y.; Grosvenor, J.H.; Janezic, M.D. *Striple Resonator for Electromagnetic Measurements of Materials*; NIST Technical Note 1505, National Institute of Standard and Technology: Boulder, Colorado, July 1998.
53. Jones, C.A. Permittivity and permeability measurements using strip-line resonator cavities—A comparison. IEEE Trans. Instrum. Meas. **Aug. 1999**, *IM-40*(4), 843–848.
54. Fenske K.; Misra, D. Dielectric materials at microwave frequencies. Applied Microwave & Wireless **Oct. 2000**, *12*(10), 92–100.
55. Land, D.V.; Campbell, A.M. A quick accurate method for measuring the microwave dielectric properties of small tissue samples. Phys. Med. Biol. **1992**. *37*(1), 183–192.
56. Carter, R.G. Accuracy of microwave cavity perturbation measurements. IEEE Trans. Microwave Theory Tech. **May 2001**, *MTT-49*(5), 918–923.
57. Tanabe, E.; Joines, W.T. A nondestructive method for measuring the complex permittivity of dielectric materials at microwave frequencies using an open transmission line resonator. IEEE Trans. Instrum. Meas. **Sept. 1976**, *IM-25*(3), 222–226.

58. Xu, D.; Liu, L.; Jiang, Z. Measurement of the dielectric properties of biological substances using and improved open-ended coaxial line resonator method. IEEE Trans. Microwave Theory Tech. **Dec. 1987**, *MTT-35*(12), 1424–1428.
59. Marcuvitz, N. *Waveguide Handbook*; McGraw-Hill: New York, 1951.
60. Jones, R.G. Precise dielectric measurements at 35 GHz using an open microwave resonator. Proc. IEE **Apr. 1976**, *123*(4), 285–290.
61. Clarke, R.N.; Rosenberg, C.B. Fabry-Perot and open resonators at microwave and millimeter wave frequencies, 2- 300 GHz. J. Phys. E.: Sci. Instrum. **1982**, *15*, 9–24.
62. Hirvonen, T.M.; Vainikainen, P.; Lozowski, A.; Raisanen, A. Measurement of dielectrics at 100 GHz with an open resonator connected to a network analyzer. IEEE Trans. Microwave Theory Tech. **Aug. 1996**, *MTT-45*(4), 780–786.
63. Afsar, M.N.; Huachi, X. An automated 60-GHz open resonator system for precision dielectric measurement. IEEE Trans. Microwave Theory Tech. **Dec. 1990**, *MTT-38*(12), 1845–1853.
64. Afsar, M.N.; Ding, H.; Tourshan, K. A new 60 GHz open-resonator technique for precision permittivity and loss-tangent measurement. IEEE Trans. Instrum. Meas. **Apr. 1999**, *IM-48*(2), 626–630.
65. Afsar, M.N.; Ding, H. A novel open-resonator system for precise measurement of permittivity and loss tangent. IEEE Trans. Instrum. Meas. **Apr. 2001**, *IM-50*(2), 402–405.
66. Stuchly, S.S.; Rzepecka, M.A.; Iskander, M.F. Permittivity measurements at microwave frequencies using lumped elements. IEEE Trans. Instrum. Meas. **Mar. 1975**, *IM-23*(1), 57–62.
67. Rzepecka, M.A.; Stuchly, S.S. A lumped element capacitance method for the measurement of the permittivity and conductivity in the frequency and time domain—A further analysis. IEEE Trans. Instrum. Meas. **Mar. 1974**, *IM-24*(1), 27–32.
68. Iskander, M.F.; Stuchly, S.S. Fringing field effect in lumped-capacitance method for permittivity measurement. IEEE Trans. Instrum. Meas. **Mar. 1978**, *IM-27*, 107–109.
69. Bianco, B.; Corana, L.; Gogioso, L.; Ridella, S.; Parodi, M. Open-circuited coaxial lines as standards for microwave measurements. Electron. Lett. **1980**, *16*(10), 373–374.
70. Kraszewski, A.; Stuchly, S.S.; Stuchly, M.A.; Symons, S. On measurement accuracy of the tissue permittivity in-vivo. IEEE Trans. Instrum. Meas. **Mar. 1983**, *IM-32*(1), 37–42.
71. Burdette, E.C.; Cain, F.L.; Seals, J. In vivo probe measurement technique for determining dielectric properties at VHF through microwave frequencies. IEEE Trans. Microwave Theory Tech. **Apr. 1980**, *MTT-28*(4), 414–424.
72. Burdette, E.C.; Cain, F.L.; Seals, J. In-situ tissue permittivity at microwave frequencies: perspective, techniques, results. In: *Medical Application of Microwave Imaging*; Larsen, L.E., Jacobi, J.H. Eds.; IEEE Press: New York, 1986, 13–40.
73. Mosig, J.R.; Besson, J.E.; Gex-Fabry, M.; Gardiol, F.E. Reflection of an open-ended coaxial line and application to nondestructive measurement of materials. IEEE Trans. Instrum. Meas. **March 1981**, *IM-30*(1), 46–51.
74. Athey, T.W.; Stuchly, M.A.; Stuchly, S.S. Measurement of radio frequency permittivity of biological tissues with an open-ended coaxial line: Part I. IEEE Trans. Microwave Theory Tech. **Jan. 1982**, *MTT-30*(1), 82–86.
75. Kraszewski, A.; Stuchly, M.A.; Stuchly, S.S. ANA calibration methods for measurement of dielectric properties. IEEE Trans. Instrum. Meas. **June 1983**, *IM-32*(2), 385–386.
76. Kraszewski A.; Stuchly, S.S. Capacitance of open-ended dielectric field coaxial lines—experimental results. IEEE Trans. Instrum. Meas. **Dec. 1983**, *IM-32*(4), 517–519.
77. Gajda, G.; Stuchly, S.S. An equivalent circuit of an open-ended coaxial line. IEEE Trans. Microwave Theory Tech. **May 1983**, *MTT-31*(5), 380–384.
78. Misra, D.K. A quasi-static analysis of open coaxial lines. IEEE Trans. Microwave Theory Tech. **1987**, *MTT-35*, 925–938.
79. Staebell, K.F.; Misra, D. An experimental technique for in vivo permittivity measurement of materials at microwave frequencies. IEEE Trans. Microwave Theory Tech. **March 1990**, *MTT-38*(3), 337–339.

80. Misra, D.M.; Chabbra, M.; Epstein, B.R.; Mirotznik, M.; Foster, K. R. Noninvasive electrical characterization of materials at microwave frequencies using an open-ended coaxial line: Test of an improved calibration technique. IEEE Trans. Microwave Theory Tech. **Jan 1990**, *MTT-38*(1), 8–13.
81. Nyshadham, A.; Sibbald, C.L.; Stuchly, S.S. Permittivity measurements using open-ended sensors and reference liquid calibration—An uncertainty analysis. IEEE Trans. Microwave Theory Tech. **Feb. 1992**, *MTT-40*(2), 305–314.
82. Stuchly, S.S.; Sibbald, C.L.; Anderson, J.M. A new aperture admittance model for open-ended waveguides. IEEE Trans. Microwave Theory Techn. **Feb. 1994**, *MTT-42*(2), 192–198.
83. Anderson, J.M.; Sibbald, C.L.; Stuchly, S.S. Dielectric measurements using a rational function model. IEEE Trans. Microwave Theory Techn. **Feb. 1994**, *MTT-42*(2), 199–204.
84. Baker-Jarvis, J.; Janezic, M.D.; Domich, P.D.; Geyer, R.G. Analysis of an open-ended coaxial probe with lift off for nondestructive testing. IEEE Trans. Instrum. Meas. **Oct. 1994**, *IM-43*(5), 711–718.
85. Gabriel, C.; Chan, T.Y.A.; Grant, E.H. Admittance models for open-ended coaxial probes and their place in dielectric spectroscopy. Phys. Med. Biol. **1994**, *39*, 2183–2199.
86. Okoniewski, O.; Anderson, J.A.; Okoniewska, E.; Gupta, K.; Stuchly, S.S. Further analysis of open-ended sensors. IEEE Trans. Microwave Theory Techn. **Aug. 1995**, *MTT-43*(8), 1986–1989.
87. Berube, D.; Ghannouchi, F.M.; Savard, P. A comparative study of four open-ended coaxial probe models for permittivity measurements of lossy dielectric/biological material at microwave frequencies. IEEE Trans. Microwave Theory Tech. **Oct. 1996**, *MTT-44*(10), 1928–1934.
88. Bao, J.; Lu, S.; Hurt, W. D. Complex dielectric measurements and analysis of brain tissues in the radio and microwave frequencies. IEEE Trans. Microwave Theory Tech. **Oct. 1997**, *MTT-45*(10), 1730–1740.
89. Hoshina, S.; Kanai, Y.; Miakawa, M. A numerical study of the measurement region of an open-ended coaxial probe used for complex permittivity measurement. IEEE Trans. Magnetics. **Sep. 2001**, *MTT-37*(5), 3311–3314.
90. Itoh, T. (Ed.). *Numerical Techniques for Microwave and Millimeter-Wave Passive Structures*, Wiley: New York, 1989.
91. Jin, J. *The Finite Element Method in Electromagnetics*; Wiley: New York, 1993.
92. Press, W.H. *Numerical Recipes in C*: *The Art of Scientific Computing*; Cambridge University Press: New York, 1992.
93. Friedsam, G.L.; Biebl, E.M. Precision free-space measurements of complex permittivity of polymers in the W-band. 1997 IEEE MTT-S International Microwave Symposium Digest, Denver, CO, **June 1997**, *3*, 1351–1354.
94. Afsar, M.N.; Tkachov, I.I.; Kocharyan, K.N. A novel W-band spectrometer for dielectric measurements. IEEE Trans. Microwave Theory Tech. **Dec. 2000**, *MTT-48*(12), 2637–2643.
95. Ghodgaonkar, D.K.; Varadan, V.V.; Varadan, V.K. Free-space measurement of complex permittivity and complex permeability of magnetic materials at microwave frequencies. IEEE Trans. Instrum. Meas. **Apr. 1990**, *IM-39*(2), 387–394.
96. Eagen, G.F.; Hoer, C.A. Thru-reflect-line: An improved technique for calibrating the dual six-port automatic network analyzer. IEEE Trans. Microwave Theory Tech. **Dec. 1979**, *MTT-27*(12), 987–993.
97. Parisot, M.; Soares, R. S parameter measurements and their use in circuit design. In *GaAs MESFET Circuit Design*; Soares, R. Ed.; Artech House: Norwood, MA, 1988, Chapter 3.
98. Soares, R.A.; Gouzien, P.; Legaud, P.; Follot, G. A unified approach to two-port calibration techniques and some applications. IEEE Trans. Microwave Theory Tech. **Nov. 1989**, *MTT-37*(11), 1669–1673.
99. Belhadj-Tahar, N.; Fourrier-Lamer, A.; Chanterac, H. Broadband simultaneous measurement of complex permittivity and permeability using a coaxial discontinuity. IEEE Trans. Microwave Theory Tech. **Jan. 1990**, *MTT-38*(1), 1–7.

100. Abdulnour, J.; Akyel, C.; Wu, K. A Generic approach for permittivity of dielectric materials using a discontinuity in a rectangular waveguide or a microstrip line. IEEE Trans. Microwave Theory Tech. **May 1995**, *MTT-43*(5), 1060–1066.
101. Queffelec, P.; Gelin, P. Influence of higher order modes on the measurements of complex permittivity and permeability of materials using a microstrip discontinuity. IEEE Trans. Microwave Theory Tech. **June 1996**, *MTT-44*(6), 814–824.
102. Abbas, Z.; Pollard, R.D.; Kelsall, R.W. Complex permittivity measurements at Ka-band using rectangular dielectric waveguide. IEEE Trans. Instrum. Meas. **Oct. 2001**, *IM-50*(5), 1334–1342.
103. Duhamel, F.; Huynen, I.; Vander Vorst, A. Measurements of complex permittivity of biological and organic liquids up to 110 GHz. 1997 IEEE MTT-S International Microwave Symposium Digest, Denver, CO, **June 1997**, *1*, 107–110.
104. Fossion, M.; Huynen, I.; Vanhoenacker, D.; Vander Vorst, A. A new and simple calibration method for measuring planar lines parameter up to 40 GHz. Proc. 22nd European Microwave Conference: Espoo, Finland, Aug. 1992, 180–185.
105. Tofighi, M.R. Design and Implementation of a Two-Port Microstrip Test Fixture for Complex Permittivity Characterization and Near-Field Imaging of Biological Materials up to 50 GHz, PhD Thesis, Drexel University: Philadelphia, PA, 2001.
106. Tofighi, M.R.; Daryoush, A.S. Characterization of biological tissues up to millimeter wave: Test fixture design, 2000 IEEE MTT-S International Microwave Symposium Digest, Boston, MA, **June 2000**, *2*, 1041–1044.
107. Tofighi, M.R.; Daryoush, A.S. Characterization of the complex permittivity of brain tissues up to 50 GHz utilizing a two-port microstrip test fixture. IEEE Trans. Microwave Theory Tech. **Oct. 2002**, *MTT-50*(10), 2217–2225.
108. Tofighi, M.R.; Daryoush, A.S. Comparison of two post-calibration correction methods for complex permittivity measurement of biological tissues up to 50 GHz. IEEE Trans. Instrum. Meas. **Dec. 2002**, *51*(6), 1170–1176.
109. Tofighi, M.R.; Daryoush, A.S. Study of the activity of neurological cell solutions using complex permittivity measurement, 2002 IEEE MTT-S International Microwave Symposium Digest, Seattle, WA, **June 2002**, *2*, 1763–1766.
110. Janezic, M.D.; Jargon, J.A. Complex permittivity determination from propagation constant measurements. IEEE Microwave Guided Wave Lett. **Feb. 1999**, *9*(2), 76–78.
111. Wan, C.; Nauwelaers, B.; De Raedt, W.; Van Rossum, M. Two new measurements methods for explicit determination of complex permittivity. IEEE Trans. Microwave Theory Tech. **Nov. 1998**, *MTT-46*(11), 1614–1619.
112. Marks, R.B. A Multiline method of network analyzer calibration. IEEE Trans. Microwave Theory Tech. **July 1991**, *MTT-39*(7), 1205–1215.
113. Nicolson, A.M.; Ross, G.F. Measurement of the intrinsic properties of materials by time-domain techniques. IEEE Trans. Instrum. Meas. **Nov. 1970**, *IM-19*(4), 377–382.
114. Courtney, C.C. Time-Domain measurement of the electromagnetic properties of materials. IEEE Trans. Microwave Theory Tech. **May 1998**, *MTT-46*(5), 517–522.
115. Jargon, J.; Janezic, M.D. Measuring complex permittivity and permeability using time-domain network analysis, 1996 IEEE MTT-S International Microwave Symposium Digest, San Francisco, CA, **June 2002**, *2*, 1407–1409.
116. Afsar, M.N. Dielectric measurements of millimeter-wave materials. IEEE Trans. Microwave Theory Tech. **Dec. 1984**, *MTT-32*(12), 1598–1609.
117. Arjavalingam, G.; Pastol, Y.; Halbout, J.; Kopcsay, G.V. Broad band microwave measurements with transient radiation from optoelectronically pulsed antennas. IEEE Trans. Microwave Theory Tech. **May 1990**, *MTT-38*(5), 615–621.
118. Cook, H.F. The dielectric behavior of some types of human tissues at microwave frequencies. Brit. J. Appl. Phys. **1951**, *2*, 295–300.
119. Kraszewski, A.; Stuchly, M.A.; Stuchly, S.S.; Smith, M. In vivo and in vitro dielectric properties of animal tissues at radio frequencies. Bioelectromagnetics **1982**, *3*, 421–432.

120. Stuchly M.A.; Stuchly, S.S. Dielectric properties of biological substances—tabulated. J. Microwave Power **1980**, *1*(15), 19–26.

121. Duck, F.A. *Physical Properties of Tissue*: *A Comprehensive Reference Book*; Academic Press: London, 1990.

122. Gabriel, C.; Gabriel, S.; Corthout, E. The dielectric properties of biological tissues: I. Literature survey. Phys. Med. Biol. **1996**, *41*, 2231–2249.

123. Gabriel, S.; Lau, R.W;. Gabriel, C. The dielectric properties of biological tissues: III. Parametric models for the dielectric spectrum of tissues. Phys. Med. Biol. **1996**, *41*, 2271–2293.

124. Malmberg C.G.; Maryott, A.A. Dielectric constant of water from 0 to 100°C. J. Res. Nat. Bureau Stand. **Jan. 1956**, *56*(1), 1–7.

125. Tables of Dielectric Dispersion Data for Pure Liquids and Dilute Solutions, National Bureau of Standards Circular 589, Nov. 1958.

126. Grant, E.H.; Buchanan, T.J.; Cook, H.F. Dielectric behavior of water at microwave frequencies. J. Chem. Phys. **1957**, *26*, 156–161.

127. Von Hippel, A. The Dielectric relaxation spectra of water, ice, and aqueous solutions, and their interpretation: 1. Critical survey of the status quo for water. IEEE Trans. Electrical Insulation, **Oct. 1988**, *23*(5), 801–816.

128. Von Hippel, A. The Dielectric relaxation spectra of water, ice, and aqueous solutions, and their interpretation: 2. Tentative interpretation of the relaxation spectrum of water in the time and frequency domain. IEEE Trans. Electric. Insulation **Oct. 1988**, *23*(5), 817–823.

129. Stogryn, A. Equations for calculating the dielectric constant of saline water. IEEE Trans. Microwave Theory Tech. **Aug. 1971**, *MTT-19*(8), 733–736.

130. Jordan, B.P.; Sheppard, R.J.; Szwarnowski, S. The Dielectric properties of formamide, ethanol, and methanol. J. Phy. D: Appl. Phys. **1978**, *11*, 695–701.

131. Model 37XXXC Vector Network Analyzer Operational Manual, ANRITSU P/N: 10410–00226, June 2000.

132. 37100C/37200C/37300C Vector Network Analyzers Technical Data Sheet, ANRITSU P/N: 11410–00247, June 2000.

133. Rehnmark, S. On the calibration process of the automatic network analyzer. IEEE Trans. Microwave Theory Tech. **Apr. 1974**, *MTT-22*, 457–458.

134. Taflove, A. *Computational Electrodynamics: The Finite Difference Time-Domain Method*, Artech House: Dedham, MA, **1995**.

135. Yee, K.S. Numerical Solution of initial boundary value problems involving Maxwell's equations in isotropic media. IEEE Trans. Antennas Propagat. **May 1966**, *AP-14*, 302–307.

Appendix A

Some Useful Constants

Permittivity of free space (ε_0) = 8.854×10^{-12} F/m
Permeability of free space (μ_0) = $4\pi \times 10^{-7}$ H/m
Speed of electromagnetic waves in free space (c) = 3×10^8 m/s
Impedance of free space (Z_0 or η_0) = 376.7 Ω
Boltzmann's constant (k) = 1.38×10^{-23} J/K
Charge of electron (e or q_e) = -1.602×10^{-19} C

Appendix B
Some Units and Conversions

Quantity	SI[a] unit	Conversion factor
Length	meter (m)	= 39.37 in
Mass	kilogram (kg)	= 2.21 pound-mass (lb_m)
Time	second (s)	
Frequency	hertz (Hz)	= 1 cycle/s
Force	newton (N)	= 0.2248 pound-force (lb_f)
Charge	coulomb (C)	
Charge density	coulomb/meter^3 (C/m^3)	
Current	ampere (A)	
Current density	ampere/ meter^2 (A/m^2)	
Electric field	volt/meter (V/m)	
Electric flux density	coulomb/meter^2 (C/m^2)	
Magnetic field	ampere/meter (A/m)	
Magnetic flux density	tesla (T)	= 10,000 G
	or weber/meter^2 (Wb/m^2)	= 10,000 G
Resistance	ohm (Ω)	
Conductivity	siemens/meter (S/m)	
	or mho/m	
Capacitance	farad (F)	
Permittivity	farad/meter (F/m)	
Inductance	henry (H)	
Permeability	henry/meter (H/m)	

[a] SI = International System of Units.

Appendix C
Review of Vector Analysis and Coordinate Systems

Since the formulation and application of various electromagnetic laws is greatly facilitated by the use of vector analysis, this appendix presents a concise review of vector analysis and the principal coordinate systems.

C.1. SCALARS AND VECTORS

A *scalar* quantity can be expressed as a single real number. (It can be positive, negative, or zero.) For example, voltage and current are scalar quantities. In ac analysis it is mathematically convenient to use *phasors* to represent sinusoidally varying voltages and currents. Phasors are referred to as *complex scalars*, since they require complex numbers (either magnitude and phase or real and imaginary parts) for their specification.

A *vector* quantity (e.g., the electric field) requires both a magnitude and a direction for its specification. The magnitude is always positive (it may be zero).

C.2. THE RECTANGULAR COORDINATE SYSTEM

The rectangular coordinate system (Fig. C.1a) locates a point P in three-dimensional space by assigning to it the coordinates (x_1, y_1, z_1) within a frame of reference defined by three mutually orthogonal (perpendicular) axes: the x axis, the y axis, and the z axis. It is conventional to choose a *right-handed* coordinate system (and we will do so throughout this handbook). This choice simply means that if we first point the fingers of the right hand along the x axis and then curl them to point along the y axis, the extended thumb will align with the z axis.

To deal with vectors, we define a set of three unit vectors $\mathbf{a}_x$, $\mathbf{a}_y$, and $\mathbf{a}_z$ (each with a magnitude equal to one) aligned with (parallel to) the three axes. An arbitrary vector $\mathbf{A}$ may now be expressed as $\mathbf{A} = A_x\ \mathbf{a}_x + A_y\ \mathbf{a}_y + A_z\ \mathbf{a}_z$, where A_x, A_y, A_z are said to be its scalar components along the three axes. The vector $\mathbf{A}$ has a magnitude $A = [A_x^2 + A_y^2 + A_z^2]^{1/2}$. Figure C.1b shows a differential volume $dV = dxdydz$. The surfaces have differential areas, ds, of $dx\,dy$, $dy\,dz$, and $dz\,dx$.

C.3. SCALAR AND VECTOR FIELDS

The concepts of scalars and vectors introduced in Sec. C.1 can be extended to define scalar and vector fields. A scalar field associates a scalar quantity with every point in a

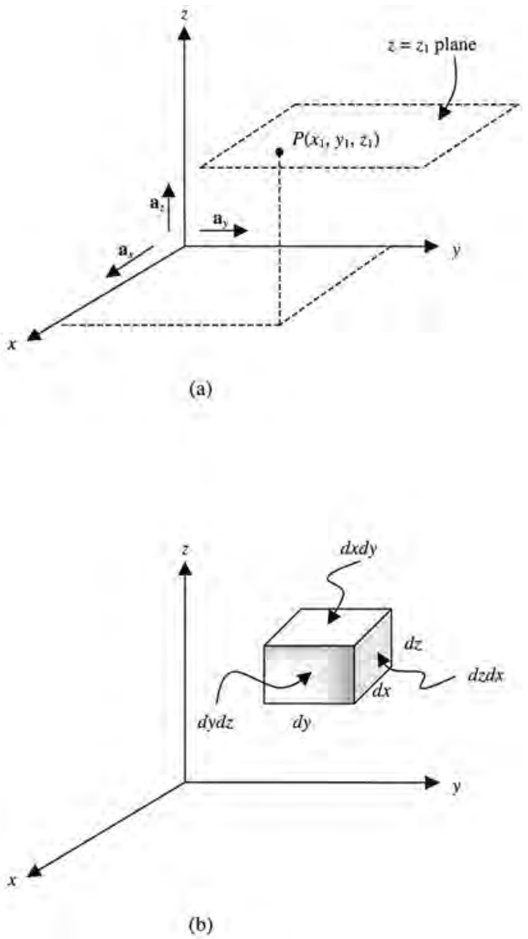


Figure C.1 The rectangular coordinate system: (a) the coordinates of a point and the unit vectors and (b) differential elements.

region of space. If we set up a rectangular coordinate system to identify various points in the 3D space in a room, we may describe the temperature distribution (scalar field) as some function $T = f(x, y, z)$ so that at the point (x_1, y_1, z_1) the temperature $T(x_1, y_1, z_1)$ is given by the value of the function $f(x_1, y_1, z_1)$. In a similar fashion, if we associate a vector with every point in a region, we will have a vector field. In the rectangular coordinate system, we can write a vector field in terms of its three components, each of which is a scalar field. For example, the velocity distribution in a river may be expressed as $\mathbf{v} = v_x(x, y, z)\mathbf{a}_x + v_y(x, y, z)\mathbf{a}_y + v_z(x, y, z)\mathbf{a}_z$.

C.4. VECTOR ADDITION AND SUBTRACTION

Two vectors **A** and **B** may be added together graphically by the familiar *parallelogram rule* shown in Fig. C.2. The addition can also be performed by adding the corresponding components of the two vectors.

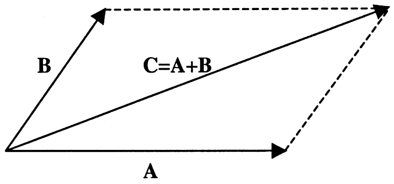


Figure C.2 Vector addition via the parallelogram rule.

If $\mathbf{A} = A_x\mathbf{a}_x + A_y\mathbf{a}_y + A_z\mathbf{a}_z$ and $\mathbf{B} = B_x\mathbf{a}_x + B_y\mathbf{a}_y + B_z\mathbf{a}_z$, their sum is a vector **C**, given as

$$\mathbf{C} = \mathbf{A} + \mathbf{B} = (A_x + B_x)\mathbf{a}_x + (A_y + B_y)\mathbf{a}_\mathrm{y} + (A_z + B_z)\mathbf{a}_z$$

Vector addition always obeys the following laws:

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \qquad \text{(commutative)}$$

$$\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C} \qquad \text{(associative)}$$

Vector subtraction, $\mathbf{A} - \mathbf{B}$, is accomplished by reversing the direction of **B** to obtain another vector $-\mathbf{B}$ and then adding it to the vector **A**. Thus we have

$$\mathbf{D} = \mathbf{A} - \mathbf{B} = \mathbf{A} + (- \mathbf{B})$$

or

$$\mathbf{D} = (A_x - B_x)\mathbf{a}_x + (A_y - B_y)\mathbf{a}_y + (A_z - B_z)\mathbf{a}_z$$

where **A** and **B** have been expressed in terms of their rectangular components.

In dealing with vector fields, it is important to realize that we should be adding and subtracting only those vectors that are defined at the same point in space.

C.5. POSITION AND DISTANCE VECTORS

The position vector associated with a point P, which has the rectangular coordinates (x_1, y_1, z_1), is the vector extending from the origin O(0, 0, 0) to the point P. It may be expressed as (Fig. C.3)

$$\mathbf{OP} = x_1\mathbf{a}_\mathrm{x} + y_1\mathbf{a}_\mathrm{y} + z_1\mathbf{a}_\mathrm{z} \tag{C.1}$$

The distance vector **PQ** extends from the point $P(x_1, y_1, z_1)$ to the point $Q(x_2, y_2, z_2)$ and can be expressed as

$$\begin{aligned}\mathbf{PQ} &= \mathbf{OQ} - \mathbf{OP}\\ &= (x_2\mathbf{a}_x + y_2\mathbf{a}_y + z_2\mathbf{a}_z) - (x_1\mathbf{a}_x + y_1\mathbf{a}_y + z_1\mathbf{a}_z)\\ &= (x_2 - x_1)\mathbf{a}_x + (y_2 - y_1)\mathbf{a}_y + (z_2 - z_1)\mathbf{a}_z\end{aligned} \tag{C.2}$$

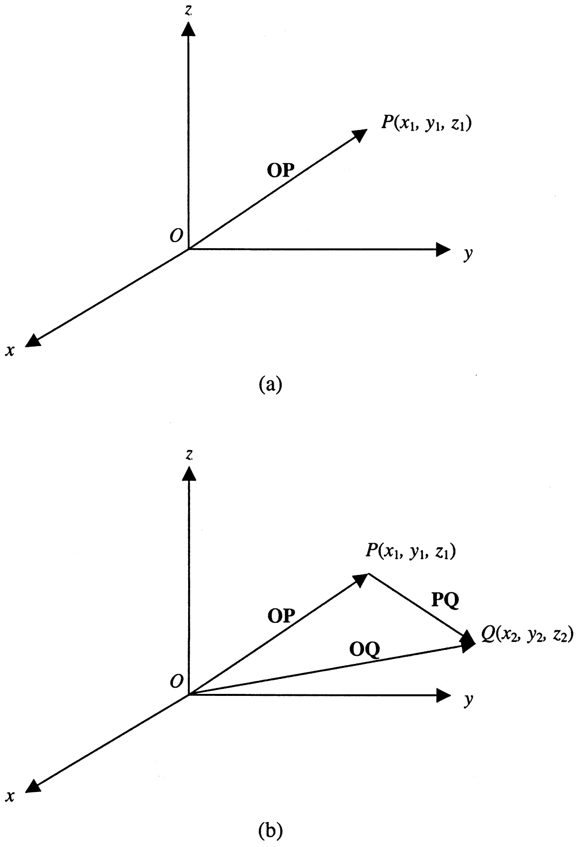


Figure C.3 (a) The position vector **OP** extends from the origin O to the point P. (b) The distance vector **PQ** extends from P to Q.

The scalar distance PQ is given by the magnitude of the vector **PQ**. Thus,

$$PQ = |\mathbf{PQ}| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \tag{C.3}$$

C.6. VECTOR DIVISION AND MULTIPLICATION

The operation **A**/**B** is *not* defined. However, a vector can be divided by a scalar.

Two forms of vector to vector multiplication are useful in our work.

C.6.1. Scalar (Dot) Product

The scalar product of the two vectors **A** and **B** is represented symbolically as **A** • **B** (hence the alternate name, the *dot product*).

$$\mathbf{A} \bullet \mathbf{B} = |\mathbf{A}||\mathbf{B}| \cos\theta_{AB} \tag{C.4}$$

where θ_{AB} is the smaller angle between **A** and **B**. Also,

$$\mathbf{A} \bullet \mathbf{B} = A_x B_x + A_y B_y + A_z B_z \tag{C.5}$$

The scalar product is *commutative*, i.e., $\mathbf{A} \bullet \mathbf{B} = \mathbf{B} \bullet \mathbf{A}$. Also note that

$$\mathbf{A} \bullet \mathbf{a}_x = A_x$$

$$\mathbf{A} \bullet \mathbf{a}_y = A_y$$

$$\mathbf{A} \bullet \mathbf{a}_z = A_z$$

C.6.2. Vector (Cross) Product

The vector product between **A** and **B** is a vector represented as $\mathbf{A} \times \mathbf{B}$ and is given by

$$\mathbf{A} \times \mathbf{B} = |\mathbf{A}||\mathbf{B}||\sin\theta_{AB}|\mathbf{a}_n \tag{C.6}$$

where θ_{AB} is the smaller angle between **A** and **B**, and $\mathbf{a}_n$ is a unit vector normal to the plane containing **A** and **B**. (Since each plane has two normal vectors, it is important to note that $\mathbf{a}_n$ is the one obtained by the right-hand rule. If the fingers of the right hand are extended in the direction of **A** and then curled towards vector **B**, the direction of the outstretched thumb is the direction of $\mathbf{a}_n$.)

Also,

$$\mathbf{A} \times \mathbf{B} = (A_y B_z - A_z B_y)\mathbf{a}_x + (A_z B_x - A_x B_z)\mathbf{a}_y + (A_x B_y - A_y B_x)\mathbf{a}_z \tag{C.7}$$

C.7. The Cylindrical Coordinate System

While much of our work is carried out conveniently in the familiar rectangular coordinate system (introduced in Sec. C.2), some physical situations have a natural symmetry which makes the cylindrical coordinate system easier to use. Examples include a coaxial cable and an optical fiber.

The cylindrical coordinate system we will use is a natural extension of the two dimensional (xy plane) polar coordinates (ρ,ϕ) to three dimensions (ρ,ϕ,z). Figure C.4 shows the geometric relationship between the rectangular coordinates (x, y, z) of a point P and its cylindrical coordinates (ρ,ϕ,z). You will notice that the z coordinate is common to both systems, while ρ and ϕ are related to x and y as follows

$$\rho = +\sqrt{x^2 + y^2} \qquad \phi = \tan^{-1}\left(\frac{y}{x}\right) \tag{C.8}$$

$$x = \rho\cos\phi \qquad y = \rho\sin\phi, \tag{C.9}$$

where ρ may be thought of as the "*horizontal*" radial distance from the origin to the point P, while ϕ measures the "*azimuthal*" angle from the x axis in a counterclockwise (toward the y axis) direction.

Just as, in the rectangular coordinate system, a point $P(x_1, y_1, z_1)$ corresponds to the intersection of the three mutually orthogonal *planar* surfaces: $x = x_1$, $y = y_1$, $z = z_1$,

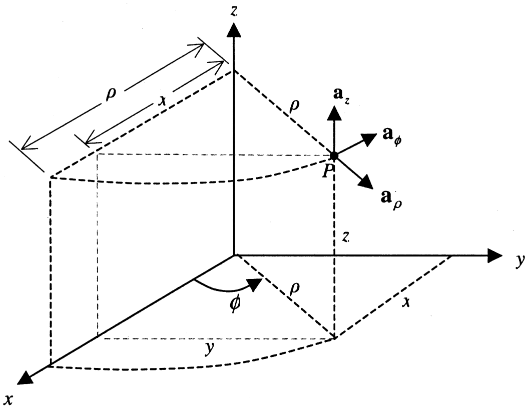


Figure C.4 The cylindrical coordinate system.

in the cylindrical system, $P(\rho_1,\phi_1,z_1)$ is located at the intersection of the three orthogonal surfaces:

$\rho = \rho_1$ (cylinder)

$\phi = \phi_1$ (plane)

$z = z_1$ (plane)

A vector **A** may be expressed in the cylindrical system as

$$\mathbf{A} = A_\rho \mathbf{a}_\rho + A_\phi \mathbf{a}_\phi + A_z \mathbf{a}_z \tag{C.10}$$

with $|\mathbf{A}| = (A_\rho^2 + A_\phi^2 + A_z^2)^{1/2}$, where $\mathbf{a}_\rho$, $\mathbf{a}_\phi$, and $\mathbf{a}_z$, are mutually orthogonal unit vectors as shown in Fig. C.5. $\mathbf{a}_\rho$ points in the direction of increasing "horizontal" radial distance ρ, $\mathbf{a}_\phi$ also lies in a "horizontal" plane (parallel to the xy plane) and points in the direction of increasing ϕ, and finally $\mathbf{a}_z$ is parallel to the positive z axis (as before). Also note the right-hand rule relationship among $\mathbf{a}_\rho$, $\mathbf{a}_\phi$, and $\mathbf{a}_z$, i.e.,

$$\begin{aligned} \mathbf{a}_\rho \times \mathbf{a}_\phi &= \mathbf{a}_z \\ \mathbf{a}_\phi \times \mathbf{a}_z &= \mathbf{a}_\rho \\ \mathbf{a}_z \times \mathbf{a}_\rho &= \mathbf{a}_\phi \end{aligned} \tag{C.11}$$

A differential volume element dV in the cylindrical coordinate system is

$$dV = (d\rho)(\rho d\phi)(dz) \tag{C.12}$$

Note that $\rho d\phi$ (and not $d\phi$ by itself) represents an incremental distance in the direction of $\mathbf{a}_\phi$, since $d\phi$ is a dimensionless angular measure.

C.8. THE SPHERICAL COORDINATE SYSTEM

If one wishes to analyze the scattering of microwave radar signals from raindrops or the electromagnetic interaction between a cell phone and the human head, the spherical

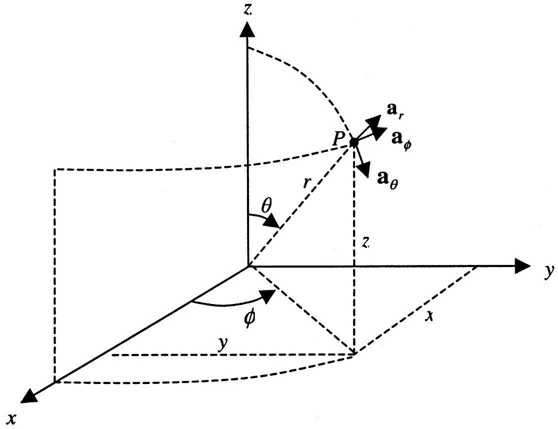


Figure C.5 The spherical coordinate system.

coordinate system may facilitate setting up the mathematical problem. In the spherical coordinate system (Fig. C.5), a point $P(r_1,\theta_1,\phi_1)$ lies at the intersection of the three mutually orthogonal surfaces

$$r = r_1 \qquad \text{(sphere)}$$

$$\theta = \theta_1 \qquad \text{(cone)}$$

$$\phi = \phi_1 \qquad \text{(plane)}$$

r represents the three-dimensional distance between the origin and the point, θ $(0 \leq \theta < \pi)$ is the "*elevation*" angle measured from the positive z axis, and ϕ $(0 \leq \phi < 2\pi)$ is the "*azimuthal*" angle measured from the positive x axis (as in the cylindrical coordinate system). They are related to (x, y, z) as follows:

$$r = [x^2 + y^2 + z^2]^{1/2} \qquad \theta = \cos^{-1}\left(\frac{z}{\sqrt{x^2 + y^2 + z^2}}\right) \qquad \phi = \tan^{-1}\left(\frac{y}{x}\right) \tag{C.13}$$

$$x = r \sin\theta \cos\phi \qquad y = r \sin\theta \sin\phi \qquad z = r \sin\theta \tag{C.14}$$

A vector **A** in the spherical coordinate system is expressed as

$$\mathbf{A} = A_r\mathbf{a}_r + A_\theta\mathbf{a}_\theta + A_\phi\mathbf{a}_\phi \tag{C.15}$$

with $|\mathbf{A}| = (A_r^2 + A_\theta^2 + A_\phi^2)^{1/2}$.

The unit vectors $\mathbf{a}_r$, $\mathbf{a}_\theta$, and $\mathbf{a}_\phi$, are mutually orthogonal and follow the right-hand rule relationship embedded in

$$\begin{aligned} \mathbf{a}_r \times \mathbf{a}_\theta &= \mathbf{a}_\phi \\ \mathbf{a}_\theta \times \mathbf{a}_\phi &= \mathbf{a}_r \\ \mathbf{a}_\phi \times \mathbf{a}_r &= \mathbf{a}_\theta \end{aligned} \tag{C.16}$$

A differential volume element dV is written as

$$dV = (dr)(r\,d\theta)(r\sin\theta\,d\phi) = r^2\sin\theta\,dr\,d\theta\,d\phi \tag{C.17}$$

C.9. COORDINATE AND VECTOR TRANSFORMATION

In working with the various coordinate systems, we may need to convert parameters given in one coordinate system into parameters in another coordinate system.

Vector Interconversion Strategy

The most common conversions in practice are those between rectangular and cylindrical coordinate systems and those between rectangular and spherical coordinate systems. Both types can be accomplished easily with the help of Table C.1.

Example C.1. Vector transformation (rectangular to cylindrical): Convert $\mathbf{A} = z\mathbf{a}_x + x\mathbf{a}_y$ to cylindrical coordinates.

We start by writing $\mathbf{A} = A_\rho\mathbf{a}_\rho + A_\phi\mathbf{a}_\phi + A_z\mathbf{a}_z$. Then

$$\begin{aligned}
A_\rho &= \mathbf{A} \bullet \mathbf{a}_\rho \\
&= (z\mathbf{a}_x + x\mathbf{a}_y) \bullet \mathbf{a}_\rho \\
&= z(\mathbf{a}_x \bullet \mathbf{a}_\rho) + x(\mathbf{a}_y \bullet \mathbf{a}_\rho) \\
&= z\cos\phi + (\rho\cos\phi)(\sin\phi) \\
A_\phi &= \mathbf{A} \bullet \mathbf{a}_\phi \\
&= (z\mathbf{a}_x + x\mathbf{a}_y) \bullet \mathbf{a}_\phi \\
&= z(\mathbf{a}_x \bullet \mathbf{a}_\phi) + x(\mathbf{a}_y \bullet \mathbf{a}_\phi) \\
&= z(-\sin\phi) + (\rho\cos\phi)(\cos\phi)
\end{aligned}$$

Table C.1 Unit Vector Transformation

(a) Dot products of unit vectors (rectangular/cylindrical)

	$\mathbf{a}_\rho$	$\mathbf{a}_\phi$	$\mathbf{a}_z$
$\mathbf{a}_x$	$\cos\phi$	$-\sin\phi$	0
$\mathbf{a}_y$	$\sin\phi$	$\cos\phi$	0
$\mathbf{a}_z$	0	0	1

(b) Dot products of unit vectors (rectangular/spherical)

	$\mathbf{a}_r$	$\mathbf{a}_\theta$	$\mathbf{a}_\phi$
$\mathbf{a}_x$	$\sin\theta\cos\phi$	$\cos\theta\cos\phi$	$-\sin\phi$
$\mathbf{a}_y$	$\sin\theta\sin\phi$	$\cos\theta\sin\phi$	$\cos\phi$
$\mathbf{a}_z$	$\cos\theta$	$-\sin\theta$	0

$$\begin{aligned} A_z &= \mathbf{A} \bullet \mathbf{a}_z \\ &= (z\mathbf{a}_x + x\mathbf{a}_y) \bullet \mathbf{a}_z \\ &= z(\mathbf{a}_x \bullet \mathbf{a}_z) + x(\mathbf{a}_y \bullet \mathbf{a}_z) \\ &= z \cdot 0 + x \cdot 0 \end{aligned}$$

Therefore,

$$\mathbf{A} = (z\cos\phi + \rho\cos\phi\sin\phi)\mathbf{a}_\rho + (-z\sin\phi + \rho\cos^2\phi)\mathbf{a}_\phi$$

Example C.2. Vector transformation (spherical to rectangular): Convert $\mathbf{E} = E_{\text{o}}/(r^2)\ \mathbf{a}_r$ into rectangular coordinates.

We start by writing $\mathbf{E} = E_x\mathbf{a}_x + E_y\mathbf{a}_y + E_z\mathbf{a}_z$. Then,

$$\begin{aligned} E_x = \mathbf{E} \bullet \mathbf{a}_x = \left(\frac{E_o}{r^2}\right)(\mathbf{a}_r \bullet \mathbf{a}_x) = \frac{E_o}{r^2}\sin\theta\cos\phi = \frac{E_o}{r^2}\frac{r\sin\theta\cos\phi}{r} = \frac{E_o}{r^3}r\sin\theta\cos\phi \\ = \frac{E_0 x}{(x^2+y^2+z^2)^{3/2}} \end{aligned}$$

Similarly,

$$E_y = \frac{E_o y}{(x^2+y^2+z^2)^{3/2}} \qquad \text{and} \qquad E_z = \frac{E_o z}{(x^2+y^2+z^2)^{3/2}}$$

C.10. VECTOR DIFFERENTIAL OPERATORS

Maxwell equations and the associated relationships are expressed in terms of vector differential operators. Therefore, we have tabulated the expressions for the various differential operators in all the coordinate systems below:

C.10.1. Divergence

Rectangular

$$\nabla \bullet \mathbf{D} = \frac{\partial D_x}{\partial x} + \frac{\partial D_y}{\partial y} + \frac{\partial D_z}{\partial z}$$

Cylindrical

$$\nabla \bullet \mathbf{D} = \frac{1}{\rho}\frac{\partial}{\partial \rho}\left(\rho D_\rho\right) + \frac{1}{\rho}\frac{\partial D_\phi}{\partial \phi} + \frac{\partial D_z}{\partial z}$$

Spherical

$$\nabla \bullet \mathbf{D} = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 D_r\right) + \frac{1}{r\sin\theta}\frac{\partial}{\partial \theta}(D_\theta \sin\theta) + \frac{1}{r\sin\theta}\frac{\partial D_\phi}{\partial \phi}$$

C.10.2. Gradient

Rectangular

$$\nabla V = \frac{\partial V}{\partial x}\mathbf{a}_x + \frac{\partial V}{\partial y}\mathbf{a}_y + \frac{\partial V}{\partial z}\mathbf{a}_z$$

Cylindrical

$$\nabla V = \frac{\partial V}{\partial \rho}\mathbf{a}_\rho + \frac{1}{\rho}\frac{\partial V}{\partial \phi}\mathbf{a}_\phi + \frac{\partial V}{\partial z}\mathbf{a}_z$$

Spherical

$$\nabla V = \frac{\partial V}{\partial r}\mathbf{a}_r + \frac{1}{r}\frac{\partial V}{\partial \theta}\mathbf{a}_\theta + \frac{1}{r\sin\theta}\frac{\partial V}{\partial \phi}\mathbf{a}_\phi$$

C.10.3. Curl

Rectangular

$$\nabla \times \mathbf{E} = \left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}\right)\mathbf{a}_x + \left(\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x}\right)\mathbf{a}_y + \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right)\mathbf{a}_z$$

Cylindrical

$$\nabla \times \mathbf{E} = \left(\frac{1}{\rho}\frac{\partial E_z}{\partial \phi} - \frac{\partial E_\phi}{\partial z}\right)\mathbf{a}_\rho + \left(\frac{\partial E_\rho}{\partial z} - \frac{\partial E_z}{\partial \rho}\right)\mathbf{a}_\phi + \frac{1}{\rho}\left(\frac{\partial}{\partial \rho}\rho E_\phi - \frac{\partial E_\rho}{\partial \phi}\right)\mathbf{a}_z$$

Spherical

$$\begin{aligned}\nabla \times \mathbf{E} = {} & \frac{1}{r\sin\theta}\left[\frac{\partial}{\partial \theta}\left(E_\phi \sin\theta\right) - \frac{\partial E_\theta}{\partial \phi}\right]\mathbf{a}_r + \frac{1}{r}\left(\frac{1}{\sin\theta}\frac{\partial E_r}{\partial \phi} - \frac{\partial}{\partial r}rE_\phi\right)\mathbf{a}_\theta \\ & + \frac{1}{r}\left(\frac{\partial}{\partial r}rE_\theta - \frac{\partial E_r}{\partial \theta}\right)\mathbf{a}_\phi\end{aligned}$$