Engineering Electromagnetics

Kenneth R. Demarest

ENGINEERING ELECTROMAGNETICS

Kenneth R. Demarest
Professor of Electrical Engineering
The University of Kansas

Prentice Hall
Upper Saddle River, New Jersey 07458

Library of Congress Cataloging-in-Publication Data

Demarest, Kenneth Robert.
Engineering electromagnetics / by Kenneth Demarest.
p.   cm.
Includes bibliographical references and index.
ISBN 0-02-328521-4
1. Electromagnetism.   2. Electric engineering.   I. Title.
QC760.D39   1998
621.3—dc21   97-26067
CIP

Acquisitions editor: *Eric Svendsen*
Editorial/production supervisor: *Rose Kernan*
Editor-in-chief: *Marcia Horton*
Managing editor: *Bayani Mendoza de Leon*
Director of production and manufacturing: *David W. Riccardi*
Manufacturing buyer: *Julia Meehan*
Cover designer: *Bruce Kenselaar*
Editorial assistant: *Andrea Au*
Art manager: *Gus Vibal*
Composition: *Preparé / Emilcomp*

The author and publisher of this book have used their best efforts in preparing this book. These efforts include the development, research, and testing of the theories and programs to determine their effectiveness. The author and publisher make no warranty of any kind, expressed or implied, with regard to these programs or the documentation contained in this book. The author and publisher shall not be liable in any event for incidental or consequential damages in connection with, or arising out of, the furnishing, performance, or use of these programs.

Printed in the United States of America

10 9 8 7 6 5 4 3 2

ISBN 0-02-328521-4

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada Inc., *Toronto*
Prentice-Hall Hispanoamericana, S.A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Simon & Schuster Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

Contents

Preface

I. Why I wrote this text

This book is about my favorite academic subject: electromagnetics.

The genesis of this book goes back to a student get-together I attended during the third year of my teaching career. I struck up a conversation with a good student who had taken an introductory electromagnetics course from me the previous semester. When I asked him what he remembered from the course, he very politely told me that he really couldn't remember much, except that it involved a lot of mathematics and theory. Worst of all, he seemed unaware of any way in which the material in that course would apply to his career as an electrical engineer. Apparently, he had gotten the message that, outside of "niche" areas like antennas, electromagnetics had little to do with the practice of electrical engineering.

This conversation caused me to do some serious soul searching about the way I had been teaching electromagnetics. In the end, I concluded that this student had been largely correct. While I had very accurately presented the laws of electromagnetics, I made no particular effort to integrate the material into the larger context of electrical engineering. My class in electromagnetics was probably not too different from the one taught across campus in the physics department. Given this, many of my students were convinced that electromagnetics was relevant to their careers in electrical engineering only because they needed to pass the course in order to obtain a degree.

In reviewing a selection of electromagnetics textbooks intended for electrical engineers, I realized that the problem was more than just my own teaching style. Upon closer examination, I concluded that most of these textbooks could be divided into two classes. Textbooks in the first class tended to integrate electromagnetics nicely into engineering practice, but were often technically weak. On the other hand, textbooks in the second class typically gave solid, mathematically oriented discussions of electromagnetic theory, but little or no physical insight or engineering application.

As my teaching skills matured, I gradually found ways to convince my students that electromagnetic theory connects with all aspects of electrical engineering, while still approaching the material with the rigor needed for solving real problems. For instance, I found that students are more apt to see the relevance of electromagnetics when I remind them that the job of an electrical engineer is to constrain electric and magnetic fields so that they perform a desired task. This way of thinking is just as applicable for computer circuit designers as it is for antenna and microwave engineers.

I also found that topics that motivated the past generation of students to think seriously about electromagnetics no longer have the same resonance with the current generation. Topics like antennas and radar, while still extremely important in engineering practice, don't touch these students lives as much as computers and consumer electronics do. Fortunately, there are a whole range of new topics that work just as well as the old ones. As an example, students quickly take notice when it is pointed out to them that the motherboard of their PC, while a digital system, is in many ways a microwave network. Even better is to show them how electrical interference between digital networks is a major safety and economic issue that can only be solved using electromagnetic techniques.

This textbook represents my vision of how to present electromagnetics to undergraduate students - one that emphasizes the physical processes and applications of electromagnetics in engineering, while at the same time presenting the material with the rigor necessary to approach real problems in engineering practice.

II. Features of This Text

Once I started writing, I discovered that a project of this magnitude appears to take on a life of its own. Although my goals for the text remained fixed, the features of the book went through a metamorphosis as the years of writing and testing in the classroom unfolded. What finally grew out of this process was a text that I feel is readable and understandable for the student, easy to integrate in the classroom, and, possibly most important, displays electromagnetics as a mainstream subject in electrical engineering.

1. This text presents electromagnetics using what most would consider a modified "traditional" or "historical" approach. After a brief overview (see point 3), static electric and magnetic fields are presented separately, followed by the time-varying case. This approach provides the greatest insight into the physical relevance of electromagnetics and allows students to grapple with electric and magnetic fields separately, before tackling the more difficult time-varying case.

2. Whenever possible, concepts and results are presented both mathematically and in plain English. My experience has shown that instructors usually prefer purely mathematical developments, but students like (and often need) word descriptions to give meaning to these developments. In addition, these physical descriptions help develop the student's engineering intuition. This is important for the students transition into engineering practice, where the problems don't look like textbook problems.
3. Chapter 3, "Sources, Forces, and Fields" provides a broad overview of the experiments and theories that led up to the definitive equations of electromagnetics—Maxwell's equations. This chapter is designed to show the basic connections between electromagnetic sources and the fields and forces they produce. The physical laws are presented in their historical order, culminating in Maxwell's equations.
4. Chapters 4–6 and 7–9 present electric and magnetic fields in parallel fashion. Each sequence starts with fields in free space, followed by material effects, and finally energy relations. This parallel construction allows students to see clearly what is similar and dissimilar about low-frequency electric and magnetic fields.
5. Graphical solution techniques are presented for both low frequency electric and magnetic fields. These techniques are useful in gaining insight into many practical problems and are also fun.
6. Unlike most traditional texts, this text presents transmission lines *before* plane waves. I have done this for two reasons. The first is that students find it easier to make the transition into propagating waves by first considering scalar voltage and current waves. Once these scalar waves are mastered, the jump into the more general case of space waves is easier. The second reason is that this choice makes it easier to introduce network-analyzer based laboratory experiments in conjunction with the last 4 chapters (transmission lines, plane waves, waveguides, and antennas and radiation).
7. Chapter 11, "Transmission Lines," covers both time-varying and frequency domain analysis. Thévenin equivalent concepts are used throughout this chapter to clearly show the student how electromagnetic and circuit theory are complementary. A great effort was made to make the time-domain section relevant to students whose primary interests are digital and computer circuits. Advanced topics in this section include microstrip transmission lines, reactive and nonlinear loads, and rise-time calculations.
8. This text covers a number of topics associated with electric and magnetic shielding, electromagnetic compatibility (EMC), and electromagnetic interference (EMI). These topics are discussed in both the static and dynamic sections of the text. My experience has shown that students very quickly see the relation between these topics and consumer electronics.
9. Chapter 13, "Waveguides," discusses both copper-based and dielectric based waveguides. This chapter includes a sizable discussion of optical fibers and systems.
10. In addition to a fundamental discussion of radiation and antenna principles, Chapter 14, "Antennas and Radiation," also presents a broad overview of the major classes of antennas encountered in engineering practice.

III. How to Use This Text

This textbook is intended for junior level electrical engineering students and can be used in either a one- or two-semester format. The following table gives suggested schedules for one- and two-semester courses.

Chapter	Title	2-Semester Lecture Hours	1-Semester Lecture Hours
1	Introduction	1	1
2	Vector Analysis	2	2
3	Electromagnetic Sources, Forces, and Fields	4	2
4	Electrostatic Fields in Free Space	5	5
5	Electrostatic Fields in Material Media	9	5
6	Capacitance and Electric Energy	3	3
7	Magnetostatic Fields in Free Space	4	4
8	Magnetostatic Fields in Material Media	5	3
9	Magnetic Inductance, Energy, and Forces	5	4
10	Time-Varying Electromagnetic Fields	4	4
11	Transmission Lines	14	6
12	Plane Waves	10	3
13	Waveguides	8	
14	Radiation and Antennas	10	
	Totals	**84**	**42**

These schedules assume that students have had some previous experience with vector calculus, presumably from their calculus sequence. When this is the case, I have found that two lectures on the material in chapter 2 is a sufficient review, and students typically refer back to chapter 2 throughout the course to review various aspects of this material. For curricula where vector calculus has not yet been encountered, a longer exposure to chapter 2 would provide students with all the necessary background.

For a two-semester course sequence, I have found that there is ample time to cover nearly all the material in this text. For a one-semester course, the suggested schedule provides a more concentrated coverage of the electrostatic and magnetostatic topics, while still allowing time for 13 lectures on the most important time-varying topics.

IV. Acknowledgments

In writing this book, I have drawn from every instructor and textbook that I have encountered throughout my career as a student and a professor. Without them, I could not possibly have written this textbook. Even though the writing style that emerged is uniquely my own, I am indebted to all of them.

My greatest thanks go the students at the University of Kansas who endured the many drafts of this text in class. I was genuinely surprised by the enthusiasm they displayed for being a part of this writing process. Their comments were very insightful, and I made many changes in the manuscript in response to their comments. Although

space does not allow me to recognize all the students who made significant comments, I could not rest without recognizing Mr. Scott Filion, who provided me with unbelievably detailed and helpful comments on the first half of the book.

I want to thank the reviewers who reviewed the manuscript of this text at various stages: Dr. Paul O. Berrett (Professor Emeritus) - Brigham Young University, Paul R. Melsaac - Cornell University, John R. Cogdell - University of Texas at Austin, Dr. Stuart A. Long - University of Houston, Walter J. Gajda, Jr. - University of Missouri - Rolla, Dennis P. Nyquist - Michigan State University, Kai Chang - Texas A & M University, Warren L. Stutzman - Virginia Polytechnic Institute and State University, Markus Zahn - Massachusetts Institute of Technology, Dr. David Rogers - North Dakota State University, Bruce Mc Leod - Montana State University, Jeffrey P. Mills - Illinois Institute of Technology, Robert York - University of California, Santa Barbara, Robert C. Owens - Santa Clara University, Vladimir Mitin - Wayne State University, Steven Scott Gearhart - University of Wisconsin - Madison. Their many comments caused me to rethink many aspects of the book and point me in the right direction. I appreciated their honesty and directness in telling me exactly what they did and did not like about the various manuscripts.

I also want to thank a number of colleagues and friends at the University of Kansas who assisted me in various ways during the writing of this text. Included in these are: Professors James Roberts, James Rowland and K. Sam Shanmugan, who gave me encouragement throughout this project, and Ms. Donnis Graham, who assisted in the final manuscript revisions. I would also like to thank Professor Ruth Miller at Kansas State University for several helpful discussions about teaching philosophies and electromagnetics.

Finally, I am most thankful to my wife Susan and my children, Eric and Rebecca, for standing by me during the over seven years it took to complete this text. Their concern and compassion for me during this time was greatly appreciated, and I could not have endured the process without them. I only hope I can follow through with my promise that life in the Demarest house will be easier now that this text is finally finished!

Kenneth Demarest
University of Kansas

1

Background and Motivation

1-1 Introduction

Electromagnetics is both the oldest and most basic of all the branches of electrical engineering. Stated simply, electromagnetics deals with four questions: What is electricity, how does it behave, what can it do, and how can we control it? So fundamental are the questions that it addresses that it is not an exaggeration to say that electromagnetics is at the heart of *everything* that is done with electricity. As a result, an understanding of electromagnetics is essential in order to fully understand the operation of many (if not most) electrical devices and effects.

At its most basic level, electromagnetics concerns itself with the forces that charged particles exert upon each other. These forces are important for two reasons. The first is that they determine how electric charges and currents distribute themselves in electrical devices. The second is that it is such forces that make electricity useful to us, since they make other things move and allow us to detect the presence of charges and currents. Indeed, applications such as telecommunications, electrical machines, and computers would not be possible were it not for electromagnetic forces.

Although there are many applications where electromagnetic forces are our primary interest (as in the case of electric motors), we are usually more interested in how those forces cause the charges and currents in circuits and devices to distribute themselves. For instance, in the case of electrical computing, electronic memory is accomplished by moving packets of charge into discrete locations in semiconductor chips and later sensing their presence (or absence). Similar processes are used in a large number of applications whereby information is stored or transmitted by controlling the flow of charges throughout a system or device. Examples include radio, television, radar, and sound reproduction, among many others.

Possibly the most useful and remarkable property of electricity is its ability to produce effects between two devices or circuits when there is no material connection between them. This is unlike mechanical systems, which must always have some sort of mechanical linkage in order for there to be any effect.[1] Electricity is capable of producing measurable effects over very large distances, even through great distances in vacuum. This allows us to routinely use electricity in such applications as wireless communications, radar, and remote sensing, as well as many others. The key to devising these applications is an understanding of the physical quantities that are responsible for the interactions involved: electric and magnetic fields. Once these concepts are understood, the range of applications in electrical devices and systems is limited only by our imagination, our knowledge of the properties of materials, and our manufacturing ability.

1-2 A Little History

Electrical systems and devices are so common in our lives that it is difficult to envision an age when electricity and magnetism were simply mysterious curiosities. But up until the early 1800s, that is exactly what they were. How these phenomena were discovered, understood, and harnessed is one of the greatest feats in the history of science and engineering.

"Electromagnetics" is a word that was coined in the late 1800s to denote a newly discovered phenomenon that was the combination of what previously had been thought to be completely separate phenomena: electricity and magnetism. Electric effects were the first to be discovered. History records that the ancient Greeks discovered that when an amber rod was rubbed with fur, the amber would attract bits of dust, straw, and other small objects. Nearly 2,000 years passed before William Gilbert realized in the early 1600s that this same effect could be observed when rubbing a variety of substances together. It was he who coined the term "electric," using the Greek word for amber, *elektron*. About the same time, Niccolo Cabeo also discovered that the electric effect could result in both attractive and repulsive forces between electrified (i.e., charged) objects.

The first indications that electricity can move from one place to another came from experiments conducted by Stephen Gray in 1729. He found that when two objects were connected by a tube, both could be electrified when only one was rubbed. This discovery led J.T. Desaguliers in 1739 to the discovery of a class of materials he called *conductors* that pass electricity easily.

[1] Although mechanical systems are, in theory, coupled by gravity, this coupling is so weak as to render it essentially useless in most applications.

As interesting as these discoveries were, they did not explain how these electric effects occurred. This started to change in the mid-1700s when a number of investigators began to suspect that the forces between charges could be described as an inverse-square law that was similar to the universal gravitational law proposed by Sir Isaac Newton in the late 1600s. Although Benjamin Franklin, Joseph Priestley, John Robison, and Henry Cavendish all made significant contributions to the discovery of this law, it was Charles Augustin de Coulomb who attracted the most attention, so we now call the law *Coulomb's law of force.* The discovery of Coulomb's law was the first step towards finding a comprehensive theory of electromagnetics.

Like the electric properties of amber, the magnetic properties of a mineral called lodestone were also known to the ancients, who knew that the mineral could attract iron and would point towards north when allowed to float on water. As time progressed, several other materials were found to possess similar characteristics. Also, it was discovered that artificial magnets could be made from naturally occurring ones. The first quantitative theories of magnetism were advanced in the 18th century. In 1750, John Michell theorized that permanent magnets have north and south poles that attract or repel each other according to an inverse-square law that is similar to Coulomb's law of force.

The pace of discovery of both the electric and magnetic effects quickened with the onset of the 19th century. In the year 1800, Volta developed the first chemical battery, which consisted of strips of dissimilar metals immersed in a weak acid electrolyte. This invention enabled the flow of steady currents and fostered numerous experiments involving chemical effects, heating, and material studies. One of the most important series of experiments was performed by George Simon Ohm in 1826; Ohm showed that when a constant voltage is applied to a conductor, the resulting current is proportional to the conductor's cross-sectional area and inversely proportional to its length. This is *Ohm's law*, which is one of the most important laws of circuit theory.

The first evidencc that electric and magnetic phenomena are related came from Hans Christian Oersted, who, in 1819, discovered that a steady current could move a compass needle, just as a permanent magnet can. This was closely followed by André-Marie Ampère's discovery that electric currents exert attractive and repulsive forces on each other. Ampère discovered that the force exerted by current segments varies inversely with the square of the distance between them and is perpendicular to the line that connects them. We call this law *Ampère's law of force*, which is the magnetic analog of Coulomb's law of force.

Another important experimental connection between electric and magnetic effects was discovered by Michael Faraday in 1831. He conducted an experiment whereby two insulated wires were wrapped around an iron core. Faraday found that when the current in one winding was switched, a voltage was induced in the other. This discovery of transformer action led Faraday to a series of experiments in which he was able to conclude that a voltage is produced in a circuit whenever a time-varying magnetic field is present—either because the current is time varying or because the circuit or source are in motion. We call this *Faraday's law of induction* (often simply called *Faraday's law*).

With the discovery of Faraday's law, the stage was set for the development of a complete theory of electromagnetism. This was accomplished by James Clerk

Maxwell, a professor of experimental physics at Cambridge University. In 1873 he published *A Treatise on Electricity and Magnetism.* In this work, he proposed that just as time-varying magnetic fields can produce electric fields, the opposite is also true. Adding this conjecture to what was already known about electricity and magnetism, Maxwell produced his now-famous system of equations, called *Maxwell's equations.* These equations relate electric and magnetic fields to each other and their sources. In addition to his work in electromagnetics, Maxwell is known for several other contributions to physics, including thermodynamics (where a set of equations also bears his name) and the first workable theory of the rings of Saturn.

The definitive experimental verification of Maxwell's theory came in 1886 through a series of experiments conducted by Heinrich Hertz. These experiments showed that electromagnetic waves can be propagated, reflected, and focused, just as light waves can. This discovery completely validated Maxwell's theory and ushered in the era of modern electromagnetic theory and applications.

1-3 Engineering Applications of Electromagnetics

One might think that engineering applications of electromagnetics occurred only after Maxwell's theories were presented and Hertz validated them with his experiments. The fact is, however, that there was a thriving electrical industry *before* the complete theory of electromagnetics was laid out. This activity started around 1834 with the introduction of the telegraph by Charles Wheatstone, William Cook, and Samuel Morse, among others. The first undersea telegraph cable was laid in 1851, and nearly 100,000 miles of cable had been laid worldwide by 1885. Also invented before Hertz's experiments were the telephone (1876) and the electric lightbulb (1879).

The most dramatic application of the new electromagnetic theory came in 1901 when Guglielmo Marconi sent the first wireless telegraph signals across the Atlantic Ocean. The next two decades saw a host of developments in antennas, amplifying devices, and modulation techniques, culminating in the first commercial radio broadcasts in the early 1920s. Television soon followed in the early 1930s, followed by radar in the late 1930s.

Wireless communication is probably the most conspicuous application of electromagnetics, since it involves the propagation of electromagnetic waves through air or space. Nevertheless, Maxwell's electromagnetic theory has been equally important in the development of a host of other engineering applications. This should come as no surprise, since electromagnetics is the comprehensive theory of what electricity is and how to control it. Other devices and systems in whose development electromagnetic theory played an important part include the vacuum tube (1906), the magnetron (a high-frequency amplifier and oscillator used in microwave systems; 1940), the transistor (1950), the laser (1960), and fiber-optic systems (late 1970s). In fact, it is safe to say that electromagnetic theory has been an essential ingredient in the development of every electrical device or system that we now take for granted.

Some of the major engineering applications of electromagnetics can be classed into the following areas:

Semiconductor devices: Electromagnetic theory and quantum semiconductor theory are the keys to understanding how charges and currents are manipulated within semiconductors.

High-speed circuits: Ordinary circuit theory is fine for low-frequency circuits, but breaks down when the circuit dimensions and frequency reach the point where propagation delay times can no longer be ignored. This is particularly true for microwave circuits and high-speed digital circuits.

Antennas: These devices launch and capture electromagnetic waves, so electromagnetic theory is essential to their operation. Even though antennas have been used for decades, recent advances in wireless communication systems have created the need for smaller and more efficient antennas.

Electromechanical machines: The forces that currents (and sometimes charges) exert on each other are used to make machines and devices that are capable of generating forces and torques.

Fiber-optic systems: Since the development of low-loss optical fibers in the 1970s, the number of fiber-optic communication links has grown steadily, to the point where fiber-optic transmission is now the standard in many industries. Electromagnetic theory is used to describe light propagation on the fibers, as well as the operation of the laser diodes and detectors.

Bioelectronics: In many respects, the human body can be considered as a massive collection of electrical circuits. This understanding has spawned the bioelectronics industry, which supplies instruments and systems that measure and modify various biological functions in humans and animals. Electromagnetics plays an essential part in understanding how these instruments interact with the body.

Electromagnetic interference (EMI) and compatibility (EMC): Even when a circuit or system is not intended to radiate or receive energy, these effects may still occur. This is particularly troublesome when digital and analog circuits are present in the same device, since the high current levels and fast switching times of the digital circuits often radiate unwanted energy towards the analog circuits. These problems can be controlled by using electromagnetic techniques.

Superconductors: When cooled below their critical temperature, superconductors exhibit zero resistance and repel magnets. The discovery of high-temperature superconductors has rekindled interest in using superconductors in a number of engineering applications, including power transmission and magnetic levitation.

1-4 In This Text ...

The goal of this text is to lay out the electromagnetic theory in the context of its engineering applications. This discussion starts with a review of vector calculus, which is the best "language" for describing electromagnetic effects. This is followed by a chapter that presents an overview of electromagnetic effects and Maxwell's equations. This chapter is intended to provide a broad view of the relationship between electric and magnetic fields and the sources that produce them. From there, the next three chapters discuss various aspects of low-frequency electric fields when magnetic effects are negligible. These chapters are followed by three chapters that discuss low-frequency magnetic fields when electric effects are negligible.

Chapter 10 begins the transition from low-frequency electromagnetic effects to high-frequency effects, where electric and magnetic fields are interdependent. In this chapter, Maxwell's equations are described in detail for both transient and time-harmonic cases.

The final four chapters of the text discuss four electromagnetic topics that are important to the operation of high-frequency devices and systems designed and used by electrical engineers. The first topic is transmission lines, which are used to transport signals and energy in electric circuits. The second topic is plane waves, which are the waves launched into space by sources such as antennas and lasers. The third topic is waveguides and cavities. Like transmission lines, waveguides are also used to transport signals and energy. A common example of waveguides is optical fibers, which are commonly used in communication systems. The final topic is radiation and antennas. In this chapter, the general theory of how sources radiate is discussed, and many practical aspects of antennas are examined.

2

Vector Analysis

2-1 Introduction

Vector analysis is the branch of mathematics that was developed to describe quantities that are both directional in nature and distributed over regions of space. The reason for starting our study of electromagnetics with vector analysis is simple: Vector analysis is the language best suited for describing electromagnetic effects.

In this chapter, we will discuss the elements of vector analysis that are directly applicable to electromagnetic phenomena. Our discussion will start by defining the concept of a physical quantity and then identifying the properties of scalar and vector fields. The remainder of our discussion will be devoted to a development of the algebra and calculus of vector fields.

2-2 Physical Quantities and Units

Electromagnetics deals with phenomena and entities that can be perceived and measured. We call entities that can be measured ***physical quantities***. In physics, physical quantities are always defined in terms of the measurement procedure used to perceive them:

The definition of a physical quantity is the description of the operational procedure used to measure it.[1]

This kind of definition is called an *operational definition*, since it defines a physical quantity in terms of the process used to measure it.

To help us understand this definition, let us consider a common physical quantity, distance. We can define the distance between two points as the total number of measuring objects that can be laid end to end on a straight line between the points. The measuring object can be anything, such as a rock, a twig, or a meterstick. Since we have defined distance by telling how to measure it, any number of people can measure the distance between two points and obtain the same answer. Obviously, the accuracy to which a physical quantity can be measured depends upon how carefully one follows the measurement procedure.

Any measurement is a comparison of what is being measured with some standard. These standards are called ***units***. In the case of the distance example just presented, the unit is the object whose size is used as the measurement standard (i.e., the rock, twig, or meter stick). In order for the specification of a physical quantity to have meaning, the unit must be well defined. For instance, if a rock is used as the unit, the particular rock must be clearly identified, along with how it is oriented during the measurement process.

Because they are defined in terms of how they are measured, physical quantities are always specified by one or more numbers, each with its associated unit. The number of numbers needed to specify a particular physical quantity depends on the way the quantity is defined. For instance, only one number is required to specify a distance. On the other hand, three numbers (called coordinates) are necessary to specify a position in three-dimensional space. All naturally occurring physical quantities can be represented by real numbers. However, we sometimes find it convenient to create complex-valued physical quantities from naturally occurring physical quantities. A common example is in circuit analysis, where complex phasors are used to represent sinusoidal steady-state voltages and currents. In these cases, complex-valued quantities can be considered to represent two quantities—one real, the other imaginary.

The unit of a physical quantity can be any well-defined standard, but it is usually desirable to limit the number of units used in a measurement to as few as possible. Certain sets of quantities are, by convention, regarded as ***fundamental quantities***, specified in internationally accepted ***fundamental units***. All other units can be derived in terms of these fundamental units. By far the most accepted unit system among electrical engineers is the MKSA system.[2] The fundamental quantities in this system are *length*, *mass*, *time*, and *current*, specified in *meters*, *kilograms*, *seconds*, and *amperes*, respectively. The MKSA system is a subset of the ***International System of Units*** (SI), which also includes the *candela* (a unit of *luminous intensity*) and the *Kelvin* (a unit of *temperature*).

[1] See Shortley and Williams, *Elements of Physics*, 4th ed. (Englewood Cliffs, NJ: Prentice-Hall, 1965), p. 4.

[2] For a complete discussion of the unit systems used in electromagnetics see *Handbook of Chemistry and Physics* (Boca Raton, FL: The Chemical Rubber Co., 1991).

Derived units are units that are specified in terms of the fundamental units. The newton, for example, is a derived unit of force, defined as 1.0 [kg · m/sec^2]. The most common derived units used in electromagnetic analysis are summarized in Appendix A.

The dimensions of a physical quantity are specified by the powers of the fundamental physical quantities that occur in its definition.

For example, since the unit of speed is a length unit divided by a time unit, its dimensions are (length)/(time). Similarly, the dimensions of the newton are (mass) · (length)/(time)^2. The dimensions of physical quantities are important, because two physical quantities can be added or subtracted if, and only if, they have the same dimensions. Thus, apples can be equated with apples, but not oranges. Any equation in which the units of the left- and right-hand sides do not agree is simply wrong.

2-2-1 DISCRETE AND FIELD QUANTITIES

The physical quantities used in electromagnetics can be either discrete or field quantities. The simplest are discrete quantities.

Discrete quantities are defined over regions of space or at single points, but not on a point-by-point basis throughout a region.

The definition of the average temperature of a room has nothing to do with the position of the observer, so it is a discrete quantity. Similarly, the distance from Kansas City to New York is a quantity that is independent of the position of an observer.

Most of the physical quantities encountered in electromagnetics are field quantities.

Field quantities are defined on a point-by-point basis throughout a region of space.

The temperature in a room is a field quantity, since it is defined uniquely for each point in the room. Another field quantity is wind speed, which is also a function of the position at which it is measured.

The distinction between discrete and field quantities is important, because the procedures used to describe them are different. One needs only ordinary algebra to balance a checkbook or to compare the weights of two objects, since these operations involve only discrete quantities. Analyzing the characteristics of the temperature distribution inside a room is more difficult. This is because the temperature is a continuous function of position and requires vector analysis to describe it fully.

2-2-2 SCALARS AND VECTORS

We have just seen that physical quantities can be classified as either discrete or field quantities. They can also be classified according to the number of numbers needed to specify them. The simplest are scalars.

A ***Scalar*** is a quantity that can be specified by a single number and its associated unit.

Temperature, altitude, and weight are all scalar quantities, since each can be specified by a single number. Throughout this text, scalar quantities are represented as nonboldface symbols, in italics, such as D and ρ. Also, the units of all scalars will be written in brackets, such as [kg/m].

There are also physical quantities that have both a size and a direction associated with them. These are called vectors and are defined as follows:

A ***Vector*** is a quantity that can be specified by a ***magnitude*** and a ***direction***.

Examples of vector quantities are velocity and force. Throughout this text, boldface alphabetic characters, such as **A**, **E**, and **h**, are used to denote vector quantities.[3]

The magnitude and direction of a vector are very different entities. The magnitude is a positive-valued scalar, which includes its associated unit. We will represent the magnitude of a vector **A** as either $|\mathbf{A}|$ or A. The direction, as its name implies, is a spatial orientation. For instance, the wind velocity at a point may be specified as 2.5 [m/s] in the southeast direction.

By convention, any vector can be represented graphically by a line extending from a tail to a head. An arrow is placed at the head and points in the direction of the vector. The distance from the tail to the head represents the vector's magnitude. The graphical representation of a vector **A** is shown in Figure 2-1.

Discrete vectors are associated with regions of space, but not a specific point. An important example is the directed distance between two points.

The ***directed distance*** $\mathbf{R}_{ab}$ between the points a and b is a vector whose magnitude equals the distance between these points and whose direction is parallel to the line directed from a to b.

Even though this definition of $\mathbf{R}_{ab}$ involves the points a and b, this vector is not defined to exist at any particular point. As a result, its representation can be translated freely to any point, as long as its magnitude and direction are not changed. This is depicted in Figure 2-2.

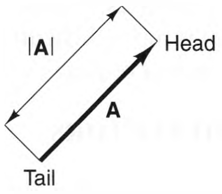


Figure 2-1 A graphical representation of a vector.

[3] In handwritten work, vectors are typically written as $\bar{A}$ or $\vec{A}$, since boldface characters are difficult to draw by hand.

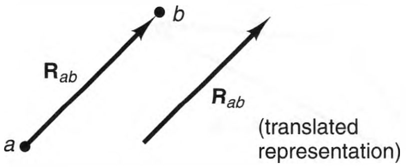


Figure 2-2 A discrete vector, shown at two different locations in space.

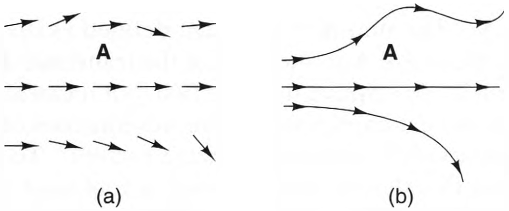


Figure 2-3 Graphical representations of a vector field: a) quiver plot, b) streamlines.

Figure 2-3a shows the graphical representation of a ***vector field***. Here, the vector is represented at equally spaced points in a ***quiver plot*** (sometimes called a ***needle plot***). This type of diagram is helpful in that it conveys both the magnitude and direction of the vector at a number of points. On the other hand, such a diagram does not easily convey the sense of the vector's "flow." By flow, we mean the path that a particle would take if it were pushed by the vector (assuming that the vector represented a force field). Flow is best represented by a streamline plot, such as that shown in Figure 2-3b. Here, continuous lines called ***streamlines*** are drawn that are tangent to the vector's direction at each point. These streamlines are the paths that the vector would "push" a particle. Magnitude information is not directly conveyed by the streamlines. Nevertheless, one can usually infer this information by measuring the spacings between the streamlines, since vector magnitudes are usually strongest where the streamlines are the closest. This can be seen by comparing Figure 2-3a and Figure 2-3b.

2-3 Vector Algebra

Having defined scalars and vectors we will now define several operations involving them. These operations are essential, for without them we would have no way to formulate the mathematical equations that describe the physical processes found in electromagnetics. Three classes of operations are possible in vector algebra: scalar-scalar, scalar-vector, and vector-vector. Since the operations from the first class are already known from ordinary algebra, our discussion will be limited to operations involving vectors.

2-3-1 ADDITION AND SUBTRACTION OF VECTORS

The sum of any two vectors **A** and **B** is itself a vector, defined by the graphical process depicted in Figure 2-4a. Here, the vector sum **A** + **B** is defined as the vector that completes the parallelogram formed by **A** and **B**.

An equivalent definition of the sum **A** + **B**, called the ***addition tail-to-head rule***, is depicted in Figure 2-4b, where the representation of **B** has been translated so that its

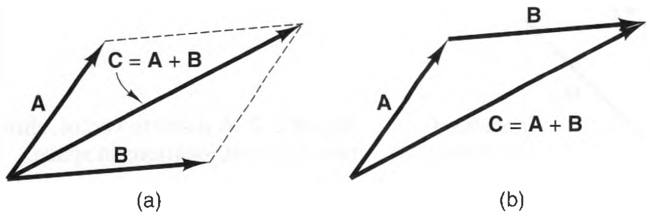


Figure 2-4 Vector addition: a) completing the parallelogram, b) the head-to-tail rule.

tail lies at the head of **A**. The sum **A** + **B** is then defined as the vector whose representation extends from the tail of **A** to the head of the translated **B**. This definition is, in some ways, more visually descriptive than the first, but it can also be somewhat misleading for field vectors, since it implies that the representations of a vector field can be moved to any point in space as if they were discrete vectors. To counter this illusion, one must remember that this sliding process is only a tool used to define **A** + **B**. In reality, **A**, **B**, and **A** + **B** are all defined at *exactly* the same point.

Vector addition satisfies the associative and commutative laws:

Associative law: $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$ (2.1a)

Commutative law: $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}.$ (2.1b)

Both proofs are straightforward from the definition of vector addition and are left as an exercise for the reader.

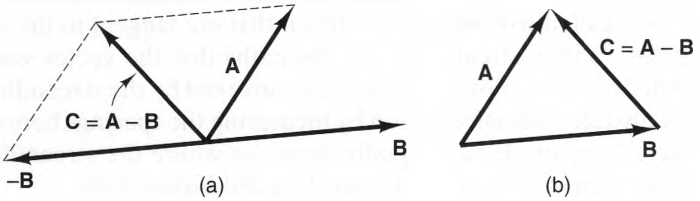


Figure 2-5 Vector subtraction: a) completing the parallelogram, b) head-to-tail rule.

Vector subtraction is defined in terms of vector addition by

$$\mathbf{C} = \mathbf{A} - \mathbf{B} \equiv \mathbf{A} + (-\mathbf{B}), \tag{2.2}$$

where the symbol "≡" means "equals by definition." The vector −**B** is called the ***negative*** of **B**; it has the same magnitude as **B**, but opposite direction. Figure 2-5a shows the graphical representation of **C** = **A** + (−**B**). Figure 2-5b shows that **C** = **A** − **B** can also be represented using the ***subtraction tail-to-head rule***, where the representation of **A** − **B** extends from the tip of **B** to the tip of **A**. When using this rule for vector fields, however, it must be remembered that **A**, **B**, and **C** all exist at the same point, even though **C** has been translated by the graphical procedure.

2-3-2 MULTIPLICATION OF A VECTOR BY A SCALAR

The product $a\mathbf{B}$ is defined as a vector with the same direction as **B** and magnitude equal to $|a|\,|\mathbf{B}|$. If the sign of a is negative, the direction of the vector $a\mathbf{B}$ is opposite to that of **B**. Figure 2-6 depicts the scalar product $a\mathbf{B}$.

The product of a scalar and a vector obeys the commutative and distributive laws:

Commutative law: $a\mathbf{B} = \mathbf{B}a$ (2.3)

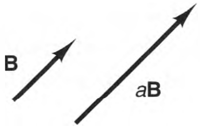


Figure 2-6 Multiplication of a scalar and a vector.

Distributive law: $$a(\mathbf{B} + \mathbf{C}) = a\mathbf{B} + a\mathbf{C}. \quad (2.4)$$

The quotient $\frac{\mathbf{B}}{a}$ can be defined in terms of the scalar-vector product

$$\frac{\mathbf{B}}{a} \equiv a^{-1}\mathbf{B} = \frac{1}{a}\mathbf{B}. \quad (2.5)$$

We can use the scalar-vector product to represent an arbitrary vector **A** in the form

$$\mathbf{A} = |\mathbf{A}|\hat{\mathbf{a}}_A = A\hat{\mathbf{a}}_A, \quad (2.6)$$

where $|\mathbf{A}|$ and A are the magnitude of **A** and $\hat{\mathbf{a}}_A$ is a ***unit vector*** that has the same direction as **A** and a magnitude of unity (i.e., 1.0). Multiplying both sides of Equation (2.6) by $|\mathbf{A}|^{-1}$, we obtain the following expression for the unit vector $\hat{\mathbf{a}}_A$:

$$\hat{\mathbf{a}}_A = \frac{\mathbf{A}}{|\mathbf{A}|}. \quad (2.7)$$

As its name implies, a unit vector has unit magnitude.

2-3-3 THE SCALAR (OR DOT) PRODUCT OF TWO VECTORS

There are two multiplication operators that involve two vectors. The first is called the ***scalar product***, because it produces a scalar. The scalar product of two vectors is defined as a scalar whose value is given by

$$\mathbf{A} \cdot \mathbf{B} \equiv |\mathbf{A}|\,|\mathbf{B}| \cos\theta_{AB}. \quad (2.8)$$

Here, the angle θ_{AB} is defined as the smaller angle between **A** and **B** (i.e., $\theta_{AB} \leqslant 180°$), and $|\mathbf{A}|$ and $|\mathbf{B}|$ are the magnitudes of **A** and **B**, respectively. The expression $\mathbf{A} \cdot \mathbf{B}$ is read as "**A** dot **B**", and the terms "scalar product" and "dot product" are used interchangeably.

When we take the dot product of a vector with itself, we obtain

$$\mathbf{A} \cdot \mathbf{A} = |\mathbf{A}|\,|\mathbf{A}| \cos\theta_{AA} = |\mathbf{A}|\,|\mathbf{A}| = |\mathbf{A}|^2. \quad (2.9)$$

Thus, the magnitude of any vector can be written in terms of its dot product:

$$|\mathbf{A}| = \sqrt{\mathbf{A} \cdot \mathbf{A}|}. \quad (2.10)$$

The dot product satisfies the commutative and distributive laws:

Commutative law $$\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A} \quad (2.11)$$

Distributive law $$\mathbf{A} \cdot (\mathbf{B} + \mathbf{C}) = \mathbf{A} \cdot \mathbf{B} + \mathbf{A} \cdot \mathbf{C}. \quad (2.12)$$

The commutative law follows directly from the symmetry of the dot product. The proof of the distributive law is straightforward and is left as an exercise.

The definitions of perpendicular and collinear vectors are derived from the dot product:

Two vectors **A** and **B** are ***perpendicular*** (or ***orthogonal***) if $\mathbf{A} \cdot \mathbf{B} = 0$. Vectors are ***collinear*** if $|\mathbf{A} \cdot \mathbf{B}| = |\mathbf{A}|\,|\mathbf{B}|$. Collinear vectors are ***parallel*** if $\mathbf{A} \cdot \mathbf{B} = |\mathbf{A}|\,|\mathbf{B}|$ and are ***antiparallel*** if $\mathbf{A} \cdot \mathbf{B} = -|\mathbf{A}|\,|\mathbf{B}|$.

The dot product is a convenient tool for finding the component of a vector along particular direction. Figure 2-7 shows a vector **A** and reference line that is parallel to the direction of the unit vector $\hat{\mathbf{a}}_B$. Also shown is the right triangle formed by the reference line, the vector **A**, and the line that extends from the tip of **A** and intersects the reference line at a right angle. When $\theta_{AB} \leqslant 90°$, the component A_B of the vector **A** along the direction $\hat{\mathbf{a}}_B$ is defined as the length of the side of the right triangle that lies along the reference line. If $\theta_{AB} > 90°$, then the component A_B is the negative of this distance. From this definition and Figure 2-7, it follows that the component of **A** in the direction $\hat{\mathbf{a}}_B$ is

$$A_B = |\mathbf{A}| \cos \theta_{AB}.$$

But from Equation (2.8), we find that $|\mathbf{A}| \cos \theta_{AB}$ can be written as the dot product $\mathbf{A} \cdot \hat{\mathbf{a}}_B$. Hence, we can write

$$A_B = \mathbf{A} \cdot \hat{\mathbf{a}}_B = |\mathbf{A}| \cos \theta_{AB}. \tag{2.13}$$

The dot product can be used to expand any vector as the sum of perpendicular component vectors. Consider the vector **A**, shown in Figure 2-8, which exists in three-dimensional space.

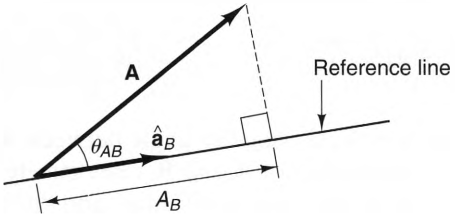


Figure 2-7 The projection of a vector along a reference line.

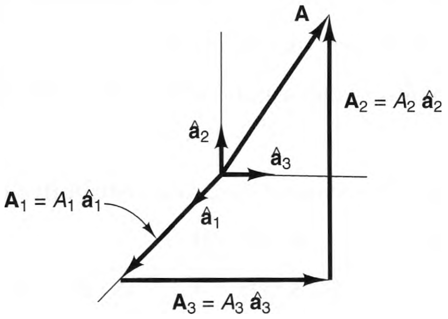

A
$\mathbf{A}_2 = A_2\,\hat{\mathbf{a}}_2$
$\hat{\mathbf{a}}_2$
$\hat{\mathbf{a}}_3$
$\mathbf{A}_1 = A_1\,\hat{\mathbf{a}}_1$
$\hat{\mathbf{a}}_1$
$\mathbf{A}_3 = A_3\,\hat{\mathbf{a}}_3$

Figure 2-8 An arbitrary vector **A**, shown as the sum of three mutually orthogonal component vectors.

If the unit vectors $\hat{\mathbf{a}}_1$, $\hat{\mathbf{a}}_2$, and $\hat{\mathbf{a}}_3$ are mutually orthogonal (perpendicular), we can express $\mathbf{A}$ as

$$\mathbf{A} = A_1\hat{\mathbf{a}}_1 + A_2\hat{\mathbf{a}}_2 + A_3\hat{\mathbf{a}}_3. \tag{2.14}$$

The scalars A_1, A_2, and A_3 are the components of $\mathbf{A}$ along the directions $\hat{\mathbf{a}}_1$, $\hat{\mathbf{a}}_2$, and $\hat{\mathbf{a}}_3$, respectively. Remembering that the dot products of perpendicular vectors are zero, we can find the components A_1, A_2, and A_3 simply by taking the dot products of $\mathbf{A}$ with $\hat{\mathbf{a}}_1$, $\hat{\mathbf{a}}_2$, and $\hat{\mathbf{a}}_3$, respectively, obtaining

$$A_i = \mathbf{A} \cdot \hat{\mathbf{a}}_i, \quad i = 1, 2, \text{or } 3. \tag{2.15}$$

2-3-4 THE VECTOR (OR CROSS) PRODUCT OF TWO VECTORS

The second product of vectors is the ***vector*** (or ***cross***) ***product.*** Unlike the scalar product, which produces a scalar from two vectors, the vector product of two vectors produces another vector, defined by

$$\mathbf{A} \times \mathbf{B} \equiv \hat{\mathbf{a}}_n |\mathbf{A}|\,|\mathbf{B}| \sin(\theta_{AB}), \tag{2.16}$$

where θ_{AB} is defined as the smallest angle between $\mathbf{A}$ and $\mathbf{B}$. The expression $\mathbf{A} \times \mathbf{B}$ is read as "**A** cross **B**," and the terms "vector product" and "cross product" are used interchangeably. Figure 2-9 shows the relationship between **A**, **B**, and $\mathbf{A} \times \mathbf{B}$. The unit vector $\hat{\mathbf{a}}_n$ is specified by a convention called the ***right-hand rule.*** This rule states that $\hat{\mathbf{a}}_n$ is perpendicular to both **A** and **B** and points in the direction of a right-hand thumb when the other fingers point along the arc that **A** would follow if it were rotated into **B** through the smallest angle between them.

The cross product obeys the distributive law:

$$\textbf{Distributive law:} \quad \mathbf{A} \times (\mathbf{B} + \mathbf{C}) = \mathbf{A} \times \mathbf{B} + \mathbf{A} \times \mathbf{C}. \tag{2.17}$$

This can be proved directly from the definition of the cross product. On the other hand, the cross product obeys neither the commutative nor the associative laws, which can be seen from the inequalities

$$\mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A} \neq \mathbf{B} \times \mathbf{A} \tag{2.18}$$

and

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) \neq (\mathbf{A} \times \mathbf{B}) \times \mathbf{C}. \tag{2.19}$$

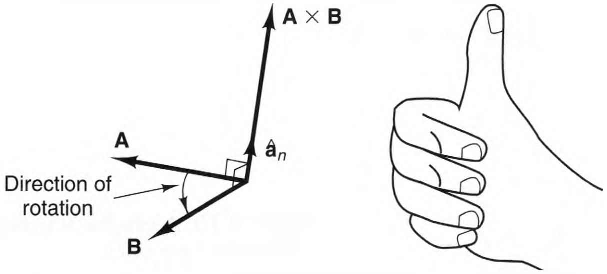


Figure 2-9 The cross product of two vectors.

Equation (2.18) is a direct result of the right-hand rule. Equation (2.19) is most easily proved by observing that the vector **A** × (**B** × **C**) lies in the plane formed by **B** and **C** (since **B** × **C** lies perpendicular to that plane), whereas the vector (**A** × **B**) × **C** lies in the plane containing **A** and **B**.

2-3-5 PRODUCTS OF THREE VECTORS

There are two combinations of products that involve three vectors. These are the ***scalar triple product*** and the ***vector triple product***, so named because they produce a scalar and a vector, respectively. The simplest of the two is the scalar triple product. For three vectors **A**, **B**, and **C**, the scalar triple product $\mathbf{A} \cdot \mathbf{B} \times \mathbf{C}$ has the following cyclic property:

$$\mathbf{A} \cdot \mathbf{B} \times \mathbf{C} = \mathbf{B} \cdot \mathbf{C} \times \mathbf{A} = \mathbf{C} \cdot \mathbf{A} \times \mathbf{B}. \tag{2.20}$$

This identity is easily proved by referring to Figure 2-10, which shows a parallelepiped formed by the vectors **A**, **B**, and **C**. From solid geometry, the volume of the parallelepiped is $|\mathbf{A}|\,|\mathbf{B}|\,|\mathbf{C}| \sin\theta_1 \cos\theta_2$, which can be expressed as $\mathbf{A} \cdot \mathbf{B} \times \mathbf{C}$. Similar reasoning yields the other two expressions in Equation (2.20).

In addition, the vector triple product **A** × (**B** × **C**) satisfies the following identity:

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = \mathbf{B}(\mathbf{A} \cdot \mathbf{C}) - \mathbf{C}(\mathbf{A} \cdot \mathbf{B}). \tag{2.21}$$

This identity can be proven by expanding the vectors in Cartesian coordinates (which will be discussed shortly).

2-4 Orthogonal Coordinate Systems

Our discussion of vectors has been hindered thus far by our inability to specify positions and directions, except through graphical representations. We will now introduce the concept of a coordinate system, which provides the framework necessary to describe these quantities without graphical representations.

Coordinate systems provide two attractive features that aid in vector operations. The first is the ability to specify positions in space by a sequence of scalars, called ***coordinates***. Coordinates identify the position of a point with respect to a coordinate center (or origin). The minimum number of scalars needed to uniquely specify a point in a particular domain (or space) determines the dimension of the space. Lines are one dimensional, surfaces are two dimensional, and volumes are three dimensional.

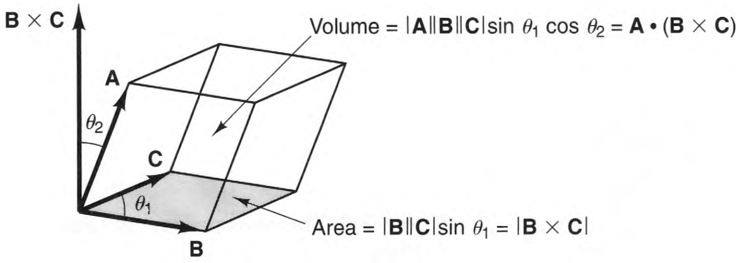

B × C
Volume = |A||B||C|sin θ1 cos θ2 = A • (B × C)
A
θ2
C
θ1
B
Area = |B||C|sin θ1 = |B × C|

Figure 2-10 A graphical depiction of the scalar triple product.

Coordinate systems also provide a convenient way to specify vectors at any point. This is accomplished through the use of an orthogonal set of vectors, called ***base vectors***, which are defined at each point. Any vector can be uniquely defined in terms of its components in the base vector directions. The number of base vectors defined by a coordinate system equals the ***dimension*** of the space. Each base vector is defined at a point in space in terms of the position coordinates used to identify the point:

In a coordinate system in which a point P is described by the coordinates $P(u_1, u_2, u_3)$, the ith base unit vector $\hat{\mathbf{a}}_i$ at P has a direction parallel to a line through P along which only u_i varies, and points towards increasing values of u_i.

Three coordinate systems are discussed in this section: Cartesian (or rectangular), cylindrical, and spherical. Although there are many others, these three are sufficient to model all of the electromagnetic configurations discussed in this text. Why do we need more than one? The reason is that no one coordinate system is best suited to all situations.

2-4-1 THE CARTESIAN COORDINATE SYSTEM

In the Cartesian coordinate system, three mutually perpendicular axes are used that intersect at a point, called the ***origin***. These axes are typically called the x-, y-, and z-axes, respectively, and are oriented according to the right-hand rule: The rotation of the positive x-axis into the positive y-axis would cause a right-handed screw at the origin to thread along the positive z-axis.

In this coordinate system, a point is identified by its three position coordinates: $u_1 = x$, $u_2 = y$, and $u_3 = z$, each defined as the perpendicular distance from the point to the x-, y-, and z-axes, respectively. As shown in Figure 2-11, any point can be envisioned as the point of intersection of three planes: x = constant, y = constant, and z = constant, where any of the three position coordinates can have any real value between $-\infty$ and $+\infty$.

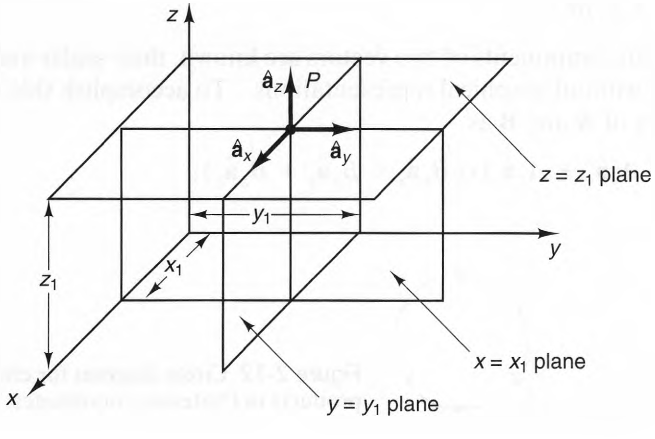


Figure 2-11 Position coordinates and base vectors in the Cartesian coordinate system.

The base vectors of the Cartesian coordinate system are particularly simple. At any point P, the unit vector $\hat{\mathbf{a}}_x$ is directed towards points having increasing values of x and perpendicular to the y = constant and z = constant planes. This direction is always parallel to the x-axis, regardless of the location of P. The definitions of $\hat{\mathbf{a}}_y$ and $\hat{\mathbf{a}}_z$ are similar, and are shown in the figure. From these definitions, it follows that the base vectors have the following product relationships:

$$\hat{\mathbf{a}}_x \bullet \hat{\mathbf{a}}_x = \hat{\mathbf{a}}_y \bullet \hat{\mathbf{a}}_y = \hat{\mathbf{a}}_z \bullet \hat{\mathbf{a}}_z = 1 \tag{2.22a}$$

$$\hat{\mathbf{a}}_x \bullet \hat{\mathbf{a}}_y = \hat{\mathbf{a}}_x \bullet \hat{\mathbf{a}}_z = \hat{\mathbf{a}}_y \bullet \hat{\mathbf{a}}_z = 0 \tag{2.22b}$$

$$\hat{\mathbf{a}}_x \times \hat{\mathbf{a}}_x = \hat{\mathbf{a}}_y \times \hat{\mathbf{a}}_y = \hat{\mathbf{a}}_z \times \hat{\mathbf{a}}_z = 0 \tag{2.22c}$$

$$\hat{\mathbf{a}}_x \times \hat{\mathbf{a}}_y = \hat{\mathbf{a}}_z \tag{2.22d}$$

$$\hat{\mathbf{a}}_y \times \hat{\mathbf{a}}_z = \hat{\mathbf{a}}_x \tag{2.22e}$$

$$\hat{\mathbf{a}}_z \times \hat{\mathbf{a}}_x = \hat{\mathbf{a}}_y. \tag{2.22f}$$

These product relations are simple to derive, but the cross products are somewhat difficult to remember. Fortunately, there is a simple way to remember them. Looking closely at these four relationships (2.22c through 2.22f), we notice a sequence between the unit vectors that can be represented by the circle shown in Figure 2-12.

To determine the cross product between any two base vectors, start on the circle at the coordinate symbol of the first vector in the product, and progress past the coordinate symbol of the second vector by the shortest route. The next symbol encountered along that route is the coordinate symbol of the resulting unit vector. The sign of this unit vector is positive if the progression is clockwise (i.e., along the arrows) and negative if it is counterclockwise.

Any vector can be expanded at any point in terms of its components in the base vectors:

$$\mathbf{A} = A_x\hat{\mathbf{a}}_x + A_y\hat{\mathbf{a}}_y + A_z\hat{\mathbf{a}}_z, \tag{2.23}$$

where the scalars A_x, A_y, and A_z are the x, y, and z components of the vector $\mathbf{A}$, respectively. Using Equation (2.15), we can find these components by taking the dot product of both sides of Equation (2.23) with each of the base vectors, yielding

$$A_i = \mathbf{A} \bullet \hat{\mathbf{a}}_i \quad i = x, y, \text{or } z. \tag{2.24}$$

Once the Cartesian components of two vectors are known, their scalar and vector products can be found without graphical representations. To accomplish this, we first express the dot product of $\mathbf{A}$ and $\mathbf{B}$ as

$$\mathbf{A} \bullet \mathbf{B} = (A_x\hat{\mathbf{a}}_x + A_y\hat{\mathbf{a}}_y + A_z\hat{\mathbf{a}}_z) \bullet (B_x\hat{\mathbf{a}}_x + B_y\hat{\mathbf{a}}_y + B_z\hat{\mathbf{a}}_z).$$

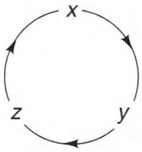


Figure 2-12 Circle diagram for cross products in Cartesian coordinates.

Using the orthogonality properties of the base vectors, this becomes

$$\mathbf{A} \cdot \mathbf{B} = A_x B_x + A_y B_y + A_z B_z. \tag{2.25}$$

From Equation 2.10, $|\mathbf{A}|$ can be expressed as

$$|\mathbf{A}| = \sqrt{\mathbf{A} \cdot \mathbf{A}} = \sqrt{A_x^2 + A_y^2 + A_z^2}, \tag{2.26}$$

which can also be derived from the Pythagorean theorem.

Similarly, the cross product of **A** and **B** can be expressed in terms of components:

$$\mathbf{A} \times \mathbf{B} = (A_x \hat{\mathbf{a}}_x + A_y \hat{\mathbf{a}}_y + A_z \hat{\mathbf{a}}_z) \times (B_x \hat{\mathbf{a}}_x + B_y \hat{\mathbf{a}}_y + B_z \hat{\mathbf{a}}_z).$$

Using the cross-product relations of the base vectors, this becomes

$$\mathbf{A} \times \mathbf{B} = (A_y B_z - A_z B_y)\hat{\mathbf{a}}_x + (A_z B_x - A_x B_z)\hat{\mathbf{a}}_y + (A_x B_y - A_y B_x)\hat{\mathbf{a}}_z. \tag{2.27}$$

Each term in this formula can be evaluated using the circle aid in Figure 2-12. For instance, the term $A_y B_z$ is positive because $\hat{\mathbf{a}}_y$ crossed into $\hat{\mathbf{a}}_z$ yields $+\hat{\mathbf{a}}_x$. Similarly, $\hat{\mathbf{a}}_z$ crossed into $\hat{\mathbf{a}}_y$ yields $-\hat{\mathbf{a}}_x$, so the sign of the term $A_z B_y$ is negative. Equation (2.27) can also be written as a determinant:

$$\mathbf{A} \times \mathbf{B} = \begin{vmatrix} \hat{\mathbf{a}}_x & \hat{\mathbf{a}}_y & \hat{\mathbf{a}}_z \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix}. \tag{2.28}$$

Expanding this determinant by minors yields

$$\mathbf{A} \times \mathbf{B} = \begin{vmatrix} A_y & A_z \\ B_y & B_z \end{vmatrix} \hat{\mathbf{a}}_x - \begin{vmatrix} A_x & A_z \\ B_x & B_z \end{vmatrix} \hat{\mathbf{a}}_y + \begin{vmatrix} A_x & A_y \\ B_x & B_y \end{vmatrix} \hat{\mathbf{a}}_z. \tag{2.29}$$

Another attractive feature of coordinate systems is that they provide simple expressions for the differential quantities needed to evaluate integrals of vector and scalar fields. Figure 2-13 shows the differential volume traced about a point when its

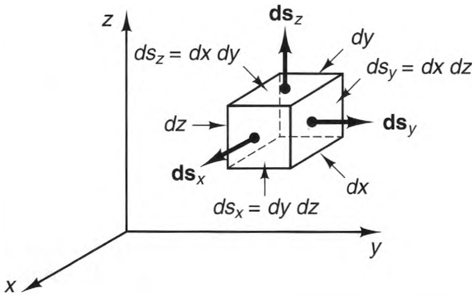


Figure 2-13 Differential volume and surface elements in the Cartesian coordinate system.

position coordinates x, y, and z are varied by the differential amounts dx, dy, and dz, respectively. The enclosed volume is

$$dv = dx\, dy\, dz. \tag{2.30}$$

Figure 2-13 also shows three differential surfaces traced when two coordinates at a point are varied by differential amounts and the third is held constant. Each of the surfaces is named according to the direction of its ***normal*** (i.e., perpendicular) direction, which is defined as the unit vector that is perpendicular to each vector (or line) that lies on that surface. From Figure 2-13 we see that the normal to each of these surfaces is the base vector corresponding to the coordinate that is constant on the surface. The area of each differential surface is

$$ds_x = dy\, dz \quad (\text{when } dx = 0) \tag{2.31a}$$

$$ds_y = dx\, dz \quad (\text{when } dy = 0) \tag{2.31b}$$

$$ds_z = dx\, dy \quad (\text{when } dz = 0). \tag{2.31c}$$

We can also define ***differential surface vectors*** for each of these differential surfaces. The magnitude of each differential surface vector equals the differential surface area, and its direction is normal to the surface. The three differential surface vectors shown in Figure 2-13 can be expressed as

$$\mathbf{ds}_x = ds_x\, \hat{\mathbf{a}}_x = dy\, dz\, \hat{\mathbf{a}}_x \quad (\text{when } dx = 0) \tag{2.32a}$$

$$\mathbf{ds}_y = ds_y\, \hat{\mathbf{a}}_y = dx\, dz\, \hat{\mathbf{a}}_y \quad (\text{when } dy = 0) \tag{2.32b}$$

$$\mathbf{ds}_z = ds_z\, \hat{\mathbf{a}}_z = dx\, dy\, \hat{\mathbf{a}}_z \quad (\text{when } dz = 0). \tag{2.32c}$$

Notice in these expressions that of the two possible normal directions for each surface, the direction outward from the enclosed volume is chosen in each case. This convention is always followed whenever a differential surface is part of a larger surface that completely encloses a volume. Such surfaces are called ***closed surfaces***.

When integrating along a line of points, it is necessary to define a differential vector that represents the magnitude and direction of each segment of the path. Consider the path shown in Figure 2-14. We define the ***differential displacement vector*** $\mathbf{d\ell}$ at point $P(x, y, z)$ to be the directed distance from $P(x, y, z)$ to $P'(x + dx, y + dy, z + dz)$. Along any path, the differential displacement vector can be represented as

$$\mathbf{d\ell} = dx\,\hat{\mathbf{a}}_x + dy\,\hat{\mathbf{a}}_y + dz\,\hat{\mathbf{a}}_z. \tag{2.33}$$

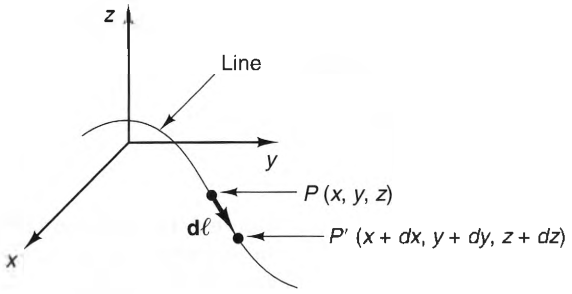


Figure 2-14 A differential displacement vector $\mathbf{d\ell}$ along an arbitrary path (line).

Here, it is important to note that dx, dy, and dz are *not* independent quantities, since each is a measure of how rapidly the x, y, and z coordinates, respectively, are varying at the point.

A frequently used method of finding dx, dy, and dz is to write the position coordinates of the line using a single, common variable called a ***parametric variable***. Thus, when a line can be represented by $P[x(s), y(s), z(s)]$, where $x(s)$, $y(s)$, and $z(s)$ are functions of the parametric variable s, the differentials dx, dy, and dz can be obtained from the relations

$$dx = \frac{dx(s)}{ds}\,ds \tag{2.34a}$$

$$dy = \frac{dy(s)}{ds}\,ds \tag{2.34b}$$

$$dz = \frac{dz(s)}{ds}\,ds\,. \tag{2.34c}$$

Example 2-1

For the vectors $\mathbf{A} = -\hat{\mathbf{a}}_x - 2\hat{\mathbf{a}}_y + 4\hat{\mathbf{a}}_z$ and $\mathbf{B} = -\hat{\mathbf{a}}_x + 3\hat{\mathbf{a}}_y - 2\hat{\mathbf{a}}_z$, find the smallest angle θ_{AB} between $\mathbf{A}$ and $\mathbf{B}$ and the unit vector $\hat{\mathbf{a}}_n$ that points along the direction of $\mathbf{A} \times \mathbf{B}$.

Solution:

We can find θ_{AB} by using the dot product. Solving Equation (2.8) for θ_{AB}, we have

$$\theta_{AB} = \cos^{-1}\frac{\mathbf{A}\cdot\mathbf{B}}{|\mathbf{A}|\,|\mathbf{B}|}\,.$$

Using Equations (2.25) and (2.26), we have

$$\mathbf{A}\cdot\mathbf{B} = (-1)(-1) + (-2)(3) + (4)(-2) = -13$$

$$|\mathbf{A}| = \sqrt{(-1)^2 + (-2)^2 + (4)^2} = \sqrt{21}$$

$$|\mathbf{B}| = \sqrt{(-1)^2 + (3)^2 + (-2)^2} = \sqrt{14}\,.$$

Substituting, we find

$$\theta_{AB} = \cos^{-1}\frac{-13}{\sqrt{21}\,\sqrt{14}} = 139.3^\circ\,.$$

To find $\hat{\mathbf{a}}_n$, we first solve Equation (2.16) for $\hat{\mathbf{a}}_n$, yielding

$$\hat{\mathbf{a}}_n = \frac{\mathbf{A}\times\mathbf{B}}{|\mathbf{A}|\,|\mathbf{B}|\,\sin\theta_{AB}}\,.$$

We can find $\mathbf{A}\times\mathbf{B}$ by using Equation (2.27):

$$\mathbf{A}\times\mathbf{B} = (4 - 12)\hat{\mathbf{a}}_x + (-4 - 2)\hat{\mathbf{a}}_y + (-3 - 2)\hat{\mathbf{a}}_z = -8\hat{\mathbf{a}}_x - 6\hat{\mathbf{a}}_y - 5\hat{\mathbf{a}}_z\,.$$

Substituting this into the expression for $\hat{\mathbf{a}}_n$, we obtain

$$\hat{\mathbf{a}}_n = \frac{-8\hat{\mathbf{a}}_x - 6\hat{\mathbf{a}}_y - 5\hat{\mathbf{a}}_z}{\sqrt{21}\,\sqrt{14}\,\sin(139.3^\circ)} = -0.716\hat{\mathbf{a}}_x - 0.537\hat{\mathbf{a}}_y - 0.447\hat{\mathbf{a}}_z\,.$$

Example 2-2

Find an expression for the differential displacement vector $\mathbf{d\ell}$ at any point on the half-circle path shown in Figure 2-15. Assume that the direction along the circle is counterclockwise.

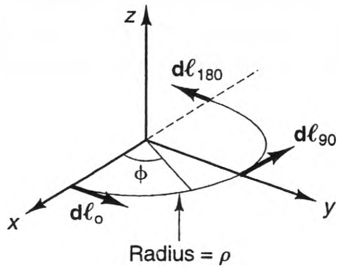


Figure 2-15 Differential displacement vectors along a counterclockwise, semicircular path of radius a.

Solution:

Since the half circle has a unit radius, the position coordinates (x, y) can be written in terms of the parametric variable ϕ as

$$x = \rho \cos \phi$$

$$y = \rho \sin \phi .$$

Using Equations (2.34a) and (2.34b), we find that

$$dx = \frac{dx}{d\phi} d\phi = -\rho \sin \phi \, d\phi$$

and

$$dy = \frac{dy}{d\phi} d\phi = \rho \cos \phi \, d\phi .$$

Hence, the differential displacement vector at any point can be written as

$$\mathbf{d\ell} = \rho(-\sin \phi \, \hat{\mathbf{a}}_x + \cos \phi \, \hat{\mathbf{a}}_y) d\phi .$$

To see if this result makes any sense, let us evaluate $\mathbf{d\ell}$ at the points $\phi = 0, 90°$, and $180°$. Substituting, we obtain

$$\mathbf{d\ell}_0 = \hat{\mathbf{a}}_y \rho d\phi \quad \text{at } \phi = 0,$$

$$\mathbf{d\ell}_{90} = -\hat{\mathbf{a}}_x \rho d\phi \quad \text{at } \phi = 90°,$$

$$\mathbf{d\ell}_{180} = -\hat{\mathbf{a}}_y \rho d\phi \quad \text{at } \phi = 180° .$$

These vectors are shown with amplified lengths (so that they can be seen) in Figure 2-15. Notice that each vector is tangent to the circle and has magnitude $\rho d\phi$.

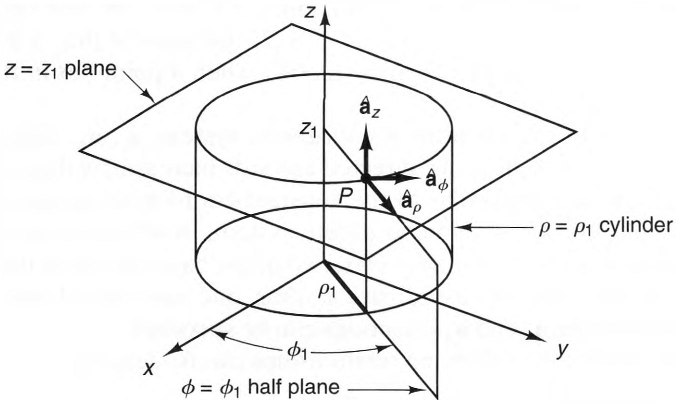


Figure 2-16 Position coordinates and base vectors in the cylindrical coordinate system.

2-4-2 THE CYLINDRICAL COORDINATE SYSTEM

The cylindrical coordinate system is a three-dimensional version of the polar coordinate system used in two-dimensional analysis.[4] Referring to Figure 2-16, we see that the position coordinates of a point P in this system are $u_1 = \rho$, $u_2 = \phi$, and $u_3 = z$. Here, ρ is defined as the perpendicular projection from the point to the z-axis, and ϕ is the angle that this projection makes with respect to the x-axis. The z-coordinate is the same as in Cartesian coordinates. All points are uniquely specified by the intersection of ρ = constant, ϕ = constant, and z = constant surfaces, where $0 < \rho < \infty$, $0 < \phi < 2\pi$, and $-\infty < z < \infty$. Using Figure 2-16, we can easily show that cylindrical and Cartesian coordinates are related by

$$\rho = \sqrt{x^2 + y^2} \tag{2.35a}$$

$$\phi = \tan^{-1}\left\{\frac{y}{x}\right\} \tag{2.35b}$$

$$z = z \tag{2.35c}$$

and

$$x = \rho \cos \phi \tag{2.36a}$$

$$y = \rho \sin \phi \tag{2.36b}$$

$$z = z. \tag{2.36c}$$

[4] There are many different cylindrical coordinate systems, such as circular cylindrical coordinates, elliptical cylindrical coordinates, and parabolic cylindrical coordinates. Throughout this text, however, we will refer to the circular cylindrical coordinate system as simply the cylindrical coordinate system.

Care must be taken when using Equation (2.35b), since $0 < \phi < 2\pi$, and the $\tan^{-1}$ function has a principal-value range of $-\pi/2 < \phi < \pi/2$. Because of this, π must be added[5] to the ϕ value of ϕ specified by Equation (2.35b) when a point lies in the second or third quadrants (i.e., $x < 0$).

The base unit vectors of the cylindrical coordinate system, $\hat{\mathbf{a}}_\rho$, $\hat{\mathbf{a}}_\phi$, and $\hat{\mathbf{a}}_z$, are depicted in Figure 2-16. These vectors are directed towards increasing values of ρ, ϕ, and z, respectively, and are perpendicular to the constant-coordinate surfaces of the other coordinates. Unlike the Cartesian coordinate system, in which all three base vectors maintain the same orientations at all points, two of the base vectors in the cylindrical coordinate system vary with the coordinate ϕ; thus, one must first define the ϕ coordinate of a point before the $\hat{\mathbf{a}}_\rho$ and $\hat{\mathbf{a}}_\phi$ directions can be specified.

From basic trigonometry, the following relationships can be derived:

$$\hat{\mathbf{a}}_\rho \cdot \hat{\mathbf{a}}_\rho = \hat{\mathbf{a}}_\phi \cdot \hat{\mathbf{a}}_\phi = \hat{\mathbf{a}}_z \cdot \hat{\mathbf{a}}_z = 1 \tag{2.37a}$$

$$\hat{\mathbf{a}}_\rho \cdot \hat{\mathbf{a}}_\phi = \hat{\mathbf{a}}_\rho \cdot \hat{\mathbf{a}}_z = \hat{\mathbf{a}}_\phi \cdot \hat{\mathbf{a}}_z = 0 \tag{2.37b}$$

$$\hat{\mathbf{a}}_\rho \times \hat{\mathbf{a}}_\rho = \hat{\mathbf{a}}_\phi \times \hat{\mathbf{a}}_\phi = \hat{\mathbf{a}}_z \times \hat{\mathbf{a}}_z = 0 \tag{2.37c}$$

$$\hat{\mathbf{a}}_\rho \times \hat{\mathbf{a}}_\phi = \hat{\mathbf{a}}_z \tag{2.37d}$$

$$\hat{\mathbf{a}}_\phi \times \hat{\mathbf{a}}_z = \hat{\mathbf{a}}_\rho \tag{2.37e}$$

$$\hat{\mathbf{a}}_z \times \hat{\mathbf{a}}_\rho = \hat{\mathbf{a}}_\phi . \tag{2.37f}$$

The cross products between the cylindrical base vectors can be symbolized using the aid shown in Figure 2-17. A vector $\mathbf{A}$ at any point can be represented by its components in the base vectors at that point:

$$\mathbf{A} = A_\rho \hat{\mathbf{a}}_\rho + A_\phi \hat{\mathbf{a}}_\phi + A_z \hat{\mathbf{a}}_z , \tag{2.38}$$

where the scalars A_ρ, A_ϕ, A_z are the ρ, ϕ, and z components of $\mathbf{A}$, respectively. Using Equation (2.15), we can find these components by taking the dot product of $\mathbf{A}$ with each of the base vectors:

$$A_i = \mathbf{A} \cdot \hat{\mathbf{a}}_i \quad i = \rho, \phi, \text{or } z . \tag{2.39}$$

The dot product of two vectors $\mathbf{A}$ and $\mathbf{B}$ can be expressed in terms of their components as

$$\mathbf{A} \cdot \mathbf{B} = (A_\rho \hat{\mathbf{a}}_\rho + A_\phi \hat{\mathbf{a}}_\phi + A_z \hat{\mathbf{a}}_z) \cdot (B_\rho \hat{\mathbf{a}}_\rho + B_\phi \hat{\mathbf{a}}_\phi + B_z \hat{\mathbf{a}}_z).$$

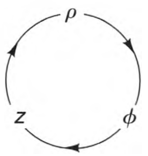


Figure 2-17 Circle diagram for cross products in cylindrical coordinates.

[5] Most calculators have a polar-to-rectangular function that automatically performs this function when x and y are specified separately.

Using the orthogonality properties of the base vectors, this becomes

$$\mathbf{A} \cdot \mathbf{B} = A_\rho B_\rho + A_\phi B_\phi + A_z B_z. \tag{2.40}$$

Since $|\mathbf{A}| = \sqrt{\mathbf{A} \cdot \mathbf{A}}$, it follows that

$$|\mathbf{A}| = \sqrt{A_\rho^2 + A_\phi^2 + A_z^2}, \tag{2.41}$$

which can also be derived from the Pythagorean theorem.

Similarly, the cross product of two vectors can be expressed as

$$\mathbf{A} \times \mathbf{B} = (A_\rho \hat{\mathbf{a}}_\rho + A_\phi \hat{\mathbf{a}}_\phi + A_z \hat{\mathbf{a}}_z) \times (B_\rho \hat{\mathbf{a}}_\rho + B_\phi \hat{\mathbf{a}}_\phi + B_z \hat{\mathbf{a}}_z),$$

which, using the properties of the base vectors, can be simplified to read

$$\mathbf{A} \times \mathbf{B} = (A_\phi B_z - A_z B_\phi)\hat{\mathbf{a}}_\rho + (A_z B_\rho - A_\rho B_z)\hat{\mathbf{a}}_\phi + (A_\rho B_\phi - A_\phi B_\rho)\hat{\mathbf{a}}_z. \tag{2.42}$$

This can also be written in shorthand as the determinant

$$\mathbf{A} \times \mathbf{B} = \begin{vmatrix} \hat{\mathbf{a}}_\rho & \hat{\mathbf{a}}_\phi & \hat{\mathbf{a}}_z \\ A_\rho & A_\phi & A_z \\ B_\rho & B_\phi & B_z \end{vmatrix}. \tag{2.43}$$

As shown in Figure 2-18, a differential volume dv is traced about a point when its coordinates are varied by the amounts $d\rho$, $d\phi$, and dz, respectively. In the limit as $d\rho$, $d\phi$, and dz approach zero, the enclosed volume dv can be written as

$$dv = \rho d\rho d\phi dz. \tag{2.44}$$

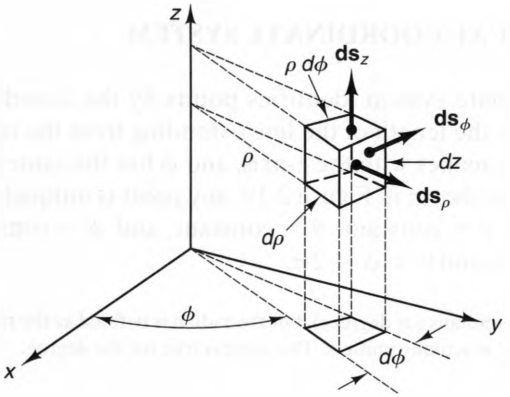


Figure 2-18 Differential volume and surface elements in the cylindrical coordinate system.

Notice that the factor ρ is necessary because the distance traced as the angular coordinate varies from ϕ to $\phi + d\phi$ equals $\rho d\phi$. This factor also makes the equation dimensionally correct since, strictly speaking, $d\phi$ is unitless.[6]

As can be deduced from Figure 2-18, the surface areas traced when two of the three coordinates at a point vary by differential amounts are

$$ds_\rho = \rho d\phi\, dz \quad (\text{when } d\rho = 0) \tag{2.45a}$$

$$ds_\phi = d\rho\, dz \quad (\text{when } d\phi = 0) \tag{2.45a}$$

$$ds_z = \rho d\rho\, d\phi \quad (\text{when } dz = 0). \tag{2.45a}$$

The differential surface vectors associated with these surfaces are found from these by adding the appropriate unit vectors:

$$\mathbf{ds}_\rho = ds_\rho \hat{\mathbf{a}}_\rho = \rho d\phi\, dz\, \hat{\mathbf{a}}_\rho \quad (\text{when } d\rho = 0) \tag{2.46a}$$

$$\mathbf{ds}_\phi = ds_\phi \hat{\mathbf{a}}_\phi = d\rho\, dz\, \hat{\mathbf{a}}_\phi \quad (\text{when } d\phi = 0) \tag{2.46b}$$

$$\mathbf{ds}_z = ds_z \hat{\mathbf{a}}_z = \rho d\rho\, d\phi \hat{\mathbf{a}}_z \quad (\text{when } dz = 0). \tag{2.46a}$$

Finally, the differential displacement vector that represents the directed distance from $P(\rho, \phi, z)$ to $P'(\rho + d\rho, \phi + d\phi, z + dz)$ along a line contour is

$$\mathbf{d\ell} = d\rho \hat{\mathbf{a}}_\rho + \rho d\phi \hat{\mathbf{a}}_\phi + dz\, \hat{\mathbf{a}}_z. \tag{2.47}$$

If the coordinates along the line are defined by $\rho(s)$, $\phi(s)$, and $z(s)$, then $d\rho$, $d\phi$, and dz can be found from the relations

$$d\rho = \frac{d\rho(s)}{ds} ds \tag{2.48a}$$

$$d\phi = \frac{d\phi(s)}{ds} ds \tag{2.48b}$$

$$dz = \frac{dz(s)}{ds} ds. \tag{2.48c}$$

2-4-3 THE SPHERICAL COORDINATE SYSTEM

The spherical coordinate system identifies points by the coordinates $u_1 = r$, $u_2 = \theta$, and $u_3 = \phi$, where r is the length of the line extending from the origin to the point, θ is the angle that this line makes with the z-axis, and ϕ has the same definition as in cylindrical coordinates. As shown in Figure 2-19, any point is uniquely defined as the point of intersection of the r = constant, θ = constant, and ϕ = constant surfaces, where $0 < r < \infty$, $0 < \theta < \pi$, and $0 < \phi < 2\pi$.

[6] The units of ϕ and $d\phi$ are radians (or degrees), but the radian is defined as the ratio of arc length to the circumference of a circle, so it is actually unitless. The same is true for the degree.

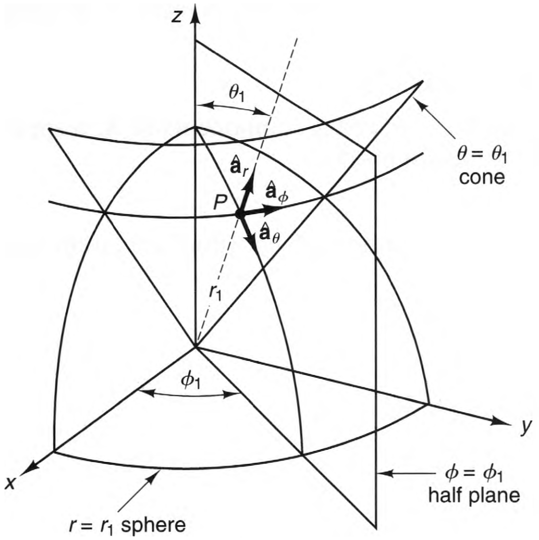


Figure 2-19 Position coordinates and base vectors in the spherical coordinate system.

The spherical and Cartesian coordinates of a point are related by

$$r = \sqrt{x^2 + y^2 + z^2} \tag{2.49a}$$

$$\theta = \cos^{-1}\left\{\frac{z}{\sqrt{x^2 + y^2 + z^2}}\right\} \tag{2.49b}$$

$$\phi = \tan^{-1}\left\{\frac{y}{x}\right\} \tag{2.49c}$$

and

$$x = r \sin\theta \cos\phi \tag{2.50a}$$

$$y = r \sin\theta \sin\phi \tag{2.50b}$$

$$z = r \cos\theta. \tag{2.50c}$$

As in the case of cylindrical coordinates, care must be exercised when using Equation (2.49c) to ensure that the calculated angle ϕ lies in the correct quadrant.

The base unit vectors in the cylindrical coordinate system, $\hat{\mathbf{a}}_r$, $\hat{\mathbf{a}}_\theta$, and $\hat{\mathbf{a}}_\phi$, are directed towards increasing values of r, θ, and ϕ, respectively. As can be seen from Figure 2-19, all three of these vectors are functions of the coordinates θ and ϕ. Thus, the coordinates of a point must be specified before the base unit vectors can be specified.

At any point, a vector **A** can be expressed in terms of its components in the base vector directions as

$$\mathbf{A} = A_r\hat{\mathbf{a}}_r + A_\theta\hat{\mathbf{a}}_\theta + A_\phi\hat{\mathbf{a}}_\phi, \tag{2.51}$$

where A_r, A_θ, and A_ϕ are the r, θ, and ϕ components of **A**, respectively. These components can be found via the dot product

$$A_i = \mathbf{A}\cdot\hat{\mathbf{a}}_i \quad i = r, \theta, \text{or } \phi. \tag{2.52}$$

The base vectors of the spherical coordinate system satisfy the following relationships:

$$\hat{\mathbf{a}}_r\cdot\hat{\mathbf{a}}_r = \hat{\mathbf{a}}_\theta\cdot\hat{\mathbf{a}}_\theta = \hat{\mathbf{a}}_\phi\cdot\hat{\mathbf{a}}_\phi = 1 \tag{2.53a}$$

$$\hat{\mathbf{a}}_r\cdot\hat{\mathbf{a}}_\theta = \hat{\mathbf{a}}_r\cdot\hat{\mathbf{a}}_\phi = \hat{\mathbf{a}}_\theta\cdot\hat{\mathbf{a}}_\phi = 0 \tag{2.53b}$$

$$\hat{\mathbf{a}}_r\times\hat{\mathbf{a}}_r = \hat{\mathbf{a}}_\theta\times\hat{\mathbf{a}}_\theta = \hat{\mathbf{a}}_\phi\times\hat{\mathbf{a}}_\phi = 0 \tag{2.53c}$$

$$\hat{\mathbf{a}}_r\times\hat{\mathbf{a}}_\theta = \hat{\mathbf{a}}_\phi \tag{2.53d}$$

$$\hat{\mathbf{a}}_\theta\times\hat{\mathbf{a}}_\phi = \hat{\mathbf{a}}_r \tag{2.53e}$$

$$\hat{\mathbf{a}}_\phi\times\hat{\mathbf{a}}_r = \hat{\mathbf{a}}_\theta. \tag{2.53f}$$

The cross-product relationships between the spherical base vectors can be symbolized using the aid shown in Figure 2-20. The dot product of any two vectors can be expressed as

$$\begin{aligned}\mathbf{A}\cdot\mathbf{B} &= (A_r\hat{\mathbf{a}}_r + A_\theta\hat{\mathbf{a}}_\theta + A_\phi\hat{\mathbf{a}}_\phi)\cdot(B_r\hat{\mathbf{a}}_r + B_\theta\hat{\mathbf{a}}_\theta + B_\phi\hat{\mathbf{a}}_\phi)\\ &= A_rB_r + A_\theta B_\theta + A_\phi B_\phi.\end{aligned} \tag{2.54}$$

Also, since $|\mathbf{A}| = \sqrt{\mathbf{A}\cdot\mathbf{A}}$, it follows that

$$|\mathbf{A}| = \sqrt{A_r^2 + A_\theta^2 + A_\phi^2}, \tag{2.55}$$

which can also be derived from the Pythagorean theorem.

Similarly, the cross product of two vectors can be expressed as

$$\mathbf{A}\times\mathbf{B} = (A_r\hat{\mathbf{a}}_r + A_\theta\hat{\mathbf{a}}_\theta + A_\phi\hat{\mathbf{a}}_\phi)\times(B_r\hat{\mathbf{a}}_r + B_\theta\hat{\mathbf{a}}_\theta + B_\phi\hat{\mathbf{a}}_\phi),$$

which reduces to

$$\mathbf{A}\times\mathbf{B} = (A_\theta B_\phi - A_\phi B_\theta)\hat{\mathbf{a}}_r + (A_\phi B_r - A_rB_\phi)\hat{\mathbf{a}}_\theta + (A_rB_\theta - A_\theta B_r)\hat{\mathbf{a}}_\phi. \tag{2.56}$$

This expression can be written in shorthand form as the determinant

$$\mathbf{A}\times\mathbf{B} = \begin{vmatrix}\hat{\mathbf{a}}_r & \hat{\mathbf{a}}_\theta & \hat{\mathbf{a}}_\phi\\ A_r & A_\theta & A_\phi\\ B_r & B_\theta & B_\phi\end{vmatrix}. \tag{2.57}$$

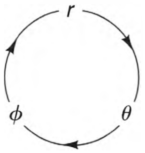


Figure 2-20 Circle diagram for cross products in spherical coordinates.

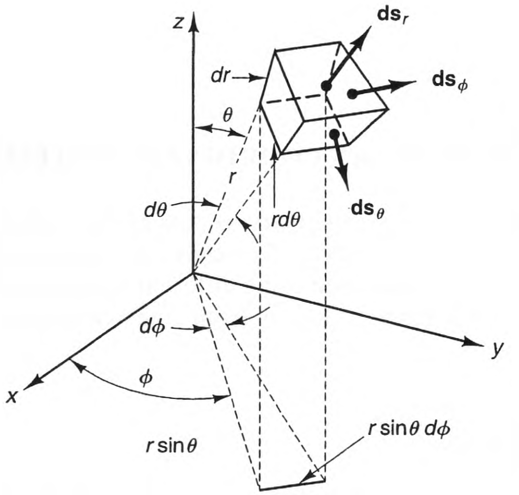


Figure 2-21 Differential volume and surface elements in the spherical coordinate system.

As shown in Figure 2-21, a differential volume dv is traced about a point $P(r, \theta, \phi)$ when its coordinates are varied by the amounts dr, $d\theta$, and $d\phi$, respectively. In the limit as dr, $d\phi$, and dz approach zero, the enclosed volume is

$$dv = r^2 \sin\theta\, dr\, d\theta\, d\phi. \tag{2.58}$$

In this expression, the multiplier $r^2 \sin\theta$ is necessary because the lengths traced by differential changes in θ and ϕ are $rd\theta$ and $r\sin\theta\, d\phi$, respectively. Also, the r^2 makes the expression dimensionally correct.

The surfaces traced when two of the three coordinates are varied by differential amounts are also shown in Figure 2-21. They have areas given by

$$ds_r = r^2 \sin\theta\, d\theta\, d\phi \quad (\text{when } dr = 0) \tag{2.59a}$$

$$ds_\theta = r\sin\theta\, dr\, d\phi \quad (\text{when } d\theta = 0) \tag{2.59b}$$

$$ds_\phi = r\, dr\, d\theta \quad (\text{when } d\phi = 0), \tag{2.59c}$$

and their associated differential surface vectors are

$$\mathbf{ds}_r = ds_r \hat{\mathbf{a}}_r = r^2 \sin\theta\, d\theta\, d\phi\, \hat{\mathbf{a}}_r \quad (\text{when } dr = 0) \tag{2.60a}$$

$$\mathbf{ds}_\theta = ds_\theta \hat{\mathbf{a}}_\theta = r\sin\theta\, dr\, d\phi\, \hat{\mathbf{a}}_\theta \quad (\text{when } d\theta = 0) \tag{2.60b}$$

$$\mathbf{ds}_\phi = ds_\phi \hat{\mathbf{a}}_\phi = r\, dr\, d\theta\, \hat{\mathbf{a}}_\phi \quad (\text{when } d\phi = 0)\ . \tag{2.60c}$$

Finally, the differential displacement vector that represents the directed distance from $P(r, \theta, \phi)$ to $P'(r + dr, \theta + d\theta, \phi + d\phi)$ along a line contour is

$$\mathbf{d\ell} = dr\, \hat{\mathbf{a}}_r + r\, d\theta\, \hat{\mathbf{a}}_\theta + r\sin\theta\, d\phi\, \hat{\mathbf{a}}_\phi. \tag{2.61}$$

If the coordinates along the line are given by $r(s)$, $\theta(s)$, and $\phi(s)$, then dr, $d\theta$, and $d\phi$ can be found from

$$dr = \frac{dr(s)}{ds} ds \tag{2.62a}$$

$$d\theta = \frac{d\theta(s)}{ds}ds \tag{2.62b}$$

$$d\phi = \frac{d\phi(s)}{ds}ds. \tag{2.62c}$$

2-4-4 CONVERSIONS BETWEEN COORDINATE SYSTEMS

There are many times when it is necessary to change the representation of a vector from one coordinate system to another. Typically, this is done when different aspects of a problem are most easily described using different coordinate representations.

Changing a vector's representation from one coordinate system to another requires two steps:

- Convert the coordinates
- Convert the components

The position coordinates in the new system are found simply by applying the appropriate coordinate transformations. The components in the new system are found by taking the dot product of the vector with each of the base vectors in the new system:

$$A_i = \mathbf{A} \cdot \hat{\mathbf{a}}_i, \tag{2.63}$$

where i is set equal to each of the coordinate variables in the new system.

Appendix B contains a number of tables that are helpful when converting the representation of a vector from one coordinate system to another. Table B-1 contains the relationships between the coordinate variables of the three coordinate systems. Table B-2 contains the dot products of the base vectors of the three coordinate systems. Finally, Table B-3 summarizes the relationships between vector components in these coordinate systems.

Example 2-3

Find the representation of $\mathbf{C} = \rho\,\hat{\mathbf{a}}_\phi$ in Cartesian coordinates.

Solution:

Using Equation (2.63) in conjunction with the values in Table B-2, we find that the Cartesian components of **C** are

$$C_x = \mathbf{C} \cdot \hat{\mathbf{a}}_x = \rho\hat{\mathbf{a}}_\phi \cdot \hat{\mathbf{a}}_x = -\rho\sin\phi$$

$$C_y = \mathbf{C} \cdot \hat{\mathbf{a}}_y = \rho\hat{\mathbf{a}}_\phi \cdot \hat{\mathbf{a}}_y = \rho\cos\phi$$

$$C_z = \mathbf{C} \cdot \hat{\mathbf{a}}_z = \rho\hat{\mathbf{a}}_\phi \cdot \hat{\mathbf{a}}_z = 0\,.$$

Next, using $x = \rho\cos\phi$ and $y = \rho\sin\phi$, we obtain

$$\mathbf{C} = C_x\hat{\mathbf{a}}_x + Cy\hat{\mathbf{a}}_y + C_z\hat{\mathbf{a}}_z = -\rho\sin\phi\,\hat{\mathbf{a}}_x + \rho\cos\phi\,\hat{\mathbf{a}}_y$$

$$= -y\hat{\mathbf{a}}_x + x\hat{\mathbf{a}}_y\,.$$

Example 2-4

Find the representation of $\mathbf{A} = 3y\hat{\mathbf{a}}_x$ in the spherical coordinate system.

Solution:

Knowing that $y = r\sin\theta\sin\phi$, we can write $\mathbf{A}$ as

$$\mathbf{A} = 3r\sin\theta\sin\phi\,\hat{\mathbf{a}}_x.$$

Thus, $A_x = 3r\sin\theta\sin\phi$, and $A_y = A_z = 0$. Using Table B-3, we obtain the spherical components of $\mathbf{A}$:

$$A_r = A_x\sin\theta\cos\phi = 3r\sin^2\theta\sin\phi\cos\phi$$

$$A_\theta = A_x\cos\theta\cos\phi = 3r\sin\theta\cos\theta\sin\phi\cos\phi = \frac{3}{4}r\sin2\theta\sin2\phi$$

$$A_\phi = -A_x\sin\phi = -3r\sin\theta\sin^2\phi.$$

Thus, the representation of $\mathbf{A}$ in spherical coordinates is

$$\mathbf{A} = 3r\sin^2\theta\sin\phi\cos\phi\,\hat{\mathbf{a}}_r + \frac{3}{4}r\sin2\theta\sin2\phi\,\hat{\mathbf{a}}_\theta - 3r\sin\theta\sin^2\phi\,\hat{\mathbf{a}}_\phi.$$

2-4-5 THE POSITION VECTOR

We have already seen that any point is uniquely defined by its position coordinates. We can also identify a point by its ***position vector***:

The position vector of a point is defined as the directed distance from the origin to the point and is represented by the symbol **r**.

Every point has a unique position vector that identifies it. This vector is denoted by the symbol **r** and is depicted in Figure 2-22 for an arbitrary point P. The position vector of an arbitrary point has the following representations in the Cartesian, cylindrical, and spherical coordinate systems:

$$\mathbf{r} = \begin{cases} x\hat{\mathbf{a}}_x + y\hat{\mathbf{a}}_y + z\hat{\mathbf{a}}_z & \text{(Cartesian)} \quad (2.64a) \\ \rho\hat{\mathbf{a}}_\rho + z\hat{\mathbf{a}}_z & \text{(cylindrical)} \quad (2.64b) \\ r\hat{\mathbf{a}}_r & \text{(spherical)} \quad (2.64c) \end{cases}$$

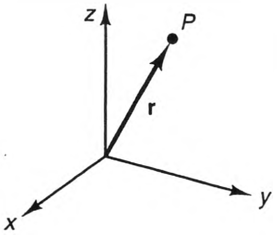


Figure 2-22 The position vector.

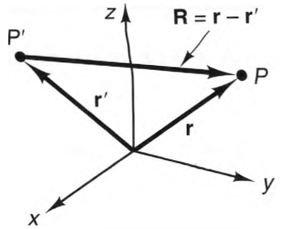


Figure 2-23 The directed distance **R** between two arbitrary points.

The spherical representation follows directly from the definition of **r**. The other two representations can be found through coordinate transformations.

The directed distance between two points can be expressed in terms of their position vectors. Referring to Figure 2-23, consider the points P and P', represented by the position vectors **r** and **r**′, respectively. If we let **R** denote the directed distance from P' to P, it follows from Figure 2-23 that

$$\mathbf{R} = \mathbf{r} - \mathbf{r}'$$

We can also write **R** as

$$\mathbf{R} = R\hat{\mathbf{a}}_R, \tag{2.65}$$

where

$$R = |\mathbf{r} - \mathbf{r}'| \tag{2.66}$$

and

$$\hat{\mathbf{a}}_R = \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|}. \tag{2.67}$$

In Cartesian coordinates, **R** is represented by

$$\mathbf{R} = \mathbf{r} - \mathbf{r}' = (x - x')\hat{\mathbf{a}}_x + (y - y')\hat{\mathbf{a}}_y + (z - z')\hat{\mathbf{a}}_z. \tag{2.68}$$

In coordinate systems other than Cartesian coordinates, it is important to remember that the base unit vectors at **r** and **r**′ are not, in general, the same. This is demonstrated in the following example.

Example 2-5

Find the directed distance from $P'(3, 30°, 1)$ to $P(1, 90°, 2)$. Write the representation of this vector at both points using the cylindrical coordinate system.

Solution:

Since the values of ϕ at these two points are different, it is best to start by finding $\mathbf{R} = \mathbf{r} - \mathbf{r}'$ in the Cartesian coordinate system. Converting P and P' to Cartesian coordinates yields

$$P(1, 90°, 2) = P(\cos 90°, \sin 90°, 2) = P(0, 1, 2)$$

$$P'(3, 30°, 1) = P'(3\cos 30°, 3\sin 30°, 1) = P'(2.6, 1.5, 1).$$

The corresponding position vectors are

$$\mathbf{r} = \hat{\mathbf{a}}_y + 2\hat{\mathbf{a}}_z$$

$$\mathbf{r}' = 2.6\hat{\mathbf{a}}_x + 1.5\hat{\mathbf{a}}_y + \hat{\mathbf{a}}_z.$$

so the directed distance from P' to P is

$$\mathbf{R} = \mathbf{r} - \mathbf{r}' = -2.6\hat{\mathbf{a}}_x - 0.5\hat{\mathbf{a}}_y + \hat{\mathbf{a}}_z.$$

Using Table B-3, we can express $\mathbf{R}$ in cylindrical coordinates at any point as

$$\mathbf{R} = R_\rho\hat{\mathbf{a}}_\rho + R_\phi\hat{\mathbf{a}}_\phi + R_z\hat{\mathbf{a}}_z,$$

where

$$R_\rho = -2.6\cos\phi - 0.5\sin\phi$$

$$R_\phi = 2.6\sin\phi - 0.5\cos\phi$$

$$R_z = 1,$$

and ϕ is the position coordinate at the point. Substituting, $\phi = 90°$ at P and $\phi = 30°$ at P', we find that

$$\mathbf{R} = -0.5\hat{\mathbf{a}}_\rho + 2.6\hat{\mathbf{a}}_\phi + \hat{\mathbf{a}}_z \quad \text{at } P$$

and

$$\mathbf{R} = -2.5\hat{\mathbf{a}}_{\rho'} + 0.87\hat{\mathbf{a}}_{\phi'} + \hat{\mathbf{a}}_z \quad \text{at } P'$$

where $\hat{\mathbf{a}}_\rho$ and $\hat{\mathbf{a}}_\phi$ are base unit vectors at P, and $\hat{\mathbf{a}}_{\rho'}$ and $\hat{\mathbf{a}}_{\phi'}$ are the base vectors at P'. Notice that although $\mathbf{R}$ is the same vector at P and P', its representations at these two points are different.

2-5 The Calculus of Scalar and Vector Fields

Most of the physical quantities of interest in electromagnetics are field quantities. Because they are functions of position, it is important that we be able to characterize the functional behaviors of field quantities over both large and small regions of space. This is accomplished through various integral and differential operators.

2-5-1 INTEGRALS OF SCALAR AND VECTOR FIELDS

Electromagnetic phenomena are often described in terms of integrals of vector or scalar quantities over a volume, a surface, or a line. Examples of the kinds of integrals encountered in electromagnetic analysis are

$$\int_V Q\,dv \tag{2.69}$$

$$\int_V \mathbf{J}\,dv \tag{2.70}$$

$$\int_S \mathbf{D}\cdot\mathbf{ds} \tag{2.71}$$

$$\int_C \mathbf{E}\cdot\mathbf{d\ell}. \tag{2.72}$$

The first two of these integrals are called ***volume integrals***, because they take place throughout a specified volume. Likewise, the third and fourth integrals are called ***surface*** and ***line*** (or ***contour***) ***integrals***, respectively, because they are evaluated over a surface and a line, respectively.

In spite of the obvious differences between the integrals given in Equations (2.69–2.72), each is simply the summation of a differential quantity (either scalar or vector) over a range of points. The basic steps for evaluating any of these integrals are:

1. Choose the coordinate system that will be used during the integration process.
2. Determine which position coordinates vary during the integration process.
3. Select the appropriate differential quantity.
4. If the integrand is a vector, make sure that all unit vectors are constants with respect to the variable(s) of integration.
5. Integrate over the appropriate limits of the position coordinates.

The three examples that follow demonstrate the general procedure for evaluating integrals of field quantities.

Example 2-6

Evaluate the integral $\int_V \mathbf{P}dv$, where $\mathbf{P} = r\cos\phi\,\hat{\mathbf{a}}_r$ and V is a sphere of unit radius centered at the origin.

Solution:

The spherical coordinate variables are the most convenient for this problem. To cover all points within the volume, the range of the coordinates must be $0 < r < 1, 0 < \theta < \pi, 0 < \phi < 2\pi$. Also, $dv = r^2\sin\theta\, dr d\theta d\phi$ at all points within the volume. Substituting into the integral, we have

$$\int_V \mathbf{P}\,dv = \int_0^{2\pi}\int_0^{\pi}\int_0^1 r^3\cos\phi\,\sin\theta\,\hat{\mathbf{a}}_r dr\,d\theta\,d\phi.$$

This integral is not as easy to evaluate as it may first appear, because of the presence of the unit vector $\hat{\mathbf{a}}_r$, which varies with the position variables θ and ϕ. Using Table B-3, however, we can represent $\hat{\mathbf{a}}_r$ in Cartesian components as

$$\hat{\mathbf{a}}_r = \sin\theta\cos\phi\,\hat{\mathbf{a}}_x + \sin\theta\sin\phi\,\hat{\mathbf{a}}_y + \cos\theta\,\hat{\mathbf{a}}_z.$$

Substituting, we obtain

$$\int_V \mathbf{P}\,dv = \hat{\mathbf{a}}_x\int_0^{2\pi}\int_0^{\pi}\int_0^1 r^3\sin^2\theta\cos^2\phi\,drd\theta d\phi + \hat{\mathbf{a}}_y\int_0^{2\pi}\int_0^{\pi}\int_0^1 r^3\sin^2\theta\sin\phi\cos\phi\,drd\theta d\phi$$
$$+ \hat{\mathbf{a}}_z\int_0^{2\pi}\int_0^{\pi}\int_0^1 r^3\sin\theta\cos\theta\cos\phi\,dr\,d\theta\,d\phi.$$

The second and third integrals on the right-hand side of this expression are zero, since $\int_0^{2\pi}\sin\phi\cos\phi\,d\phi = 0$ and $\int_0^{2\pi}\cos\phi\,d\phi = 0$, respectively, leaving

$$\int_V \mathbf{P}\,dv = \hat{\mathbf{a}}_x\int_0^{2\pi}\int_0^{\pi}\int_0^1 r^3\sin^2\theta\cos^2\phi dr\,d\theta\,d\phi = \frac{\hat{\mathbf{a}}_x}{4}\int_0^{2\pi}\int_0^{\pi}\sin^2\theta\cos^2\phi\,d\theta\,d\phi$$

$$= \frac{\pi\hat{\mathbf{a}}_x}{8}\int_0^{2\pi}\cos^2\phi\,d\phi = \frac{\pi^2}{8}\hat{\mathbf{a}}_x.$$

Example 2-7

Evaluate the integral $\oint_S \mathbf{F} \cdot \mathbf{ds}$, where $\mathbf{F} = x\hat{\mathbf{a}}_x$ and S is the closed circular cylinder shown in Figure 2-24.

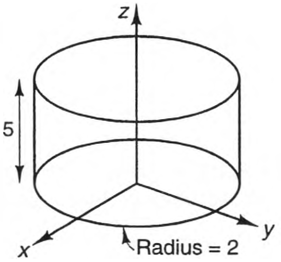


Figure 2-24 A circular cylinder.

Solution:

From Figure 2-24, we see that S consists of two discs at $z = 0$ and $z = 5$, respectively, and an open cylinder $\rho = 2$ for $0 < \phi < 2\pi$. On the discs, $\mathbf{ds} = \pm \rho \, d\rho \, d\phi \, \hat{\mathbf{a}}_z$, where the upper and lower signs correspond to the upper and lower discs, respectively. Since $\mathbf{F}$ has no z component, $\mathbf{F} \cdot \mathbf{ds} = 0$ everywhere on both discs.

On the open cylinder, $\mathbf{ds} = \rho \, d\phi \, dz \, \hat{\mathbf{a}}_\rho = 2 \, d\phi \, dz \, \hat{\mathbf{a}}_\rho$. Remembering that $x = \rho \cos \phi$, we can write $\mathbf{F} \cdot ds$ as

$$\mathbf{F} \cdot \mathbf{ds} = x\hat{\mathbf{a}}_x \cdot 2 \, d\phi \, dz \, \hat{\mathbf{a}}_\rho = 4\cos^2 \phi \, d\phi \, dz ,$$

where the value of $\hat{\mathbf{a}}_x \cdot \hat{\mathbf{a}}_\rho = \cos \phi$ and $x = \rho\cos \phi$ were obtained from Tables B-2 and B-1, respectively. Substituting into the integral yields

$$\oint_S \mathbf{F} \cdot \mathbf{ds} = \int_0^5 \int_0^{2\pi} 4 \cos^2 \phi \, d\phi \, dz = \int_0^5 4\pi \, dz = 20\pi .$$

Example 2-8

For $\mathbf{F} = y\hat{\mathbf{a}}_x - x\hat{\mathbf{a}}_y$, evaluate the line integral $\int_C \mathbf{F} \cdot \mathbf{d\ell}$ along two paths shown in Figure 2-25, each starting at (0, 0, 0) and ending at (1, 2, 4).

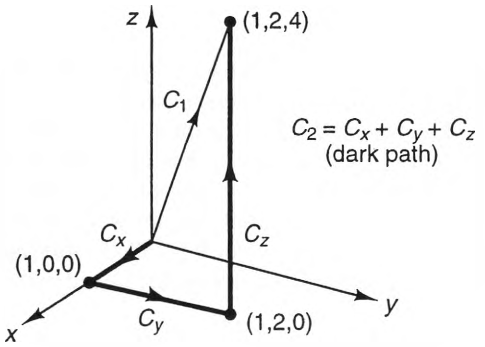


Figure 2-25 Two paths connecting the points (0, 0, 0) and (1, 2, 4).

Solution:

For either path, the differential displacement vector can be written in the form $\mathbf{d\ell} = dx\,\hat{\mathbf{a}}_x + dy\,\hat{\mathbf{a}}_y + dz\,\hat{\mathbf{a}}_z$. For the vector $\mathbf{F}$ given in this problem, the dot product $\mathbf{F}\cdot\mathbf{d\ell}$ is

$$\mathbf{F}\cdot\mathbf{d\ell} = ydx - xdy.$$

a) The path C_1 is a straight line, which can be described by the equations

$$y = 2x$$

$$z = 4x.$$

Since both y and z can be written as functions of x, we can consider x as the parametric variable for use in Equations (2.34a) through (2.34c). Using these equations, we obtain

$$dx = \frac{dx}{dx}dx = dx$$

$$dy = \frac{dy(x)}{dx}dx = 2dx$$

$$dz = \frac{dz(x)}{dx}dx = 4dx.$$

Substituting these expressions for dx, dy, and dz yields

$$\mathbf{F}\cdot\mathbf{d\ell} = ydx - xdy = 2xdx - 2xdx = 0dx.$$

This means that $\mathbf{F}$ is perpendicular to $\mathbf{d\ell}$ at every point along the path C_1. Integrating, we obtain the result

$$\int_{C_1}\mathbf{F}\cdot\mathbf{d\ell} = \int_0^1 0dx = 0.$$

b) Path C_2 is actually a collection of three straight-line paths: C_a from (0, 0, 0) to (1, 0, 0), C_b from (1, 0, 0) to (1, 2, 0), and C_c from (1, 2, 0) to (1, 2, 4). We have

$$\int_{C_2}\mathbf{F}\cdot\mathbf{d\ell} = \int_{C_x}\mathbf{F}\cdot\mathbf{d\ell} + \int_{C_y}\mathbf{F}\cdot\mathbf{d\ell} + \int_{C_z}\mathbf{F}\cdot\mathbf{d\ell}.$$

These line integrals are simple to evaluate, since only one position variable varies along each path. Thus, along the paths C_x, C_y and C_z, we have $\mathbf{d\ell} = dx\,\hat{\mathbf{a}}_x$, $\mathbf{d\ell} = dy\,\hat{\mathbf{a}}_y$, and $\mathbf{d\ell} = dz\,\hat{\mathbf{a}}_z$, respectively. Substituting these into the integrals and taking the dot products with $\mathbf{F}$, we obtain

$$\int_{C_2}\mathbf{F}\cdot\mathbf{d\ell} = \int_0^1 y\,dx\bigg|_{y=0} - \int_0^2 x\,dy\bigg|_{x=1} - \int_0^4 0\,dz = -2.$$

Since different answers were obtained when integrating $\mathbf{F}\cdot\mathbf{d\ell}$ along two different paths that connect the same endpoints, $\mathbf{F}$ is called a ***nonconservative*** vector field. This name comes from mechanics, where, if $\mathbf{F}$ represents a force, the integral $\int_C \mathbf{F}\cdot\mathbf{d\ell}$ equals the work done on an object as it moves along the path C. For the vector $\mathbf{F}$ in this problem, the net work done in moving the object from the origin to the point (1, 2, 4) along C_2 and back to the origin along path C_1 would be $-2 - 0 = -2 \neq 0$, which means that the net work done along this closed path is nonzero. Since work is not conserved, the vector $\mathbf{F}$ is called a ***nonconservative vector.*** On the other hand, vectors for which $\oint_C \mathbf{F}\cdot\mathbf{d\ell} = 0$ for all possible closed paths C are called ***conservative vectors.***

2-5-2 THE GRADIENT OF A SCALAR FIELD

Up to this point we have described scalar fields in terms of the rules that determine their values at each point in space. Often, however, the *rate* at which a scalar changes close to a point is more important than its value at the point itself. When walking up an incline, for example, one is usually more concerned about the change in altitude encountered with each step than with the altitude of each point relative to sea level. The gradient operation provides this kind of information.

To start our discussion, let us consider the change in the value of an arbitrary scalar field f as we move from (x, y, z) to $(x + dx, y + dy, z + dz)$. We will denote this change as df. From ordinary multivariable calculus,

$$df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial z}dz. \tag{2.73}$$

This expression can be written as the following dot product between two vectors:

$$df = \left(\frac{\partial f}{\partial x}\hat{\mathbf{a}}_x + \frac{\partial f}{\partial y}\hat{\mathbf{a}}_y + \frac{\partial f}{\partial z}\hat{\mathbf{a}}_z\right) \bullet (dx\,\hat{\mathbf{a}}_x + dy\,\hat{\mathbf{a}}_y + dz\,\hat{\mathbf{a}}_z). \tag{2.74}$$

The vector on the far right is simply the differential displacement vector $\mathbf{d\ell}$ along the path of movement (Equation (2.33)), so we can write Equation (2.74) in the form

$$df = \left(\frac{\partial f}{\partial x}\hat{\mathbf{a}}_x + \frac{\partial f}{\partial y}\hat{\mathbf{a}}_y + \frac{\partial f}{\partial z}\hat{\mathbf{a}}_z\right) \bullet \mathbf{d\ell}. \tag{2.75}$$

The vector quantity in parentheses is called the ***gradient*** of f and is denoted symbolically by **grad** f. Hence,

$$df = \mathbf{grad}\, f \bullet \mathbf{d\ell}, \tag{2.76}$$

where, in the Cartesian coordinate system,

$$\mathbf{grad}\, f = \frac{\partial f}{\partial x}\hat{\mathbf{a}}_x + \frac{\partial f}{\partial y}\hat{\mathbf{a}}_y + \frac{\partial f}{\partial z}\hat{\mathbf{a}}_z \qquad \text{(Cartesian coordinates).} \tag{2.77}$$

We can also write Equation (2.77) in the shorthand form,

$$\mathbf{grad}\, f = \boldsymbol{\nabla} f = \left(\frac{\partial f}{\partial x}\hat{\mathbf{a}}_x + \frac{\partial f}{\partial y}\hat{\mathbf{a}}_y + \frac{\partial f}{\partial z}\hat{\mathbf{a}}_z\right) \qquad \text{(Cartesian coordinates)} \tag{2.78}$$

where $\boldsymbol{\nabla}$ is called the ***del operator*** and is defined by

$$\boldsymbol{\nabla} \equiv \frac{\partial}{\partial x}\hat{\mathbf{a}}_x + \frac{\partial}{\partial y}\hat{\mathbf{a}}_y + \frac{\partial}{\partial z}\hat{\mathbf{a}}_z \qquad \text{(Cartesian coordinates).} \tag{2.79}$$

Strictly speaking, the del (∇) operator is not a true vector, since its components are operators, rather than numbers. Nevertheless, it is convenient to treat it like a vector in product equations such as Equation (2.79) and several others in this chapter. Throughout the remainder of this text, we always represent the vector **grad** f as ∇f.

Before we derive the representations of ∇f in the other coordinate systems, let us determine general properties of the gradient operation. Using the definition of the dot product (Equation (2.8)), we can write Equation (2.76) in the form

$$df = \nabla f \cdot \mathbf{d\ell} = |\nabla f| d\ell \cos\theta, \tag{2.80}$$

where θ is the angle between ∇f and $\mathbf{d\ell}$. Dividing both sides by $d\ell$ yields

$$\frac{df}{d\ell} = |\nabla f| \cos\theta. \tag{2.81}$$

When the direction of the path is parallel to ∇f, $\cos\theta = 1$. Along such a path, $df/d\ell$ attains its maximum value. Thus,

$$\left.\frac{df}{d\ell}\right|_{\max} = |\nabla f|. \tag{2.82}$$

Using Equation (2.82), we can now define ∇f as

$$\nabla f \equiv \left.\frac{\partial f}{\partial \ell}\right|_{\max} \hat{\mathbf{a}}_n, \tag{2.83}$$

where $\hat{\mathbf{a}}_n$ points in the direction of maximum increase in f. This definition is valid in all coordinate systems. Thus, the gradient ∇f is a vector that points in the direction of maximum rate of increase of the function f.

From Equation (2.80) we see that $df = 0$ whenever $\mathbf{d\ell}$ is perpendicular to ∇f. Thus, ∇f is always perpendicular to surfaces over which f is constant. This can be seen from Figure 2-26, which shows several surfaces of constant value for a scalar function f. In the figure, ∇f is perpendicular to each of these surfaces and points towards increasing values of f.

Representations of ∇f can also be found in the cylindrical and spherical coordinate systems. In cylindrical coordinates, the total differential of a scalar function f is

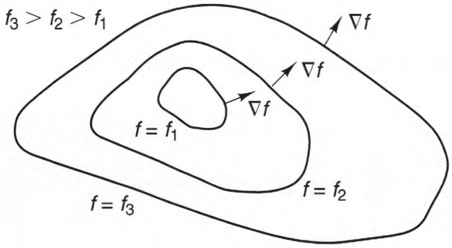


Figure 2-26 Equivalue surfaces and gradient vectors for an arbitrary function f.

$$df = \frac{\partial f}{\partial \rho} d\rho + \frac{\partial f}{\partial \phi} d\phi + \frac{\partial f}{\partial z} dz,$$

which can be rewritten as the dot product

$$df = \left(\frac{\partial f}{\partial \rho}\hat{\mathbf{a}}_\rho + \frac{1}{\rho}\frac{\partial f}{\partial \phi}\hat{\mathbf{a}}_\phi + \frac{\partial f}{\partial z}\hat{\mathbf{a}}_z\right) \bullet (d\rho\hat{\mathbf{a}}_\rho + \rho d\phi\hat{\mathbf{a}}_\phi + dz\hat{\mathbf{a}}_z).$$

Since the second vector on the right side of this equation is the differential displacement vector $\mathbf{d\ell}$ in cylindrical coordinates (see Equation (2.47)), df can be in the form

$$df = \left(\frac{\partial f}{\partial \rho}\hat{\mathbf{a}}_\rho + \frac{1}{\rho}\frac{\partial f}{\partial \phi}\hat{\mathbf{a}}_\phi + \frac{\partial f}{\partial z}\hat{\mathbf{a}}_z\right) \bullet \mathbf{d\ell}.$$

Comparing this with Equation (2.80), we see that the vector in the parentheses must be ∇f. Thus, we have

$$\nabla f = \left(\frac{\partial f}{\partial \rho}\hat{\mathbf{a}}_\rho + \frac{1}{\rho}\frac{\partial f}{\partial \phi}\hat{\mathbf{a}}_\phi + \frac{\partial f}{\partial z}\hat{\mathbf{a}}_z\right) \quad \text{(cylindrical coordinates).} \qquad (2.84)$$

Similarly, in spherical coordinates, we can write

$$df = \frac{\partial f}{\partial r} dr + \frac{\partial f}{\partial \theta} d\theta + \frac{\partial f}{\partial \phi} d\phi,$$

or

$$df = \left(\frac{\partial f}{\partial r}\hat{\mathbf{a}}_r + \frac{1}{r}\frac{\partial f}{\partial \theta}\hat{\mathbf{a}}_\theta + \frac{1}{r\sin\theta}\frac{\partial f}{\partial \phi}\hat{\mathbf{a}}_\phi\right) \bullet (dr\hat{\mathbf{a}}_r + rd\theta\hat{\mathbf{a}}_\theta + r\sin\theta\, d\phi\hat{\mathbf{a}}_\phi).$$

The second vector on the right side of this expression is $\mathbf{d\ell}$ (see Equation (2.61)), so this expression can be rewritten as

$$df = \left(\frac{\partial f}{\partial r}\hat{\mathbf{a}}_r + \frac{1}{r}\frac{\partial f}{\partial \theta}\hat{\mathbf{a}}_\theta + \frac{1}{r\sin\theta}\frac{\partial f}{\partial \phi}\hat{\mathbf{a}}_\phi\right) \bullet \mathbf{d\ell}.$$

Comparing this with Equation (2.80), we can finally write

$$\nabla f = \frac{\partial f}{\partial r}\hat{\mathbf{a}}_r + \frac{1}{r}\frac{\partial f}{\partial \theta}\hat{\mathbf{a}}_\theta + \frac{1}{r\sin\theta}\frac{\partial f}{\partial \phi}\hat{\mathbf{a}}_\phi \quad \text{(spherical coordinates).} \qquad (2.85)$$

Example 2-9

Find the gradient of the scalar field $f = x^2 + y^2$ in a) Cartesian and b) cylindrical coordinates.

Solution:

a) The necessary partial derivatives dictated by Equation (2.79) are

$$\frac{\partial f}{\partial x} = 2x \qquad \frac{\partial f}{\partial y} = 2y \qquad \frac{\partial f}{\partial z} = 0\,.$$

Therefore,

$$\nabla f = 2x\hat{\mathbf{a}}_x + 2y\hat{\mathbf{a}}_y.$$

b) The representation of ∇f in cylindrical coordinates can be obtained either by transforming the preceding expression by the normal rules of vector transformations or by using the cylindrical coordinate expression for ∇f directly. Choosing the latter, we must first express f in cylindrical coordinates:

$$f = x^2 + y^2 = \rho^2\cos^2\phi + \rho^2\sin^2\phi = \rho^2.$$

Next, the necessary partial derivatives of f are

$$\frac{\partial f}{\partial \rho} = 2\rho \qquad \frac{\partial f}{\partial \phi} = 0 \qquad \frac{\partial f}{\partial z} = 0\,.$$

Substituting these into Equation (2.84), we obtain

$$\nabla f = 2\rho\hat{\mathbf{a}}_\rho.$$

It is left as an exercise for the reader to show that the two expressions for ∇f found in parts *a* and *b* are indeed the same vector.

2-5-3 THE DIVERGENCE OF A VECTOR FIELD

As with scalars, a knowledge of how a vector field changes about a point is often more important than the value of the field at that point. When piloting an airplane, for instance, it is often more important to know whether the airflow at a point is smooth or swirling than it is to know its velocity at a particular point. For vectors, two different indications of their rates of change are necessary to completely characterize the changes. The first of these, called ***divergence***, is discussed in this section. The second, called ***curl***, will be discussed in the section that follows.

The divergence of a vector **A** at a point *P* is a scalar quantity, defined as

$$\text{div}\,\mathbf{A} \equiv \lim_{\Delta v \to 0} \frac{\oint_S \mathbf{A} \cdot \mathbf{ds}}{\Delta v}\,. \qquad (2.86)$$

According to this definition, *S* is the surface that bounds the volume Δv, and **ds** always points outward from Δv. The value of surface integral $\oint_S \mathbf{A} \cdot \mathbf{ds}$ indicates whether there is a net tendency for **A** to point outward from *P*. Integrals of this type are called ***flux integrals***.

Figure 2-27a Figure 2-27b show two vectors that have nonzero divergence. In the case of Figure 2-27a, the positive divergence of the vector at the origin is easy to see, since all the vector streamlines are directed away from the origin. For this case, $\mathbf{A} \cdot \mathbf{ds}$ is positive at all points on a surface that surrounds the origin, resulting in a net positive flux.

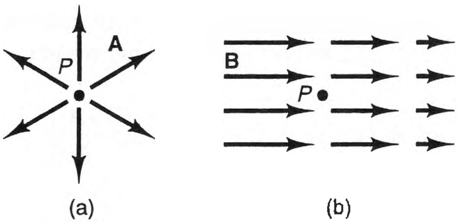


Figure 2-27 Two vector fields that have nonzero divergence at a point P.

The negative divergence of the vector shown in Figure 2-27b is less obvious to the eye, however, since this vector maintains a general left-to-right direction on both sides of the point P. Nevertheless, the flux entering[7] a surface surrounding P from the left is greater than that which leaves on the right, resulting in a negative divergence at P.

Even though the divergence of a vector is defined in terms of a surface integral, we will now show that it can be represented in terms of derivatives of the components of the vector. We will start by evaluating the flux integral $\oint_S \mathbf{A}\cdot ds$ about the rectangular surface shown in Figure 2-28. Here, a small rectangular volume of dimensions Δx, Δy, and Δz surrounds the point $P(x_o, y_o, z_o)$, which is shown in the center of the volume. The integral over this closed surface can be written as the sum of six open surface integrals:

$$\oint_S \mathbf{A}\cdot\mathbf{ds} = \int_{\substack{\text{front}\\ \text{face}}} \mathbf{A}\cdot\mathbf{ds} + \int_{\substack{\text{back}\\ \text{face}}} \mathbf{A}\cdot\mathbf{ds} + \int_{\substack{\text{right}\\ \text{face}}} \mathbf{A}\cdot\mathbf{ds} + \int_{\substack{\text{left}\\ \text{face}}} \mathbf{A}\cdot\mathbf{ds} + \int_{\substack{\text{top}\\ \text{face}}} \mathbf{A}\cdot\mathbf{ds} + \int_{\substack{\text{bottom}\\ \text{face}}} \mathbf{A}\cdot\mathbf{ds}\,. \quad (2.87)$$

On the front face, $x = x_o + (\Delta x)/2$, $\mathbf{ds} = dy\,dz\,\hat{\mathbf{a}}_x$, and $\mathbf{A}\cdot\mathbf{ds} = A_x\,dydz$. Substituting, we find that the integral over this face becomes

$$\int_{\substack{\text{front}\\ \text{face}}} \mathbf{A}\cdot\mathbf{ds} = \int_{y_o-\Delta y/2}^{y_o+\Delta y/2} \int_{z_o-\Delta z/2}^{z_o+\Delta z/2} A_x\left(x_o + \frac{\Delta x}{2}, y, z\right) dy\,dz\,. \quad (2.88)$$

Since Δx, Δy, and Δz are all assumed to be small, we can use Taylor's theorem to expand $A_x(x_o + \Delta x/2, y, z)$ about the point $P(x_o, y_o, z_o)$. Using the first two terms of the Taylor's expansion for each coordinate, we obtain

$$A_x\left(x_o + \frac{\Delta x}{2}, y, z\right) \cong A_x(x_o, y_o, z_o) + \frac{\Delta x}{2}\left.\frac{\partial A_x}{\partial x}\right|_P + (y - y_o)\left.\frac{\partial A_x}{\partial y}\right|_P + (z - z_o)\left.\frac{\partial A_x}{\partial z}\right|_P, \quad (2.89)$$

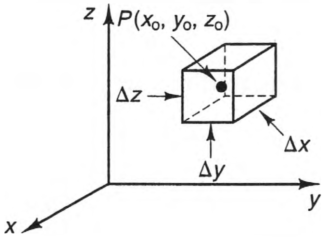


Figure 2-28 A small rectangular surface surrounding a point P.

[7] It is common to speak of flux as if it is something that moves through the surface, even if the vector in question does not represent a quantity of motion (such as a force).

where the notation $|_P$ indicates that the derivatives are evaluated at $P(x_o, y_o, z_o)$. Substituting Equation (2.89) into (2.88) and integrating, we get

$$\int_{\substack{\text{front}\\\text{face}}} \mathbf{A} \cdot \mathbf{ds} \cong \Delta y\, \Delta z\, A_x(x_o, y_o, z_o) + \frac{\Delta x}{2} \Delta y\, \Delta x \left.\frac{\partial A_x}{\partial x}\right|_P . \tag{2.90}$$

Similarly, on the back face we have $x = x_o - \Delta x/2$, $\mathbf{ds} = -dydz\, \hat{\mathbf{a}}_x$, $\mathbf{A} \cdot \mathbf{ds} = -A_x dy\, dz$, and

$$A_x\left(x_o - \frac{\Delta x}{2}, y, z\right) \cong A_x(x_o, y_o, z_o) - \frac{\Delta x}{2} \left.\frac{\partial A_x}{\partial x}\right|_P + (y - y_0) \left.\frac{\partial A_x}{\partial y}\right|_P + (z - z_o) \left.\frac{\partial A_x}{\partial z}\right|_P .$$

Integrating over this back face, we obtain

$$\int_{\substack{\text{back}\\\text{face}}} \mathbf{A} \cdot \mathbf{ds} \cong -\Delta y\, \Delta z\, A_x(x_o, y_o, z_o) + \frac{\Delta x}{2} \Delta y\, \Delta z \left.\frac{\partial A_x}{\partial x}\right|_P . \tag{2.91}$$

The sum of the flux contributions from the front and back faces is found by adding Equations (2.90) and (2.91):

$$\int_{\substack{\text{front}\\\text{face}}} \mathbf{A} \cdot \mathbf{ds} + \int_{\substack{\text{back}\\\text{face}}} \mathbf{A} \cdot \mathbf{ds} \cong \Delta x\, \Delta y\, \Delta z \left.\frac{\partial A_x}{\partial x}\right|_P . \tag{2.92}$$

Using similar steps, we can also be show that

$$\int_{\substack{\text{right}\\\text{face}}} \mathbf{A} \cdot \mathbf{ds} + \int_{\substack{\text{left}\\\text{face}}} \mathbf{A} \cdot \mathbf{ds} \cong \Delta x\, \Delta y\, \Delta z \left.\frac{\partial A_y}{\partial y}\right|_P \tag{2.93}$$

and

$$\int_{\substack{\text{top}\\\text{face}}} \mathbf{A} \cdot \mathbf{ds} + \int_{\substack{\text{bottom}\\\text{face}}} \mathbf{A} \cdot \mathbf{ds} \cong \Delta x\, \Delta y\, \Delta z \left.\frac{\partial A_z}{\partial z}\right|_P . \tag{2.94}$$

Summing all the contributions to $\oint_S \mathbf{A} \cdot \mathbf{ds}$ and noting that $\Delta x\, \Delta y\, \Delta z = \Delta v$, we find that

$$\oint_S \mathbf{A} \cdot \mathbf{ds} \cong \left\{ \left.\frac{\partial A_x}{\partial x}\right|_P + \left.\frac{\partial A_y}{\partial y}\right|_P + \left.\frac{\partial A_z}{\partial z}\right|_P \right\} \Delta v.$$

This expression becomes exact in the limit as $\Delta v \to 0$. Comparing the expression with the definition of divergence (Equation (2.86)), we find that

$$\text{div}\,\mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}, \tag{2.95}$$

where we have dropped the notation $|_P$, since the volume Δv collapses to the point P as $\Delta v \to 0$. This equation can also be written as a dot product,

$$\text{div}\,\mathbf{A} = \left(\frac{\partial}{\partial x}\hat{\mathbf{a}}_x + \frac{\partial}{\partial y}\hat{\mathbf{a}}_y + \frac{\partial}{\partial z}\hat{\mathbf{a}}_z\right) \bullet (A_x\hat{\mathbf{a}}_x + A_y\hat{\mathbf{a}}_y + A_z\hat{\mathbf{a}}_z) = \nabla \bullet \mathbf{A},$$

where ∇ is the del operator, defined by Equation (2.79). Thus, the notation $\nabla \bullet \mathbf{A}$ has the same meaning as div $\mathbf{A}$. In the Cartesian coordinate system, we can write

$$\nabla \bullet \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} \quad \text{(Cartesian coordinates).} \tag{2.96}$$

Expressions for $\nabla \bullet \mathbf{A}$ also can also be derived in the cylindrical and spherical coordinate systems. This can be accomplished in either of two ways. The first is to transform the Cartesian coordinate expression into these coordinate systems by the standard transformation rules outlined in Section 2-4-4. This procedure is straightforward but tedious, since the chain rule must be used repeatedly to transform the variables in the partial derivatives. The second method is to evaluate $\oint_S \mathbf{A} \bullet \mathbf{ds}$ directly in the cylindrical and spherical coordinate systems using a procedure similar to what we used in Cartesian coordinates.[8] By either technique, it can be shown that

$$\nabla \bullet \mathbf{A} = \frac{1}{\rho}\left[\frac{\partial}{\partial \rho}(\rho A_\rho)\right] + \frac{1}{\rho}\frac{\partial A_\phi}{\partial \phi} + \frac{\partial A_z}{\partial z} \quad \text{(cylindrical coordinates)} \tag{2.97}$$

and

$$\nabla \bullet \mathbf{A} = \frac{1}{r^2}\left[\frac{\partial}{\partial r}(r^2 A_r)\right] + \frac{1}{r\sin\theta}\left[\frac{\partial}{\partial \theta}(A_\theta \sin\theta)\right] + \frac{1}{r\sin\theta}\frac{\partial A_\phi}{\partial \phi} \quad \text{(spherical coordinates).} \tag{2.98}$$

Example 2-10

Find the divergence of $\mathbf{A} = x\hat{\mathbf{a}}_x$ at any point using a) Cartesian coordinates and b) cylindrical coordinates.

[8] See Plonsey and Collin, *Principles and Applications of Electromagnetic Fields*, New York: McGraw-Hill, 1961.

Solution:

a) From Equation (2.96)

$$\nabla \cdot \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} = \frac{\partial}{\partial x}(x) = 1 \qquad \text{(at all points).}$$

b) Transforming **A** into cylindrical coordinates, we find that

$$\mathbf{A} = \rho \cos^2 \phi \, \hat{\mathbf{a}}_\rho - \rho \cos \phi \sin \phi \, \hat{\mathbf{a}}_\phi .$$

Using Equation (2.97), we get

$$\nabla \cdot \mathbf{A} = \frac{1}{\rho}\left[\frac{\partial}{\partial \rho}(\rho^2 \cos^2 \phi)\right] - \frac{1}{\rho}\frac{\partial}{\partial \phi}(\rho \cos \phi \sin \phi)$$

$$= 2\cos^2 \phi - (-\sin^2 \phi + \cos^2 \phi) = 1 .$$

As expected, this result is the same as was obtained using the Cartesian coordinate system.

Before leaving the subject of divergence, we will derive an important theorem called the ***divergence theorem***. Consider the volume integral $\int_V \nabla \cdot \mathbf{A}\, dv$. Using the definition of divergence, we can write this integral as

$$\int_V \nabla \cdot \mathbf{A}\, dv = \int_V \lim_{\Delta v \to 0} \frac{\oint_S \mathbf{A} \cdot \mathbf{ds}}{\Delta v} dv .$$

Expressing the right-hand integral as an infinite sum of infinitesimal volumes, we obtain

$$\int_V \nabla \cdot \mathbf{A}\, dv = \sum_k \lim_{\Delta v_k \to 0} \frac{\oint_{S_k} \mathbf{A} \cdot \mathbf{ds}}{\Delta v_k} \Delta v_k ,$$

where Δv_k is the kth differential subvolume, which is surrounded by the closed surface S_k.

The right-hand side of the above expression allows us to interpret $\int_V \nabla \cdot \mathbf{A}\, dv$ as the sum of the fluxes emanating from each point within V. But as can be seen from Figure 2-29, flux contributions from adjacent points within V cancel, since the outward flux from one volume is at the same time inward flux to its neighbor. All flux contributions in the integral cancel, except those at points on the surface bounding V. Thus,

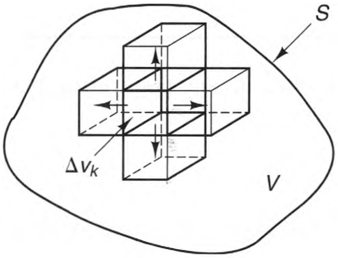


Figure 2-29 Geometry for deriving the divergence theorem.

$$\int_V \nabla \cdot \mathbf{A}\, dv = \sum_k \lim_{\Delta v_k \to 0} \oint_{S_k} \mathbf{A} \cdot \mathbf{ds} = \oint_S \mathbf{A} \cdot \mathbf{ds}.$$

The sum on the right-hand side is the integral $\oint_S \mathbf{A} \cdot \mathbf{ds}$, where S is the closed surface that bounds the volume V. Hence, we obtain the ***divergence theorem***:

$$\int_V \nabla \cdot \mathbf{A}\, dv = \oint_S \mathbf{A} \cdot \mathbf{ds} \qquad \text{(divergence theorem).} \qquad (2.99)$$

This theorem is useful for transforming equations involving vector integrals into simpler forms.

Example 2-11

Given $\mathbf{A} = r\hat{\mathbf{a}}_r + \sin\theta\, \hat{\mathbf{a}}_\theta$, verify the divergence theorem over the spherical volume of radius $r = 1$, centered about the origin.

Solution:

For the surface integral,

$$\mathbf{ds} = \mathbf{ds}_r = r^2 \sin\theta\, d\theta\, d\phi\, \hat{\mathbf{a}}_r$$

and

$$\mathbf{A} \cdot \mathbf{ds} = r^3 \sin\theta\, d\theta\, d\phi\, .$$

Substituting, we find that the surface integral becomes

$$\oint_S \mathbf{A} \cdot \mathbf{ds} = \int_0^{2\pi} \int_0^{\pi} r^3 \sin\theta\, d\theta\, d\phi \bigg|_{r=1} = 2\pi \int_0^{\pi} \sin\theta\, d\theta = 4\pi\, .$$

To evaluate the volume integral, we must first evaluate the divergence of $\mathbf{A}$:

$$\nabla \cdot \mathbf{A} = \frac{1}{r^2}\left[\frac{\partial}{\partial r}(r^3)\right] + \frac{1}{r\sin\theta}\left[\frac{\partial}{\partial \theta}(\sin^2\theta)\right]$$

$$= 3 + \frac{2\cos\theta}{r}\, .$$

Substituting, the volume integral becomes

$$\int_V \nabla \cdot \mathbf{A}\, dv = \int_0^{2\pi} \int_0^{\pi} \int_0^{1} \left(3 + \frac{2\cos\theta}{r}\right) r^2 \sin\theta\, dr\, d\theta\, d\phi$$

$$= \int_0^{2\pi} \int_0^{\pi} \int_0^{1} 3r^3 \sin\theta\, dr\, d\theta\, d\phi + \int_0^{\pi} \int_0^{2\pi} \int_0^{1} 2r \sin\theta \cos\theta\, dr\, d\theta\, d\phi$$

$$= 4\pi + 0 = 4\pi\, .$$

The two integrals are indeed equal, just as predicted by the divergence theorem.

2-5-4 THE CURL OF A VECTOR FIELD

The ***curl*** of a vector **A** is an indication of the tendency of **A** to "push" or "pull" along a closed path that encircles a point. By this, we mean that a vector **A** has curl at a point if the line integral $\oint_C \mathbf{A} \cdot \mathbf{d\ell}$ is nonzero and C is a differential path that encircles the point. This tendency to push or pull around a path is called ***circulation***. There are three perpendicular planes that such a path can lie in about a point, so the curl is defined as a vector quantity, denoted by the symbol "curl **A**" or "$\nabla \times \mathbf{A}$." Referring to Figure 2-30, we define the component of $\nabla \times \mathbf{A}$ in the direction $\hat{\mathbf{a}}_i$ by

$$(\text{curl } \mathbf{A})_i \equiv (\nabla \times \mathbf{A})_i \equiv \lim_{\Delta s_i \to 0} \frac{\oint_{C_i} \mathbf{A} \cdot \mathbf{d\ell}}{\Delta s_i}, \tag{2.100}$$

where Δs_i is a small surface that is bounded by the contour (i.e., path) C_i and has unit normal $\hat{\mathbf{a}}_i$. The direction of C_i is governed by the right-hand rule, which says that when the right-hand thumb is placed along the path, the remaining fingers "poke" through the surface Δs_i in the direction of $\hat{\mathbf{a}}_i$.

Since $\nabla \times \mathbf{A}$ is a vector, we can represent it by its magnitude and direction, which we will denote as $|\nabla \times \mathbf{A}|$ and $\hat{\mathbf{a}}_n$, respectively. To find $|\nabla \times \mathbf{A}|$, we notice from Equation (2.100) that the values of the components of $|\nabla \times \mathbf{A}|$ vary with the orientations of the integration paths C_i. Since the maximum value that any component of a vector can attain equals the vector's magnitude, we can conclude that

$$|\nabla \times \mathbf{A}| = \left[\lim_{\Delta s \to 0} \frac{\oint_C \mathbf{A} \cdot \mathbf{d\ell}}{\Delta s} \right]_{\max}, \tag{2.101}$$

where C is the differential path that maximizes the circulation integral. Thus, we can write **curl A** as

$$\nabla \times \mathbf{A} = \hat{\mathbf{a}}_n \left[\lim_{\Delta s \to 0} \frac{\oint_C \mathbf{A} \cdot \mathbf{d\ell}}{\Delta s} \right]_{\max}, \tag{2.102}$$

where $\hat{\mathbf{a}}_n$ is perpendicular to the surface bounded by C and points in the direction determined by the right-hand rule.

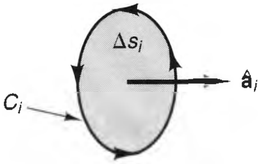


Figure 2-30 A surface Δs_i with unit normal $\hat{\mathbf{a}}_i$, bounded by the contour C_i.

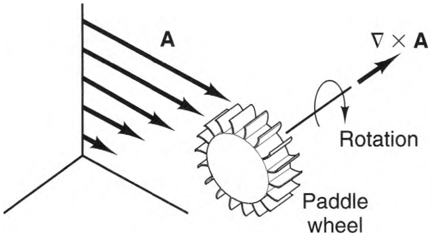


Figure 2-31 "Paddle wheel" analogy of the curl of a vector.

Figure 2-31 is helpful in understanding the meaning of the vector $\nabla \times \mathbf{A}$. Here, a paddle wheel is placed in a fluid whose velocity is represented by the vector **A**. A torque will be exerted on the paddle wheel whenever there is a nonzero circulation of **A** about the paddle wheel axis. According to Equation (2.102), maximum torque is produced when the axis of the wheel is in the direction of $\nabla \times \mathbf{A}$. If no torque is produced at a point for any orientation of the wheel, **A** has no curl there.

The curl of a vector can be calculated by evaluating partial derivatives of the components of **A** with respect to the coordinate variables. To show this, let us first find the x component of $\nabla \times \mathbf{A}$, which requires that we evaluate $\oint_{C_x} \mathbf{A} \cdot \mathbf{d\ell}$ along the contour C_x, shown in Figure 2-32. This integral can be written as the sum of four line integrals:

$$\oint_{C_x} \mathbf{A} \cdot \mathbf{d\ell} = \int_{\text{right}} \mathbf{A} \cdot \mathbf{d\ell} + \int_{\text{top}} \mathbf{A} \cdot \mathbf{d\ell} + \int_{\text{left}} \mathbf{A} \cdot \mathbf{d\ell} + \int_{\text{bottom}} \mathbf{A} \cdot \mathbf{d\ell}. \tag{2.103}$$

Along the right and left contours, $\mathbf{d\ell} = dz\, \hat{\mathbf{a}}_z$. Similarly, $\mathbf{d\ell} = dy\, \hat{\mathbf{a}}_y$ along the top and bottom contours. Substituting these into the contour integrals, we find that

$$\oint_{C_x} \mathbf{A} \cdot \mathbf{d\ell} = \int_{z_o - \frac{\Delta z}{2}}^{z_o + \frac{\Delta z}{2}} A_z\left(x_o, y_o + \frac{\Delta y}{2}, z\right) dz + \int_{y_o + \frac{\Delta y}{2}}^{y_o - \frac{\Delta y}{2}} A_y\left(x_o, y, z_o + \frac{\Delta z}{2}\right) dy$$
$$+ \int_{z_o + \frac{\Delta z}{2}}^{z_o - \frac{\Delta z}{2}} A_z\left(x_o, y_o - \frac{\Delta y}{2}, z\right) dz + \int_{y_o - \frac{\Delta y}{2}}^{y_o + \frac{\Delta y}{2}} A_y\left(x_o, y, z_o - \frac{\Delta z}{2}\right) dy. \tag{2.104}$$

Note that the limits of integration are such that the path of integration is counterclockwise, which is consistent with the right-hand rule.

Since both Δx and Δy are small, each of the integrands on the right-hand side of Equation (2.104) can be expanded in a Taylor's series about $P(x_o, y_o, z_o)$. For the first integral, we can write

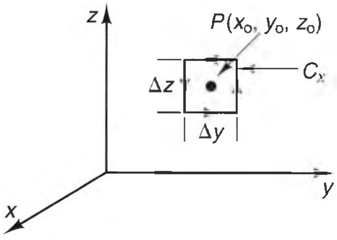


Figure 2-32 A contour C_x in the yz-plane about an arbitrary point P.

$$A_z\left(x_o, y_o + \frac{\Delta y}{2}, z\right) \cong A_z(x_o, y_o, z_o) + \frac{\Delta y}{2}\frac{\partial A_z}{\partial y}\bigg|_P + (z - z_o)\frac{\partial A_z}{\partial z}\bigg|_P. \tag{2.105}$$

Integrating, we obtain

$$\int_{\text{right}} \mathbf{A}\cdot d\boldsymbol{\ell} \cong \Delta z A_z(x_o, y_o, z_o) + \frac{\Delta y}{2}\Delta z\frac{\partial A_z}{\partial y}\bigg|_P. \tag{2.106}$$

Similarly, for the integral over the left segment of C_x, we can express the integrand as

$$A_z\left(x_o, y_o - \frac{\Delta y}{2}, z\right) \cong A_z(x_o, y_o, z_o) - \frac{\Delta y}{2}\frac{\partial A_z}{\partial y}\bigg|_P + (z - z_o)\frac{\partial A_z}{\partial z}\bigg|_P, \tag{2.107}$$

which yields

$$\int_{\text{left}} \mathbf{A}\cdot d\boldsymbol{\ell} \cong -\Delta z\, A_z(x_o, y_o, z_o) + \frac{\Delta y}{2}\Delta z\frac{\partial A_z}{\partial y}\bigg|_P. \tag{2.108}$$

Summing the contributions from the "right" and "left" portions of the contour yields

$$\int_{\text{right}} \mathbf{A}\cdot d\boldsymbol{\ell} + \int_{\text{left}} \mathbf{A}\cdot d\boldsymbol{\ell} \cong \Delta y\,\Delta z\frac{\partial A_z}{\partial y}\bigg|_P. \tag{2.109}$$

Similar analysis of the "top" and "bottom" portions of the contour results in

$$\int_{\text{top}} \mathbf{A}\cdot d\boldsymbol{\ell} + \int_{\text{bottom}} \mathbf{A}\cdot d\boldsymbol{\ell} \cong -\Delta y\,\Delta z\frac{\partial A_y}{\partial z}\bigg|_P. \tag{2.110}$$

Substituting Equations (2.109) and (2.110) into Equation (2.103), we have

$$\oint_{C_x} \mathbf{A}\cdot d\boldsymbol{\ell} \cong \Delta y\,\Delta z\left[\frac{\partial A_z}{\partial y}\bigg|_P - \frac{\partial A_y}{\partial z}\bigg|_P\right], \tag{2.111}$$

which becomes exact in the limit as $\Delta s_x = \Delta y\,\Delta z \to 0$. Comparing this expression with Equation (2.100), we can conclude that

$$(\nabla \times \mathbf{A})_x = \lim_{\Delta s_x \to 0}\frac{\oint_{C_x}\mathbf{A}\cdot d\boldsymbol{\ell}}{\Delta s_x} = \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}, \tag{2.112}$$

where the notation $|_P$ has been dropped from the partial derivatives, since the surface has collapsed to a point.

The y- and z-components of $\nabla \times \mathbf{A}$ can be found by evaluating Equation (2.100) around the contours C_y and C_z, which lie in the $y = y_o$ and $z = z_o$ planes, respectively.

Evaluating the resulting circulation integrals using the same procedure as used for $(\nabla \times \mathbf{A})_x$, we finally obtain

$$\nabla \times \mathbf{A} = \left[\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right]\hat{\mathbf{a}}_x + \left[\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right]\hat{\mathbf{a}}_y + \left[\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right]\hat{\mathbf{a}}_z \quad \text{(Cartesian coordinates).} \tag{2.113}$$

This formula can also be written in shorthand form as the determinant

$$\nabla \times \mathbf{A} = \begin{vmatrix} \hat{\mathbf{a}}_x & \hat{\mathbf{a}}_y & \hat{\mathbf{a}}_z \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ A_x & A_y & A_z \end{vmatrix}, \tag{2.114}$$

which shows why the symbol "$\nabla \times \mathbf{A}$" and "curl $\mathbf{A}$" are used interchangeably.

Corresponding expressions for $\nabla \times \mathbf{A}$ exist in the cylindrical and spherical coordinate systems. We have

$$\nabla \times \mathbf{A} = \left[\frac{1}{\rho}\frac{\partial A_z}{\partial \phi} - \frac{\partial A_\phi}{\partial z}\right]\hat{\mathbf{a}}_\rho + \left[\frac{\partial A_\rho}{\partial z} - \frac{\partial A_z}{\partial \rho}\right]\hat{\mathbf{a}}_\phi + \frac{1}{\rho}\left[\frac{\partial}{\partial \rho}(\rho A_\phi) - \frac{\partial A_\rho}{\partial \phi}\right]\hat{\mathbf{a}}_z \quad \text{(cylindrical coordinates)} \tag{2.115}$$

and

$$\begin{aligned} \nabla \times \mathbf{A} &= \frac{1}{r \sin\theta}\left[\frac{\partial}{\partial \theta}(A_\phi \sin\theta) - \frac{\partial A_\theta}{\partial \phi}\right]\hat{\mathbf{a}}_r \\ &+ \frac{1}{r}\left[\frac{1}{\sin\theta}\frac{\partial A_r}{\partial \phi} - \frac{\partial}{\partial r}(rA_\phi)\right]\hat{\mathbf{a}}_\theta + \frac{1}{r}\left[\frac{\partial}{\partial r}(rA_\theta) - \frac{\partial A_r}{\partial \theta}\right]\hat{\mathbf{a}}_\phi \quad \text{(spherical coordinates).} \end{aligned} \tag{2.116}$$

These expressions can be derived either by transforming the components and coordinates of Equation (2.113) into the new coordinate system or by evaluating the circulation integrals of Equation (2.100) directly in the new coordinate systems.[9]

Example 2-12

Calculate the curl of $\mathbf{A} = y\hat{\mathbf{a}}_x$ at all points using a) Cartesian coordinates and b) spherical coordinates.

[9] See Plonsey and Collin, *Principles and Applications of Electromagnetic Fields* (New York: McGraw-Hill, 1961).

Solution:

a) Of the six partial derivatives present in the expression for $\nabla \times \mathbf{A}$, only one is nonzero, since $A_y = A_z = 0$ and A_x is a function only of y. Thus,

$$\nabla \times \mathbf{A} = -\frac{\partial A_x}{\partial y}\,\hat{\mathbf{a}}_z = -\hat{\mathbf{a}}_z.$$

b) Transforming $\mathbf{A}$ into spherical coordinates, we find that

$$\mathbf{A} = r\sin^2\theta\sin\phi\cos\phi\,\hat{\mathbf{a}}_r + r\sin\theta\cos\theta\sin\phi\cos\phi\,\hat{\mathbf{a}}_\theta - r\sin\theta\sin^2\phi\,\hat{\mathbf{a}}_\phi.$$

From Equation (2.116), we have

$$(\nabla \times \mathbf{A})_r = \frac{1}{r\sin\theta}\left[\frac{\partial}{\partial\theta}\left(-r\sin^2\theta\sin^2\phi\right) - \frac{\partial}{\partial\phi}\left(r\sin\theta\cos\theta\sin\phi\cos\phi\right)\right] = -\cos\theta$$

$$(\nabla \times \mathbf{A})_\theta = \frac{1}{r}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\phi}\left(r\sin^2\theta\sin\phi\cos\phi\right) - \frac{\partial}{\partial r}\left(-r^2\sin\theta\sin^2\phi\right)\right] = \sin\theta$$

$$(\nabla \times \mathbf{A})_\phi = \frac{1}{r}\left[\frac{\partial}{\partial r}\left(r^2\sin\theta\cos\theta\sin\phi\cos\phi\right) - \frac{\partial}{\partial\theta}\left(r\sin^2\theta\sin\phi\cos\phi\right)\right] = 0.$$

Thus, $\nabla \times \mathbf{A} = -\cos\theta\,\hat{\mathbf{a}}_r + \sin\theta\,\hat{\mathbf{a}}_\theta$. It is left as an exercise to the reader to show that this result is equivalent to the one found in part *a*).

A useful theorem that involves the curl operation is ***Stokes's theorem***. To derive this theorem, consider the integral $\int_S (\nabla \times \mathbf{A}) \cdot \mathbf{ds}$ over an open surface S. From the properties of the dot product, we can write the integrand as

$$(\nabla \times \mathbf{A}) \cdot \mathbf{ds} = (\nabla \times \mathbf{A}) \cdot ds\,\hat{\mathbf{a}}_n = (\nabla \times \mathbf{A})_n\,ds, \tag{2.117}$$

where $\hat{\mathbf{a}}_n$ is the outward normal to the differential surface and $(\nabla \times \mathbf{A})_n$ is the component of $\nabla \times \mathbf{A}$ in the $\hat{\mathbf{a}}_n$ direction. Using Equation (2.117), we can write $\int_S (\nabla \times \mathbf{A}) \cdot \mathbf{ds}$ in the form

$$\int_S (\nabla \times \mathbf{A}) \cdot \mathbf{ds} = \int_S (\nabla \times \mathbf{A})_n\,ds.$$

Substituting Equation (2.100) into the right-hand side of this expression, we obtain

$$\int_S (\nabla \times \mathbf{A}) \cdot \mathbf{ds} = \int_S \lim_{\Delta s \to 0}\left[\frac{\oint_{\Delta C} \mathbf{A} \cdot \mathbf{d\ell}}{\Delta s}\right] ds,$$

where the contour ΔC bounds the surface $\Delta\mathbf{s} = \Delta s\,\hat{\mathbf{a}}_n$ in a right-handed sense. Expressing the right-hand integral as an infinite sum of differential surface elements, we obtain

$$\int_S (\nabla \times \mathbf{A}) \cdot \mathbf{ds} = \sum_k \lim_{\Delta s_k \to 0} \frac{\oint_{C_k} \mathbf{A} \cdot \mathbf{d\ell}}{\Delta s_k}\,\Delta s_k.$$

Canceling the Δs_k terms, we find that

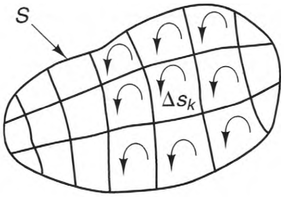


Figure 2-33 Geometry for deriving Stokes's theorem.

$$\int_S (\nabla \times \mathbf{A}) \cdot \mathbf{ds} = \sum_k \lim_{\Delta s_k \to 0} \oint_{C_k} \mathbf{A} \cdot \mathbf{d\ell}.$$

As can be seen from Figure 2-33, the line integral contributions from adjacent cells cancel, since the directions of integration along these paths are opposite. As a result, all of the line integral contributions cancel, except those along the contour that bounds S. Thus,

$$\int_S (\nabla \times \mathbf{A}) \cdot \mathbf{ds} = \oint_C \mathbf{A} \cdot \mathbf{d\ell} \qquad \text{(Stokes's theorem)}, \tag{2.118}$$

where C is the contour that bounds S in a right-handed sense. If S is a closed surface, it has no bounding contour, so

$$\oint_S \nabla \times \mathbf{A} \cdot \mathbf{ds} = 0. \tag{2.119}$$

An important consequence of Stokes's theorem is that there is more than one surface that corresponds to a particular contour C. This is depicted in Figure 2-34, where the surfaces S_1 and S_2 both have the same bounding contour C. Since both surfaces are bounded by the closed contour C, it follows from Stokes's theorem that both surface integrals have the same value:

$$\int_{S_1} (\nabla \times \mathbf{A}) \cdot \mathbf{ds}_1 = \int_{S_2} (\nabla \times \mathbf{A}) \cdot \mathbf{ds}_2. \tag{2.120}$$

The appropriate orientation of the differential surface vector $\mathbf{ds}_1$ on S_1 is easy to visualize from the right-hand rule, since S_1 is flat. Because S_2 is curved, however, the correct orientation of $\mathbf{ds}_2$ is not as obvious. An aid that is helpful here is to imagine that S_1 is an elastic membrane that, when stretched, assumes the shape of S_2. During this process, we simply allow **ds** at each point to remain perpendicular to the surface as the membrane transforms from S_1 to S_2.

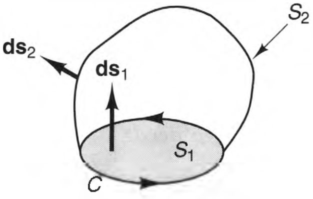


Figure 2-34 Two surfaces bounded by the same contour C.

Example 2-13

For $\mathbf{A} = \rho z\, \hat{\mathbf{a}}_\phi$, evaluate both sides of Equation (2.118), Stokes's theorem, for the contour C and the surfaces S_1 and S_2 shown in Figure 2-35.

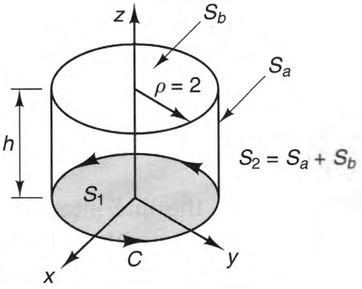


Figure 2-35 Circular contour C that bounds two open surfaces S_1 and S_2.

Solution:

Both S_1 and S_2 are bounded by the contour C, which is described by $\rho = 2, 0 < \phi < 2\pi$, and $z = 0$. Along this contour,

$$\mathbf{d\ell} = \rho d\phi\, \hat{\mathbf{a}}_\phi \Big|_{\rho=2} = 2d\phi\, \hat{\mathbf{a}}_\phi.$$

Substituting this $\mathbf{d\ell}$ into the contour integral, we obtain

$$\oint_C \mathbf{A} \cdot \mathbf{d\ell} = \oint_C 2\rho z\, \hat{\mathbf{a}}_\phi \cdot \hat{\mathbf{a}}_\phi d\phi \Big|_{\substack{\rho=2 \\ z=0}} = \int_0^{2\pi} 0 d\phi = 0.$$

To evaluate the surface integrals, we must first calculate $\nabla \times \mathbf{A}$. Since $\mathbf{A}$ has only a ϕ component, we have

$$\nabla \times \mathbf{A} = -\frac{\partial A_\phi}{\partial z} \hat{\mathbf{a}}_\rho + \frac{1}{\rho} \frac{\partial}{\partial \rho}(\rho A_\phi) \hat{\mathbf{a}}_z = -\rho \hat{\mathbf{a}}_\rho + 2z\, \hat{\mathbf{a}}_z.$$

For $\int_{S_1} (\nabla \times \mathbf{A}) \cdot \mathbf{ds}$, we note that S_1 is a circle of radius $\rho = 2$. Since the direction of C is counterclockwise, the right-hand rule requires that $\mathbf{ds} = \rho\, d\rho\, d\phi\, \hat{\mathbf{a}}_z$. But, since the surface is in the $z = 0$ plane, $(\nabla \times \mathbf{A}) \cdot \mathbf{ds}|_{z=0} = 0$, yielding

$$\int_{S_2} (\nabla \times \mathbf{A}) \cdot \mathbf{ds} = \int_{S_1} 0\, d\rho d\phi = 0.$$

The surface S_2 consists of two simple surfaces—a cylinder S_a and its end cap S_b. From the right-hand rule, the differential surface vectors on S_a and S_b are $\mathbf{ds}_a = \rho\, d\phi\, dz\, \hat{\mathbf{a}}_\rho$ and $\mathbf{ds}_b = \rho\, d\rho\, \phi\, \hat{\mathbf{a}}_z$, respectively. Given these, the surface integral over S_2 becomes

$$\int_{S_2} (\nabla \times \mathbf{A}) \cdot \mathbf{ds} = \int_0^h \int_0^{2\pi} -\rho^2 d\phi\, dz \Big|_{\rho=2} + \int_0^{2\pi} \int_0^2 2z\rho\, d\rho\, d\phi \Big|_{z=h} = -8\pi h + 8\pi h = 0.$$

2-5-5 THE LAPLACIAN OPERATOR

There are many occasions in vector analysis where a gradient operation is followed by a divergence operation. This combined operation is called the Laplacian and is denoted by the symbol

$$\nabla^2 \equiv \nabla \cdot \nabla. \tag{2.121}$$

The application of the Laplacian to scalar fields is straightforward. In Cartesian coordinates, we have

$$\nabla^2 f = \nabla \cdot \nabla f = \nabla \cdot \left(\frac{\partial f}{\partial x}\hat{\mathbf{a}}_x + \frac{\partial f}{\partial y}\hat{\mathbf{a}}_y + \frac{\partial f}{\partial z}\hat{\mathbf{a}}_z\right), \tag{2.122}$$

which yields

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} \qquad \text{(Cartesian coordinates).} \tag{2.123}$$

Similarly, the Laplacian of a scalar field can be expanded in cylindrical and spherical coordinates to yield

$$\nabla^2 f = \frac{1}{\rho}\frac{\partial}{\partial \rho}\left(\rho\frac{\partial f}{\partial \rho}\right) + \frac{1}{\rho^2}\frac{\partial^2 f}{\partial \phi^2} + \frac{\partial^2 f}{\partial z^2} \qquad \text{(cylindrical coordinates)} \tag{2.124}$$

and

$$\nabla^2 f = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial f}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta\frac{\partial f}{\partial \theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 f}{\partial \phi^2} \qquad \text{(spherical coordinates).} \tag{2.125}$$

The Laplacian operator can also be applied to vector fields. To see how this is possible, let us consider the Laplacian of a vector **A** that is represented in Cartesian coordinates:

$$\nabla^2 \mathbf{A} = \nabla^2(A_x\hat{\mathbf{a}}_x + A_y\hat{\mathbf{a}}_y + A_z\hat{\mathbf{a}}_z).$$

Since the unit vectors $\hat{\mathbf{a}}_x$, $\hat{\mathbf{a}}_y$, and $\hat{\mathbf{a}}_z$ and are not functions of position, they are constants with respect to the ∇^2 operator. Thus, we can conclude that the Laplacian of a vector is also a vector, with Cartesian components given by

$$\nabla^2 \mathbf{A} \equiv \hat{\mathbf{a}}_x\nabla^2 A_x + \hat{\mathbf{a}}_y\nabla^2 A_y + \hat{\mathbf{a}}_z\nabla^2 A_z \qquad \text{(Cartesian components).} \tag{2.126}$$

If a vector is expressed in non-Cartesian components, its Laplacian cannot be evaluated so simply. To derive a general expression for $\nabla^2\mathbf{A}$, we note that the right-hand side of Equation (2.126) can be written as

$$\nabla^2\mathbf{A} = \hat{\mathbf{a}}_x\nabla^2 A_x + \hat{\mathbf{a}}_y\nabla^2 A_y + \hat{\mathbf{a}}_z\nabla^2 A_z = \nabla(\nabla\cdot\mathbf{A}) - \nabla\times\nabla\times\mathbf{A}\,. \tag{2.127}$$

The proof of this identity is straightforward in Cartesian coordinates and is left as an exercise for the reader. Since the divergence and curl operations are well defined in all coordinate systems, the right side of Equation (2.127) can be evaluated in any coordinate system. Thus, the Laplacian of a vector can be expressed in all coordinate systems as

$$\nabla^2\mathbf{A} = \nabla(\nabla\cdot\mathbf{A}) - \nabla\times\nabla\times\mathbf{A}. \tag{2.128}$$

2-5-6 HELMHOLTZ'S THEOREM

An important question in vector analysis is, "What kind of information is necessary to completely characterize a vector field over some region of space?" The answer to this question is important for two reasons. First, it allows us to judge whether a particular set of specifications uniquely defines a vector within some region. Second, a knowledge of the minimum information necessary to uniquely specify a vector quantity can simplify the work necessary to solve a given problem.

The key to determining the behavior of any quantity over a region is knowing how it changes from point to point. For scalars, the gradient operation supplies all of the information necessary. The following theorems make it clear that for vectors, two operations are needed: the divergence and the curl.

Theorem I: Any vector field that is continuously differentiable in some volume V can be uniquely determined if its divergence and curl are known throughout the volume and its value is known on the surface S that bounds the volume:

$$\mathbf{A}(\mathbf{r}) = -\nabla\left[\int_V \frac{\nabla'\cdot\mathbf{A}(\mathbf{r}')}{4\pi|\mathbf{r}-\mathbf{r}'|}\,dv' - \oint_S \frac{\mathbf{A}(\mathbf{r}')\cdot\hat{\mathbf{a}}_{n'}}{4\pi|\mathbf{r}-\mathbf{r}'|}\,ds'\right]$$

$$+\nabla\times\left[\int_V \frac{\nabla'\times\mathbf{A}(\mathbf{r}')}{4\pi|\mathbf{r}-\mathbf{r}'|}\,dv' - \oint_S \frac{\mathbf{A}(\mathbf{r}')\times\hat{\mathbf{a}}_{n'}}{4\pi|\mathbf{r}-\mathbf{r}'|}\,ds'\right]. \tag{2.129}$$

In this expression, the unit vector $\hat{\mathbf{a}}_{n'}$ points outward from S. Also, inside the integrals, the dummy integration position variable is

$$\mathbf{r}' = x'\hat{\mathbf{a}}_x + y'\hat{\mathbf{a}}_y + z'\hat{\mathbf{a}}_z$$

and the del operator ∇' is given by

$$\nabla' = \frac{\partial}{\partial x'}\hat{\mathbf{a}}_x + \frac{\partial}{\partial y'}\hat{\mathbf{a}}_y + \frac{\partial}{\partial z'}\hat{\mathbf{a}}_z.$$

This relationship is called ***Helmholtz's theorem*** and is proved in a number of advanced electromagnetics and mathematics texts.[10]

For most vectors found in electromagnetics, the surface integrals in Equation (2.29) vanish when the volume V is chosen to be all of space. This means that these vectors can be uniquely specified when their curl and divergences are known at all points in space.

Theorem II: Any vector field that is continuously differentiable in some region can be expressed at every point in the region as the sum of an irrotational vector and a solenoidal vector. Thus,

$$\mathbf{A} = \nabla f + \nabla \times \mathbf{G}, \tag{2.130}$$

where f is a scalar field and $\mathbf{G}$ is a vector field. This identity follows directly from Helmholtz's theorem.

Theorem III: If $\nabla \times \mathbf{A} = 0$ throughout a region, then $\mathbf{A}$ can be represented as

$$\mathbf{A} = \nabla f \tag{2.131}$$

throughout the region, where f is a scalar field. Vectors for which $\nabla \times \mathbf{A} = 0$ are called ***irrotational vectors***. This theorem follows from Helmholtz's theorem and the identity $\nabla \times \nabla f = 0$ (Equation (B.9) in Appendix B).

Theorem IV: If $\nabla \cdot \mathbf{A} = 0$ throughout a region, then $\mathbf{A}$ can be represented as

$$\mathbf{A} = \nabla \times \mathbf{G} \tag{2.132}$$

throughout the region, where $\mathbf{G}$ is a vector field. Vectors for which $\nabla \cdot \mathbf{A} = 0$ are called ***solenoidal vectors***. This theorem follows from Helmholtz's theorem and the identity $\nabla \cdot \nabla \times \mathbf{G} = 0$ (Equation (B.8) in Appendix B).

2-6 Summation

In this chapter, we have presented the basic concepts of vector analysis. While these concepts are firmly rooted in mathematics, our interest in them is solely in their ability to describe physical processes that involve scalar and vector quantities. In the chapters to follow, we will use these concepts freely as we develop the basic equations that define electromagnetics. Vector analysis will also form the basic framework of our analysis and design of electromagnetic systems.

PROBLEMS

2-1 If $\mathbf{A} = 3\hat{\mathbf{a}}_x + 2\hat{\mathbf{a}}_y - 4\hat{\mathbf{a}}_z$ and $\mathbf{B} = -2\hat{\mathbf{a}}_x + \hat{\mathbf{a}}_y + 2\hat{\mathbf{a}}_z$, find:

(a) $|\mathbf{A}|$

(b) $|\mathbf{B}|$

[10] For instance, see Robert Plonsey and Robert Collin, *Principles and Applications of Electromagnetic Fields* (New York: McGraw-Hill, 1961).

(c) $\hat{\mathbf{a}}_B$
(d) $\mathbf{A} + \mathbf{B}$
(e) $\mathbf{A} \cdot \mathbf{B}$
(f) the minimum angle θ_{AB} between **A** and **B**

2-2 If $\mathbf{A} = -\hat{\mathbf{a}}_x + 3\hat{\mathbf{a}}_y - 2\hat{\mathbf{a}}_z$ and $\mathbf{B} = 2\hat{\mathbf{a}}_x + 3\hat{\mathbf{a}}_y - 2\hat{\mathbf{a}}_z$, find:
(a) $|\mathbf{A}|$
(b) $|\mathbf{B}|$
(c) $\mathbf{A} - \mathbf{B}$
(d) $\mathbf{A} \times \mathbf{B}$
(e) the minimum angle θ_{AB} between **A** and **B**

2-3 If $\mathbf{A} = 2\hat{\mathbf{a}}_\rho - \hat{\mathbf{a}}_\phi - 2\hat{\mathbf{a}}_z$, $\mathbf{B} = 3\hat{\mathbf{a}}_\rho + 2\hat{\mathbf{a}}_\phi + 4\hat{\mathbf{a}}_z$, and $\mathbf{C} = \hat{\mathbf{a}}_\rho + 2\hat{\mathbf{a}}_\phi + \hat{\mathbf{a}}_z$, find:
(a) $\mathbf{A} \cdot \mathbf{B}$
(b) minimum angle θ_{AB} between **A** and **B**
(c) $\mathbf{A} \times \mathbf{B}$
(d) the unit vector $\hat{\mathbf{a}}_n$ that points in the direction of $\mathbf{A} \times \mathbf{B}$
(e) $\mathbf{C} \cdot \mathbf{A} \times \mathbf{B}$
(f) $\mathbf{C} \times (\mathbf{A} \times \mathbf{B})$

2-4 Using the Cartesian coordinate system, prove that the following properties of vector addition are true for all vectors:

$$\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C} \qquad \text{(associative law)}$$

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \qquad \text{(commutative law)}$$

$$\mathbf{A} \cdot (\mathbf{B} + \mathbf{C}) = \mathbf{A} \cdot \mathbf{B} + \mathbf{A} \cdot \mathbf{C} \qquad \text{(distributive law)}$$

2-5 If $\mathbf{A} = 2\hat{\mathbf{a}}_x - 3\hat{\mathbf{a}}_y + 2\hat{\mathbf{a}}_z$ at all points P,
(a) find the expression for **A** in the cylindrical coordinate system.
(b) evaluate this expression at the points $P_1(1, 60°, 2)$ and $P_2(2, 30°, 4)$.

2-6 If $\mathbf{A} = 2r\hat{\mathbf{a}}_r - 3r \sin\phi\, \hat{\mathbf{a}}_\theta$, find the representation of **A** in the Cartesian coordinate system.

2-7 The representation of a vector **C** using the Cartesian coordinate system base vectors is $\mathbf{C} = 3\hat{\mathbf{a}}_x + \hat{\mathbf{a}}_y - 3\hat{\mathbf{a}}_z$. Find its representation using the following base vectors:

$$\hat{\mathbf{a}}_1 = \frac{1}{\sqrt{2}}[\hat{\mathbf{a}}_x + \hat{\mathbf{a}}_z]$$

$$\hat{\mathbf{a}}_2 = \frac{1}{\sqrt{2}}[\hat{\mathbf{a}}_x - \hat{\mathbf{a}}_z]$$

$$\hat{\mathbf{a}}_3 = \hat{\mathbf{a}}_y$$

2-8 A force $\mathbf{F} = 10\hat{\mathbf{a}}_x - 8\hat{\mathbf{a}}_y$ [N] is applied to an object that is constrained to travel towards increasing values of x along the path defined by $y = x^2$, $z = 0$. Find the component of **F** that is tangent to this path at the point (2, 4, 0).

2-9 Using integration, calculate the triangular area shown in Figure P2-9.

2-10 Using integration, find the volume of the right pyramid shown in Figure P2-10.

2-11 Evaluate the integral $\int_S \mathbf{D} \cdot \mathbf{ds}$ when $\mathbf{D} = r \sin\theta\, \hat{\mathbf{a}}_r + r \sin\theta\, \hat{\mathbf{a}}_\theta$ and S is a unit sphere centered at the origin.

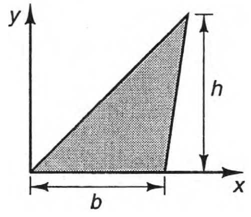


Figure P2-9

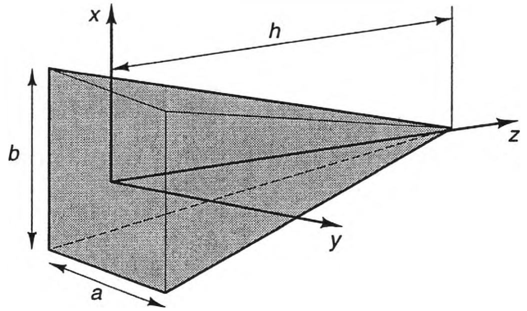


Figure P2-10

2-12 Consider the integral $\int_C \mathbf{F} \cdot \mathbf{d\ell}$, where $\mathbf{F} = \rho \hat{\mathbf{a}}_\rho + z^2 \hat{\mathbf{a}}_\phi$.

(a) Calculate this integral from $P(1, 0°, 0)$ to $P(1, 90°, 2)$ along the path C_1 shown in Figure P2-12, which consists of the arc $\rho = 1, 0 < \phi < \pi/2, z = 0$, followed by the straight line $\rho = 1, \phi = \pi/2, 0 < z < 2$.

(b) Calculate this integral from $P(1, 0°, 0)$ to $P(1, 90°, 2)$ along the path C_2 shown in Figure P2-12 that is defined by the arc $\rho = 1, 0 < \phi < \pi/2, z = 4\phi/\pi$.

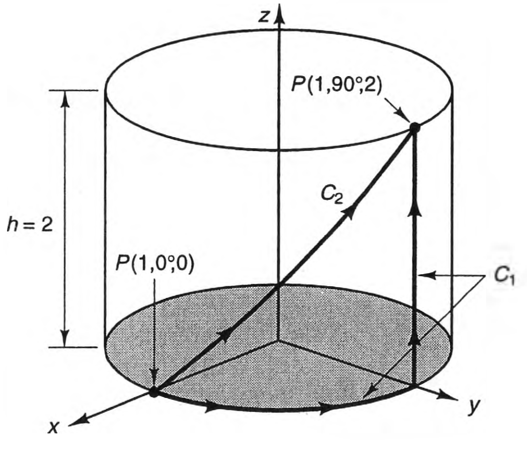


Figure P2-12

2-13 Consider the line integral $\int_C \mathbf{E} \cdot \mathbf{d\ell}$, where $\mathbf{E} = x \hat{\mathbf{a}}_x + 2xy \hat{\mathbf{a}}_y + 3 \hat{\mathbf{a}}_z$.

(a) Calculate this integral along the path C_1 that extends from the origin to the point $P(1, 1, 1)$ along the straight-line segments that sequentially pass through the points $P(0, 0, 0)$, $P(1, 0, 0)$, $P(1, 1, 0)$, and $P(1, 1, 1)$.

(b) Calculate this integral along the path C_2 that extends from the origin to the point $P(1, 1, 1)$ along a single straight line.

2-14 Evaluate the volume integral $\int_V Q dv$, where $Q = 2x^3 z$ when x and z are specified in meters and V is the cube shown Figure P2-14.

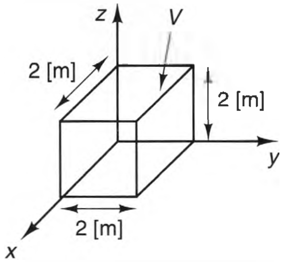


Figure P2-14

2-15 Evaluate the surface integral $\int_S g ds$ over the sector shown in Figure P2-15 if $g = 2\rho \cos \phi$.

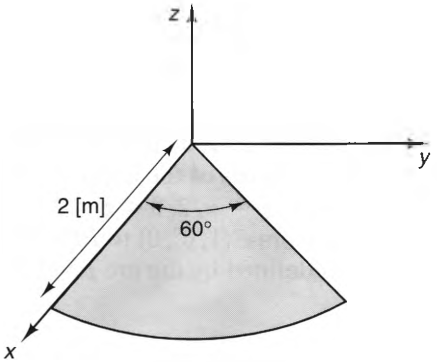


Figure P2-15

2-16 If $\mathbf{F} = xy\hat{\mathbf{a}}_x - y\hat{\mathbf{a}}_y$, calculate the value of the line integral $\int \mathbf{F} \bullet d\boldsymbol{\ell}$ from P_1 to P_2 in Figure P2-16

(a) along a straight line from P_1 to P_2,

(b) along the path P_1 to P_3 to P_2.

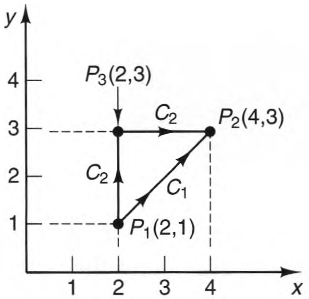


Figure P2-16

2-17 A family of surfaces is defined by the equation

$$2x^2 y + xz = C,$$

where each surface corresponds to a different value of the constant C. Find the unit vector $\hat{\mathbf{a}}_n$ that is directed outward from the surface at the point $P(1, 2, -1)$.

2-18 For the scalar function $g = 2xy + z^2$, find

(a) the magnitude and direction of the maximum rate of change of g at the point $P(1, 3, 2)$.

(b) the rate of change of g along the line directed from $P(1, 3, 2)$ to $P(2, 2, -1)$, evaluated at $P(1, 3, 2)$.

2-19 Consider the line integral $W = \int_{P_1}^{P_2} \mathbf{F} \cdot d\boldsymbol{\ell}$, where $\mathbf{F} = 4y\hat{\mathbf{a}}_y$.

(a) Using the properties of the gradient, prove that the value of W is independent of the path chosen between the endpoints P_1 and P_2.

(b) Find the value of W when the endpoints are $P(1, 0, 0)$ and $P_2(2, -1, 4)$.

2-20 For the function $f = 2xy$,

(a) calculate ∇f in Cartesian coordinates.

(b) express f in cylindrical coordinates and calculate ∇f in cylindrical coordinates.

(c) show that ∇f is the same vector in both coordinate systems by transforming the vector found in a) into cylindrical coordinates.

2-21 If the representation of a vector $\mathbf{A}$ in spherical coordinates is $\mathbf{A} = r\hat{\mathbf{a}}_r$,

(a) calculate$\nabla \cdot \mathbf{A}$ in the spherical coordinate system.

(b) find the representation of $\mathbf{A}$ in the Cartesian coordinate system and then calculate$\nabla \cdot \mathbf{A}$. Is it the same value as found in part a)? Why or why not?

2-22 Evaluate the integral $\oint_S \mathbf{D} \cdot \mathbf{ds}$ over the surface bounding the cube shown in Figure P2-14 when $\mathbf{D} = 2y\hat{\mathbf{a}}_x + xz\hat{\mathbf{a}}_y + z\hat{\mathbf{a}}_z$. Show that the same result is obtained using the divergence theorem by integrating $\nabla \cdot \mathbf{D}$ througout the volume.

2-23 Consider the line integral $\oint_C \mathbf{B} \cdot d\boldsymbol{\ell}$, where $\mathbf{B} = y\hat{\mathbf{a}}_x + z\hat{\mathbf{a}}_y$ and C is a square path in the $z = 0$ plane with sides $x = -1, x = 1, y = -1$ and $y = 1$. Assume that the direction of the path is counterclockwise when looking downward from the $+z$ axis.

(a) Calculate the line integral directly.

(b) Calculate the line integral by using Stokes's theorem and integrating $\nabla \times \mathbf{B}$ over the square surface in the $z = 0$ plane that is bounded by C.

2-24 Find $\nabla \cdot \mathbf{B}$ and $\nabla \times \mathbf{B}$ if

(a) $\mathbf{B} = \rho z\hat{\mathbf{a}}_\rho + \rho^2\hat{\mathbf{a}}_\phi + 2z^2\hat{\mathbf{a}}_z$

(b) $\mathbf{B} = 2xy\hat{\mathbf{a}}_x + 3y\hat{\mathbf{a}}_z$

(c) $\mathbf{B} = 4r\sin\theta\hat{\mathbf{a}}_r + 3r\cos\phi\hat{\mathbf{a}}_\theta$

2-25 Given that $f = r\sin\theta\cos\phi$, calculate

(a) ∇f

(b) $\nabla \times \nabla f$

(c) $\nabla \cdot \nabla f$

2-26 In Figure P2-26, S_1 is a circular disk with unit radius, centered in the $z = 0$ plane, and S_2 is a hemisphere for $z > 0$, centered at the origin with unit radius. If $\mathbf{A} = 3r\hat{\mathbf{a}}_\phi$, calculate $\int \nabla \times \mathbf{A} \cdot \mathbf{ds}$ on S_1 and then on S_2. Assume that the normal direction to both surfaces has a positive z component. Do these integrals have the same values? Why or why not?

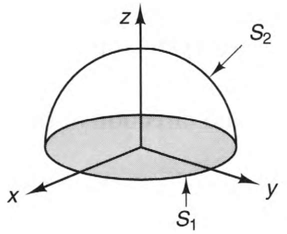


Figure P2-26

2-27 Using the Cartesian coordinate system, verify that the identity $\nabla \times \nabla V = 0$ is valid for all scalar fields V.

2-28 Using the Cartesian coordinate system, verify that the identity $\nabla \cdot \nabla \times \mathbf{A} = 0$ is valid for all vector fields $\mathbf{A}$.

2-29 Using the identity $\nabla^2 \mathbf{A} = \hat{\mathbf{a}}_x \nabla^2 A_x + \hat{\mathbf{a}}_y \nabla^2 A_y + \hat{\mathbf{a}}_z \nabla^2 A_z$, prove that $\nabla^2 \mathbf{A} = \nabla(\nabla \cdot \mathbf{A}) - \nabla \times \nabla \times \mathbf{A}$.

3

Electromagnetic Sources, Forces, and Fields

3-1 Introduction

From the most fundamental point of view, electromagnetics is the study of how electric charges and currents interact with each other. This simple definition may seem somewhat surprising, since we often think of electromagnetics in terms of the things we can do with it, things like communications, control, computing, lighting, and electromechanical applications. But each of these applications, ultimately, is linked to the interactions between the charged particles found in matter. Thus, the starting point to all applications of electromagnetic effects is to understand the nature of charges and current, where and when they occur in materials, and how they can be controlled to produce desired effects.

In this chapter, we will introduce the basic constituents of all electromagnetic effects: sources, forces, and fields. The sources of electromagnetic effects are charges and currents. These quantities are already familiar from circuit analysis. But whereas they are always considered as discrete quantities in circuit analysis, we will usually treat them as field quantities. This will allow us to analyze many kinds of phenomena, devices, and systems that cannot be modeled using ordinary circuit analysis.

The forces that charges and currents exert on each other are important for two reasons. First, these forces are responsible for the way that currents and charges distribute themselves throughout electrical devices and systems. Second, it is these forces that, in the final analysis, are important to people, for without those forces we would never be able to detect the presence, or absence, of electricity. To see why this is important, try to envision a stereo system without speakers or headphones. Even the best stereo system that money could buy would be of no use to us if it were not capable of creating mechanical sound waves that we can hear.

The third topic discussed in this chapter is the electric and magnetic fields that are generated by charges and currents. These fields are the agents through which charges and currents interact with each other, even when they are separated from each other by large distances. We will also introduce the basic equations that relate the fields to their sources. These equations, called Maxwell's equations, are the foundation for all electromagnetic analysis and design.

3-2 Charge and Charge Density

The elemental charged particles are the electron and the proton,[1] which have charges that are equal in magnitude but opposite in sign. The charge of the electron is e, where

$$e = -1.60210 \times 10^{-19} \qquad [\mathrm{C}] \tag{3.1}$$

and "C" is the abbreviation for the basic unit of charge, the coulomb. From this expression, we see that it takes many electrons to make 1 [C] of charge. Electrons can usually be considered to be ***point charges***, since they possess a finite charge within an exceedingly small volume.

Charges produce effects that are a function not only of how much charge is present in a region, but also of how it is distributed. Because of this, it is often necessary to describe charge distributions on a point-by-point basis. For charge distributed throughout a volume, we define the ***volume charge density*** as

$$\rho_v \equiv \lim_{\Delta v \to 0} \frac{\Delta Q}{\Delta v} \qquad [\mathrm{C/m^3}], \tag{3.2}$$

where, as shown in Figure 3-1a, ΔQ is the total charge contained within the volume Δv. Even though there is always some space between the charges in a volume charge distribution, the distances are usually small enough so that the charge can be considered to be a continuous distribution.

There are many situations in which charge is confined to a thin layer. For example, when charge is deposited on a conductor, it is always drawn to the surface. In cases like this, it is convenient to model these charge distributions as ***surface charge***

[1] Although it has been proposed that the quark may be a more fundamental basic building block of matter and possesses a fractional electron charge, it does not appear likely that quarks will ever be observed in nonrelativistic environments.

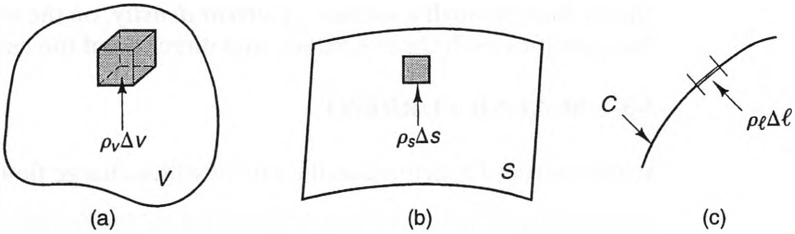


Figure 3-1 Geometries used to define a) volumetric, b) surface, and c) line charge distributions.

distributions, where the charge is assumed to lie within an infinitesimal depth. Referring to Figure 3-1b, we define the ***surface charge density*** as

$$\rho_s \equiv \lim_{\Delta s \to 0} \frac{\Delta Q}{\Delta s} \qquad [\mathrm{C/m^2}], \tag{3.3}$$

where ΔQ is the charge contained within the surface Δs.

There are also situations where charge is confined to lines with small cross sections, such as in a wire or the electron beam in a cathode ray tube (CRT). Since the volume charge density for these distributions is extremely large within the lines and zero outside, it is usually more convenient to consider these distributions as ***line charge distributions***, where the charge is assumed to lie within an infinitesimal cross section along a line. Referring to Figure 3-1c, we define the ***line charge density*** as

$$\rho_\ell \equiv \lim_{\Delta \ell \to 0} \frac{\Delta Q}{\Delta \ell} \qquad [\mathrm{C/m}], \tag{3.4}$$

where ΔQ is the charge that lies within the length $\Delta \ell$.

The total charge contained within a volume, surface, or line can be determined in terms of the volume, surface, or line charge densities by integrating Equations (3.2)–(3.4), respectively. The resulting expressions are

$$Q = \int_V \rho_v \, dv \qquad (\rho_s = \rho_\ell = 0 \text{ inside } V) \tag{3.5}$$

$$Q = \int_S \rho_s \, ds \qquad (\rho_\ell = 0 \text{ inside } S) \tag{3.6}$$

$$Q = \int_C \rho_\ell \, d\ell \tag{3.7}$$

3-3 Current and Current Density

Current is charge in motion. We can specify current by using either a vector or a scalar quantity. The scalar quantity, called ***scalar current*** (or simply ***current***), should be already familiar from circuit analysis. It is useful when it is enough to know the rate of charge flow through a surface. ***Current density***, on the other hand, is a vector quantity that specifies both the magnitude and direction of the current flow[2] at any point.

3-3-1 SCALAR CURRENT

Scalar current I is defined as the rate at which charge flows through a specified surface,

$$I \equiv \frac{dQ}{dt} \qquad [\text{C/s or A}], \tag{3.8}$$

where "A" is the abbreviation for ampere. In this definition, dQ is the charge that passes through a surface S in the time dt. The sign of dQ depends on the direction of this flow with respect to the surface normal $\hat{\mathbf{a}}_n$; positive charge moving through S in the direction indicated by $\hat{\mathbf{a}}_n$ constitutes a positive current, as does negative charge passing through S in the opposite sense. In circuits, scalar current is indicated on diagrams and schematics by showing the numerical value of I, together with an arrow that defines the direction of positive current. As an example, consider the positive and negative ions flowing in the pipe shown in Figure 3-2. Here, both the positive and negative ions impart positive contributions to I, since $\mathbf{u}_+ \cdot \hat{\mathbf{a}}_n > 0$ and $\mathbf{u}_- \cdot \hat{\mathbf{a}}_n < 0$.

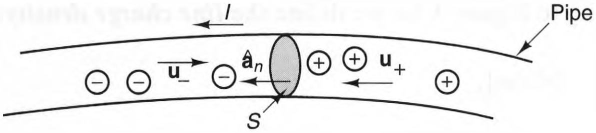


Figure 3-2 Positive and negative charge carriers flowing in a pipe.

3-3-2 CURRENT DENSITY

Describing a current using the scalar current is acceptable when the direction of the flow is obvious, such as when current flows on wires in a low-frequency circuit. But there are many times when the direction of the current and its magnitude vary throughout a volume. In these cases it is best to represent the current as a ***volume current density***, which is a vector quantity.

Figure 3-3 shows several streamlines that indicate the paths of moving charges. We define the volume current density at a point by

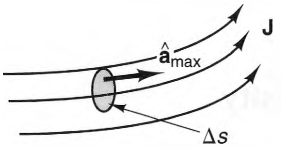


Figure 3-3 Current streamlines flowing past a small surface.

[2] Although the term "current flow" is redundant, it is nevertheless customary to use the terms "current," "current flow," and "charge flow" interchangeably.

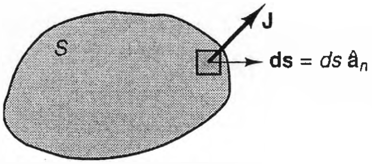


Figure 3-4 Geometry for determining the current passing through a surface.

$$\mathbf{J} = \lim_{\Delta s \to 0} \left. \frac{\Delta I}{\Delta s} \right|_{\max} \hat{\mathbf{a}}_{\max} \quad [\mathrm{A/m^2}], \tag{3.9}$$

where $\hat{\mathbf{a}}_{\max}$ is perpendicular to the surface Δs and points in the direction that maximizes the current ΔI flowing through Δs.

To see how the current density $\mathbf{J}$ and the current I are related, consider the situation shown in Figure 3-4. Here, a $\mathbf{J}$ current passes through a surface S. We can find the total current I flowing outward through S by summing the contributions dI that pass through each differential surface element $\mathbf{ds} = ds\,\hat{\mathbf{a}}_n$, where $\hat{\mathbf{a}}_n$ points outward from the surface. When $\mathbf{J}$ and $\mathbf{ds}$ are perpendicular, $\mathbf{J}$ has no tendency to flow through the surface, so dI is zero. On the other hand, when $\mathbf{J}$ and $\mathbf{ds}$ are parallel, $dI = Jds$. Hence, we can write

$$dI = \mathbf{J} \cdot \mathbf{ds}. \tag{3.10}$$

Integrating dI over all of S, we obtain the total current I passing through this surface:

$$I = \int_S \mathbf{J} \cdot \mathbf{ds} \quad [\mathrm{A}]. \tag{3.11}$$

When S is a closed surface, $\mathbf{ds}$ is chosen by convention to point away from the enclosed volume, and thus, I is defined as an outward flow.

There are many situations in which a current flows within a thin layer. For instance, at high frequencies, current flows within a thin layer under the surface of a good conductor. We can model these current distributions as ***surface current distributions***, where the current is assumed to flow within a layer of infinitesimal depth. Referring to Figure 3-5a, we define surface current density as

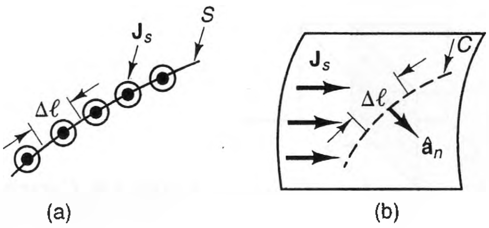


Figure 3-5 A surface current: a) cross-sectional view, b) top view.

$$\mathbf{J}_s = \lim_{\Delta\ell\to 0} \left.\frac{\Delta I}{\Delta\ell}\right|_{\max} \hat{\mathbf{a}}_{\max} \qquad [\text{A/m}], \tag{3.12}$$

where $\Delta\ell$ is measured perpendicular to the direction $\hat{\mathbf{a}}_{\max}$ that maximizes the ratio $\Delta I/\Delta\ell$. In this figure, both $\hat{\mathbf{a}}_{\max}$ and $\mathbf{J}_s$ are directed out of the paper.

The scalar current flowing past an arbitrary contour C on a surface can be found by noting that the current flowing past the segment in Figure 3-5b is

$$\Delta I = \mathbf{J}_s \cdot \Delta\ell\hat{\mathbf{a}}_n, \tag{3.13}$$

where $\hat{\mathbf{a}}_n$ is in the plane of the surface current and perpendicular to the differential path $\Delta\ell$. Integrating over the entire contour C, we find that the total current that crosses the contour C is

$$I = \int_C \mathbf{J}_s \cdot \hat{\mathbf{a}}_n \, d\ell \qquad [\text{A}]. \tag{3.14}$$

3-3-3 THE CURRENT DENSITY OF A MOVING CHARGE DISTRIBUTION

Since current is charge in motion, it is often convenient to specify it in terms of the velocity of the charge movement. These currents are often called ***convection currents***. Referring to Figure 3-6, let us consider a volume charge distribution ρ_v that moves with velocity $\mathbf{u} = \text{u}_x\hat{\mathbf{a}}_x$. In the time interval Δt, each elemental charge in this distribution moves a distance $\Delta\ell = u_x\Delta t$. Thus, charge moves through the surface Δs at the rate

$$\Delta I = \frac{\Delta Q}{\Delta t} = \rho_v u_x \Delta s. \tag{3.15}$$

Also, from Equation (3.10), we have

$$\Delta I = \mathbf{J} \cdot \Delta\mathbf{s} = J_x \Delta s. \tag{3.16}$$

Substituting Equation (3.15) into Equation (3.16), we find that the convection current density in Figure 3-6 is

$$\mathbf{J} = \rho_v u_x \hat{\mathbf{a}}_x \qquad [\text{A/m}^2]. \tag{3.17}$$

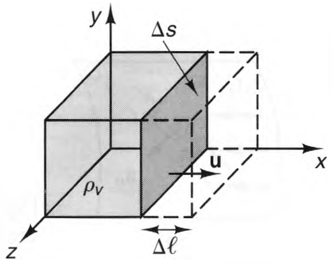


Figure 3-6 Current resulting from a moving charge distribution.

Similar expressions are obtained when the charge distribution has velocity components in the y- and z-directions. Summing these velocity components, we obtain the general expression for a charge distribution with an arbitrary velocity $\mathbf{u}$:

$$\mathbf{J} = \rho_v \mathbf{u} \qquad [\mathrm{A/m^2}]. \tag{3.18}$$

If a moving charge distribution is confined to a surface, similar analysis shows that the surface current density is

$$\mathbf{J}_s = \rho_s \mathbf{u} \qquad [\mathrm{A/m}], \tag{3.19}$$

where $\mathbf{u}$ is the velocity of ρ_s in the plane of the surface. Likewise, when charge flows along a line ℓ with velocity $\mathbf{u} = u\hat{\mathbf{a}}_\ell$, the scalar current along the line is

$$I = \rho_\ell u \qquad [\mathrm{A}]. \tag{3.20}$$

There are many situations in which two or more kinds of charge carriers move with different velocities within a material. For instance, this happens in semiconductors, where electrons and holes[3] typically move in opposite directions at different speeds. For cases where there are N kinds of charge carriers present, we can generalize Equation (3.18) to read

$$\mathbf{J} = \sum_{i=1}^{N} \rho_{vi} \mathbf{u}_i \qquad [\mathrm{A/m^2}], \tag{3.21}$$

where ρ_{vi} and $\mathbf{u}_i$ are the charge density and velocity of the ith charge distribution, respectively. An important point to note about this expression is that a net current can exist even when the net charge density is zero (i.e., when $\Sigma_{i=1}^{N} \rho_{vi} = 0$). This occurs when different types of charge move in different directions. The following example demonstrates the point.

Example 3-1

Consider the bulk piece of germanium with cross section 1 [cm^2] shown in Figure 3-7. Assume that the electron charge density is $\rho_{v-} = -4 \times 10^{-6}$ [C/cm^3] and moves with velocity

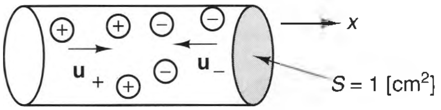


Figure 3-7 Current flow in a section of bulk germanium.

[3] Although there is no such particle as a hole, it has been found that the net actions of many valence electrons in a semiconductor can be accurately modeled by a single fictitious hole particle possessing a charge of +|e|.

$\mathbf{u}_- = -15\hat{\mathbf{a}}_x$ [cm/s]. Also, the hole charge density is $\rho_{v+} = +4 \times 10^{-6}$ [C/cm^3] and has velocity $\mathbf{u}_+ = +10\hat{\mathbf{a}}_x$ [cm/s]. Find the current through the semiconductor cross section.

Solution:

From Equation (3.21), the current density inside the semiconductor is

$$\mathbf{J} = 4 \times 10^{-6} \times 10\hat{\mathbf{a}}_x - 4 \times 10^{-6} \times (-15)\hat{\mathbf{a}}_x = 1.0 \times 10^{-4}\, \hat{\mathbf{a}}_x \text{ [A/cm}^2\text{]}.$$

Using Equation (3.11) and noting that **J** and **ds** are parallel, we find that the current flowing through the semiconductor cross section is

$$I = \int_S \mathbf{J} \cdot \mathbf{ds} = 1 \times 10^{-4} \text{ [A/cm}^2\text{]} \times 1.0 \text{ [cm}^2\text{]} = 0.1 \text{ [mA]}.$$

3-4 The Law of Charge Conservation

Now that we have discussed the concepts of current and charge, we are ready to introduce the law of charge conservation. This principle is one of the basic postulates upon which all electromagnetic theory rests. As with all experimental laws, it is based on observation and is accepted as true because no contradictory evidence has ever been found.

The ***law of charge conservation*** states that the charge contained in a closed system remains constant for all time.

A closed system is a system in which charge can neither enter nor leave. A corollary of this law states that if the total charge contained within a region changes, it must be accompanied by a net current flow either into or out of the region.

To see what constraints this law places on charge and current distributions, consider the current I passing outward through the closed surface S shown in Figure 3-8:

$$I = \oint_S \mathbf{J} \cdot \mathbf{ds} = \frac{dQ_{\text{out}}}{dt}, \tag{3.22}$$

where dQ_{out} is the charge passing outward through S in time dt. According to the law of charge conservation, dQ_{out} cannot be created spontaneously at S, but rather must come from within the volume V that is bounded by the closed surface S. This means that the total charge Q_{enc} enclosed within V must decrease at exactly the same rate at which charge passes outward through S. Thus,

$$\frac{dQ_{\text{out}}}{dt} = -\frac{dQ_{\text{enc}}}{dt}. \tag{3.23}$$

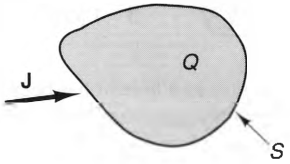


Figure 3-8 Geometry for deriving the continuity equation.

Substituting Equation (3.23) into Equation (3.22) yields

$$\oint_S \mathbf{J}\cdot d\mathbf{s} = -\frac{dQ_{\text{enc}}}{dt}. \tag{3.24}$$

This expression is called the ***continuity equation***, because it states that the charge contained in any region is constant when no current flows out of the region.

We can derive a point form of the continuity equation by remembering that the charge inside S can be expressed as a volume integral of the charge density. Thus,

$$\oint_S \mathbf{J}\cdot d\mathbf{s} = -\frac{dQ_{\text{enc}}}{dt} = -\frac{d}{dt}\int_V \rho_v\, dv,$$

where V is the volume enclosed by S. If V is constant in time, the order of differentiation and integration can be interchanged, yielding

$$\oint_S \mathbf{J}\cdot d\mathbf{s} = -\int_V \frac{\partial \rho_v}{\partial t}\, dv.$$

Next, we can use the divergence theorem to write the integral on the left as a volume integral, yielding

$$\oint_S \mathbf{J}\cdot d\mathbf{s} = \int_V \nabla\cdot\mathbf{J}dv = -\int_V \frac{\partial \rho_v}{\partial t}\, dv.$$

Finally, since this expression is valid for all volumes V, it must be valid as $V \to 0$. Thus, the integrands of both volume integrals must be equal at all points, yielding

$$\nabla\cdot\mathbf{J} = -\frac{\partial \rho_v}{\partial t}. \tag{3.25}$$

This expression is called the differential (or point) form of the continuity equation.

Example 3-2

Suppose that a current density throughout a region is specified by $\mathbf{J} = re^{-(r/\alpha)}\hat{\mathbf{a}}_r$ [A/m^2], where α is a constant that is measured in meters. Find the corresponding charge density.

Solution:

Since $\mathbf{J}$ has only an r component, we have, from the continuity equation,

$$\nabla\cdot\mathbf{J} = \frac{1}{r^2}\left[\frac{\partial}{\partial r}(r^2 J_r)\right] = \left[3 - \frac{r}{\alpha}\right]e^{-(r/\alpha)} = -\frac{\partial \rho_v}{\partial t}.$$

Integrating this result over time, we obtain

$$\rho_v = -\int_{t_o}^{t}\left[3 - \frac{r}{\alpha}\right]e^{-(r/\alpha)}\,dt' = [t - t_o]\left[\frac{r}{\alpha} - 3\right]e^{-(r/\alpha)} + \rho_v(t_o),$$

where $\rho_v(t_o)$ is the charge density at $t = t_o$, where t_o can be any value of time.

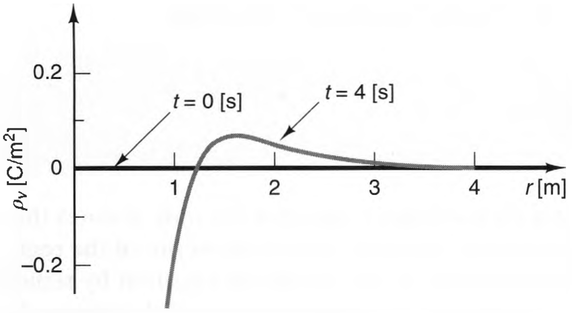


Figure 3-9 Charge density as a function of position at two points in time corresponding to an outward-flowing current.

Figure 3-9 shows ρ_v as a function of r for two values of t for the case when $\alpha = 0.4$ [m] and $\rho_v(t = 0) = 0$ for all values of r. Here we see that ρ_v decreases with time for small values of r as time progresses. This occurs because **J** is directed away from the origin, which means that charge is depleted there. On the other hand, more charge is entering the region $r \approx 2$ [m] than is leaving, so the charge density there increases with time.

3-4-1 KIRCHHOFF'S CURRENT LAW

When the charge Q enclosed by a closed surface S is time invariant, $\dfrac{\partial Q}{\partial t} = 0$, and the continuity equation becomes

$$\oint_S \mathbf{J} \cdot d\mathbf{s} = 0 \qquad \left(\frac{\partial Q_{\text{enc}}}{\partial t} = 0\right). \tag{3.26}$$

This expression is the integral form of ***Kirchhoff's current law*** (KCL), which is a foundational principle in circuit theory. At low frequencies, it follows from Equation (3.26) that the currents in the N-wire junction shown in Figure 3-10a satisfy the relation

$$\sum_{i=1}^{N} I_i = 0 \quad \text{(low frequencies)}, \tag{3.27}$$

where each current I_i is defined outward from the junction.

At higher frequencies, the charge density at the junction will often vary with time. This is the result of the finite capacitance between the conductors in the network, which is often called ***stray capacitance.*** This capacitance can be modeled as one or

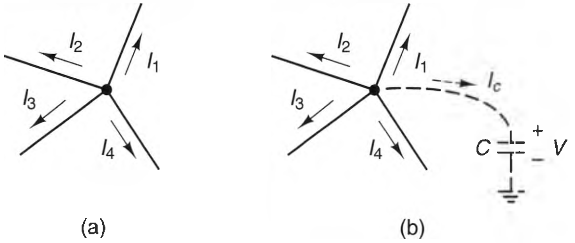


Figure 3-10 Circuit diagrams for Kirchhoff's current law at a point. a) At low frequencies, stray capacitance can be ignored. b) At high frequencies, stray capacitance cannot be ignored.

more lumped capacitances, such as the one shown in Figure 3-10b. Here, the wires connecting the capacitor to the node and ground are shown as dotted lines, indicating that they are not physical wires. When the effects of this capacitance are included, the continuity equation yields

$$\sum_{i=1}^{N} I_i = -\frac{\partial Q}{\partial t} = -I_c = -C\frac{\partial V}{\partial t} \quad \text{(high frequencies)}, \tag{3.28}$$

where V is the voltage between the junction and the ground conductors, C is the junction capacitance to ground, and I_c is the current flowing through this capacitance. Notice that this is the same KCL expression as would be obtained if the node-to-ground capacitance C were an actual lumped capacitor. Later in this text we will show that the current I_c is an example of something called displacement current, which occurs whenever there is a time-varying charge density at a point.

3-5 Two Action-at-a-Distance Force Laws

Throughout the 19th century, studies of electromagnetic effects were conducted by many investigators. As the experimental evidence mounted, laws were proposed to explain these effects. The first laws that were proposed viewed the effects as point-by-point interactions of charges and currents and are called ***action-at-a-distance*** laws. The simplest of these laws are Coulomb's force law, which predicts the force between two charges, and Ampère's force law, which predicts the force between two currents. We will discuss these laws first, because they are simple and easy to visualize.

3-5-1 COULOMB'S LAW OF FORCE

In its simplest form, Coulomb's law of force describes the force between stationary point charges suspended in free space.[4] Figure 3-11 depicts two point charges, Q_1 and Q_2, located at $\mathbf{r}_1$ and $\mathbf{r}_2$, respectively.

Coulomb's law states that the force exerted on a point charge Q_1 by another point charge Q_2 is

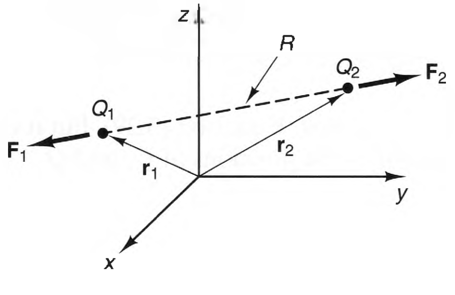


Figure 3-11 Two point charges exerting forces on each other.

[4] *Free space* is a term commonly used in electromagnetics to descibe a medium that contains no charged particles, such as a vacuum.

$$\mathbf{F}_1 = \frac{1}{4\pi\epsilon_o}\frac{Q_1Q_2}{R^2}\hat{\mathbf{a}}_{21} \quad \text{[N] (charges in free space)}, \tag{3.29}$$

where R is the distance from Q_2 to Q_1, $\hat{\mathbf{a}}_{21}$ is the unit vector directed from Q_2 to Q_1, and "N" is the abbreviation for the newton, the fundamental unit of force. Also, ϵ_o is a physical constant of proportionality called the ***permittivity of free space***, whose value is

$$\epsilon_o = 8.854 \times 10^{-12} \approx \frac{10^{-9}}{36\pi} \quad \left[\frac{\text{C}^2}{\text{N} \cdot \text{m}^2} \text{ or F/m}\right], \tag{3.30}$$

where "F" is the abbreviation for the farad. As its unit implies, the permittivity of free space plays an important role in capacitance, which we will discuss in Chapter 5.

As can be seen from Equation (3.29), $\mathbf{F}_1$ is directed along the line extending from Q_1 to Q_2. The magnitude of $\mathbf{F}_1$ is proportional to the product of the magnitudes of the two charges and inversely proportional to the square of the distance between them. The sign of the product Q_1Q_2 determines whether the force is attractive or repulsive; like charges repel and unlike charges attract. Also, since $\hat{\mathbf{a}}_{21} = -\hat{\mathbf{a}}_{12}$, we find that $\mathbf{F}_1 = -\mathbf{F}_2$. Thus, Coulomb's law is consistent with Newton's third law of mechanics, which states that every action has an equal and opposite reaction.

Coulomb's law of force can also be expressed in terms of the position vectors of the two charges. Using Equation (2.67), we have

$$\hat{\mathbf{a}}_{21} = \frac{(\mathbf{r}_1 - \mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|}. \tag{3.31}$$

Also, $R = |\mathbf{r}_1 - \mathbf{r}_2|$. Substituting these into Equation (3.29), we find that $\mathbf{F}_1$ can be expressed as

$$\mathbf{F}_1 = \frac{Q_1Q_2}{4\pi\epsilon_o}\frac{(\mathbf{r}_1 - \mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|^3} \quad \text{[N] (charges in free space)}. \tag{3.32}$$

This form of $\mathbf{F}_1$ is less elegant than Equation (3.29), but it is more suitable for calculations, since it clearly identifies the positions of Q_1 and Q_2. This can be seen in the following example.

Example 3-3

The point charges Q_1 and Q_2 are located at $P_1 = (1, -2, 3)$ [m] and $P_2 = (2, 3, -4)$ [m], respectively. Calculate the force that Q_2 exerts on Q_1 if $Q_2 = +1$ [C] and $Q_1 = -2$ [C].

Solution:

From the given values of P_1 and P_2,

$$\mathbf{r}_1 = \hat{\mathbf{a}}_x - 2\hat{\mathbf{a}}_y + 3\hat{\mathbf{a}}_z \text{ [m]}, \quad \mathbf{r}_2 = 2\hat{\mathbf{a}}_x + 3\hat{\mathbf{a}}_y - 4\hat{\mathbf{a}}_z \text{ [m]},$$

and

$$|\mathbf{r}_1 - \mathbf{r}_2| = \sqrt{1^2 + 5^2 + 7^2} = \sqrt{75} \text{ [m]}.$$

Substituting into Equation (3.32), we obtain

$$\mathbf{F}_1 = \frac{-2}{4\pi\epsilon_0}\frac{(-\hat{\mathbf{a}}_x - 5\hat{\mathbf{a}}_y + 7\hat{\mathbf{a}}_z)}{75^{3/2}} = -2.77 \times 10^7(-\hat{\mathbf{a}}_x - 5\hat{\mathbf{a}}_y + 7\hat{\mathbf{a}}_z) = -2.40 \times 10^8\,\hat{\mathbf{a}}_{21} \quad \text{[N]},$$

where

$$\hat{\mathbf{a}}_{21} = \frac{-\hat{\mathbf{a}}_x - 5\hat{\mathbf{a}}_y + 7\hat{\mathbf{a}}_z}{\sqrt{75}} = -\hat{\mathbf{a}}_{12}.$$

3-5-2 AMPÈRE'S LAW OF FORCE

Ampère's law of force describes the force exerted by one current upon another when both currents are time invariant. Such currents are called ***steady currents***. When steady currents flow in complete loops, no net charge is transported, and the charge density is everywhere constant (i.e., static) in time.

Figure 3-12 shows two differential-length filaments $\mathbf{d\ell}_1$ and $\mathbf{d\ell}_2$ that carry steady currents I_1 and I_2, respectively. In practice, such filaments must be sections of complete circuits or loops, but we will for now consider only these short sections.

Through a series of cleverly devised experiments,[5] André Marie Ampère (1775–1836) deduced that in free space, the current segment $I_2\mathbf{d\ell}_2$ exerts a force on the current segment $I_1\mathbf{d\ell}_1$ that is given by

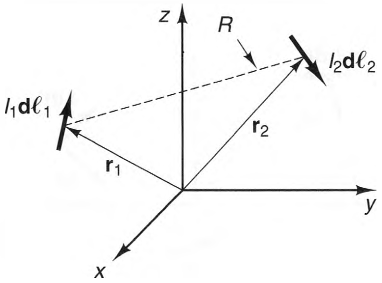


Figure 3-12 Two current filaments exerting forces on each other.

[5] For a complete description of Ampère's experiments, see James Clerk Maxwell, *Electricity and Magnetism* (New York: Dover, 1954), Volume 2, Part 4, Chapter 2. In Chapter 3 of Part 4, Maxwell states, "The experimental investigation by which Ampere established the laws of the mechanical action between electric currents is one of the most brilliant achievements in science. The whole, theory and experiment, seems as if it had leaped, full grown and full armed, from the brain of the 'Newton of electricity.'"

$$\mathbf{dF}_1 = \frac{\mu_o}{4\pi}\frac{I_1\mathbf{d\ell}_1 \times (I_2\mathbf{d\ell}_2 \times \hat{\mathbf{a}}_{21})}{R^2} \quad [\mathrm{N}] \quad \text{(currents in free space)}, \tag{3.33}$$

where the unit vector $\hat{\mathbf{a}}_{21}$ points from segment 2 to segment 1 and R is the distance between the segments. In addition, μ_o is a physical constant of proportionality called the ***permeability of free space***:

$$\mu_o = 4\pi \times 10^{-7} \quad [\mathrm{N/A^2 \text{ or } H/m}], \tag{3.34}$$

where "H" is the abbreviation for the henry. The cross-product expression can be rewritten in terms of dot products by using Equation (B.2) in Appendix B, yielding the alternate form

$$\mathbf{dF}_1 = \frac{\mu_o}{4\pi} I_1 I_2 \frac{(\mathbf{d\ell}_1 \cdot \hat{\mathbf{a}}_{21})\mathbf{d\ell}_2 - (\mathbf{d\ell}_1 \cdot \mathbf{d\ell}_2)\hat{\mathbf{a}}_{21}}{R^2}. \tag{3.35}$$

Comparing Equations (3.33) and (3.35) with Equation (3.32), we see that the forces between infinitesimal current filaments bear some similarities with the forces between point charges. In particular, both forces vary inversely with the square of the distance R between the sources. Also, both forces are proportional to the products of the source values, Q_1Q_2 for charges and I_1I_2 for currents. But whereas the force exerted by one static charge upon another is always directed along the line between them, the direction of the force exerted by one steady-current filament upon another depends upon their orientations relative to each other.

Figure 3-13 shows the relationship between the orientations of currents and the direction of the resulting force for three different cases. From Equation (3.35) we see that $\mathbf{dF}_1$ is in the plane that contains the vectors $\mathbf{d\ell}_2$ and $\hat{\mathbf{a}}_{21}$ and is also perpendicular to $\mathbf{d\ell}_1$. This is depicted in Figure 3-13a. When $\mathbf{d\ell}_1$ is perpendicular to $\hat{\mathbf{a}}_{21}$, $\mathbf{dF}_1$ and $\hat{\mathbf{a}}_{21}$ are collinear (just as in the case of the force between static charges).

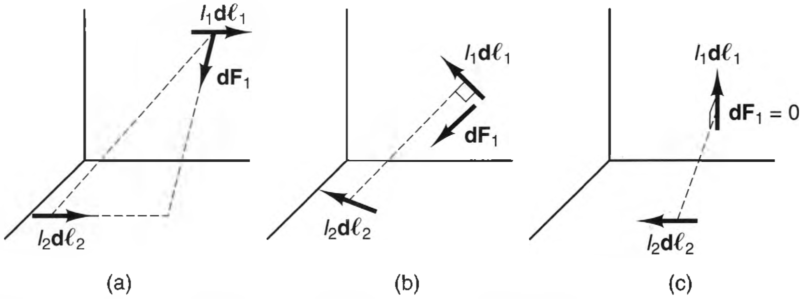


Figure 3-13 Graphical depiction of the force between current filaments. a) Parallel filaments. b) Filament #1 is perpendicular to the line connecting the filaments. c) Filament #1 is perpendicular to the line connecting the filaments, and the filaments are perpendicular to each other.

This is depicted in Figure 3-13b. Figure 3-13c shows that the interaction force $\mathbf{dF}_1$ is zero when the current filaments are perpendicular to each other and $\mathbf{d\ell}_1$ is perpendicular to $\hat{\mathbf{a}}_{21}$.

Example 3-4

Figure 3-14 shows two parallel current filaments $I_1\,\mathbf{d\ell}_1$ and $I_2\mathbf{d\ell}_2$. Both filaments are of length $d\ell$ and are directed perpendicular to the line that connects them. Calculate the force acting on I_1.

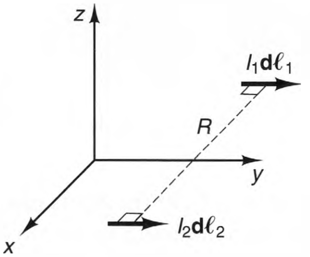


Figure 3-14 Two parallel current filaments that are perpendicular to the line connecting them.

Solution:

As can be seen from Figure 3-14, both current segments are directed along the y-axis, so we can write $I_1\mathbf{d\ell}_1 = I_1 d\ell\hat{\mathbf{a}}_y$ and $I_2\mathbf{d\ell}_2 = I_2 d\ell\hat{\mathbf{a}}_y$. The filaments are parallel, so $\mathbf{d\ell}_1 \cdot \mathbf{d\ell}_2 = (d\ell)^2$. Also, both filaments are perpendicular to the line that connects them, so $\mathbf{d\ell}_1 \cdot \hat{\mathbf{a}}_{21} = \mathbf{d\ell}_2 \cdot \hat{\mathbf{a}}_{21} = 0$. Substituting these expressions into Equation (3.35), we find that the force acting on $I_1\,\mathbf{d\ell}_1$ is

$$\mathbf{dF}_1 = -\frac{\mu_o}{4\pi} I_1 I_2 \frac{(d\ell)^2}{R^2}\hat{\mathbf{a}}_{21} \quad [\mathrm{N}], \tag{3.36}$$

where R is the distance between the elements. Knowing that $\hat{\mathbf{a}}_{21}$ is directed from element 2 to element 1, we conclude that $\mathbf{dF}_1$ is an attractive force when both currents flow in the same direction (i.e., when $I_1 I_2 > 0$) and a repulsive force when the currents are oppositely directed.

3-6 The Lorentz Force Law and the Field Concept of Electromagnetics

The action-at-a-distance force laws discussed in the previous section are fundamental postulates of physics whose experimental validity is unquestioned. Important as they are, however, our discussion of electromagnetics would be severely limited if we were to model all electromagnetic forces and effects with action-at-a-distance laws. This is because such laws require that we know the values and locations of all the currents and charges in a system at all times. But in many cases this information is not available, at least not entirely. Imagine, for instance, the difficulty in measuring the current distribution on an antenna in city A when you are located in city B. In situations like this, a more useful view of electromagnetic effects is one that assumes that each electromagnetic source emits "something" throughout all space that interacts with other sources. This kind of view is called a ***field theory***, because fields are entities that exist continuously throughout regions of space.

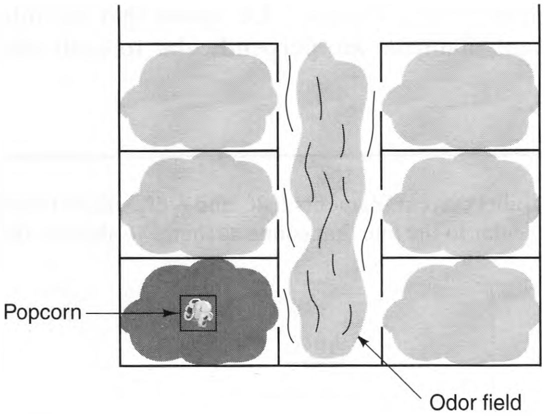


Figure 3-15 An example of a field quantity: popcorn odor.

To get an idea of what a field theory is like, let us start by considering a common example: odors. Figure 3-15 depicts a situation known to anyone who has ever lived in a dormitory. As any student knows, it is hard to keep a bowl of warm popcorn a secret. Try as you might, all kinds of people become your "friend" when they know that you have some popcorn. There are two ways in which we can understand how people in other rooms are drawn to the popcorn. The first is to say that they are drawn directly by the popcorn itself. This is an action-at-a-distance view of the interaction. The other view is to say that what really attracts people to warm popcorn is its odor, which can be sensed at large distances from the popcorn itself. This is a field view, since the odor density is a field quantity that exists at each point throughout the building. The advantage of this view is that it is not necessary to know the position of the popcorn to predict its effect. All that is needed is a knowledge of the odor field, which can be measured remotely. Experience tells us that we can tell a lot about the source of an odor simply by sniffing around.

It was Michael Faraday (1791–1867) who first proposed a field theory of electromagnetics. Faraday speculated that "something" propagates outward from charges and currents and manifests itself in two vector fields: the ***electric field intensity*** **E** and the ***magnetic flux density*** **B**. This was a bold proposal in the 1830s, since no one had yet speculated on the existence of photons, which we now know are the agents responsible for these fields.

The starting point of the field theory of electromagnetics is the ***Lorentz force law***, which states that the total electromagnetic force acting on a point "test charge" of value Q can always be expressed as

$$\mathbf{F} = Q(\mathbf{E} + \mathbf{u} \times \mathbf{B}) \quad [\mathrm{N}], \tag{3.37}$$

where **E** is the electric field intensity at Q, **B** is the magnetic flux density at Q, and **u** is the velocity of Q with respect to the "laboratory" frame of reference. The electric field intensity **E** is measured in units of newtons per coulomb [N/C], which is equivalent to volts per meter [V/m]. The magnetic flux density **B** is measured in units of newton • seconds per coulomb • meter [N • s/(C • m)], which is equivalent to webers per square meter [Wb/m^2] or tesla [T]. In Equation (3.37), both **E** and **B** are measured in the laboratory reference frame.[6]

The Lorentz force law was named in honor of Hendrik Lorentz (1853–1928), although it also contains contributions from Joseph Thomson (1856–1940) and Oliver Heaviside (1850–1925). Like the Coulomb and Ampère force laws, this law is also based on experimental evidence. But it differs from the action-at-a-distance force laws in that the force on a charge is described in terms of the E- and B-fields generated by other sources whose locations and magnitudes need not be known. Also, the Lorentz force law is valid when time-varying charge and current distributions are present, while the Coulomb and Ampère force laws are valid only for time-invariant sources.

According to the Lorentz force law, there are two distinct kinds of force that other currents and charges can exert upon a test charge. The first is the ***electric force***, which is defined as

$$\mathbf{F}_e = Q\,\mathbf{E} \quad [\mathrm{N}]. \tag{3.38}$$

This force depends only on the magnitude Q of the test charge and the magnitude and direction of the electric field **E**. Also, its direction is always collinear with the direction of **E**. The other force predicted by the Lorentz force law is called the ***magnetic force***, which is defined as

$$\mathbf{F}_m = Q\,\mathbf{u} \times \mathbf{B} \quad [\mathrm{N}]. \tag{3.39}$$

Like the electric force $\mathbf{F}_e$, the magnetic force $\mathbf{F}_m$ is also proportional to the magnitude Q of the test charge. These forces differ, however, in that the electric force is independent of the velocity of the test charge, whereas the magnetic force is present only when the test charge is moving. Also, the direction of the magnetic force is perpendicular to both the direction of the test charge velocity **u** and the direction of the magnetic flux density **B**.

Although charged particles can experience both electric and magnetic forces, only the electric force can do work on a charge and change its kinetic energy. This is because the magnetic force is always perpendicular to the motion of the charge. Magnetic forces can, however, change the direction of a *moving* charge.

Example 3-5

A particular source distribution produces the following E- and B-fields at the location of a test charge Q:

$$\mathbf{E} = 2\hat{\mathbf{a}}_x - 3\hat{\mathbf{a}}_y + 4\hat{\mathbf{a}}_z \quad [\mu\mathrm{V/m\ or\ }\mu\mathrm{N/C}]$$

$$\mathbf{B} = 6\hat{\mathbf{a}}_x + 8\hat{\mathbf{a}}_y - 4\hat{\mathbf{a}}_z \left[\mu\mathrm{T\ or\ }\frac{\mu\mathrm{N}\cdot\mathrm{s}}{\mathrm{C}\cdot\mathrm{m}}\right].$$

Find the total force acting on Q if its velocity is $\mathbf{u} = 4\hat{\mathbf{a}}_x - 3\hat{\mathbf{a}}_y$ [m/s] and $Q = +2$ [μC].

[6] In Chapter 9 we will show that **E** is different when measured in moving and stationary reference frames.

Solution:

According to the Lorentz force law, the electric force

$$\mathbf{F}_e = Q\,\mathbf{E} = 2 \times 10^{-6} \times (2\hat{\mathbf{a}}_x - 3\hat{\mathbf{a}}_y + 4\hat{\mathbf{a}}_z) \times 10^{-6}$$
$$= 4\hat{\mathbf{a}}_x - 6\hat{\mathbf{a}}_y + 8\hat{\mathbf{a}}_z \quad [\text{pN}].$$

The magnetic force

$$\mathbf{F}_m = Q\,\mathbf{u} \times \mathbf{B} = 2 \times 10^{-6} \times (4\hat{\mathbf{a}}_x - 3\hat{\mathbf{a}}_y) \times (6\hat{\mathbf{a}}_x + 8\hat{\mathbf{a}}_y - 4\hat{\mathbf{a}}_z) \times 10^{-6}$$
$$= 24\hat{\mathbf{a}}_x + 32\hat{\mathbf{a}}_y + 100\hat{\mathbf{a}}_z \quad [\text{pN}].$$

The total force acting on Q is thus

$$\mathbf{F} = \mathbf{F}_e + \mathbf{F}_m = 28\hat{\mathbf{a}}_x + 26\hat{\mathbf{a}}_y + 108\hat{\mathbf{a}}_z \quad [\text{pN}].$$

The Lorentz force law does more than just predict the forces on charges. It also serves to define the electric and magnetic fields themselves. For instance, solving Equation (3.38) for **E**, we find that

$$\mathbf{E} = \frac{\mathbf{F}_e}{Q} \quad [\text{N/C or V/m}], \tag{3.40}$$

which means that we can test, or measure, the value of **E** at any point simply by measuring the force acting on a stationary test charge. In order for the test charge not to disturb the currents and charges responsible for **E**, Q must be as small as possible. Thus, we *define* the E-field at a point to be

$$\mathbf{E} \equiv \lim_{Q\to 0} \frac{\mathbf{F}_e}{Q} \quad [\text{N/C or V/m}], \tag{3.41}$$

where $\mathbf{F}_e$ is the total force acting on Q when it is at rest.

In a similar way we can use Equation (3.39) to define the B-field at a point. If we take the cross product of both sides of that equation with **u**, we obtain

$$\mathbf{F}_m \times \mathbf{u} = Q(\mathbf{u} \times \mathbf{B}) \times \mathbf{u} = Q\,\mathbf{B}|\mathbf{u}|^2 - Q\,\mathbf{u}(\mathbf{u} \cdot \mathbf{B}), \tag{3.42}$$

where we have used Equation (B.2) in Appendix B to expand the triple cross product. To solve this expression for **B**, we need to choose the "test velocity" **u** such that $\mathbf{u} \cdot \mathbf{B} = 0$. This choice of the direction of **u** also maximizes $\mathbf{F}_m$, since $\mathbf{u} \times \mathbf{B}$ is maximized when **u** and **B** are perpendicular to each other. Solving Equation (3.42) for **B** under this condition yields

$$\mathbf{B} = \frac{1}{Q}\left[\frac{\mathbf{F}_m \times \mathbf{u}}{|\mathbf{u}|^2}\right]_{\max}. \tag{3.43}$$

Finally, since the purpose of the test charge Q is to measure the field produced by other sources without affecting them, we take the limit of Equation (3.43) as $Q \to 0$, yielding

$$\mathbf{B} \equiv \lim_{Q \to 0} \left\{ \frac{1}{Q} \left[\frac{\mathbf{F}_m \times \mathbf{u}}{|\mathbf{u}|^2} \right]_{\max} \right\} \left[\frac{\mathrm{N} \cdot \mathrm{s}}{\mathrm{C} \cdot \mathrm{m}} \text{ or T} \right]. \tag{3.44}$$

Example 3-6

An unknown source distribution exerts a force $\mathbf{F}_1 = -8\hat{\mathbf{a}}_x + 3\hat{\mathbf{a}}_y$ [pN] on a $Q = 2$ [pC] test charge that is at rest at a point P. However, when the charge moves with a speed of 2 [m/s] through P, the change in the force is greatest when the direction of the velocity $\mathbf{u}$ is $\hat{\mathbf{a}}_{\max} = (1/\sqrt{2})(\hat{\mathbf{a}}_x - \hat{\mathbf{a}}_y)$. If the force acting on Q in this case is $\mathbf{F}_2 = 2\hat{\mathbf{a}}_x - \hat{\mathbf{a}}_y + 4\hat{\mathbf{a}}_z$ [pN], find both $\mathbf{E}$ and $\mathbf{B}$ at P.

Solution:

Since $\mathbf{u} = 0$ when $\mathbf{F}_1$ was measured, $\mathbf{F}_e = \mathbf{F}_1$. Using Equation (3.41), we have

$$\mathbf{E} = \frac{\mathbf{F}_e}{Q} = \frac{(-8\hat{\mathbf{a}}_x + 3\hat{\mathbf{a}}_y) \times 10^{-12}}{2 \times 10^{-12}} = -4\hat{\mathbf{a}}_x + 1.5\hat{\mathbf{a}}_y \text{ [V/m]}.$$

To find $\mathbf{B}$, we first note that the magnetic force $\mathbf{F}_m$ is the difference between the force when Q is moving and its rest force:

$$\mathbf{F}_m = \mathbf{F}_2 - \mathbf{F}_1 = 10\hat{\mathbf{a}}_x - 4\hat{\mathbf{a}}_y + 4\hat{\mathbf{a}}_z \text{ [pN]}.$$

Using Equation (3.44), we obtain

$$\mathbf{B} = \frac{1}{Q}\frac{\mathbf{F}_m \times \mathbf{u}}{|\mathbf{u}|^2} = \frac{10^{-12}(10\hat{\mathbf{a}}_x - 4\hat{\mathbf{a}}_y + 4\hat{\mathbf{a}}_z) \times \frac{2}{\sqrt{2}}(\hat{\mathbf{a}}_x - \hat{\mathbf{a}}_y)}{2 \times 10^{-12} \times 2^2}$$

$$= \frac{1}{\sqrt{2}}(\hat{\mathbf{a}}_x + \hat{\mathbf{a}}_y - 1.5\hat{\mathbf{a}}_z) \quad \text{[T]}.$$

When charges are moving, it is often more convenient to express the magnetic component of the force exerted on them in terms of a ***test current***, rather than a test charge. Figure 3-16 shows a thin wire that carries a test current I. We can find the force $\mathbf{dF}_m$ that acts on the differential segment $I\,\mathbf{d\ell}$ by noting that

$$I\mathbf{d\ell} = \frac{dQ}{dt}\mathbf{d\ell} = \frac{dQ}{dt}\mathbf{u}dt = dQ\mathbf{u}. \tag{3.45}$$

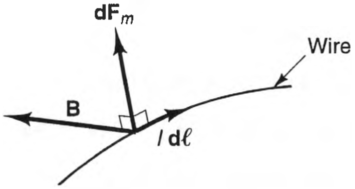


Figure 3-16 The magnetic field acting on a current element.

where dQ is the charge contained within the length $d\ell$. Substituting Equation (3.45) into Equation (3.39), we find that the magnetic force is given

$$\mathbf{dF}_m = I\, \mathbf{d\ell} \times \mathbf{B}. \tag{3.46}$$

Similarly, the magnetic forces acting on differential surface and volume elements are

$$\mathbf{dF}_m = \mathbf{J}_s \times \mathbf{B}ds \tag{3.47}$$

and

$$\mathbf{dF}_m = \mathbf{J} \times \mathbf{B}dv, \tag{3.48}$$

respectively.

The total force $\mathbf{dF}$ acting on a differential volume dv of a moving charge distribution equals the sum of the electric and magnetic forces. Using Equations (3.40) and (3.48), we obtain

$$\mathbf{dF} = \mathbf{dF}_e + \mathbf{dF}_m = dQ\,\mathbf{E} + \mathbf{J} \times \mathbf{B}dv$$

If we replace dQ with $\rho_v dv$, this expression becomes

$$\mathbf{dF} = (\rho_v \mathbf{E} + \mathbf{J} \times \mathbf{B})dv. \tag{3.49}$$

For surface and line charge distributions, Equation (3.49) becomes

$$\mathbf{dF} = (\rho_s \mathbf{E} + \mathbf{J}_s \times \mathbf{B})ds \tag{3.50}$$

and

$$\mathbf{dF} = \rho_\ell \mathbf{E}d\ell + I\,\mathbf{d\ell} \times \mathbf{B}, \tag{3.51}$$

respectively.

Example 3-7

A large sheet of charge lies in the $z = 0$ plane. The charge density $\rho_s = 4$ [C/m^2] is uniform and moves with a constant velocity, $\mathbf{u} = 2\hat{\mathbf{a}}_x$ [m/s]. If $\mathbf{E} = 2\hat{\mathbf{a}}_x - 3\hat{\mathbf{a}}_z$ [V/m] and $\mathbf{B} = 3\hat{\mathbf{a}}_y$ [T] at all points in the $z = 0$ plane, find the force per unit area acting on the sheet.

Solution:

In order to use Equation (3.50), we must first find the surface current density on the sheet. Using Equation (3.19), we have

$$\mathbf{J}_s = \rho_s\,\mathbf{u} = 8\hat{\mathbf{a}}_x \quad [\text{A/m}],$$

and

$$\mathbf{J}_s \times \mathbf{B} = 8\hat{\mathbf{a}}_x \times 3\hat{\mathbf{a}}_y = 24\hat{\mathbf{a}}_z \quad [\text{N/m}^2].$$

Substituting this into Equation (3.50), we find that

$$\mathbf{dF}/ds = \rho_s\,\mathbf{E} + \mathbf{J}_s \times \mathbf{B} = 4(2\hat{\mathbf{a}}_x - 3\hat{\mathbf{a}}_z) + 24\hat{\mathbf{a}}_z = 8\hat{\mathbf{a}}_x + 12\hat{\mathbf{a}}_z \quad [\text{N/m}^2],$$

which is the force per unit area acting on the sheet.

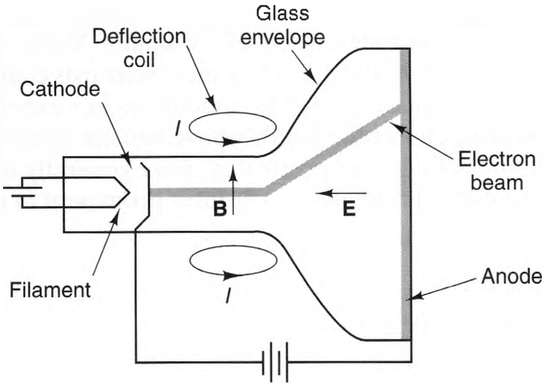


Figure 3-17 Schematic diagram of a cathode-ray tube.

A common device that demonstrates the Lorentz force law is the cathode-ray tube (CRT). A simplified illustration of a CRT is shown in Figure 3-17. Here, electrons are emitted by the cathode, which is heated by the filament. These electrons are then accelerated towards the anode by an E-field that is directed along the axis of the tube. This field is the result of a large voltage (roughly 15 kilovolts for a monochrome CRT) between the cathode and the anode. The anode is coated with luminescent phosphors that glow when struck by high-velocity electrons.

The vertical and horizontal scanning of the electron beam in a CRT is accomplished by using current-carrying coils. These coils are housed in the ***deflection yoke*** outside the tube and create B-fields that are perpendicular to the electron beam. Two pairs of coils are used to allow independent horizontal and vertical deflections, which are controlled by the magnitude and direction of the coil currents.

3-7 Maxwell's Equations

Until now, our discussion of electromagnetics has centered mainly on the forces that charges and currents exert upon each other. This is a fitting place to start, since these forces are ultimately responsible for all the things that we use electricity and electromagnetism for. It may come as a surprise to the reader, then, that the majority of this text is not devoted to discussing forces, but rather is given over to the relationship between electromagnetic sources (i.e., currents and charges) and the electric and magnetic fields they produce. There are two reasons for this. The first is that we often are not aware of these electromagnetic forces. For instance, even though the memory function of an integrated circuit (IC) chip is accomplished by applying forces to small packets of charge and moving them between various locations within the chip, we are not aware of these forces, since they are inside the chip. The second reason is that the performance of electrical devices is determined by how the electric and magnetic fields distribute themselves. In the case of an IC memory chip, this means that we can determine its operating characteristics once we know how **E** and **B** distribute themselves for all the possible input configurations.

The task of determining the definitive set of equations that describe the relationship between electromagnetic sources and fields was a laborious one that involved many of the most distinguished scientists of the 18th and 19th centuries. The list of names that are prominent in this history includes those of Coulomb, Volta, Poisson, Frankin, Oersted, Ampère, Lorentz, and Faraday. Each of these scientists contributed to our basic understanding of electromagnetics, either by conducting key experiments or by postulating theories of how electromagnetic fields behave and are related to their sources. By the mid-1800s, a number of laws (or postulates) were generally accepted that each identified certain aspects of electromagnets. The most prominent of these were:

1. The Lorentz force law
2. Coulomb's law of force
3. Ampère's law of force
4. The law of charge conservation
5. Faraday's law of induction

We have already discussed the first four laws. The fifth law, Faraday's law of induction, was proposed in 1831 in order to explain why a time-varying current in one circuit produces voltages in other circuits. This law explains the operation of inductors and transformers.

As useful as the preceding five laws are, they are relatively unrelated, and some are applicable only under restricted circumstances. For instance, we have already seen that Coulomb's law of force is applicable only for charges at rest. Similarly, Ampère's law of force is applicable only for steady currents. As a result, the five laws did not constitute an all-inclusive system of equations that described the behavior of electromagnetic sources and fields under all circumstances.

In 1873, James Clerk Maxwell published his now famous *A Treatise on Electricity and Magnetism* that put forth a theory of electromagnetics that accounted for all that was then known about electromagnetism, plus a bold postulate that was all his own. This theory could be summarized with four equations that relate the relationship between the electric and magnetic field vectors, **E** and **B**, respectively, with the two fundamental sources of these fields—current density **J** and charge density ρ_v:

$$\left.\begin{aligned}
\nabla \times \mathbf{B} &= \mu_o \mathbf{J} + \mu_o \epsilon_o \frac{\partial \mathbf{E}}{\partial t} && (3.52)\\
\nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} && (3.53)\\
\nabla \cdot \mathbf{E} &= \frac{\rho_v}{\epsilon_o} && (3.54)\\
\nabla \cdot \mathbf{B} &= 0 && (3.55)
\end{aligned}\right\} \text{(Maxwell's equations in free space),}$$

where "$\nabla \times$ " and "$\nabla \cdot$" are the curl and divergence operators, respectively. Even though these equations embody many contributions from all the investigators up till that time, these equations have been named ***Maxwell's equations***, because it was he who cast them into a complete, unified theory of electromagnetics. The equations are consistent with the five experimentally based postulates that were listed earlier in this section. More importantly, the equations describe *every* electromagnetic experiment, device, or result that has been encountered since Maxwell first proposed them.[7]

Even though Maxwell's equations are now universally accepted as the most fundamental and general postulates of electromagnetics, their acceptance did not come as quickly as one might think. This was due in part to the limited range of experimental evidence that was available at that time. Another reason is that Maxwell justified one of the key terms in his equations not on the basis of experimental evidence, but rather on his belief that electromagnetism and light are manifestations of the same physical phenomenon. This proposal mystified many of the scientists of that day, because they could see no link between the behavior of low-frequency electromagnetic effects and the properties of light.

The controversy ended in the late 1880s when Heinrich Hertz (1857–1894) published the results of his now famous series of experiments.[8] These showed clearly that electromagnetic fields do indeed have the same characteristics (scaled for the difference in frequency) as light waves. Hertz created electromagnetic waves by exciting short-duration, broadband currents on wires. The currents were created by charging a Leyden jar (an early type of capacitor) to a high voltage and touching its terminals to both transmitting wires, as shown in Figure 3-18.

This voltage caused a spark to jump across the gap between the transmitting wires, resulting in a short burst of current on the wires. The fields launched by the transmitting wires created a voltage across the spark gap of a receiving loop (also shown in the figure), creating a visible spark. Using this simple transmit–receive system, Hertz was able to demonstrate that electromagnetic fields exhibit lightlike properties such as propagation, reflection, polarization, resonance, and focusing, at frequencies we now call microwave frequencies.

The connections between Maxwell's equations and the five experimental postulates that we summarized earlier in this section are not immediately obvious, except for one—the law of charge conservation. We can see this connection by first taking the divergence of both sides of Equation (3.52):

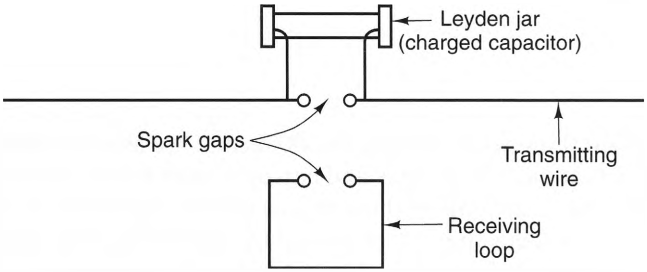


Figure 3-18 Schematic diagram of Hertz's experiment for generating and detecting electromagnetic waves.

[7] Apart from experiments involving subatomic particles, where some quantum mechanical adjustments must be made.

[8] For an excellent review of Hertz's experiments, see John Kraus, "Heinrich Hertz—Theorist and Experimentor," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 36, No. 5, May 1988.

$$\nabla \cdot \nabla \times \mathbf{B} = \mu_o \nabla \cdot \mathbf{J} + \mu_o \epsilon_o \frac{\partial}{\partial t} \nabla \cdot \mathbf{E},$$

where we have interchanged the order of the time and spatial differentiation in the last term. According to Equation (B.8) in Appendix B, however, the left side of this equation is identically zero, which leaves us with

$$\nabla \cdot \mathbf{J} = -\epsilon_o \frac{\partial}{\partial t} \nabla \cdot \mathbf{E}.$$

But from Equation (3.54), we also have $\nabla \cdot \mathbf{E} = \rho_v/\epsilon_o$. Substituting, we obtain

$$\nabla \cdot \mathbf{J} = -\frac{\partial \rho_v}{\partial t},$$

which is the law of charge conservation (see Equation (3.25)). Ironically, Maxwell appears to have been unaware of the connection between his equations and the law of charge conservation. Had he known of this, his equations might have been universally accepted within his lifetime.

To ease our transition into describing electromagnetic effects using Maxwell's equations, we will start by considering time-invariant (i.e., dc) fields. One reason for doing this is that the time-invariant case is much simpler, since time is not a factor. Another reason is that electric and magnetic fields are independent of each other for that case, which will allow us to discuss the characteristics of electric and magnetic fields separately. This is helpful, since electric and magnetic fields behave differently around their sources and interact differently with materials. Also, many of the characteristics of these time-invariant fields are maintained when the sources are time varying.

Following our discussion of time-invariant fields (Chapters 4–9), we will move on to the general time-varying case, where electric and magnetic fields always appear together. This will allow us to discuss an important effect that can occur only when sources are time-varying: propagation. We will discuss three types of applications that utilize electromagnetic propagation: transmission lines, free-space plane waves, and waveguides, all of which are useful for transporting power and information over large and small distances. Also, the final chapter will introduce the closely related subject of radiation and antennas.

3-8 Summation

In this chapter we identified the sources, forces, and fields that are associated with electromagnetics. We found that electric charges and currents exert forces on each other. According to the Lorentz force law, all of these forces can be explained by the existence of electric and magnetic fields that are produced by electromagnetic sources.

Maxwell's equations were also introduced, which relate the electric and magnetic fields to the electromagnetic sources that produce them. These equations are now considered to be the fundamental postulates of electromagnetics. In one way or another, Maxwell's equations (or equations derived from them) are the foundational design tools for electrical engineers.

Problems

3-1 The charge density throughout a region is given by $\rho_v = 10e^{-3r}$ [μC/m^3], where r is measured in meters. Find the total charge Q contained in a sphere centered about the origin that has radius 2 meters.

3-2 The surface charge density throughout the $z = 0$ plane is $\rho_s = 1/\rho$ [C/m^2], where ρ is specified in meters. Find the total charge Q that is contained in a circular disk of radius 1.0 [m] that is centered about the origin in the same plane.

3-3 The current density in a region is $\mathbf{J} = xy\,\hat{\mathbf{a}}_z$ [A/m^2], where x and y are specified in meters. Find the current I flowing towards increasing values of z through the surface $-1 < x < 1$, $-1 < y < 1$, $z = 0$.

3-4 A surface current $\mathbf{J}_s = x\,\hat{\mathbf{a}}_x + xy\,\hat{\mathbf{a}}_y$ [A/m] flows on the $z = 0$ plane, where x and y are specified in meters. Find the magnitude of the current I that flows towards increasing values of x, past the line that extends from $P(0, 0, 0)$ to $P(1, 1, 0)$.

3-5 Find the current I that flows out of a sphere of radius 2 [m] that is centered at the origin if the current density throughout the region is $\mathbf{J} = re^{-r}\,\hat{\mathbf{a}}_r$ [μA/m^2], where r is specified in meters.

3-6 Find an expression for the charge density at each point in space at an arbitrary time t if the current density is $\mathbf{J} = \rho\,\hat{\mathbf{a}}_\rho$ [mA/m^2], where ρ is measured in meters. Assume that $\rho_v = 0$ at all points at $t = 0$.

3-7 Find the force exerted on a point charge Q_1 by the point charges Q_2 and Q_3. Assume that $Q_1 = 2.5$ [nC], $Q_2 = -12.0$ [nC], and $Q_3 = -1.7$ [nC] are located at $(0, -1, 2)$ [m], $(1.5, 2, -1)$ [m], and $(-1, -1.5, 2)$ [m], respectively.

3-8 Figure P3-8 shows two short, parallel current segments, each of length 0.001[m] and carrying currents $I_1 = 2$ [A] and $I_2 = 4$ [A], respectively. If $R = 1$ [cm] and $\theta = 30°$, calculate the force acting on I_1.

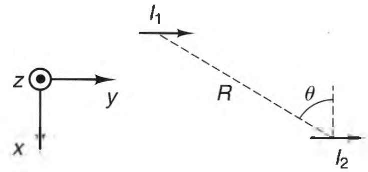


Figure P3-8

3-9 A 1 [μC] test charge has velocity $\mathbf{u} = 3\hat{\mathbf{a}}_x - 2.5\hat{\mathbf{a}}_y + 1.5\hat{\mathbf{a}}_z$ [m/s] at the point $P(1, -2, 1)$ [m] in the presence of the fields $\mathbf{E} = x\,\hat{\mathbf{a}}_x + xy\,\hat{\mathbf{a}}_z$ [V/m] and $\mathbf{B} = y\,\hat{\mathbf{a}}_x + x\,\hat{\mathbf{a}}_y$ [T], where x and y are specified in meters. Find the total force $\mathbf{F}$ acting on the test charge at $P(1, -2, 1)$ [m].

3-10 A point charge of value +3 [μC] moves with velocity $\mathbf{u} = 2.5\hat{\mathbf{a}}_x - 3.2\hat{\mathbf{a}}_y + 4.0\hat{\mathbf{a}}_z$ [m/s] in the presence of an electric and magnetic field. Find $\mathbf{E}$ if the force acting on the electron is $\mathbf{F} = -320\hat{\mathbf{a}}_x + 150\hat{\mathbf{a}}_y$ [μN] and it is known that $\mathbf{B} = 200\hat{\mathbf{a}}_y$ [T].

3-11 The charge density in a region is $\rho_v = e^{-r}e^{-t}$ [C/m^3], where r and t are measured in meters and seconds, respectively. Find the current density $\mathbf{J}$ throughout this region. (*Hint:* Assume that $\mathbf{J}$ is spherically symmetric.)

3-12 Calculate the final velocity of the electron beam in a CRT as it impinges upon the screen if the voltage between the anode and the cathode is 24 [kV] and they are spaced by 0.4 [m]. Assume that the electron velocity is zero at the cathode, the E-field is directed from the anode to the cathode with a magnitude of $E = 24/0.4$ [kV/m], and there is no magnetic field present.

3-13 A point charge with charge Q and mass m is placed in a region with a uniform B-field, $\mathbf{B} = B_o \hat{\mathbf{a}}_z$. If the charge's position and velocity at $t = 0$ are $\mathbf{u}(0) = u_o \hat{\mathbf{a}}_x$ [m/s], $x(0) = x_o$ and $y(0) = y_o$, find $x(t)$ and $y(t)$ for all $t > 0$. What kind of path is this?

4

Electrostatic Fields in Free Space

4-1 Introduction

We have seen in the previous chapter that current and charge distributions generate electric and magnetic fields. In general, time-varying charge distributions cause both electric and magnetic fields. The same is true for time-varying current distributions. However, static charge distributions generate only electric fields, called ***electrostatic fields***, and steady current distributions cause only magnetic fields, called ***magnetostatic fields***. These two special cases are much simpler than the general time-varying case, so we will start our discussion of electric and magnetic fields with them. In the next three chapters, we will identify the various characteristics of electrostatic fields. Later, we will do the same for magnetostatic fields. Not only will this ease our eventual transition into time-varying fields, but we will also see along the way that there are a host of practical applications of electrostatic and magnetostatic fields.

In this chapter, we will investigate the electrostatic fields generated by static charge distributions in free space (i.e., in a vacuum). We will start our discussion by specializing Maxwell's equations for the electrostatic case and then deriving Coulomb's law, which explicitly specifies the E-field generated by any known, static charge distribution. Using

Coulomb's law and Maxwell's equations, we will determine the E-fields generated by several classes of charge distributions.

We will also introduce a new field quantity, called the electrostatic potential. Unlike the E-field, which is a vector, the electrostatic potential is a scalar, which makes it much easier to deal with. This potential is directly related to the potential differences and voltages used in circuit analysis. We will show that using the electric potential often simplifies E-field calculations.

4-2 Maxwell's Equations for Electrostatics in Free Space

We stated in the previous chapter that Maxwell's equations are considered to be the fundamental postulates of all electromagnetic phenomena. In free space, these equations read

$$\nabla \cdot \mathbf{B} = 0 \qquad \nabla \times \mathbf{B} = \mu_o \mathbf{J} + \mu_o \epsilon_o \frac{\partial \mathbf{E}}{\partial t}$$

$$\nabla \cdot \mathbf{E} = \frac{\rho_v}{\epsilon_o} \qquad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}.$$

The preceding equations are written in what is called ***differential*** (or ***point***) ***form***, since they contain derivatives that are evaluated at individual points. For the time-invariant case, all derivatives with respect to time vanish. Thus, for static charge and steady current distributions, Maxwell's equations become

$$\nabla \cdot \mathbf{B} = 0 \qquad (4.1) \qquad\qquad \nabla \times \mathbf{B} = \mu_o \mathbf{J} \qquad (4.2)$$

$$\nabla \cdot \mathbf{E} = \frac{\rho_v}{\epsilon_o} \qquad (4.3) \qquad\qquad \nabla \times \mathbf{E} = 0. \qquad (4.4)$$

With the time-derivative terms gone, **E** and **B** now appear in separate equations. Equations (4.1) and (4.2) show that magnetostatic fields depend only upon the current density **J**, whereas Equations (4.3) and (4.4) show that electrostatic fields depend only upon the charge density ρ_v. As a result, we can discuss the characteristics of electrostatic fields without considering whether magnetostatic fields are also present.

Equations (4.3) and (4.4) are called the point form of ***Maxwell's equations for electrostatics in free space***. Since they specify both the divergence and curl of **E** at every point, we can conclude from the Helmholtz's theorem (see Section 2-5-6) that no additional equations are needed to uniquely specify **E** for any given charge distribution ρ_v. We can also derive integral representations of these equations. To do this, let us first take the dot product of both sides of Equation (4.4) with a differential surface vector **ds** and integrate over an arbitrary *open* surface *S*, yielding

$$\int_S \nabla \times \mathbf{E} \cdot \mathbf{ds} = 0.$$

Using Stokes's theorem, we can express the surface integral as a line integral over the *closed* contour C that bounds S, resulting in

$$\oint_C \mathbf{E} \cdot d\boldsymbol{\ell} = 0. \tag{4.5}$$

We can obtain an integral representation of Equation (4.3) by first multiplying both sides by the differential volume element dv and integrating over some volume V, yielding

$$\int_V \nabla \cdot \mathbf{E}\, dv = \frac{1}{\epsilon_o} \int_V \rho_v dv.$$

The integral on the right side of this equation is simply the total charge Q contained in V. Also, we can use the divergence theorem to write the integral on the left as a surface integral over the closed surface S that surrounds V, giving

$$\oint_S \mathbf{E} \cdot d\mathbf{s} = \frac{Q}{\epsilon_o}. \tag{4.6}$$

Maxwell's equations for electrostatic fields in free space are summarized as follows in both point and integral forms:

MAXWELL'S EQUATIONS FOR ELECTROSTATICS IN FREE SPACE

Point Form		Integral Form	
$\nabla \cdot \mathbf{E} = \frac{\rho_v}{\epsilon_o}$	(4.7)	$\oint_S \mathbf{E} \cdot d\mathbf{s} = \frac{Q}{\epsilon_o}$	(4.8)
$\nabla \times \mathbf{E} = 0$	(4.9)	$\oint_C \mathbf{E} \cdot d\boldsymbol{\ell} = 0$	(4.10)

Regardless of the form they are written in, these equations completely define the behavior of electrostatic fields in free space.

Equations (4.7) and (4.8) define the relationship between charge distributions and the electrostatic E-fields they generate. These equations are representations of ***Gauss's law***, named in honor of Karl Friedrich Gauss (1777–1855). Gauss's law states that the total outward flux of the electric field intensity over any closed surface in free space equals the total charge enclosed by the surface, divided by the permittivity of free space. In this chapter we will use Gauss's law to find the electric field intensities of several classes of charge distributions.

Equations (4.9) and (4.10) state another important property of electrostatic fields, namely that the electric field intensity **E** is a conservative vector field. As we saw in Chapter 2, a conservative vector field has that property that the scalar line integral around any closed path is always zero. Conservative vectors are always irrotational (i.e., have zero curl), since a vector with zero rotation at each point cannot have a net circulation around a closed path. Equations (4.9) and (4.10) are also the basis of another important law of electrostatics: ***Kirchoff's voltage law***. This law states that the sum of all the voltage drops around any closed circuit is zero whenever electrostatic fields are present.

4-3 Coulomb's Law

Maxwell's equations for electrostatics contain all the information necessary to characterize the fields generated by static charge distributions, but they are usually not the most convenient starting point for actual calculations. This is because Maxwell's equations describe **E** implicitly, either in terms of differential equations (Equations (4.7) and (4.9)) or as integral equations (Equations (4.8) and (4.10)). However, we can use Maxwell's equations to derive an explicit expression for **E**, called ***Coulomb's law***, that allows for direct calculations of **E** when the charge distribution ρ_v is known.

We can derive Coulomb's law directly from Maxwell's equations for electrostatics, but the proof is tedious. Rather, we will start by using Coulomb's law of force (discussed in Chapter 3) to find an expression for the E-field generated by a single point charge. Once we have determined that this expression satisfies Maxwell's equations for electrostatics, we will then generalize the formula to predict the E-field generated by an arbitrary charge distribution. Figure 4-1 shows a point charge Q located at the origin. According to Coulomb's law of force (Equation (3.29)), this charge will induce a force **F** on a positive test charge q that is given by

$$\mathbf{F} = \frac{1}{4\pi\epsilon_0}\frac{qQ}{r^2}\hat{\mathbf{a}}_r \qquad [\mathrm{N}],$$

where r is the distance from Q to q and $\hat{\mathbf{a}}_r$ is parallel to the position vector **r** that represents the position of q. Remembering that **E** is defined as the force per unit test charge, we can write

$$\mathbf{E} = \frac{\mathbf{F}}{q} = \frac{1}{4\pi\epsilon_0}\frac{Q}{r^2}\hat{\mathbf{a}}_r \qquad [\mathrm{N/C\ or\ V/m}], \tag{4.11}$$

which states that the charge Q at the origin generates an E-field that is directed radially outward and decays proportional to $1/r^2$. Even though **E** is defined in terms of the electric force on a test charge, it is usually more convenient to specify it in units of volts per meter, as we shall see later in the chapter.

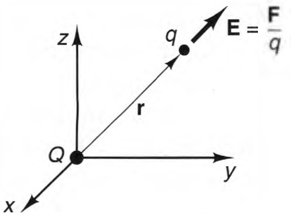


Figure 4-1 Geometry for determining the E-field generated by a point charge at the origin.

To see if Equation (4.11) is consistent with Maxwell's equations, let us first substitute Equation (4.11) into Equation (4.9). Since **E** has only one component, E_r, which is a function only of r, we find that

$$\nabla \times \mathbf{E} = \frac{Q}{4\pi\epsilon_o} \nabla \times \left[\frac{\hat{\mathbf{a}}_r}{r^2}\right] = 0.$$

Next, if we substitute Equation (4.11) into Equation (4.8) and integrate around a sphere that is centered at the origin, we obtain $\mathbf{E} \cdot \mathbf{ds} = E_r r^2 \sin\theta \, d\theta d\phi$ and

$$\oint_S \mathbf{E} \cdot \mathbf{ds} = \frac{Q}{4\pi\epsilon_o} \int_0^{2\pi} \int_0^{\pi} \frac{1}{r^2} r^2 \sin\theta \, d\theta \, d\phi = \frac{Q}{\epsilon_o},$$

which is valid for all values of r. Hence, Equation (4.11) is consistent with Maxwell's equations of electrostatics in free space.

We can generalize Equation (4.11) by first recognizing that the E-field generated by a point charge is always directed outward from the charge, regardless of whether it is located at the origin or not. For the point charge Q in Figure 4-2 that is located away from the origin, we can represent the E-field that it generates at a field point P by

$$\mathbf{E} = \frac{1}{4\pi\epsilon_o} \frac{Q}{R^2} \hat{\mathbf{a}}_R \qquad [\text{V/m}], \tag{4.12}$$

where R is the distance from the observer to the charge and $\hat{\mathbf{a}}_R$ points outward from the source point (Q) towards the field point (P). If the field and source points are represented by the position vectors **r** and **r**′, respectively, we can write

$$R = |\mathbf{r} - \mathbf{r}'|$$

and

$$\hat{\mathbf{a}}_R = \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|}.$$

Using these, we can write **E** in the form

$$\mathbf{E} = \frac{Q}{4\pi\epsilon_o} \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \qquad [\text{V/m}]. \tag{4.13}$$

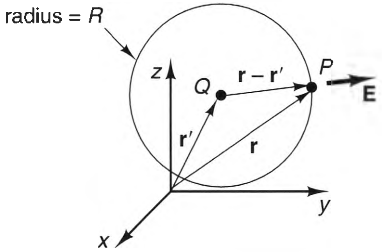


Figure 4-2 Geometry for determining the E-field generated by a point charge located at an arbitrary point.

This expression may not appear as easy to use as Equation (4.12), but just the opposite is true, since the expression makes it more clear what information must be in hand in order to actually calculate **E**.

Example 4-1

A point charge $Q = 1$ [nC] is located at $(1, -1, 0)$ [m]. Find the electric field intensity at the point $(0, 0, 3)$ [m].

Solution:

The positions of Q and the observation point are represented by

$$\mathbf{r}' = \hat{\mathbf{a}}_x - \hat{\mathbf{a}}_y$$
$$\mathbf{r} = 3\hat{\mathbf{a}}_z,$$

respectively. Thus,

$$\mathbf{r} - \mathbf{r}' = 3\hat{\mathbf{a}}_z - \hat{\mathbf{a}}_x + \hat{\mathbf{a}}_y$$
$$|\mathbf{r} - \mathbf{r}'| = \sqrt{3^2 + 1^2 + 1^2} = \sqrt{11}.$$

Substituting these values into Equation (4.13), we obtain

$$\mathbf{E} = \frac{10^{-9}}{4\pi\epsilon_0}\frac{(3\hat{\mathbf{a}}_z - \hat{\mathbf{a}}_x + \hat{\mathbf{a}}_y)}{11^{3/2}}$$
$$= -0.246\hat{\mathbf{a}}_x + 0.246\hat{\mathbf{a}}_y + 0.739\hat{\mathbf{a}}_z \qquad [\text{V/m}].$$

Figure 4-3 shows a volume V that contains a volumetric distribution of charge with charge density ρ_v. The charge that is contained within the differential volume element dv' is $dQ = \rho_v dv'$, where ρ_v is the charge density within the differential volume. The field **dE** that this charge generates at the field point P is[1]

$$\mathbf{dE} = \frac{\rho_v dv'}{4\pi\epsilon_0}\frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \qquad [\text{V/m}],$$

where $\mathbf{r}'$ is the position vector that represents the position of the differential volume dv' and $\mathbf{r}$ is the position vector of P. Since **dE** is proportional to the charge density ρ_v

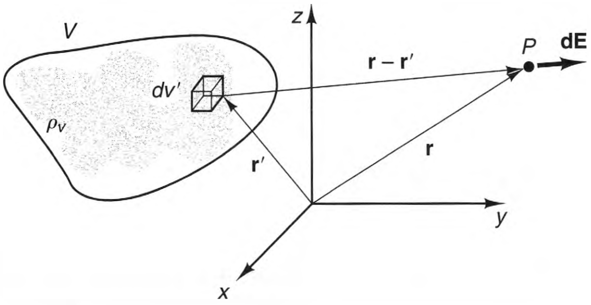


Figure 4-3 Geometry for determining the E-field of a volumetric charge distribution.

[1] Here, we have drawn **r** outside the volume V for the sake of visual clarity, but **r** can also lie inside V.

and ϵ_o is a constant, we can use the superposition principle, which states that the total response in a linear medium due to a number of sources equals the the sum of the responses due to each source individually. Integrating the contributions from each differential charge element $\rho_v dv'$, we obtain

$$\mathbf{E} = \frac{1}{4\pi\epsilon_o}\int_V \rho_v \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3}\, dv' \quad \text{[V/m]} \quad \text{(Coulomb's law for volume-charge distributions),} \tag{4.14}$$

where ρ_v is a function of the primed coordinates. During the integration process, the dummy position vector $\mathbf{r}'$ sweeps through all the points within V where the charge density ρ_v is nonzero. Equation (4.14) is the volumetric form of ***Coulomb's law***.

Equation (4.14) is not convenient when surface charge distributions are present, since the volume charge density is infinite on these surfaces. For such situations, we can replace the term $\rho_v dv'$ with $\rho_s ds'$ to obtain

$$\mathbf{E} = \frac{1}{4\pi\epsilon_o}\int_S \rho_s \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3}\, ds' \quad \text{[V/m]} \quad \text{(Coulomb's law for surface-charge distributions).} \tag{4.15}$$

Similarly, for a line-charge distribution, we can replace $\rho_v dv'$ with $\rho_\ell d\ell'$, yielding

$$\mathbf{E} = \frac{1}{4\pi\epsilon_o}\int_C \rho_\ell \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3}\, d\ell' \quad \text{[V/m]} \quad \text{(Coulomb's law for line-charge distributions).} \tag{4.16}$$

The three expressions of Coulomb's law given by Equations (4.14) to (4.16) are typical of many integral expressions we will encounter throughout this text in that they contain two position vectors, $\mathbf{r}$ and $\mathbf{r}'$. Whenever these vectors appear in a formula, their interpretation is *always* the same. By convention, $\mathbf{r}$ always represents the position (x, y, z) of the field point, which can be thought of as the position of an observer who is measuring the field. Conversely, vector $\mathbf{r}'$ is a dummy position vector that sweeps during the integration through all points (x', y', z') where the integrand is nonzero. Because the primed coordinate variables are dummy variables of integration, the final result after the integration ($\mathbf{E}$, in this case) is a function only of the unprimed, field-point position variables, (x, y, z).

4-4 E-Field Calculations Using Coulomb's Law

Coulomb's law can be used to calculate the E-field generated by any charge distribution, provided that the charge density is known at all points. The resulting integrals, however, are usually complicated and, in most cases, must be evaluated numerically. Fortunately, there are a several special cases where the integrals can be evaluated analytically. The point charge is one such example. In the discussion that follows, we will

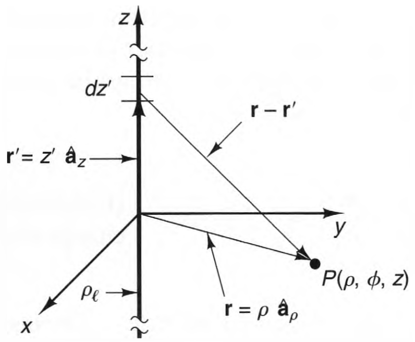


Figure 4-4 Geometry for determining the E-field of an infinite line of charge.

derive solutions from Coulomb's law for three charge distributions: an infinite line of charge, a circular charged disk, and an infinite sheet of charge.

4-4-1 THE UNIFORM, INFINITE LINE CHARGE

Figure 4-4 shows an infinite line of charge with a uniform line-charge density ρ_ℓ. This charge distribution is often called an infinite line source and is a good approximation of the charge distributions found in many practical situations. Examples include the electron beam in a cathode-ray tube and the charges on long, thin wires.

Since the line is infinitely long, the E-field at a point $P(\rho, \phi, z)$ will be the same for all values of z. Thus, for simplicity, let us choose our field point to lie in the $z = 0$ plane. Using the cylindrical coordinate system, we find that the position vectors representing the field and source points are

$$\mathbf{r} = \rho\hat{\mathbf{a}}_\rho$$

$$\mathbf{r}' = z'\hat{\mathbf{a}}_{z'} = z'\hat{\mathbf{a}}_z,$$

respectively. Hence,

$$\mathbf{r} - \mathbf{r}' = \rho\hat{\mathbf{a}}_\rho - z'\hat{\mathbf{a}}_z$$

and

$$|\mathbf{r} - \mathbf{r}'|^3 = [\rho^2 + z'^2]^{3/2}.$$

Also,

$$d\ell' = dz'.$$

Substituting these into Equation (4.16), we obtain

$$\mathbf{E} = \frac{\rho_\ell}{4\pi\epsilon_0}\int_{-\infty}^{\infty} \frac{(\rho\hat{\mathbf{a}}_\rho - z'\hat{\mathbf{a}}_z)}{[\rho^2 + z'^2]^{3/2}}\,dz'$$
$$= \frac{\rho_\ell\rho\hat{\mathbf{a}}_\rho}{4\pi\epsilon_0}\int_{-\infty}^{\infty} \frac{1}{[\rho^2 + z'^2]^{3/2}}\,dz' - \frac{\rho_\ell\hat{\mathbf{a}}_z}{4\pi\epsilon_0}\int_{-\infty}^{\infty} \frac{z'}{[\rho^2 + z'^2]^{3/2}}\,dz'.$$

Here, we are able to move both unit vectors and the charge density ρ_ℓ outside the integrals, since they are not functions of the integration variable z'. Evaluating the integrals yields

$$\mathbf{E} = \frac{\rho_\ell\,\rho}{4\pi\epsilon_o}\left[\frac{z'\,\hat{\mathbf{a}}_\rho}{\rho^2\sqrt{z'^2+\rho^2}} + \frac{\hat{\mathbf{a}}_z}{\rho\sqrt{z'^2+\rho^2}}\right]\Bigg|_{z'=-\infty}^{z'=\infty} = \frac{\rho_\ell\,\rho\,\hat{\mathbf{a}}_\rho}{4\pi\epsilon_o}\left[\frac{1}{\rho^2}+\frac{1}{\rho^2}\right] + 0\hat{\mathbf{a}}_z.$$

Simplifying this expression, we obtain

$$\mathbf{E} = \frac{\rho_\ell}{2\pi\epsilon_o\rho}\hat{\mathbf{a}}_\rho \qquad [\mathrm{V/m}] \tag{4.17}$$

This result is noteworthy for two reasons. The first is that the E-field generated by an infinite, uniform line charge is directed radially out from the line. This occurs because the field at any observation point can be considered as the sum of contributions from an infinite number of point-charge pairs. One such pair is shown in Figure 4-5a. As can be seen, the symmetric locations of these charges results in a net field that has only a radial component. Using the superposition principle, we find that the field due to all charges has this same property.

The second significant aspect of the E-field generated by a uniform, infinite line charge is that it varies as ρ^{-1} and not ρ^{-2}. To see why this occurs, we first note that only those charges that lie within a finite field of view of an observer have significant contributions to **E**. This is because those charges farthest from the observer contribute almost no radial component to **E**. Figure 4-5b shows a 90° field of view of an observer located a distance ρ from a line charge with charge density ρ_ℓ. As can be seen, the charge Q contained in this field of view is proportional to $\rho\rho_\ell$. Thus, the ρ^{-2} rate of decay of the field produced by each point charge in the line is compensated for by a total charge Q "seen" by an observer that is proportional to ρ, resulting in a net E-field that decays proportionally to ρ^{-1}.

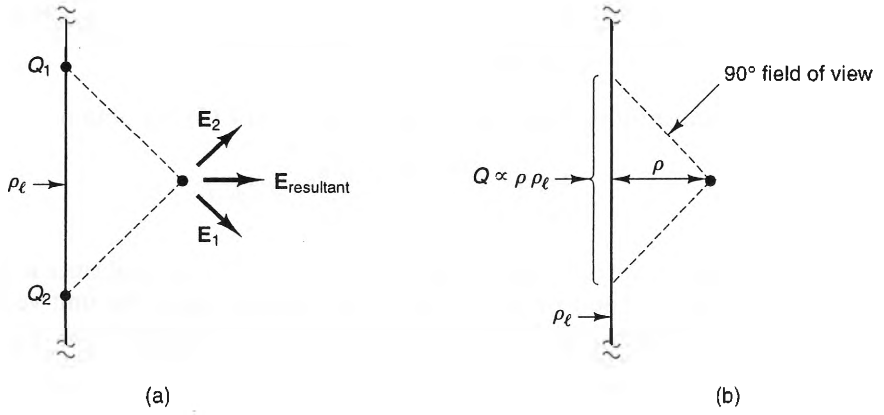


Figure 4-5 The E-field of an infinite line of charge. a) The cancellation of the tangential components of **E**, b) the $1/\rho$ variation of $|\mathbf{E}|$.

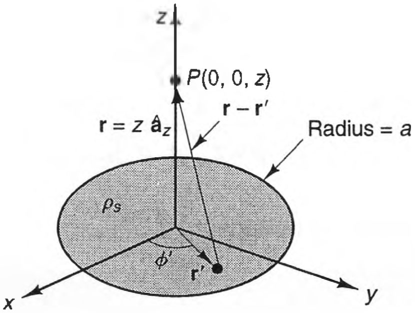


Figure 4-6 Geometry for determining the E-field generated by a uniform disk of charge.

4-4-2 THE UNIFORM DISK OF CHARGE

Consider the uniform circular disk of charge, shown in Figure 4-6. This disk has surface charge density ρ_s, has radius a, and lies in the $z = 0$ plane. Using the cylindrical coordinate system, we can represent the position vectors of an arbitrary field point $P(0, 0, z)$ on the z-axis and the dummy source position vector on the disk by

$$\mathbf{r} = z\hat{\mathbf{a}}_z$$
$$\mathbf{r}' = \rho'\hat{\mathbf{a}}_{\rho'},$$

respectively, noting that the direction of the unit vector $\hat{\mathbf{a}}_{\rho'}$ depends on the value of the source coordinate ϕ'. Thus,

$$\mathbf{r} - \mathbf{r}' = z\hat{\mathbf{a}}_z - \rho'\hat{\mathbf{a}}_{\rho'}.$$

Since $\hat{\mathbf{a}}_z$ and $\hat{\mathbf{a}}_{\rho'}$ are always perpendicular,

$$|\mathbf{r} - \mathbf{r}'|^3 = [z^2 + \rho'^2]^{3/2}.$$

Also,

$$ds' = \rho' d\rho' d\phi'.$$

Substituting these factors into Equation (4.15), we obtain

$$\mathbf{E} = \frac{\rho_s}{4\pi\epsilon_0}\int_0^a \int_0^{2\pi} \frac{z\hat{\mathbf{a}}_z - \rho'\hat{\mathbf{a}}_{\rho'}}{[z^2 + \rho'^2]^{3/2}} \rho' d\phi' d\rho', \tag{4.18}$$

where we note that ρ_s can be taken out of the integral since it is a constant. Care must be exercised when evaluating this integral, since the unit vector $\hat{\mathbf{a}}_{\rho'}$ is not a constant with respect to the integration variable ϕ'. However, we can express $\hat{\mathbf{a}}_{\rho'}$ in Cartesian components as

$$\hat{\mathbf{a}}_{\rho'} = \cos\phi'\,\hat{\mathbf{a}}_x + \sin\phi'\,\hat{\mathbf{a}}_y.$$

Substituting this into the integral, we obtain

$$\mathbf{E} = \frac{\rho_s \hat{\mathbf{a}}_z}{4\pi\epsilon_o}\int_0^a\int_0^{2\pi}\frac{z\rho'}{[z^2+\rho'^2]^{3/2}}\,d\phi'd\rho' - \frac{\rho_s \hat{\mathbf{a}}_x}{4\pi\epsilon_o}\int_0^a\int_0^{2\pi}\frac{\rho'^2\cos\phi'}{[z^2+\rho'^2]^{3/2}}\,d\phi'd\rho'$$

$$-\frac{\rho_s \hat{\mathbf{a}}_y}{4\pi\epsilon_o}\int_0^a\int_0^{2\pi}\frac{\rho'^2\sin\phi'}{[z^2+\rho'^2]^{3/2}}\,d\phi'd\rho',$$

where we note that the unit vectors can be taken out of the integrals, since they are all constants. The x and y components of $\mathbf{E}$ are zero, as they both involve integrals of either sin ϕ' or cos ϕ' over the range $0 < \phi' < 2\pi$. Evaluating the remaining integrals, we obtain

$$\mathbf{E} = \frac{\rho_s \hat{\mathbf{a}}_z}{4\pi\epsilon_o}\int_0^a\int_0^{2\pi}\frac{z\rho'}{[z^2+\rho'^2]^{3/2}}\,d\phi'd\rho' = \frac{\rho_s \hat{\mathbf{a}}_z}{2\epsilon_o}\int_0^a\frac{z\rho'}{[z^2+\rho'^2]^{3/2}}\,d\rho'$$

$$= \frac{\rho_s \hat{\mathbf{a}}_z}{2\epsilon_o}\left[\frac{-z}{\sqrt{z^2+\rho'^2}}\right]\Bigg|_{\rho'=0}^{\rho'=a},$$

which can be expressed as

$$\mathbf{E} = \frac{\rho_s}{2\epsilon_o}\left[\pm 1 - \frac{z}{\sqrt{z^2+a^2}}\right]\hat{\mathbf{a}}_z \quad \text{(uniform, circular disk of charge)}, \qquad (4.19)$$

where the upper sign is used for $z > 0$ and the lower sign is used for $z < 0$.

Figure 4-7 shows how E_z varies with the height z above a disk with $a = 1$[m] that contains a total charge Q. Also shown is $|\mathbf{E}|$ for a point charge at the origin with the same charge Q. As can be seen, E_z is finite just above the disk, whereas the field due to the point charge is unbounded as z approaches zero. We also see that, for large values of z, the fields of the disk and the point charge are nearly identical. This means that a charged disk looks more and more like a point charge when observed from increasingly large distances. This can also be seen from Equation (4.19) by using the binomial expansion to approximate the square root for large values of z.

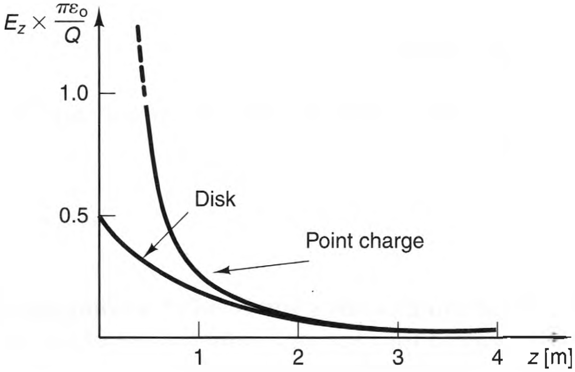


Figure 4-7 Comparison of the E-fields generated by a uniform disk of charge (with radius $a = 1$ [m]) and a point charge. Both contain the same total charge Q.

4-4-3 THE INFINITE SHEET OF CHARGE

When viewed from small distances, flat surfaces appear to an observer to be very large. Because of this, it is often convenient to approximate large surfaces of charge as infinite surfaces. We can obtain the E-field generated by an infinite, uniformly charged sheet from Coulomb's law, but we can obtain the same result from the expression derived in the previous section for the field generated by a circular disk. When the disk radius a approaches infinity, the disk becomes an infinite sheet in the $z = 0$ plane. For this case, Equation (4.19) yields

$$\mathbf{E} = \begin{cases} \dfrac{\rho_s}{2\epsilon_0}\hat{\mathbf{a}}_z & z > 0 \\ -\dfrac{\rho_s}{2\epsilon_0}\hat{\mathbf{a}}_z & z < 0 \end{cases} \quad \text{(infinite surface of uniform charge).} \tag{4.20}$$

Here we see that the field generated by an infinite surface of uniform charge does not vary with height above (or below) the surface and has opposite directions above and below the surface.

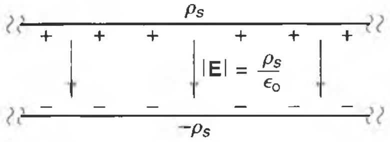


Figure 4-8 Two infinite sheets of uniform charge.

An important application of this result is presented in Figure 4-8, which shows two large sheets with opposite surface charge densities. This approximates the charge distributions found on parallel-plate capacitors with metal plates. The top sheet has a surface charge density of ρ_s and the bottom surface has a charge density of $-\rho_s$.

Between the sheets, the fields generated by each sheet add, producing a uniform field. Outside the sheets, the fields generated by each sheet cancel, yielding no net field in this region.

4-5 Field Computation Using Gauss's Law

Gauss's law states that the net flux passing through any closed surface is proportional to the charge Q_{enc} enclosed by that surface:

$$\oint_S \mathbf{E} \cdot \mathbf{ds} = \frac{Q_{\text{enc}}}{\epsilon_0}. \tag{4.21}$$

The surface S used when evaluating Gauss's law is called a ***Gaussian surface.*** Since Gauss's law is an explicit equation for Q_{enc}, one common use of the law is to find the total charge that is contained within some closed surface.

Example 4-2

Find the total charge contained within the cubic surface shown in Figure 4-9. Assume that the electric field intensity within the cube is given by the expression $\mathbf{E} = E_o[1 - e^{-|\alpha y|}]\hat{\mathbf{a}}_y$, where $E_o = 2[\mu\text{V/m}]$ and $\alpha = 1\ [\text{m}^{-1}]$.

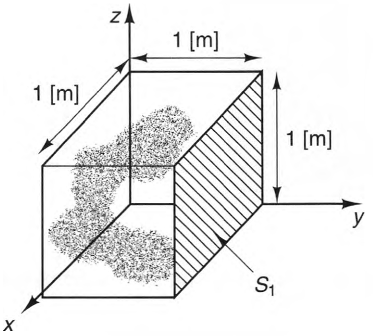


Figure 4-9 A cubic volume containing an unknown charge.

Solution:

Since **E** has only a y-component, $\mathbf{E} \bullet \mathbf{ds}$ is nonzero only on the y-directed faces. Also, since $\mathbf{E} = 0$ for $y = 0$, the integral receives a contribution only from the $y = 1$ face (S_1). Evaluating Gauss's law, we obtain

$$Q_{\text{enc}} = \epsilon_o \oint_S \mathbf{E} \bullet \mathbf{ds} = \epsilon_o \int_{S_1} \mathbf{E} \bullet \mathbf{ds} = E_o \epsilon_o \int_0^1 \int_0^1 [1 - e^{-1}]dx\, dz$$
$$= \epsilon_o E_o (1 - e^{-1}) S_1 = 1.12 \times 10^{-17} \qquad [\text{C}].$$

Gauss's law is an implicit equation of the variable **E**, which means that it cannot, in general, be solved explicitly for **E**. Nevertheless, it is possible to use this law to solve for the E-fields generated by certain classes of charge distributions. This is so whenever a Gaussian surface can be found on which **E** is constant and perpendicular to the surface. When such a surface exists, **E** can be pulled outside the integral and solved for directly. In the sections that follow, we will use Gauss's law to find the E-fields generated by charge distributions with cylindrical and spherical symmetry.

4-5-1 CYLINDRICALLY SYMMETRIC CHARGE DISTRIBUTIONS

One important class of charge distributions that can be easily analyzed using Gauss's law is those with rotational symmetry about an axis. These charge distributions are infinite in one dimension (along the z-axis) and are rotationally symmetric about this axis. Included in this class are the infinite line charge, hollow and solid circular cylinders, and coaxial cylinders (which we will usually refer to as coaxial lines).

Figure 4-10a depicts a rotationally symmetric charge distribution ρ_v. To show that Gauss's law can be used to find **E**, we must first determine how many components **E** has and what position variables they are functions of.

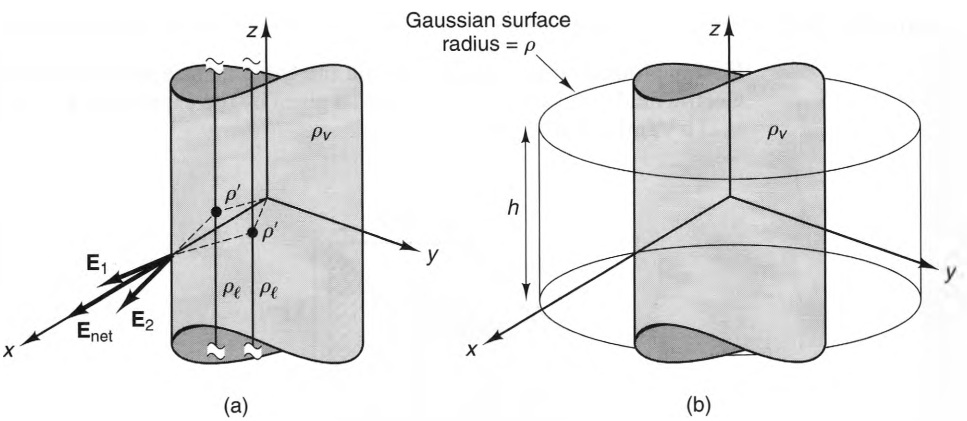


Figure 4-10 Rotationally symmetric charge distributions: a) cancellation of all components except E_ρ, b) a Gaussian surface.

We can accomplish this by recognizing that any charge distribution in this class can be considered as a collection of uniform infinite line charges, each parallel to the z-axis, such as the pair shown in the figure. Since both lines are the same radial distance ρ' from the z-axis, they have the same line-charge density. Their net E-field has only a radial component, since their ϕ components cancel. This same cancellation occurs for all of the complementary line-charge pairs that make up the entire charge distribution. Also, this charge distribution "looks" the same when observed at a constant radial coordinate ρ for all values of ϕ and z, so we can conclude that **E** can be a function only of the radial coordinate ρ. Hence, we can write **E** in the form

$$\mathbf{E} = E_\rho(\rho)\hat{\mathbf{a}}_\rho. \tag{4.22}$$

A family of Gaussian surfaces that can exploit the symmetric properties of this class of rotationally symmetric charge distributions consists of circular cylinders of arbitrary radius ρ and length h, centered about the z-axis. One such surface S is shown in Figure 4-10b. The radius ρ and height h can have any positive value, and charge can lie both inside and outside the surface. For any of these Gaussian surfaces, Gauss's law reads

$$\oint_S \mathbf{E}\cdot\mathbf{ds} = \underset{\text{cylinder}}{\int}\mathbf{E}\cdot\mathbf{ds} + \underset{\substack{\text{upper}\\\text{end cap}}}{\int}\mathbf{E}\cdot\mathbf{ds} + \underset{\substack{\text{lower}\\\text{end cap}}}{\int}\mathbf{E}\cdot\mathbf{ds} = \frac{Q_{\text{enc}}}{\epsilon_0}, \tag{4.23}$$

where Q_{enc} is the total charge contained inside the cylinder. The integrals along the end caps are both zero, since **E** is perpendicular to both end-cap surfaces. On the cylindrical portion of S, $\mathbf{E}\cdot\mathbf{ds} = \rho E_\rho(\rho)d\phi\,dz$. Since $E_\rho(\rho)$ is constant on this portion of S, it can be pulled outside the integral, yielding

$$\oint_S \mathbf{E}\cdot\mathbf{ds} = \int_0^h\int_0^{2\pi} E_\rho(\rho)\rho\,d\phi\,dz = 2\pi h\rho E_\rho(\rho) = \frac{Q_{\text{enc}}}{\epsilon_0}.$$

Solving for $E_\rho(\rho)$, we obtain

$$E_\rho(\rho) = \frac{Q_{\text{enc}}}{2\pi\epsilon_0 h\rho}. \tag{4.24}$$

As a result, we can write **E** in the form

$$\mathbf{E} = \frac{Q_{\text{enc}}}{2\pi h\epsilon_0\rho}\hat{\mathbf{a}}_\rho \quad \text{[V/m]}. \tag{4.25}$$

This expression may look the same as the field generated by a point charge, but it should be remembered that here Q_{enc} is the total charge contained within a cylinder of radius ρ and height h.

To demonstrate how to use Equation (4.25), consider the uniform, infinite cylinder of charge shown in Figure 4-11. The cylinder has radius a and has a constant volume charge density ρ_v. The charge contained within a cylinder of radius $\rho < a$ and height h is

$$Q_{\text{enc}} = 2\pi\int_0^h\int_0^\rho \rho_v\rho'\,d\rho'\,dz' = 2\pi h\rho_v\int_0^\rho \rho'\,d\rho' = \pi h\rho^2\rho_v.$$

Since there is no charge beyond $\rho = a$, we have

$$Q_{\text{enc}} = \begin{cases} \pi h\rho^2\rho_v & \rho < a \\ \pi ha^2\rho_v & \rho > a. \end{cases} \tag{4.26}$$

Substituting Equation (4.26) into Equation (4.25), we obtain

$$\mathbf{E} = \begin{cases} \dfrac{\rho\rho_v}{2\epsilon_0}\hat{\mathbf{a}}_\rho & \rho < a \\ \dfrac{a^2\rho_v}{2\epsilon_0\rho}\hat{\mathbf{a}}_\rho & \rho > a \end{cases} \quad \text{(solid, uniformly charged cylinder).} \tag{4.27}$$

Figure 4-12b shows E_ρ vs. ρ for a uniformly charged cylinder. Inside the cylinder, E_ρ grows linearly with increasing distance from the z-axis. Outside the cylinder, E_ρ drops off as ρ^{-1}, just like the field of an infinite line charge. Figure 4-12a shows E_ρ for an infinite line charge that has the same charge per unit length as the solid cylinder does.

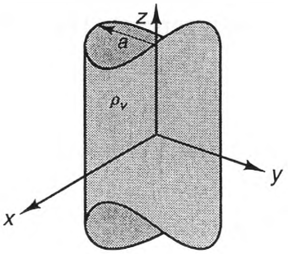


Figure 4-11 An infinite, solid cylinder of charge with uniform charge density.

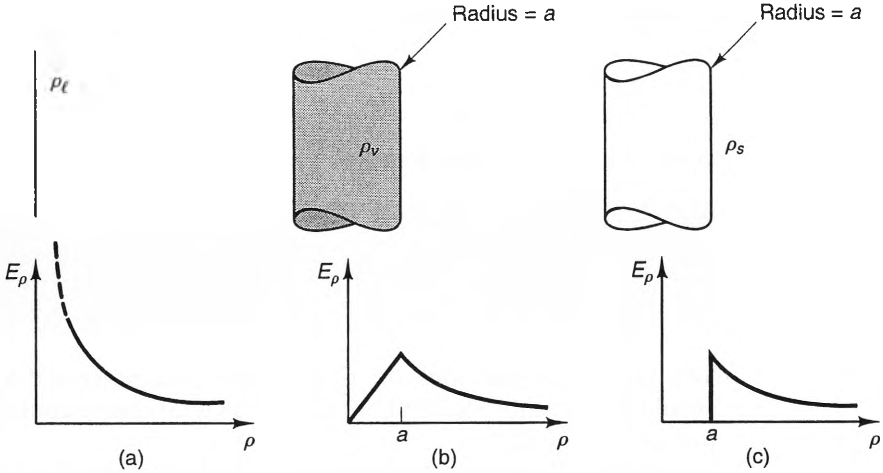


Figure 4-12 Comparison of the E-fields generated by three cylindrical charge distributions that carry the same charge per unit length: a) line charge, b) solid cylinder, c) hollow cylinder.

We can also use Gauss's law to find the E-field generated by a hollow, uniformly charged cylinder. If the surface charge density is ρ_s and the cylinder radius is a, the charge contained within a cylinder length h is

$$Q_{\text{enc}} = \begin{cases} 0 & \rho < a \\ 2\pi h a \rho_s & \rho > a. \end{cases} \tag{4.28}$$

Substituting Equation (4.28) into Equation (4.25), we obtain

$$\mathbf{E} = \begin{cases} 0 & \rho < a \\ \dfrac{a\rho_s}{\epsilon_0 \rho}\,\hat{\mathbf{a}}_\rho & \rho > a \end{cases} \quad \text{(hollow, uniformly charged cylinder).} \tag{4.29}$$

The variation of this field with distance ρ is depicted in Figure 4-12c. Notice in this figure that the field outside a hollow cylinder is identical to the field generated by a solid cylinder that contains the same charge per unit length.

Coaxial Cylinders. Figure 4-13a shows two infinite coaxial surface charge distributions. The cylinders have radii a and b, respectively, and uniform surface charge densities ρ_{sa} and ρ_{sb}, respectively. We can find the field generated by these coaxial cylinders simply by adding the fields generated by each cylinder individually. Using Equation (4.29) for the fields generated by both the inner and outer cylinders, we obtain

$$\mathbf{E} = \begin{cases} 0 & \rho < a \\ \dfrac{a\rho_{sa}}{\epsilon_0 \rho}\,\hat{\mathbf{a}}_\rho & a < \rho < b \\ \dfrac{a\rho_{sa} + b\rho_{sb}}{\epsilon_0 \rho}\,\hat{\mathbf{a}}_\rho & \rho > b \end{cases} \quad \text{(coaxial cylinders of charge).} \tag{4.30}$$

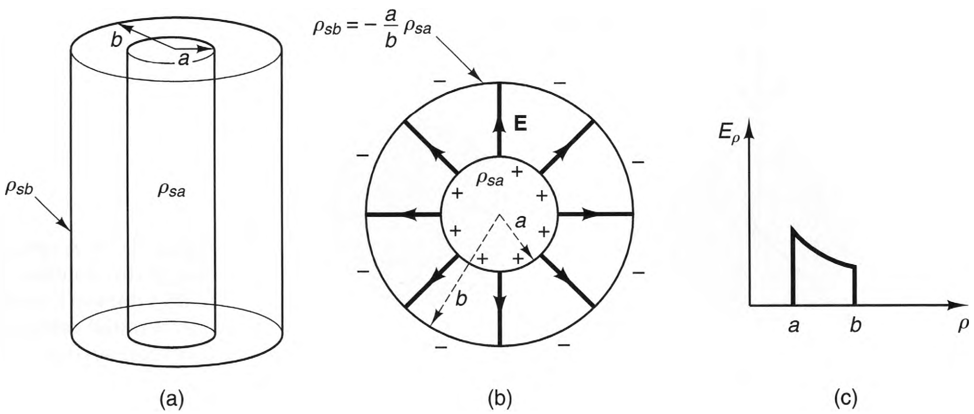


Figure 4-13 Coaxial cylinders of charge that contain opposite charges per unit length: a) side view, b) cross-sectional view, c) variation of E_ρ with radial position ρ.

Here, we see that the field between the cylinders is controlled only by the charge density on the inner cylinder and the field outside is controlled by the charges on both cylinders. An important special case of Equation (4.30) occurs when the charges per unit length on the inner and outer cylinders are exactly opposite. For this case, we have $a\rho_{sa} = -b\rho_{sb}$, and

$$\mathbf{E} = \begin{cases} \dfrac{a\rho_{sa}}{\epsilon_0 \rho}\hat{\mathbf{a}}_\rho & a < \rho < b \\ 0 & \text{elsewhere} \end{cases} \qquad \text{(balanced, coaxial cylinders).} \qquad (4.31)$$

Figure 4-13b shows the streamlines of **E** and Figure 4-13c shows E_ρ as a function of ρ, both for the balanced case. As can be seen, the field outside the outer cylinder is zero. This is why the outer conductor of a coaxial line (cable) is sometimes called a shield. Coaxial cables are often used when interference between adjacent circuits is undesirable.

4-5-2 SPHERICALLY SYMMETRIC CHARGE DISTRIBUTIONS

Gauss's law can also be used to evaluate the fields generated by spherically symmetric charge distributions. Included in this class are solid, hollow, and layered spheres of charge. Figure 4-14a depicts an arbitrary spherically symmetric charge distribution. Since ρ_v varies only with the radial coordinate r, it can be considered as a collection of complementary point-charge pairs, such as the pair shown in the figure. These two charges have the same r and θ coordinates, but their ϕ coordinates differ by 180°. Since both points are equidistant from the origin, their net E-field along the z-axis has only a radial component. Our choice of the direction of the z-axis was arbitrary, so we can conclude that the E-field generated by spherically symmetric charge distribution has only a radial component at all field points. Also, since the entire charge distribution "looks" the same when observed at a constant radius for all values of θ and ϕ, **E** is independent of these coordinates. Hence, we can write **E** in the form

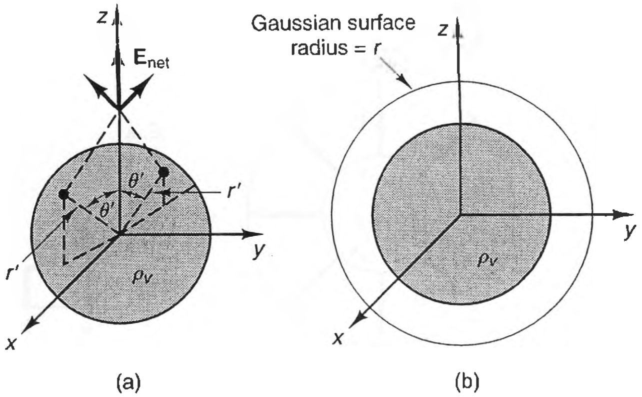


Figure 4-14 A spherically symmetric charge distribution. a) Field cancellations of all symmetric point-charge pairs yield only a radial component of **E**. b) A Gaussian surface.

$$\mathbf{E} = E_r(r)\hat{\mathbf{a}}_r. \tag{4.32}$$

Considering the spherically symmetric characteristics of the E-field given by Equation (4.32), it should come as no surprise that the Gaussian surfaces that exploit these symmetric properties are spheres centered at the origin. One such sphere is shown in Figure 4-14b. Knowing that $\mathbf{ds} = r^2 \sin\theta \, d\theta \, d\phi \, \hat{\mathbf{a}}_r$ at all points on a sphere that is centered at the origin, we can substitute $\mathbf{E} \cdot \mathbf{ds} = E_r(r)\, r^2 \sin\theta \, d\theta \, d\phi$ into Gauss's law, obtaining

$$\frac{Q_{\text{enc}}}{\epsilon_o} = \oint_S \mathbf{E} \cdot \mathbf{ds} = \int_0^{2\pi} \int_0^{\pi} E_r(r) r^2 \sin\theta \, d\theta \, d\phi = 4\pi r^2 E_r(r),$$

where Q is the charge contained inside a sphere of radius r. Solving this expression for $E_r(r)$ and substituting the latter into Equation (4.32), we obtain

$$\mathbf{E} = \frac{Q_{\text{enc}}}{4\pi\epsilon_o r^2}\hat{\mathbf{a}}_r \quad [\text{V/m}] \quad \text{(spherically symmetric charge distributions).} \tag{4.33}$$

We can use Equation (4.33) to find the E-field generated by a solid, uniformly charged sphere, as shown in Figure 4-15. If the sphere has radius a and charge density ρ_v, the charge contained within a spherical Gaussian surface of radius $r < a$ is

$$Q_{\text{enc}} = 4\pi \int_0^r r'^2 \rho_v \, dr' = 4\pi\rho_v \int_0^r r'^2 dr' = \frac{4}{3}\pi r^3 \rho_v.$$

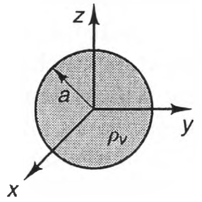


Figure 4-15 A solid sphere of charge.

Since no more charge is enclosed inside the Gaussian surface as r is increased beyond $r = a$, $Q_{\text{enc}} = (4/3)\ \pi a^3 \rho_v$ for $r > a$. Substituting these expressions for Q_{enc} into Equation (4.33), we obtain

$$\mathbf{E} = \begin{cases} \dfrac{r\rho_v}{3\epsilon_o}\,\hat{\mathbf{a}}_r & r < a \\ \dfrac{a^3\rho_v}{3\epsilon_o r^2}\,\hat{\mathbf{a}}_r & r > a \end{cases} \qquad \text{(uniformly charged sphere).} \tag{4.34}$$

We can also express this result in terms of the total charge $Q_{\text{T}} = (4/3)\ \pi a^3 \rho_v$ contained within the sphere:

$$\mathbf{E} = \begin{cases} \dfrac{rQ_{\text{T}}}{4\pi\epsilon_o a^3}\,\hat{\mathbf{a}}_r & r < a \\ \dfrac{Q_{\text{T}}}{4\pi\epsilon_o r^2}\,\hat{\mathbf{a}}_r & r > a \end{cases} \qquad \text{(uniformly charged sphere).} \tag{4.35}$$

Figure 4-16b shows the variation of E_r with the radial coordinate r for a uniform, solid sphere of charge. As can be seen, E_r is proportional to r inside the sphere and falls off as r^{-2} outside. Comparing Figures 4-16a and b, we see that the field outside the sphere is indistinguishable from that of a point charge of value Q_{T} placed at the origin.

We can also use Gauss's law to find the E-field generated by a hollow, uniformly charged sphere. If the surface charge density is ρ_s, the charge on the sphere is simply ρ_s times the surface area of the sphere, so Q_{enc} contained within a Gaussian surface of radius r is

$$Q_{\text{enc}} = \begin{cases} 0 & r < a \\ 4\pi a^2 \rho_s & r > a, \end{cases} \tag{4.36}$$

where a is the radius of the sphere. Substituting Equation (4.36) into Equation (4.33), we obtain

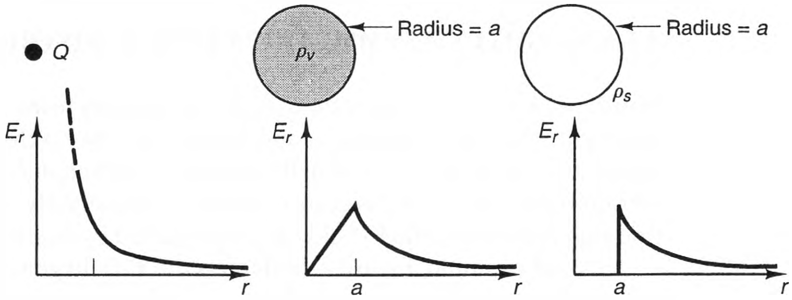


Figure 4-16 Comparison of the E-fields generated by three spherically symmetric charge distributions with total charge Q: a) a point charge, b) a solid sphere, c) a hollow sphere.

$$\mathbf{E} = \begin{cases} 0 & r < a \\ \dfrac{a^2\rho_s}{\epsilon_0 r^2}\,\hat{\mathbf{a}}_r = \dfrac{Q_T}{4\pi\epsilon_0 r^2}\,\hat{\mathbf{a}}_r & r > a \end{cases} \quad \text{(hollow, uniformly charged sphere)}, \tag{4.37}$$

where Q_T is the total charge on the sphere. The variation of this field with distance r is depicted in Figure 4-16c. Notice in this figure that the field outside a hollow cylinder is identical to the field generated by a solid sphere that contains the same charge.

Example 4-3

Figure 4-17 shows two hollow, concentric spheres. If the inner and outer spheres each have uniform surface charge distributions and have radii a and b, respectively, find the generated E-field when the total charge on the inner and outer spheres is Q and $-Q$, respectively.

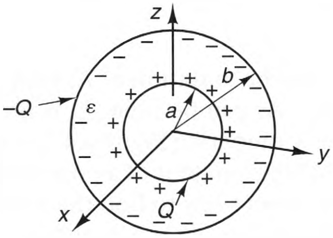


Figure 4-17 Concentric spheres of charge.

Solution:

By using the superposition principle, we can express the total E-field as the sum of the fields generated by each sphere individually. Using Equation (4.33), we can write

$$\mathbf{E} = \begin{cases} \dfrac{Q}{4\pi\epsilon_0 r^2}\,\hat{\mathbf{a}}_r & a < r < b \\ 0 & \text{otherwise.} \end{cases} \tag{4.38}$$

Here we see that E_r between the spheres is the same as that generated by the inner sphere alone, whereas the field outside the outer sphere is exactly zero.

4-5-3 SLIGHTLY ASYMMETRIC CHARGE DISTRIBUTIONS

Before we leave our discussion of field calculations using Gauss's law, it is important to emphasize that this technique is applicable *only* when a charge distribution has sufficient symmetry. Any deviation from this symmetry will change the character of **E**, sometimes substantially. When a charge distribution is only slightly asymmetric, however, we often find that portions of the E-field can be predicted by neglecting the asymmetry.

As an example, Figure 4-18 depicts the E-field generated by a coaxial cable when a slot is cut into the outer conductor. In this case, the charge distribution does not have perfect cylindrical symmetry, even when the total charge per unit length on the two conductors is balanced. This is because no charges exist in the slot and the surface charge densities along the remaining surfaces are not perfectly uniform. Compar-

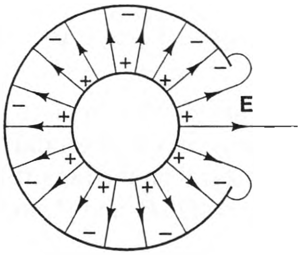


Figure 4-18 The E-field distribution generated by a coaxial cable with a longitudinal slit in the outer conductor.

ing these streamlines with those generated by a pair of balanced coaxial cylinders, we see both similarities and differences. Far from the slot the differences are minimal. Near the slot, however, the differences are more obvious. This is particularly true outside the outer surface, where the field is no longer zero.

4-6 Voltage and Electric Potential

Our discussion thus far has concentrated on the relationship between electrostatic charge distributions and the E-fields that they generate. This is fitting, since the E-field is responsible for the forces that charges exert upon each other. There is, however, an important scalar quantity that is also associated with electrostatic fields, called the electric potential. We will start our discussion of electric potential by remembering that the E-field generated by a static charge distribution has zero curl at all points in space:

$$\nabla \times \mathbf{E} = 0.$$

In integral form, this expression reads:

$$\oint_C \mathbf{E} \cdot \mathbf{d}\boldsymbol{\ell} = 0.$$

Both of these equations state that an electrostatic E-field is a conservative vector field. According to Equation (2.131), any conservative vector field can be represented at any point as the gradient of a scalar function, so we can represent **E** as:

$$\mathbf{E} = -\nabla V, \tag{4.39}$$

where V is called the ***electric potential function***,[2] measured in units of volts. According to this equation, there is a one-to-one correspondence between the vector **E** at a point and the behavior of the scalar field V at the same point. The minus sign is chosen in this expression so that **E** always points towards *decreasing* values of V.

We can develop a more physical interpretation of the electric potential function by integrating Equation (4.39) between two points. Taking the dot product of both sides with a differential displacement vector $\mathbf{d}\boldsymbol{\ell}$ and integrating along a path between two endpoints P_a and P_b, we obtain

[2] The electric potential function is also called the ***electrostatic potential function*** or the ***scalar potential function***.

$$-\int_b^a \mathbf{E}\cdot\mathbf{d\ell} = \int_b^a \nabla V\cdot\mathbf{d\ell} = \int_b^a dV = V_a - V_b. \tag{4.40}$$

Here, we have used Equation (2.80) to replace the dot product $\nabla V\cdot\mathbf{d\ell}$ with the differential dV, which is the change in the potential V between the endpoints of the differential path $\mathbf{d\ell}$. The line integral on the far left-hand side of Equation (4.40) is defined as the ***voltage*** between the points P_a and P_b and is denoted by the symbol V_{ab}:

$$V_{ab} \equiv -\int_b^a \mathbf{E}\cdot\mathbf{d\ell} = \int_a^b \mathbf{E}\cdot\mathbf{d\ell} \qquad \text{(electrostatic and time-varying fields).} \tag{4.41}$$

This definition is valid for both electrostatic and time-varying fields. However, for the special case of electrostatic fields, we have from Equation (4.40) that V_{ab} can also be expressed as the difference of the electric potentials at the endpoints of the path of integration:

$$V_{ab} = -\int_b^a \mathbf{E}\cdot\mathbf{d\ell} = V_a - V_b \qquad \text{(electrostatic fields),} \tag{4.42}$$

where V_a and V_b are the potentials at the points P_a and P_b, respectively. In this case V_{ab} equals the difference of the potentials at the endpoints, so the terms *voltage* and *potential difference* are interchangable for electrostatic fields. This means that the voltage between two points in an electrostatic field is independent of the path of integration chosen between the endpoints.[3]

The voltage between two points is a measure of the work necessary to move a charge between the points. To show this, consider the two points P_a and P_b, shown in Figure 4-19. Also shown are two paths leading from P_a to P_b. The force acting on a test charge Q is $\mathbf{F} = Q\mathbf{E}$, so the work per unit charge done by the field on the charge when moving it from P_a and P_b is

$$\frac{W_{ab}}{Q} = \int_a^b \mathbf{E}\cdot\mathbf{d\ell} = V_{ab}. \tag{4.43}$$

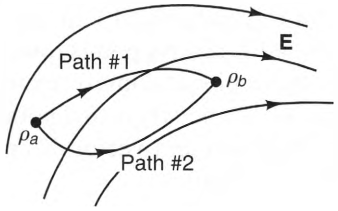


Figure 4-19 Two different paths with common endpoints in an electrostatic E-field.

[3] We will see in Chapter 9 that voltages can be path dependant when time-varying magnetic fields are present.

Using this result, we can now offer the following interpretation of the voltage between two points.

The voltage V_{ab} between two points is the work per unit charge done by the electric field when a positive test charge is moved from P_a to P_b along a given path.

This interpretation of voltage is valid for both electrostatic and time-varying fields. However, in the case of electrostatic fields, the voltage between two points is unique (i.e., not dependant on the path of integration). Also, the energy it takes to move a charge between two points in an electrostatic field is independant of the path chosen.

It is important to specify voltages with the correct sign. By convention, the potential difference V_{ab} is indicated by denoting a "+" sign at the first point (P_a) and a "−" sign at the second point (P_b). When this convention is used (as in circuit analysis), it is possible to drop the subscripts and simply use a symbol such as "V", since this notation clearly indicates that V is the voltage between the "+" and "−" points. In this case, however, the symbol V is *not* the electrostatic potential function (defined in Equation (4.39)), but rather the difference of the electrostatic potential functions at the "+" and "−" points.

Example 4-4

Find the voltage V between the two infinite sheets of charge shown in Figure 4-20. Assume that the top sheet has a positive surface charge density of ρ_s and the bottom sheet has a negative charge density $-\rho_s$.

Solution:

We found earlier (see Section 4.4.3) that the E-field between balanced sheets of charge has a magnitude of ρ_s/ϵ_o and is directed from the top (positive) sheet to the bottom (negative) sheet. Since the "+" sign is located at the top sheet, we have, from Equation (4.41), that

$$V = -\int_{\text{bottom}}^{\text{top}} \mathbf{E} \cdot d\boldsymbol{\ell}.$$

The simplest path between the "+" and "−" points is straight down along the z-axis, so

$$V = -\frac{\rho_s}{\epsilon_o}\int_d^0 dz = \frac{\rho_s}{\epsilon_o}\int_0^d dz = \frac{\rho_s d}{\epsilon_o} \qquad [\text{V}].$$

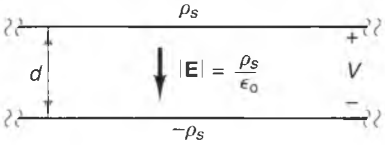


Figure 4-20 Two uniform, infinite sheets of charge with opposite charges, separated by a distance d.

Here we see that V is positive when $\mathbf{E}$ is directed from the "+" sign to the "−" sign. Also, since $\mathbf{E}$ is perpendicular to both surfaces, we find the same potential difference V between any point on the upper surface and any point on the lower surface.

So far we have defined the electric potential V in terms of $\mathbf{E}$, but it is also possible to derive an expression for V in terms of the charge distribution that generates $\mathbf{E}$. We will do this by first finding the potential function for a point charge and then generalizing this expression for an arbitrary charge distribution. Figure 4-21 shows a point charge Q and two points a and b, located radial distances R_a and R_b from Q, respectively.

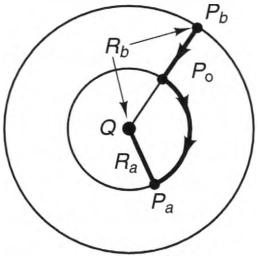


Figure 4-21 Geometry for determining the potential difference between two points due to the E-field of a point charge.

The potential difference V_{ab} between these points is

$$V_{ab} = -\int_b^a \mathbf{E} \cdot d\boldsymbol{\ell}.$$

The simplest path between the points is the one shown in the figure, from P_b, to P_o, to P_a. Since $\mathbf{E}$ is at all points perpendicular to the path from P_o to P_a, the contribution to V_{ab} along this portion of the path is zero. Thus, using the E-field of a point charge (Equation (4.11)), we can write

$$V_{ab} = -\int_{R_b}^{R_a} \frac{1}{4\pi\epsilon_o}\frac{Q}{R^2}\hat{\mathbf{a}}_R \cdot \hat{\mathbf{a}}_R dR = \frac{Q}{4\pi\epsilon_o}\left[\frac{1}{R_a} - \frac{1}{R_b}\right],$$

where the unit vector $\hat{\mathbf{a}}_R$ points outward from Q. If we choose R_b to be infinity, we obtain the following expression for the potential V at a radial distance R from a point charge Q, referenced to infinity:

$$V = \frac{Q}{4\pi\epsilon_o R} \quad [\text{V}]. \tag{4.44}$$

From this expression, we see that the potential rises as one approaches a positive point charge.

We can generalize Equation (4.44) to find the potential function generated by an arbitrary charge distribution. First, for a collection of N point charges, we find

$$V = \frac{1}{4\pi\epsilon_o}\sum_{k=1}^{N}\frac{Q_k}{|\mathbf{r} - \mathbf{r}_k|}, \tag{4.45}$$

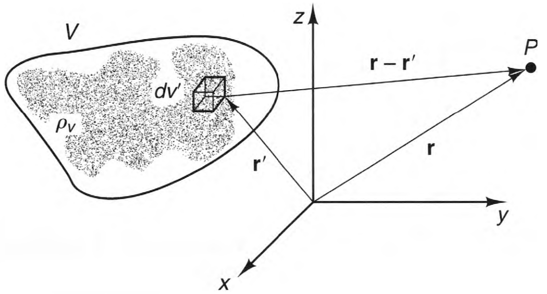


Figure 4-22 Geometry for determining the potential field generated by an arbitrary charge distribution.

where $\mathbf{r}$ is the position vector that represents the field point and $\mathbf{r}_k$ is the position vector of the charge Q_k. Next, for the volume-charge distribution shown in Figure 4-22, the contribution of the charge $\rho_v dv'$ in the volume dv' to the potential at an arbitrary point is

$$dV = \frac{\rho_v dv'}{4\pi\epsilon_o |\mathbf{r} - \mathbf{r}'|},$$

where $\mathbf{r}$ and $\mathbf{r}'$ are the position vectors of the field point P and the differential volume dv', respectively. Summing the contributions from all the charges, we obtain

$$V = \frac{1}{4\pi\epsilon_o} \int_{\text{Vol.}} \frac{\rho_v dv'}{|\mathbf{r} - \mathbf{r}'|} \quad [\text{V}]. \tag{4.46}$$

In this integral, the primed variables are the dummy integration variables. Also, the integration takes place only at locations where charge is present. Using a similar sequence of steps, we can find the following potential functions for surface- and line-charge distributions:

$$V = \frac{1}{4\pi\epsilon_o} \int_S \frac{\rho_s ds'}{|\mathbf{r} - \mathbf{r}'|} \quad [\text{V}], \tag{4.47}$$

where the surface-charge distribution ρ_s is contained in the surface S, and

$$V = \frac{1}{4\pi\epsilon_o} \int_C \frac{\rho_\ell d\ell'}{|\mathbf{r} - \mathbf{r}'|} \quad [\text{V}], \tag{4.48}$$

where the line-charge distribution ρ_ℓ lies on the contour C.

Example 4-5

For the uniform circular loop of charge shown in Figure 4-23, find the E-field it generates at an arbitrary point $P(0, 0, z)$ along the z-axis by first finding the potential V. Assume that the radius of the loop is a and the line-charge density is ρ_ℓ.

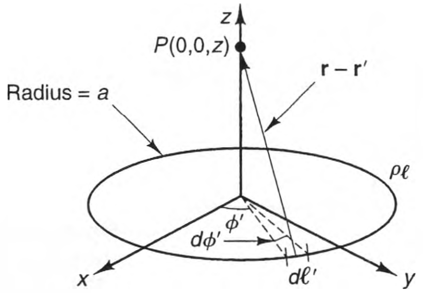


Figure 4-23 A uniformly charged circular loop.

Solution:

Substituting $\mathbf{r} = z\hat{\mathbf{a}}_z$, $\mathbf{r}' = a\hat{\mathbf{a}}_\rho$, and $d\ell' = a\,d\phi'$ into Equation (4.48), the potential V at an arbitrary point along the z-axis is

$$V = \frac{1}{4\pi\epsilon_0}\oint_C \frac{\rho_\ell}{|\mathbf{r} - \mathbf{r}'|}\,d\ell' = \frac{\rho_\ell}{4\pi\epsilon_0}\int_0^{2\pi} \frac{1}{\sqrt{z^2 + a^2}}\,a\,d\phi' = \frac{a\rho_\ell}{2\epsilon_0\sqrt{z^2 + a^2}}.$$

The E-field generated by the loop is related to V by

$$\mathbf{E} = -\nabla V = -\frac{\partial V}{\partial \rho}\hat{\mathbf{a}}_\rho - \frac{1}{\rho}\frac{\partial V}{\partial \phi}\hat{\mathbf{a}}_\phi - \frac{\partial V}{\partial z}\hat{\mathbf{a}}_z.$$

This equation implies that $\mathbf{E}$ may have up to three components, but we can surmise that the ρ and ϕ components of $\mathbf{E}$ are zero at all points on the z-axis, since this charge distribution is symmetric about that axis. Thus, the z-component of $\mathbf{E}$ is

$$E_z = -\frac{\partial V}{\partial z} = -\frac{\partial}{\partial z}\left[\frac{a\rho_\ell}{2\epsilon_0\sqrt{z^2 + a^2}}\right] = \frac{az\rho_\ell}{2\epsilon_0[z^2 + a^2]^{3/2}}.$$

Hence,

$$\mathbf{E} = \frac{az\rho_\ell}{2\epsilon_0[z^2 + a^2]^{3/2}}\hat{\mathbf{a}}_z.$$

It is left as an exercise to show that this same result could be obtained using Coulomb's law.

4-6-1 ABSOLUTE AND RELATIVE POTENTIALS

The potentials given by Equations (4.46) through (4.48) are called ***absolute potentials***, because each represents the potential difference between a field (i.e., observation) point and a point at infinity. Choosing infinity as a reference point is convenient for many field calculations, because the potentials of all charge distributions with finite dimensions approach zero at points that are very far from the charges. We can see this by taking the limit of Equation (4.46) as $\mathbf{r} \to \infty$. If $\rho_v \to 0$ at infinity, we have

$$V(\infty) = \frac{1}{4\pi\epsilon_0}\int_{\text{Vol.}} \lim_{\mathbf{r}\to\infty}\left[\frac{\rho_v}{|\mathbf{r} - \mathbf{r}'|}\right] dv' = 0.$$

The same thing occurs for surface- and line-charge distributions when they are of finite extent (i.e., contain no charges at infinity).[4]

We can also define ***relative potentials*** that are referenced to points other than at infinity. This is particularly convenient when there exists a constant potential surface near the system in question, such as a metal chassis. We can define a relative potential function V' as

$$V' = V - V_{\text{ref}}, \tag{4.49}$$

where V is the absolute potential and V_{ref} is a reference potential. Comparing this expression with Equation (4.42), we see that V' is the potential difference between a point and a point (or surface) with absolute potential V_{ref}. Since V_{ref} is a constant, $\mathbf{E}$ is related to both V and V' by the same formula:

$$\mathbf{E} = -\nabla V = -\nabla V'. \tag{4.50}$$

4-6-2 THE ELECTRIC DIPOLE

An electric dipole consists of two point charges of equal magnitude and opposite sign, separated by a distance d. Figure 4-24a shows an electric dipole located on the z-axis. Although the E-field of a dipole at a point P can be found by summing the fields of the individual point charges, the resulting expression is cumbersome and difficult to simplify. A better approach is to first find the potential function associated with this charge pair and then simplify it for the case where the distance r from the observer to the dipole is much larger than the dipole spacing d.

Using Equation (4.44) for each charge, we can write the potential generated by the charge pair as

$$V = \frac{Q}{4\pi\epsilon_0 R_+} - \frac{Q}{4\pi\epsilon_0 R_-} = \frac{Q}{4\pi\epsilon_0}\left(\frac{1}{R_+} - \frac{1}{R_-}\right), \tag{4.51}$$

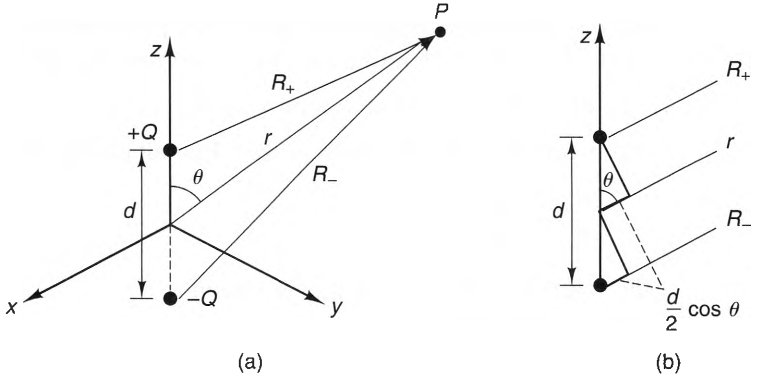


Figure 4-24 a) The dimensions of an electric dipole, b) the relationship between the radial distances for a far-zone observer.

[4] This is not true for infinite sheets of charge or for infinite line-sources, since they contain charges out to infinity.

where R_+ and R_- are the distances from P to the positive and negative charges, respectively. From Figure 4.24b, it follows that when $r \gg d$,

$$R_{\pm} \approx r \mp \frac{d}{2} \cos\theta.$$

Applying the binomial expansion $1/(x \mp a) \approx (1/x) \pm (a/x^2)$ when $x \gg a$, we can write

$$\frac{1}{R_{\pm}} \approx \frac{1}{r} \pm \frac{d}{2r^2} \cos\theta.$$

From this, we obtain

$$\frac{1}{R_+} - \frac{1}{R_-} \approx \frac{d\cos\theta}{r^2}. \tag{4.52}$$

Substituting this approximation into Equation (4.51), we obtain

$$V \approx \frac{Qd\cos\theta}{4\pi\epsilon_0 r^2} \quad r >> d. \tag{4.53}$$

Finally, the E-field is obtained by using

$$\mathbf{E} = -\nabla V = -\left(\frac{\partial V}{\partial r}\hat{\mathbf{a}}_r + \frac{1}{r}\frac{\partial V}{\partial \theta}\hat{\mathbf{a}}_\theta + \frac{1}{r\sin\theta}\frac{\partial V}{\partial \phi}\hat{\mathbf{a}}_\phi\right).$$

Evaluating the partial derivatives, we find that

$$\mathbf{E} \approx \frac{Qd}{4\pi\epsilon_0 r^3}(2\cos\theta\,\hat{\mathbf{a}}_r + \sin\theta\,\hat{\mathbf{a}}_\theta) \quad r >> d. \tag{4.54}$$

Expressions for each of the E-field streamlines can be obtained by remembering that the differential path vector $\mathbf{d}\ell$ at each point on a streamline is parallel to $\mathbf{E}$ at that point. In spherical coordinates, we have, from Equation (2.61) that

$$\mathbf{d}\ell = dr\,\hat{\mathbf{a}}_r + r\,d\theta\,\hat{\mathbf{a}}_\theta + r\sin\theta\,d\phi\,\hat{\mathbf{a}}_\phi.$$

This means that in order to have $\mathbf{d}\ell \propto \mathbf{E}$ at all points, we must require that

$$\frac{E_\theta}{E_r} = \frac{r\,d\theta}{dr} = \frac{\sin\theta}{2\cos\theta}.$$

The solution of this differential equation is

$$r = C\sin^2\theta, \tag{4.55}$$

where C is an arbitrary constant; each value of C corresponds to a different streamline. The several streamlines and constant potential surfaces of an electric dipole are plotted in Figure 4-25.

A distinctive characteristic of the E-field of a dipole is that it decays proportionally to r^{-3}, whereas the field of a single point charge decays as r^{-2}. The reason for this

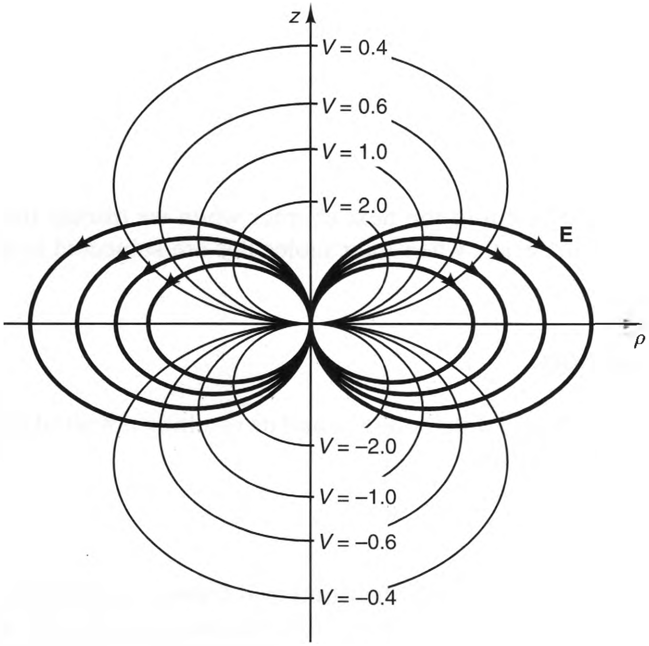


Figure 4-25 Field lines of an electrostatic dipole.
▬ E-field streamlines
— Constant potential surfaces

is simple: The farther away an observer gets, the more the two charges appear to lie at the same point. Also, the field very close to the dipole center is more intense than that due to a single point charge. Thus, the field generated by an electric dipole is very localized.

We can derive a simple expression for the potential V of a dipole whose center is at the origin, but is not necessarily aligned with the z-axis. This can be accomplished by noticing that the term $Qd \cos \theta$, which appears in Equation (4.53), can be written as a dot product:

$$Qd \cos \theta = \mathbf{p} \cdot \hat{\mathbf{a}}_r.$$

In this expression, $\mathbf{p}$ is called the ***dipole moment***, defined by

$$\mathbf{p} = Q\mathbf{d} \qquad [\mathrm{C \cdot m}], \tag{4.56}$$

where $\mathbf{d}$ is the directed distance from the negative charge to the positive charge. From this definition, the potential of an electric dipole located at the origin can be written as

$$V = \frac{\mathbf{p} \cdot \hat{\mathbf{a}}_r}{4\pi\epsilon_0 r^2} \qquad [\mathrm{V}].$$

Finally, if the center of the dipole is located at $\mathbf{r}'$, V can be expressed by

$$V = \frac{\mathbf{p} \cdot \hat{\mathbf{a}}_R}{4\pi\epsilon_0 R^2} = \frac{\mathbf{p} \cdot (\mathbf{r} - \mathbf{r}')}{4\pi\epsilon_0 |\mathbf{r} - \mathbf{r}'|^3} \qquad [\mathrm{V}], \tag{4.57}$$

where

$$\hat{\mathbf{a}}_R = \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|} \tag{4.58}$$

and

$$R = |\mathbf{r} - \mathbf{r}'|. \tag{4.59}$$

We will find Equation (4.59) useful in the next chapter when we discuss the electric dipoles that are induced in materials when their molecules are subjected to an externally generated E-field.

4-7 Poisson's and Laplace's Equations

We have already seen that a charge distribution ρ_v and its resulting potential function V are related by the integral expression

$$V = \frac{1}{4\pi\epsilon_o} \int_{\text{Vol.}} \frac{\rho_v \, dv'}{|\mathbf{r} - \mathbf{r}'|}. \tag{4.60}$$

This equation is useful whenever a charge distribution is known everywhere. There are, however, many situations in which the charge distribution is known only in certain regions, but the potential V is also known along certain boundaries. In these cases, it is far more useful to have a differential equation that relates V and ρ_v.

We can find this differential equation by manipulating Equation (4.60), but a much simpler way is by starting with the point form of Gauss's law,

$$\nabla \cdot \mathbf{E} = \frac{\rho_v}{\epsilon_o}.$$

Substituting $\mathbf{E} = -\nabla V$, we can write

$$-\nabla \cdot \nabla V = -\nabla^2 V = \frac{\rho_v}{\epsilon_o},$$

where ∇^2 denotes the Laplacian operator. Thus,

$$\nabla^2 V = -\frac{\rho_v}{\epsilon_o}. \tag{4.61}$$

This differential equation is called ***Poisson's equation***. A special case of Poisson's equation occurs in source-free regions (i.e., where $\rho_v = 0$):

$$\nabla^2 V = 0 \qquad (\rho_v = 0). \tag{4.62}$$

This is called ***Laplace's equation***.

Example 4-6

Prove that the potential function for a point charge, given by Equation (4.44), satisfies Poisson's equation.

Solution:

For a point charge Q located at the origin,

$$V = \frac{Q}{4\pi\epsilon_o r}.$$

Taking the gradient of this expression, we find that

$$\nabla V = \frac{Q}{4\pi\epsilon_o}\frac{\partial}{\partial r}\left(\frac{1}{r}\right) = -\frac{Q}{4\pi\epsilon_o r^2}\hat{\mathbf{a}}_r.$$

Next, taking the divergence, we obtain

$$\nabla \cdot \nabla V = \nabla^2 V = -\frac{Q}{4\pi\epsilon_o}\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \times \frac{1}{r^2}\right) = \frac{0}{r^2}$$

which is zero at all points $r = 0$, which it should be, since the point charge exists only at the origin.

To see if V satisfies Poisson's equation at the origin, let us integrate $\nabla^2 V = -(\rho_v)/(\epsilon_o)$ throughout a small spherical volume, centered about $r = 0$:

$$\int_{\text{Vol.}} \nabla^2 V\, dv = -\frac{1}{\epsilon_o}\int_{\text{Vol.}} \rho_v\, dv.$$

The integral on the right-hand side is simply the enclosed charge Q, so

$$\int_{\text{Vol.}} \nabla^2 V\, dv = -\frac{Q}{\epsilon_o}.$$

We cannot evaluate the integral in its present form, since $\nabla^2 V$ is undefined at $r = 0$. However, remembering that $\nabla^2 V = \nabla \cdot \nabla V$, we can write

$$\int_{\text{Vol.}} \nabla^2 V\, dv = \oint_S \nabla V \cdot \mathbf{ds},$$

where S is a small spherical surface centered about the origin. Using $\nabla V = -\frac{Q}{4\pi\epsilon_o r^2}\hat{\mathbf{a}}_r$ and $\mathbf{ds} = r^2 \sin\theta\, d\theta\, d\phi\, \hat{\mathbf{a}}_r$ we obtain

$$\int_{\text{Vol.}} \nabla^2 V\, dv = -\frac{Q}{4\pi\epsilon_o}\int_0^{2\pi}\int_0^{\pi} \sin\theta\, d\theta d\phi = -\frac{Q}{\epsilon_o},$$

which shows that V satisfies Poisson's equation at $r = 0$.

In the next chapter we will use Poisson's and Laplace's equations to solve a number of practical problems where little is known a priori about the complete charge distribution, but where the potential V is known on the surface bounding some region. Problems of this sort are called ***boundary value problems*** and are an important part of electrostatic analysis.

4-8 Summation

We started this chapter by specializing Maxwell's equations for static charge distributions in free space. For this case, only an electric field is produced, which is governed by Maxwell's two equations of electrostatics. From these, we were also able to derive Coulomb's law, which is an explicit equation for the E-field generated by any charge distribution. Using these relations, we were able to calculate the E-fields that are generated by a number of simple charge distributions.

We also found that electrostatic calculations can be accomplished by using the electric potential function. Since it is a scalar, the use of this potential greatly simplifies certain electrostatic problems. The potential is also directly related to energy and is easily measured.

In the next chapter, we will expand our electrostatic analysis by including the effects of material media on the fields generated by electric sources.

PROBLEMS

4-1 A point charge of value 2 [C] is located at the point (2, −1, 4) [m]. Calculate the electric field **E** at the point $P(-1, 2, 2)$ [m].

4-2 Point charges of value −0.1 [μC] and 3 [μC] are located at the points (2, − 1, 4) [m] and (−1, 2, 0) [m], respectively. Calculate the electric field **E** at the point $P(2, -1, 3)$ [m].

4-3 Figure P4-3 shows a line-charge ring with uniform charge density ρ_ℓ and radius a, lying in the $z = 0$ plane. Use Coulomb's law to find the electric field **E** at any point on the z-axis.

4-4 Repeat Problem 3 for the case where the line-charge density is $\rho_\ell = \rho_{\ell o} \cos \phi$ [C/m].

4-5 Figure P4-5 shows a uniform line-charge segment with charge density ρ_ℓ, with length $2h$, and centered along the z-axis. Find an expression for the E-field at any point $P(0, y, 0)$ on the y-axis. Compare this with the field of an infinite line charge (i.e., $h \to \infty$) by plotting $|\mathbf{E}|$ vs. y for both cases.

4-6 Use Equations (4.19) and (4.20) to find the range of heights above a uniform disk of charge with radius a where $|\mathbf{E}|$ is within 5% of the E-field of an infinite surface charge with the same charge density.

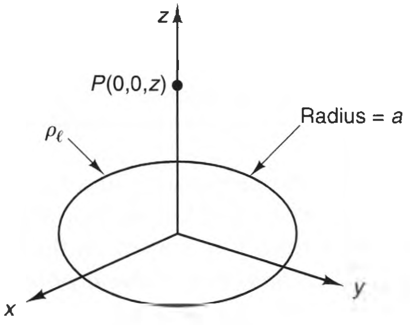


Figure P4-3

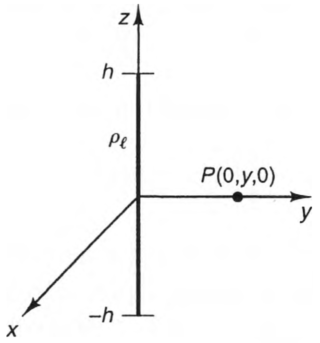


Figure P4-5

4-7 For an infinite line charge along the z-axis with uniform charge density ρ_ℓ,

(a) using either Equation (4.42) or Equation (4.48), show that the absolute potential function (i.e., referenced at infinity) is of the form

$$V = \frac{\rho_\ell}{2\pi\epsilon_0}[\ln(\infty) - \ln\rho].$$

(b) show that this potential function yields the correct E-field, using $\mathbf{E} = -\nabla V$.

(c) speculate as to why this potential has an infinite value for all finite values of ρ.

(d) To circumvent the problem of infinite absolute potentials, the relative potential

$$V' = -\frac{\rho_\ell}{2\pi\epsilon_0}\ln\rho$$

is often used. Show that this potential yields the same **E**.

4-8 Find the E-field at an arbitrary point (ρ, ϕ, z) generated by the charge distribution $\rho_v = (1/\rho)\,e^{-\rho}$ [C/m^3], where ρ is measured in meters. (*Hint:* Is there any symmetry here that can be exploited?)

4-9 For the potential function $V(x, y) = V_1 \sin(\pi x/a)\sinh(\pi y/a)$ [V],

(a) show that this potential satisfies Laplace's equation.

(b) find the E-field that follows from this potential.

4-10 Calculate the potential difference V_{ab} between two concentric spherical shells of charge. Assume that the inner shell has radius a and charge Q, and the outer shell has radius b and charge $-Q$.

4-11 Use Gauss's law to find the E-field generated by an infinite line charge with charge density ρ_ℓ.

4-12 Use Gauss's law to find the E-field generated by an infinite sheet of charge with uniform surface charge density ρ_s.

4-13 A spherically symmetric charge distribution has a charge density

$$\rho_v = \rho_{vo}\frac{e^{-r}}{r} \qquad [\text{C/m}^3],$$

where r is measured in meters.

(a) Use Gauss's law to determine **E** at any point.

(b) What form does **E** have in the limit as $r \to \infty$? Does this make sense? Why or why not?

4-14 If measurements have shown that the E-field in an all of space is

$$\mathbf{E} = \left[\frac{1}{2}\ln r - \frac{1}{4}\right]\hat{\mathbf{a}}_r \quad [\text{V/m}],$$

find the charge distribution that generated this E-field.

4-15 If it is known that an electric dipole is located at the origin and that its E-field at the point $P(2[\text{m}], 30°, 90°)$ is $\mathbf{E} = 4\hat{\mathbf{a}}_\theta$ [V/m], find the magnitude and direction of the dipole moment **p**.

4-16 If point charges of value +2 [nC] and −3 [nC] are located at the points $P_1(2, 2, -1)$[m] and $P_2(1, -2, 1)$[m], respectively, find the voltage V_{ab} between points $P_a(2, 3, 1)$[m] and $P_b(-1, 2, 2)$[m] by using the superposition principle and the potential expression for point charges (Equation (4.44)).

4-17 Find the expression for the E-field at any point along the z-axis due to the solid, uniformly charged circular cylinder shown in Figure P4-17. Assume that the cylinder has charge density ρ_v, has radius a, and extends between $z = -h/2$ and $z = h/2$. Derive the expression for the E-field by treating the cylinder as an infinite stack of circular disks.

4-18 Two infinite sheets of charge lie parallel to each other, separated by a distance d. The upper and lower sheets have surface charge densities ρ_{sa} and ρ_{sb} [C/m^2], respectively. Find the voltage V_{ab} from the top to the bottom surface.

4-19 Calculate the E-field generated at all points (r, θ, ϕ) by two concentric spheres of surface charge that are centered at the origin. Assume that inner and outer spheres have radii r_a and r_b and surface charge densities ρ_{sa} and ρ_{sb}, respectively.

4-20 Prove that for a point charge Q at the origin, the integral form of Coulomb's law (Equation (4.14)) collapses to the familiar form

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_o}\frac{Q}{r^2}\hat{\mathbf{a}}_r,$$

by

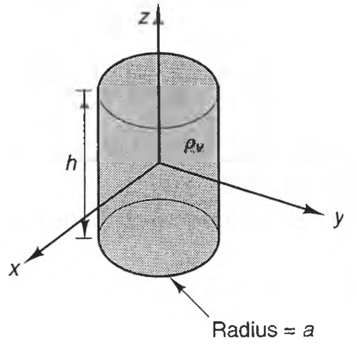


Figure P4-17

(a) showing that $\rho_v(\mathbf{r}) = Q\,\delta(x)\,\delta(y)\delta(z)$ [C/m^3] represents a point charge at the origin of value Q, where $\delta(x) = 0$ for $x \neq 0$ and $\int_{-\infty}^{\infty} \delta(x)\,dx = 1$.

(b) substituting $\rho_v(\mathbf{r})$ of part a) into Equation (4.12) and integrating.

4-21 Prove that $\nabla \dfrac{1}{|\mathbf{r} - \mathbf{r}'|} = -\dfrac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3}$ by expanding **r** and **r**′ in Cartesian coordinates and performing the gradient operation.

5

Electrostatic Fields In Material Media

5-1 Introduction

Our discussion of electric fields thus far has treated charged particles as if they were somehow suspended in a vacuum, with no apparent means of support. Although there are times when this is indeed the case (such as the electron beam inside a cathode-ray tube), the vast majority of engineering applications of electromagnetics involve materials, chosen for their electrical, magnetic, or mechanical properties. In fact, most of the advances in electrical engineering have been the result of the discoveries of new materials and the interactions these materials have with electromagnetic fields.

Materials interact with electric fields because they are composed, in part, of charged particles. When subjected to an electric field, these charges experience electric forces that cause them to move. How the charges move depends upon the nature of the material. Some materials, such as metals, possess charges that are free to move about the material as conduction currents. These materials are called ***conductors***. ***Dielectrics***, on the other hand, are composed of charges that are tightly bound to individual nuclei. These charges can move only small distances, but they can generate secondary electric fields that can substantially alter the total electric field, both inside and outside the material.

In this chapter, we will discuss the ways in which electric fields interact with material media. We will start this discussion by identifying the relationship between electric fields and currents in conducting materials. This will allow us to calculate the resistance of simple devices and develop Kirchhoff's laws for dc circuits. We will then discuss the relationship between the bound charges and electric fields in dielectric materials.

The last part of the chapter is devoted to the development of solution techniques for finding the electrostatic fields generated by systems that contain constant potential surfaces. Problems of this sort are called ***boundary value problems***, and we will discuss how they can be solved using analytical, graphical, and numerical techniques.

5-2 Conductors

When some of the electrons in a material are free to move from molecule to molecule, a ***conduction*** (or ***drift***) ***current*** will flow when an electric field is applied to the material. These free charges can be electrons, holes, or ions. For most materials, the current density **J** and the electric field **E** are related by

$$\mathbf{J} = \sigma \mathbf{E} \qquad [\mathrm{A/m^2}], \tag{5.1}$$

where σ is called the ***conductivity***, which is measured in units of siemens per meter [S/m] or inverse ohms per meter $[\Omega^{-1}\mathrm{m}^{-1}]$. Equation (5.1) is called a ***constitutive equation***, because it relates **J** and **E** in terms of the material-dependent parameter, called a ***constitutive parameter***. This equation is also called the point form of ***Ohm's law***.

The conductivity of a material is a direct indication of the ease with which its free charges can be moved by an electric field. Materials with large conductivites are considered good conductors, or simply conductors, and those with small conductivities are called poor conductors, or insulators. The values of σ for a number of materials are given in Table C-2 of Appendix C.

We can derive an expression for the conductivity of a material in terms of the properties of its free-charge carriers. To accomplish this, we first remember from Equation (3.21) that the current density at a point can be expressed in terms of the charge densities ρ_{vi} and drift velocities $\mathbf{u}_i$ of the different charge carrier types; that is,

$$\mathbf{J} = \sum_{i}^{m} \rho_{vi} \mathbf{u}_i, \tag{5.2}$$

where m is the total number of different types of charge carriers present in the material. For simple materials, the drift velocity $\mathbf{u}_i$ of each type of charge carrier is related to **E** by

$$\mathbf{u}_i = \pm \mu_i \mathbf{E}, \tag{5.3}$$

where μ_i is the mobility of the ith charge distribution.[1] Mobility is measured in units of $[\mathrm{m^2 \cdot V^{-1} \cdot s^{-1}}]$. By convention, mobility is always a positive number; the plus sign is used in Equation (5.3) for charges that move parallel to **E** (typically, positive charges), and the minus sign is used for charges that move antiparallel to **E** (typically, negative

[1] By convention, the same symbol μ is used to to represent both mobility and permeability. Even so, it is usually easy to tell which meaning is intended in an expression by looking at the context of the expression.

charges).[2] Substituting Equations (5.2) and (5.3) into Equation (5.1), and solving for σ, we obtain

$$\sigma = \sum_{i}^{m} \pm \mu_i \rho_{vi}. \tag{5.4}$$

Example 5-1

At 300° K (80°F), the electron and hole mobilities in pure silicon are $\mu_n = 1350$ [cm^2/(V • s)] and $\mu_p = 480$ [cm^2/V • s], respectively. If the electron and hole densities are both $N = 1.5 \times 10^{10}$ [cm^{-3}], find the conductivity σ.

Solution:

From Equation (5.4), we have

$$\begin{aligned} \sigma &= (eN)(-\mu_n) + (-eN)(\mu_p) \\ &= (1.6 \times 10^{-19}[\mathrm{C}]) \times (1.5 \times 10^{10}[\mathrm{cm}^{-3}]) \times (1350 + 480)[\mathrm{cm}^2/\mathrm{V} \cdot \mathrm{s}]) \\ &= 4.39 \times 10^{-6}\,[\mathrm{S/cm}] = 4.39 \times 10^{-4}\,[\mathrm{S/m}]. \end{aligned}$$

This result compares well with the value given in Table C-2 of Appendix C.

Materials are often classed according to whether or not they are ***homogeneous***, ***linear***, or ***isotropic***. The definitions of these classes are:

1) A homogeneous material is one in which the constitutive parameters do not vary from point to point throughout the material.

2) A linear material is one whose constitutive parameters are not functions of the field strength or the current density.

3) An isotropic material is one that has no preferred directions. In the case of conductivity, this means that **J** and **E** are always collinear, and σ has the same value for all orientations of **E**.

Materials that are homogeneous, linear, and isotropic are called ***simple materials***.

5-2-1 RESISTANCE FORMULAS FOR SIMPLE CIRCUIT ELEMENTS

When a conducting material is placed between two constant potential surfaces (often called terminals), the relationship between the current I flowing through the element and the potential difference V_{ab} between the surfaces is governed by ***Ohm's law***,

$$V_{ab} = IR, \tag{5.5}$$

[2] A notable exception is the valence electrons in semiconductors, which have a negative effective mass and actually move in the same direction as **E**. See Ben G. Streetman, *Solid State Electronic Devices*, 3rd edition, Englewood Cliffs, NJ, 1990, (Prentice Hall), pp. 63–64.

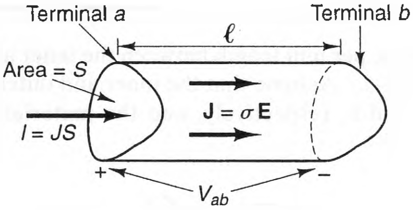


Figure 5-1 A homogeneous section of conducting material.

where R is the resistance of the element, measured in ohms $[\Omega]$.

To see how the resistance of an element is related to its dimensions and its conductivity, consider the simple element shown in Figure 5-1. Here, a homogeneous, cylinder of material has length ℓ, cross-sectional area S, and conductivity σ. If $\mathbf{E}$ and $\mathbf{J}$ are uniform throughout the element, the voltage between the two terminals is

$$V = -\int_b^a \mathbf{E} \cdot d\boldsymbol{\ell} = E\ell. \tag{5.6}$$

The current I flowing into the positive terminal can be found by integrating the current density $\mathbf{J} = \sigma\mathbf{E}$ over the cross section of the element:

$$I = \int_S \mathbf{J} \cdot \mathbf{ds} = \int_S \sigma\mathbf{E} \cdot \mathbf{ds} = \sigma E S. \tag{5.7}$$

Substituting Equations (5.6) and (5.7) into Equation (5.5), we find that the resistance of this simple element is

$$R = \frac{\ell}{\sigma S} \quad [\Omega]. \tag{5.8}$$

We can use Equation (5.8) to derive a general expression for the resistance R of an element in terms of the E-field alone. Using $V_{ab} = -\int_b^a \mathbf{E} \cdot d\boldsymbol{\ell}$, $I = \int_S \mathbf{J} \cdot \mathbf{ds}$ and $\mathbf{J} = \sigma\mathbf{E}$, we obtain

$$R = \frac{V}{I} = \frac{-\int_b^a \mathbf{E} \cdot d\boldsymbol{\ell}}{\int_S \sigma\mathbf{E} \cdot \mathbf{ds}} \quad [\Omega], \tag{5.9}$$

where a and b are the positive and negative terminals, respectively, and S is any cross-sectional surface of the element whose normal vector is parallel to the current flow. If the material has a linear conductivity, R is a function only of the material parameters and the dimensions of the element. Also, the conductance G of an element is defined as the reciprocal of its resistance, R:

$$G = R^{-1} = \frac{I}{V} \qquad [\Omega^{-1} \text{ or S}]. \tag{5.10}$$

Example 5-2

Determine the conductance per unit length between the inner and outer conductors of the coaxial cable shown in Figure 5-2. Assume that the inner and outer conductors are constant potential surfaces of radii a and b, respectively, and the material between the conductors has a constant conductivity σ.

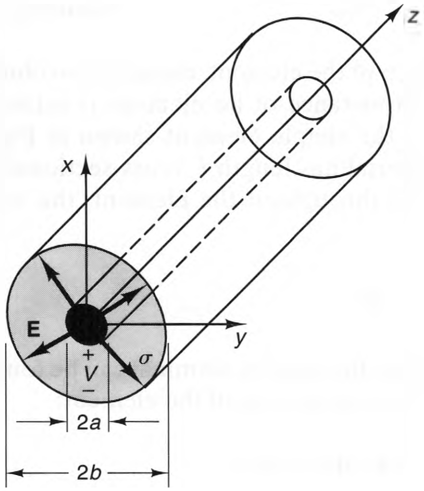


Figure 5-2 A coaxial cable, filled with a conducting dielectric.

Solution:

Even though a conducting medium lies between the inner and outer cylinders, the geometry still has perfect cylindrical symmetry. This means that the E-field has the same general form as a cylindrically symmetric charge distribution in free space. (See Equation (4.31).) If the conducting medium is homogeneous, **E** can be expressed as

$$\mathbf{E} = \frac{E_o}{\rho}\hat{\mathbf{a}}_\rho \quad a < \rho < b,$$

where E_o is a constant. Using this E-field expression, we find that the voltage between the inner and outer conductors is

$$V = -\int_b^a \frac{E_o}{\rho}\hat{\mathbf{a}}_\rho \cdot \hat{\mathbf{a}}_\rho d\rho = -\int_b^a \frac{E_o}{\rho} d\rho = E_o \ln\frac{b}{a} \qquad [\text{V}].$$

We can find the current per meter I that passes between the inner and outer cylinders by integrating $\mathbf{J} \cdot \mathbf{ds}$ on a circular cylinder of radius ρ and unit length that surrounds the inner conductor. Using $\mathbf{J} = \sigma\mathbf{E}$, we obtain

$$I = \int_0^{2\pi} \frac{\sigma E_o}{\rho}\hat{\mathbf{a}}_\rho \cdot \hat{\mathbf{a}}_\rho \rho \, d\phi = 2\pi\sigma E_o \qquad [\text{A/m}].$$

Substituting these expressions for V and I into Equation (5.10), we get

$$G = \frac{I}{V} = \frac{2\pi\sigma}{\ln\frac{b}{a}} \qquad [\text{S/m}].$$

5-2-2 METALS AND PERFECT CONDUCTORS

Metals have large conductivities, because they are composed of atoms (or molecules) that each contain at least one loosely bound electron. These electrons constitute a highly mobile ***electron cloud*** that can flow in response to a small electric field. Nonmetals can also exhibit large conductivities. A good example is saltwater, which has a large concentration of ions that can drift with an applied E-field.

When the conductivity of a material is high enough, it is often an excellent approximation to model it as a ***perfect conductor***, where $\sigma \rightarrow \infty$. Perfect conductors are particularly simple to model, since the E-field inside them is always zero. To see why, let us take the limit of $\mathbf{E} = \mathbf{J}/\sigma$ as $\sigma \rightarrow \infty$. Since **J** remains finite in a conductor, even when $\sigma \rightarrow \infty$, we obtain[3]

$$\mathbf{E} = \lim_{\sigma \to \infty} \frac{\mathbf{J}}{\sigma} = 0 \qquad \text{(perfect conductor).} \tag{5.11}$$

Hence, since $V = \int \mathbf{E} \cdot d\boldsymbol{\ell}$, the potential at each point on and within a homogeneous, perfect conductor is constant. Later in this chapter, we will show that the tangential electric field above the surface of a perfect conductor is also zero.

The conductivities of most metals vary inversely with temperature. This is because the average number of collisions experienced by the electrons increases with the thermal activity of the metal, which, in turn, slows the drift velocity of the electrons. For example, the conductivity of copper decreases roughly 0.4% for each 1°C increase in temperature.

Some metals, called ***superconductors***, exhibit zero resistivity when cooled below a certain ***critical temperature***. For instance, aluminum becomes perfectly conducting at temperatures below 1.18°K. This change of state happens very abruptly and can be explained only by quantum mechanics. Figure 5-3 shows the variation of the resistivity ($1/\sigma$) of a typical superconductor with temperature. The critical temperatures of metallic superconductors range from roughly 1.1–23°K, which means that their usefulness is limited to those situations where liquid helium (boiling temperature = 4°K) can be used as a coolant.

One application where the benefits of metal superconductors outweigh the problems associated with helium cooling is in the particle accelerators used in high-energy physics experiments. Here, enormous magnetic fields are required to confine the particles. These magnetic fields must be generated by large currents in coils, but the ohmic losses of ordinary wire limit the maximum B-fields that can be generated. However, superconducting coils made from niobium wires with a tin coating have a critical

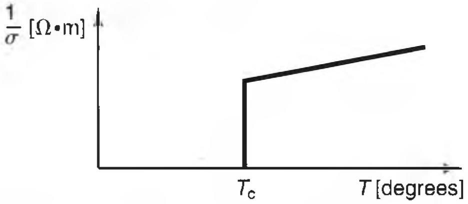


Figure 5-3 The resistivity of a superconductor as a function of temperature.

[3] Collisions between charge carriers limit drift velocities and currents, even in good conductors.

temperature of approximately 9°K and have been used successfully in particle accelerators for many years.

More recently, a new class of compounds has been discovered that exhibits superconductivity at much higher temperatures. Unlike the earliest superconductors, these materials are ceramics, not metals. An example is the compound $YBa_2Cu_3O_7$, which has a critical temperature of approximately 83°K. Ceramics are brittle, which means that they cannot be extruded into wires. But current research has indicated that thin layers of these materials can be deposited onto long, flexible tapes. In time, such superconducting tapes may be used routinely for a variety of applications.

5-2-3 KIRCHHOFF'S VOLTAGE LAW

We will now derive Kirchhoff's voltage law. In a sense, there is nothing new here, since the reader is probably familiar with this law from circuit theory. We will take the time to deal with it here because circuit analysis is rooted in electromagnetic theory. In particular, the laws of dc circuit analysis can be derived directly from the laws of electrostatics.

We will start by showing that a steady current cannot flow in a circuit if the only force field present is an electrostatic field. Figure 5-4 shows a circuit that consists of two resistors, R_1 and R_2, connected in a loop with a conducting wire. This circuit is subjected to an electrostatic field $\mathbf{E}$, generated by a static charge distribution. Since the E-field is conservative, $\oint_C \mathbf{E} \cdot \mathbf{d}\ell = 0$ around the circuit path. Using Ohm's law, we find

$$\oint_C \mathbf{E} \cdot \mathbf{d}\ell = I(R_1 + R_2) = 0, \tag{5.12}$$

from which we conclude that $I = 0$. Thus, no steady current can flow in a circuit when it is subjected only to an E-field produced by a static charge distribution.

Having proved that a steady current cannot be supported in a circuit via electrostatic forces alone (or any other conservative force field, for that matter), it is logical to conclude that the only way a steady current can be induced in a circuit is for there to be a nonconservative force field present somewhere in the circuit. Sources of nonconservative force fields include:

1. Electric batteries, which produce chemical forces that act on electrons.
2. Magnetic induction, where forces are caused by time-varying magnetic fields or the movement of conductors in the presence of a magnetic field.
3. Thermocouples, which convert thermal energy to electric forces.

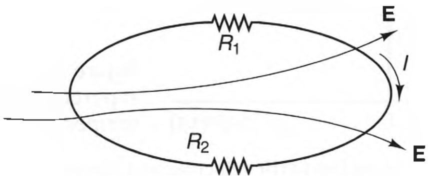


Figure 5-4 A passive circuit subjected to an electrostatic field.

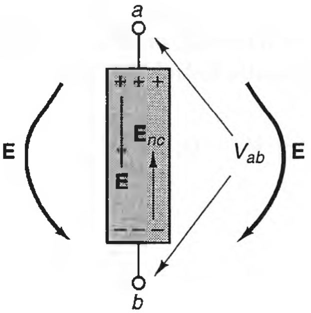


Figure 5-5 Cutaway view of a battery, showing the conservative electrostatic field **E** inside and outside the battery and the nonconservative field $\mathbf{E}_{\text{nc}}$ that exists only inside the battery.

4. Photovoltaic cells, which convert light energy to electric forces within semiconductors.

To understand how these nonconservative forces can produce steady currents in a circuit, consider the schematic of a battery, shown in Figure 5-5. All batteries have an internal resistance, but we will assume that this resistance can be modeled as an external resistor when the battery is connected in a circuit. Inside the battery, chemical forces produce a nonconservative electric field $\mathbf{E}_{nc}$ on the free charges. Initially, this force causes positive and negative charges to congregate at the positive and negative terminals, a and b, respectively. This process continues until the resulting electrostatic E-field inside the battery exactly counterbalances the chemical force, resulting in a net zero force on the charges.

Unlike the chemically induced field $\mathbf{E}_{nc}$, which is present only inside the battery, the electrostatic field **E** is present both inside and outside the battery and gives rise to a voltage V_{ab} that can be measured between the terminals as

$$V_{ab} = -\int_b^a \mathbf{E} \cdot \mathbf{d\ell} = \int_b^a \mathbf{E}_{nc} \cdot \mathbf{d\ell}.$$

In this expression, the integration involving **E** can take place on a path either inside or outside the battery, whereas the integration involving $\mathbf{E}_{nc}$ must take place inside the battery, since $\mathbf{E}_{nc} = 0$ outside the battery. As this voltage is caused by a nonconservative force, it is called an ***electromotive force*** (***emf***), specified in volts [V]. This name sometimes causes some confusion with units, but it emphasizes the fact that the power required to drive a steady current in a circuit must come from a force that is not electrostatic.

Now that we have shown that batteries (or any other dc source) produce an electrostatic field and a voltage between their terminals, it is simple to derive Kirchhoff's law for lumped-element, dc circuits. Consider the circuit shown in Figure 5-6, which

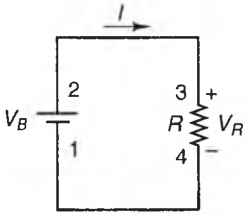


Figure 5-6 A simple circuit for deriving Kirchhoff's voltage law.

consists of a battery with voltage V_B and a lumped resistor R, connected by perfectly conducting wires. For a closed, clockwise path C around the circuit, the conservative property of the electrostatic E-field gives us

$$\oint_C \mathbf{E} \cdot d\boldsymbol{\ell} = -V_B + V_R = 0,$$

where

$$V_B = -\int_1^2 \mathbf{E} \cdot d\boldsymbol{\ell}$$

and

$$V_R = -\int_4^3 \mathbf{E} \cdot d\boldsymbol{\ell}.$$

Also, from Ohm's law, we have $V_R = IR$, and the preceding expression can be written as

$$V_B = IR. \tag{5.13}$$

Thus, if $V_B > 0$, $I > 0$, which agrees with the common notion that current flows out of the positive terminal of a battery when it is connected to a resistive load.

Finally, we can generalize Equation (5.13) for the case where N voltage sources and resistors are connected in series. If V_i is the voltage across the ith element while traversing the circuit in a particular direction (say, CW or CCW), we obtain the familiar circuit form of Kirchhoff's voltage law,

$$\sum_{i=0}^{N} V_i = 0. \tag{5.14}$$

Here we see that voltage drops across the sources are treated just the same as the voltages across the resistors.

5-2-4 OHMIC POWER DISSIPATION—JOULE'S LAW

It follows from the definition of potential difference that the energy dissipated when a charge dQ moves through a potential difference V is $dW = VdQ$. Since a current consists of a stream of charges, the power P that must be expended in order to maintain a steady current I through a lumped circuit element is

$$P = \frac{dW}{dt} = V\frac{dQ}{dt} = VI, \tag{5.15}$$

where V is the potential difference (voltage) across the resistor and I is defined into the most positive terminal of the element. When $P > 0$, the element is absorbing power, and when $P < 0$, the element is supplying power. When the element is a resistor, $V = IR$, and Equation (5.15) can be expressed in the form

$$P = I^2R = \frac{V^2}{R}. \tag{5.16}$$

This equation is the lumped-load version of ***Joule's law***.

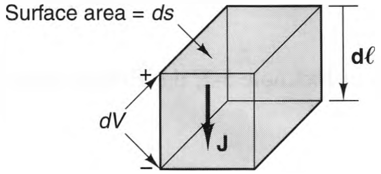


Figure 5-7 Geometry for deriving Joule's law.

We can also determine the power dissipated in an arbitrary volume of resistive material. To accomplish this, consider the differential volume shown in Figure 5-7. If this volume is oriented such that $\mathbf{J}$ is parallel to the displacement vector $\mathbf{d\ell}$, the current flowing through the volume is $dI = Jds$, and the voltage across the element is $dV = \mathbf{E} \cdot \mathbf{d\ell}$. Substituting this into Equation (5.15), we obtain

$$dP = Jds\, \mathbf{E} \cdot \mathbf{d\ell}.$$

Since $\mathbf{J}$ is parallel to $\mathbf{d\ell}$, $J\,\mathbf{d\ell} = \mathbf{J}d\ell$, so we can write dP as

$$dP = \mathbf{E} \cdot \mathbf{J}dv \qquad [\text{W}],$$

where $dv = d\ell ds$ is the volume of the differential element. Thus, $\mathbf{E} \cdot \mathbf{J}$ is the dissipated energy density, measured in units of watts per cubic meter [W/m^3]. The total power dissipated within a volume can be found by integrating both sides of this expression over the entire volume, yielding

$$P = \int_{\text{Vol.}} \mathbf{E} \cdot \mathbf{J}dv \qquad [\text{W}]. \tag{5.17}$$

This is the integral version of Joule's law. For isotropic media, $\mathbf{J} = \sigma\mathbf{E}$, so

$$P = \int_{\text{Vol.}} \frac{|\mathbf{J}|^2}{\sigma} dv \tag{5.18}$$

or

$$P = \int_{\text{Vol.}} \sigma|\mathbf{E}|^2 dv. \tag{5.19}$$

Media that dissipate power are called ***lossy media***. According to Equations (5.18) and (5.19), the conductivities of lossy media are greater than zero but less than infinity: $0 < \sigma < \infty$. There are two types of ***lossless media*** that do not dissipate power. The first are insulators, with $\sigma = 0$. Insulators are open circuits to current ($\mathbf{J} = 0$), so

$P = 0$. On the other hand, $\mathbf{E} = 0$ inside perfect conductors, yet $\mathbf{J}$ remains finite because of electron collisions. Thus, $P = 0$ in perfect conductors.

Example 5-3

Use Joule's law to calculate the power dissipated per meter in the coaxial resistor discussed in Example 5-2. Compare this result with the value obtained from the lumped form of Joule's law.

Solution:

From the solution of Example 5-2, the E-field between the conductors can be expressed in the form

$$\mathbf{E} = \frac{V}{\rho \ln \dfrac{b}{a}} \hat{\mathbf{a}}_\rho,$$

where a and b are the radii of the inner and outer conductors, respectively, and V is the voltage between the conductors. Substituting this into Equation (5.19) and integrating over a 1 [m] length, we obtain

$$P = \frac{\sigma V^2}{\left[\ln \dfrac{b}{a}\right]^2} \int_0^1 \int_0^{2\pi} \int_a^b \frac{1}{\rho^2} \rho \, d\rho \, d\phi \, dz = \frac{2\pi\sigma V^2}{\ln \dfrac{b}{a}} \qquad [\mathrm{W/m}].$$

Also from Example 5-2, the resistance per meter of this resistor is

$$R = \frac{1}{2\pi\sigma} \ln \frac{b}{a} \qquad [\Omega\text{•m}].$$

Substituting this into the lumped form of Joule's law (Equation (5.16)), we obtain

$$P = \frac{2\pi\sigma V^2}{\ln \dfrac{b}{a}},$$

which is the same result.

5-3 Dielectrics

Dielectrics contain charges that are tightly bound to individual nuclei. Although these ***bound charges*** can move only a fraction of an atomic distance away from their equilibrium positions, they can produce large charge imbalances throughout these materials that can significantly affect the total electric field, both inside and outside the dielectric material. In most dielectrics, the nuclei are unable to move because of molecular lattice forces, but the electron orbits can be distorted when an electric field is applied. Figure 5-8a is a simplified illustration of an atom (or molecule) when no externally applied electric field (which we will call a ***polarizing field***) is present. Here the negative electron cloud and the positive nucleus form two concentric spheres of charge that produce no net electric field outside the atom. In Figure 5-8b, the electron cloud is

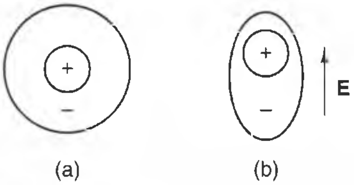


Figure 5-8 Simplified model of molecular charges. a) An unpolarized molecule. b) A polarized molecule.

distorted (or polarized) by a polarizing field. In this case, the positive and negative charge centers are offset and form an electric dipole with dipole moment **p**. This induced dipole generates an E-field that, when added to the polarizing field, changes the total field both inside and outside the dielectric.

The molecule depicted in Figure 5-8a is a ***nonpolar*** molecule, since it has no dipole moment when there is no polarizing field. Some molecules, such as water (H_2O), are ***polar***, meaning that their electron clouds are not symmetrically located about the positive nuclei.[4] Polar molecules produce a net electric field even when no polarizing E-field is present. Even so, the dipole moments of most polar materials tend to orient themselves randomly when no polarizing field is present, producing a macroscopic dipole moment of zero.

Some dielectrics exhibit a permanent, macroscopic dipole moment, even in the absence of a polarizing field. These materials, called ***electrets***, are the electrical analog of permanent magnets. Electrets contain polar molecules that, when heated, can be aligned by a polarizing field. As they cool, these molecules are locked in the aligned state, giving rise to a permanent dipole moment.

Piezoelectric materials exhibit time-varying dipole moments and voltages when they are subjected to a time-varying mechanical stress. This effect, which can also act in reverse, is used in devices such as microphones and ultrasonic transmitters. Many materials that exhibit the piezoelectric effect also exhibit the ***pyroelectric effect***, where macroscopic time-varying dipole moments and voltages can be induced by sudden temperature changes. Pyroelectric detectors are often used for measuring the output powers of high-infrared lasers, such as carbon dioxide (CO_2) lasers.[5]

5-3-1 DIELECTRIC SUSCEPTIBILITY

The electric dipoles formed throughout a dielectric are discrete, but from a macroscopic point of view, they appear to be continuously distributed. Thus, it is convenient to define the ***dipole moment per unit volume*** or ***polarization vector*** as

$$\mathbf{P} \equiv \lim_{\Delta v \to 0} \frac{\sum_{k=1}^{N\Delta v} \mathbf{p}_k}{\Delta v} \qquad [\mathrm{C/m^2}], \tag{5.20}$$

where N is the number of dipoles per unit volume and $\mathbf{p}_k = Q_k \mathbf{d}_k$ is the dipole moment of the kth dipole. Unlike the discrete quantity $\mathbf{p}_k$, $\mathbf{P}$ is much easier to work with, since it is a continuous function of position.

[4] The dipole moment of a water molecule has a magnitude of 6.15×10^{-30} [C • m].

[5] See J. T. Verdeyen, *Laser Electronics*, 2d ed. (Englewood Cliffs, NJ: Prentice Hall, 1989).

In simple (i.e., linear, homogeneous, and isotropic) dielectric media, **P** is always proportional to **E**; thus,

$$\mathbf{P} = \epsilon_o \chi_e \mathbf{E}, \tag{5.21}$$

where χ_e, called the ***electric susceptibility***, is a unitless constant of proportionality that is a measure of the ease with which dipoles can be formed in the dielectric. In inhomogeneous materials, the value of χ_e varies with position. In nonlinear media, χ_e is a function of the magnitude of **E**. Most dielectrics are linear when the polarizing fields are small, but some materials exhibit substantial nonlinearities when the polarizing fields are large.

Anisotropic dielectrics possess the property that **P** and **E** do not always point in the same direction. Crystalline solids often have this property, since their crystal lattices allow the electron clouds to stretch more easily in certain directions than in others. For these materials, the electric susceptibility χ_e must be represented as a matrix (often called a ***tensor***), since the product of a vector and a scalar cannot change the direction of the vector.

5-3-2 POLARIZATION CHARGE DISTRIBUTIONS IN DIELECTRICS

To find the relationship between a bound charge distribution in a dielectric and the polarization **P**, let us start by recalling from Equation (4.57) that the potential function V_k for a single dipole of moment $\mathbf{p}_k$ is given by

$$V_k = \frac{\mathbf{p}_k \cdot (\mathbf{r} - \mathbf{r}')}{4\pi\epsilon_o |\mathbf{r} - \mathbf{r}'|^3},$$

where the position vectors **r** and **r**′ represent field and source points, respectively. For a volume filled with dipoles, such as that shown in Figure 5-9, the resulting potential function can be found by replacing **p** with **P**dv' and integrating throughout the volume, to obtain

$$V = \frac{1}{4\pi\epsilon_o} \int_{\text{Vol.}} \mathbf{P} \cdot \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} dv', \tag{5.22}$$

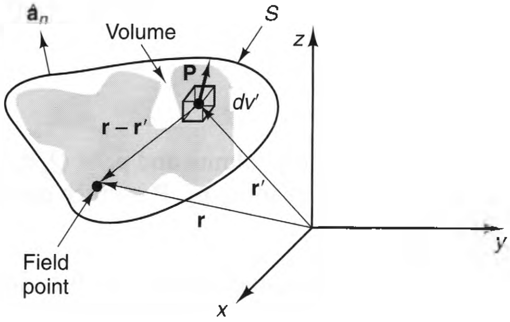


Figure 5-9 Geometry for determining the potential field of a charge distribution in terms of the polarization vector.

where **P** is a function of the primed coordinates. In the discussion that follows, we will show that the polarization **P** within a dielectric gives rise to charge distributions on and within the dielectric. These are called ***polarization charge distributions***.

We can manipulate Equation (5.22) into a form that makes the relationship between **P** and the induced volume and surface charge densities on and within the dielectric more obvious. This can be accomplished by first noting that the second term in the integrand can be written as

$$\frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3} = \nabla' \frac{1}{|\mathbf{r}-\mathbf{r}'|},$$

where the notation ∇' indicates that the derivatives associated with the del operator are with respect to the primed coordinates. The proof of this identity is straightforward when **r** and **r**′ are expressed in Cartesian coordinates. Substituting, we can write

$$V = \frac{1}{4\pi\epsilon_0}\int_{\text{Vol.}} \mathbf{P}\cdot\nabla' \frac{1}{|\mathbf{r}-\mathbf{r}'|}\,dv'\,.$$

Also, using Equation (B.3) of Appendix B, we can write the foregoing integrand as

$$\mathbf{P}\cdot\nabla'\frac{1}{|\mathbf{r}-\mathbf{r}'|} = -\frac{\nabla'\cdot\mathbf{P}}{|\mathbf{r}-\mathbf{r}'|} + \nabla'\cdot\frac{\mathbf{P}}{|\mathbf{r}-\mathbf{r}'|}\,.$$

Substituting this identity into the integral, we obtain

$$V = -\frac{1}{4\pi\epsilon_0}\int_{\text{Vol.}} \frac{\nabla'\cdot\mathbf{P}}{|\mathbf{r}-\mathbf{r}'|}\,dv' + \frac{1}{4\pi\epsilon_0}\int_{\text{Vol.}} \nabla'\cdot\frac{\mathbf{P}}{|\mathbf{r}-\mathbf{r}'|}\,dv'.$$

Finally, the divergence theorem can be used to transform the second volume integral into a surface integral, yielding

$$V = \frac{1}{4\pi\epsilon_0}\int_{\text{Vol.}} \frac{-\nabla'\cdot\mathbf{P}}{|\mathbf{r}-\mathbf{r}'|}\,dv' + \frac{1}{4\pi\epsilon_0}\oint_S \frac{\mathbf{P}\cdot\hat{\mathbf{a}}_n}{|\mathbf{r}-\mathbf{r}'|}\,ds', \tag{5.23}$$

where $\hat{\mathbf{a}}_n$ is the outward surface normal to the surface S that bounds the volume.

To see what this potential expression tells us about the charge distribution on and within a polarized dielectric, let us compare it with the potential field generated by a charge distribution that contains a charge distribution ρ_v in a volume and surface-charge distribution ρ_s on the surrounding surface. Using Equations (4.46) and (4.47), we see that the electric potential function of such a charge distribution is given by

$$V = \frac{1}{4\pi\epsilon_0}\int_{\text{Vol.}} \frac{\rho_v\,dv'}{|\mathbf{r}-\mathbf{r}'|} + \frac{1}{4\pi\epsilon_0}\oint_S \frac{\rho_s}{|\mathbf{r}-\mathbf{r}'|}\,ds', \tag{5.24}$$

where ρ_v and ρ_s are both functions of the primed coordinates. Comparing the volume integral of Equation (5.24) with that of Equation (5.23), we can conclude that $-\nabla'\cdot\mathbf{P}$ represents a volumetric polarization charge density ρ_{vp}. Similarly, equating the integrands of the surface integrals, we can conclude that $\mathbf{P}\cdot\hat{\mathbf{a}}_n$ represents a polarization surface charge density ρ_{sp}. Thus, the volume and surface charge densities caused by **P** can be expressed as

$$\rho_{vp} = -\nabla \cdot \mathbf{P} \qquad [\mathrm{C/m^3}], \tag{5.25}$$

$$\rho_{sp} = \mathbf{P} \cdot \hat{\mathbf{a}}_n \qquad [\mathrm{C/m^2}], \tag{5.26}$$

In Equation (5.25), we have dropped the prime notation from the divergence operation to indicate ρ_{vp} as a function of unprimed coordinates. Also, the subscript "p" indicates that these are polarization charge distributions, resulting from the displacement of bound molecular charges.

Equation (5.26) is valid at the interface between a dielectric and free space. For the case where one dielectric is adjacent to another, this expression can be modified to take into account the polarization charge deposited on S from both sides of the interface, that is

$$\rho_{sp} = (\mathbf{P}_1 - \mathbf{P}_2) \cdot \hat{\mathbf{a}}_{12n} \qquad [\mathrm{C/m^2}], \tag{5.27}$$

where $\mathbf{P}_1$ and $\mathbf{P}_2$ are the values of **P** on both sides of the surface and $\hat{\mathbf{a}}_{12n}$ points from region 1 towards region 2.

According to Equations (5.25) and (5.26), a polarization charge distribution exists wherever the polarization **P** has a nonzero divergence inside a dielectric or is discontinuous across a dielectric interface. For instance, consider the uniformly polarized dielectric shown in Figure 5-10. Here, the charge density within the volume is zero (in a macroscopic sense), since the positive and negative charges of adjacent dipoles cancel at each interior point. This cancellation does not happen at the right- and left-hand faces, however, where positive and negative surface charge densities accumulate, respectively. Figure 5-11, on the other hand, depicts a situation in which a volume

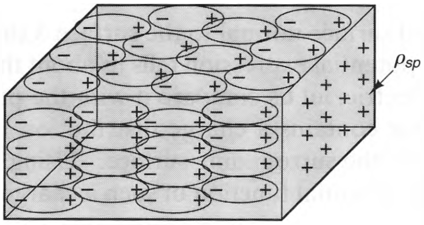


Figure 5-10 A uniformly polarized dielectric, showing the cancellation of charges everywhere inside the volume and a surface charge at the edges.

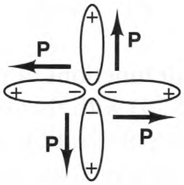


Figure 5-11 A polarized dielectric, where a point charge is created at a point where **P** has nonzero divergence.

charge distribution does exist inside a dielectric. Here, the molecular dipoles are aligned so that **P** diverges away from the point at the center, resulting in a negative polarization charge density there.

5-3-3 ELECTRIC FLUX DENSITY

The polarization charge distributions that are induced on and within dielectric materials generate secondary electric fields that must be accounted for in the design and analysis of electromagnetic systems. To accomplish this, let us start by considering the point form of Gauss' law inside a dielectric region, namely,

$$\nabla \cdot \epsilon_o \mathbf{E} = \rho_{vT} = \rho_v + \rho_{vp},$$

where ρ_v, ρ_{vp}, and ρ_{vT} are the free, polarization, and total charge densities, respectively. Remembering that $\rho_{vp} = -\nabla \cdot \mathbf{P}$, we can write

$$\nabla \cdot \epsilon_o \mathbf{E} = \rho_v - \nabla \cdot \mathbf{P},$$

which can be written in the form

$$\nabla \cdot (\epsilon_o \mathbf{E} + \mathbf{P}) = \rho_v.$$

Comparing the previous two equations, we notice that the right-hand side of the latter one is much simpler, since the free charge in most systems usually exists only on the surfaces of conductors. Thus, we now define the following new physical parameter called the electric flux density:

$$\mathbf{D} \equiv \epsilon_o \mathbf{E} + \mathbf{P} \qquad [\mathrm{C/m^2}]. \tag{5.28}$$

Using this definition, we obtain another form of Gauss' law[6]:

$$\nabla \cdot \mathbf{D} = \rho_v. \tag{5.29}$$

The right-hand side of this form of Gauss' law looks identical to the free-space form (Equation (4.7)), but the reader should note that ρ_v is the free-charge density, *not* the total-charge density.

The electric flux density **D** in most dielectrics is proportional to **E**, since **P** is itself proportional to **E**. By remembering that **P** and **E** are related by $\mathbf{P} = \epsilon_o \chi_e \mathbf{E}$, we can write Equation (5.28) in the form

$$\mathbf{D} = \epsilon_o(1 + \chi_e)\mathbf{E} \tag{5.30}$$

or

$$\mathbf{D} = \epsilon \mathbf{E} = \epsilon_o \epsilon_r \mathbf{E}, \tag{5.31}$$

[6] For the remainder of this text, the symbols ρ_v, ρ_s, and ρ_ℓ will represent *free* charge distributions.

where

$$\epsilon = \epsilon_o \epsilon_r = \epsilon_o(1 + \chi_e) \tag{5.32}$$

and

$$\epsilon_r = (1 + \chi_e). \tag{5.33}$$

Here, ϵ is a constitutive parameter called the ***permittivity*** of the medium and ϵ_r is the ***relative permittivity*** (or ***dielectric constant***). By substituting Equation (5.32) into Equation (5.21), we can also write the relationship between **P** and **E** as

$$\mathbf{P} = \epsilon_o(\epsilon_r - 1)\mathbf{E}. \tag{5.34}$$

Table C-3 of Appendix C lists the values of ϵ_r for a number of materials commonly used in engineering applications.

5-3-4 DIELECTRIC STRENGTH

When a dielectric is subjected to a strong electric field, the forces on the electrons can reach the point where they are stripped away from the nuclei, ionizing some or all of the atoms in the dielectric. This phenomenon is called ***dielectric breakdown***. The minimum electric field intensity E_{min} at which it occurs is called the ***dielectric strength*** of the material. The dielectric strengths of a number of common materials are given in Table C-3 in Appendix C.

Dielectric breakdown can be either desirable or undesirable, depending on where it occurs. In the case of the zener diode, dielectric breakdown is desirable. As depicted in Figure 5-12a, electrons are liberated at a *pn* junction when the reverse-bias field exceeds the dielectric strength of the semiconductor. These electrons, in turn, are accelerated by the field and collide with other molecules, liberating even more electrons.

Once initiated, a chain reaction liberates a large concentration of free charge in the junction region. This charge is free to drift as a conduction current and results in a terminal current that changes rapidly with small changes in the reverse-bias voltage.

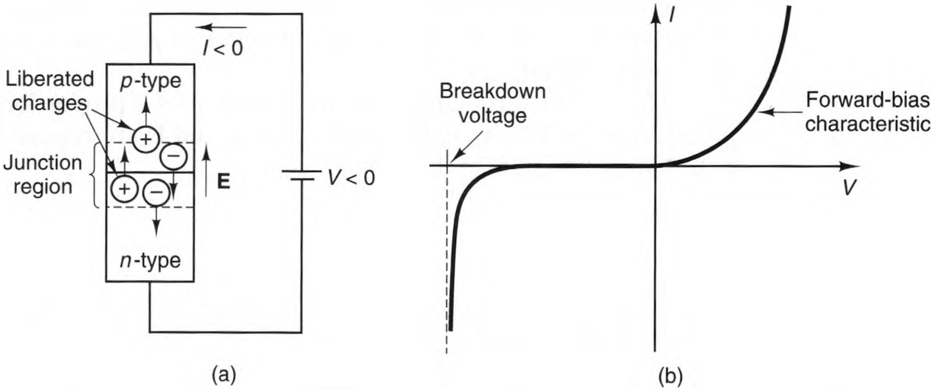


Figure 5-12 a) A circuit containing a semiconductor *pn* junction. b) The *V-I* curve for a typical *pn* junction.

Because of the nature of the process, this current is called an ***avalanche current***. The *V-I* curve of a typical zener diode[7] is shown in Figure 5-12b. As long as the avalanche current is kept within acceptable limits, zener diodes can act as effective voltage regulators. If, however, the reverse-bias current is not restricted by the external circuit, the avalanche process can quickly destroy the diode.

Lightning is a more dramatic example of dielectric breakdown. During a thunderstorm, two physical processes work in tandem to create large regions of charge within the clouds: the frictional forces associated with the high winds and the electrochemical reactions associated with condensing water vapor. When the electric field produced by these charges exceeds the dielectric strength of the air, an avalanche ionization process is initiated that creates a series of ***stepped leaders***, either between parts of the cloud (cloud-to-cloud lightning) or from the cloud to the ground (cloud-to-ground lightning). This is depicted in Figure 5-13. Each stepped leader is a relatively thin line of ionized air (i.e., a plasma). The leaders from the cloud are called "stepped" because they are created sequentially, in short segments (or steps). If a stepped-leader channel connects large pockets of opposite charge, a ***return stroke*** current is initiated. Not only are these return stroke currents extremely dangerous (with peak values in the range of millions of amperes), but they also create time-varying electric and magnetic fields that can interfere with communication systems.

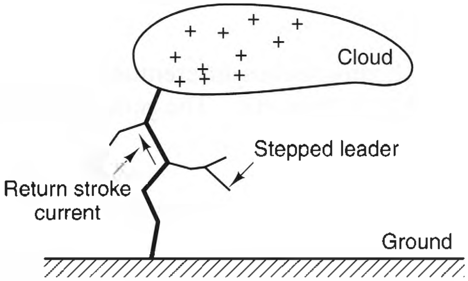


Figure 5-13 A cloud-to-ground lightning channel and stepped leaders.

5-3-5 DIELECTRIC RELAXATION

When free charge is injected into a material medium, the medium reacts to counteract the resulting charge imbalance. This reaction is called the ***relaxation process*** and involves both the conduction and dielectric properties of the medium.

To understand the relaxation process, let us suppose that free charge has somehow appeared in a homogeneous dielectric. This could occur by injecting free charge into the medium or as the result of the random, thermal motion of the molecular charges. From Gauss' law (which is valid for both static and time-varying sources), we know that

$$\nabla \cdot \mathbf{D} = \nabla \cdot \epsilon \mathbf{E} = \rho_v,$$

[7] Some zener diodes have this *V-I* characteristic because of the ***Zener effect***, but most zener diodes actually utilize the avalanche breakdown effect.

where ρ_v is the volumetric, free-charge density within the dielectric. Since $\mathbf{J} = \sigma \mathbf{E}$, the preceding expression can be written as

$$\nabla \cdot \frac{\epsilon}{\sigma} \mathbf{J} = \rho_v.$$

When the medium is homogeneous, both ϵ and σ are constants and can be taken out of the divergence operator, yielding

$$\nabla \cdot \mathbf{J} = \frac{\sigma}{\epsilon} \rho_v.$$

But from the law of charge conservation (Equation (3.25)), we also know that $\mathbf{J}$ and ρ_v are related by the expression

$$\nabla \cdot \mathbf{J} = -\frac{\partial \rho_v}{\partial t}.$$

Equating these two expressions for $\nabla \cdot \mathbf{J}$, we obtain

$$\frac{\partial \rho_v}{\partial t} + \frac{\sigma}{\epsilon} \rho_v = 0,$$

which is a linear, first-order, homogeneous differential equation that describes the charge density ρ_v at each point within the material. The general solution at any point P is

$$\rho_v(t) = \rho_v(0)\, e^{-t/\tau} \qquad t > 0, \tag{5.35}$$

where $\rho_v(0)$ is the volume charge density at $t = 0$ and

$$\tau = \frac{\epsilon}{\sigma} \qquad [\mathrm{s}] \tag{5.36}$$

is called the ***relaxation time constant.***

In words, Equation (5.35) states that after an initial disturbance, the volume free-charge density at each point within a homogeneous dielectric decays exponentially at a rate dictated by the relaxation time constant. Where does the excess charge go? The answer is, to the surface of the dielectric, where it appears as a surface charge distribution. For good conductors, τ is extremely small. In copper, $\tau = 1.5 \times 10^{-19}$ [s]. Insulators, on the other hand, can have long relaxation times. Porcelain, for instance, has a relaxation time of $\tau = 252$ [s].

Materials with positive relaxation times are called passive materials, since they tend to damp out charge disturbances, such as thermally induced disturbances. Some materials, however, can be made to exhibit negative conductivities when they are properly biased by an external source (such as a battery). These are called active materials. In this case, the relaxation time constant is negative, so thermally induced charge fluctuations in the material are amplified, rather than damped out. This phenomenon occurs in Gunn diodes, which are often used as microwave oscillators and amplifiers.[8]

Not only does the free-charge density inside a homogeneous dielectric approach zero after an initial disturbance, but the same is true for the polarization charge density ρ_{vp}. To show this, we note from Equation (5.25) that

$$\rho_{vp} = -\nabla \cdot \mathbf{P},$$

where, using Equations (5.31) and (5.34), we can write

$$\mathbf{P} = \epsilon_o(\epsilon_r - 1)\mathbf{E} = \left(\frac{\epsilon_r - 1}{\epsilon_r}\right)\mathbf{D}.$$

If the dielectric is homogeneous, ϵ_r is a constant, and we can write

$$\rho_{vp} = -\nabla \cdot \left(\frac{\epsilon_r - 1}{\epsilon_r}\mathbf{D}\right) = -\frac{1}{\epsilon_r}(\epsilon_r - 1)\nabla \cdot \mathbf{D} = -\frac{1}{\epsilon_r}(\epsilon_r - 1)\rho_v,$$

since $\nabla \cdot \mathbf{D} = \rho_v$. Using $\rho_v(t) = \rho_v(0)\, e^{-t/\tau}$ (from Equation (5.35)), we finally obtain

$$\rho_{vp} = -\frac{1}{\epsilon_r}(\epsilon_r - 1)\rho_v(0)\, e^{-t/\tau}. \tag{5.37}$$

Hence, we see that the volumetric polarization charge density ρ_{vp} also decays exponentially to zero with the same rate as does the volumetric free-charge density inside a homogeneous dielectric.

We can summarize the conclusions of Equations (5.35) and (5.37) by the following statement:

Under normal circumstances, the steady-state charge density inside a homogeneous dielectric region is zero.

This is true even if current flows in the dielectric, such as in the case of a resistor. On the other hand, it *is* possible for there to be a surface charge distribution at the interface between two dissimilar dielectrics. If one or both regions are conducting (i.e., $\sigma \neq 0$), this surface charge can be made up of free charge, polarization charge, or both, depending upon whether a current is flowing between materials. If both materials are insulators, only a polarization surface charge can exist at the interface.

There are two major exceptions to the preceding statement, where a homogeneous dielectric region can support a volumetric charge, often called a ***space charge***. The first is when charge is injected into free space, such as occurs in a vacuum tube. The other is when the charges are subjected to quantum forces, in addition to electric forces. This occurs at semiconductor *pn* junctions and results in opposite space charge regions on both sides of the junction.

[8] See S. Y. Liao, *Microwave Devices and Circuits*, 3d ed. (Englewood Cliffs, NJ: Prentice Hall, 1990).

5-3-6 FIELD EQUATIONS IN DIELECTRICS

Now that we have determined how the bound charges in a dielectric distribute themselves in response to an applied field, we are in a position to derive the equations that model the relationship between electrostatic fields and their sources when dielectrics are present. We can start by remembering that for static charges in free space (i.e., a vacuum), the electric field intensity satisfies the following equations:

$$\left.\begin{aligned} \nabla \times \mathbf{E} &= 0 \\ \nabla \cdot \mathbf{E} &= \frac{\rho_v}{\epsilon_0} \end{aligned}\right\} \text{(Electrostatic equations in free space).}$$

These equations can also be used when dielectrics are present if all the charges are accounted for, both free and bound. For this case the curl equation is unchanged, and Gauss' law can be written as

$$\nabla \cdot \mathbf{E} = \frac{\rho_{vT}}{\epsilon_0}.$$

where the total charge density ρ_{vT} is the sum of the free and polarization charge densities, ρ_v and ρ_{vp}, respectively. Unfortunately, this equation is much more complicated than the free-space case, since the total charge density ρ_{vT} depends on the distribution of the bound charges, which depends upon **E** itself.

We can simplify matters considerably by remembering that Gauss' law can also be written in terms of **D** as

$$\nabla \cdot \mathbf{D} = \rho_v.$$

where ρ_v is the free-charge density. The right-hand side of this equation is much simpler than the corresponding E-field equation, because more is known a priori about the free charge in a device or system than is known about the total-charge density. For example, free charge usually exists only on the surface of conductors. Also, the free-charge density on and within perfect dielectrics is usually zero, since perfect dielectrics are insulators.

Because of this simplification, Maxwell's equations for electrostatics are best written in the form

$$\left.\begin{aligned} \nabla \times \mathbf{E} &= 0 \\ \nabla \cdot \mathbf{D} &= \rho_v \end{aligned}\right\} \text{(Electrostatic equations in dielectrics).} \qquad \begin{aligned} &(5.38) \\ &(5.39) \end{aligned}$$

Since both **E** and **D** appear in these equations, we also need the constitutive relation

$$\mathbf{D} = \epsilon \mathbf{E}. \tag{5.40}$$

Taken as a set, these equations are sufficient to model all electrostatic fields in dielectrics. Equations (5.38) and (5.39) can also be expressed in integral form, as

$$\left.\begin{aligned}\oint_C \mathbf{E}\cdot d\boldsymbol{\ell} &= 0 \qquad (5.41)\\ \oint_S \mathbf{D}\cdot d\mathbf{s} &= Q \qquad (5.42)\end{aligned}\right\} \text{(Electrostatic equations in dielectrics),}$$

where C is a closed path and Q is the free charge enclosed by the closed surface S.

The electrostatic potential V can still be used when dielectrics are present. Since $\nabla \times \mathbf{E} = 0$ in both free space and dielectrics, we still have

$$\mathbf{E} = -\nabla V. \qquad (5.43)$$

To see what equation V satisfies in dielectrics, we can substitute Equations (5.40) and (5.43) into Equation (5.39), which yields

$$\nabla \cdot (\epsilon \nabla V) = -\rho_v.$$

Using the identity $\nabla \cdot (\psi \mathbf{A}) = \psi \nabla \cdot \mathbf{A} + \mathbf{A} \cdot \nabla \psi$, we can write this expression as

$$\nabla^2 V + \frac{1}{\epsilon}\nabla V \cdot \nabla \epsilon = -\frac{\rho_v}{\epsilon}. \qquad (5.44)$$

In a homogeneous dielectric, $\nabla \epsilon = 0$, so Equation (5.44) becomes

$$\nabla^2 V = -\frac{\rho_v}{\epsilon} \qquad \text{(Homogeneous dielectrics).} \qquad (5.45)$$

Furthermore, in homogeneous regions where $\rho_v = 0$, V satisfies Laplace's equation,

$$\nabla^2 V = 0 \qquad \text{(Charge-free, homogeneous dielectrics).} \qquad (5.46)$$

As we saw in Section 5-3-5, the volumetric charge density ρ_v within a homogeneous region normally decays to zero after a short relaxation time, unless free charge is somehow embedded in the dielectric by another physical process (such as occurs in a semiconductor *pn* junction, which will be discussed shortly).

Example 5-4

Calculate the E-field generated by a point charge placed in an infinite, homogeneous region with permittivity ϵ.

Solution:

For convenience, let us place the point charge at the origin, as shown in Figure 5-14.

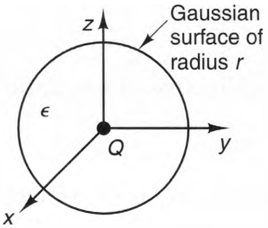


Figure 5-14 A point charge surrounded by a Gaussian surface.

Since this charge distribution has spherical symmetry, we can assume that $\mathbf{D} = D_r\hat{\mathbf{a}}_r$. For a spherical, Gaussian surface of radius r, Gauss' law becomes

$$\oint_S \mathbf{D} \cdot d\mathbf{s} = 4\pi r^2 D_r = Q.$$

Thus,

$$\mathbf{D} = \frac{Q}{4\pi r^2}\hat{\mathbf{a}}_r.$$

Finally, since $\mathbf{D} = \epsilon\mathbf{E}$, we have

$$\mathbf{E} = \frac{Q}{4\pi\epsilon r^2}\hat{\mathbf{a}}_r.$$

This example has shown that the E-field expression for a point charge in an infinite, homogeneous dielectric with permittivity ϵ is the same as for the same charge in free space, except that ϵ_o is replaced by ϵ. By the superposition principle, we can deduce that the same is true for *any* charge distribution in an infinite, homogeneous dielectric. Thus, Coulomb's law for a free-charge distribution ρ_v in an infinite, homogeneous dielectric with permittivity ϵ reads,

$$\mathbf{D} = \epsilon\mathbf{E} = \frac{1}{4\pi}\int_{\text{Vol.}} \rho_v \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3}\,dv' \quad \text{(Infinite, homogeneous dielectric).} \qquad (5.47)$$

In a similar way, the potential function generated by a free-charge distribution in an infinite, homogeneous dielectric can be written as

$$V = \frac{1}{4\pi\epsilon}\int_{\text{Vol.}} \frac{\rho_v}{|\mathbf{r} - \mathbf{r}'|}\,dv' \qquad \text{(Infinite, homogeneous dielectric).} \qquad (5.48)$$

Example 5-5

Calculate the E- and D-fields generated by the two concentric, uniformly charged spheres shown in Figure 5-15. The inner sphere has radius a and charge Q, and the outer sphere has radius b

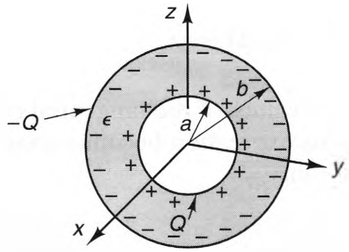


Figure 5-15 Concentric, oppositely charged spheres, separated by a uniform dielectric.

and charge $-Q$. Assume that the dielectric between the spheres is homogeneous and has permittivity ϵ.

Solution:

This problem has perfect spherical symmetry, so we can use Gauss' law. We have

$$\oint_S \mathbf{D} \cdot \mathbf{ds} = 4\pi r^2 D_r = Q_{\text{enc}},$$

where Q_{enc} is the charge enclosed by a Gaussian sphere of radius r. Because the charges on the spheres are balanced, $Q_{\text{enc}} = Q$ when $a < r < b$, and $Q_{\text{enc}} = 0$ otherwise. Thus, we obtain

$$\mathbf{D} = \epsilon \mathbf{E} = \begin{cases} \dfrac{Q}{4\pi r^2}\hat{\mathbf{a}}_r & a < r < b \\ 0 & \text{otherwise} \end{cases}.$$

5-3-7 DIELECTRIC BOUNDARY CONDITIONS

In order to calculate the E-fields when more than one kind of dielectric is present, one must know the ***boundary conditions*** that these fields exhibit across the material discontinuities. Such a situation is depicted in Figure 5-16, which shows the boundary surface between two dissimilar dielectric regions.

The behavior of the tangential components of **E** and **D** at the interface can be determined by using the conservative property of **E**, namely,

$$\oint_C \mathbf{E} \cdot \mathbf{d\ell} = 0.$$

Here we will choose the contour C shown in the figure that has depth Δh and length $\Delta\ell$, and that straddles the boundary between the two regions. In the limit as $\Delta h \to 0$, the contributions from the left and right portions of the path become negligible. Thus,

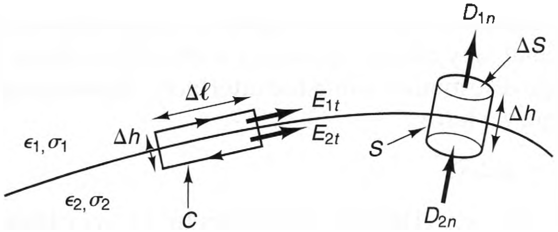


Figure 5-16 The surface and contour used to determine the boundary conditions at the interface between two dissimilar dielectrics.

$$\oint_C \mathbf{E} \cdot \mathbf{d\ell} \approx E_{1t}\,\Delta\ell - E_{2t}\,\Delta\ell \approx 0,$$

where E_{1t} and E_{2t} are the values of the tangential component of **E** just inside regions 1 and 2, respectively. This expression becomes exact in the limit as $\Delta\ell \rightarrow 0$. Dividing both sides by $\Delta\ell$, we obtain

$$E_{1t} = E_{2t}. \tag{5.49}$$

Also, since $\mathbf{D} = \epsilon\mathbf{E}$, we have $D_t = \epsilon E_t$. Substituting this into Equation (5.49), we find that

$$\frac{1}{\epsilon_1} D_{1t} = \frac{1}{\epsilon_2} D_{2t}, \tag{5.50}$$

or

$$\frac{D_{1t}}{D_{2t}} = \frac{\epsilon_1}{\epsilon_2}. \tag{5.51}$$

Thus, although the tangential components of **E** are always continuous across a dielectric interface, the tangential components of **D** are discontinuous when the permittivities differ.

The behaviors of the normal components of **E** and **D** at a dielectric interface can be found using Gauss' law,

$$\oint_S \mathbf{D} \cdot \mathbf{ds} = Q.$$

In this case, we will choose the surface S to be the "pillbox" surface shown in Figure 5-16, with height Δh and end-cap area ΔS. If we let $\Delta h \rightarrow 0$, the area of the cylindrical portion of the surface becomes zero, so the only contributions to the integral come from the bottom and top end-cap surfaces, and we have

$$\oint_S \mathbf{D} \cdot \mathbf{ds} \approx D_{1n}\,\Delta S - D_{2n}\,\Delta S \approx Q,$$

where D_{1n} and D_{2n} are the normal components of **D** in regions 1 and 2, respectively, each defined as extending from region 2 to region 1. Since the volume enclosed by this surface is infinitesimal, any charge contained within this volume can only be the result of a surface-charge distribution along the interface. Substituting $Q = \rho_s \Delta S$ into the preceding expression, we have

$$D_{1n}\,\Delta S - D_{2n}\,\Delta S \approx \rho_s \Delta S,$$

which becomes exact as $\Delta S \rightarrow 0$. Dividing both sides by ΔS, we obtain

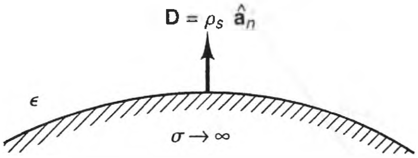

$\mathbf{D} = \rho_s \hat{\mathbf{a}}_n$
ϵ
$\sigma \rightarrow \infty$

Figure 5-17 The D-field at the interface between a dielectric and a perfect conductor.

$$D_{1n} - D_{2n} = \rho_s. \tag{5.52}$$

Also, since $D_n = \epsilon E_n$, Equation (5.52) can be written in terms of the normal components of **E**:

$$\epsilon_1 E_{1n} - \epsilon_2 E_{2n} = \rho_s. \tag{5.53}$$

When the media on both sides of the interface are insulators, ρ_s is typically zero. For this case, Equations (5.52) and (5.53) become

$$D_{2n} = D_{1n} \qquad (\rho_s = 0) \tag{5.54}$$

and

$$\epsilon_2 E_{2n} = \epsilon_1 E_{1n} \qquad (\rho_s = 0). \tag{5.55}$$

An important special case of these boundary conditions occurs when one of the regions is a perfect conductor. Such an interface is shown in Figure 5-17. For this case, both **E** and **D** are zero everywhere inside a perfect conductor, so Equations (5.51), (5.52), (5.54), and (5.55) yield:

$$\left.\begin{aligned} E_t = D_t &= 0 \\ D_n = \epsilon E_n &= \rho_s \end{aligned}\right\} \text{At the surface of a perfect conductor.} \qquad \begin{aligned}(5.56)\\(5.57)\end{aligned}$$

Thus, even though both components of **E** and **D** are zero inside a perfect conductor, only the tangential components must be zero on the surface. The condition $E_{\tan} = 0$ should not be surprising, since E-field lines are always perpendicular to surfaces of constant potential.

Example 5-6

Figure 5-18 shows the interface between two perfect dielectrics. Find the magnitude of $\mathbf{E}_2$ and angle it makes with the surface normal if the magnitude of $\mathbf{E}_1$ and the angle it makes with respect to the surface normal are both known.

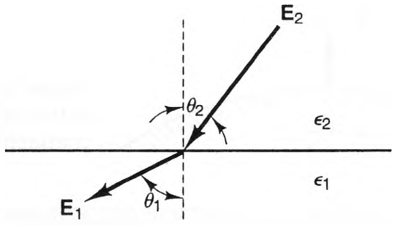


Figure 5-18 The E-field at the interface between two perfect dielectrics.

Solution:

The tangential and normal components of $\mathbf{E}_1$ are

$$E_{1_t} = E_1 \sin \theta_1$$

and

$$E_{1_n} = E_1 \cos \theta_1,$$

respectively. Using Equations (5.49) and (5.55), we see that the tangential and normal components of $\mathbf{E}_2$ just above the interface are given by

$$E_{2_t} = E_{1_t} = E_1 \sin \theta_1 = E_2 \sin \theta_2$$

and

$$E_{2_n} = \frac{\epsilon_1}{\epsilon_2} E_{1_n} = \frac{\epsilon_1}{\epsilon_2} E_1 \cos \theta_1 = E_2 \cos \theta_2,$$

where ϵ_1 and ϵ_2 are the permittivities of regions 1 and 2, respectively. Using these, we find that the magnitude and angle of $\mathbf{E}_2$ are

$$E_2 = E_1\left[1 + \left(\frac{\epsilon_1^2}{\epsilon_2^2} - 1\right) \cos^2 \theta_1\right]^{1/2} \tag{5.58}$$

and

$$\theta_2 = \tan^{-1}\left[\frac{\epsilon_2}{\epsilon_1} \tan \theta_1\right]. \tag{5.59}$$

From these expressions, we can conclude that if $\epsilon_2 > \epsilon_1$, then $\theta_2 > \theta_1$ and $E_2 < E_1$. Notice also that we always have $\theta_2 = \theta_1$ when $\theta_1 = 0$ or $\pi/2$, so E-field streamlines do not bend at a dielectric interface when they are parallel to or perpendicular to the interface.

At the interface between two media that have nonzero conductivities, the law of charge continuity places one more constraint on the fields. For the time-invariant case, $\frac{\partial Q}{\partial t} = 0$, so the continuity equation becomes

$$\oint_S \mathbf{J} \cdot \mathbf{ds} = 0.$$

Evaluating this integral around the pillbox surface shown in Figure 5-16 yields the following relationship between the normal components of **J** on each side of the interface:

$$J_{1n} = J_{2n}, \tag{5.60}$$

Here, both J_{1n} and J_{2n} are defined as pointing from region 2 to region 1. Also, since $\mathbf{J} = \sigma\mathbf{E}$ and $E_{1t} = E_{2t}$ across any interface, we have

$$\frac{J_{1t}}{\sigma_1} = \frac{J_{2t}}{\sigma_2}. \tag{5.61}$$

Finally, we can use Equations (5.52) and (5.60) to determine the surface charge density that is present at the interface between two conducting media. Substituting Equation (5.60) into Equation (5.52) and using $\mathbf{D} = \dfrac{\epsilon}{\sigma}\mathbf{J}$, we obtain

$$\rho_s = J_n\left(\frac{\epsilon_1}{\sigma_1} - \frac{\epsilon_2}{\sigma_2}\right) \quad \text{(interface between two conducting media),} \tag{5.62}$$

where J_n is the normal component of **J**, extending from region 2 to region 1.

5-4 Electrostatic Boundary Value Problems

Now that we have developed the differential equations and boundary conditions that govern the behavior of electrostatic fields in the presence of material media, we can discuss an important class of electrostatic problems in which little (or possibly nothing) is known a priori about the charge distribution. Instead, the potentials along one or more of the bounding surfaces are known. This type of situation occurs whenever dielectrics and conductors are present in a system and known voltages are impressed between the conductors. Problems of this sort are solved using Laplace's and Poisson's equations and are called ***electrostatic boundary value problems***.

There are three classes of methods used to solve electrostatic boundary value problems: analytical techniques, numerical techniques, and graphical techniques. Each method has its own advantages and disadvantages, and the choice of which to use usually depends upon the exact nature of the problem and the solution accuracy that is needed. Although a complete discussion of all of the electrostatic boundary value methods in common use is beyond the scope of this text, we will now present examples in each of the three major classes that show the kinds of solutions that can be obtained.

5-4-1 THE UNIQUENESS PRINCIPLE

The uniqueness principle states that there is one and only one solution to Laplace's and Poisson's equations for a given set of sources and boundary conditions. Although this may seem obvious, it is worth proving, since the proof shows us clearly what must be specified in order to make a solution unique.

The simplest way to prove the uniqueness principle is by proving that if an electrostatic system has two solutions, they must be identical. This kind of proof is called a ***proof by contradiction***. We start by assuming that V_1 and V_2 both satisfy Poisson's equation in a region that is bounded by a surface S where the potential is known. To keep things simple, we'll assume that the dielectric is uniform throughout the region, so V_1 and V_2 both satisfy the homogeneous Poisson's equation; that is,

$$\nabla^2 V_1 = -\frac{\rho_v}{\epsilon} \qquad \text{and} \qquad \nabla^2 V_2 = -\frac{\rho_v}{\epsilon},$$

where ρ_v is the charge density function throughout the region. Also, V_1 and V_2 satisfy the same boundary conditions on the surface S that bounds the volume V, so we have in addition,

$$V_1 = V_2 \qquad \text{everywhere on } S.$$

If we define V_d as the difference between V_1 and V_2, i.e.,

$$V_d = V_1 - V_2,$$

then V_d satisfies the conditions

$$\nabla^2 V_d = 0$$

and

$$V_d = 0 \text{ on } S.$$

To show that V_d must be a null solution (i.e., $V_d = 0$ at all points), let us consider the following volume and surface integrals that are related by the divergence theorem:

$$\int_{\text{Vol.}} \nabla \cdot [V_d \nabla V_d]\, dv = \oint_S [V_d \nabla V_d] \cdot \mathbf{ds} = 0.$$

(S is the surface surrounding the volume.) The surface integral is obviously zero, since $V_d = 0$ on S, so the volume integral must also be zero. From the vector identity

$$\nabla \cdot [V_d \nabla V_d] = V_d \nabla^2 V_d + \nabla V_d \cdot \nabla V_d = V_d \nabla^2 V_d + |\nabla V_d|^2$$

and the fact that $\nabla^2 V_d = 0$ throughout the volume, we can write

$$\int_{\text{Vol.}} \nabla \cdot [V_d \nabla V_d]\, dv = \int_{\text{Vol.}} |\nabla V_d|^2\, dv = 0.$$

Since $|\nabla V_d|^2$ cannot be negative, we must have $\nabla V_d = 0$ everywhere in the volume. This can happen only when V_d is a constant throughout the volume, so

$$V_d = V_1 - V_2 = \text{a constant everywhere in the volume.}$$

But $V_1 = V_2$ on the bounding surface S, so this constant must be zero; hence, we have

$$V_1 = V_2 \text{ everywhere in the volume and on } S,$$

and the proof is complete.

Example 5-7

Figure 5-19 shows a perfect conductor that completely encloses a source-free (i.e., $\rho_v = 0$) region. If the conductor is maintained at a potential V_o with respect to infinity, what are the potential and the E-field inside the enclosed region?

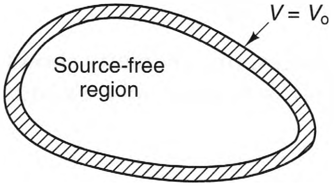


Figure 5-19 A source-free region surrounded by a constant-potential surface.

Solution:

Since the interior region is source free, the potential in the source-free region bounded by the surface satisfies Laplace's equation, $\nabla^2 V = 0$. Certainly, the potential function $V = V_o$ satisfies this equation. Since it also satisfies the boundary condition imposed by the conducting surface that surrounds the region, we can conclude from the uniqueness principle that the potential throughout the interior region is

$$V = V_o.$$

Also, using $\mathbf{E} = -\nabla V$, we find that the E-field throughout the interior region is

$$\mathbf{E} = 0.$$

Thus, we can conclude that the potential inside any source-free region that is surrounded by a constant-potential surface is itself constant, and the E-field is zero, *regardless of what sources exist outside the surface*. This is an example of ***electric shielding***.

5-4-2 ANALYTICAL SOLUTIONS

Analytical solutions are solutions that can be expressed mathematically, usually in terms of sums and products of simple functions. Only a small percentage of all electrostatic problems can be solved using analytical techniques. Nevertheless, they are important, because they are exact solutions which can provide insight into other, more complicated problems that cannot be solved analytically.

5-4-2-1 Some Simple Cases. Let us start by considering the geometry shown in Figure 5-20, which consists of two flat, perfectly conducting plates, separated by a perfect dielectric of permittivity ϵ and thickness d. The voltage between the plates is V_o.

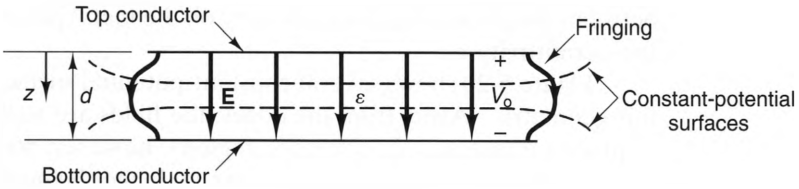


Figure 5-20 A parallel-plate capacitor.

Since the dielectric is homogeneous and insulating, the volume charge density ρ_v is zero and the potential V satisfies Laplace's equation. In Cartesian coordinates, Laplace's equation reads

$$\nabla^2 V = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = 0.$$

Away from the plate edges, this geometry shows variations only in the z direction, so we have $\partial V/\partial x = \partial V/\partial y = 0$, and Laplace's equation becomes

$$\frac{\partial^2 V}{\partial z^2} = 0.$$

The general solution of this differential equation is

$$V = C_1 z + C_2,$$

where C_1 and C_2 are constants.

The only boundary condition that must be enforced on this solution is that the potential difference between the plates is V_o. Hence, we require

$$V(z = 0) - V(z = d) = C_2 - C_1 d - C_2 = V_o,$$

which means that $C_1 = -V_o/d$. Thus, we obtain

$$V = -\frac{V_o}{d} z + C_2,$$

where C_2 can have any value, depending upon the choice of the reference potential. Since $\mathbf{E} = -\nabla V$, we have

$$\mathbf{E} = -\frac{\partial V}{\partial z}\hat{\mathbf{a}}_z = \frac{V_o}{d}\hat{\mathbf{a}}_z. \tag{5.63}$$

We can also find the free-charge density on the conducting plates by using the relationship between $\mathbf{D}$ and the free-charge density at the surface of a perfect conductor. Using Equation (5.57), we can express the surface charge density on the inner surfaces of the plates in terms of the normal component of $\mathbf{D}$ at the surface; that is,

$$\rho_s = \frac{\epsilon V_o}{d} \qquad [\text{C/m}^2], \tag{5.64}$$

where ρ_s is the surface charge density on the lower surface of the upper plate and $-\rho_s$ is the charge density on upper surface of the lower plate. Hence, for a fixed voltage V_o between the plates, the charge density on the metal plates is proportional to the dielectric permittivity ϵ.

Figure 5-20 shows several constant-potential surfaces and E-field streamlines for this geometry. Away from the edges, the fields are well predicted by our solution of Laplace's equation. In the edge regions, however, streamlines fringe, because the $(\partial^2 V)/(\partial x^2)$ and $(\partial^2 V)/(\partial y^2)$ terms can no longer be neglected. A full analytical solution in these regions is very cumbersome, but we will show later in this chapter how these fringing fields can be modeled graphically. It is also worth noting that because of the characteristics of the fields generated by infinite sheets of charge, the charge densities on the outer surfaces of the plates (i.e., the upper and lower surfaces of the top and bottom plates, respectively) must have the same sign in order to maintain $\mathbf{E} = 0$ inside the conductors. Hence, when the total charges on the upper and lower plates are opposite (as occurs when the plates are charged using a closed circuit), the charge on these outer surfaces is zero, and thus, the fields are negligible above and below the top and bottom plates, respectively, except near the edges.

Another simple geometry is depicted in Figure 5-21, which shows a coaxial line (cable) with inner radius a and an outer conductor of radius b. The region between the conductors is filled with a homogeneous dielectric with permittivity ϵ, and the voltage between the inner and outer conductors is V_o. In cylindrical coordinates, Laplace's equation reads

$$\nabla^2 V = \frac{1}{\rho}\frac{\partial}{\partial \rho}\left(\rho \frac{\partial V}{\partial \rho}\right) + \frac{1}{\rho^2}\frac{\partial^2 V}{\partial \phi^2} + \frac{\partial^2 V}{\partial z^2} = 0.$$

Since this cable has perfect cylindrical symmetry, it is reasonable to assume that V is a function only of the cylindrical coordinate ρ. Hence, for $a < \rho < b$, Laplace's equation becomes

$$\nabla^2 V = \frac{1}{\rho}\frac{\partial}{\partial \rho}\left(\rho \frac{\partial V}{\partial \rho}\right) = 0.$$

Integrating twice, we obtain

$$V = C_1 \ln \rho + C_2.$$

Since $V(\rho = a) - V(\rho = b) = V_o$, we have

$$C_1 = \frac{V_o}{\ln\left(\dfrac{a}{b}\right)} = -\frac{V_o}{\ln\left(\dfrac{b}{a}\right)}.$$

Thus,

$$V = -\frac{V_o}{\ln\left(\dfrac{b}{a}\right)} \ln \rho + C_2 \qquad a < \rho < b,$$

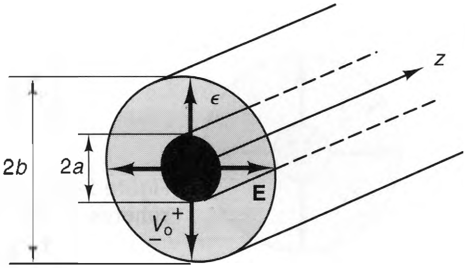


Figure 5-21 A coaxial cable.

where the value C_2 is determined by the reference potential (which was not specified). From $\mathbf{E} = -\nabla V$, the electric field intensity between the conductors is

$$\mathbf{E} = \frac{V_o}{\rho \ln\left(\frac{b}{a}\right)} \hat{\mathbf{a}}_\rho \qquad a < \rho < b. \tag{5.65}$$

We know from our discussion in the previous chapter on coaxial surface-charge distributions that the E-field beyond the outer surface will be zero if the charges per unit length on both cylinders are opposite. Laplace's equation gives us some insight into when this occurs. In this outer region, the general solution for V has the same form as it does in the interior region:

$$V = C_3 \ln \rho + C_4 \qquad (\rho > b).$$

If the absolute potential of the outer cylinder and infinity is zero, C_3 and C_4 are both zero (since the potential at infinity is zero), so $\mathbf{E} = 0$ everywhere for $\rho > b$. For this case, the E-field is confined to the interior of a coaxial cable, and the exterior region is ***shielded*** from the interior charges. In practice, we can accomplish this by connecting the outer conductor to a zero-potential surface. This practice is called ***grounding*** and is depicted in Figure 5-22.

Practically speaking, the best ground is usually an earth ground, since the potential of the earth is usually negligible. Another common scheme is to connect the outer conductor to a metal chassis. This is less optimum, however, since the chassis potential may well vary significantly from zero.

Another simple problem is shown in Figure 5-23, which depicts two hollow, concentric, conducting spheres with a homogeneous dielectric in between. The spheres have radii a and b, respectively, and the voltage between them is V_o. In spherical coordinates, Laplace's equation reads,

$$\nabla^2 V = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial V}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial V}{\partial \theta}\right) + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2 V}{\partial \phi^2} = 0.$$

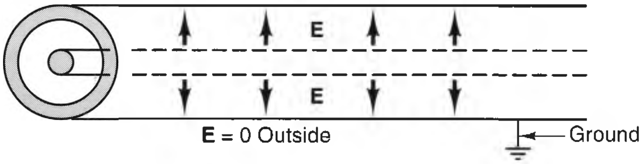


Figure 5-22 A grounded coaxial cable.

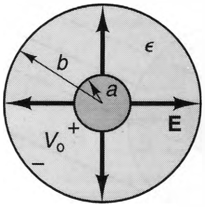


Figure 5-23 Concentric, conducting spheres, separated by a uniform dielectric.

Since the dielectric is homogeneous, the symmetry of this problem implies that V should vary only with r. Thus, Laplace's equation becomes

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial V}{\partial r}\right) = 0.$$

Multiplying both sides by r^2 and integrating twice with respect to r yields

$$V = \frac{C_1}{r} + C_2.$$

Because the potential difference between the inner and outer spheres is V_o, we have

$$V(r = a) - V(r = b) = C_1\left[\frac{1}{a} - \frac{1}{b}\right] = V_o,$$

so

$$V = \frac{V_o}{r\left[\frac{1}{a} - \frac{1}{b}\right]} + C_2 \qquad a < r < b.$$

Also, since $\mathbf{E} = -\nabla V$, and V is independent of θ and ϕ, we also have

$$\mathbf{E} = -\frac{\partial V}{\partial r}\hat{\mathbf{a}}_r = \frac{V_o}{r^2\left[\frac{1}{a} - \frac{1}{b}\right]}\hat{\mathbf{a}}_r \qquad a < r < b. \tag{5.66}$$

As in the case of the coaxial cable, the E-field will be zero for $r > b$ if the potential difference between the outer sphere and infinity is zero. Otherwise, $\mathbf{E}$ will decay in this region, proportional to r^{-2}.

5-4-2-2 Solutions Involving Inhomogeneous Dielectrics. As a group, geometries that contain inhomogeneous dielectrics are harder to analyze than those that have homogeneous dielectrics. This is because the inhomogeneous forms of Laplace's and Poisson's equations (Equation (5.44)) are more difficult to solve than their homogeneous counterparts. Even when the dielectric consists of a collection of homogeneous regions, the boundary conditions that the E-field must satisfy at the interfaces between these regions can be difficult to handle analytically.

There are, however, some problems with inhomogeneous dielectrics that are easy to analyze. These problems occur whenever the contours of constant dielectric permittivity lie either parallel to or perpendicular to E-field streamlines that would be present if the dielectric were homogeneous. For this case, the E-field streamline contours are not affected by the inhomogenieties, since E-field streamlines do not bend when they are directed either perpendicular to or parallel to a dielectric interface. (See Example 5-6.)

Figure 5-24 shows two such cases that are variations of the parallel-plate geometry we considered earlier. In Figure 5-24a, two perfect dielectrics are stacked between two conducting plates. In Figure 5-24b, two perfect dielectrics are placed side by side

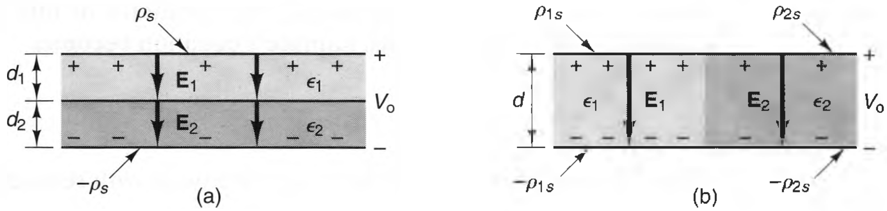


Figure 5-24 Two parallel-plate capacitors with inhomogeneous dielectrics.

between two conducting plates. Since the E-field for the homogeneous dielectric case is directed straight from the top plate to the bottom plate, neither of the inhomogeneities shown in Figures 5-24a and b will cause the E-field streamlines to bend, greatly simplifying the analysis.

To model the geometry shown in Figure 5-24a, we first note that since each dielectric is uniform, the potential function in each dielectric region will have the same form as it does for the uniform dielectric case. Hence, we can write

$$V_1(z) = C_a z + C_b$$
$$V_2(z) = C_d z + C_e,$$

where $V_1(z)$ and $V_2(z)$ are the potential functions in the upper and lower dielectric regions, respectively, and z is measured from the top plate. Because $\mathbf{E} = -\nabla V$, we find that $\mathbf{E}_1$ and $\mathbf{E}_2$ are both uniform vectors that are directed perpendicular to the plates.

To find the magnitudes of $\mathbf{E}_1$ and $\mathbf{E}_2$, we first require that these fields satisfy the necessary boundary condition (Equation (5.53)) at the dielectric interface; thus,

$$\epsilon_1 E_1 = \epsilon_2 E_2,$$

since $\rho_s = 0$ at the interface between two perfect dielectrics. Second, because the voltage between the plates is V_o, we must also have

$$V_o = d_1 E_1 + d_2 E_2 = \left[d_1 + \frac{\epsilon_1}{\epsilon_2} d_2\right] E_1 .$$

Solving this expression for E_1 and E_2, we obtain

$$E_1 = \frac{\epsilon_2 V_o}{\epsilon_1 d_2 + \epsilon_2 d_1}, \qquad E_2 = \frac{\epsilon_1 V_o}{\epsilon_1 d_2 + \epsilon_2 d_1}. \tag{5.67}$$

These expressions tell us that $\mathbf{E}$ is strongest in the dielectric with the smallest permittivity. Also, since $\mathbf{D} = \epsilon \mathbf{E}$, we see that $\mathbf{D}$ is the same in both regions:

$$D = D_1 = D_2 = \frac{\epsilon_1 \epsilon_2 V_o}{\epsilon_1 d_2 + \epsilon_2 d_1}.$$

Because $\rho_s = D_n$ at perfectly conducting surfaces, the free-charge densities on the inside portions of the upper and lower plates are opposite, with

$$\rho_s = \pm \frac{\epsilon_1 \epsilon_2 V_o}{\epsilon_1 d_2 + \epsilon_2 d_1}, \tag{5.68}$$

where the positive sign is used for the upper plate and the negative sign is used for the lower plate. Also, just as in the case where the entire dielectric is homogeneous, the charge density on the outer surfaces of the plates is negligible, except near the edges.

To find the fields present in the geometry shown in Figure 5-24b, we first note that because **E** is tangent to the interface in both regions, the dielectric boundary conditions require that **E** be continuous across the interface. Thus, it is reasonable to assume that the E-fields are the same in both dielectrics, with

$$E_1 = E_2 = \frac{V_o}{d}, \tag{5.69}$$

where E_1 and E_2 are the magnitudes of **E** in the left- and right-hand regions, respectively. Inasmuch as these fields satisfy Laplace's equation in both regions and also the boundary conditions, they are indeed the correct fields. Because the E-fields are the same in both regions, the D-fields must be different. Using $\mathbf{D} = \epsilon \mathbf{E}$, we have

$$D_1 = \frac{\epsilon_1 V_o}{d}, \qquad D_2 = \frac{\epsilon_2 V_o}{d}.$$

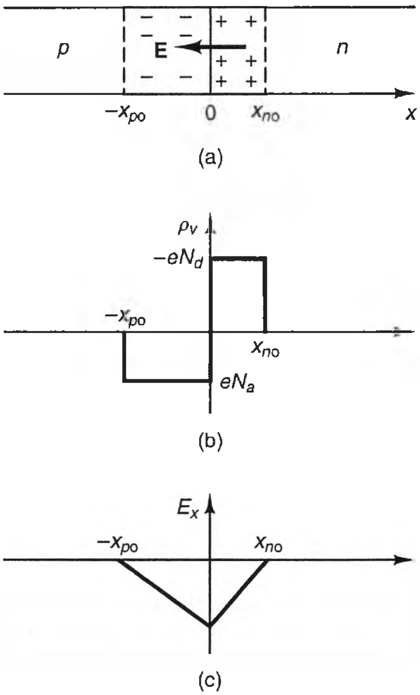
Since $\rho_s = D_n$ at the conducting surfaces, the magnitudes of the free-charge densities on the conducting plates are different in the two regions. Using Equation (5.57), we find that the charge densities on the inner sides of the plates are given by

$$|\rho_{1s}| = D_1, \quad |\rho_{2s}| = D_2, \tag{5.70}$$

where the charge densities on the inner surfaces of the top and bottom plates are positive and negative, respectively, when V_o is positive. Hence, the charge densities have the largest magnitudes on the side where the dielectric permittivity is also the largest.

A remarkable aspect of this solution is that we would have been hard pressed to have guessed the surface charge distribution that we finally found in Equation (5.70). This shows one of the greatest strengths of using a Laplace's equation method to solve field problems: If it is known that the volume charge distribution is zero, solutions can be obtained as long as the potentials of the conducting surfaces are known. This is in contrast to Coulomb's law, which requires one to know the charge distribution in advance.

5-4-2-3 A Semiconductor *pn* Junction. A simple, yet important electrostatic boundary value problem that involves Poisson's equation is the field distribution in the vicinity of a semiconductor *pn* junction. Figure 5-25a shows a simplified *pn* junction. Here, *p* and *n* doped semiconductor materials from a junction at $x = 0$. Because of the quantum properties of the dopants, electrons migrate from the donor atoms on the *n* side to the acceptor atoms on the *p* side, leaving a positive space charge on the *n* side and a negative space charge on the *p* side. A simple approximation of the resulting charge distribution is shown in Figure 5-25b, where *e* is the electron charge, and N_a and N_d are the densities of the acceptor and donor atoms on the *p* and *n* sides of the junction, respectively. The charges contained on each side of the junction have equal magnitudes and opposite signs, so the widths x_{no} and x_{po} are related to the doping densities by $N_d x_{no} = N_a x_{po}$.



(a)

(b)

(c)

Figure 5-25 A *pn* junction.
a) Simplified geometry. b) Charge distribution. c) Potential distribution.

If the total junction width $W = x_{po} + x_{no}$ is much smaller than the cross-sectional dimensions of the junction, the potential function V is independent of the y- and z-coordinates, and Poisson's equation reads

$$\nabla^2 V = \frac{\partial^2 V}{\partial x^2} = \begin{cases} -\dfrac{eN_a}{\epsilon} & -x_{p0} < x < 0 \\ \dfrac{eN_d}{\epsilon} & 0 < x < x_{no} \\ 0 & \text{Otherwise} \end{cases},$$

where ϵ is the semiconductor permittivity. Integrating this expression once with respect to x and using $\mathbf{E} = -\nabla V$ (which reduces in this case to $E_x = -\partial V/\partial x$), we obtain

$$E_x = \begin{cases} C_1 & x < -x_{po} \\ \dfrac{eN_a\,(x + x_{po})}{\epsilon} + C_1 & -x_{po} < x < 0 \\ \dfrac{-eN_d\,(x - x_{no})}{\epsilon} + C_1 & 0 < x < x_{no} \\ C_1 & x > x_{no} \end{cases},$$

where the constants of integration have been chosen to ensure that V is continuous at all points (which always occurs when $\mathbf{E}$ is finite).

To find the constant C_1, we note that the total charge contained in the p and n regions of the junction are exactly opposite, so $|\mathbf{E}|$ approaches zero as $|x| \to \infty$. Thus, $C_1 = 0$, and $\mathbf{E}$ is given by

$$\mathbf{E} = E_x \hat{\mathbf{a}}_x = \begin{cases} \dfrac{eN_a(x + x_{po})}{\epsilon} \hat{\mathbf{a}}_x & -x_{po} < x < 0 \\ \dfrac{-eN_d(x - x_{no})}{\epsilon} \hat{\mathbf{a}}_x & 0 < x < x_{no} \\ 0 & \text{otherwise} \end{cases} . \tag{5.71}$$

The value E_x is plotted as a function of x in Figure 5-25c.

The junction voltage V_{junct} can be found by integrating Equation (5.71) across the junction

$$V_{\text{junct}} = -\int_{-x_{p0}}^{x_{n0}} E_x \, dx = -\frac{eN_a x_{po}^2}{2\epsilon} - \frac{eN_d x_{no}^2}{2\epsilon},$$

where V_{junct} is measured from the n side of the junction to the p side. Remembering that $W = x_{po} + x_{no}$ and $N_a x_{po} = N_d x_{no}$, we see that this can be written as

$$V_{\text{junct}} = \frac{|e|}{2\epsilon} \frac{N_a N_d}{N_a + N_d} W^2, \tag{5.72}$$

which shows that the junction voltage is a nonlinear function of the junction width.

5-4-2-4 The Method of Images. The ***method of images*** is an analytical technique that involves replacing constant-potential surfaces with equivalent sources called ***image sources*** that generate the same fields. In many cases, a problem can be significantly simplified using this technique. Conducting boundaries that can be modeled in this way include infinite planes, spheres, infinite cylinders, and wedges. Of these, the infinite plane is the simplest to analyze.

To demonstrate the method of images, let us consider the situation depicted in Figure 5-26a. Here a point charge Q is located a distance d above an infinite, conducting plane. The ground symbol indicates that the plane is maintained at zero absolute potential, so it is called a ***ground plane***. This problem may appear simple to analyze by applying Coulomb's law, since only one point charge is present above the ground plane. However, there is also a surface-charge distribution on the ground plane, because the boundary condition at a perfect conductor requires that $\rho_s = D_n$ (See Equation (5.56).) This surface-charge distribution can be found only after $\mathbf{E}$ is found, so we cannot use Coulomb's law directly to find $\mathbf{E}$.

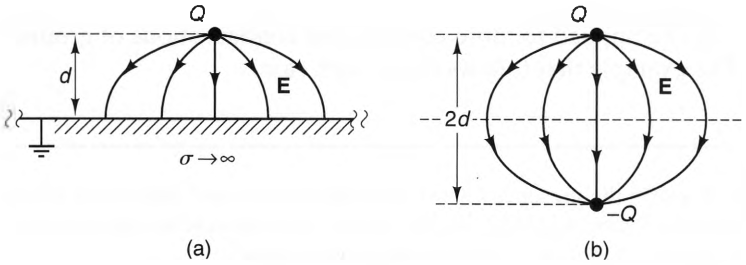


Figure 5-26 Example of the method of images. a) A point charge above an infinite, grounded conductor. b) An equivalent geometry that has the same E-field above the plane.

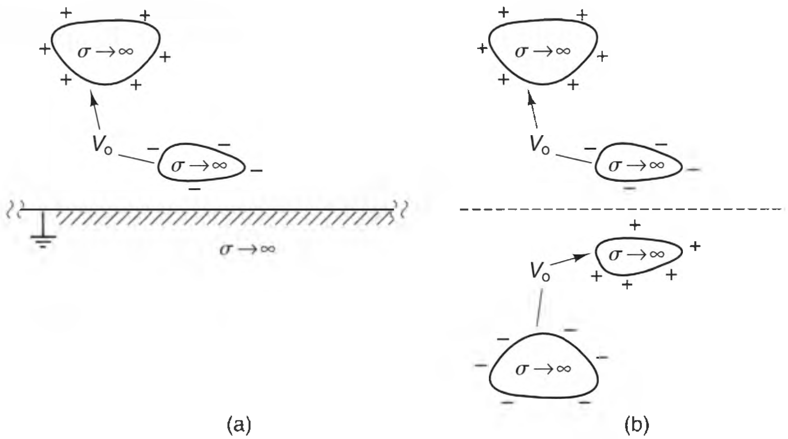


Figure 5-27 The method of images applied to a multiconductor geometry. a) The original geometry. b) An equivalent geometry.

Now, consider the configuration shown in Figure 5-26b. Here, the ground plane has been removed. In its place is a point charge of value $-Q$, located a distance $2d$ directly below the charge Q. We call this charge an ***image charge***, because the charge and its image constitute an electrostatic dipole that straddles the $z = 0$ plane, $E_t = 0$ everywhere on this plane, just as in Figure 5-26a. Hence, replacing the ground plane with this mirror-image charge maintains the same boundary condition as the ground plane did. The sources above the $z = 0$ planes for these two situations are identical, and the $E_t = 0$ boundary conditions along the planes are also identical. From the uniqueness principle, we can conclude that the fields generated by the two situations are also identical for all $z > 0$. Of course, for $z < 0$ the fields are not the same, which means that the situations are equivalent only for $z > 0$.

This procedure can be generalized for any configuration of electrostatic sources and materials above an infinite ground plane. Such a situation is depicted in Figure 5-27a. Here, two conducting surfaces are maintained at a potential difference of V_o above an infinite ground plane. In Figure 5-27b, the ground plane has been replaced by the image of the conductors, the voltage source, and the dielectric. Note that the polarity of the image voltage source must be reversed in order to maintain the correct symmetry of the charges. Since potentials above and below the $z = 0$ plane are exact mirror images of each other, they produce zero tangential E-field everywhere on the $z = 0$ plane, just as the ground plane does.

Image theory can also be applied to more complicated configurations of ground plane configurations. The example that follows is one such case.

Example 5-8

Figure 5-28a shows a point charge in the presence of two perpendicular ground planes that intersect at (0, 0, 0). If the point charge is located at $(d_1, d_2, 0)$, find an equivalent geometry that generates the same field in the range $0 < \phi < \pi/2$, without the ground planes.

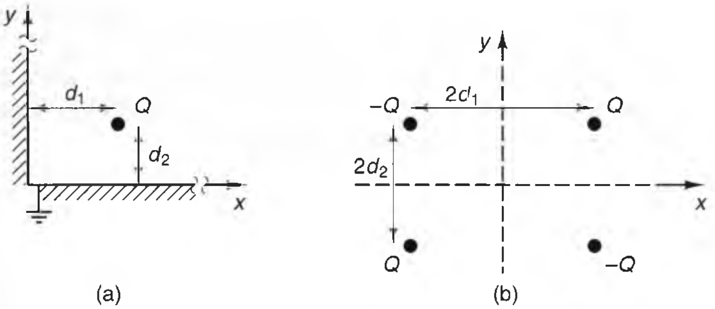


Figure 5-28 a) A point charge near a conducting corner. b) An equivalent geometry using the method of images.

Solution:

To maintain $E_t = 0$ along the entire yz-plane without the conductor present, a point charge of value $-Q$ can be placed at the point $(-d_1, d_2, 0)$. In order to obtain $E_t = 0$ along the xz-plane with the conductor removed, images of both of these charges can be placed below the xz-plane. Thus, charges of value Q and $-Q$ must be placed at $(-d_1, -d_2, 0)$ and $(d_1, -d_2, 0)$, respectively. This new configuration is shown in Figure 5-28b. The E-field generated by these four point charges can now be evaluated in the region $x > 0$ and $y > 0$ using Coulomb's law.

5-4-2-5 The Separation-of-Variables Technique. The ***separation-of-variables*** technique is a very powerful method of solving homogeneous differential equations, including Laplace's equation. Using this technique, we can represent a differential equation with N coordinate variables as N differential equations, each with one coordinate variable; hence the name "separation of variables."

To introduce the basic principles of the separation-of-variables technique, let us consider Laplace's equation in Cartesian coordinates,

$$\nabla^2 V = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = 0.$$

Next, let us look for solutions that can be expressed in the form

$$V(x,y,z) = X(x)\,Y(y)\,Z(z),$$

where $X(x)$, $Y(y)$, and $Z(z)$ are functions only of x, y, and z, respectively. Solutions of this type are called ***product solutions***. Substituting, we obtain

$$X''(x)\,Y(y)\,Z(z) + X(x)\,Y''(y)\,Z(z) + X(x)\,Y(y)\,Z''(z) = 0,$$

where $X''(x) = d^2X(x)/dx^2$, $Y''(y) = d^2Y(y)/dy^2$, and $Z''(z) = d^2Z(z)/dz^2$. Dividing both sides of this expression by the product $X(x)\,Y(y)\,Z(z)$, we obtain

$$\frac{X''(x)}{X(x)} + \frac{Y''(y)}{Y(y)} + \frac{Z''(z)}{Z(z)} = 0. \tag{5.73}$$

Taking a closer look at this equation, we see that the left-hand side is the sum of three terms, which are functions of x, y, and z, respectively. However, let's consider what happens as one moves from $x = x_o$ to $x = x_o + \delta$ while keeping y and z constant. Along this path, $Y''(y)/Y(y)$ and $Z''(z)/Z(z)$ are constant. Since the sum of all three terms in Equation (5.73) is zero throughout the region, $X''(x)/X(x)$ must be constant for all values of x. By similar reasoning, we can conclude that $Y''(y)/Y(y)$ and $Z''(z)/Z(z)$ must also be constants for all values of y and z, respectively. If we denote these constants as ***separation constants***, $-k_x^2$, $-k_y^2$, and $-k_z^2$, respectively, Equation (5.73) can be written in a form called the ***separation equation***,

$$k_x^2 + k_y^2 + k_z^2 = 0, \tag{5.74}$$

where

$$\frac{d^2X}{dx^2} + k_x^2 X = 0 \tag{5.75}$$

$$\frac{d^2Y}{dy^2} + k_y^2 Y = 0 \tag{5.76}$$

$$\frac{d^2Z}{dz^2} + k_z^2 Z = 0. \tag{5.77}$$

Equations (5.75) through (5.77) are each ordinary differential equations, as opposed to Laplace's equation, which is a partial differential equation. Thus, the separation-of-variables technique has reduced a partial differential equation in three variables to three ordinary differential equations.

To make our discussion simpler, let us now restrict ourselves to geometries where the potential V is independent of z. This is a two-dimensional case, where $Z(z)$ is constant, so $k_z = 0$. For this case, the separation equation becomes

$$k_x^2 + k_y^2 = 0,$$

or

$$k_y = \pm jk_x \equiv k,$$

where $j = \sqrt{-1}$. If k is real, the general solutions of Equations (5.75) and (5.76) are

$$X(x) = A \cos kx + B \sin kx$$

and

$$Y(y) = C \cosh ky + D \sinh ky,$$

respectively, and the product solutions are of the form

$$V(x,y) = (A \cos kx + B \sin kx)(C \cosh ky + D \sinh ky), \tag{5.78}$$

where A, B, C, D, and k are constants that must be determined to match the boundary conditions of the particular problem being solved.

One simple solution occurs when we let $A = C = 0$, $BD = V_1$, and $k = \pi/a$. For this case,

$$V(x,y) = V_1 \sin\left(\frac{\pi x}{a}\right) \sinh\left(\frac{\pi y}{a}\right). \tag{5.79}$$

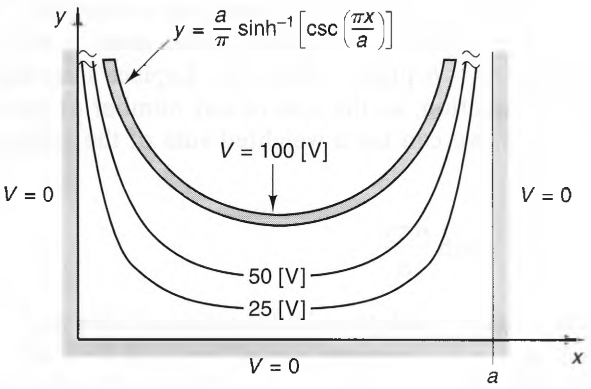

y
$y = \frac{a}{\pi} \sinh^{-1}\left[\csc\left(\frac{\pi x}{a}\right)\right]$
V = 100 [V]
V = 0
V = 0
50 [V]
25 [V]
V = 0
a
x

Figure 5-29 Potential distribution inside a conducting trough with three straight sides at ground potential and a curved top side at 100 [V] potential.

This potential function is zero valued at $x = 0$, $x = a$, and $y = 0$. Also, $V(x,y)$ has a constant value of V_1 along the surface where $\sin(\pi x/a)\sinh(\pi y/a) = 1$. Hence, this is the potential distribution for the geometry shown in Figure 5-29, where conductors have been placed along these constant-potential surfaces. This figure shows several constant potential surfaces for the case where $V_1 = 100$ [V]. Since the potential is independent of z, these surfaces are shown in cross section. Notice that the conducting surfaces extend to infinity at $x = 0$ and $x = a$ without meeting, which means that this case is interesting, but impractical to construct.

A more practical geometry is the two-dimensional rectangular trough shown in Figure 5-30. Here, conducting walls along the $x = 0$, $x = a$, and $y = 0$ planes are maintained at a potential $V = 0$, and a conducting wall along the $y = b$ plane is maintained at a potential of V_1. For this case, we notice that the solution given by Equation (5.79) matches the boundary condition $V = 0$ at $x = 0$ and $y = 0$ when $A = C = 0$. Also, this solution matches the $V = 0$ condition if we choose k such that $\sin ka = 1$, which occurs when

$$k = \frac{n\pi}{a} \qquad n = 1, 2, \dots \infty.$$

The resulting class of potential functions can be represented by

$$V_n(x,y) = V_n \sin\left(\frac{n\pi x}{a}\right) \sinh\left(\frac{n\pi y}{a}\right) \qquad n = 1, 2, \dots \infty. \tag{5.80}$$

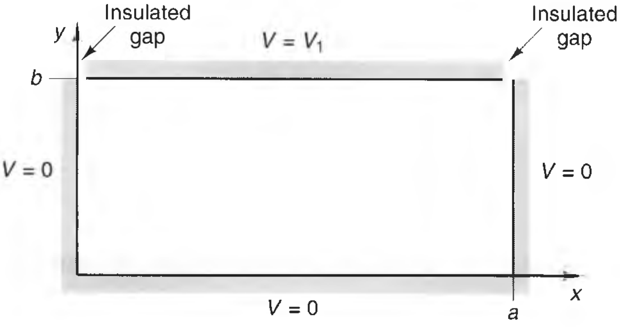


Figure 5-30 A rectangular, conducting trough.

Even though each member of this class of functions matches the boundary conditions along the $x = 0$, $x = a$, and $y = 0$ planes, none alone is able to match the $V = V_1$ condition along the $y = b$ plane. However, Laplace's equation is a linear, homogeneous differential equation, so the sum of any number of solutions is also a solution. With this in mind, we can try a weighted sum of the foregoing solutions, i.e.,

$$V(x,y) = \sum_{n=1}^{\infty} \gamma_n \sin \frac{n\pi x}{a} \sinh \frac{n\pi y}{a}, \tag{5.81}$$

where the γ_n are constants. Since each term in this series satisfies the $V = 0$ boundary conditions of Figure 5-30 at $x = 0$, $x = a$, and $y = 0$, their sum does also. However, in order to satisfy the boundary condition at $y = b$, we must also require that

$$\sum_{n=1}^{\infty} \gamma_n \sin \frac{n\pi x}{a} \sinh \frac{n\pi b}{a} = V_1 \qquad 0 < x < a.$$

This expression may at first look formidable, but a closer inspection reveals that the sum on the left is simply a Fourier sine series in the variable x. By choosing appropriate constants γ_n, this series is capable of representing any periodic function for $0 < x < a$ that has even symmetry about $x = a$, including a constant function. We can evaluate the constants γ_n by using the orthogonality properties of the sinusoidal functions. Multiplying both sides by $\sin(m\pi x/a)$ and integrating over the range $0 < x < a$, we obtain

$$\sum_{n=1}^{\infty} \gamma_n \sinh \frac{n\pi b}{a} \int_0^a \sin \frac{n\pi x}{a} \sin \frac{m\pi x}{a}\, dx = \int_0^a V_1 \sin \frac{m\pi x}{a}\, dx. \tag{5.82}$$

The integral on the left-hand side of this expression is zero for all integer values of m, except when $n = m$, where we find that

$$\int_0^a \sin \frac{m\pi x}{a} \sin \frac{m\pi x}{a}\, dx = \frac{a}{2}.$$

This means that for a given m, only the $n = m$ term remains in the sum on the left-hand side, so Equation (5.82) becomes

$$\frac{a}{2} \gamma_m \sinh \frac{m\pi b}{a} = \frac{aV_1}{m\pi}(1 - \cos m\pi) \qquad m = 1, 2, \ldots \infty.$$

Solving for the constants γ_m yields

$$\gamma_m = \frac{4V_1}{m\pi \sinh \dfrac{m\pi b}{a}} \qquad m = 1, 3, 5, \ldots \infty. \tag{5.83}$$

Now that all the constants in the series have been found, we can write the final potential solution,

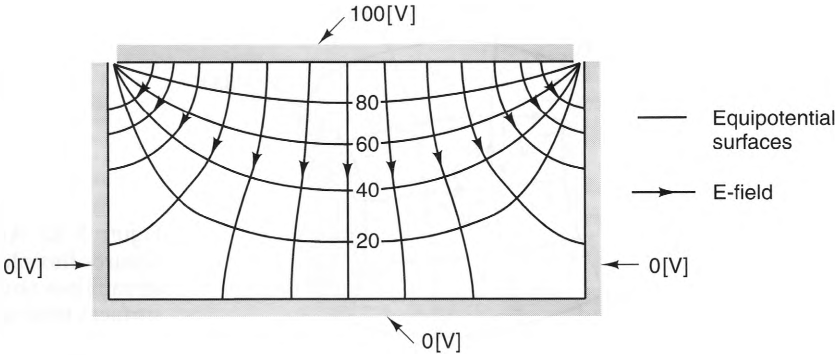


Figure 5-31 The potential distribution and E-field inside a rectangular trough.

$$V(x,y) = \frac{4V_1}{\pi} \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \sin \frac{n\pi x}{a} \frac{\sinh \frac{n\pi y}{a}}{n \sinh \frac{n\pi b}{a}} \tag{5.84}$$

Although it is not easy to envision the behavior of this solution simply by looking at it, it is a rather simple matter to calculate it using a computer or even a programmable calculator. A number of equipotential surfaces are shown in Figure 5-31 for the case where $V_1 = 100$ [V] and $a = 2b$. Also shown are several E-field streamlines, which are obtained using $\mathbf{E} = -\nabla V$.

As with all analytical techniques, the biggest advantage of the separation-of-variables technique is that it provides exact solutions to certain electrostatic problems. Unfortunately, analytical techniques can be used only on a relatively small number of problems. Also, these geometries are usually much simpler than those typically encountered in engineering practice. Fortunately, the graphical and numerical techniques discussed in the following sections can often be used to solve these more difficult problems.

5-4-3 FLUX PLOTS AND THE CURVILINEAR SQUARES TECHNIQUE

The curvilinear squares technique is a graphical method of solving two-dimensional electrostatic problems. While it lacks the accuracy of analytical and numerical techniques, it has several advantages over those techniques. Probably the most important is its ability to provide rough estimates of a solution with much less effort than is required by the other techniques. Also important is that the process of generating a graphical solution often provides insight into why the fields behave as they do. This is particularly helpful in the early stages of a design, where it is important to understand the basic operation of a system or device. If a particular configuration looks promising after a graphical analysis, a numerical technique can then be used to refine the analysis.

Figure 5-32 shows a portion of the cross section of a two-dimensional geometry that consists of a pair of charged conductors.

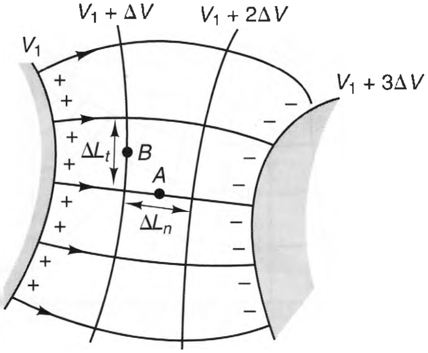


Figure 5-32 An electrostatic, curvilinear squares flux plot, showing the E-field streamlines and constant-potential surfaces between two conductors.

Two equipotential surfaces between these conductors are shown in this figure, as well as several E- (or D-) field streamlines. The tubes formed by adjacent streamlines are called ***flux tubes***, and the electric flux carried by each tube stays constant along its length, since flux enters and leaves the tubes only at the ends. The rectangles formed by the flux tubes and equipotential surfaces are called ***cells***, and the entire plot of field lines is called a ***flux plot***. In the discussion that follows, we will show that these field lines can be determined by graphically following a few simple "sketching rules."

Of all the equipotential surfaces and streamlines that could be drawn for the geometry shown in Figure 5-32, let us assume that the ones shown here were selected according to two criteria. First, the potential differences between adjacent equipotential surfaces are the same. We will call this potential difference ΔV, and for each cell we have

$$\Delta V = \int_{\Delta L_n} E \, d\ell \qquad [\text{V}], \tag{5.85}$$

where ΔL_n is the distance between the bounding potential surfaces of the cell. Second, we will assume that the spacing between the streamlines is such that the flux $\Delta\Psi$ passing through each flux tube is the same, where $\Delta\Psi$ for any cell is given by

$$\Delta\Psi = \int_{\Delta L_t} \epsilon E d\ell \qquad [\text{C/m}]. \tag{5.86}$$

Here, ΔL_t is the distance between the streamlines that bound the cell, and ϵ is the permittivity of the dielectric. Also, since $\rho_s = D$ at the surface of a perfect conductor, the charge contained at the positive and negative conductor ends of each flux tube is $+\Delta\Psi$ [C/m] and $-\Delta\Psi$ [C/m], respectively.

To develop the sketching rules for drawing the equipotential surfaces and flux tubes, let us consider the value of E_A at the point A in Figure 5-32. From Equation (5.85), we see that if the E-field is relatively constant within the cell, then E_A can be approximated as

$$E_A \approx \frac{\Delta V}{\Delta L_n}. \tag{5.87}$$

Equation (5.87) becomes exact in the limit as ΔV and ΔL_n approach zero. Similarly, at the point B, which lies in the same cell as A, we can estimate E_B using Equation (5.86) as

$$E_B = \frac{1}{\epsilon}\frac{\Delta\Psi}{\Delta L_t}, \tag{5.88}$$

where we have again assumed that **E** is nearly constant throughout the cell. This expression also becomes exact as ΔL_t and $\Delta\Psi$ approach zero.

Assuming that the number of potential surfaces and flux tubes drawn is large enough, the E-field at the points A and B will be nearly equal. Setting Equations (5.87) and (5.88) equal to each other, we obtain

$$\frac{\Delta V}{\Delta L_n} = \frac{1}{\epsilon}\frac{\Delta\Psi}{\Delta L_t}.$$

Rearranging this expression, we find that

$$\frac{\Delta L_t}{\Delta L_n} = \frac{1}{\epsilon}\frac{\Delta\Psi}{\Delta V}. \tag{5.89}$$

However, since we assumed from the onset that ΔV and $\Delta\Psi$ do not change from cell to cell, the ratio $\Delta L_t/\Delta L_n$ must also be the same for each cell. Hence,

$$\frac{\Delta L_t}{\Delta L_n} = \text{constant} = \frac{1}{\epsilon}\frac{\Delta\Psi}{\Delta V}. \tag{5.90}$$

The ratio $\Delta L_t/\Delta L_n$ is called the ***cell aspect ratio***. According to Equation (5.90), the cells of a properly drawn flux plot each have the same cell aspect ratio, regardless of their size. The simplest aspect ratio to draw is unity (1.0), in which case all the cells are square.

We can now formalize the procedure for using the curvilinear squares technique to solve two-dimensional problems. The following six-step procedure will almost always produce good flux plots in a minimum amount of time:

1. Start the initial plot by sketching just a few equipotential surfaces between the conductors. Two or three is often a good choice. Don't worry at this point about whether or not these surfaces are correct—the plot will eventually tell you this and you can quickly correct it. As a rule of thumb, the surfaces closest to the conductor boundaries will have shapes that are similar to the conductor shapes.
2. Sketch the first streamline. Any line can be chosen, but one along a line of symmetry is the best choice if such a line exists. Make sure that this line intersects each equipotential surface at right angles. Start the second line at a point that makes the first cell as square as possible. As this second streamline is extended, make sure that it intersects the equipotential surfaces at right angles, even if the resulting cells become rectangular, rather than square. Continue this process by adding streamlines until the conductors are completely surrounded. Don't worry if the last flux tube is only a partial one; this will not affect the accuracy of the flux plot.

3. Look at the sketch critically. Determine what changes in the equipotential surfaces would result in cells that are more square. Draw a new sketch that incorporates these changes. Repeat this step until you are satisfied with the "squareness" of all the cells. Place greater emphasis on accuracy in the smallest cells, since the energy density is highest in these regions.
4. In regions of the plot where the electric field is weak, the cells will be large and curved, and will possibly have more than four sides. To judge the "squareness" of these cells, subdivide them one or more times. If the original cell is correctly drawn, the subdivided cells will be more square.
5. The resolution of the plot is determined by the number of equipotential surfaces and electric field lines used. To obtain a more accurate solution, simply add more lines.
6. The process is complete when a plot is obtained that has the desired resolution and is square throughout the sketch.

A flux plot is a good way to show the fringing of the E-field lines near the edges of the plates of a parallel-plate capacitor. Figure 5-33 shows this clearly. Because of the symmetry of this problem, only the lines in the upper right-hand quadrant need be drawn. As can be seen, the cells are smallest between the plates, indicating that the E-field is strongest there. Outside the plates, the fields are weak, so the cells are large and irregularly shaped. One of these large, irregular cells has been subdivided in the figure to check its squareness. Since the subdivided cells are reasonably square, the large cell is indeed correctly drawn.

Flux plots are also useful for showing the behavior of fields near wedges and corners formed by conducting plates. Two examples are shown in Figures 5-34a and b. In Figure 5-34a, the fields near a "rooftop" conducting wedge are plotted. As can be seen, the cells are smallest near the tip, indicating that the largest fields and charge densities occur there. Therefore, this is where dielectric breakdown (arcing) is most likely to occur. Figure 5-34b shows the fields near a conductor in the shape of a right-angle corner. Here we see that the field strength and charge densities are smallest at the corner—just the opposite of the rooftop case. A quick conclusion that can be drawn by comparing the two geometries in Figure 5-34 is that a ditch is a much safer place to be during a thunderstorm than the peak of a roof!

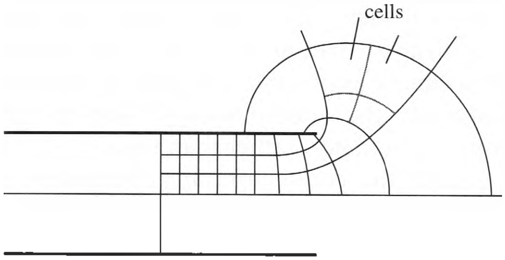


Figure 5-33 Flux plot for a parallel-plate capacitor.

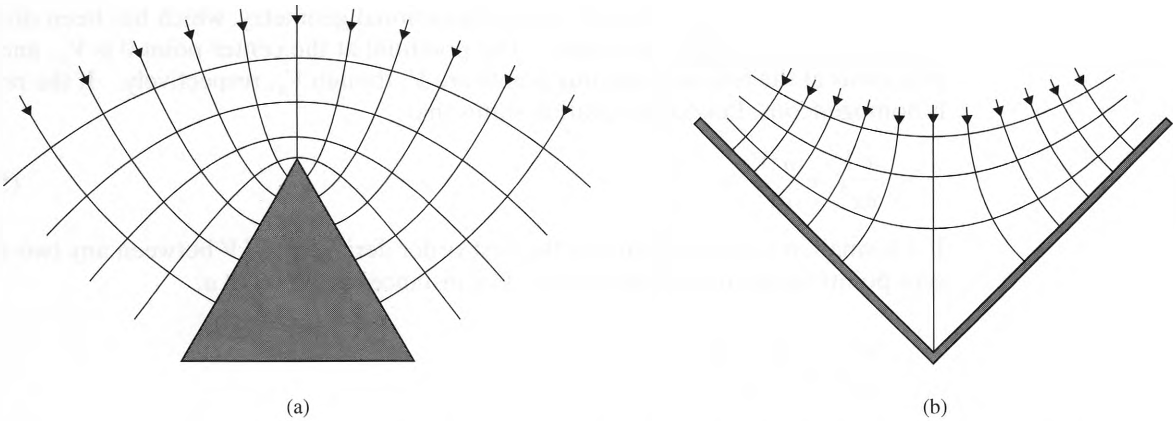


Figure 5-34 Flux plots at conducting wedges. a) An outer wedge. b) An inner wedge.

5-4-4 NUMERICAL TECHNIQUES

Most practical electrostatic boundary value problems are too complicated to be solved with a high degree of accuracy either by analytical or graphical techniques. For these situations, numerical techniques are needed. While they may lack the elegance of analytical solutions, numerical techniques are attractive because they can analyze broad classes of problems using the same solution procedure. This makes them ideal for use on computers.

Many different numerical techniques have been developed to solve electrostatic boundary value problems. This is fortunate, since no one technique is suitable for all types of problems. In general, these techniques differ from each other in their complexity, their accuracy, the range of problems they can handle, and the amount of computer resources they require to yield a solution. A presentation of all the numerical techniques used in electrostatic analysis is beyond the scope of this text. For that, the interested reader can refer to the references cited at the end of this text. Instead, we will present a relatively simple technique called the ***finite-difference*** technique that can be performed using either hand or computer calculations.

The finite-difference technique solves Laplace's and Poisson's equations by replacing the second-order derivatives in the Laplacian operator with finite-difference approximations. To see how this is accomplished, let us consider a two-dimensional problem in which the potential does not vary with the z-coordinate. Figure 5-35 shows

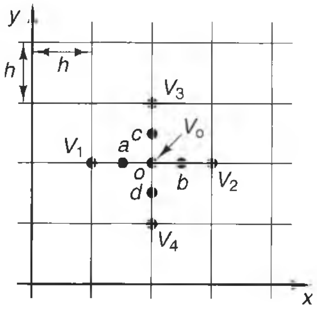


Figure 5-35 A numerical grid for the finite-difference technique of solving Laplace's and Poisson's equations.

a portion of the cross section of a two-dimensional geometry, which has been divided into squares of length h on a side. The potential at the center point 0 is V_o, and the potentials at the four surrounding points are V_1 though V_4, respectively. If the region is homogeneous, Poisson's equation states that

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = -\frac{\rho_v}{\epsilon}. \tag{5.91}$$

If h is small, we can approximate the first-order derivative of V between any two adjacent points by the finite differences. For instance, at the point a,

$$\left.\frac{\partial V}{\partial x}\right|_a \approx \frac{V_o - V_1}{h}.$$

Similarly, at the points b, c, and d, we have

$$\left.\frac{\partial V}{\partial x}\right|_b \approx \frac{V_2 - V_o}{h}$$

$$\left.\frac{\partial V}{\partial y}\right|_c \approx \frac{V_3 - V_o}{h}$$

$$\left.\frac{\partial V}{\partial y}\right|_d \approx \frac{V_o - V_4}{h}.$$

Knowing that $\partial^2 V/\partial x^2 = \partial/\partial x\,[\partial V/\partial x]$, we can approximate the second derivative of V at the point 0 by

$$\frac{\partial^2 V}{\partial x^2} \approx \frac{\left.\frac{\partial V}{\partial x}\right|_b - \left.\frac{\partial V}{\partial x}\right|_a}{h} \approx \frac{V_2 - V_o - V_o + V_1}{h^2}. \tag{5.92}$$

Similarly,

$$\frac{\partial^2 V}{\partial y^2} \approx \frac{\left.\frac{\partial V}{\partial y}\right|_c - \left.\frac{\partial V}{\partial y}\right|_d}{h} \approx \frac{V_3 - V_o - V_o + V_4}{h^2}. \tag{5.93}$$

Substituting Equations (5.92) and (5.93) into Equation (5.91), we have

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} \approx \frac{V_1 + V_2 + V_3 + V_4 - 4V_o}{h^2} \approx -\frac{\rho_v}{\epsilon}.$$

Solving this expression for V_o, we find that

$$V_o \approx \frac{1}{4}[V_1 + V_2 + V_3 + V_4] + \frac{h^2 \rho_v}{4\epsilon}. \tag{5.94}$$

Equation (5.94) can be generalized for any point in a rectangular grid, yielding

$$V_{i,j} \approx \frac{1}{4}[V_{i+1,j} + V_{i,j+1} + V_{i-1,j} + V_{i,j-1}] + \frac{h^2 \rho_{v_{i,j}}}{4\epsilon}, \tag{5.95}$$

where i and j represent the positions of the points along the x- and y-axes, respectively. This means that when the charge density at a point is zero, the potential at that point is simply the average of the potentials of the adjacent points.

We can use Equation (5.95) to solve electrostatic boundary value problems. To demonstrate this procedure, consider the geometry shown in Figure 5-36, which depicts a charge-free region surrounded by two conducting surfaces. The potentials of these surfaces are $V = 100$ and $V = 0$ volts, respectively. When a computer is used to solve for the potentials, the initial values of the unknown potentials are usually set to zero. For hand calculations, however, it is usually best to start with rough estimates, so that the number of iterations necessary to obtain an accurate solution is minimized.

For this problem, an initial estimate of V_5 can be obtained by averaging the four boundary points. We obtain $V_5 \approx (1/4)\,(0 + 0 + 0 + 100) = 25.0$. Next, V_7 lies at the center of a square bounded by corner potentials of 0, 0, 0, and $V_5 = 25$, so $V_7 \approx (1/4)$ $(25 + 0 + 0 + 0) = 6.3$ and $V_9 = V_7$. Similarly, V_1 lies at the center of a square with corner potentials of 0, 100, V_5, and the gap. If we estimate the potential in the gap to be 50 (i.e., the average of the two wall potentials), we can then estimate $V_1 = V_3 \approx (1/4)\,(50 + 0 + 100 + 25) = 43.8$. Finally, estimates of the remaining potentials can be obtained by applying Equation (5.95) directly to these nodes. These initial estimates are tabulated as the first entries in Figure 5-37.

To refine the initial estimate at the upper left-hand interior node potential V_1, let us apply Equation (5.95) to that node:

$$V_1 = \frac{1}{4}[0 + 100 + 53.2 + 18.8] = 43.0\ \text{V}.$$

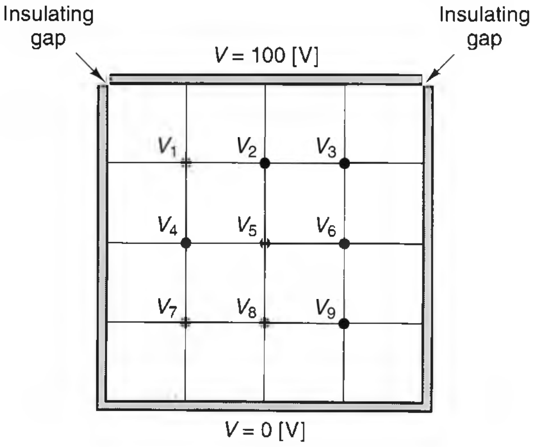


Figure 5-36 Numerical grid for determining the potential inside a rectangular trough.

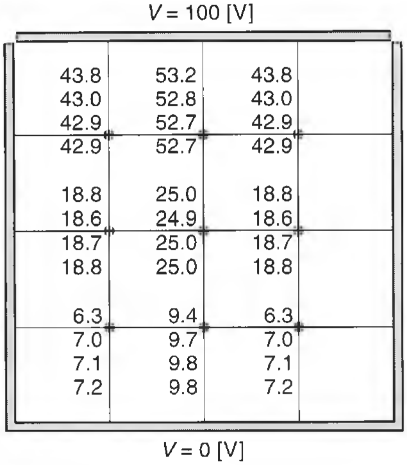


Figure 5-37 Potential values determined by successive iterations for a potential trough.

From the symmetry of the problem, this is also the next estimate of V_3. Now, we can refine the estimate of V_2 by using the best information available, which includes the values of V_1 and V_3 just calculated. Thus,

$$V_2 = \frac{1}{4}[42.6 + 100 + 25 + 42.6] = 52.8 \text{ V}.$$

We can use this same procedure at all the interior nodes in this problem, resulting in the values shown just below the initial estimates in Figure 5-37. Notice that the symmetry of the problem demands that $V_3 = V_1$, $V_6 = V_4$, and $V_9 = V_7$.

Once the first pass through the interior nodes has been completed, the procedure can be repeated until the nodal potentials have converged to the desired accuracy. Figure 5-37 shows the potential values through three complete iterations. The number of iterations necessary depends upon the complexity of the problem and the accuracy of the original potential estimates.

A more direct technique of solving for the nodal potentials using the finite-difference method involves the solution of simultaneous equations of the form

$$[\mathbf{Z}](\mathbf{V}) = (\mathbf{B}), \tag{5.96}$$

where $(\mathbf{V})$ is a column vector composed of the unknown interior potentials, $(\mathbf{B})$ is a column vector composed of the known boundary potentials, and $[\mathbf{Z}]$ is a square matrix that takes into account relative positions of the nodal points. Once the appropriate values of $[\mathbf{Z}]$ and $(\mathbf{B})$ have been determined, Equation (5.96) can be solved by standard matrix techniques. This is demonstrated in the following example.

Example 5-9

Calculate the nodal potentials for the geometry shown in Figure 5-37 using the matrix approach.

Solution:

We can reduce the number of unknowns in this problem by noting from symmetry that $V_3 = V_1$, $V_6 = V_4$, and $V_9 = V_7$. Applying Equation (5.95) at node 1 yields

$$V_1 = \frac{1}{4}[100 + V_2 + 0 + V_4],$$

or

$$4V_1 - V_2 - V_4 = 100.$$

Similarly, at nodes 2, 4, 5, 7, and 8, we obtain

$$-2V_1 + 4V_2 - V_5 = 100$$

$$-V_1 + 4V_4 - V_5 - V_7 = 0$$

$$-V_2 - 2V_4 + 4V_5 - V_8 = 0$$

$$-V_4 + 4V_7 - V_8 = 0$$

$$-V_5 - 2V_7 + 4V_8 = 0$$

These simultaneous equations can be written in matrix form as

$$\begin{bmatrix} 4 & -1 & -1 & 0 & 0 & 0 \\ -2 & 4 & 0 & -1 & 0 & 0 \\ -1 & 0 & 4 & -1 & -1 & 0 \\ 0 & -1 & -2 & 4 & 0 & -1 \\ 0 & 0 & -1 & 0 & 4 & -1 \\ 0 & 0 & 0 & -1 & -2 & 4 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_4 \\ V_5 \\ V_7 \\ V_8 \end{bmatrix} = \begin{bmatrix} 100 \\ 100 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Using a numerical matrix solver, the following values for the node voltages are obtained:

$$V_1 = 42.8 \qquad V_2 = 52.6 \qquad V_4 = 18.7$$

$$V_5 = 25.0 \qquad V_7 = 7.1 \qquad V_8 = 9.8.$$

These values agree well with the values shown in Figure 5-37.

If we were to compare the numerical solutions just obtained with the known analytical solution for this problem (see Equation (5.84)), we would see that these values are close, but they do not quite agree. The reason for this is that the values obtained using the finite-difference method are limited in accuracy by the grid size used. If more accurate results are desired, more grid points must be used. The drawback of this, of course, is that the number of calculations needed to obtain these values increases rapidly as the number of grid points increases. This is a common tradeoff when using numerical techniques; increased accuracy demands more computational resources (both in storage and time).

5-5 Summation

In this chapter we have laid the foundation for calculating electrostatic fields when materials are present. An important part of this discussion was the development of equations that relate the currents and charges within materials to the electric field intensity. The most important of these equations are

$$\mathbf{J} = \sigma \mathbf{E}$$

and

$$\mathbf{D} = \epsilon \mathbf{E}.$$

These constitutive relations provide the necessary material information to Maxwell's equations by accounting for charge and current distributions that are induced on and within materials.

The boundary value problems and solution techniques presented in this chapter were selected so as to demonstrate some of the basic characteristics of the electrostatic fields generated by charge and potential distributions. Many more techniques exist for modeling more complicated geometries, for which interested readers can refer to the references cited at the end of the chapter.

Problems

5-1 Calculate the conductivity of copper [Cu] if it is known that its free-electron charge density is $N = 8.5 \times 10^{22}$ [cm^{-3}] and its electron mobility is $\mu = 41.9$ [cm^2/V • s]. Compare this value with the one shown in Table C-2.

5-2 A sample of lightly doped GaAs has a room temperature conductivity of 4.04×10^{-9} [S/cm]. If it is known that the electron and hole mobilities are $\mu_n = 8500$ [cm^2/V • s] and $\mu_p = 400$ [cm^2/V • s], respectively, and the electron density is $N_n = 2.5 \times 10^6$ [cm^{-3}], find the hole density N_p.

5-3 Calculate the resistance of a 10 [m] length of copper wire that has a radius of 2 [mm].

5-4 Calculate the resistance between opposite sides of a solid cube of stainless steel if the cube is 1 [cm] on a side.

5-5 In later chapters it will be shown that the E-field inside a good conductor decays exponentially with increasing depth. Figure P5-5 depicts such a situation, where **E** varies in flat block of conductor according to the formula

$$\mathbf{E} = E_o e^{-\alpha z} \hat{\mathbf{a}}_x .$$

Assuming that the conductivity of the block is σ and the block is deep enough so that it can be considered to be infinitely deep, calculate the power dissipated in the entire depth beneath a square meter of the surface.

5-6 A material has a dielectric constant of $\epsilon_r = 3.0$ and has an atomic density of 10^{28} atoms per cubic meter. If only two electrons in the outer orbital shell will distort with an applied E-field, and both electrons follow the same orbital path as a pair, find the spacing between the center of the nucleus and the average location of the electrons when the applied E-field is 10,000 [V/m].

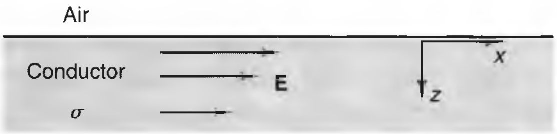


Figure P5-5

5-7 HCl is a polar molecule, with a dipole moment of 3.44×10^{-30} [C • m] per molecule. If it is known that the molecular bond distance is 0.136 [nm], find the magnitude of the net displaced charge in each molecule.

5-8 Water (H_2O) is a polar molecule, with a dipole moment of 6.15×10^{-30} [C • m] and a dielectric constant of 80. If the density of water is 33.4×10^{27} molecules per cubic meter, find the percentage of water molecules that align themselves with an applied E-field of magnitude:

(a) 10 [V/m]

(b) 10 [kV/m].

5-9 If a zener diode has a breakdown voltage of 8 [V], find the dielectric strength of the semiconducting material if it is assumed that the entire voltage is dropped across the *pn* junction, which has width 0.3 [μm]. Assume that **E** is uniform across the junction.

5-10 If the earth is considered to be a metal sphere (radius ≈ 6371 [km]), how much charge Q must be deposited on its surface in order for an arc to be established in the air. If the earth's surface was charged to this value by removing all the electrons from a volume of soil, how large would this volume be? Assume that the electron density of soil is approximately 7×10^{23} [cm^{-3}].

5-11 Determine the resistance between the inner and outer surfaces of a homogeneous cylinder. Assume that the cylinder has conductivity σ, and length ℓ, and that the inner and outer radii are a and b, respectively.

5-12 Repeat Problem 11 for the case where the conductivity of the cylinder varies inversely with increasing distance from its axis; that is, $\sigma = k/\rho$.

5-13 Figure P5-13 shows two conducting plates with surface area S, separated by two homogeneous dielectric sheets. The sheets have permittivities ϵ_1 and ϵ_2, respectively, and thickness d_1 and d_2, respectively.

(a) Calculate the surface charge densities on the upper and lower plates.

(b) Calculate the polarization surface charge densities on the upper, middle, and lower dielectric interfaces.

(c) Calculate the total (i.e., free plus polarization) charge contained on all the conductor and dielectric surfaces.

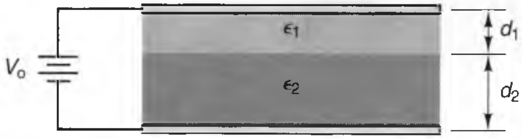


Figure P5-13

5-14 Prove the identity $\nabla' \dfrac{1}{|\mathbf{r} - \mathbf{r}'|} = \dfrac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3}$.

5-15 Two point charges a distance d apart in free space exert a force of 1.4×10^{-4} [N]. When the free space is replaced by a homogeneous dielectric medium, the force becomes 0.9×10^{-4} [N]. What is the dielectric constant ϵ_r of the medium?

5-16 Determine whether or not $V = xy$ is the correct potential function for the geometry shown in Figure P5-16 for the region $0 < x < 1$ and $0 < y < 1$. Why or why not? Assume that the region between the conducting plates has a relative permittivity of $\epsilon_r = 3.0$.

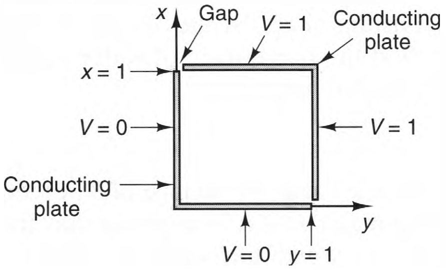


Figure P5-16

5-17 A point charge Q is located at $(0,0,d)$ above an infinite conducting plane that lies in the xy-plane and is maintained at ground potential. Find a) the surface charge density as a function of x and y on the conducting plane and b) the total charge induced on the conducting plane.

5-18 Figure P5-18 shows two conducting plates of width w, separated by a distance d, where $w >> d$. The voltage between the plates is V_o. The dielectric constant is ϵ_r between the plates, and 1.0 outside the plates.

(a) Show that, away from the edges, the charge density on the inner plate surfaces is $\pm \epsilon_0 \epsilon_r V_o/d$.

(b) If the plates are oppositely charged, show that the charge density is negligible on the outer surfaces of the plates away from the edges. (*Hint:* Use the E-field expression for uniform surfaces of charge to show that $\mathbf{E} = 0$ inside the plates only when the charge density on the outer surfaces is zero.)

(c) Show that, away from the edges, $\mathbf{E} \approx 0$ above and below the plates.

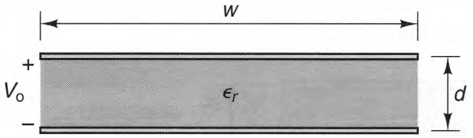


Figure P5-18

5-19 Consider a coaxial transmission line that consists of inner and outer conductors of radii a and b, and is filled with an electron cloud with a volume charge density of $\rho_v = \gamma/\rho$ [C/m^3] for $a < \rho < b$. If the outer conductor is $+V_o$ volts more positive than the inner conductor, find the E-field between the conductors.

5-20 Figure P5-20 shows two infinite conductors that extend from $\rho = 0$ to $\rho = \infty$ on the $\phi = 0$ and the $\phi = \phi_o$ planes, respectively. Solve Laplace's equation for the potential V and the electric field $\mathbf{E}$ in the regions:

(a) $0 < \phi < \phi_o$

(b) $\phi_o < \phi < 2\pi$.

(*Hint:* Start by assuming that V is independent of ρ and z.)

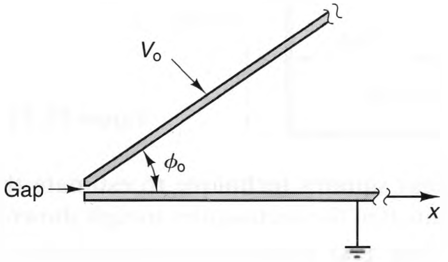


Figure P5-20

5-21 Modify the rectangular trough shown in Figure 5-30 by assuming that the potential along the wall at $y = b$ is given by $V = V_o \sin(\pi x/a)$ and the other potentials are unchanged.

(a) Find $V(x, y)$.

(b) Evaluate V and $\mathbf{E}$ at the at the center of the trough when $V_o = 100$ [V].

5-22 Use the curvilinear squares technique to find three equipotential surfaces between the inner and outer conductors of the two-dimensional geometry shown in Figure P5-22. Assume that the dielectric between the circular and triangular conductors is uniform.

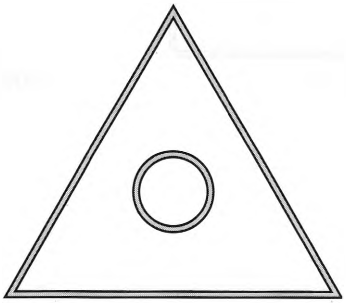


Figure P5-22

5-23 Figure P5-23 shows a rectangular trough. All the walls are maintained at zero potential, except the top wall, which is maintained at a potential of 100 [V]. Write a numerical program (using a language such as FORTRAN or C++, or a mathematical software program, such as Matlab™ or Mathcad™) that solves

Laplace's equation throughout the trough. From these calculations, specify the potential and the E-field at the point P.

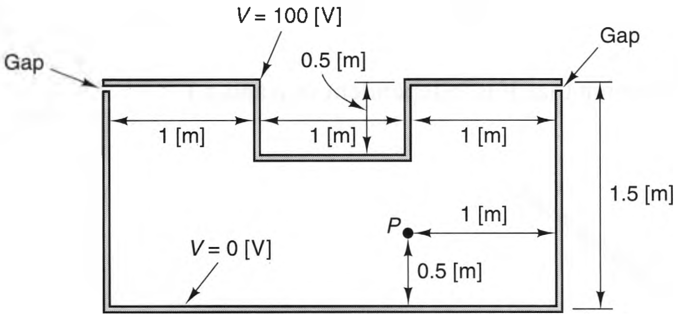


Figure P5-23

5-24 Use the curvilinear squares technique to estimate the potential V and electric field $\mathbf{E}$ at the point P in the rectangular trough shown in Figure P5-23.

5-25 Use the finite-difference technique to calculate the potentials at the points a, b, and c in the geometry shown in Figure P5-25.

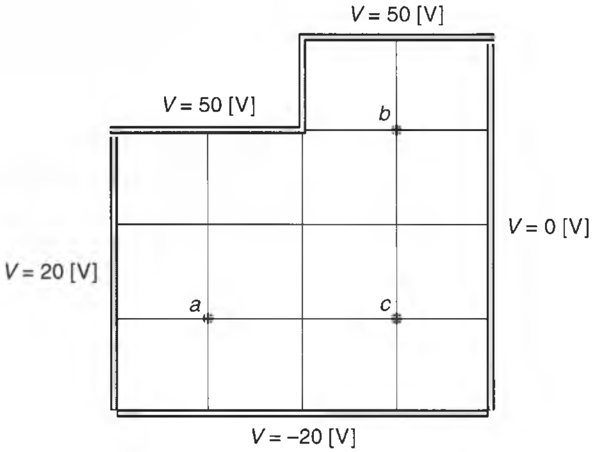


Figure P5-25

6

Capacitance and Electric Energy

6-1 Introduction

We havc already seen from the Lorentz force law that electric fields can exert forces on charges and accelerate them. We have also seen that a voltage is a measure of the work necessary to move charges through an electric field. Because of this, it should not be surprising that the E-field generated by a system of charges is directly linked to the energy necessary to assemble those charges. In this chapter, we will formalize this relationship between E-fields and energy.

An important parameter that is a measure of an electrostatic element's ability to store energy in its electric field is its capacitance. The reader is no doubt familiar with capacitance in lumped electrical circuits, but capacitance is also present in any system or element in which electric fields are present. As a result, capacitance is important in many types of distributed systems or elements, such as transmission lines and waveguides. In this chapter, we will show how the capacitances of lumped or distributed elements are governed by the geometric arrangement of their components.

6-2 Capacitance

Consider the geometry shown in Figure 6-1, which consists of two perfectly conducting surfaces, S_+ and S_-, in a dielectric medium with permittivity ϵ. If the conductors hold free charges of Q and $-Q$, respectively, a voltage V will exist between the conductors whose value depends upon the charge Q, the sizes and shapes of the conductors, and the permittivity ε of the dielectric. We define the ***mutual capacitance*** (or simply ***capacitance***) of these conductors as

$$C \equiv \frac{Q}{V} \qquad \text{[C/V or F].} \tag{6.1}$$

Capacitance is measured in coulombs per volt [C/V] or farads [F], but most capacitances found in engineering practice usually have values that are much smaller than a farad. Thus, capacitance values are typically specified in microfarads [μF] or picofarads [pF].

The definition given by Equation (6.1) is useful for understanding how capacitance is related to voltage and charge, but does not readily show what physical aspects of a system of conductors give rise to its capacitance. To actually calculate the capacitance of an element that consists of two or more distinct conductors, we can rewrite this expression in terms of the E-field generated by charges on the conductors. Referring to Figure 6-1, we can write

$$V = -\int_{S_-}^{S_+} \mathbf{E} \cdot \mathbf{d\ell}. \tag{6.2}$$

Also, since the charge density on a conductor equals the normal component of **D**, we have

$$Q = \oint_{S_+} \rho_s ds = \oint_{S_+} \epsilon\mathbf{E} \cdot \mathbf{ds}, \tag{6.3}$$

Substituting Equations (6.2) and (6.3) into Equation (6.1), we obtain

$$C = \frac{\displaystyle\oint_{S_+} \epsilon\mathbf{E} \cdot \mathbf{ds}}{\displaystyle\int_{S_+}^{S_-} \mathbf{E} \cdot \mathbf{d\ell}} \qquad \text{[F].} \tag{6.4}$$

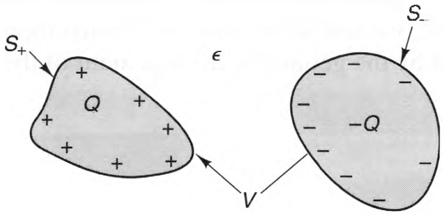


Figure 6-1 Two charged conductors in a dielectric medium.

This formula shows that the capacitance is a function of how the E-field distributes itself throughout a system. Hence, capacitance is controlled by the size, shape, and placement of conductors, as well as the permittivity and orientation of the dielectric.

6-2-1 TWO SIMPLE CAPACITANCE EXAMPLES

Figure 6-2 shows a parallel-plate capacitor, which consists of a homogeneous dielectric with permittivity ε that is sandwiched between two parallel, perfectly conducting sheets. In Section 5-4-2-1, we found that when fringing is negligible, the E-field generated inside the dielectric is directed perpendicular to the plates and has a magnitude given by Equation (5.63), i.e.,

$$E = \frac{V}{d},$$

where V is the voltage between the plates and d is the spacing between the plates. Also, we know from the previous chapter that the charge density on the outer surfaces of the plates is negligible and that the charge density on the positive plate is given by

$$\rho_s = D = \frac{\epsilon V}{d}.$$

This means that the charge on the positive plate is

$$Q = \rho_s S = \frac{\epsilon V S}{d}.$$

Finally, using $C = Q/V$, we find that the capacitance of a parallel-plate capacitor is given by

$$C = \frac{\epsilon S}{d} \text{ [F]} \qquad \text{(Parallel-plate capacitor).} \qquad (6.5)$$

Figure 6-3 shows another simple capacitor, consisting of two concentric, oppositely charged spheres of radii a and b, respectively. In Chapter 5, we found that when the dielectric between the spheres is homogeneous, the E-field between the spheres is directed radially outward from the inner sphere to the outer sphere with a magnitude given by

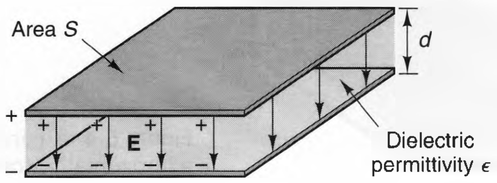


Figure 6-2 A parallel-plate capacitor with a uniform dielectric.

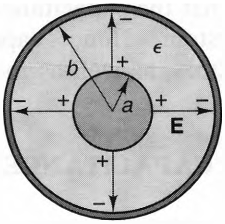


Figure 6-3 A capacitor formed out of two concentric spheres, separated by a uniform dielectric.

$$E = \frac{V}{r^2\left[\dfrac{1}{a} - \dfrac{1}{b}\right]} \qquad a < r < b,$$

where V is the voltage between the inner and outer spheres. The surface-charge density on the inner conductor equals the magnitude of $\mathbf{D} = \epsilon\mathbf{E}$ at $r = a$. Thus, the charge on the inner conductor is

$$Q = \rho_s S = \frac{4\pi\epsilon V}{\left[\dfrac{1}{a} - \dfrac{1}{b}\right]},$$

where $S = 4\pi a^2$ is the surface area of the inner sphere. Using $C = Q/V$, we obtain

$$C = \frac{4\pi\epsilon}{\left[\dfrac{1}{a} - \dfrac{1}{b}\right]} \quad \text{[F]} \qquad \text{(Concentric spheres).} \tag{6.6}$$

A special case of this formula occurs when the outer sphere has an infinite radius. For this case, Equation (6.6) yields the capacitance of an isolated sphere:

$$C_{\text{sphere}} = \lim_{b\to\infty} \frac{4\pi\epsilon}{\left[\dfrac{1}{a} - \dfrac{1}{b}\right]} = 4\pi\epsilon a. \tag{6.7}$$

6-2-2 COMPOUND CAPACITORS

It is often possible to model complex configurations of conductors and dielectrics as series or parallel arrangements of simple capacitors. For instance, consider the compound system shown in Figure 6-4.

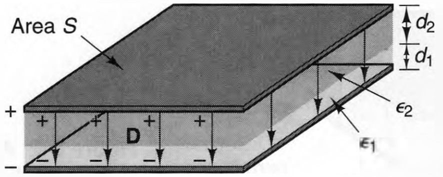


Figure 6-4 A parallel-plate capacitor with a horizontally stratified dielectric.

Here, two homogeneous dielectric slabs with widths d_1 and d_2, respectively, are sandwiched between two parallel, conducting plates that each have surface area S. This geometry was analyzed previously; we found that the D-field between the plates is uniform and given by Equation (5.68), namely,

$$D = \frac{\epsilon_1 \epsilon_2 V}{\epsilon_2 d_1 + \epsilon_1 d_2},$$

where V is the voltage between the plates. The surface-charge density on the positive plate equals the downward component of **D**. Hence, the total charge Q on the positive plate is

$$Q = \rho_s S = DS = \frac{\epsilon_1 \epsilon_2 VS}{\epsilon_2 d_1 + \epsilon_1 d_2}.$$

Using $C = Q/V$, we obtain

$$C = \frac{\epsilon_1 \epsilon_2 S}{\epsilon_2 d_1 + \epsilon_1 d_2} \qquad \text{[F]}. \tag{6.8}$$

This can also be written as

$$\frac{1}{C} = \frac{d_1}{\epsilon_1 S} + \frac{d_2}{\epsilon_2 S} = \frac{1}{C_1} + \frac{1}{C_2}, \tag{6.9}$$

where C_1 and C_2 are the capacitances of homogeneous dielectric capacitors formed by the upper and lower dielectrics, respectively. This is the familiar formula for capacitors in series that is used in circuit analysis. Notice that the same result would be obtained if a thin, conducting sheet were placed between the two dielectrics. In fact, Equation (6.9) can be applied to any series capacitive system whenever conducting sheets can be placed between the series elements without upsetting the fields in either region.

Next, let us consider the geometry shown in Figure 6-5, which consists of two different dielectric slabs placed side by side between two parallel, conducting plates. The surface areas and permittivities of the two dielectric regions are S_1, ϵ_1 and S_2, ϵ_2, respectively. When fringing is negligible, both **E** and **D** are directed normal to the plates, but this time the boundary conditions require that **E**, rather than **D**, be continuous across the dielectric interface. Thus,

$$E = \frac{V}{d} = \frac{D}{\epsilon}$$

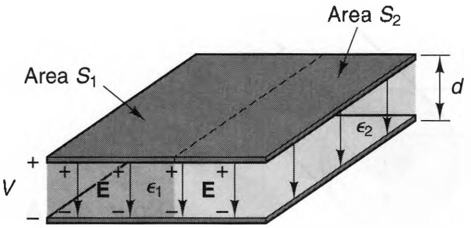


Figure 6-5 A parallel-plate capacitor with a vertically stratified dielectric.

where ϵ equals ϵ_1 in region 1 and ϵ_2 in region 2. Given that the charge density on the upper plate is $\rho_s = D$, the total charge on this plate is

$$Q = \frac{\epsilon_1 V}{d} S_1 + \frac{\epsilon_2 V}{d} S_2.$$

Using $C = Q/V$, we find that the capacitance of this device is

$$C = \frac{\epsilon_1 S_1}{d} + \frac{\epsilon_2 S_2}{d}. \tag{6.10}$$

This expression is the same as the formula for capacitors in parallel that is familiar from circuit analysis, i.e.,

$$C = C_1 + C_2, \tag{6.11}$$

where C_1 and C_2 are the capacitances of each portion of the capacitor alone. Equation (6.11) applies to any capacitive system whenever the E-field in either parallel region is unaffected by the presence of the other.

6-2-3 DISTRIBUTED CAPACITANCE ON TRANSMISSION LINES

Transmission lines are multiconductor structures that are capable of guiding electromagnetic energy. Because of this property, they are useful in a large number of applications, ranging from 60-Hz power transmission to microwave signal transmission.

One of the most important characteristics of any transmission line is its capacitance per unit length. This parameter, in conjunction with the transmission line's inductance per unit length (which we will discuss in Chapter 9), governs a number of the operating characteristics of the line. We will now calculate the capacitance properties of two popular types of transmission lines: coaxial lines and two-wire lines.

Coaxial Lines. One of the most common transmission lines is the coaxial line. In its simplest form, a coaxial line consists of a solid, conducting wire, surrounded by a uniform dielectric and a conducting cylinder. In practice, the outer conductor is usually made of braided wire, but can often be modeled accurately by a solid cylinder. This structure is shown in Figure 6-6, where the inner and outer conductors have radii of a and b, respectively. The electric field between the inner and outer conductors was derived in Chapter 5 using Laplace's equation. When the voltage V between the

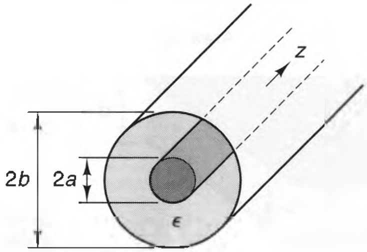


Figure 6-6 A coaxial transmission line.

inner and outer cylinders is positive, **E** and **D** are directed radially outward from the inner cylinder and have magnitudes given by

$$D = \epsilon E = \frac{\epsilon V}{\rho \ln\left(\frac{b}{a}\right)}.$$

The surface-charge density ρ_s on the inner conductor equals D at $\rho = a$, so the charge per unit length ρ_ℓ on the inner conductor equals the surface-charge density times the circumference of the inner conductor:

$$\rho_\ell = 2\pi a \rho_s = \frac{2\pi\epsilon V}{\ln\left(\frac{b}{a}\right)} \qquad [\text{C/m}].$$

The capacitance per meter equals the number of coulombs per meter ρ_ℓ divided by the voltage V, so

$$C = \frac{2\pi\epsilon}{\ln\left(\frac{b}{a}\right)} \qquad [\text{F/m}]. \tag{6.12}$$

Example 6-1

RG-58U coaxial cable has a 20-AWG solid inner conductor with a 0.406 [mm] radius, surrounded by a solid polyethelene dielectric and an outer, braided conductor of radius 1.553 [mm]. Find its capacitance per meter.

Solution:

From Table C-3 of Appendix C, the dielectric constant of polyethylene is $\epsilon_r = 2.26$. Using Equation (6.12), we find that the capacitance per meter is

$$C = \frac{2\pi \times 2.26 \times \epsilon_o}{\ln\left(\frac{1.553}{0.406}\right)} = 93.73 \qquad [\text{pF/m}].$$

Two-wire lines. In its simplest form, a two-wire transmission line consists of two round, parallel wires, each of radius *a,* spaced by a distance *D*, as shown in Figure 6-7.

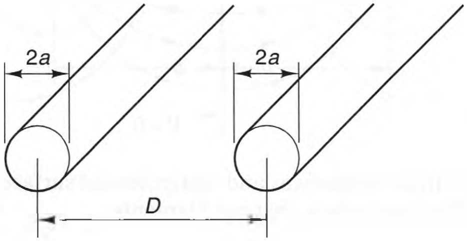


Figure 6-7 A two-wire transmission line.

In order to find the capacitance per meter of this structure, we must first determine the potential field between the conductors. This is a more difficult problem than it first might seem, since we cannot assume that the charge distribution on either wire is symmetrical about its cross section, particularly when the wires are closely spaced. However, we will now show that the image of each wire into the other is an infinite charge filament that lies parallel and inside the other wire. This derivation involves quite a few steps, but is worth it, since the final result is very simple.

Let us start by considering the two infinite line charges shown in Figure 6-8. These line charges have opposite line charge densities, ρ_ℓ and $-\rho_\ell$, respectively, and intersect the xy-plane on opposite sides of the y-axis, at $x = \pm b$, respectively. To describe the constant-potential surfaces generated by the two line charges, we first note that the E-field generated by a single infinite line charge varies inversely with the radial distance R from each line. (See Equation (4.17).)

Since $V = \int_C \mathbf{E} \cdot \mathbf{d\ell}$, the potential V at an arbitrary point P varies logarithmically with the radial distance R from the line; that is,

$$V = \frac{\rho_\ell}{2\pi\epsilon} \ln \frac{R_o}{R},$$

where ρ_ℓ is the line-charge density and R_o is the radial distance from the line charge to an arbitrary zero-potential reference point. When considering the two line charges shown in Figure 6-8, it is convenient to choose the reference point R_o equidistant from both charges, which can be any point along the y-axis. For this case, the potential V due to both line charges at any point in the xy-plane is given by

$$V = \frac{\rho_\ell}{2\pi\epsilon}\left(\ln \frac{R_o}{R_+} - \ln \frac{R_o}{R_-}\right) = \frac{\rho_\ell}{2\pi\epsilon} \ln \frac{R_-}{R_+},$$

where R_+ and R_- are the radial distances from the positive and negative lines, respectively, to the point P. In Cartesian coordinates, this can be expressed as

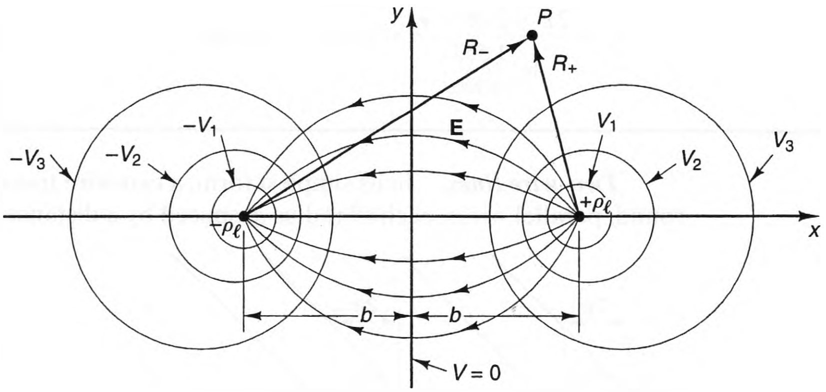


Figure 6-8 The E-field streamlines and equipotential surfaces generated by two infinite, parallel, oppositely charged filaments.

$$V = \frac{\rho_\ell}{2\pi\epsilon}\ln\left[\frac{(x+b)^2+y^2}{(x-b)^2+y^2}\right]^{1/2} = \frac{\rho_\ell}{4\pi\epsilon}\ln\left[\frac{(x+b)^2+y^2}{(x-b)^2+y^2}\right]. \tag{6.13}$$

The contours of constant potential occur when the argument of the logarithm in Equation (6.13) is constant. Thus, the constant potential contours satisfy the expression

$$\frac{(x+b)^2+y^2}{(x-b)^2+y^2} = k^2.$$

This equation can be rearranged to read

$$\left(x - \frac{k^2+1}{k^2-1}b\right)^2 + y^2 = \left(\frac{2kb}{k^2-1}\right)^2. \tag{6.14}$$

Equation (6.14) defines a family of circles in the $z = 0$ plane whose centers lie along the x-axis. Each circle corresponds to a different potential V and has a radius "a" and a center coordinate $x = c$, which are given by

$$a = \left|\frac{2kb}{k^2-1}\right| \tag{6.15}$$

and

$$c = \frac{k^2+1}{k^2-1}b, \tag{6.16}$$

respectively. Using Equations (6.15) and (6.16), it is easy to show that a, b, and c are related by the simple equation

$$a^2 + b^2 = c^2. \tag{6.17}$$

Several of these equipotential cylinders are shown in Figure 6-8, along with several E-field streamlines.

Since the circular cylinders shown in Figure 6-8 are all equipotential surfaces, any of them could be coated with a perfect conductor without changing either the potential distribution or the E-field streamlines. If we were to do this to, say, the smallest cylinders in the figure we would obtain a two-wire transmission line. If the radii of these cylinders are $a_1 = a_2 = a$, their center-to-center spacing can be expressed as $D = c_1 - c_2 = 2c_1$. The cylinders have equal and opposite potentials V_1 and $-V_1$, respectively, so the voltage between them is $V_o = 2V_1$. Substituting $x = D/2 - a$ and $y = 0$ into Equation (6.13), we obtain

$$V_o = 2\frac{\rho_\ell}{4\pi\epsilon}\ln\left[\frac{\left(\frac{D}{2} - a + b\right)^2}{\left(\frac{D}{2} - a - b\right)^2}\right] = \frac{\rho_\ell}{\pi\epsilon}\ln\left[\frac{\frac{D}{2} - a + b}{\frac{D}{2} - a - b}\right].$$

Substituting Equation (6.17) and $c_1 = D/2$, we obtain (after some rearranging)

$$V_o = \frac{\rho_\ell}{\pi\epsilon}\ln\left[\frac{D}{2a} + \sqrt{\left(\frac{D}{2a}\right)^2 - 1}\right].$$

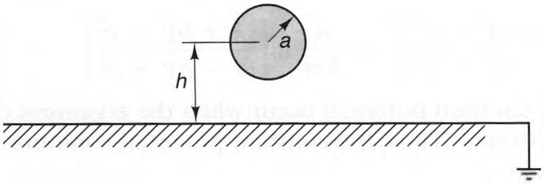


Figure 6-9 A single-wire transmission line.

The charge per unit length on the outer surface of the positive cylinder is ρ_ℓ, since all the flux lines from the positive line charge pass through the cylinder. Using $C = Q/V_0$, the capacitance per unit length of the two-wire line can be expressed as

$$C = \frac{\pi\epsilon}{\ln\left(\dfrac{D}{2a} + \sqrt{\left(\dfrac{D}{2a}\right)^2 - 1}\right)} \quad \text{[F/m]}, \tag{6.18}$$

or, using the identity $\ln[x + \sqrt{x^2 - 1}] = \cosh^{-1}x$, we can write this as

$$C = \frac{\pi\varepsilon}{\cosh^{-1}\left[\dfrac{D}{2a}\right]} \quad \text{[F/m]}. \tag{6.19}$$

We can use this result to find the capacitance of another type of transmission line, the ***single-wire line***, which consists of a wire at a constant height above an infinite, zero-potential, conducting plane. Such a line is shown in Figure 6-9. From this figure, it is obvious that the two conductors that constitute this transmission line are not identical, making this an example of an unbalanced transmission line. We can find the capacitance of this transmission line by noticing that the $x = 0$ plane between two cylinders in Figure 6-8 has zero potential and thus can be considered to be a conducting surface. The voltage between the cylinder and this plane is exactly half the voltage between the two cylinders, so the capacitance per unit length of a single-wire line is exactly double that of the corresponding two-wire line, that is,

$$C = \frac{2\pi\varepsilon}{\cosh^{-1}\left[\dfrac{h}{a}\right]} \quad \text{[F/m]}, \tag{6.20}$$

where h is the height of the wire above the conducting plane.

Example 6-2

Estimate the capacitance between the wire and the chassis shown in Figure 6-10. Assume that the wire has radius 1 [mm], and length 1 [cm], and is positioned 2 [mm] above the conducting plane.

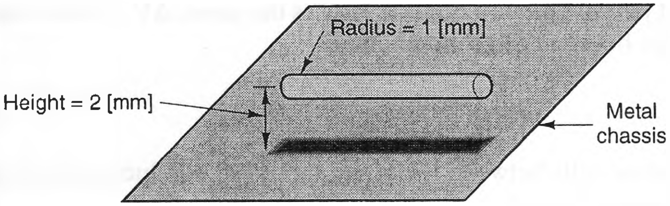


Figure 6-10 A wire mounted above a flat metal chassis.

Solution:

Assuming that the chassis can be approximated by an infinite ground plane, we have from Equation (6.20),

$$C/\text{length} = \frac{2\pi\epsilon_0}{\cosh^{-1}\left(\frac{2}{1}\right)} = 42.24 \qquad [\text{pF/m}].$$

Since the length of the wire is 1 [cm], the total capacitance between the wire and the chassis is

$$C = 42.24\ [\text{pF/m}] \times .01\ [\text{m}] = 0.422\ [\text{pF}].$$

6-2-4 ESTIMATING CAPACITANCE WITH THE CURVILINEAR SQUARES TECHNIQUE

We showed in the previous chapter that it is possible to solve two-dimensional electrostatic problems using the curvilinear squares technique. This technique can also be used to provide quick estimates of the capacitance of two-dimensional structures, such as transmission lines. Figure 6-11 shows the cross-sectional cut and flux plot of a transmission line with a triangular inner conductor and a circular outer conductor. Because of the symmetry of this geometry, only one-sixth of the flux plot need be drawn. All the information needed to estimate the capacitance of the transmission line is contained in this flux plot. To show this, we first note that according to the rules of the curvilinear squares technique (see Section 5-4-3), the potential difference

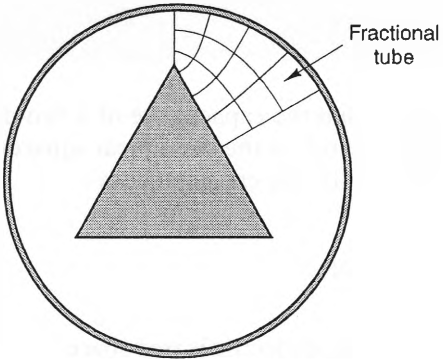


Figure 6-11 A curvilinear squares plot of a transmission line with a triangular inner conductor.

between each adjacent pair of equipotential surfaces is the same, ΔV. This means that we can write the voltage between the conductors as

$$V = N_v \Delta V, \tag{6.21}$$

where N_v is the number of cells between the inner and outer conductors along any of the flux tubes. In this case, $N_v = 3$.

The flux carried by each flux tube in a curvilinear squares flux plot is the same, $\Delta\Psi$. This means that the charge per unit length contained on the positive conductor can be expressed as

$$Q = N_Q \Delta\Psi, \tag{6.22}$$

where N_Q is the total number of flux tubes that surround the positive conductor. In this case, since there are six identical sectors that surround the inner conductor, we have $N_Q = 6 \times (3 + 0.9) \approx 23.4$, where we have estimated that the fractional tube carries approximately nine-tenths the flux of the full tubes.

Using Equations (6.21) and (6.22), we can express the capacitance per meter as

$$C = \frac{Q}{V} = \frac{N_Q}{N_V}\frac{\Delta\Psi}{\Delta V}.$$

But, according to Equation (5.90), when the cells of a flux plot are square, the ratio of $\Delta\Psi$ to ΔV is given by

$$\frac{\Delta\Psi}{\Delta V} = \epsilon\frac{\Delta L_t}{\Delta L_n},$$

where $\Delta L_t/\Delta L_n$ is the cell aspect ratio and ϵ is the permittivity of the dielectric between the conductors. When cells are square, the aspect ratio is unity, and the capacitance per unit length of a two-dimensional system is given by

$$C = \epsilon\frac{N_Q}{N_V} \quad \text{[F/m]}. \tag{6.23}$$

Thus, all that is necessary to find the capacitance of a two-dimensional structure is to count the voltage and flux squares on the curvilinear squares flux plot. For the transmission line shown in Figure 6-11, the capacitance is

$$C \approx \epsilon_0\frac{6 \times 3.9}{3.0} \approx 69 \text{ [pF/m]},$$

where we have assumed that the dielectric is free space.

6-2-5 CAPACITORS WITH LOSSY DIELECTRICS

So far we have considered only capacitors with perfect (i.e., nonconducting) dielectrics. This is often a good approximation to real-world components, but there are many cases where the conductivity of the dielectric cannot be ignored. To see how these components can be modeled, consider the geometry depicted in Figure 6-12a. Here, a parallel-plate capacitor contains a dielectric that has permittivity ϵ and conductivity σ. If a dc voltage is applied to the plates, a resistive current I_R will flow between them. Nevertheless, the potential distribution between the plates is unaffected by the nonzero σ. This is because both the free and polarization charge densities throughout a homogeneous dielectric are zero after the initial transient period. (See Equations (5.35) and (5.37).) As a result, the potential distribution between the plates still satisfies Laplace's equation and is not affected by the nonzero σ.

In order to determine the equivalent circuit of a lossy capacitor, let us now consider the voltage between the plates to be time varying. As long as the magnetic field is negligible, the E-field between the plates will still distribute itself as an electrostatic field.[1] In this case, the total current I can be considered as the sum of the resistive current I_R and the current I_C associated with the time-varying charge Q on the plates. Using the definition of capacitance, we have

$$I_C = \frac{dQ}{dt} = \frac{d}{dt}(CV) = C\frac{dV}{dt}.$$

Hence, the total current flowing into the positive terminal of this lossy capacitor is

$$I = I_C + I_R = C\frac{dV}{dt} + \frac{V}{R},$$

which means that a lossy capacitor can be modeled as the parallel combination of a resistor and capacitor, as shown in Figure 6-12b.

When a capacitor has a homogeneous dielectric, the value of the product RC is determined solely by the ratio of the dielectric permittivity ϵ to the conductivity σ. To show this, we first note that we have already shown that the capacitance C and the resistance R of a lumped element are governed by the way in which **E** distributes itself throughout the element. From Equations (6.4) and (5.9), we have

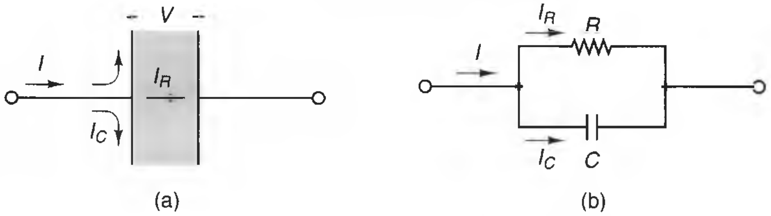


Figure 6-12 A lossy capacitor a) Physical geometry. b) Equivalent circuit.

[1] This is usually the case when the spacing between the capacitor plates is small with respect to the free-space wavelength at the operating frequency. Free space wavelength is discussed in Chapter 12.

$$C = \frac{\oint_{S_+} \epsilon \mathbf{E} \cdot \mathbf{ds}}{\int_{S_+}^{S_-} \mathbf{E} \cdot \mathbf{d\ell}}, \qquad R = \frac{\int_{S_+}^{S_-} \mathbf{E} \cdot \mathbf{d\ell}}{\oint_{S_+} \sigma \mathbf{E} \cdot \mathbf{ds}}.$$

Hence, the product RC is given by

$$RC = \frac{\oint_{S_+} \epsilon \mathbf{E} \cdot \mathbf{ds}}{\oint_{S_+} \sigma \mathbf{E} \cdot \mathbf{ds}}.$$

If the dielectric medium is homogeneous, both ϵ and σ can be taken out of the integrals, and the remaining integrals cancel, yielding

$$RC = \frac{\epsilon}{\sigma} \quad \text{(Homogeneous dielectric).} \tag{6.24}$$

Thus, if the permittivity and conductivity of a capacitor's dielectric are known, its capacitance can be calculated if its resistance is known, and vice versa.

Example 6-3

It is known that a lossy capacitor is filled with a homogeneous dielectric with permittivity $\epsilon = 2\epsilon_0$ and conductivity $\sigma = 1.0 \times 10^{-4}$ [S/m]. If measurements show that the resistance of the device is 10 [kΩ], find its capacitance.

Solution:

Using Equation (6.24), we find that

$$C = \frac{\epsilon}{R\sigma} = \frac{2 \times 8.854 \times 10^{-12}}{(10 \times 10^3)(1 \times 10^{-4})} = 17.7 \text{ [pF]}.$$

6-2-6 ELECTRIC SHIELDING

The capacitive coupling between the conductors of a system is not always a welcome effect. For instance, the coupling between wires in a telephone circuit can give rise to ***cross talk***, where the signal energy from one circuit also appears in another. Mutual capacitance decreases as the distance between the conductors increases, but it is often impractical to place the conductors far enough apart to reduce the capacitive coupling to acceptable levels. A more practical method is to surround the critical components with a grounded, metal enclosure. This procedure is called ***electric shielding,*** which is often called ***Faraday shielding***.

We can model the effects of a conducting shield by considering the multiconductor system shown in Figure 6-13a. In this system, conductors #1 and #2 are separated by a shield, conductor #3, that completely surrounds conductor #2. In practice, the

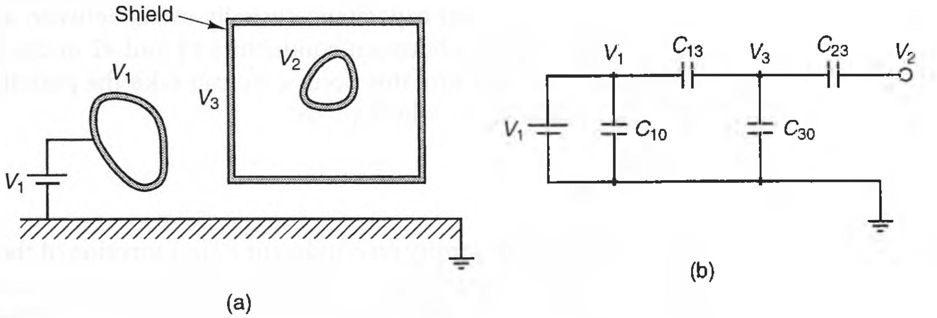


Figure 6-13 A multiconductor system with an ungrounded shield. a) Geometry, b) Equivalent circuit.

shield may be a chassis or subchassis, and conductors #1 and #2 may represent circuits or components that need to be electrically isolated. For the moment, let us assume that the shield potential V_3 is floating, which means that it is not held at a constant potential with respect to ground.

We can model the characteristics of this multiconductor system using the configuration of lumped capacitances shown in Figure 6-13b. Here, the conductors are represented by nodes and the potential differences between them are governed by lumped capacitances between these nodes. A complete verification of this equivalent circuit is beyond the scope of this text[2], but we can offer an abbreviated proof by first expressing the total charge on each of the ungrounded conductors in terms of their potentials. According to the superposition principle, these charges can be expressed as a weighted sum of the conductor potentials. If Q_1, Q_2, and Q_3 are the charges on conductors #1, #2 and #3, respectively, we can write

$$Q_1 = c_{11}V_1 + c_{12}V_2 + c_{13}V_3$$
$$Q_2 = c_{21}V_1 + c_{22}V_2 + c_{23}V_3$$
$$Q_3 = c_{31}V_1 + c_{32}V_2 + c_{33}V_3.$$

The values of the coefficients $c_{11}, c_{12}, \ldots\ c_{33}$ are governed by the geometric layout of the conductors and are called ***coefficients of capacitance***. These three expressions are linear equations with respect to the conductor potentials, so we can rearrange them so that they are functions of the potential differences between the conductors,

$$Q_1 = C_{10}V_1 + C_{12}(V_1 - V_2) + C_{13}(V_1 - V_3) \tag{6.25a}$$

$$Q_2 = C_{12}(V_2 - V_1) + C_{20}V_2 + C_{23}(V_2 - V_3) \tag{6.25b}$$

$$Q_3 = C_{13}(V_3 - V_1) + C_{23}(V_3 - V_2) + C_{30}V_3\,. \tag{6.25c}$$

The coefficients C_{10}, C_{12}, etc., are linear combinations of the coefficients of capacitance. For instance, $C_{12} = -c_{12}$, $C_{13} = -c_{13}$, and $C_{10} = c_{11} + c_{12} + c_{13}$. These new coefficients are always positive-valued and represent the lumped, ***mutual capacitances*** shown in Figure 6-13b.

[2] For a complete description, see *Principles of Electrodynamics* by Melvin Schwartz, published by McGraw Hill, 1972, pp. 54–62.

Although a mutual capacitance usually exists between any two conductors, the mutual capacitance C_{12} between conductors #1 and #2 in the system shown in figure 6-13a is zero. To see why this occurs, we can take the partial derivative of equation 6.25b with respect to V_1, which yields:

$$\frac{\partial Q_2}{\partial V_1} = -C_{12}\,.$$

We know the charge density on conductor #2 is a function of the D-field strength at the surface. However, according to the uniqueness principle, neither **D** nor **E** inside the shield are affected by potentials or charges outside the shield as long as the shield potential V_3 is held constant. Hence, the partial derivative of this expression is zero, which means that C_{12} is zero. By similar reasoning, C_{20} is also zero.

Even though the shield causes C_{12} to vanish, we can see from Figure 6-13b that the potential V_2 can still be affected by V_1 when the shield potential V_3 is allowed to float. This is because of the mutual capacitances C_{13} and C_{23} between the outer and inner conductors and the shield, respectively. Thus, a floating shield will not isolate conductor #1 from changes in potential outside the shield. On the other hand, if the shield is grounded (as shown in Figure 6-14a), the potential distribution in the region enclosed by the shield is now independent of the charges and potentials outside this region, since the surrounding potential is now constant.

The schematic for this situation is shown in Figure 6-14b. Comparing Figures 6-14b and 6-13b, we see that the shield is effective in isolating components within it only when it is connected to a constant potential surface—usually a ground.

A region can be electrostatically shielded from the E-fields generated by external charge distributions by surrounding the region with a constant-potential surface.

This is an important concept to remember when designing low-noise circuits and systems.

There are many situations where it is not possible to completely enclose circuits with a perfect shield, since openings are often needed for power supply wires and ventilation. Fortunately, even partial shields can often reduce the coupling to acceptable

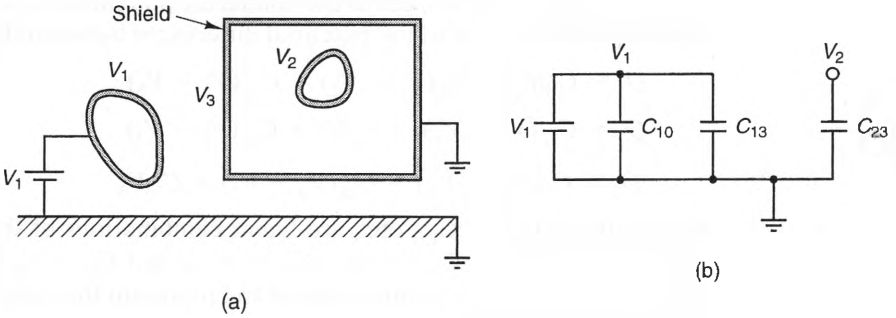


Figure 6-14 A multiconductor system with a grounded shield. a) geometry, b) equivalent circuit.

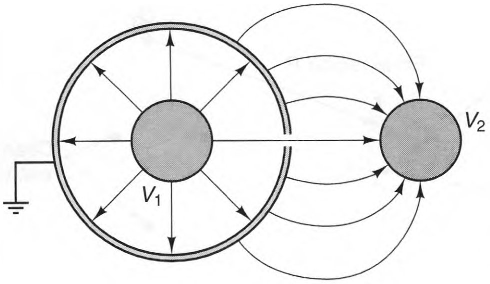


Figure 6-15 A partially shielded conductor.

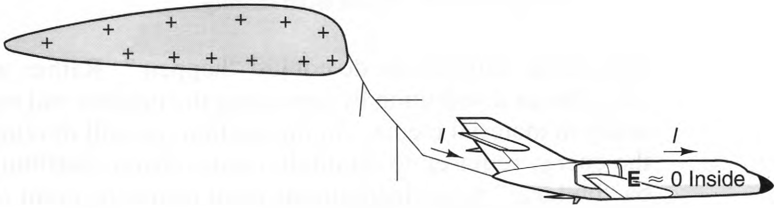


Figure 6-16 An aircraft with a lightning channel attached. The fields inside the fuselage are small duc to Faraday shielding.

levels. Figure 6-15 shows an example of a partial shield. Here, an inner conductor is partially shielded by an outer conductor with a slit. Even with the slit, the coupling between the inside and outside conductors is small, because the shield intercepts most of the E-field lines that would otherwise link the two conductors.

Faraday shields are used not only to reduce capacitive coupling in circuits, but also to protect regions from high fields and currents. For example, the metal fuselage of an airplane acts as a Faraday shield during a lightning strike. This is depicted in Figure 6-16. Here, the fuselage provides a path for the large currents and maintains an essentially field-free environment inside. Of course, this shielding is not perfect, since an aircraft fuselage is not perfectly conducting and also has openings, such as windows and door seams. Nevertheless, the shielding effectiveness afforded by an aircraft fuselage is good enough to make it safe for passengers during a lightning strike. Unfortunately, the shielding effectiveness may not be adequate to protect the digital control circuits used in modern aircraft, as a transient event can actually change the logic states of these circuits. For these types of circuits, additional shielding is often needed. This type of shielding is often called ***hardening***.

6-3 The Energy Contained in an Electrostatic Charge Distribution

Up to this point in our discussion, we have investigated the electric fields generated by charge distributions with little thought as to how these charge distributions are created. If we care only about the fields generated by a known charge distribution, there is no need to know how this charge distribution was created. But common sense tells us

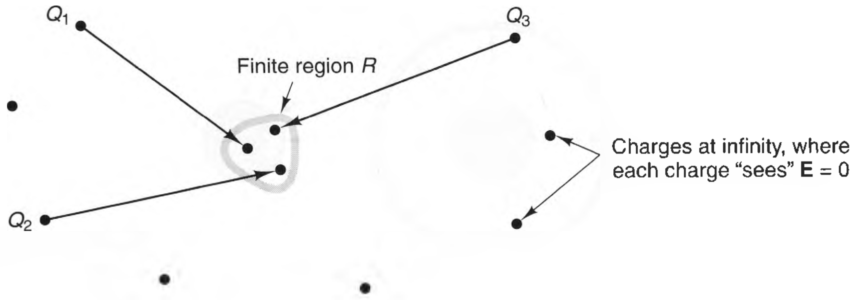


Figure 6-17 An initially charge-free region R that is charged by bringing point charges into the region from infinity.

that charge distributions do not just "happen." Rather, work must be expended to create a charge distribution by separating the positive and negative charges that occur naturally in material media. In this section, we will develop the expressions that specify the energy required to establish a static charge distribution. We will accomplish this by "building" these distributions point charge by point charge and summing the energies needed to place each charge.

Figure 6-17 shows a homogeneous, linear dielectric region R that initially contains no charge. Far outside of R there exists a pool of point charges—so far away as (infinity) that there is no **E**-field initially within R. In addition, these charges are also far enough away from each other that the forces exerted on each other are negligible. This means that for this initial state, the E-field experienced by each charge is zero.

If we move the charge Q_1 to a point within R, no energy is expended, since Q_1 experiences no E-field anywhere along the path. To move the next charge Q_2 to a point within R, a quantity W_2 of energy must be expended on the charge, since Q_2 experiences a force $\mathbf{F} = Q_1\mathbf{E}$ from the electric field generated by Q_1 as it nears and enters R. Noting that the force $-\mathbf{F}$ must be applied to move the charge Q_2, we find that

$$W_2 = \int_{\infty}^{P} (-\mathbf{F})\cdot d\boldsymbol{\ell} = -Q_2\int_{\infty}^{P} \mathbf{E}\cdot d\boldsymbol{\ell} = Q_2V_{12},$$

where V_{12} is the potential (referenced to the zero potential surface at ∞) due to the already placed Q_1 at the resting location of Q_2. Using Equation (4.44), we can express V_{12} as

$$V_{12} = \frac{Q_1}{4\pi\epsilon R_{12}},$$

where R_{12} is the distance between the resting locations of Q_1 and Q_2. Next, to move Q_3 inside R, work W_3 must be expended against the fields generated by both Q_1 and Q_2; thus,

$$W_3 = Q_3V_{13} + Q_3V_{23} = Q_3(V_{13} + V_{23}),$$

where V_{13} and V_{23} are the potentials[3] due to Q_1 and Q_2 at the location of Q_3, respectively.

[3] Since we have assumed that the medium is linear, the value of ϵ is independent of the fields already present in the system.

By now, the procedure for adding more charges to the charge distribution should be obvious. To place N charges inside R, the total work expended on the charges is

$$W_e = Q_2V_{12} + Q_3(V_{13} + V_{23}) + Q_4(V_{14} + V_{24} + V_{34}) + \ldots$$
$$\ldots + Q_N(V_{1N} + V_{2N} + \ldots + V_{N-1,N}). \tag{6.26}$$

If we note that

$$Q_iV_{ji} = \frac{Q_iQ_j}{4\pi\epsilon R_{ij}} = Q_jV_{ij},$$

we can then rewrite Equation (6.26) as

$$W_e = Q_1(V_{21} + V_{31} + \ldots) + Q_2(V_{32} + V_{42} + \ldots)$$
$$+ Q_3(V_{43} + V_{53} + \ldots) + \ldots . \tag{6.27}$$

Adding Equations (6.26) and (6.27) and regrouping, we obtain

$$2W_e = Q_1(V_{21} + V_{31} + V_{41}\ldots) + Q_2(V_{12} + V_{32} + V_{42} + \ldots)$$
$$+ Q_3(V_{13} + V_{23} + V_{43} + V_{53} + \ldots) + \ldots ,$$

which can be written as

$$W_e = \frac{1}{2}\sum_{i=1}^{N} Q_i \sum_{\substack{j=1 \\ j\neq i}}^{N} V_{ji} = \frac{1}{2}\sum_{i=1}^{N} Q_iV_i, \tag{6.28}$$

where

$$V_i = \sum_{\substack{j=1 \\ j\neq i}}^{N} V_{ji}$$

is the potential "seen" by the i^{th} charge due to all of the other charges placed in the system.

To find the energy expended to construct a continuously distributed charge distribution within a volume, we can replace Q_i in Equation (6.28) with $\rho_v dv$. In the limit as $N \rightarrow \infty$, this expression can be written as

$$W_e = \frac{1}{2}\int_{\text{Vol.}} \rho_v V dv \qquad [\text{J}], \tag{6.29}$$

where V is the absolute potential function generated by the charge distribution ρ_v. Similar expressions can be derived for the energy required to assemble surface and line charge distributions:

$$W_e = \frac{1}{2}\int_S \rho_s V ds \qquad [\text{J}], \tag{6.30}$$

and

$$W_e = \frac{1}{2}\int_C \rho_\ell V d\ell \qquad [\text{J}]. \tag{6.31}$$

From Equations (6.29)–(6.31), we can see that the energy required to create a charge distribution is a function of both the charge distribution itself and the potential distribution that it generates. We can also formulate an expression for this energy solely in terms of the field quantities **D** and **E**. To accomplish this, we first replace ρ_v with $\nabla \cdot \mathbf{D}$, which yields

$$W_e = \frac{1}{2}\int_{\text{Vol.}} (\nabla \cdot \mathbf{D}) V dv\,.$$

Using Identity B.3, we can write the integrand as

$$(\nabla \cdot \mathbf{D})V = \nabla \cdot V\mathbf{D} - \mathbf{D} \cdot \nabla V.$$

Also, substituting $\mathbf{E} = -\nabla V$, we can express W_e as

$$\begin{aligned} W_e &= \frac{1}{2}\int_{\text{Vol.}} \nabla \cdot V\mathbf{D} dv + \frac{1}{2}\int_{\text{Vol.}} \mathbf{D} \cdot \mathbf{E} dv \\ &= \frac{1}{2}\oint_S V\mathbf{D} \cdot \mathbf{ds} + \frac{1}{2}\int_{\text{Vol.}} \mathbf{D} \cdot \mathbf{E} dv, \end{aligned} \tag{6.32}$$

where the divergence theorem has been used to transform the left-hand integral into a surface integral about the closed surface S that bounds the volume in which the charge lies.

Looking closely at Equation (6.29), we see that the volume of integration can be any volume, just as long as it contains the actual volume filled by the charge distribution ρ_v. If we let the integration volume be all of space, the contribution to W_e from the surface integral over S approaches zero. This is because **E** and V decay no slower than r^{-2} and r^{-1} (respectively) at large distances from any charge distribution with finite dimensions. Thus,

$$\oint_{S_\infty} V\mathbf{D} \cdot \mathbf{ds} = \lim_{r\to\infty} \int_0^{2\pi}\int_0^{\pi} V\mathbf{D} \cdot r^2 \sin\theta d\theta d\phi \to 0.$$

Substituting this result into Equation (6.32), we obtain

$$W_e = \frac{1}{2}\int_{\text{Vol.}} \mathbf{D} \cdot \mathbf{E} dv \qquad [\text{J}], \tag{6.33}$$

where the integration takes place at all points at which the dot product $\mathbf{D} \cdot \mathbf{E}$ is nonzero. The term $\frac{1}{2}\,\mathbf{D} \cdot \mathbf{E}$ is called the ***electric energy density*** and is measured in units of joules per cubic meter. Also, if the medium is isotropic, $\mathbf{D} = \epsilon\mathbf{E}$, from which it follows that

$$W_e = \frac{1}{2}\int_{\text{Vol.}} \epsilon|\mathbf{E}|^2 dv. \tag{6.34}$$

The energy W_e required to construct a charge distribution is also the potential energy of the system once it is built. The question of where this energy resides is an interesting one, and Equations (6.29) and (6.33) offer different perspectives on this question. Whereas Equation (6.29) indicates that this energy resides in the charge distribution, Equation (6.33) indicates that it resides in the electric fields generated by charge distribution, which can exist far away from the charge distribution itself. Which view is correct? The answer is that both are correct. This is because a charge distribution and the electric field that it generates are an inseparable pair; a complete knowledge of one completely defines the other.

Example 6-4

Find the total energy contained in a system consisting of two concentric spheres of radii a and b, respectively. Assume that both spheres are perfectly conducting. The inner and outer spheres contains charges of $+Q$ and $-Q$, respectively, and the region between the spheres is filled with a dielectric of permittivity ϵ.

Solution:

The fields generated by this charge distribution have already been determined in Example 5.5:

$$\mathbf{D} = \epsilon\mathbf{E} = \begin{cases} \dfrac{Q}{4\pi r^2}\,\hat{\mathbf{a}}_r & a < r < b \\ 0 & \text{otherwise} \end{cases}$$

Substituting this into Equation (6.33), we obtain

$$W_e = \frac{1}{2}\int_{V_\infty} \mathbf{D}\cdot\mathbf{E}dv = \frac{1}{2}\int_0^{2\pi}\int_0^{\pi}\int_0^{\infty} \mathbf{D}\cdot\mathbf{E}r^2\sin\theta dr d\theta d\phi$$
$$= \frac{1}{2}\int_0^{2\pi}\int_0^{\pi}\int_a^{b} \frac{Q^2\sin\theta}{16\pi^2\epsilon r^2} dr d\theta d\phi = \frac{Q^2}{8\pi\epsilon}\left[\frac{1}{a} - \frac{1}{b}\right].$$

This same result can be obtained by using Equation (6.30). Since both conductors are equipotential surfaces, we can write

$$W_e = \frac{1}{2}\int_{S'} \rho_s V ds = \frac{V_a}{2}\int_{S_a} \rho_{sa} ds + \frac{V_b}{2}\int_{S_b} \rho_{sb} ds$$
$$= \frac{V_a}{2}Q - \frac{V_b}{2}Q,$$

where V_a and V_b are the potentials of the inner and outer spheres, respectively, and Q is the charge on the inner sphere. Since the E-field outside the outer sphere is zero, $V_b = 0$, so V_a equals the voltage between the spheres. Using the capacitance expression that we derived earlier (Equation 6.6), V_a can be written as

$$V_a = Q/C = \frac{Q}{4\pi\epsilon}\left[\frac{1}{a} - \frac{1}{b}\right].$$

Substituting this into the expression for W_e, we obtain

$$W_e = \frac{Q^2}{8\pi\epsilon}\left[\frac{1}{a} - \frac{1}{b}\right],$$

which is the same result as was obtained earlier.

6-4 Energy Storage in Capacitors

Capacitors store charge. This means that they also store energy in the electric fields generated by that charge. To calculate the energy stored in a capacitor, consider the two charged conductors shown in Figure 6-18 with a voltage V between them.

From Equation (6.30), we have

$$W_e = \frac{1}{2}\int_S \rho_s V ds = \frac{1}{2}\int_{S_+} \rho_s V ds + \frac{1}{2}\int_{S_-} \rho_s V ds,$$

where S_+ and S_- are the bounding surfaces of the positive and negative conductors, respectively. Since the potential on each surface is a constant, this expression can be written as

$$W_e = \frac{1}{2}V_+\int_{S_+} \rho_s ds + \frac{1}{2}V_-\int_{S_-} \rho_s ds = \frac{1}{2}V_+Q_+ + \frac{1}{2}V_-Q_-,$$

where V_+ and V_- are the absolute potentials of the positive and negative conductors, respectively, and Q_+ and Q_- are the charges on the positive and negative conductors, respectively. If $Q_+ = -Q_-$, the preceding expression can be written as

$$W_e = \frac{1}{2}Q[V_+ - V_-] = \frac{1}{2}QV, \tag{6.35}$$

where V is the potential difference between the conductors and Q is the capacitor charge. Finally, substituting $C = Q/V$ into Equation (6.35) yields

$$W_e = \frac{1}{2}CV^2. \tag{6.36}$$

Thus, the energy required to charge a capacitor is specified by its capacitance and the voltage to which it is charged.

We can also rearrange Equation (6.36) to provide an alternative definition of the capacitance of a lumped element:

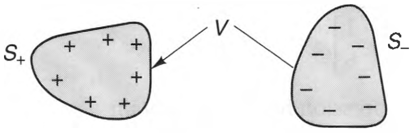


Figure 6-18 Two conductors with balanced charges.

$$C = \frac{2W_e}{V^2}. \tag{6.37}$$

This expression is equivalent to the definition of capacitance given by Equation (6.1). The following example shows how it can be applied to calculate capacitance.

Example 6-5

Find the capacitance of the parallel-plate capacitor shown in Figure 6-19 using the energy definition of capacitance. Assume that the conducting plates have surface area S and are spaced by a distance d.

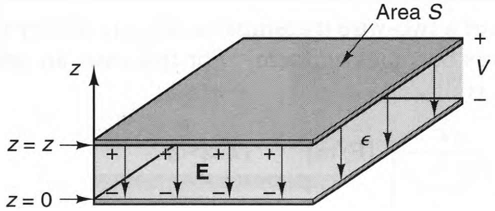


Figure 6-19 A parallel-plate capacitor.

Solution:

When fringing is negligible, the magnitude of **E** between the plates is $E = V/d$. Since the fields are zero outside the capacitor when fringing is negligible, we have, from Equation (6.33),

$$W_e = \frac{1}{2}\int_{V_\infty} \mathbf{D} \cdot \mathbf{E} dv = \frac{1}{2}\epsilon\left(\frac{V}{d}\right)^2 Sd = \frac{\epsilon V^2 S}{2d}.$$

Substituting this into Equation (6.37), we obtain

$$C = \frac{2W_e}{V^2} = \frac{\epsilon S}{d},$$

which is the same formula as was obtained earlier (Equation (6.5)).

6-5 Summation

In this chapter, we have defined the capacitance of an element both in terms of its terminal characteristics and the electric field within it. The terminal relations are simple and well known from circuit theory, but the field relations give greater insight into what physical characteristics of a geometry are responsible for its capacitance. We have also shown that the energy stored by an electrostatic system is completely determined by the charge distributions within the system and their geometric configuration. At the same time, this stored energy state can be determined from a knowledge of the electric field distribution throughout all space.

Even though our discussion of capacitance assumed that the fields were time-invariant, the vast majority of the expressions that were derived in this chapter are applicable to time-varying systems with little or no modification. As a result, we will often refer to these results once our discussion turns to time-varying fields.

PROBLEMS

6-1 A parallel-plate capacitor has two metal plates, each with surface area of 10 [cm^2], spaced by 0.2 [mm], with a uniform dielectric with a dielectric constant of 4.0. Find the capacitance C.

6-2 Calculate the capacitance per meter of a two-wire transmission line with an air dielectric if the radius of each wire is 1 [mm] and the center-to-center spacing is 1 [cm].

6-3 When the wires of a two-wire transmission line are widely spaced, the charge density on both wires becomes uniform. For this case, an approximate formula for the capacitance is

$$C \approx \frac{\pi\epsilon}{\ln\left[\dfrac{D-a}{a}\right]} \text{ [F/m]} \qquad D \gg a.$$

(a) Derive this formula by using the E-field expression for infinite, uniform cylinders of charge.

(b) Compare the results of the formula with the exact formula (Equation (6.19)) when $D/a = 10.0$ and 3.0.

6-4 Consider a capacitor formed by two coaxial metal cylinders with radii 8 [mm] and 2 [mm], respectively. Find the capacitance per meter, C, if the relative dielectric permittivity between the cylinders is $\epsilon_r = 4 + 2\rho$, where ρ is measured in millimeters.

6-5 Find the center-to-center spacing of a two-wire transmission line that has a capacitance per meter of 10 [pF/m]. Assume that the wires have radii of 0.5 [mm] and that the dielectric is air.

6-6 Calculate the capacitance of the earth by considering the earth to be a perfectly conducting sphere (radius ≈ 6371 [km]). Compare your result with the capacitance of commercially available capacitors. Is this surprising?

6-7 Use the curvilinear squares technique to estimate the capacitance of an air-filled coaxial line with an outer-conductor/inner-conductor ratio of 4.0. Compare this result with Equation (6.12).

6-8 Use the curvilinear squares technique to estimate the capacitance per meter of the strip-line transmission line shown in Figure P6-8, which consists of a narrow strip, sandwiched between two outer conductors that are both at ground potential. (*Hint*: This geometry has dual symmetry about the center point of the strip, so only the field lines in a single quadrant need to be drawn.)

6-9 Figure P6-9 shows a square cylinder surrounded by a larger square cylinder. Use the curvilinear squares technique to estimate the capacitance per meter of this

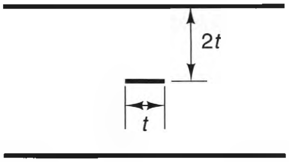


Figure P6-8

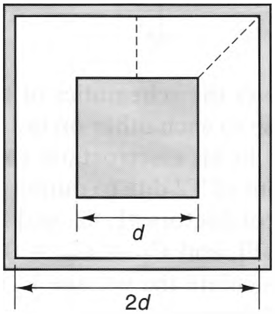


Figure P6-9

geometry. Notice that the symmetry of the geometry is such that only the lines in the upper octant need to be drawn.

6-10 A capacitor with a uniform dielectric has a capacitance of .02 [μF] and a resistance of 10 [KΩ]. If the dielectric constant is $\epsilon_r = 40$, find the conductivity σ of the dielectric.

6-11 Prove that the energy W_e necessary to assemble a uniformly charged solid sphere with total charge Q and radius a is

$$W_e = \frac{3Q^2}{20\pi\epsilon_0 a}.$$

(*Hint*: Find the E-field inside and outside the sphere, and use Equation (6.34).)

6-12 After the discovery of the electron, attempts were made to calculate the electron's radius by assuming that all of the energy stored in an electron is electromagnetic and then equating this energy with Einstein's equation, $E = mc^2$, where E is the electron mass energy, m is its rest mass, and c is the speed of light in a vacuum. Under different assumptions of how the charge within the electron is distributed, various radii were calculated, all of which contained the quantity

$$r_0 = \frac{e^2}{4\pi\epsilon_0 mc^2} = 2.81784 \times 10^{-15} \qquad [\text{m}],$$

which came to be called the *classical radius* of the electron.[4] If an electron is assumed to be a uniformly charged spherical shell, use this reasoning to show that the resulting electron radius is $r_0/2$.

[4] See Robert Leighton, *Principles of Modern Physics* (New York: McGraw-Hill, 1959).

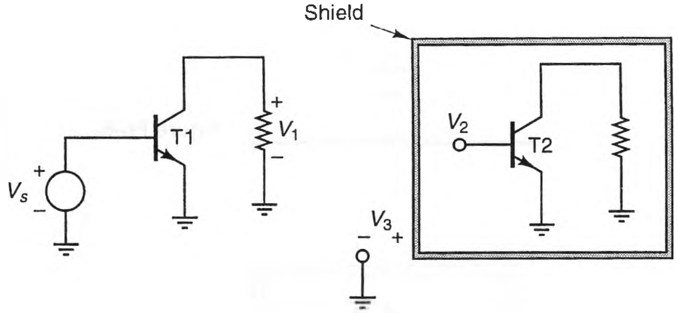


Figure P6-13

6-13 Figure P6-13 shows the schematics of two transistor amplifiers, T1 and T2, that are mounted close to each other on the same circuit board. The amplifier on the right is enclosed in an electrostatic shield in a effort to reduce the capacitive pickup at the input of T2 due to output voltage swing of T1. The mutual capacitances between conductors #1, #2, and #3 and ground are $C_{13} = 0.1$ [pF], $C_{23} = 2$ [pF], $C_{30} = 10$ [pF], and $C_{12} = C_{20} = 0$. If V_1 varies sinusoidally between $+10$ and -10 volts, calculate the voltage V_2 if:

(a) the shield is grounded,

(b) the shield is allowed to float.

7

Magnetostatic Fields in Free Space

7-1 Introduction

Just as static charge distributions generate electrostatic fields, steady current distributions generate magnetostatic fields. In addition to being a natural consequence of electric currents, magnetostatic fields are desirable in their own right, particularly for generating large forces. Devices that use magnetostatic fields include motors, generators, relays, and electromagnets.

Magnetostatic fields are a convenient starting point towards our ultimate goal of understanding the nature of all magnetic fields (both static and time varying), since the equations that describe this special case are much simpler than those describing the general time-varying case. In addition to their simplicity, many other of the characteristics of magnetostatic fields are seen in time-varying magnetic fields. As a result, we will use many of the concepts and formulas from this discussion later in the text when we discuss the time-varying case.

Just as in the case of electrostatic fields, we will start our discussion of magnetostatic fields by identifying the behavior of these fields when currents are suspended in free space. Starting here will allow us to see the relationship between the fields and their sources, without the effects of materials. In addition to developing a number of formulas that relate the magnetic flux density **B** to the currents that cause them, we will introduce two types of magnetic potentials that are useful in the analysis of magnetic systems.

7-2 Maxwell's Equations for Magnetostatics in Free Space

In Chapter 3, we found that Maxwell's equations are the fundamental postulates of all electromagnetic phenomena. In free space, the point form of these equations read

$$\nabla \cdot \mathbf{B} = 0 \qquad \nabla \times \mathbf{B} = \mu_o \mathbf{J} + \mu_o \epsilon_o \frac{\partial \mathbf{E}}{\partial t}$$

$$\nabla \cdot \mathbf{E} = \frac{\rho_v}{\epsilon_o} \qquad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}.$$

When the sources are time invariant, all derivatives with respect to t vanish, and these equations become

$$\nabla \cdot \mathbf{B} = 0 \quad (7.1) \qquad \nabla \times \mathbf{B} = \mu_o \mathbf{J} \quad (7.2)$$

$$\nabla \cdot \mathbf{E} = \frac{\rho_v}{\epsilon_o} \quad (7.3) \qquad \nabla \times \mathbf{E} = 0. \quad (7.4)$$

Equations (7.3) and (7.4) define the behavior of electrostatic fields; these were discussed in Chapter 4. Equations (7.1) and (7.2) define the behavior of magnetostatic fields and are called ***Maxwell's equations for magnetostatics in free space.*** Similar to the electrostatic equations, these equations define the magnetic flux density **B** at any point, since both the divergence and curl of **B** are defined. However, unlike the electrostatic equations, where **E** always has zero curl, in the magnetostatic equations it is the divergence of **B** that is always zero. This is the cause of many differences between the behaviors of electrostatic and magnetostatic fields.

We can derive integral representations of the magnetostatic equations. To do this, let us first take the dot product of both sides of Equation (7.2) with a differential surface vector **ds** and integrate over an arbitrary, *open* surface S, yielding

$$\int_S \nabla \times \mathbf{B} \cdot \mathbf{ds} = \mu_o \int_S \mathbf{J} \cdot \mathbf{ds}.$$

The right-hand side of this expression equals the current I passing through S in a right-handed sense. Also, we can use Stokes's theorem to express the surface integral on the left-hand side as a line integral over the *closed* contour C that bounds S, resulting in

$$\oint_C \mathbf{B} \cdot \mathbf{d\ell} = \mu_o I,$$

which is valid for any closed contour C.

Similarly, we can obtain an integral representation of Equation (7.1) by first multiplying both sides by the differential volume element dv and integrating over some volume V, yielding

$$\int_V \nabla \cdot \mathbf{B}\, dv = 0.$$

We can use the divergence theorem to write the integral on the left as a surface integral over the closed surface S that surrounds V:

$$\oint_S \mathbf{B} \cdot \mathbf{ds} = 0.$$

Maxwell's equations for magnetostatic fields in free space are summarized in the following table in both point and integral forms.

MAXWELL'S EQUATIONS FOR MAGNETOSTATICS IN FREE SPACE

Point Form:		Integral Form:	
$\nabla \cdot \mathbf{B} = 0$	(7.5)	$\oint_S \mathbf{B} \cdot \mathbf{ds} = 0$	(7.6)
$\nabla \times \mathbf{B} = \mu_o \mathbf{J}$	(7.7)	$\oint_C \mathbf{B} \cdot \mathbf{d\ell} = \mu_o I$	(7.8)

In either integral or point form, these equations completely define the behavior of magnetostatic fields in free space. Equations (7.7) and (7.8) define the relationship between steady current distributions and the magnetostatic B-fields they generate. These equations are representations of ***Ampère's circuital law*** (or simply ***Ampère's law***), named in honor of André Marie Ampère (1775–1836), who conducted a definitive series of experiments that unraveled the mystery of magnetostatic fields. In words, Ampère's law states that the magnetic flux density **B** tends to rotate around currents. We will use this law to find the magnetic fields generated by several classes of current distributions.

Equations (7.5) and (7.6) state another important property of magnetostatic fields, namely, that the net magnetic flux entering or leaving any point or surface is always zero. This is often called ***Gauss' law for magnetics***, or the ***law of conservation of magnetic flux***. An important consequence of this law is that B-field streamlines, unlike E-field streamlines, never terminate on points or surfaces. This is because unipolar magnetic charges do not exist in nature. Rather, the elemental sources of magnetic fields are electric currents, which act as bipolar sources of magnetic fields. A common example of this is permanent magnets, which always have north and south poles. Even when one is cut in half, each half has its own north and south poles.

7-3 The Biot-Savart Law and the Magnetic Vector Potential

Maxwell's equations for magnetostatics are the most compact formulation of the laws of magnetostatics, but that does not mean that they are always the most convenient starting point for actual magnetic field calculations. This is analogous to electrostatics, where we found it helpful to derive Coulomb's law from Maxwell's electrostatics equations. Using Coulomb's law, we were able to calculate the E-fields generated by a number of different types of charge distributions. In the case of magnetostatics, the Biot-Savart law is analogous to Coulomb's law in that it allows the direct calculation of the B-fields generated by a given current distribution.

Deriving the Biot-Savart law from Maxwell's magnetostatic equations is a bit more involved than deriving Coulomb's law from Maxwell's electrostatic equations. This is because the sources of magnetic fields are currents, which are vector quantities, as opposed to charges, which are scalar quantities. Because of this, the Biot-Savart law is most easily derived by first introducing the concept of the magnetic vector potential. Once this is introduced, the Biot-Savart law will follow quickly.

7-3-1 THE MAGNETIC VECTOR POTENTIAL

In electrostatics, we were able to use the fact that $\nabla \times \mathbf{E}$ is always zero to express **E** as the gradient of the electrostatic scalar potential V. In addition to its physical interpretation, we found that this potential often simplifies electrostatic calculations. In magnetostatics, Gauss' law for magnetostatics (Equation (7.5)) provides an analogous situation. Since **B** has zero divergence at all points, Theorem IV in Section 2-5-6 (Equation (2.132)) allows us to write **B** in terms of the curl of another vector, that is,

$$\mathbf{B} = \nabla \times \mathbf{A}, \tag{7.9}$$

where **A** is called the ***magnetic vector potential***, which is measured in units of webers per meter [Wb/m] or tesla-meters [T • m].

Although Equation (7.9) tells us of the existence of **A**, it does not tell us how to calculate it for a known current distribution. We can accomplish this by first substituting Equation (7.9) into Equation (7.7), which yields

$$\nabla \times \nabla \times \mathbf{A} = \mu_0 \mathbf{J}.$$

We can expand the curl-curl operation using Equation (B.10), from which we obtain

$$\nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} = \mu_0 \mathbf{J}, \tag{7.10}$$

where ∇^2 is the Laplacian operator, which is defined in Equations (2.126) and (2.128).

Before we attempt to solve Equation (7.10) for the magnetic vector potential **A**, we would benefit from a bit of reflection. In particular, we note that the left-hand side of Equation (7.10) has the term $\nabla(\nabla \cdot \mathbf{A})$. This term would certainly be zero if $\nabla \cdot \mathbf{A}$ were zero, but do we have the right to assume that this is so? The answer to this question is *yes*, since, according to Equation (7.9), the only property of **A** that is important to us is its curl. As a result, we can choose its divergence to be anything that we find convenient, as the divergence and curl of a vector are independent quantities that can be specified separately. (This is a corollary of Theorem II in Section 2-5-6.) The process of selecting a specific divergence of **A** is called "setting the gauge," and we will choose

$$\nabla \cdot \mathbf{A} = 0, \tag{7.11}$$

which is called ***Coulomb's gauge*** and reduces Equation (7.10) to

$$\nabla^2 \mathbf{A} = -\mu_0 \mathbf{J}. \tag{7.12}$$

This equation is the vector form of Poisson's equation.

The easiest way to solve the vector Poisson's equation is to use the fact that the ∇^2 operator is particularly simple when the vector it is acting on is expressed in Cartesian components. (See Equation (2.126).) Thus, the vector Poisson's equation can be written as three scalar equations,

$$\nabla^2 A_x = -\mu_0 J_x \tag{7.13a}$$

$$\nabla^2 A_y = -\mu_0 J_y \tag{7.13b}$$

$$\nabla^2 A_z = -\mu_0 J_z, \tag{7.13c}$$

where A_x, A_y, and A_z are the x, y, and z components of **A**, respectively, and J_x, J_y, and J_z are the x, y, and z components of **J**. Each of these equations is a scalar Poisson's equation, of the same form as the Poisson's equation found in electrostatics: $\nabla^2 V = -\rho_v/\epsilon_0$. We can express the solutions of these three magnetostatic Poisson's equations using the electrostatic solution (Equation (4.60)) by replacing ρ_v with J_i and $1/\epsilon_0$ with μ_0, obtaining

$$A_i = \frac{\mu_0}{4\pi} \int_{\text{Vol.}} \frac{J_i dv'}{|\mathbf{r} - \mathbf{r}'|} \qquad i = x, y, z.$$

Here, **r** is the field point (i.e., the point at which A_i is being evaluated), and $\mathbf{r}'$ is the dummy position vector that sweeps through all the points where the current density is nonzero. Since the solutions for each of the components of **A** have similar integrands, we can combine them to form the vector expression,

$$\mathbf{A} = \frac{\mu_0}{4\pi} \int_{\text{Vol.}} \frac{\mathbf{J} dv'}{|\mathbf{r} - \mathbf{r}'|}. \tag{7.14}$$

Even though we derived this solution by using the Cartesian components of the vectors, it is independent of the coordinate system chosen. Hence, the expression is the particular solution[1] of Poisson's vector equation in all coordinate systems.

[1] The particular solution of a differential equation is one that is proportional to the forcing function (in this case, **J**) and has no arbitrary constants.

We will find a number of uses for the magnetic vector potential **A** throughout this chapter. For the moment, however, our only concern is that we can use it to quickly derive the Biot-Savart law.

7-3-2 THE BIOT-SAVART LAW

Let us consider a steady current I that flows around an arbitrary, closed path C, such as the path shown in Figure 7-1. Since this is a filamentary current, we can modify Equation (7.14) to express the magnetic vector potential of this current by replacing $\mathbf{J}dv'$ with $I\mathbf{d\ell}'$, yielding

$$\mathbf{A} = \frac{\mu_o}{4\pi}\oint_C \frac{I\mathbf{d\ell}'}{|\mathbf{r}-\mathbf{r}'|},$$

where **r** and **r**′ are the position vectors of the field and source points, respectively. To keep things simple, we have chosen the direction of integration $\mathbf{d\ell}'$ and the direction of the current I to be the same. Substituting the foregoing expression into Equation (7.9), we obtain

$$\mathbf{B} = \nabla \times \frac{\mu_o}{4\pi}\oint_C \frac{I\mathbf{d\ell}'}{|\mathbf{r}-\mathbf{r}'|} = \frac{\mu_o I}{4\pi}\oint_C \nabla \times \frac{\mathbf{d\ell}'}{|\mathbf{r}-\mathbf{r}'|}. \tag{7.15}$$

In this expression, we note that we can bring the $\nabla \times$ inside the integral, because it operates only on the primed variables.

The easiest way to handle the curl operation inside the preceding integral is to use the vector identity

$$\nabla \times (f\mathbf{G}) = f\nabla \times \mathbf{G} + (\nabla f) \times \mathbf{G}.$$

If we let $f = \dfrac{1}{|\mathbf{r}-\mathbf{r}'|}$ and $\mathbf{G} = \mathbf{d\ell}'$, we can write

$$\nabla \times \frac{\mathbf{d\ell}'}{|\mathbf{r}-\mathbf{r}'|} = \frac{1}{|\mathbf{r}-\mathbf{r}'|}\nabla \times \mathbf{d\ell}' + \left(\nabla\frac{1}{|\mathbf{r}-\mathbf{r}'|}\right) \times \mathbf{d\ell}'.$$

In Cartesian coordinates, $\mathbf{d\ell}'$ can be expressed as

$$\mathbf{d\ell}' = dx'\,\hat{\mathbf{a}}_x + dy'\,\hat{\mathbf{a}}_y + dz'\,\hat{\mathbf{a}}_z.$$

This clearly shows that $\mathbf{d\ell}'$ is not a function of the unprimed variables, so $\nabla \times \mathbf{d\ell}' = 0$ and

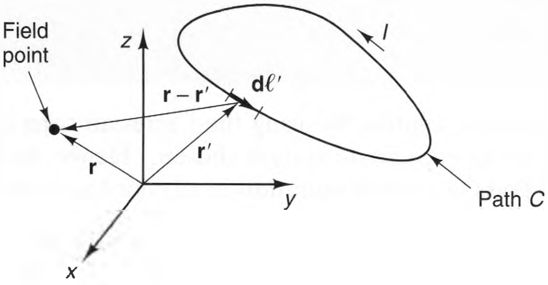


Figure 7-1 Geometry for deriving the Biot-Savart law.

$$\nabla \times \frac{d\boldsymbol{\ell}'}{|\mathbf{r}-\mathbf{r}'|} = \left(\nabla \frac{1}{|\mathbf{r}-\mathbf{r}'|}\right) \times d\boldsymbol{\ell}'. \tag{7.16}$$

Next, we can express the scalar function $1/|\mathbf{r}-\mathbf{r}'|$ in Cartesian coordinates as

$$\frac{1}{|\mathbf{r}-\mathbf{r}'|} = [(x-x')^2 + (y-y')^2 + (z-z')^2]^{-1/2}.$$

Taking the gradient of this function, we have

$$\begin{aligned}\nabla \frac{1}{|\mathbf{r}-\mathbf{r}'|} &= \hat{\mathbf{a}}_x \frac{\partial}{\partial x}\left(\frac{1}{|\mathbf{r}-\mathbf{r}'|}\right) + \hat{\mathbf{a}}_y \frac{\partial}{\partial y}\left(\frac{1}{|\mathbf{r}-\mathbf{r}'|}\right) + \hat{\mathbf{a}}_z \frac{\partial}{\partial z}\left(\frac{1}{|\mathbf{r}-\mathbf{r}'|}\right)\\ &= -\frac{(x-x')\hat{\mathbf{a}}_x + (y-y')\hat{\mathbf{a}}_y + (z-z')\hat{\mathbf{a}}_z}{[(x-x')^2 + (y-y')^2 + (z-z')^2]^{3/2}}.\end{aligned}$$

This can be written in terms of the position vectors **r** and **r**′ to read

$$\nabla \frac{1}{|\mathbf{r}-\mathbf{r}'|} = -\frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|^3} \tag{7.17}$$

Substituting this result into Equation (7.16), we obtain

$$\nabla \times \frac{d\boldsymbol{\ell}'}{|\mathbf{r}-\mathbf{r}'|} = -\frac{(\mathbf{r}-\mathbf{r}')\times d\boldsymbol{\ell}'}{|\mathbf{r}-\mathbf{r}'|^3} = \frac{d\boldsymbol{\ell}' \times (\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3}. \tag{7.18}$$

Substituting Equation (7.18) into Equation (7.15), we obtain the ***Biot-Savart law for filamentary currents***:

$$\mathbf{B} = \frac{\mu_o I}{4\pi}\oint_C \frac{d\boldsymbol{\ell}' \times (\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3} \quad [\mathrm{T}] \quad \text{(The Biot-Savart law for steady, filamentary currents in free space).} \tag{7.19}$$

The Biot-Savart law is an explicit equation for the magnetic flux density **B** in terms of a steady current I and the path C that it flows on. To better understand the nature of the fields predicted by the Biot-Savart law, let us first consider the B-field that is generated by a differential segment of a complete current loop. According to Equation (7.19), a short current element $Id\boldsymbol{\ell}'$ that is located at **r**′ generates a field

$$d\mathbf{B} = \frac{\mu_o I d\boldsymbol{\ell}' \times (\mathbf{r}-\mathbf{r}')}{4\pi|\mathbf{r}-\mathbf{r}'|^3} = \frac{\mu_o I}{4\pi R^2} d\boldsymbol{\ell}' \times \hat{\mathbf{a}}_R \qquad [\mathrm{T}], \tag{7.20}$$

where R is the distance from the current element to the observer (located at **r**) and the unit vector

$$\hat{\mathbf{a}}_R = \frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|} \tag{7.21}$$

points from the current element to the observer. Like the E-field generated by a point source, the B-field generated by a short current element of a steady current loop

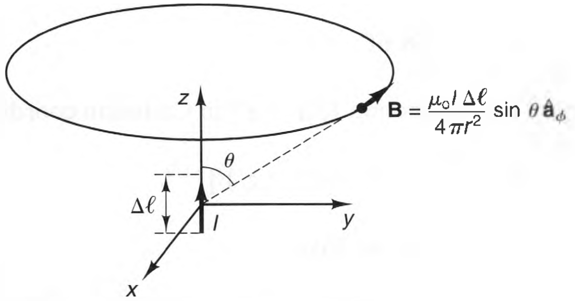

z
$\mathbf{B} = \frac{\mu_o I\,\Delta\ell}{4\pi r^2}\sin\theta\,\hat{\mathbf{a}}_\phi$
θ
Δℓ
I
y
x

Figure 7-2 The B-field generated by a short filament of current.

varies as the inverse square of the distance to the observer. But unlike the electrostatic case, where **E** is always directed outward from point sources, the B-field of a short current source is perpendicular to both the current direction and the line directed from the current to the observer.

Figure 7.2 shows a current element, directed along the z-axis at the origin, with length $\Delta\ell$. For this case, we have $\mathbf{d\ell}' = \Delta\ell\hat{\mathbf{a}}_z$ and $\hat{\mathbf{a}}_R = \hat{\mathbf{a}}_r$. so Equation (7.20) becomes

$$\mathbf{B} = \frac{\mu_o I\Delta\ell}{4\pi r^2}\sin\theta\,\hat{\mathbf{a}}_\phi \qquad [\mathrm{T}], \tag{7.22}$$

where (r, θ, ϕ) are the spherical coordinates of the observer. This is the magnetostatic equivalent of a point charge located at the origin. The resulting B-field is similar to the electrostatic E-field in that both fields are proportional to the magnitude of the source (I, in this case) and to $1/r^2$. But whereas the E-field of a point charge is directed outward from the charge, the B-field of a short current element circulates around the current according to the right-hand rule. Also, unlike the point charge, where $|\mathbf{E}|$ does not vary with the observer's angular position about the charge, $|\mathbf{B}|$ is proportional to the sine of the angle the observer makes with the filament.

We can extend the Biot-Savart law to include steady-current distributions that flow on surfaces and in volumes. For a surface current, we can replace $I\mathbf{d\ell}'$ with $\mathbf{J}_s ds'$ in Equation (7.19), yielding

$$\mathbf{B} = \frac{\mu_o}{4\pi}\int_S \frac{\mathbf{J}_s \times (\mathbf{r} - \mathbf{r}')\,ds'}{|\mathbf{r} - \mathbf{r}'|^3} \qquad [\mathrm{T}] \tag{7.23}$$

(The Biot-Savart law for steady surface currents in free space).

Notice here that the surface current density $\mathbf{J}_s$ must be placed inside the integral, since both the magnitude and direction of the current density may vary with position. Similarly, for a volumetric current, we have

$$\mathbf{B} = \frac{\mu_o}{4\pi}\int_{\text{Vol.}} \frac{\mathbf{J} \times (\mathbf{r} - \mathbf{r}')\,dv'}{|\mathbf{r} - \mathbf{r}'|^3} \qquad [\mathrm{T}] \tag{7.24}$$

(The Biot-Savart law for steady volumetric currents in free space).

The B-field of any steady-current distribution can be calculated directly from the Biot-Savart law. For most current distributions, these integrals must be performed numerically. Nevertheless, there exist several special cases in which the integrals can be performed in closed form. In the section that follows, we will use the Biot-Savart law to determine the B-fields generated by several types of steady-current distributions.

7-3-3 THE UNIFORM, INFINITE LINE OF CURRENT

Figure 7-3 shows a uniform, steady current that flows in the $+\hat{\mathbf{a}}_z$ direction along the entire z-axis. We will assume that this current is part of a circuit that is completed at infinity. To evaluate $\mathbf{B}$, we first note that the position vectors representing the field point $\mathbf{r}$ and source points $\mathbf{r}'$ can be represented in cylindrical coordinates by

$$\mathbf{r} = \rho\hat{\mathbf{a}}_\rho + z\hat{\mathbf{a}}_z$$

$$\mathbf{r}' = z'\hat{\mathbf{a}}_z,$$

from which we obtain

$$|\mathbf{r} - \mathbf{r}'|^3 = [\rho^2 + (z - z')^2]^{3/2}$$

$$\mathbf{d\ell}' = dz'\hat{\mathbf{a}}_z.$$

Substituting these into Equation (7.19), we have

$$\mathbf{B} = \frac{\mu_o I}{4\pi}\int_{-\infty}^{\infty} \frac{\hat{\mathbf{a}}_z \times [\rho\hat{\mathbf{a}}_\rho + (z - z')\hat{\mathbf{a}}_z]dz'}{[\rho^2 + (z - z')^2]^{3/2}}.$$

Since $\hat{\mathbf{a}}_z \times \hat{\mathbf{a}}_z = 0$, $\hat{\mathbf{a}}_z \times \hat{\mathbf{a}}_\rho = \hat{\mathbf{a}}_\phi$, and $\hat{\mathbf{a}}_\phi$ does not vary with z', the preceding expression can be written as

$$\mathbf{B} = \frac{\mu_o I}{4\pi}\int_{-\infty}^{\infty} \frac{\rho\hat{\mathbf{a}}_\phi dz'}{[\rho^2 + (z - z')^2]^{3/2}} = \frac{\mu_o I\rho\hat{\mathbf{a}}_\phi}{4\pi}\int_{-\infty}^{\infty} \frac{dz'}{[\rho^2 + (z - z')^2]^{3/2}}$$

$$= \frac{\mu_o I\rho\hat{\mathbf{a}}_\phi}{4\pi}\left[\frac{-(z - z')}{\rho^2\sqrt{\rho^2 + (z - z')^2}}\right]\Bigg|_{z'=-\infty}^{z'=\infty} = \frac{\mu_o I\rho\hat{\mathbf{a}}_\phi}{4\pi}\left[\frac{1}{\rho^2} + \frac{1}{\rho^2}\right].$$

Simplifying this expression, we obtain

$$\mathbf{B} = \frac{\mu_o I}{2\pi\rho}\hat{\mathbf{a}}_\phi \quad \text{(Infinite line current).} \qquad (7.25)$$

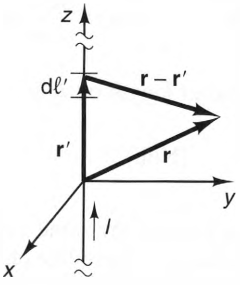


Figure 7-3 Geometry for determining the B-field generated by an infinite line of current.

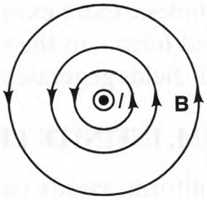


Figure 7-4 Cross-sectional view of the B-field streamlines generated by an infinite line of current.

A cross section of the B-field streamlines for an infinite line current is shown in Figure 7-4. Like the E-field generated by an infinite line charge, this field is proportional to ρ^{-1} and independent of both ϕ and z. But the B-field circulates around its source in a right-handed sense, rather than pointing away from it, as it does for the line charge.

Example 7-1

Find the force per unit length acting on the long, parallel lines of steady current shown in Figure 7-5.

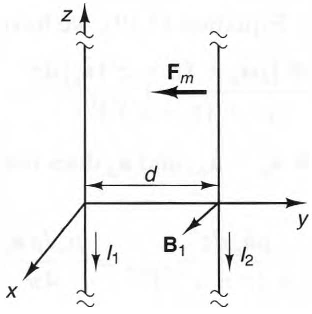


Figure 7-5 Two parallel, infinite lines of current.

Solution:

Since I_1 lies along the z-axis, we can use Equation (7.25) to evaluate the field $\mathbf{B}_1$ that this current imposes upon I_2. Noting that the $-\hat{\mathbf{a}}_\phi$ direction is the same as the $+\hat{\mathbf{a}}_x$ direction along the positive y-axis, we can write

$$\mathbf{B}_1 = \frac{\mu_0 I_1}{2\pi d}\hat{\mathbf{a}}_x$$

According to Equation (3.46), the magnetic force $\mathbf{dF}_m$ that acts on each differential element of I_2 is

$$\mathbf{dF}_m = I_2 \mathbf{d\ell}_2 \times \mathbf{B}_1 = -I_2 dz\,\hat{\mathbf{a}}_z \times \frac{\mu_0 I_1}{2\pi d}\hat{\mathbf{a}}_x = -\frac{\mu_0 I_1 I_2 dz}{2\pi d}\hat{\mathbf{a}}_y \qquad \text{[N]}.$$

Thus, the force $\mathbf{F}_m$ per meter length on I_2 is

$$\mathbf{F}_m = -\frac{\mu_0 I_1 I_2}{2\pi d}\hat{\mathbf{a}}_y \qquad \text{[N/m]}. \tag{7.26}$$

Notice that when I_1 and I_2 have the same sign, the force is attractive. This is the opposite of what occurs with point charges, where like charges repel.

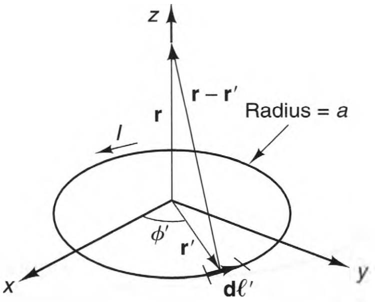


Figure 7-6 Geometry for determining the B-field of a circular loop of current.

7-3-4 CIRCULAR LOOPS

Figure 7-6 shows a loop of radius a that lies in the xy plane and carries a steady current I.

For an arbitrary observation point on the z-axis, the coordinate variables and differentials needed to evaluate the Biot-Savart law are

$$\mathbf{r} = z\hat{\mathbf{a}}_z$$
$$\mathbf{r}' = a\hat{\mathbf{a}}_{\rho'} = a\cos\phi'\hat{\mathbf{a}}_x + a\sin\phi'\hat{\mathbf{a}}_y$$
$$\mathbf{r} - \mathbf{r}' = z\hat{\mathbf{a}}_z - a\cos\phi'\hat{\mathbf{a}}_x - a\sin\phi'\hat{\mathbf{a}}_y$$
$$|\mathbf{r} - \mathbf{r}'|^3 = [z^2 + a^2]^{3/2}$$
$$\mathbf{d\ell}' = a\,d\phi'\hat{\mathbf{a}}_{\phi'} = -a\,d\phi'\sin\phi'\hat{\mathbf{a}}_x + a\,d\phi'\cos\phi'\hat{\mathbf{a}}_y.$$

Also,

$$\mathbf{d\ell}' \times (\mathbf{r} - \mathbf{r}') = (az\cos\phi'\hat{\mathbf{a}}_x + az\sin\phi'\hat{\mathbf{a}}_y + a^2\hat{\mathbf{a}}_z)\,d\phi'.$$

Substituting these terms into Equation (7.19), we obtain

$$\mathbf{B} = \frac{\mu_0 I}{4\pi}\int_0^{2\pi}\frac{(az\cos\phi'\hat{\mathbf{a}}_x + az\sin\phi'\hat{\mathbf{a}}_y + a^2\hat{\mathbf{a}}_z)\,d\phi'}{[z^2 + a^2]^{3/2}}.$$

The unit vectors in this expression are all constants with respect to ϕ' and can be taken outside the integral. Upon integrating, we find that the x and y components of **B** are zero, since the integrals of the sine and cosine functions over a complete circle are zero. The integral associated with the z component of **B** is trivial, because the integrand is independent of ϕ'. The final expression for **B** is

$$\mathbf{B} = B_z\hat{\mathbf{a}}_z = \frac{\mu_0 a^2 I}{2[z^2 + a^2]^{3/2}}\hat{\mathbf{a}}_z. \tag{7.27}$$

The variation of B_z with z is shown in Figure 7-7.

For large values of z, B_z decays proportionally to z^{-3}, rather than z^{-2}. The reason for this is that, viewed from a large distance, a loop appears to be composed of current-element pairs that are oppositely directed and very close together. As a result, their field contributions tend to cancel, much like the E-field contributions from the opposite charges of an electric dipole.

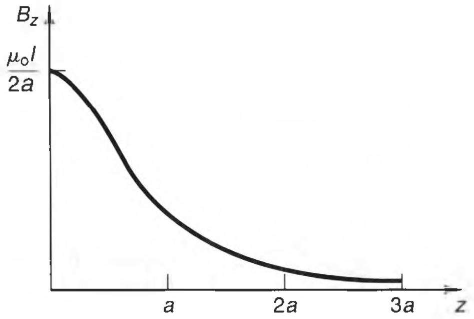


Figure 7-7 Plot of B_z vs. height above a circular loop of current of radius a.

Equation (7.27) is an exact expression, but it is valid only for observation points on the axis of the loop. For all other points, the resulting integrals are very difficult to evaluate analytically[2] Later in this chapter we will return to the current loop and develop a simple, approximate expression for **B** using the magnetic vector potential.

7-3-5 FLAT STRIPS AND SHEETS OF CURRENT

Another problem that is easily solved using the Biot-Savart law is depicted in Figure 7-8. Here, a surface current of value $\mathbf{J}_s = J_x\hat{\mathbf{a}}_x$ flows along a flat strip of width $2a$ that is centered along the x-axis. If we consider an arbitrary observation point along the z-axis, the position variables and differential quantities necessary to evaluate the Biot-Savart law are

$$\begin{aligned}
\mathbf{r} &= z\hat{\mathbf{a}}_z \\
\mathbf{r}' &= x'\hat{\mathbf{a}}_x + y'\hat{\mathbf{a}}_y \\
|\mathbf{r} - \mathbf{r}'|^3 &= [x'^2 + y'^2 + z^2]^{3/2} \\
\mathbf{J}_s \times (\mathbf{r} - \mathbf{r}') &= J_x\hat{\mathbf{a}}_x \times (z\hat{\mathbf{a}}_z - x'\hat{\mathbf{a}}_x - y'\hat{\mathbf{a}}_y) = -J_x(y'\hat{\mathbf{a}}_z + z\hat{\mathbf{a}}_y) \\
ds' &= dx'dy'.
\end{aligned}$$

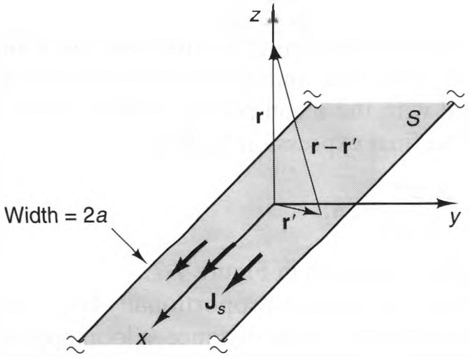


Figure 7-8 Geometry for determining the B-field of an infinite strip of current.

[2] For a complete derivation, see W. R. Smythe, *Static and Dynamic Electricity* (New York, London: McGraw-Hill, 2d. ed), 1939 p. 270.

Substituting these into Equation (7.23), we obtain

$$\mathbf{B} = \frac{-\mu_0 J_x}{4\pi} \int_{-a}^{a} \int_{-\infty}^{\infty} \frac{y'\hat{\mathbf{a}}_z + z\hat{\mathbf{a}}_y}{[x'^2 + y'^2 + z^2]^{3/2}}\, dx'dy'.$$

We can move the constant unit vectors out of the integrals, obtaining

$$\mathbf{B} = -\frac{\mu_0 J_x}{4\pi} \left[\hat{\mathbf{a}}_z \int_{-a}^{a} \int_{-\infty}^{\infty} \frac{y'}{[x'^2 + y'^2 + z^2]^{3/2}}\, dx'dy' \right.$$

$$\left. + z\hat{\mathbf{a}}_y \int_{-a}^{a} \int_{-\infty}^{\infty} \frac{1}{[x'^2 + y'^2 + z^2]^{3/2}}\, dx'dy' \right],$$

The z-component of $\mathbf{B}$ is zero, since its integrand is an odd function of y' and the limits of integration are symmetric about $y' = 0$. Evaluating the y-component of $\mathbf{B}$, we find

$$\mathbf{B} = -\frac{\mu_0 J_x z\hat{\mathbf{a}}_y}{4\pi} \int_{-a}^{a} \left[\frac{x'}{(y'^2 + z^2)\sqrt{x'^2 + y'^2 + z^2}} \Bigg|_{x'=-\infty}^{x'=\infty} \right] dy'$$

$$= -\frac{\mu_0 J_x z\hat{\mathbf{a}}_y}{2\pi} \int_{-a}^{a} \frac{dy'}{y'^2 + z^2} = -\frac{\mu_0 J_x \hat{\mathbf{a}}_y}{2\pi} \tan^{-1}\left[\frac{y'}{z}\right] \Bigg|_{y'=-a}^{y'=a}.$$

Thus, we finally obtain

$$\mathbf{B} = B_y \hat{\mathbf{a}}_y = -\frac{\mu_0 J_x \hat{\mathbf{a}}_y}{\pi} \tan^{-1}\left[\frac{a}{z}\right]. \tag{7.28}$$

Hence, the B-field directly above a uniform strip of steady current is directed perpendicular to the direction of the current. A plot of B_y vs. z is shown in Figure 7-9.

To see how B_y behaves for large values of z, we can use the approximation $\tan^{-1}\beta \approx \beta$ when β is small. Hence, as $z \to \infty$, we obtain

$$\lim_{z\to\infty} \mathbf{B} = -\frac{\mu_0 J_x \hat{\mathbf{a}}_y}{\pi} \lim_{z\to\infty} \left(\tan^{-1}\left[\frac{a}{z}\right] \right) = -\frac{\mu_0 I \hat{\mathbf{a}}_y}{2\pi z}, \tag{7.29}$$

where $I = 2aJ_x$ is the total current carried by the strip. Comparing Equation (7.29) with the field generated by an infinite line current (Equation (7.25)) we see that, for large values of z, the fields generated by a uniform strip and an infinite line current look the same.

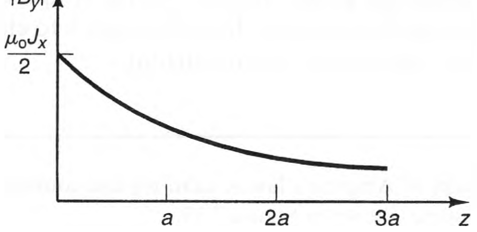


Figure 7-9 Plot of B_y vs. height above an infinite strip of current of width $2a$.

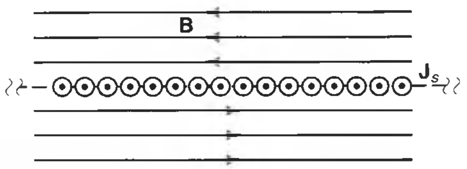


Figure 7-10 Cross-sectional view of the **B**-field streamlines generated by an infinite sheet of current.

We can also use Equation (7.28) to determine the B-field generated by a uniform, infinite sheet of current. If we take the limit as $a \to \infty$, we have

$$\lim_{a\to\infty} \mathbf{B} = -\frac{\mu_o J_x \hat{\mathbf{a}}_y}{\pi} \lim_{a\to\infty} \left(\tan^{-1}\left[\frac{a}{z}\right]\right).$$

Noting that $\tan^{-1}(\pm\infty) = \pm\frac{\pi}{2}$, we obtain

$$\mathbf{B} = \begin{cases} -\dfrac{\mu_o J_x}{2}\hat{\mathbf{a}}_y & z > 0 \\ \dfrac{\mu_o J_x}{2}\hat{\mathbf{a}}_y & z < 0 \end{cases} \quad \text{(Infinite sheet of steady current).} \tag{7.30}$$

Here, we see that the B-field generated by a uniform, infinite sheet of current is independent of height above the sheet and has opposite signs above and below the sheet. Also, as in the case of the infinite line, the direction of this magnetic field is such that it tends to circulate around the current source according to the right-hand rule. The streamlines generated by an infinite current sheet are shown in Figure 7-10. In this figure, the surface current $\mathbf{J}_s$ is directed out of the paper.

7-4 Field Calculations Using Ampère's Law

Just as Gauss' law is useful when calculating the E-fields of certain classes of symmetric charge distributions, Ampère's law is often useful for calculating the B-fields generated by symmetric current distributions. In its integral form, Ampère's law states that the line integral of **B** around any closed path C equals the product of the current I_{enc} enclosed by the path (in a right-handed sense) and the permeability of free space μ_o; that is,

$$\oint_C \mathbf{B}\cdot d\boldsymbol{\ell} = \mu_o I_{\text{enc}},$$

where C is called an ***ampèrian path.*** Before we use Ampère's law to derive B-field expressions for new current distributions, let us first use it to check the B-field of a previously modeled source—the infinite line of current.

Example 7-2

Show that the integral form of Ampère's law is valid for the counterclockwise, circular path C around the infinite line source, shown in Figure 7-11.

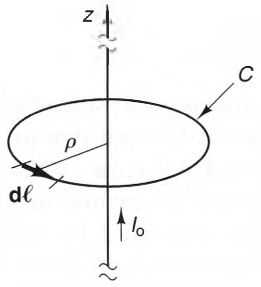


Figure 7-11 An infinite line of current and an ampèrian path.

Solution:

Since the current I_o passes through C in a right-handed sense, Ampère's law for this problem becomes

$$\oint_C \mathbf{B} \cdot \mathbf{d\ell} = \mu_o I_o.$$

To evaluate the line integral, we first note that the differential displacement vector **dℓ** along the circular path is

$$\mathbf{d\ell} = \rho d\phi \hat{\mathbf{a}}_\phi,$$

where ρ is the radius of the circle. Also, the field generated by an infinite line source is

$$\mathbf{B} = \frac{\mu_o I_o}{2\pi\rho} \hat{\mathbf{a}}_\phi.$$

Substituting these into Ampère's law, we obtain

$$\int_0^{2\pi} \frac{\mu_o I_o}{2\pi\rho} \hat{\mathbf{a}}_\phi \cdot \rho d\phi \hat{\mathbf{a}}_\phi = \frac{\mu_o I_o}{2\pi} \int_0^{2\pi} d\phi = \mu_o I_o.$$

Thus, as we expected, the field of an infinite line current satisfies Ampère's law for any circular contour whose axis lies along the line current.

Ampère's law is valid for all possible configurations of steady currents and integration contours C, but there are certain combinations of sources and contours that allow us to use it to calculate the B-fields of these sources. This is possible whenever a contour can be found along which **B** is constant and parallel to the path of integration. When such a contour is found, **B** can be taken out of the contour integral and solved for easily. In the sections that follow, we will use Ampère's law to determine the B-fields of several classes of current distributions with much less work than would be required if the Biot-Savart law were used. These classes are cylinders, solenoids, and toroids, all of which have important practical applications.

7-4-1 CYLINDRICALLY SYMMETRIC CURRENT DISTRIBUTIONS

One class of problems that is easily solved using Ampère's law consists of infinite cylinders of current that have rotational symmetry. Current distributions in this class can be expressed in the form $\mathbf{J} = J_z(\rho)\hat{\mathbf{a}}_z$, where the z axis is chosen as the axis of symmetry. To analyze this class of problems, our strategy will be to identify the general characteristics of the fields generated by these current distributions and then to use Ampère's law to find the exact expressions for **B**.

Figure 7-12a depicts the cross section of an arbitrary, cylindrically symmetric current distribution, where the direction of the current flow is perpendicular to the paper. We can determine several aspects of the B-field generated by this current distribution by treating it as a collection of infinite line currents. Two such filaments are shown in the figure that are equidistant from the z-axis and, hence, carry the same currents, $I_1 = I_2$. Since both filaments are located off center with respect to the z-axis, they each generate fields that have both ρ and ϕ components. However, these contributions add to produce only a ϕ component along a radial line that lies midway between these filaments. Since our choice of the x-axis was arbitrary and the currents are independent of z, **B** will have only a ϕ-component, which can be a function only of the radial coordinate ρ. Hence, we can express **B** in the form

$$\mathbf{B} = B_\phi(\rho)\hat{\mathbf{a}}_\phi.$$

As might be expected, a family of ampèrian paths that exploits this B-field symmetry consists of circles that are centered about the z-axis and lie in the xy-plane. One such circle is shown in Figure 7-12b. The differential displacement vector $\mathbf{d\ell}$ along this circle is

$$\mathbf{d\ell} = \rho d\phi \hat{\mathbf{a}}_\phi.$$

Substituting this into Ampère's law and integrating in a right-handed sense about the z-axis, we obtain

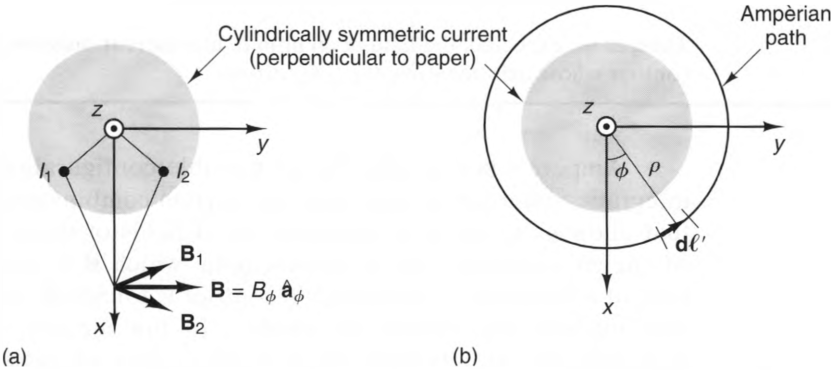


Figure 7-12 Cylindrically symmetric current distributions. a) The fields of all the complementary filaments produce only a ϕ-component of **B**. b) An ampèrian path.

$$\oint_C \mathbf{B} \cdot \mathbf{d}\boldsymbol{\ell} = \int_0^{2\pi} B_\phi(\rho)\hat{\mathbf{a}}_\phi \cdot \rho d\phi \hat{\mathbf{a}}_\phi = 2\pi\rho B_\phi(\rho) = \mu_o I_{\text{enc}}. \tag{7.31}$$

Here, I_{enc} is the total current that passes through the circle in a right-handed sense; thus

$$I_{\text{enc}} = \int_S \mathbf{J} \cdot \mathbf{ds} = \int_0^{2\pi}\int_0^{\rho} J_z(\rho')\rho' d\rho' d\phi' = 2\pi \int_0^{\rho} J_z(\rho')\rho' d\rho'.$$

Substituting this into Equation (7.31) and solving for B_ϕ, we obtain

$$\mathbf{B} = \frac{\mu_o I_{\text{enc}}}{2\pi\rho}\hat{\mathbf{a}}_\phi = \frac{\mu_o \hat{\mathbf{a}}_\phi}{\rho}\int_0^{\rho} J_z(\rho')\rho' d\rho' \quad \text{(Cylindrically symmetric current distribution).} \tag{7.32}$$

This expression can be applied to any cylindrically symmetric current distribution. We will now discuss three specific examples.

Solid Cylinders of Current. Figure 7-13a shows an infinite, solid cylinder of radius a that carries a uniform current $\mathbf{J} = J_o\hat{\mathbf{a}}_z$ [A/m^2] for $\rho < a$. This current distribution closely approximates a dc current in a solid wire. Also shown is a circular, ampèrian path of radius ρ.

The current I_{enc} that passes through the ampèrian path can be found by integrating $\mathbf{J}$ over the surface S. For $\rho < a$, we obtain

$$I_{\text{enc}} = \int_0^{2\pi}\int_0^{\rho} J_o\hat{\mathbf{a}}_z \cdot \hat{\mathbf{a}}_z \rho' d\rho' d\phi' = J_o \int_0^{2\pi}\int_0^{\rho} \rho' d\rho' d\phi' = \frac{\rho^2 J_o}{2}\int_0^{2\pi} d\phi' = \pi\rho^2 J_o \quad (\rho \le a).$$

If ρ is increased beyond a, no additional current is enclosed by S. Thus, for $\rho > a$ we have

$$I_{\text{enc}} = \pi a^2 J_o \equiv I_o \quad (\rho > a),$$

where I_o is the total current carried by the cylinder. Substituting these values for I into Equation (7.32), we obtain

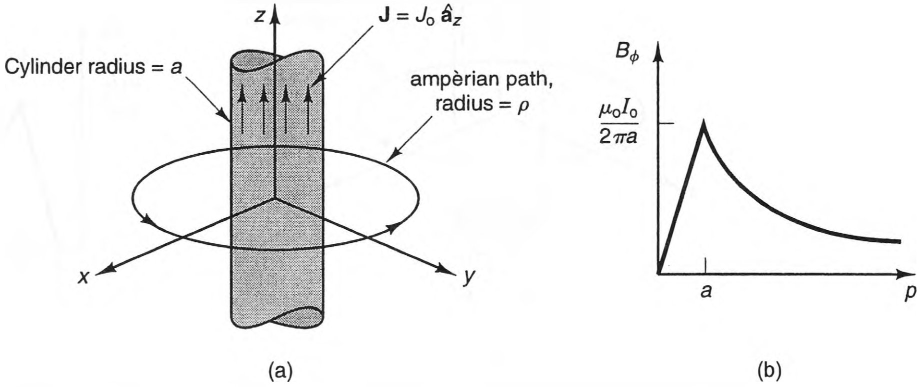

z
J = J_o $\hat{\mathbf{a}}_z$
Cylinder radius = a
ampèrian path, radius = ρ
x
y
(a)
B_ϕ
$\frac{\mu_o I_o}{2\pi a}$
a
ρ
(b)

Figure 7-13 A solid cylinder of uniform current. a) Geometry, b) Plot of B_ϕ vs. ρ.

$$\mathbf{B} = \begin{cases} \dfrac{\mu_o \rho I_o}{2\pi a^2}\,\hat{\mathbf{a}}_\phi & \rho < a \\ \dfrac{\mu_o I_o}{2\pi\rho}\,\hat{\mathbf{a}}_\phi & \rho > a \end{cases} \tag{7.33}$$

As can be seen from Figure 7-13b, B_ϕ increases linearly inside the cylinder from a value of zero on the z-axis to a maximum value of $(\mu_0 I_0)/(2\pi a)$ at the cylinder's edge. Outside the cylinder, however, B_ϕ decays proportional to ρ^{-1}, just like the field of an infinite line source. In fact, the field outside the cylinder is identical to that of an infinite line source carrying the same current I_o.

Hollow Cylinders of Current. Consider now the current distribution shown in Figure 7-14a. Here, a uniform current $\mathbf{J} = J_o\hat{\mathbf{a}}_z$ [A/m^2] within the cross section of a hollow cylinder with inner radius b and outer radius c. When $\rho < b$, the current $I(\rho)$ that passes through the ampèrian path is zero. For $b < \rho < c$, we have

$$I_{\text{enc}} = \int_0^{2\pi}\int_0^{\rho} \mathbf{J}\cdot d\mathbf{s} = \int_0^{2\pi}\int_b^{\rho} J_o\rho'\,d\rho'\,d\phi' = 2\pi J_o \int_b^{\rho} \rho' d\rho' = \pi J_o(\rho^2 - b^2) \quad b < \rho < c.$$

When $\rho \geq c$, I equals the total current I_o carried by the cylinder: $I_o = \pi J_o(c^2 - b^2)$. Substituting these values into Equation (7.32), we obtain

$$\mathbf{B} = \begin{cases} 0 & \rho < b \\ \dfrac{\mu_o I_o(\rho^2 - b^2)}{2\pi\rho(c^2 - b^2)}\,\hat{\mathbf{a}}_\phi & b < \rho < c. \\ \dfrac{\mu_o I_o}{2\pi\rho}\,\hat{\mathbf{a}}_\phi & \rho > c \end{cases} \tag{7.34}$$

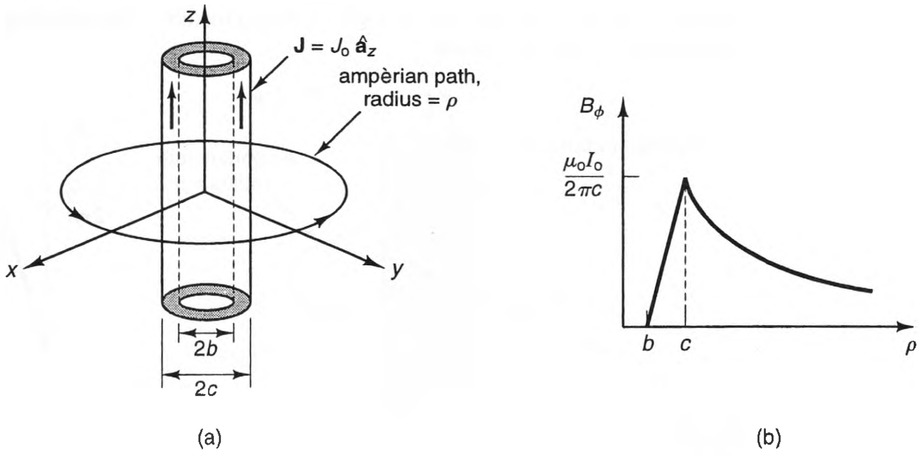


Figure 7-14 Hollow cylinder of current. a) Geometry. b) Plot of B_ϕ vs. ρ.

As can be seen from Figure 7-14b, B_ϕ increases with increasing values of ρ inside the cylinder. Outside the cylinder, the field is identical to that of an infinite line source along the cylinder axis that carries the same total current I_o.

Coaxial lines. We can combine the results of the previous two sections to find the B-field generated by the current distribution shown in Figure 7-15a. Here, a solid cylinder of radius a is surrounded by a hollow cylinder of inner radius b and outer radius c. The solid cylinder carries a total current I_o in the $+\hat{\mathbf{a}}_z$ direction, and the hollow cylinder carries the same current in the opposite direction. This is the type of current distribution that flows on a coaxial cable when it is operated in a balanced mode.

The fields generated by both the inner and outer cylinders have already been determined (see Equations (7.33) and (7.34)), so we can use the superposition principle to find the total B-field. Since the current flowing in the outer conductor is $-I_o$, we have

$$\mathbf{B} = \begin{cases} \dfrac{\mu_o \rho}{2\pi a^2} I_o \hat{\mathbf{a}}_\phi & \rho < a \\ \dfrac{\mu_o I_o}{2\pi\rho} \hat{\mathbf{a}}_\phi & a < \rho < b \\ \dfrac{\mu_o I_o}{2\pi\rho} \dfrac{(c^2 - \rho^2)}{(c^2 - b^2)} \hat{\mathbf{a}}_\phi & b < \rho < c \\ 0 & \rho > c \end{cases} \tag{7.35}$$

A plot of B_ϕ vs. ρ is shown in Figure 7-15b. As can be seen, **B** is zero outside a coaxial line as long as the currents on the inner and outer cylinders are equal. This is one of the attractive features of coaxial lines (cables) and is an example of ***magnetic shielding***, in which the fields generated by two currents cancel.

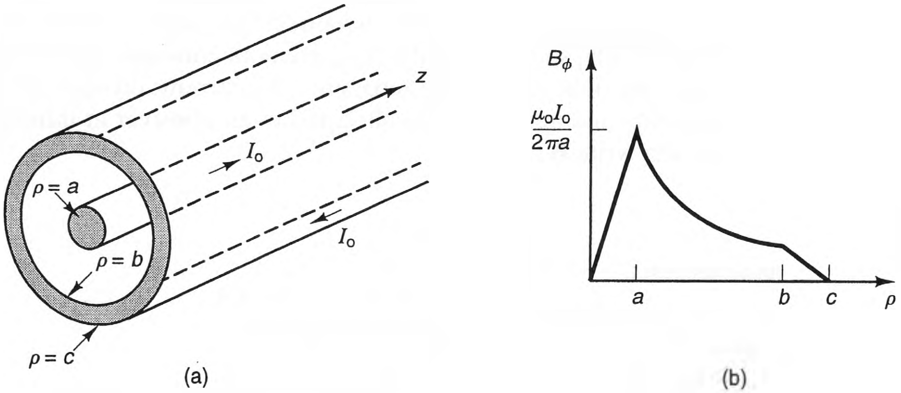


Figure 7-15 A coaxial line of current. a) Geometry. (b) Plot of B_ϕ vs. ρ.

7-4-2 SOLENOIDS

A solenoid is a cylinder with current flowing axially around it. Solenoids are used for a large range of devices, including relays, speakers, microphones, and electromagnets. Practical solenoids are made by wrapping many turns of wire around a cylindrical form and are capable of producing large magnetic fields from relatively small volumes of wire.

To model the fields produced by practical solenoids, let us start by considering the infinite solenoid shown in Figure 7-16. Here, a uniform surface current $J_s\hat{\mathbf{a}}_\phi$ [A/m] flows around a solenoid of radius a. In order to use Ampère's law to determine the B-field generated by this solenoid, it is first necessary to find an appropriate ampèrian path. To do this, however, we must know in advance what components **B** has and what coordinates they are functions of.

First, we note that **B** is independent of the ϕ and z-coordinates, since the current is rotationally symmetric and infinite in extent. Second, this current distribution is simply an infinite series of current loops, each carrying the same current. This means that the field along the axis of the solenoid has only a z-component. (This is seen from Equation (7.27).) The behavior of **B** at points off the z-axis is not as obvious, but it can be shown that at points both inside and outside the solenoid, **B** is of the form

$$\mathbf{B} = B_z(\rho)\hat{\mathbf{a}}_z.$$

This can be derived directly from the Biot-Savart law by noticing that all the field contributions generated by complementary pairs of loops on both sides of the observer cancel, except for the z-component.

An ampèrian path C that exploits the symmetry of this field is shown in Figure 7-16. Evaluating Ampère's law along this path, we find that

$$\oint_C \mathbf{B}\cdot\mathbf{d\ell} = [B_z(\rho_1) - B_z(\rho_2)]\,\Delta\ell = \mu_o I_{\text{enc}},$$

where I is the current enclosed by the path. When $\rho_1 < a$ and $\rho_2 > a$, the current is $I_{\text{enc}} = J_s\Delta\ell$, so

$$B_z(\rho_1) - B_z(\rho_2) = \mu_o J_s \ \ (\rho_1 < a \text{ and } \rho_2 > a).$$

This expression is valid for all values of ρ_1 and ρ_2 inside and outside the solenoid, respectively, so $B_z(\rho_1)$ and $B_z(\rho_2)$ are both constants and their difference equals $\mu_o J_s$. Also, the field at $\rho_2 \to \infty$ is zero, since the currents on opposite sides of the solenoid are opposite and appear to be collocated to an observer at infinity. Thus, $B_z(\rho_2) = 0$, so we can write **B** as

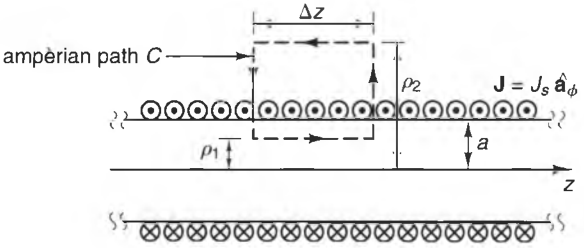


Figure 7-16 An infinite solenoid of current.

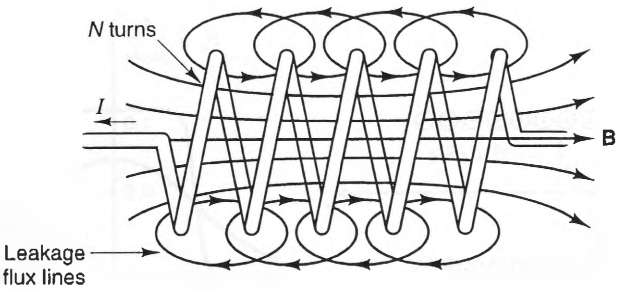


Figure 7-17 B-field streamlines of a typical solenoid.

$$\mathbf{B} = \begin{cases} \mu_0 J_s \hat{\mathbf{a}}_z & \text{(Inside solenoid)} \\ 0 & \text{(Outside solenoid)} \end{cases}. \tag{7.36}$$

Practical solenoids have finite lengths and current distributions that are not quite uniform, since the current flows through wires, rather than as a continuous sheet. Figure 7-17 shows a wire-wound solenoid. Here, the surface current density is approximately NI/L [A/m], where N is the total number of turns, I is the current flowing in each turn, and L is the solenoid length. Using this, we can approximate the B-field in a practical solenoid as

$$\mathbf{B} \approx \begin{cases} \dfrac{\mu_0 NI}{L}\hat{\mathbf{a}}_z & \text{(Inside solenoid)} \\ 0 & \text{(Outside solenoid)} \end{cases}. \tag{7.37}$$

There are two aspects of the B-field streamlines of practical solenoids that are not predicted by Equation (7.37). The first is the "leakage" flux lines between the turns. This is the result of the slightly nonuniform surface current density, which keeps the fields outside the solenoid from exactly canceling. Also, the streamlines spread out near the ends of the solenoid, weakening **B** at the ends. This occurs because the solenoid does not look infinite there.

To see how **B** varies in intensity along the axis of a finite-length solenoid, we can treat this surface current distribution as an infinite collection of circular loops. Figure 7-18a shows a solenoid of length L and diameter d, with a surface current $\mathbf{J}_s = J_s \hat{\mathbf{a}}_\phi$. The current on a loop at z' is $J_s dz'$. Using Equation (7.27) and $d = 2a$, we can write the differential field it generates at a point z as

$$dB_z = \frac{\mu_0 d^2 J_s dz'}{8[(z - z')^2 + d^2/4]^{3/2}},$$

where d is the diameter of the solenoid (and the loop). Integrating from $-L/2 < z' < L/2$, we obtain

$$\begin{aligned} B_z &= \frac{\mu_0 d^2 J_s}{8} \int_{-L/2}^{L/2} \frac{dz'}{[(z - z')^2 + d^2/4]^{3/2}} \\ &= \frac{\mu_0 J_s}{2}\left[\frac{L - 2z}{\sqrt{(L - 2z)^2 + d^2}} + \frac{L + 2z}{\sqrt{(L + 2z)^2 + d^2}}\right]. \end{aligned} \tag{7.38}$$

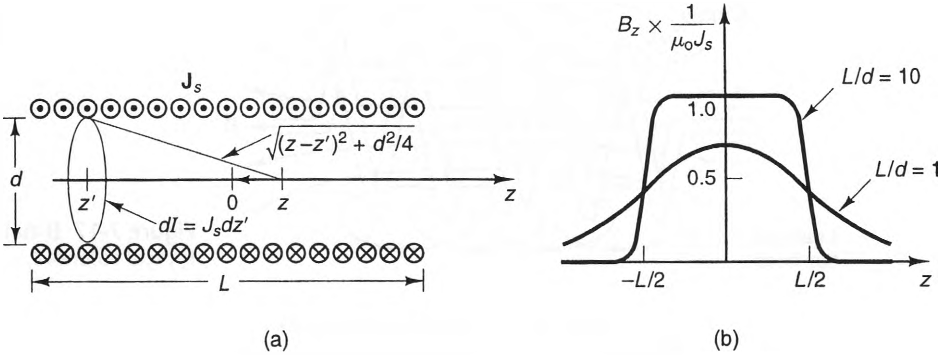


Figure 7-18 A finite-length solenoid. a) Geometry. b) Plot of B_z vs. z for two length-to-diameter ratios.

Figure 7-18b shows a plot of B_z vs. z for two solenoids, one with $L/d = 10$ and the other for $L/d = 1$. As can be seen, the field inside the solenoid is relatively constant away from the ends when $L \gg d$. For this case, the B-field is approximately the same as that generated by an infinite solenoid with the same surface current. Notice also that when $L/d \gg 1$, the field strength at the solenoid ends is approximately half its value deep within the solenoid.

7-4-3 TOROIDS

A ***toroid*** is formed when a solenoid is bent so that it forms a complete circle. Figure 7-19 shows a toroid with cross-sectional radius a and radius of revolution ρ_o. Since the inner and outer radii of the toroid are different, the surface current density is slightly greater on the inner wall than on the outer wall. We will assume that the current density at $\rho = \rho_o - a$ and $z = 0$ is $J_s\hat{\mathbf{a}}_z$ [A/m].

Once again, in order to use Ampère's law, we must first determine what components of **B** exist and then find an appropriate ampèrian path (if one exists). We can determine the general form of the B-field by considering the toroidal current to be an infinite collection of loops. Using this line of reasoning, we can show that the B-field in the xy-plane is of the form

$$\mathbf{B} = B_\phi(\rho)\hat{\mathbf{a}}_\phi.$$

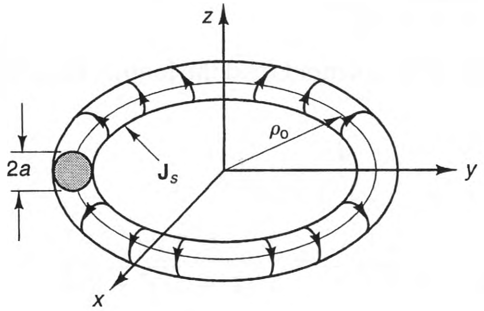


Figure 7-19 A toroid of current.

Applying Ampère's law along a circle of radius ρ, we obtain

$$\oint_C \mathbf{B} \cdot d\boldsymbol{\ell} = 2\pi\rho B_\phi(\rho) = \mu_o I_{\text{enc}},$$

where I is the current that passes through the path. When the perimeter of the path lies inside the toroid, current passes through the surface bounded by the ampèrian path once, yielding $I_{\text{enc}} = 2\pi(\rho_o - a)J_s$. Conversely, when the circle lies outside the toroid, the total current I passing through the surface is zero, either because the current never crosses the bounded surface or because it passes through it twice in opposite directions. Hence,

$$\mathbf{B} = \begin{cases} \mu_o J_s \dfrac{(\rho_o - a)}{\rho} \hat{\mathbf{a}}_\phi & \text{(Inside toroid)} \\ 0 & \text{(Outside toroid)} \end{cases} . \tag{7.39}$$

When $\rho_o \gg a$, $\mathbf{B} \approx \mu_o J_s \hat{\mathbf{a}}_\phi$ throughout the interior of the toroid.

Practical toroids are constructed by wrapping many turns of wire around a toroidal form. If the total number of turns of wire is N and $\rho_o \gg a$, then $J_s \approx (NI)/[2\pi(\rho_o - a)]$, for which we obtain

$$\mathbf{B} \approx \begin{cases} \dfrac{\mu_o NI}{2\pi\rho_o} \hat{\mathbf{a}}_\phi & \text{(Inside toroid)} \\ 0 & \text{(Outside toroid)} \end{cases} . \tag{7.40}$$

Just as in the case of practical solenoids, there is some leakage flux between the windings, which means that there is a small B-field outside a practical toroid.

7-5 Magnetic Potentials

Just as the electrostatic potential function simplifies the analysis of many electrostatic problems, potentials are also useful in magnetostatic analysis. But whereas only one kind of potential function is used in electrostatics, two kinds of potentials are useful in magnetostatics. The most general is the magnetic vector potential, which we introduced earlier. We can also use a *scalar* magnetic potential in regions where no current is flowing. After discussing the properties of these potentials, we will show how they can be used to model small current loops.

7-5-1 MAGNETIC VECTOR POTENTIAL

We have already seen that because $\nabla \cdot \mathbf{B} = 0$ at all points, we can always write $\mathbf{B}$ in terms of the magnetic vector potential $\mathbf{A}$, namely,

$$\mathbf{B} = \nabla \times \mathbf{A}, \tag{7.41}$$

where

$$\mathbf{A} = \frac{\mu_o}{4\pi}\int_{\text{Vol.}} \frac{\mathbf{J}}{|\mathbf{r} - \mathbf{r}'|}\,dv' \qquad [\text{Wb/m}]. \tag{7.42}$$

Here, **r** represents the observation point, and the integration is over all points **r**′ where the current density **J** is nonzero. When the source is a surface or line current, we can replace $\mathbf{J}dv'$ with $\mathbf{J}_s ds'$ and $Id\ell'$, respectively, yielding

$$\mathbf{A} = \frac{\mu_o}{4\pi}\int_S \frac{\mathbf{J}_s}{|\mathbf{r} - \mathbf{r}'|}\,ds' \qquad [\text{Wb/m}], \tag{7.43}$$

and

$$\mathbf{A} = \frac{\mu_o I}{4\pi}\int_C \frac{d\boldsymbol{\ell}'}{|\mathbf{r} - \mathbf{r}'|} \qquad [\text{Wb/m}]. \tag{7.44}$$

Another important relationship between **B** and **A** can be derived from Equation (7.41) by integrating both sides over an arbitrary surface S:

$$\int_S \mathbf{B}\cdot d\mathbf{s} = \int_S \nabla\times\mathbf{A}\cdot d\mathbf{s}.$$

We can use Stokes's theorem to write the right-hand integral as a contour integral,

$$\int_S \mathbf{B}\cdot d\mathbf{s} = \int_C \mathbf{A}\cdot d\boldsymbol{\ell}, \tag{7.45}$$

where C is the closed path that bounds S. Thus, the line integral of **A** around a closed path C equals the magnetic flux passing through the surface bounded by C.

Example 7-3

Find the magnetic vector potential in the region shown in Figure 7-20 if measurements show that $\mathbf{B} = B_o\hat{\mathbf{a}}_z$ throughout the region.

Solution:

Evaluating Equation (7.45) when S is a circular disc of radius ρ in the $z = 0$ plane and C is the bounding circle, we obtain

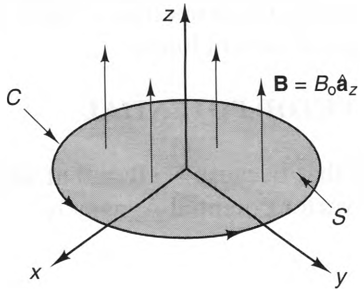


Figure 7-20 A circular region with uniform **B**.

$$\int_S \mathbf{B} \cdot \mathbf{ds} = B_0 \pi \rho^2 = \oint_C \mathbf{A} \cdot \mathbf{d\ell} = \int_0^{2\pi} \rho A_\phi d\phi.$$

Since this system "looks" the same for all values of ϕ, let us for the moment assume that A_ϕ is independent of ϕ and later check to see if this assumption is valid. Using this assumption, we obtain

$$\int_0^{2\pi} \rho A_\phi d\phi = 2\pi\rho A_\phi = \pi\rho^2 B_o.$$

Thus, **A** is given by

$$\mathbf{A} = \frac{\rho B_o}{2} \hat{\mathbf{a}}_\phi.$$

To show that this expression for **A** is indeed correct, we can take the curl of **A**, which yields

$$\mathbf{B} = \nabla \times \mathbf{A} = \frac{1}{\rho}\frac{\partial}{\partial \rho}(\rho A_\phi)\hat{\mathbf{a}}_z = B_o \hat{\mathbf{a}}_z.$$

7-5-2 MAGNETIC SCALAR POTENTIAL

In regions where the current density **J** is zero, Ampère's law becomes

$$\nabla \times \mathbf{B} = 0 \qquad \text{(regions where } \mathbf{J} = 0\text{)}, \tag{7.46}$$

which means that **B** is an irrotational vector. According to Theorem III in Section 2-5-6 (Equation (2.131)), a vector that is irrotational in a region can always be represented as the gradient of a scalar function. Thus, we can express **B** as

$$\mathbf{B} = -\mu_o \nabla V_m \qquad \text{(regions where } \mathbf{J} = 0\text{)}, \tag{7.47}$$

where V_m is called the ***magnetic scalar potential***, measured in ampères [A]. The magnetic scalar potential is applicable only in regions where $\mathbf{J} = 0$, but this is not as severe a restriction as one might think, since we are usually more interested in the fields at positions that are removed from their sources.

An important characteristic of the magnetic scalar potential is that it satisfies Laplace's equation, just as the electrostatic potential V does in regions that are charge free. To show this, let us take the divergence of both sides of Equation (7.47), obtaining

$$\nabla \cdot \mathbf{B} = -\mu_o \nabla \cdot \nabla V_m = -\mu_o \nabla^2 V_m. \tag{7.48}$$

But since $\nabla \cdot \mathbf{B} = 0$ at all points, Equation (7.48) becomes

$$\nabla^2 V_m = 0. \tag{7.49}$$

An integral relationship between **B** and V_m can be obtained by multiplying both sides of Equation (7.47) by $\mathbf{d\ell}$ and integrating over a contour with endpoints A and B:

$$\int_b^a \mathbf{B} \cdot d\boldsymbol{\ell} = -\mu_o \int_b^a \nabla V_m \cdot d\boldsymbol{\ell}.$$

Using the properties of the gradient operation (see Equation (2.76)), we have $\nabla V_m \cdot d\boldsymbol{\ell} = dV_m$, so the preceding expression can be written as

$$\int_b^a \mathbf{B} \cdot d\boldsymbol{\ell} = -\mu_o \int_b^a dV_m = -\mu_o(V_{ma} - V_{mb}), \tag{7.50}$$

where V_{ma} and V_{mb} are the scalar magnetic potentials at the points a and b, respectively. This relationship between a magnetic field and the magnetic potential difference it generates between two points is used to model the operation of magnetic circuits, which are discussed in Chapter 8.

7-5-3 THE MAGNETIC DIPOLE

Small loops of current are often called ***magnetic dipoles*** because, as we will soon see, they generate B-fields that have the same form as the E-fields generated by electrostatic dipoles. We have already used the Biot-Savart law to calculate the B-field along the axes of current loops. In this section we will use the magnetic potentials to obtain a more complete model of the fields generated by this important type of source.

Figure 7-21 shows a magnetic dipole, which has radius a and carries a current I in the counterclockwise direction. Since this is a filamentary current, the magnetic vector potential $\mathbf{A}$ can be written as

$$\mathbf{A} = \frac{\mu_o I}{4\pi} \oint_C \frac{d\boldsymbol{\ell}'}{|\mathbf{r} - \mathbf{r}'|}. \tag{7.51}$$

For this source, the terms in the integrand are,

$$d\boldsymbol{\ell}' = a d\phi' \hat{\mathbf{a}}_{\phi'} = (a \cos \phi' \hat{\mathbf{a}}_y - a \sin \phi' \hat{\mathbf{a}}_x) d\phi' \tag{7.52}$$

$$\mathbf{r} = r\hat{\mathbf{a}}_r = x\hat{\mathbf{a}}_x + y\hat{\mathbf{a}}_y + z\hat{\mathbf{a}}_z$$

$$\mathbf{r}' = a\mathbf{a}_{\rho'} = a \cos \phi' \hat{\mathbf{a}}_x + a \sin \phi' \hat{\mathbf{a}}_y$$

$$\mathbf{r} - \mathbf{r}' = (x - a \cos\phi')\hat{\mathbf{a}}_x + (y - a \sin \phi')\hat{\mathbf{a}}_y + z\hat{\mathbf{a}}_z$$

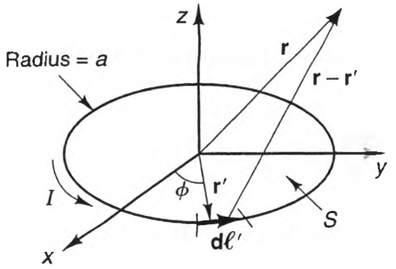


Figure 7-21 Geometry for determining the B-field of a magnetic dipole.

Also,

$$\frac{1}{|\mathbf{r}-\mathbf{r}'|} = [(x - a\cos\phi')^2 + (y - a\sin\phi')^2 + z^2]^{-1/2}$$
$$= [r^2 + a^2 - 2a(x\cos\phi' + y\sin\phi')]^{-1/2}. \qquad (7.53)$$

Rather than substituting Equation (7.53) directly into Equation (7.51), we would do well to first simplify Equation (7.53). We can do this by expanding it in a Maclaurin series about $a = 0$:

$$\frac{1}{|\mathbf{r}-\mathbf{r}'|} = f(a) = f(0) + af'(0) + \frac{1}{2!}a^2 f''(0) + \ldots$$

The first two terms are

$$f(0) = \frac{1}{r}$$
$$f'(0) = \frac{(x\cos\phi' + y\sin\phi')}{r^3}.$$

In the "far zone" (i.e., at large distances), where $r \gg a$, only the first two terms are needed in the expansion. Thus,

$$\frac{1}{|\mathbf{r}-\mathbf{r}'|} \approx \frac{1}{r} + \frac{a(x\cos\phi' + y\sin\phi')}{r^3}. \qquad (7.54)$$

Substituting Equations (7.52) and (7.54) into the integral in Equation (7.51), we find that

$$\oint_C \frac{d\boldsymbol{\ell}'}{|\mathbf{r}-\mathbf{r}'|} = \int_0^{2\pi}\left[\frac{1}{r} + \frac{a(x\cos\phi' + y\sin\phi')}{r^3}\right](a\cos\phi'\hat{\mathbf{a}}_y - a\sin\phi'\hat{\mathbf{a}}_x)\,d\phi'$$
$$= \frac{\pi a^2}{r^3}(x\hat{\mathbf{a}}_y - y\hat{\mathbf{a}}_x).$$

Also, since $S = \pi a^2$ is the area of the loop and $(x\hat{\mathbf{a}}_y - y\hat{\mathbf{a}}_x) = r\sin\theta\hat{\mathbf{a}}_\phi$, this expression can be further simplified to read

$$\oint_C \frac{d\boldsymbol{\ell}'}{|\mathbf{r}-\mathbf{r}'|} = \frac{S}{r^2}\sin\theta\,\hat{\mathbf{a}}_\phi \quad r \gg a \qquad (7.55)$$

Substituting Equation (7.55) into Equation (7.51), we obtain

$$\mathbf{A} = \frac{\mu_0 SI}{4\pi r^2}\sin\theta\,\hat{\mathbf{a}}_\phi \quad r \gg a. \qquad (7.56)$$

Now that we have found **A**, calculating **B** is straightforward. Using $\mathbf{B} = \nabla \times \mathbf{A}$, we find that

$$\mathbf{B} = \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(A_\phi\sin\theta)\,\hat{\mathbf{a}}_r - \frac{1}{r}\frac{\partial}{\partial r}(rA_\phi)\,\hat{\mathbf{a}}_\theta.$$

Evaluating the derivatives, we obtain the following far-zone expression:

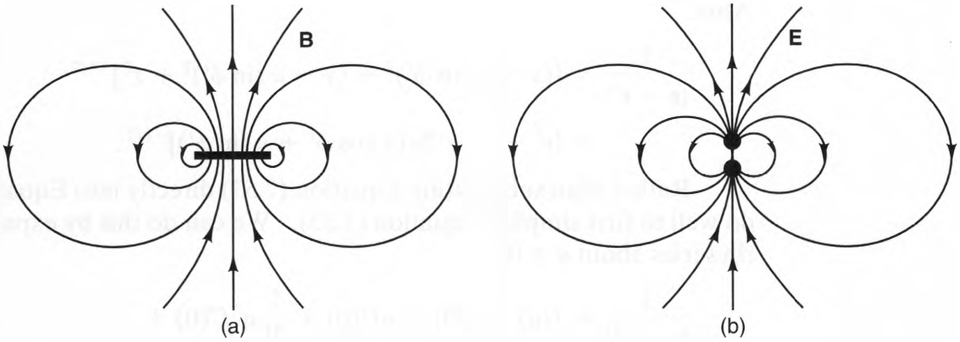


Figure 7-22 a) A magnetic dipole. b) An electric dipole.

$$\mathbf{B} = \frac{\mu_o SI}{4\pi r^3}(2\cos\theta\,\hat{\mathbf{a}}_r + \sin\theta\,\hat{\mathbf{a}}_\theta) \quad r \gg a \qquad [\mathrm{Wb/m^2}]. \tag{7.57}$$

Comparing Equation (7.57) with the far-zone E-field generated by an electrostatic dipole (Equation (4.54)), we see that, when viewed at large distances, they are of *exactly* the same form. This can also be seen from Figure 7-22, which compares the B- and E-fields generated by magnetic and electric dipoles. Here we see that these fields differ only at points close to their sources, where the B-field streamlines pass through the loop and the E-field streamlines terminate on the charges. Because of this similarity, small current loops are often called ***magnetic dipoles.***

We can extend the analogy between electric and magnetic dipoles by using the magnetic scalar potential V_m. Given that $\mathbf{B} = -\mu_o \nabla V_m$ has the same form as $\mathbf{E} = -\nabla V$ and that the E-field of an electric dipole and the **B**-field of a magnetic dipole differ only by a constant, their potentials V_m and V should differ by only the same constant. In Chapter 4 we showed that the potential function of an electric dipole at the origin and directed along the z-axis is

$$V = \frac{p\cos\theta}{4\pi\epsilon_o r^2} \quad r \gg d, \tag{7.58}$$

where $p = Qd$ is the electric dipole moment. Using Equations (4.53) and (4.54), we conclude that the magnetic scalar potential that is associated with the B-field of Equation (7.57) must be

$$V_m = \frac{m\cos\theta}{4\pi r^2} \quad r \gg d, \tag{7.59}$$

where m is the ***magnetic dipole moment***, measured in units of amperes times meters squared [A • m^2]. Taking the gradient of this expression and multiplying by $-\mu_o$, we obtain

$$\mathbf{B} = \frac{\mu_o m}{4\pi r^3}(2\cos\theta\,\hat{\mathbf{a}}_r + \sin\theta\,\hat{\mathbf{a}}_\theta) \quad r \gg a \qquad [\mathrm{Wb/m^2}]. \tag{7.60}$$

Comparing this expression with Equation (7.57), we see that they are the same if the magnetic dipole moment m equals the product of the loop current I and area S:

$$m = SI = \pi a^2 I \qquad [\mathrm{A \cdot m^2}]. \tag{7.61}$$

We can also use our experience with electric dipoles to generalize our expressions for magnetic dipoles to the case where they are located away from the origin and oriented along an axis other than the z-axis. To accomplish this, let us make the magnetic dipole moment into a vector quantity,

$$\mathbf{m} = SI\hat{\mathbf{a}}_n, \tag{7.62}$$

where S and I are respectively the area and current carried by the loop and $\hat{\mathbf{a}}_n$ points along the axis of the loop in a right-handed sense according to the positive direction of the current. Using a sequence of steps similar to the ones used in Section 4-6-2, we can write the scalar potential of a magnetic dipole located at a point $\mathbf{r}'$ as

$$V_m = \frac{\mathbf{m} \cdot \hat{\mathbf{a}}_R}{4\pi R^2} = \frac{\mathbf{m} \cdot (\mathbf{r} - \mathbf{r}')}{4\pi |\mathbf{r} - \mathbf{r}'|^3} \qquad [\mathrm{V}], \tag{7.63}$$

where R is the distance from the center of the loop to an observer at the point $\mathbf{r}$ and $\hat{\mathbf{a}}_R$ points along that line; that is,

$$\hat{\mathbf{a}}_R = \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|} \tag{7.64}$$

and

$$R = |\mathbf{r} - \mathbf{r}'|. \tag{7.65}$$

In a similar manner, we can generalize the magnetic vector potential expression for a magnetic dipole. From Equation (7.56), we see that $\mathbf{A}$ for a magnetic dipole is directed perpendicular to both the axis of the dipole and the line extending from the observer to the center of the loop. We can write the vector potential $\mathbf{A}$ for an arbitrarily located and oriented magnetic dipole as

$$\mathbf{A} = \frac{\mu_o \mathbf{m} \times \hat{\mathbf{a}}_R}{4\pi R^2} = \frac{\mu_o \mathbf{m} \times (\mathbf{r} - \mathbf{r}')}{4\pi |\mathbf{r} - \mathbf{r}'|^3}. \tag{7.66}$$

The analogy between the fields generated by electric and magnetic dipoles also suggests that we can consider the source of a magnetic dipole to be a pair of fictitious magnetic point charges. Figure 7-23a shows a small current loop that has a radius a and

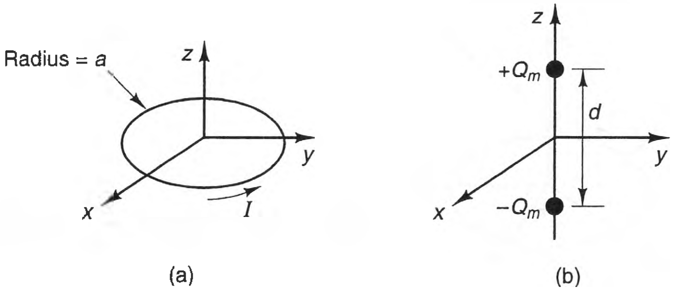


Figure 7-23 A magnetic dipole. a) Physical geometry. b) Equivalent geometry using fictitious magnetic charges.

carries a current I. Figure 7-23b shows an equivalent source (for $r \gg a$), consisting of magnetic point charges Q_m and $-Q_m$, separated by a distance d. In order for the B-field generated by this equivalent source to be the same as for the loop, they must have the same magnetic dipole moments. Thus, we must have

$$m = Q_m d = \pi a^2 I = IS. \tag{7.68}$$

We also see from this expression that if magnetic charges are ever found in nature, they would have units of amperes times meters [A • m].

In the next chapter, we will find that the spinning and orbiting electrons in atoms and molecules appear as tiny magnetic dipoles. Although the B-field generated by each magnetic dipole is very small, large fields can be generated when even a small fraction of these dipoles point in a common direction. The following example shows that a B-field applied to a magnetic dipole exerts a torque on the dipole that tends to align it with the applied field.

Example 7-4

Find the torque exerted by a uniform B-field on the magnetic dipole shown in Figure 7-24. The loop has radius a, lies in the $z = 0$ plane, and carries a current I in the counterclockwise direction.

Solution:

From mechanics, we know that the torque applied to each element of an object about some coordinate origin equals the cross product of the position vector $\mathbf{r} \times \mathbf{F}$, where $\mathbf{r}$ is the position vector of the element and force $\mathbf{F}$ is the applied force. Thus, the torque contributed by each differential element of the loop is

$$\mathbf{dT}_m = \mathbf{r} \times \mathbf{dF}$$

where $\mathbf{r}$ is given by

$$\mathbf{r} = a\hat{\mathbf{a}}_\rho = a(\cos\phi\,\hat{\mathbf{a}}_x + \sin\phi\,\hat{\mathbf{a}}_y),$$

and

$$\mathbf{dF} = I\mathbf{d\ell} \times \mathbf{B} = I\,a\,d\phi\hat{\mathbf{a}}_\phi \times \mathbf{B} = I\,a\,d\phi(\cos\phi\,\hat{\mathbf{a}}_y - \sin\phi\hat{\mathbf{a}}_x) \times \mathbf{B}$$

$$= I\,a\,d\phi[B_z\cos\phi\,\hat{\mathbf{a}}_x + B_z\sin\phi\,\hat{\mathbf{a}}_y - (B_x\cos\phi - B_y\sin\phi)\hat{\mathbf{a}}_z].$$

Taking the cross product of $\mathbf{r}$ and $\mathbf{dF}_m$, we find

$$\mathbf{dT}_m = \mathbf{r} \times \mathbf{dF} = Ia^2d\phi(B_x\cos\phi + B_y\sin\phi)(-\sin\phi\,\hat{\mathbf{a}}_x + \cos\phi\,\hat{\mathbf{a}}_y).$$

Integrating the differential torque contributions around the entire loop yields

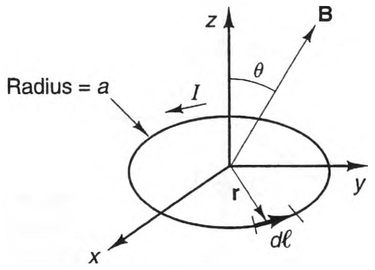


Figure 7-24 Geometry for calculating the torque on a magnetic dipole due to a uniform B-field.

$$\mathbf{T}_m = \int_0^{2\pi} d\mathbf{T}_m = \pi a^2 I(B_x \hat{\mathbf{a}}_y - B_y \hat{\mathbf{a}}_x),$$

which can be written as

$$\mathbf{T}_m = IS\hat{\mathbf{a}}_z \times \mathbf{B} = \mathbf{m} \times \mathbf{B}, \tag{7.69}$$

where S is the area of the loop, and $\mathbf{m} = \pi a^2 I\hat{\mathbf{a}}_z$ is the magnetic dipole moment of the loop. From this expression we see that maximum torque is induced when **m** and **B** are perpendicular and the direction of the torque is such that it tends to align **m** with **B**.

When a magnetic dipole is allowed to rotate under the influence of the Earth's magnetic field, it's magnetic moment **m** will point from geographic south to geographic north, since the earth's magnetic field points in this same direction along the Earth's surface. The same kind of torque is exerted on a permanent magnet when it is subjected to the earth's magnetic field, so that north and south poles of the magnet point towards the north and south directions, respectively.

Because of this similarity between magnetic dipoles and permanent magnets, the positive and negative magnetic charges of a magnetic dipole are called its ***north*** and ***south*** poles, respectively.

Figure 7-25 shows the Earth's magnetic field near its surface. Notice that the axis of the magnetic dipole lines is inclined 12° from the Earth's rotational axis. The northernmost and southernmost intersections of this axis with the Earth's surface are called the north and south magnetic poles, respectively. The offset between the geometric

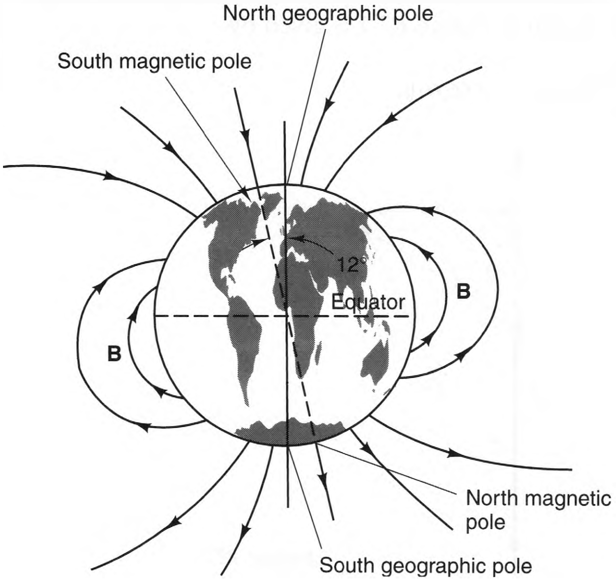


Figure 7-25 The earth's magnetic field.

and magnetic poles is called the ***magnetic declination*** and must be compensated for when making precise geographic measurements with a compass. The horizontal component of the magnetic field on the Earth's surface varies from 35 [μT] (0.35 gauss) at the equator to zero at the magnetic poles; the vertical component varies from zero at the equator to 70 [μT] (0.7 gauss) at the magnetic poles.

The Earth's magnetic field is produced in its core, which acts as a permanent magnet. We will see in the following chapter that the B-fields produced by permanent magnets (such as the Earth's core) are the result of circulating molecular currents. Interestingly, since the B-field streamlines are directed from south to north, the north pole of this permanent magnet is located in the southern hemisphere, whereas the magnet's south pole is located in the northern hemisphere.

7-6 Summation

In this chapter, we developed the fundamental equations that describe the behavior of magnetostatic fields in free space. We started this process by specializing Maxwell's equations for the case of steady currents and then proceeded to develop a number of different ways to model magnetostatic fields. Using these techniques, we were able to calculate the B-fields generated by a number of simple source configurations that are good approximations of sources encountered in engineering practice.

In the next chapter, we will expand this discussion by including the effects of material media on the fields generated by magnetic sources.

PROBLEMS

7-1 Show that the B-field generated at the point (ρ, ϕ, z) by the straight, finite-length current filament shown in Figure P7-1 is given by

$$\mathbf{B} = \frac{\mu_0 I \hat{\mathbf{a}}_\phi}{4\pi\rho}(\cos\alpha_2 - \cos\alpha_1),$$

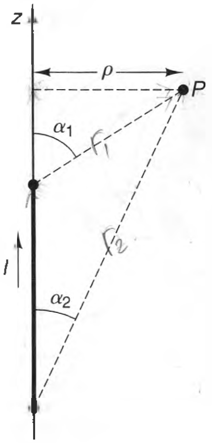


Figure P7-1

where α_1 and α_2 are the angles that the lines extending from the endpoints of the filament to the observer make with the z-axis. (*Hint:* Use the derivation leading up to Equation (7.25), except with finite integration limits.)

7-2 Use the result of Problem 7.1 to show that the B-field generated by a finite-length current filament is nearly the same as an infinite one when viewed close to the current and away from the endpoints.

7-3 Derive the expression for the B-field generated by a uniform, infinite sheet of current (Equation (7.30)) using Ampère's law. (*Hint:* Use the fact that the field has odd symmetry on both sides of the sheet).

7-4 Derive the expressions for the B-field inside and outside an infinite solenoid with radius a and current density $\mathbf{J}_s = J_\phi \hat{\mathbf{a}}_\phi + J_z \hat{\mathbf{a}}_z$ [A/m] on its surface. (*Hint:* Use the superposition principle to account for the two components of $\mathbf{J}_s$ separately.)

7-5 An infinite slab of thickness d contains a uniform current $\mathbf{J} = J_o \hat{\mathbf{a}}_x$ [A/m^2], as shown in Figure P7-5. Find $\mathbf{B}$ at all points in space. (*Hint:* Use Ampère's law.)

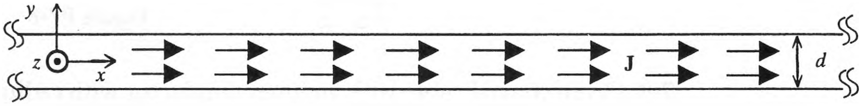


Figure P7-5

7-6 Find the B-field at the center of the square loop shown in Figure P7-6. The loop has width w and carries a filamentary current I. (*Hint:* Use the result of Problem 7-1).

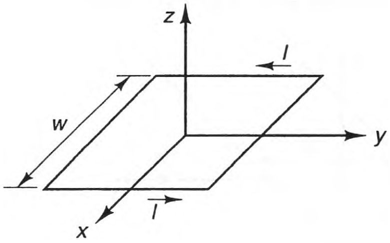


Figure P7-6

7-7 At high frequencies, the current density in wires decays exponentially beneath the surface. (This is discussed in Chapter 12.) Find the B-field inside and outside the long, straight wire in Figure P7-7, which has radius a and current density $\mathbf{J} = J_o e^{-\alpha(a-\rho)} \hat{\mathbf{a}}_z$ [A/m^2] for $\rho < a$. Assume that $\alpha \gg 1/a$.

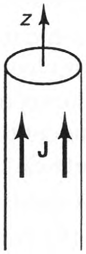


Figure P7-7

7-8 Figure P7-8 shows two coils spaced by a distance d, each with N turns, radius a, and current I flowing in the same direction. Coils in this configuration are called ***Helmholtz coils*** and produce a nearly uniform B-field in the region between them.
(a) Find an expression for $\mathbf{B}$ at all points along the z-axis.
(b) Show that $\partial\mathbf{B}/\partial z = 0$ at $z = 0$ for all values of a and d.
(c) For a given a, find the coil spacing d that yields $\partial^2\mathbf{B}/\partial z^2 = 0$ at $z = 0$.

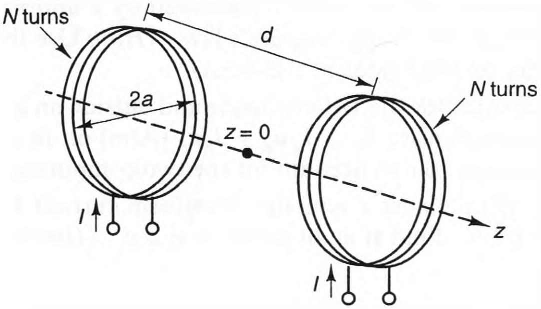


Figure P7-8

7-9 Given coaxial cable with an inner conductor with radius a and an outer conductor with radius b, find $\mathbf{B}$ for $\rho > b$ when the inner and outer currents are +100 [mA] and −90 [mA], respectively, and both conductors are centered along the z-axis. Assume that the positive direction for both currents is the $+z$-direction.

7-10 Figure P7-10 shows two infinite line currents that each lie symmetrically about the x-axis in the $z = 0$ plane and are spaced by a distance $d = 1$ [cm]. If $I_1 = 4$ [mA], find I_2 such that $\mathbf{B} = 0$ at $P(0, 1, 0)$ [cm].

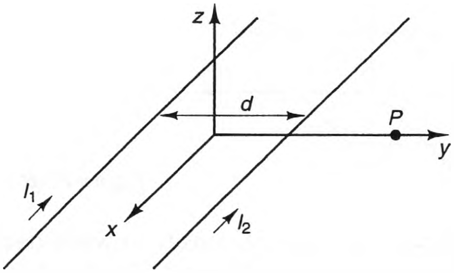


Figure P7-10

7-11 How many turns of wire are required to produce a B-field whose magnitude is greater than or equal to 10^{-3} [T] inside an air-wound solenoid that is 2 [cm] long and carries a current of 100 [mA].

7-12 A toroid with a mean radius of revolution of 4.25 [cm] and a cross-sectional radius of 0.25 [cm] has two separate windings. The first has 400 turns and carries 100 [mA], and the second contains 300 turns and carries 150 [mA]. Calculate the B-field along the center of the toroid if the currents are a) in the same direction and b) in opposite directions.

7-13 Prove that the B-field at the end of a long, narrow solenoid is approximately one-half its value in the center of the solenoid.

7-14 Figure P7-14 shows the cross section of a long, circular cylinder of radius b that carries a uniform current density $\mathbf{J} = J_o\hat{\mathbf{a}}_z$. If a cylindrical cavity of radius a and offset d is cut out of this uniform current, find the B-field inside the cavity and show that it is uniform. (*Hint:* Use the superposition principle, and consider this current distribution as the sum of a solid axial current $J_o\hat{\mathbf{a}}_z$ with radius b and an oppositely directed current $-J_o\hat{\mathbf{a}}_z$ in the cavity region.)

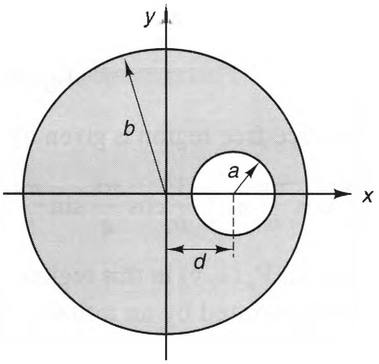


Figure P7-14

7-15 Figure P7-15 shows a short current element of length ℓ that carries a current I.
(a) Find an approximate expression for the vector potential **A** for $r \gg \ell$.
(b) Use this **A** to find an expression for **B**.
(c) Compare the result in part b) with Equation (7.22).

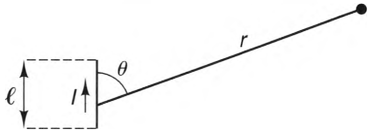


Figure P7-15

7-16 Prove that the magnetic vector potential inside an infinite, cylindrical conductor that carries a uniform current density $\mathbf{J} = J_o\hat{\mathbf{a}}_z$ is given by

$$\mathbf{A} = \frac{-\mu_o J_o \rho^2}{4}\hat{\mathbf{a}}_z \qquad (\rho < \text{radius}).$$

(*Hint:* Use the B-field inside the cylinder, given by Equation (7.33)).

7-17 Figure P7-17 shows two small current loops. Loop #1 is located at the origin and has a magnetic moment $\mathbf{m}_1$, directed along the z-axis. Loop #2 is located a distance d away from loop #1 and has a magnetic moment $\mathbf{m}_2$ that lies in the plane of the paper and is also perpendicular to the line that connects the centers of both loops. If the distance between the loops is large enough so that the B-field generated by one loop is essentially constant at all points on the other, show that the torque **T** exerted on loop #2 is

$$\mathbf{T} = \frac{\mu_o m_1 m_2}{2\pi d^3}\cos\theta\hat{\mathbf{a}}_\phi = \frac{\mu_0}{2\pi d^3}(\mathbf{m}_2 \mathbf{X}\, \mathbf{m}_1).$$

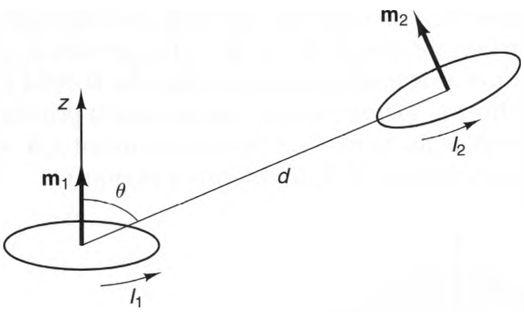


Figure P7-17

7-18 If the B-field in a source-free region is given by

$$\mathbf{B} = \frac{1}{a}\sin\frac{\pi x}{a}\cos\frac{\pi y}{b}\,\hat{\mathbf{a}}_x + \frac{1}{b}\cos\frac{\pi x}{a}\sin\frac{\pi y}{b}\,\hat{\mathbf{a}}_y,$$

find the scalar potential $V_m(x, y)$ in this region.

7-19 Prove that the field generated by an infinite, uniform solenoid that is centered along the z-axis is of the form $\mathbf{B} = B_z(\rho)\hat{\mathbf{a}}_z$ by considering the current distribution as symmetric pairs of loops.

8

Magnetostatic Fields In Material Media

8-1 Introduction

Now that we have discussed the magnetic fields generated by currents in free space, we are ready to see how materials can alter these fields. Magnetic fields interact with materials because materials are composed of charges in motion that are acted upon by forces when a magnetic field is applied. These forces create current distributions on and within those materials that can substantially alter the total magnetic field. By choosing the right materials, it is usually possible to manipulate the magnetic fields generated by a system to fit the requirements of a specific engineering application.

In this chapter, we will discuss two kinds of mechanisms in which magnetic fields interact with materials. The first is called the Hall effect, which involves the movement of free charge when a magnetic field is applied to a material. The second is the interaction of an applied magnetic field with the orbital and spin currents that naturally occur in atoms and molecules. As we shall see, this latter phenomenon allows certain types of materials to greatly enhance the magnetic fields created by other sources, or even act as independent sources (as in the case of permanent magnets).

After discussing the basic types of interactions between materials and magnetic fields, we will address how these material effects can be incorporated into magnetic field calculations. Problems of this sort are called magnetostatic boundary value problems, and we will discuss both analytical and graphical methods for solving them. We will also discuss an important class of magnetostatic configurations called magnetic circuits. These networks are encountered often in engineering practice and can be analyzed using a circuit theory that is analogous to electric circuit theory.

8-2 The Hall Effect

Just as the free charges in materials can move in response to an electric field, they can also move under the influence of a magnetic field. For example, Figure 8-1 shows a conducting strip of width w and thickness t that carries a bias current I_o, flowing in the $+x$ direction. According to the Lorentz force law, the total force $\mathbf{dF}$ acting on each differential volume dv of free charge in the strip is the sum of an electric force $\mathbf{dF}_e$ and a magnetic force $\mathbf{dF}_m$; that is,

$$\mathbf{dF} = \mathbf{dF}_e + \mathbf{dF}_m = \rho_v dv(\mathbf{E} + \mathbf{u} \times \mathbf{B}),$$

where ρ_v is the volume charge density of the mobile charge carriers and $\mathbf{u}$ is the velocity of the moving charges. When $\mathbf{B} = 0$, the only E-field present is the $+x$ directed field that is caused by the battery. Since the bias current I_o flows through a cross-sectional area $S = wt$, the current density is I_o/wt. Using $\mathbf{J} = \rho_v \mathbf{u}$ (Equation (3.18)), we can express the velocity $\mathbf{u}$ of the free charges in the strip as

$$\mathbf{u} = \frac{I_o}{\rho_v wt}\hat{\mathbf{a}}_x. \tag{8.1}$$

According to Equation (8.1), negative charge carriers move in the $-\hat{\mathbf{a}}_x$ direction, since $\rho_v < 0$. The opposite is true for positive charge carriers, such as holes in semiconductors. When a magnetic field $\mathbf{B} = B_o\hat{\mathbf{a}}_z$ is applied, each charge in the sample experiences a magnetic force $\mathbf{dF}_m = \rho_v \mathbf{u} \times \mathbf{B}dv$. Substituting, we obtain

$$\mathbf{dF}_m = -\frac{I_o B_o dv}{wt}\hat{\mathbf{a}}_y,$$

which shows that the force acting on the moving charge carriers is in the same direction for both positive and negative charges. When the moving charge carriers are negative, this force results in a buildup of negative charge on the left-hand edge of the sample and a corresponding positive charge on the right-hand edge. Conversely, when the moving charge carriers are positive, the charge buildup on the edges of the sample is

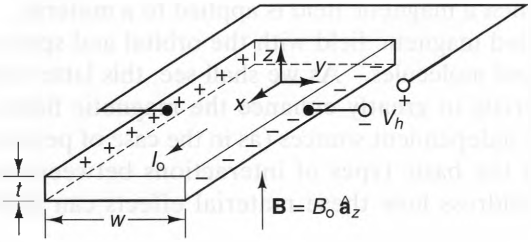


Figure 8-1 The Hall effect.

just the opposite. (This is the case depicted in Figure 8-1.) For both cases, the migration of charges towards the sample edges continues until the electric field that they generate exactly counteracts the $\mathbf{u} \times \mathbf{B}$ force. This electric field $\mathbf{E}_h$ is called a ***Hall field*** and, at equilibrium, is given by

$$\mathbf{E}_h = -\mathbf{u} \times \mathbf{B} = \frac{I_o B_o}{\rho_v w t} \hat{\mathbf{a}}_y . \tag{8.2}$$

The Hall field produces a ***Hall voltage*** V_h across the strip that is given by

$$V_h = E_h w = \frac{I_o B_o}{\rho_v t} , \tag{8.3}$$

where the left-hand edge is assumed to be the positive terminal.

As can be seen from the preceding expression, the polarity of the Hall voltage V_h is determined by the sign of the volume charge density ρ_v of the mobile charge carriers. It is for this reason that the Hall effect was instrumental in proving the existence of holes in semiconductors.[1] For n-type semiconductors, ρ_v is negative, since the free-charge carriers are conduction band electrons. For p-type semiconductors, missing valence electrons can be treated as positive-valued particles called ***holes***, resulting in a positive, mobile charge density. Thus, it is possible to distinguish whether a semiconductor sample is n or p type by measuring the polarity of the induced Hall voltage when the sample is subjected to a magnetic field that is perpendicular to the current flow.

Hall sensors are routinely used to detect magnetic fields with intensities from 10^{-11} to 1 tesla. The following example considers the Hall voltage produced in a semiconductor by the earth's magnetic field.

Example 8-1

The earth's magnetic field has a nominal value of 0.5×10^{-4} [T], or 0.5 gauss, on the earth's surface. Calculate the maximum Hall voltage V_h that could be induced in a silicon strip of thickness $t = 0.1$ [mm] and that carries a bias current of 100 [mA]. Assume that the silicon has been doped with acceptor atoms, resulting in a hole charge density of +160 [C/m³] and a negligible electron density.

Solution

Substituting the preceding values into Equation (8.3), we obtain

$$V_h = \frac{I_o B_o}{\rho_v t} = \frac{(100 \times 10^{-3}\,[\mathrm{A}])(0.5 \times 10^{-4}\,[\mathrm{T}])}{(160\,[\mathrm{C/m^3}])(0.1 \times 10^{-3}\,[\mathrm{m}])} = 0.313\,[\mathrm{mV}].$$

Notice that this voltage is well within the range of measurable voltages.

Sensitive Hall compasses can be made by placing two identical Hall sensors at right angles to each other and biasing them with identical currents, as shown in Figure

[1] See William Shockley, *Electrons and Holes in Semiconductors,* (D. Van Nostrand Company, 1950.) New York

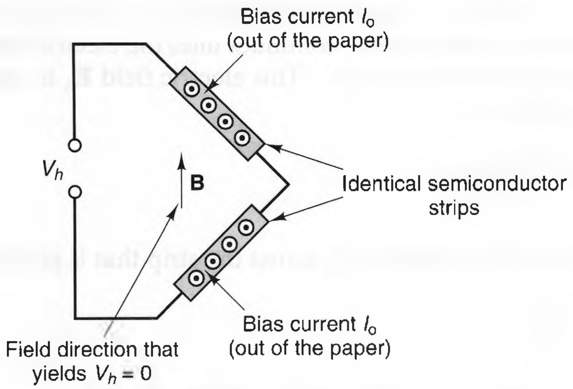


Figure 8-2 A Hall-sensor compass.

8-2. When the compass is aligned either parallel or antiparallel with the earth's magnetic field, the Hall voltages of the two sensors are opposite, producing a null output voltage. When such a null voltage is obtained, the heading (north or south) can be determined by measuring the sign of either of the Hall sensor outputs.

8-3 Magnetic Materials

Just as electric fields affect the bound charges in material media, so do magnetic fields. The physical mechanism of interaction, however, is very different, because magnetic fields interact only with moving charges. A comprehensive description of these atomic currents and their interactions with magnetic fields demands full use of the quantum theory. Fortunately, classical models of atoms are adequate for predicting most properties of materials.

8-3-1 ORBITAL AND SPIN CURRENTS

Currents and magnetic dipole moments are naturally present within atoms and molecules because of two types of electron motion:[2] orbital motion and spin motion. Figure 8-3a shows an electron orbiting a nucleus. The current flowing along the orbital path is

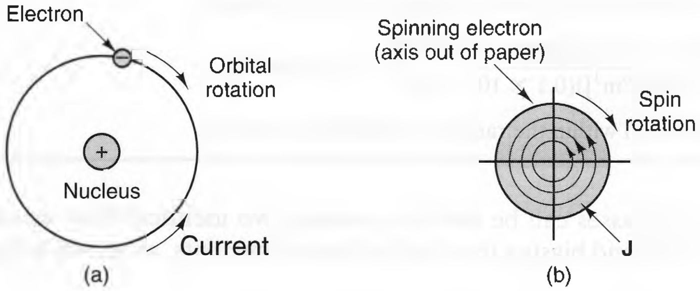


Figure 8-3 Molecular currents.
a) Orbital current. b) Spin current.

[2] There is also a contribution due to the spin motion of the nucleus, but it is orders of magnitude smaller than the electron orbit and spin contributions.

$$I_o = \frac{\text{charge/orbit}}{\text{time/orbit}} = \frac{e}{2\pi/\omega_o} = \frac{e\omega_o}{2\pi} \qquad [\text{A}], \tag{8.4}$$

where ω_o is the angular velocity of the electron and e is the electron charge. Substituting Equation (8.4) into Equation (7.62), we find that the magnetic dipole moment of this current loop is

$$\mathbf{m}_o = I_o S_o \hat{\mathbf{a}}_z = \frac{e\omega_o r_o^2}{2} \hat{\mathbf{a}}_z \qquad [\text{A} \cdot \text{m}^2], \tag{8.5}$$

where S_o and r_o are the surface area and radius of the orbit, respectively.

Electrons have another angular momentum and a related magnetic dipole moment that are independent of their orbital motion. The physical structure of the electron that gives rise to these properties is not completely understood and is an area of intense research, but it can be accounted for by assuming that electrons are solid particles of charge that spin on an axis. Figure 8-3b shows this classical model of a spinning electron. Here, the circulation of a charge about an axis of rotation constitutes an infinite number of current loops whose magnetic moments sum to create a total spin dipole moment of $\mathbf{m}_s$. The magnitude of $\mathbf{m}_s$ always has a value of $eh/4\pi m_e$, where h is Plank's constant (6.6262×10^{-34} [J • s]) and m_e is the mass of the electron. The direction of $\mathbf{m}_s$ for each electron aligns itself either parallel or antiparallel with an applied B-field, depending on the quantum state of the electron.

In the absence of an applied magnetic field from external sources (which we will hereafter call a ***magnetizing field***), the spin and orbital magnetic moments within most materials are randomly oriented and produce no net B-field. In most of the materials this cancellation takes place within the atoms and molecules themselves, resulting in a net zero magnetic moment for each atom or molecule.

In some materials, each atom or molecule has a nonzero magnetic moment, but these moments normally distribute themselves randomly so as to cancel when no magnetizing field is present. However, it is sometimes possible to lock at least some of these dipoles into aligned orientations, resulting in a net magnetic moment throughout the material. When this occurs, the result is a permanent magnet.

Two things can happen to the atomic orbital and spin dipole moments when a magnetizing field is applied to a material:

(1) In materials composed of atoms or molecules that naturally have nonzero dipole moments when no magnetizing field is present, a magnetizing field exerts a torque on each dipole and tends to align it with the field. This produces a net magnetic dipole moment that is parallel to the magnetizing field. The dipole moments increase as the field increases, until saturation occurs, where all the dipoles are completely aligned. Because the atomic magnetic moments align themselves with the magnetizing field, the net B-field within the material increases.

(2) In materials whose atoms or molecules ordinarily do not have a net dipole moment, the application of a magnetizing field will slightly decrease each orbital magnetic moment, leaving the spin magnetic moments unchanged. This upsets the balance between the orbital and spin moments and results in a net magnetic

moment that is antiparallel to the magnetizing field. The net result is a reduction of the net B-field within the material.

Regardless of the physical mechanisms by which they are formed, the magnetic dipoles induced (or aligned) within a material change the net B-field both inside and outside the material. Thus, both the free currents between the molecules and bound currents within them must be known in order to calculate the B-field produced by a given device or system. Fortunately, this is not as difficult as it seems, since the molecular currents are usually proportional to the magnetizing field. In the sections that follow, we will develop the methodology that makes these calculations possible.

8-3-2 MAGNETIC SUSCEPTIBILITY AND EQUIVALENT MAGNETIZATION CURRENTS

The first step to modeling the macroscopic properties of magnetic materials is to introduce a vector called the ***magnetization vector*** **M**, which indicates the net magnetic dipole moment per unit volume at each point throughout a magnetic material. This vector is obtained by taking the limit of the sum of all the dipole moments within a differential volume; that is,

$$\mathbf{M} \equiv \lim_{\Delta v \to 0} \frac{\sum_{k=1}^{N\Delta v} \mathbf{m}_k}{\Delta v} \qquad [\text{A/m}], \tag{8.6}$$

where N is the total number of dipoles in the volume Δv and $\mathbf{m}_k$ is the magnetic dipole moment of the k^{th} dipole. The magnetization of a material is directly related to the movement of its bound charges. This type of current is called ***magnetization current***. We will show in the development which follows that the magnetization current in a material is related to **M** in a way similar to the relationship between the polarization charge density and the polarization **P** in an electrically polarized material.

Figure 8-4 shows a volume V that contains a magnetic material with a magnetization **M** defined at each point. Each differential volume dv' within the material can be considered as a differential magnetic dipole of moment $\mathbf{dM} = \mathbf{M}(\mathbf{r}')dv'$, where the notation $\mathbf{M}(\mathbf{r}')$ indicates that the magnetization is evaluated at the field point $\mathbf{r}'$ which

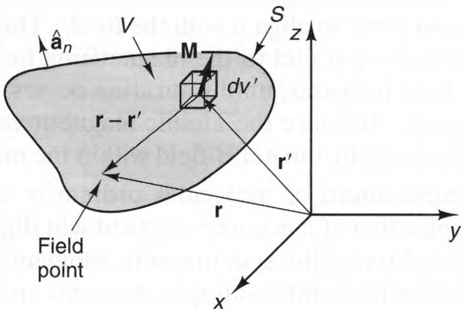


Figure 8-4 Geometry for determining the magnetic vector potential **A** of an arbitrary magnetostatic current distribution in terms of the magnetization vector **M**.

lies inside the differential volume dv'. Using Equation (7.66), we can write the vector-potential contribution $\mathbf{dA}$ due to an arbitrary point in this volume as

$$\mathbf{dA} = \frac{\mu_o \mathbf{M}(\mathbf{r}') \times (\mathbf{r} - \mathbf{r}')\, dv'}{4\pi |\mathbf{r} - \mathbf{r}'|^3} = \frac{\mu_o}{4\pi} \mathbf{M}(\mathbf{r}') \times \nabla' \frac{1}{|\mathbf{r} - \mathbf{r}'|}\, dv'. \tag{8.7}$$

Here, we have used the identity $\nabla' [1/|\mathbf{r} - \mathbf{r}'|] = (\mathbf{r} - \mathbf{r}')/|\mathbf{r} - \mathbf{r}'|^3$, which is the same as Equation (7.17), except that the differentiation is with respect to the primed variables. Integrating Equation (8.7) over the volume V, we obtain

$$\mathbf{A} = \frac{\mu_o}{4\pi} \int_{\text{Vol.}} \mathbf{M}(\mathbf{r}') \times \nabla' \left(\frac{1}{|\mathbf{r} - \mathbf{r}'|} \right) dv'. \tag{8.8}$$

Equation (8.8) can be manipulated into a form where the relationship between the magnetization $\mathbf{M}$ and the volume and surface magnetization currents can be clearly seen. Since the derivation takes several steps, let us start by simply stating the final result, namely,

$$\mathbf{A} = \frac{\mu_o}{4\pi} \int_{\text{Vol.}} \frac{\nabla' \times \mathbf{M}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\, dv' + \frac{\mu_o}{4\pi} \oint_S \frac{\mathbf{M} \times \hat{\mathbf{a}}_n}{|\mathbf{r} - \mathbf{r}'|}\, ds', \tag{8.9}$$

where S is the surface that bounds the volume V and the unit vector $\hat{\mathbf{a}}_n$ points outward from S. To see what these integrals mean, let us compare them with the magnetic vector potential of a current distribution that consists of volumetric current $\mathbf{J}$ in a volume V and surface current $\mathbf{J}_s$ on the surface S that surrounds V. Using Equations (7.42) and (7.43), we see that the magnetic vector potential of such a current distribution is given by

$$\mathbf{A} = \frac{\mu_o}{4\pi} \int_{\text{Vol.}} \frac{\mathbf{J}}{|\mathbf{r} - \mathbf{r}'|}\, dv' + \frac{\mu_o}{4\pi} \oint_S \frac{\mathbf{J}_s}{|\mathbf{r} - \mathbf{r}'|}\, ds', \tag{8.10}$$

where, as we recall, $\mathbf{J}$ and $\mathbf{J}_s$ are both functions of the primed coordinates. Comparing the volume integral of Equation (8.9) with that of Equation (8.10), we can conclude that $\nabla' \times \mathbf{M}$ represents a magnetization current density $\mathbf{J}_m$. Similarly, equating the integrands of the surface integrals, we can conclude that $\mathbf{M} \times \hat{\mathbf{a}}_n$ represents a magnetic surface current density $\mathbf{J}_{sm}$. Thus, the volume and surface currents caused by the magnetization $\mathbf{M}$ are given by

$$\mathbf{J}_m = \nabla \times \mathbf{M} \qquad [\text{A/m}^2] \tag{8.11}$$

$$\mathbf{J}_{sm} = \mathbf{M} \times \hat{\mathbf{a}}_n \qquad [\text{A/m}]. \tag{8.12}$$

In Equation (8.12), we have dropped the prime notation from the curl operation to indicate the $\mathbf{J}_m$ as a function of unprimed coordinates. Also, the subscript m indicates that these are magnetization currents, resulting from the movement of bound molecular charges.

To derive Equation (8.9) from Equation (8.8), we start by using Equation (B.4) to expand the integrand of Equation (8.8):

$$\mathbf{M}(\mathbf{r}') \times \nabla' \left(\frac{1}{|\mathbf{r} - \mathbf{r}'|} \right) = \frac{1}{|\mathbf{r} - \mathbf{r}'|} \nabla' \times \mathbf{M}(\mathbf{r}') - \nabla' \times \left(\frac{\mathbf{M}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \right).$$

Substituting, we obtain

$$\mathbf{A} = \frac{\mu_o}{4\pi} \int_{\text{Vol.}} \frac{\nabla' \times \mathbf{M}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} dv' - \frac{\mu_o}{4\pi} \int_{\text{Vol.}} \nabla' \times \left(\frac{\mathbf{M}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \right) dv'. \tag{8.13}$$

The second volume integral in this expression can be written as a surface integral by using the vector identity

$$\int_{\text{Vol.}} \nabla' \times \mathbf{F} dv' = -\oint_S \mathbf{F} \times d\mathbf{s}', \tag{8.14}$$

where the volume is bounded by the closed surface S. This identity is similar to Stokes's theorem and can be proved using the vector theorems in Appendix B. By means of Equation (8.14), Equation (8.13) becomes

$$\mathbf{A} = \frac{\mu_o}{4\pi} \int_{\text{Vol.}} \frac{\nabla' \times \mathbf{M}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} dv' + \frac{\mu_o}{4\pi} \oint_S \frac{\mathbf{M} \times \hat{\mathbf{a}}_n}{|\mathbf{r} - \mathbf{r}'|} ds', \tag{8.15}$$

which is the desired result.

The relationships given by Equations (8.9) and (8.10) show that magnetization currents are present whenever **M** has curl, or a nonzero tangential component, at the edge of a magnetized material. Figure 8-5a depicts a situation in which **M** has curl at the point P. As can be seen, the counterclockwise circulation of **M** results in a net current flow out of the paper at P. Conversely, Figure 8-5b depicts a situation in which **M** is uniform throughout the material. Since $\nabla \times \mathbf{M}$ is zero inside the material, there is a net cancellation of the circulating atomic currents at each interior point, and thus,

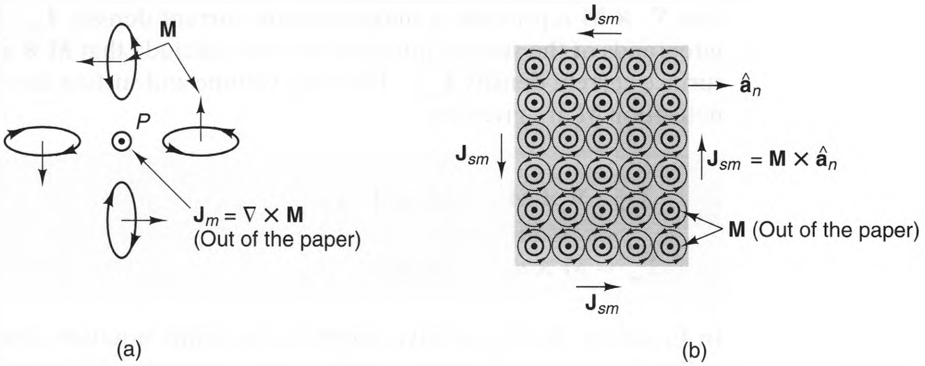


Figure 8-5 a) A volumetric magnetization current produced by a medium where $\nabla \times \mathbf{M} \neq 0$. b) A surface magnetization current produced by a uniform magnetization **M**.

$\mathbf{J}_m = 0$. At the material edges, however, there is a surface current $\mathbf{J}_{sm} = \mathbf{M} \times \hat{\mathbf{a}}_n$, since the tangential currents are not canceled there.

8-3-3 THE MAGNETIC FIELD INTENSITY

The magnetization currents on and within magnetic materials generate secondary magnetic fields that can be substantial and must be accounted for. To accomplish this, we will start with Ampère's law, taking into account both free and magnetization currents:

$$\nabla \times \mathbf{B} = \mu_o \mathbf{J}_T = \mu_o(\mathbf{J} + \mathbf{J}_m), \tag{8.16}$$

Here $\mathbf{J}$, $\mathbf{J}_m$, and $\mathbf{J}_T$ are the free, magnetization, and total current densities, respectively.[3] Equation (8.16) is always valid, but the term $\mu_o \mathbf{J}_m$ is troublesome, because it represents an induced current density and, hence, is usually a function of $\mathbf{B}$. We can eliminate $\mathbf{J}_m$ from this equation, however, by substituting $\nabla \times \mathbf{M}$ for $\mathbf{J}_m$ and collecting terms. This yields

$$\nabla \times \left(\frac{\mathbf{B}}{\mu_o} - \mathbf{M}\right) = \mathbf{J}. \tag{8.17}$$

The right-hand side of Equation (8.17) involves only free current, so it is a more attractive differential equation to solve than Equation (8.16). Because of this, the quantity in parentheses on the left-hand side of Equation (8.17) is defined as a new electromagnetic quantity, which we call the *magnetic field intensity*,

$$\mathbf{H} \equiv \frac{\mathbf{B}}{\mu_o} - \mathbf{M} \qquad [\text{A/m}], \tag{8.18}$$

which is measured in units of [A/m]. Using this definition, we can now write Ampère's law as

$$\nabla \times \mathbf{H} = \mathbf{J}. \tag{8.19}$$

Writing Ampère's law in terms of $\mathbf{H}$ greatly simplifies the handling of bound currents within magnetic materials, since the right-hand side is now simply the free current density $\mathbf{J}$, rather than the total current density.

The magnetization $\mathbf{M}$ within most materials is directly linked to the magnetic torques produced by the B-fields acting on the orbiting and spinning electrons in each atom or molecule. (See example 7-4.) In linear, isotropic media, $\mathbf{M}$ is proportional to $\mathbf{B}$, which means (according to Equation (8.18)) that $\mathbf{H}$ is also proportional to $\mathbf{B}$. Thus, we can write

$$\mathbf{M} = \chi_m \mathbf{H}, \tag{8.20}$$

[3] Hereafter, the symbol $\mathbf{J}$ will always denote the *free* current density.

where χ_m is a unitless scalar called the ***magnetic susceptibility*** and is a measure of the ease with which magnetic dipoles are formed in a medium. Substituting Equation (8.20) into Equation (8.18) yields the following constitutive relation:

$$\mathbf{B} = \mu_o(1 + \chi_m)\mathbf{H}. \tag{8.21}$$

The unitless quantity $1 + \chi_m$ is called the ***relative permeability*** of the medium and is denoted by the symbol μ_r. Thus, the constitutive relationship between **B** and **H** can also be written as

$$\mathbf{B} = \mu_r\mu_o\mathbf{H} = \mu\mathbf{H}, \tag{8.22}$$

where $\mu = \mu_r\mu_o$ is the permeability of the medium.

As in the case of electric materials, we define a simple medium as one that is homogeneous, linear, and isotropic. The meanings of these three terms are essentially unchanged. For instance, μ_r is a constant with respect to position in a homogeneous material. Likewise, μ_r is independent of the field level in a linear material. Finally, **B** and **H** always point in the same direction in isotropic materials, which means that μ_r is a scalar in these materials.

8-3-4 THE PHYSICAL PROPERTIES OF MAGNETIC MATERIALS

Table C-5 in Appendix C lists the relative permeabilities of a number of materials commonly used in engineering practice. As can be seen from this table, there is a wide range of values. Typically, magnetic materials are classified according to the nature of their response to a magnetizing field. The major classes are diamagnetic, paramagnetic, ferromagnetic, antiferromagnetic, and ferrimagnetic materials.

We will now discuss the physical mechanisms responsible for the magnetic behavior of each class of materials. We do this by describing the response of the orbital and spin magnetic dipoles to the presence of a uniform magnetizing field $\mu_o\mathbf{H}_m$ that is generated by external sources (such as the free current in the wires of a solenoid). During this discussion, we will assume that the magnetic material sample does not upset the direction of the the B-field when it is placed in the magnetizing field. For this case, when the material is present, $\mathbf{B} = \mu_o(\mathbf{H}_m + \mathbf{M})$, so $\mu_o\mathbf{M} = \mu_o(\mu_r - 1)\mathbf{H}_m$ can be thought of as the response of the material to the magnetizing field.[4]

8-3-4-1 Diamagnetic Materials. In *diamagnetic materials*, all of the orbital and spin magnetic moments pair off in the absence of a magnetizing field in such a way that each atom has no net magnetic moment. When a magnetizing field is applied, each

[4] This cannot be said when the direction of **B** changes when the material is present in the magnetizing field.

orbiting electron experiences an outward Lorentz force that slightly decreases the electron orbital velocities. This reduces the orbital magnetic moments and leaves the spin moments unchanged, producing a small, net magnetic moment in each molecule that is antiparallel to the magnetizing field. The net effect is a slight reduction of the B-field inside the material, which means that μ_r is less than unity for diamagnetic materials. Examples of diamagnetic materials include the inert gases, hydrogen, copper, gold, silicon, germanium, graphite, and bismuth. Of these, bismuth has the most pronounced diamagnetic effect, with $\mu_r = 0.9999833$.

Since they have permeabilities that are less than the permeability of free space, diamagnetic materials are repelled by permanent magnets, although the repulsive force is small. This repulsion was first discovered by Michael Faraday, who found that bismuth is repelled by a strong bar magnet. We will discuss the cause of the repulsive force in Chapter 9.

Superconductors exhibit ***perfect diamagnetism*** when they are in their perfectly conducting state. This means that $\chi_m = -1$, $\mu_r = 0$, and $\mathbf{B} = 0$ inside a superconductor when operated below its critical temperature T_c. Figures 8-6 a & b show the B-field streamlines of a typical superconductor when placed in a uniform B-field at temperatures above and below the critical temperature, respectively. When $T > T_c$, the material acts as an ordinary diatomic material, with $\mu_r \approx 1.0$. Thus, the B-field lines of Figure 8-6a are uniform.

Figure 8-6b shows the B-field when the temperature of the material is below T_c. Here, the streamlines do not enter the material, since the field inside is zero. Although it may look like the superconductor somehow expels the B-field lines, what really happens is that the magnetic dipoles within the superconductor align themselves antiparallel with the magnetizing field and generate a B-field that exactly cancels the magnetizing field inside.

Superconductors are strongly repulsed by permanent magnets, which makes them useful for suspending objects without mechanical constraints. Figure 8-7 shows a permanent magnet suspended above a superconducting disc that is maintained below the critical temperature. Unlike the force between two permanent magnets, which can be either repulsive or attractive (depending upon the orientations of their north and south poles), a superconductor experiences a repulsive force from a permanent magnet regardless of its orientation.

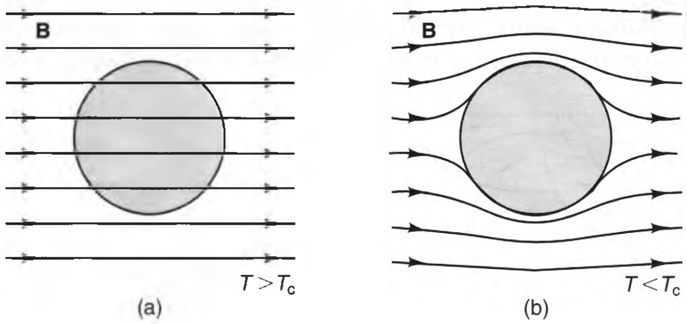


Figure 8-6 B-field streamlines of a superconductor placed in a uniform B-field. a) Above the critical temperature. b) Below the critical temperature.

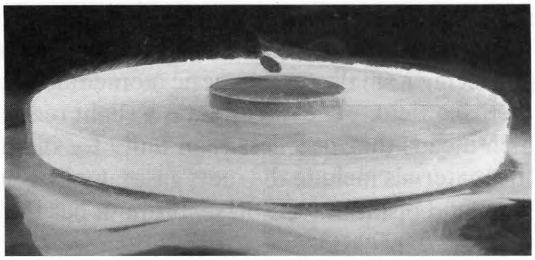


Figure 8-7 A small permanent magnet is suspended above a superconducting disc of barium-yttrium-copper-oxygen compound, which is cooled with liquid nitrogen. Courtesy of Edmund Scientific Co.

8-3-4-2 Paramagnetic Materials. Each atom or molecule of a ***paramagnetic material*** has a net magnetic moment because its outer electron shells are not filled. In the absence of a magnetizing field, the spin magnetic moments assume random orientations and the average magnetic moment is zero. However, when a magnetizing field is applied, quantum effects cause slightly more than half of the magnetic moments to align themselves parallel to this field, and the rest antiparallel. Hence, a net magnetic moment is induced in the material parallel to the magnetizing field. This net magnetization increases as the magnetizing field increases. The diamagnetic effect is also present in these materials, but it is masked, because the decrease of the orbital magnetic moments is much smaller than the increase of the average spin magnetic moments due to the paramagnetic effect.

Paramagnetic materials have relative permeabilities that are slightly greater than unity. One of the strongest paramagnetic materials is nickel chloride, which has $\mu_r = 1.00004$. Since the paramagnetic effect results in a small increase in the B-field within the material, these materials are attracted to permanent magnets, although the attractive force is usually small. Other examples of paramagnetic materials are air, aluminum, potassium, and tungsten.

8-3-4-3 Ferromagnetic Materials. Like paramagnetic materials, ***ferromagnetic materials*** are composed of atoms and molecules that each have a spin magnetic moment resulting from the incomplete cancellation of electron spins in the outer orbital shells. But unlike paramagnetic materials, in which only slightly more than half of the spin magnetic moments align themselves with a magnetizing field, ferromagnetic materials exhibit a quantum effect called an ***exchange force*** that causes all the magnetic moments in small regions throughout the material to "lock" in the same direction, even in the absence of a magnetizing field. These regions are called ***ferromagnetic domains*** and are depicted in Figure 8-8. Between adjacent ferromagnetic domains are thin, disordered layers called ***domain walls***. When a magnetizing field is applied to a sample that is initially unmagnetized (i.e., the domain moments are randomly oriented),

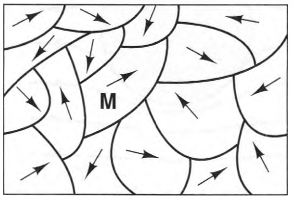


Figure 8-8 Magnetic domains in a ferromagnetic material.

the domain walls tend to shift so that those domains already aligned with the magnetizing field grow at the expense of the others. For small fields, this wall movement is reversible, meaning that they will return to their original positions when the magnetizing field is removed. As the magnetizing field is increased, however, this wall movement becomes irreversible, requiring a negative magnetizing field to bring the walls to their original positions. As the field is further increased, the remaining unaligned domains eventually rotate, making the material appear as a single magnetized domain. At this point, the material is ***saturated***: Further increases in the magnetizing field produce no further increases in **M**.

Since the ferromagnetic domains can maintain their alignments even when the magnetizing field is removed, these materials make excellent permanent magnets. Once magnetized, they can maintain their magnetized state over long periods of time as long as they are kept at temperatures below the ***Currie temperature***. Above this temperature, the random thermal energy forces overshadow the exchange forces, and the material returns to an unmagnetized state, behaving as a paramagnetic material. The Currie temperatures of most ferromagnetic materials lie in the range from 150 to 1000 degrees Celsius.

Figure 8-9 shows the relationship between B and H inside a typical ferromagnetic material as the magnetizing field $\mu_0 H$ (produced by external currents) oscillates between the values $\pm\mu_0 H_{\max}$. If the sample is initially unmagnetized, B follows the dotted curve (called the initial magnetization curve) as H increases toward $H_{\max}$. This curve starts out linear, but eventually levels off as saturation is reached. As H decreases from this point, B at first falls off slowly due to the "memory" of the material. Later, the decrease in B is more rapid, as the domain walls begin to shift. When $H = 0$, B has a positive value B_r, called the ***residual flux density***, but it is not until $H = -H_c$ (the ***coercive field intensity***) that $B = 0$. As H decreases further towards $-H_{\max}$, B continues to decrease and approaches saturation in the negative direction. From this point, the material goes through a similar process as H increases from $-H_{\max}$ to $H_{\max}$ and B follows the bottom curve shown in the figure.

The S-shaped magnetization curve shown in Figure 8-9 is called a ***hysteresis loop***. The area of the hysteresis loop corresponds to the energy that must be expended to bring the material through one cycle. This energy is called ***hysteresis loss***. "Soft" materials have narrow hysteresis loops and are well suited for motors, transformers, and magnetic-recording read/write heads, because the loss per cycle is relatively low. Other applications make use of the memory characteristics of materials with wide hys-

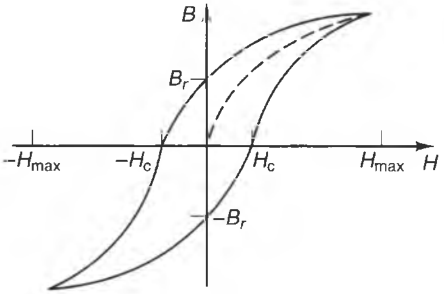


Figure 8-9 A hysteresis loop.

teresis loops. Permanent magnets and recording emulsions for tapes and disks are two examples. Materials with wide hysteresis loops are called "hard" materials.

Ferromagnetic materials exhibit the largest relative permeabilities of all materials and are strongly attracted to permanent magnets. Examples include iron, cobalt, nickel, and steel. Relative permeablities exceeding 10^5 are not uncommon for ferromagnetic materials, although values in the range of 100–20,000 are more typical. These permeabilities are strong functions of frequency, however. At frequencies above a few tens of kilohertz, the relative permeabilities of most ferromagnetic materials are essentially unity (1.0). As a result, ferromagnetic materials are usually used for low-frequency applications, such as power frequency devices (50–60 [Hz]).

Another characteristic of ferromagnetic materials is that they are moderate conductors. Unfortunately, this quality is usually unattractive, because it gives rise to ***eddy currents*** and their associated losses when time-varying fields are present. ***Eddy current losses*** are discussed in Chapter 9. As a group, ferromagnetic materials exhibit the largest eddy current losses of all magnetic materials.

8-3-4-4 Antiferromagnetic and Ferrimagnetic Materials. The atoms and molecules in ***antiferromagnetic materials*** each have a net magnetic moment due to uncompensated spin moments, but the interaction forces between adjacent atoms are such that their moments align antiparallel. Thus, there is no net magnetic moment within an antiferromagnetic material, resulting in a relative permeability of unity.

Ferrimagnetic materials are similar to antiferromagnetic materials in that adjacent atomic moments align antiparallel, but the magnitudes of these adjacent moments are not equal, so they do not cancel. The net magnetic moments found in ferrimagnetic materials are smaller than those found in ferromagnetic materials, but they can still be substantial.

An important subset of the ferrimagnetic materials is the ***ferrites***, which have conductivities that are several orders of magnitude lower than those found in ferromagnetic materials. The resulting low eddy current losses of these materials make them attractive for high-frequency applications, where losses can significantly reduce the efficiency or the Q (quality factor) of tuned circuits. Iron oxide magnetite and nickel–zinc ferrite are examples of ferrite materials. Relative permeabilities in the range of 1,500 are typical for ferrites. Also, unlike ferromagnetic materials, many ferrites exhibit relatively large permeabilities at high-frequencies. For example, nickel–zinc (NiZn) exhibits a relative permeability of nearly 100 at frequencies approaching 100 [MHz]. This high-frequency characteristic makes ferrites attractive for radio frequency (RF) and microwave applications.

Ferrites exhibit anisotropic behavior when subjected to high-frequency magnetic fields. This occurs because electrons have an angular momentum that is always antiparallel to their magnetic moment. As a magnetizing field attempts to align the magnetic moments of these electrons, their spin angular momentum causes the electrons to precess about the direction of $\mathbf{H}_m$, just like a gyroscope precesses about the gravitational field. Thus, $\mathbf{M}$ precesses around $\mathbf{H}_m$, causing $\mathbf{B}$ and $\mathbf{H}_m$ to have different directions.

This phenomenon is useful in the design of many high-frequency components, such as isolators, phase shifters, and oscillators.[5]

8-3-4-5 Superparamagnetic Materials. ***Superparamagnetic*** materials consist of ferromagnetic particles that are suspended in a dielectric binder. The binder breaks the exchange forces between adjacent particles, which allows each particle to be magnetized independently of all other particles. This composite material can then be deposited as a thin film on a flexible tape, such as Mylar$^{\text{TM}}$. When the particles are very small, the magnetic state of the tape can change rapidly as a function of position, which allows these tapes to store large amounts of information in a small volume. This makes superparamagnetic materials attractive for use in audio-, video-, and data-storage tapes.

8-3-5 FIELD EQUATIONS IN MAGNETIC MATERIALS

Now that we have determined how the orbital and spin currents in a magnetic material distribute themselves in response to an applied magnetic field, we are in a position to derive the equations that model the relationship between magnetostatic fields and their sources when magnetic materials are present. We can start by remembering that for steady currents in free space (i.e., in a vacuum), the magnetic flux density satisfies the equations

$$\left.\begin{aligned}\nabla \times \mathbf{B} &= \mu_0 \mathbf{J}_T \\ \nabla \cdot \mathbf{B} &= 0\end{aligned}\right\} \text{ (Magnetostatic equations in free space),}$$

where $\mathbf{J}_T$ is the total current density at a point. This version of Ampère's law is also valid when magnetic materials are present, as long as $\mathbf{J}_T$ includes both the free and the magnetization current densities, $\mathbf{J}$ and $\mathbf{J}_m$, respectively. This complicates matters, however, since magnetization currents are usually functions of the magnetic field **B**.

We can simplify things considerably by remembering that Ampère's law can also be written in terms of **H** as

$$\nabla \times \mathbf{H} = \mathbf{J},$$

where **J** is the free current density. The right-hand side of this equation is much simpler than the corresponding B-field equation, because more is usually known a priori about the free current in a device or system than is known about the total current density. Using this simplification, we can now write Maxwell's equations for magnetostatics as

$$\left.\begin{aligned}\nabla \times \mathbf{H} &= \mathbf{J} \\ \nabla \cdot \mathbf{B} &= 0\end{aligned}\right\} \begin{aligned}&\text{(Magnetostatic equations in} \\ &\text{magnetic materials)}\end{aligned} \qquad \begin{aligned}(8.23)\\(8.24)\end{aligned}$$

Since both **B** and **H** appear in these equations, we also need the constitutive relation

[5] For more on this subject, see R.E. Collin, *The Theory of Guided Waves,* 2d ed. (New York 1994), IEEE Press.

$$\mathbf{B} = \mu\mathbf{H}. \tag{8.25}$$

Taken as a set, Equations (8.23) and (8.24) are sufficient to model all magnetostatic fields when magnetic materials are present. They can also be expressed in integral form, as

$$\left.\begin{aligned} &\oint_C \mathbf{H}\cdot d\boldsymbol{\ell} = I \qquad (8.26)\\ &\oint_S \mathbf{B}\cdot d\mathbf{s} = 0 \qquad (8.27)\end{aligned}\right\} \text{(Magnetostatic equations in magnetic materials)}$$

where S is a closed surface and I is the free current enclosed in a right-handed sense by the closed path C.

Both the vector and scalar magnetic potentials can be used when magnetic materials are present. For instance, $\nabla\bullet\mathbf{B}$ is zero even when magnetic materials are present, so we still can write

$$\mathbf{B} = \nabla \times \mathbf{A}. \tag{8.28}$$

Substituting this expression into Ampère's law and using Coulomb's gauge, we find that in a homogeneous region, $\mathbf{A}$ satisfies the vector equation

$$\nabla^2\mathbf{A} = -\mu\mathbf{J} \qquad \text{(Homogeneous regions).} \tag{8.29}$$

The expression for the magnetic scalar potential in magnetic media follows from the property that $\nabla \times \mathbf{H} = 0$ in regions where $\mathbf{J} = 0$. Thus, in source-free regions, $\mathbf{H}$ and $\mathbf{B}$ can be respectively written as

$$\mathbf{H} = -\nabla V_m \qquad \text{(Regions where } \mathbf{J} = 0\text{),} \tag{8.30}$$

and

$$\mathbf{B} = -\mu\nabla V_m \qquad \text{(Regions where } \mathbf{J} = 0\text{).} \tag{8.31}$$

Substituting Equation (8.31) into Equation (8.24), we find that V_m satisfies

$$\nabla\cdot\mu\nabla V_m = 0 \qquad \text{(Regions where } \mathbf{J} = 0\text{).} \tag{8.32}$$

In homogeneous media, Equation (8.32) becomes

$$\nabla^2 V_m = 0 \qquad \text{(Homogeneous regions where } \mathbf{J} = 0\text{),} \tag{8.33}$$

which is Laplace's equation.

Example 8-2

Find the B-field generated by the solenoid shown in Figure 8-10 that is filled with a magnetic material with permeability μ.

Solution:

We solved this problem earlier for the case where $\mu = \mu_o$ by using Ampère's law to find $\mathbf{B}$ directly. (See Section 7-4-2.) We can employ a similar procedure here, this time using Ampère's law to find $\mathbf{H}$ and then the constitutive relation $\mathbf{B} = \mu\mathbf{H}$ to find $\mathbf{B}$.

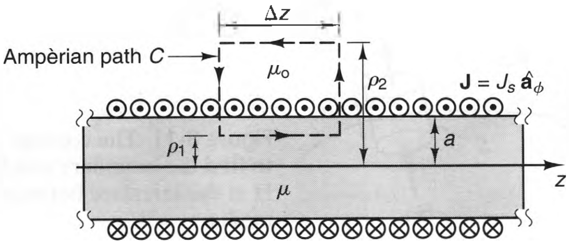


Figure 8-10 An infinite solenoid, filled with a homogeneous magnetic material.

Even with the magnetic material present, the geometry still has perfect cylindrical symmetry. Thus, it is reasonable to assume that both **B** and **H** have only a z-component, which can vary just with the radial coordinate ρ. For the ampèrian path shown in the figure, we can write

$$\oint_C \mathbf{H}\cdot\mathbf{d\ell} = [H_z(\rho_1) - H_z(\rho_2)]\,\Delta\ell = I_{\text{enc}}$$

where I_{enc} is the current enclosed by the path. When $\rho_1 < a$ and $\rho_2 > a$, the enclosed current is $I_{\text{enc}} = J_s\Delta\ell$. Thus,

$$H_z(\rho_1) - H_z(\rho_2) = J_s \qquad (\rho_1 < a \text{ and } \rho_2 > a).$$

Since this expression is valid for all $\rho_1 < a$, we can conclude that $H_z(\rho_1)$ is a constant. Using similar reasoning, we can conclude the same thing about $H_z(\rho_2)$. We can argue that $H_z(\rho_2) = 0$, since the field at infinity is zero. Thus, $H_z(\rho_1) = J_s$, and we can write **B** as

$$\mathbf{B} = \begin{cases} \mu J_s\hat{\mathbf{a}}_z & \text{(Inside solenoid)} \\ 0 & \text{(Outside solenoid)} \end{cases}. \tag{8.34}$$

8-3-6 MAGNETIC FIELD BOUNDARY CONDITIONS

In order to model systems that contain different kinds of magnetic materials, it is necessary to know the boundary conditions that **B** and **H** exhibit across these material discontinuities. These boundary conditions can be obtained using the integral form of Maxwell's equations for magnetostatics:

$$\oint_C \mathbf{H}\cdot\mathbf{d\ell} = I \tag{8.35}$$

$$\oint_S \mathbf{B}\cdot\mathbf{ds} = 0. \tag{8.36}$$

Let us consider the interface between two magnetic media shown in Figure 8-11, which have permeabilities μ_1 and μ_2, respectively. The contour C has length $\Delta\ell$, has height Δh, and straddles the interface. If $\Delta h \to 0$, we can write

$$\oint_C \mathbf{H}\cdot\mathbf{d\ell} \approx H_{1t}\Delta\ell - H_{2t}\Delta\ell \approx I,$$

where H_{1t} and H_{2t} are the tangential components of **H** just inside regions 1 and 2, respectively.

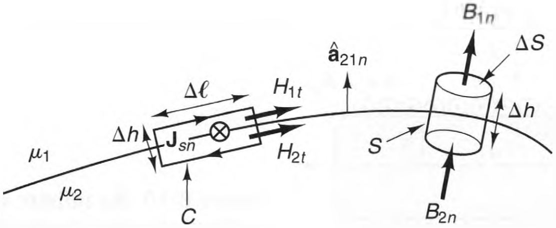


Figure 8-11 The contour and surface used to find the boundary conditions of **B** and **H** at the interface between two dissimilar media.

Since $\Delta h \to 0$, the only finite current I that can pass through the surface bounded by C is a surface current $\mathbf{J}_s$ at the interface. If we replace I with $J_s\Delta\ell$, where J_s is the component of the surface current density that is perpendicular to the loop in a right-handed sense, we find that

$$H_{1t}\Delta\ell - H_{2t}\Delta\ell \approx J_s\Delta\ell.$$

This expression becomes exact in the limit as $\Delta\ell \to 0$. Dividing both sides by $\Delta\ell$, we finally obtain

$$H_{1t} - H_{2t} = J_s. \tag{8.37}$$

Also, $B_t = \mu H_t$, so we can write

$$\frac{1}{\mu_1}B_{1t} - \frac{1}{\mu_2}B_{2t} = J_s. \tag{8.38}$$

The integration contour in Figure 8-11 was chosen to lie in the plane of the paper for simplicity. This choice resulted in a relationship between components of **B** and **H** that lie in the same plane. We can obtain similar expressions for the tangential components of **B** and **H** that lie perpendicular to the paper by choosing a path that extends out of the paper. These expressions can be combined into a single vector expression,

$$\hat{\mathbf{a}}_{21n} \times (\mathbf{H}_1 - \mathbf{H}_2) = \mathbf{J}_s, \tag{8.39}$$

where the unit vector $\hat{\mathbf{a}}_{21n}$ is perpendicular to the interface and points from region 2 towards region 1, and $\mathbf{J}_s$ is the surface current density at the interface. When using Equations (8.37)–(8.39), it is important to note that a surface current $\mathbf{J}_s$ can flow at a material interface only when one of the media is perfectly conducting or when there is an infinitesimally thin conducting film between the two media.

To find the relationship between the normal components of **B** and **H** at a magnetic material interface, we can evaluate Equation (8.36) using the "pillbox" surface shown in Figure 8-11, which has height Δh and top and bottom surface area ΔS. If $h \to 0$, the surface area of the cylindrical portion of the surface becomes zero. Hence, the only contributions to the integral come from the bottom and top surfaces, and we have

$$\oint_S \mathbf{B} \cdot \mathbf{ds} \approx B_{1n} \Delta S - B_{2n} \Delta S = 0,$$

which becomes exact as $\Delta S \to 0$. Dividing both sides by ΔS, we conclude that

$$B_{1n} = B_{2n}. \tag{8.40}$$

Also, since $B_n = \mu H_n$, Equation (8.40) can be written as

$$\mu_1 H_{1n} = \mu_2 H_{2n}. \tag{8.41}$$

Example 8-3

At the material interface shown in Figure 8-12, derive the relationship between the magnitudes and angles of **B** on both sides of the interface. Assume that neither region is perfectly conducting, so there is no surface current.

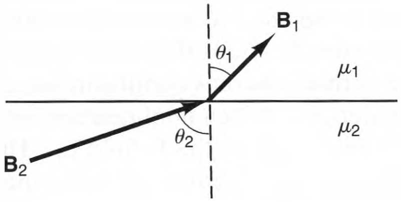


Figure 8-12 An interface between two dissimilar magnetic media.

Solution:

The normal components of $\mathbf{B}_1$ and $\mathbf{B}_2$ are $B_1 \cos\theta_1$ and $B_2 \cos\theta_2$, respectively. Substituting these into Equation (8.40), we find that

$$B_1 \cos\theta_1 = B_2 \cos\theta_2. \tag{8.42}$$

Similarly, the tangential components of $\mathbf{B}_1$ and $\mathbf{B}_2$ are $B_1 \sin\theta_1$ and $B_2 \sin\theta_2$, respectively. Substituting these into Equation (8.38) (with $J_{sn} = 0$, since neither medium is perfectly conducting), we obtain

$$\frac{1}{\mu_1} B_1 \sin\theta_1 = \frac{1}{\mu_2} B_2 \sin\theta_2.$$

Dividing this expression into the previous one and solving for θ_2, we get

$$\theta_2 = \tan^{-1}\left[\frac{\mu_2}{\mu_1} \tan\theta_1\right]. \tag{8.43}$$

Similarly, substituting Equation (8.43) into Equation (8.42) and solving for B_2 results in

$$B_2 = B_1\sqrt{\left(\frac{\mu_2}{\mu_1}\right)^2 \sin^2\theta_1 + \cos^2\theta_1}\,. \tag{8.44}$$

An important special case of Equation (8.43) occurs when $\mu_1 \gg \mu_2$ and $\theta_1 \neq \pm 90°$. Under this condition, we have

$$\lim_{\mu_1\to\infty} \theta_2 = \lim_{\mu_1\to\infty} \tan^{-1}\left[\frac{\mu_2}{\mu_1}\tan\theta_1\right] = 0. \tag{8.45}$$

Thus, B-field lines in a very low-permeability material approach the interface with a high-permeability material at right angles to the surface.

8-4 Magnetostatic Boundary Value Problems

When different kinds of magnetic materials are present in a system, solutions are often found by solving the appropriate differential or integral equations in the region(s) of interest and then enforcing the correct behavior of these solutions at the boundaries of the materials. Problems of this sort are called ***magnetostatic boundary value problems.***

The magnetic vector and scalar potentials satisfy Poisson's and Laplace's equations, respectively, so the techniques used to solve magnetostatic boundary value problems are very similar to those used to solve electrostatic boundary value problems. There are, however, two complications that make magnetostatic problems more difficult to solve. The first is that, whereas nonlinearities are rare in dielectrics, they are common in magnetic materials. When nonlinearities are present, the value of μ must be considered as an unknown, just as the B-field is. The second complication is that constant (magnetic) potential surfaces often do not coincide with the material surfaces, because magnetic conductors do not exist in nature. Thus, unlike E-fields, which vanish inside good conductors, B-fields tend to penetrate magnetic materials.

Magnetostatic boundary value problems can be solved using analytical, graphical, or numerical solution methods. These techniques are very similar to those used for electrostatic fields, so we will present only a few. In the paragraphs that follow, we examine two analytical solutions and a graphical technique. Interested readers can find discussions of more advanced techniques in the references at the end of the text.

8-4-1 ANALYTICAL SOLUTIONS

Just as in the electrostatic case, some types of magnetostatic boundary value problems can be solved analytically in terms of simple expressions. These problems all deal with objects with high symmetry, such as solenoids, toroids, spheres, etc. In many cases, the solutions can be found almost by inspection, since the uniqueness principle ensures that only the correct solution can satisfy both the appropriate differential equations and the boundary conditions. We will demonstrate this type of procedure in the following two examples.

Example 8-4

Figure 8-13 shows an infinite solenoid of radius b. The solenoid is partially filled with a solid, cylindrical core with permeability μ and radius a that is centered about the solenoid's axis. If the surface current density around the solenoid is $\mathbf{J}_s = J_s\hat{\mathbf{a}}_\phi$, find the B-field both inside and outside the cylinder.

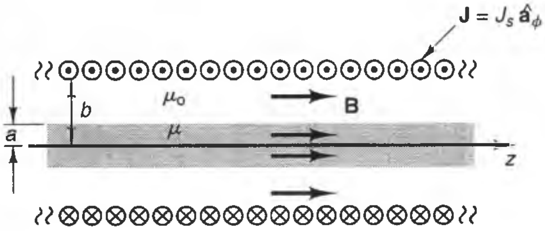


Figure 8-13 An infinite solenoid with a partial magnetic core.

Solution:

We can approach this problem by remembering that the H-field inside the solenoid with the bar absent is (see Section 7-4-2)

$$\mathbf{H} = \begin{cases} J_s\hat{\mathbf{a}}_z & \rho < b \\ 0 & \rho > b \end{cases}.$$

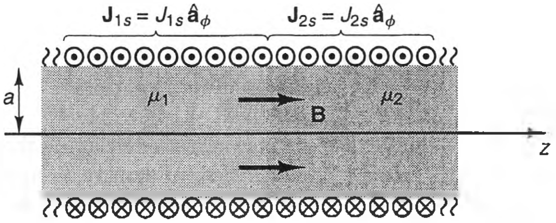
This solution satisfies $\nabla \times \mathbf{H} = 0$ inside and outside the solenoid and also satisfies the boundary condition (Equation (8.37)) across the surface current. The boundary condition imposed by the bar is that the tangential H-field must be continuous across the air-bar interface. However, this H-field already satisfies the condition, so the bar does not impose any new boundary conditions on $\mathbf{H}$. Also, the bar does not add any new free currents to the geometry, so it is reasonable to assume that $\mathbf{H}$ does not change with the addition of the bar. Finally, using $\mathbf{B} = \mu\mathbf{H}$, we obtain

$$\mathbf{B} = \begin{cases} \mu J_s\hat{\mathbf{a}}_z & \rho < a \\ \mu_0 J_s\hat{\mathbf{a}}_z & a < \rho < b. \\ 0 & \rho > b \end{cases}$$

Notice that this expression satisfies the necessary boundary conditions (Equation (8.38)) at $\rho = a$ and $\rho = b$. From this expression, we see that the B-field is greatly enhanced in the magnetic core when $\mu \gg \mu_0$, but is unchanged elsewhere.

Example 8-5

Figure 8-14 depicts a long solenoid that contains two different magnetic materials. The materials



$\mathbf{J}_{1s} = J_{1s}\hat{\mathbf{a}}_\phi$
$\mathbf{J}_{2s} = J_{2s}\hat{\mathbf{a}}_\phi$
a
μ_1
$\mathbf{B}$
μ_2
z

Figure 8-14 An infinite solenoid with an inhomogeneous core and nonuniform surface current.

on the left and right sides of the solenoid have permeabilities μ_1 and μ_2, respectively, and each has the same radius as the solenoid. Find the surface current densities on each side of the solenoid that generate a uniform, axially directed B-field throughout the inside of the solenoid and no field outside.

Solution:

We will start by assuming that $\mathbf{B} = B_o\hat{\mathbf{a}}_z$ throughout the interior of the solenoid and find the surface currents $\mathbf{J}_{1s}$ and $\mathbf{J}_{2s}$ that support this field. The interface between the two magnetic media poses no problem with this assumed B-field, since the normal component of **B** is always continuous across a material interface.

The tangential components of **B** must satisfy Equation (8.38) at the outer radius of the solenoid. Thus, in the left-hand medium, we have

$$\frac{1}{\mu_1}B_o = J_{1s}.$$

Similarly, in the right-hand medium, we have

$$\frac{1}{\mu_2}B_o = J_{2s}.$$

Hence, the required surface current densities along the left- and right-hand sides of the solenoid are

$$\mathbf{J}_{1s} = \frac{1}{\mu_1}B_o\hat{\mathbf{a}}_\phi \qquad \text{[A/m]}$$

and

$$\mathbf{J}_{2s} = \frac{1}{\mu_2}B_o\hat{\mathbf{a}}_\phi \qquad \text{[A/m]},$$

respectively.

We can check this result by recognizing that this B-field is the same as would be generated by a uniform solenoidal surface current in free space. Thus, the total (i.e., free plus magnetization) current density on the solenoid boundary must be uniform along the entire solenoid. In the left-hand region, the total surface current density is

$$\mathbf{J}_{1T} = \mathbf{J}_{1s} + \mathbf{J}_{1sm},$$

where $\mathbf{J}_{1sm}$ is the surface magnetization current density. According to Equation (8.12), $\mathbf{J}_{1sm}$ is given by $\mathbf{J}_{1sm} = \mathbf{M}_1 \times \hat{\mathbf{a}}_n$, where $\mathbf{M}_1$ is the magnetization in the left-hand region. Using Equations (8.18) and (8.22), we find that $\mathbf{M}_1 = [(\mu_1 - \mu_o)B_o]/(\mu_o\mu_1)\hat{\mathbf{a}}_z$, so

$$\mathbf{J}_{1sm} = \frac{(\mu_1 - \mu_o)B_o}{\mu_o\mu_1}\hat{\mathbf{a}}_\phi.$$

Consequently, we can write $\mathbf{J}_{1T}$ as

$$\mathbf{J}_{1T} = \frac{1}{\mu_1}B_o\hat{\mathbf{a}}_\phi + \frac{(\mu_1 - \mu_o)B_o}{\mu_o\mu_1}\hat{\mathbf{a}}_\phi = \frac{B_o}{\mu_o}\hat{\mathbf{a}}_\phi \qquad \text{[A/m]}.$$

A similar sequence of steps shows that the total surface current density on the right-hand side of the solenoid $\mathbf{J}_{2T}$ has the same value. Hence, the *total* current density along the solenoid is uniform, which produces a uniform B-field throughout the interior of the solenoid.

The two previous examples share the characteristic that the B-field streamlines are simple. There are, of course, many situations where this does not occur, which means that we must often depend upon graphical or numerical solutions for most of the problems encountered in engineering practice. The numerical methods used to solve magnetostatic problems are similar to those used for electrostatic problems and, for that reason, will not be discussed separately here. The interested reader can find discussions of these techniques in several of the references mentioned at the text.

8-4-2 MAGNETIC FLUX PLOTS

There are many situations where quick, approximate solutions of magnetostatic problems are needed in the preliminary stages of a design. In these situations, graphical solutions are attractive, because they are reasonably accurate and very easy to obtain. The curvilinear squares technique provides accurate estimates of two-dimensional geometries.

To see how the curvilinear squares technique can be applied to magnetostatic fields, let us first compare the scalar potential equations for both cases. For electrostatic fields, we have

$$\mathbf{E} = -\nabla V,$$

and

$$\nabla^2 V = 0,$$

whereas for the magnetostatic case, we have

$$\mathbf{H} = -\nabla V_m$$

and

$$\nabla^2 V_m = 0.$$

These equations are the same when V is replaced by V_m and $\mathbf{E}$ is replaced by $\mathbf{H}$. As a result, the rules for drawing a correct magnetic field flux plot of H-field streamlines and magnetic equipotential surfaces are the same as for the electrostatic case.

Figure 8-15 shows a section of a magnetostatic flux plot. Here, each tube "carries"

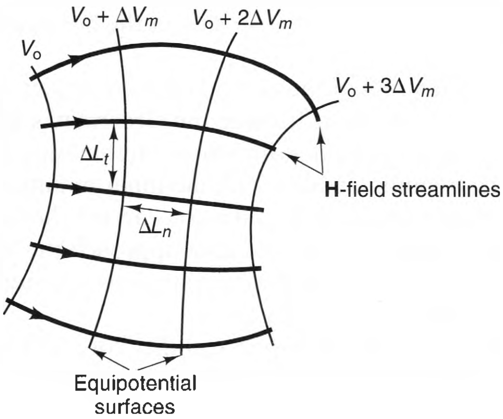


Figure 8-15 A magnetostatic, curvilinear squares flux plot.

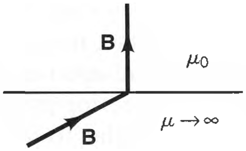


Figure 8-16 The behavior of **B** at the interface between high and low-permeability media.

the same flux $\Delta\Phi$, and the magnetic potential difference ΔV_m between each equipotential surface is also the same from cell to cell. Thus, we can write

$$\Delta\Phi = \int_{\Delta L_t} \mu H d\ell \approx \mu H \Delta L_t \qquad [\mathrm{Wb/m}] \tag{8.46}$$

and

$$\Delta V_m = \int_{\Delta L_n} H d\ell \approx H \Delta L_n \qquad [\mathrm{A}], \tag{8.47}$$

where ΔL_t and ΔL_n are the spacings between adjacent streamlines and equipotential surfaces, respectively.

From Equations (8.46) and (8.47), it follows that

$$\frac{\Delta\Phi}{\Delta V_m} \approx \mu \frac{\Delta L_t}{\Delta L_n}, \tag{8.48}$$

where $\Delta L_t/\Delta L_n$ is the cell aspect ratio. As in the electrostatic case, this aspect ratio is usually chosen to be unity, so the cells are square.

The rules for drawing the flux and potential lines of magnetostatic and electrostatic flux plots are the same in homogeneous regions, but their boundary conditions are different. In electrostatic flux plots, conductors act as equipotential surfaces that are not penetrated by E-field streamilines. In magnetostatic flux plots, B-field lines tend to concentrate in high-permeability materials. Moreover, the magnetic potential is constant on the outside of high-permeability materials but not inside. This is illustrated in Figure 8-16, which shows the interface between a high-permeability material and air (free space). As we found in Example 8-2, the B-field streamlines enter the low-permeability region at right angles to the surface, except when **B** is nearly parallel to the interface in the high-permeability material. This makes the interface look like a constant magnetic potential surface on the air side of the boundary, but not necessarily on the material side.

Figure 8-17 shows a flux plot of the B-field near an air gap in a high-permeability core. The plot was obtained by first drawing the streamlines and constant-potential surfaces in the air region and then extending these streamlines into the core. In theory, the angle that each streamline makes with the inner boundary should be adjusted such that the normal component of **B** is continuous across the interface. In practice, however, most of the streamlines inside the core approach the air gap at nearly right angles. As can be seen in the figure, B-field streamlines are guided by a high-permeability core and tend to fringe near an air gap.

Figure 8-18 shows a bend in a high-permeability core. The field outside the bar is almost nonexistent, since there is no gap in the core, so the streamlines follow the edges of the core.

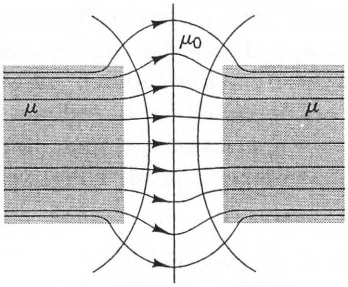


Figure 8-17 B-field streamlines near the gap of a high-permeability core.

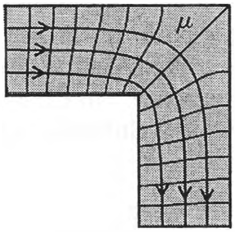


Figure 8-18 B-field streamlines near a bend in a high-permeability core.

As can be seen from this flux plot, the streamlines are uniformly spaced away from the corner, but they tend to bunch up near the inside of the corner, increasing the field strength there. The tends to increase the eddy curent losses in the region around the bend. The effect can be minimized by rounding the inside corner; the more gradual the bend, the less the B-field lines are bunched.

8-5 Permanent Magnets and Magnetic Recording

The hysteresis (i.e., memory) of magnetic materials is usually a disadvantage in high-flux networks, such as magnetic circuits. This is because energy is lost in time-varying circuits as the magnetizing field works against the residual field. On the other hand, there are many applications where this memory is very desirable. Two such applications are permanent magnets and magnetic memories.

The techniques for determining the fields generated by permanent magnets are somewhat different from those discussed in the previous sections, since **B** and **H** are not linearly proportional to each other inside a material that has residual magnetization **M**. To illustrate this, consider the permanent magnet shown in Figure 8-19.

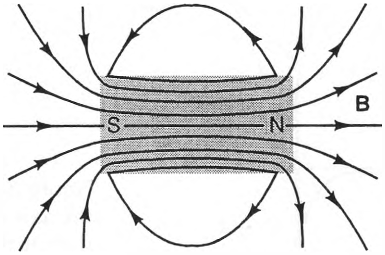


Figure 8-19 Cross section of a uniformly magnetized cylindrical rod.

Here, a cylindrical rod has been magnetized so that **M** is uniform throughout its interior; $\mathbf{M} = M_o \hat{\mathbf{a}}_z$, where $\hat{\mathbf{a}}_z$ is directed along the axis of the cylinder. Since there is no free current present anywhere in this geometry, it is not obvious how to proceed to find the B-field using any of the analytical techniques that we used for linear materials. However, since **M** is known, we can use Equations (8.11) and (8.12) to find the magnetization currents that are present. Substituting, we have

$$\mathbf{J}_m = \nabla \times (M_o \hat{\mathbf{a}}_z) = 0$$

throughout the interior of the cylinder and

$$\mathbf{J}_{sm} = (M_o \hat{\mathbf{a}}_z) \times \hat{\mathbf{a}}_\rho = M_o \hat{\mathbf{a}}_\phi \qquad [\mathrm{A/m}]$$

on the surface of the cylinder. From these expressions, we see that the uniform magnetization **M** produces a uniform surface current that circulates around the cylinder. Since we now know the current distribution, we can treat these currents as if they were simply suspended in free space. In Chapter 7 we calculated the field along the center of a finite-length solenoid. Substituting the preceding value of $\mathbf{J}_{sm}$ into Equation (7.38), we obtain

$$B_z = \frac{\mu_o M_o}{2}\left[\frac{L - 2z}{\sqrt{(L - 2z)^2 + d^2}} + \frac{L + 2z}{\sqrt{(L + 2z)^2 + d^2}}\right]. \tag{8.49}$$

This expression is valid at all points along the z-axis, both inside and outside the cylinder, and is plotted as the solid curve in Figure 8.20 for the case where $L/d = 5$.

To see what the relationship between **B** and **H** is inside and outside the cylinder, we can use

$$\mathbf{B} = \mu_o(\mathbf{H} + \mathbf{M}).$$

Solving for **H**, we obtain

$$\mathbf{H} = \frac{\mathbf{B}}{\mu_o} - \mathbf{M}.$$

Since **M** is uniform inside the cylinder and zero outside, **H** is discontinuous at the cylinder end caps. This is evident in Figure 8-20, where H is dark. We also notice from these plots that B_z and H_z are proportional to each other outside the cylinder, but definitely not inside. This is because of the residual magnetization M_o of the material.

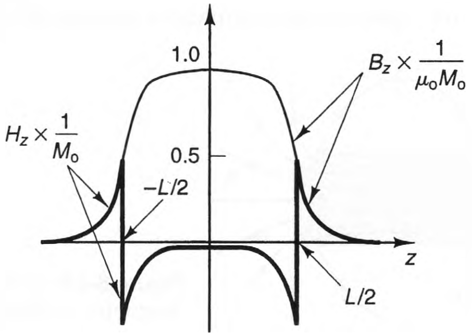


Figure 8-20 Plot of B_z and H_z vs. z along the axis of the permanently magnetized rod for the case $L/d = 5$.

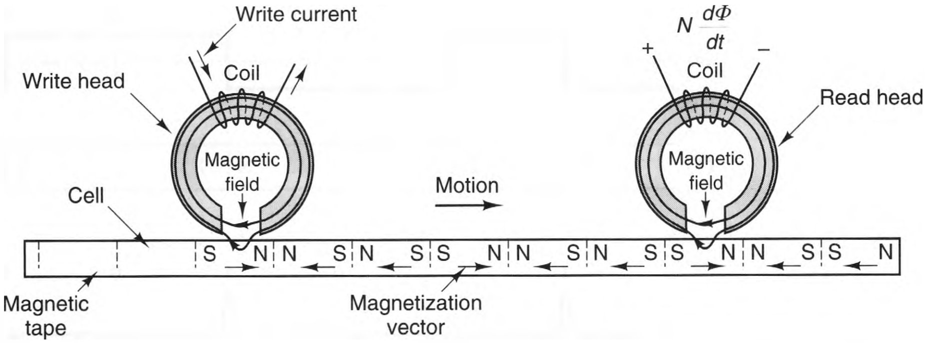


Figure 8-21 A magnetic tape moving past recording and playback heads.

One of the most important uses of hysteresis is in magnetic recording or memory applications. Here, information is stored on a magnetic tape or disc by magnetizing it with a magnetizing field. The information can later be retrieved by sensing the magnetization pattern. Figure 8-21 shows a simplified binary (i.e. digital) recording configuration. In the figure, a magnetic tape or disc moves past both a write head and a read head. In their simplest forms, these heads consist of toroids of high-permeability material with an air gap. In many systems, the same head is used for both the read and write operations.

Information is encoded in the magnetic medium by passing a current in the write head coil as the medium moves by. A B-field is induced that fringes outside the gap, which magnetizes the material as it passes by. The medium is typically divided into racks that run along the direction of motion, and each track is divided into cells. Each cell stores one bit of data. The magnetization direction of each cell is determined by the polarity of the write current while the cell passes by the write head.

As the magnetized medium passes by the gap of the read head, B-fields are induced in the read-head gap and core whose direction is determined by the magnetization vector of the cell nearest to the gap. As the cell polarities change from cell to cell, a time varying flux in the core induces a voltage $N(d\Phi/dt)$ across the terminals of the read coil. This is a result of Faraday's law, which we discussed briefly in Chapter 3 and will discuss more fully in Chapter 9. The voltage waveform in the read head is the time derivative of the magnetization pattern passing by the head. Figure 8-22 shows sample bit sequences of the write current, the magnetization, and the read voltage waveforms.

The quest for achieving higher and higher data densities on magnetic tapes and disks presents several interesting challenges for designers. First, the heads must be constructed so that the gaps are very small. Not only this, but the B-fields must drop off quickly away from the gaps, to ensure that the field directed towards one cell does not also magnetize an adjacent cell. The characteristics of the magnetic recording material are also important. For instance, the magnetic domains must be very small and not vulnerable to "bleed-through" from the magnetization of adjacent domains. At present, the maximum achievable memory density on a disk is approximately 10^8 bits per square centimeter.

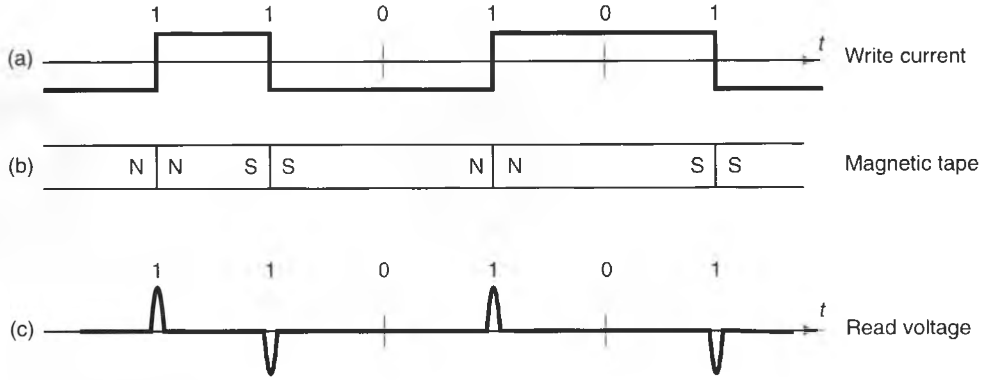


Figure 8-22 Magnetic tape recording. a) Write head current vs. time. b) Tape magnetization. c) Read head voltage vs. time.

8-6 Magnetic Circuits

We have already seen that high-permeability materials tend to concentrate B-fields within themselves, just as conductors do with currents, making high-permeability materials ideal for guiding large amounts of magnetic flux around well-defined paths. Networks that accomplish this are called ***magnetic circuits***. As their name implies, magnetic circuits are in many respects analogs of electric circuits. Magnetic circuits are used in a variety of applications, such as transformers, motors, loudspeakers, and relays.

Figure 8-23a shows N turns of wire, wrapped around a core that has permeability μ_c and cross-sectional area S. In addition, there is an air gap in the core. If the core permeability μ_c is large and the gap width L_g is small, the B-field will be nearly uniform throughout the cross section of the core, and fringing in the gap will be negligible. This means that the magnetic flux Φ is constant at all points around the loop (including the gap) and can be obtained by integrating over the cross section of the loop. We obtain

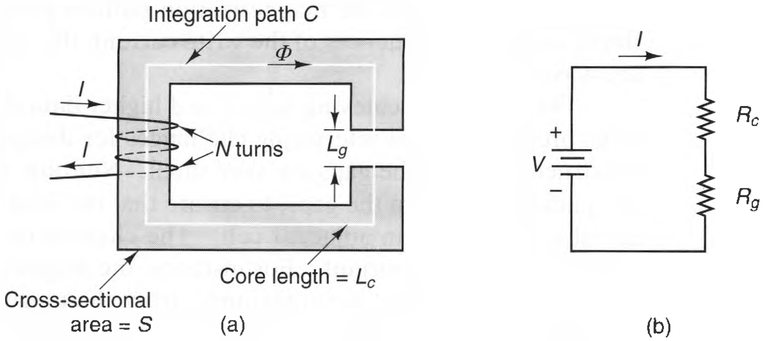


Figure 8-23 A magnetic circuit. a) Geometry. b) Equivalent circuit.

$$\Phi = \int_S \mathbf{B} \cdot d\mathbf{s} = BS,$$

where B is the magnetic flux density inside the core and gap. We can find the relationship between Φ and the current I by evaluating Ampère's law around the mean path through the center of the core and the gap. Integrating clockwise and noting that the current I links this path N times in a right-handed sense, we obtain

$$\oint_C \mathbf{H} \cdot d\boldsymbol{\ell} = \oint_C \frac{1}{\mu} \mathbf{B} \cdot d\boldsymbol{\ell} = NI,$$

where μ equals μ_c in the core and μ_0 in the gap. Since the streamlines are parallel to the path C, we can replace $\mathbf{B} \cdot d\boldsymbol{\ell}$ with $B\,d\ell$. Also, B is constant throughout the loop, so we can take it outside the integral, which yields

$$B \oint_C \frac{1}{\mu}\, d\ell = B\left[\frac{L_c}{\mu_c} + \frac{L_g}{\mu_0}\right] = NI,$$

where L_c and L_g are the lengths of the core and gap, respectively. Using $B = \Phi/S$ and solving for Φ, we obtain

$$\Phi = \frac{NI}{\dfrac{L_c}{\mu_c S} + \dfrac{L_g}{\mu_0 S}}.$$

This can be written in the form

$$\Phi = \frac{V_m}{\Re_c + \Re_g}, \qquad (8.50)$$

where

$$V_m = NI \qquad [\text{A}] \qquad (8.51)$$

is the ***magnetomotive force*** (mmf) of the circuit windings and $\Re_c$ and $\Re_g$ are the ***reluctances*** of the core and gap, respectively:

$$\Re_c = \frac{L_c}{\mu_c S} \qquad [\text{A/Wb or H}^{-1}] \qquad (8.52)$$

$$\Re_g = \frac{L_g}{\mu_0 S} \qquad [\text{A/Wb or H}^{-1}]. \qquad (8.53)$$

Figure 8-23b shows an electrical analog of this magnetic circuit. The magnetomotive force V_m is represented by the voltage V, the flux Φ is represented by the current I, and the reluctances $\Re_c$ and $\Re_g$ are represented by the series resistances R_c and R_g, respectively. The analogous quantities are summarized as follows:

Magnetic Circuits	**Electric Circuits**
magnetomotive force (mmf), V_m [A]	electromotive force (emf), V [V]
magnetic flux, Φ [Wb] or [$\mathrm{T \cdot m^2}$]	electric current, I [A]
reluctance, $\mathfrak{R}$ [$\mathrm{H^{-1}}$]	Resistance, R [Ω]

This analog can be used whenever the magnetic flux is confined to flow within a well-defined path. All that is needed to find the equivalent electric circuit is to identify the sources of magnetomotive force and to determine the reluctances of the various flux paths. The sum of all the magnetic voltage drops around each path is zero, and the magnetic voltage drop across each reluctance satisfies the magnetic equivalent of Ohm's law:

$$V_m = \mathfrak{R}\Phi. \tag{8.54}$$

Example 8-6

For the magnetic circuit shown in Figure 8-24a, find the flux in the gap region, assuming that fringing and leakage can be neglected. Assume also that the permeability of the ferromagnetic μ_c core is constant and that the currents I_1 and I_2 are steady.

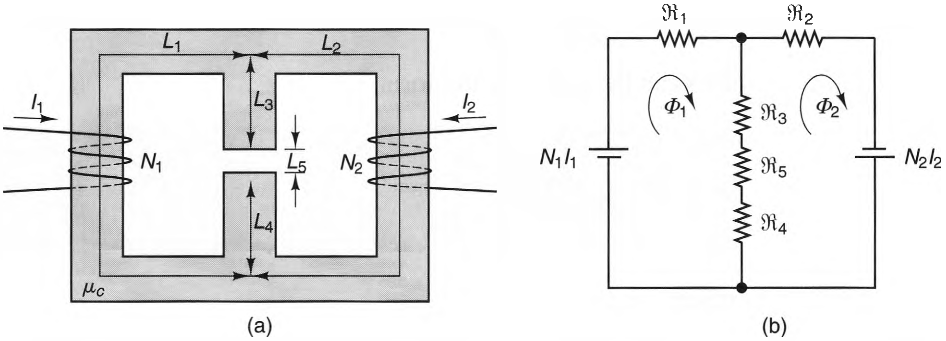


Figure 8-24 A magnetic circuit. a) Geometry. b) Equivalent circuit.

Solution:

The sections along L_1, L_2, L_3, L_4, and L_5 are each simple circuit elements, since they have constant cross sectional areas.[6] On the basis of these average path lengths, we have

$$\mathfrak{R}_1 = \frac{L_1}{\mu_c S}, \quad \mathfrak{R}_3 = \frac{L_3}{\mu_c S}, \quad \text{and} \quad \mathfrak{R}_5 = \frac{L_5}{\mu_o S}, \text{ when } S \text{ is the cross sectional area of the core.}$$

Also, since $L_2 = L_1$ and $L_3 = L_4$,

$$\mathfrak{R}_1 = \mathfrak{R}_2 \text{ and } \mathfrak{R}_3 = \mathfrak{R}_4.$$

[6] This is obviously not true near the corners of the two side regions, but if the volume of these regions is small, their effect on the total reluctance is small.

The equivalent circuit for this configuration is shown in Figure 8-24b. Using the right-hand rule, we see that the voltage sources in the equivalent circuit have values $V_1 = N_1 I_1$ and $V_2 = -N_2 I_2$, respectively.

Using standard loop analysis, we obtain the following loop equations:

$$\Phi_1 \Re_1 + (\Phi_1 - \Phi_2)(\Re_3 + \Re_5 + \Re_4) = N_1 I_1$$

$$\Phi_2 \Re_2 + (\Phi_2 - \Phi_1)(\Re_3 + \Re_5 + \Re_4) = N_2 I_2.$$

Solving these equations, we obtain

$$\Phi_1 = \frac{N_1 I_1 (\Re_2 + \Re_3 + \Re_5 + \Re_4) + N_2 I_2 (\Re_3 + \Re_5 + \Re_4)}{(\Re_1 + \Re_2)(\Re_3 + \Re_5 + \Re_4) + \Re_1 \Re_2}$$

$$\Phi_2 = \frac{N_1 I_1 (\Re_3 + \Re_5 + \Re_4) + N_2 I_2 (\Re_1 + \Re_3 + \Re_5 + \Re_4)}{(\Re_1 + \Re_2)(\Re_3 + \Re_5 + \Re_4) + \Re_1 \Re_2},$$

and the flux in the gap region is

$$\Phi_{\text{gap}} = \Phi_1 - \Phi_2 = \frac{N_1 I_1 - N_2 I_2}{\Re_1 + 4\Re_3 + 2\Re_5},$$

where we have used $\Re_2 = \Re_1$ and $\Re_4 = \Re_3$.

When nonlinear ferromagnetic materials are present, it is necessary to use the magnetization characteristics of the material to determine the actual value of the permeability. This procedure is demonstrated in the following example.

Example 8-7

Find the flux Φ in the rectangular circuit shown in Figure 8-23a if the ferromagnetic core has a nonlinear permeability with a magnetization characteristic given by the solid curve in Figure 8-25. Assume that $NI = 1{,}600$ [A] and the circuit dimensions are $L_c = 63$ [cm], $S = 1$ [cm^2] and $L_g = 1$ [mm].

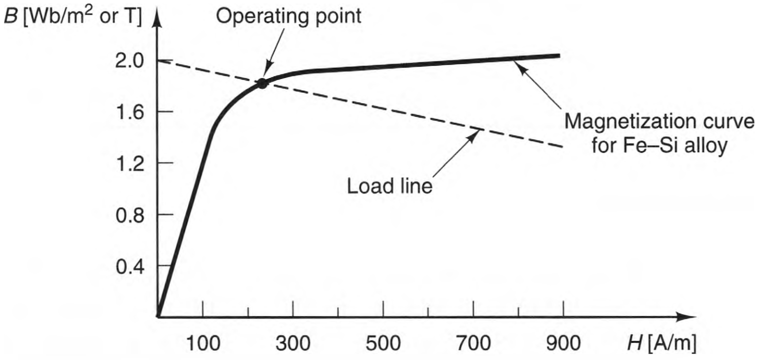


Figure 8-25 The magnetization curve for Fe–Se alloy[7] and the load line for the magnetic circuit of Figure 8-23a.

[7] See D.C. Heck, *Magnetic Materials and Their Applications* (Crane, Russak & Co., 1974), pp. 365.

Solution:

The reluctances of the core and gap are given by

$$\Re_c = \frac{L_c}{\mu_c S}$$

$$\Re_g = \frac{L_g}{\mu_0 S},$$

where the value of μ_c is, for the moment, unknown. The loop equation around the circuit is thus

$$\Phi\frac{L_c}{\mu_c S} + \Phi\frac{L_g}{\mu_0 S} = NI = 1{,}600\ [\mathrm{A}].$$

Using $\Phi \approx B_c S$, where B_c is the magnetic flux density inside the core, we can write this equation in terms of B_c:

$$\frac{B_c}{\mu_c}L_c + \frac{B_c}{\mu_0}L_g = 1{,}600\ [\mathrm{A}].$$

Also, since the B-field is continuous across the gap, $H_c = B_c/\mu_c$. Substituting this into the foregoing equation yields

$$H_c L_c + \frac{B_c}{\mu_0}L_g = 1{,}600\ [\mathrm{A}],$$

which is the equation of a straight line in the variables B_c and H_c. For the values specified in this problem, this equation becomes

$$H_c + 1{,}263B_c = 2{,}540,$$

which is shown as the dotted curve in Figure 8-25 and can be considered as the *load line*. Along with the magnetization curve of the material, this load line determines the operating point of the circuit. The correct values of B_c and H_c lie at the intersection of the load line with the magnetization curve, where we find that

$$B_c = 1.83\ [\mathrm{T}]\ \text{or}\ [\mathrm{Wb/m^2}]$$

and

$$H_c = 225\ [\mathrm{A/m}].$$

Thus, we finally obtain

$$\Phi = B_c S = 1.8\ [\mathrm{Wb/m^2}] \times 10^{-4}\ [\mathrm{m^2}] = 1.8 \times 10^{-4}\ [\mathrm{Wb}].$$

8-7 Summation

In this chapter, we have discussed two interactions between magnetic fields and currents in material media. In the Hall effect, voltages are induced across material samples whenever a magnetic field is impressed perpendicular to the flow of current. The voltages produced are generally small, but are useful in a variety of applications.

The interaction between an applied magnetic field and the spin and orbital currents is very pronounced in certain materials. Extremely large magnetic fields can be produced via this interaction with relatively small currents. It is for this reason that most electrical machines generate forces using magnetic fields, rather than electric fields. We will talk more about these forces in the next chapter.

PROBLEMS

8-1 A homogeneous strip of width w = 0.5 [cm] and thickness t = .01 [cm] carries a 200 [mA] current that flows lengthwise through the strip. A 1.5×10^{-4} [T] B-field is directed perpendicular to width of the strip, resulting in a Hall voltage across its width of 3 [μV]. Find the net free-charge density of mobile charge carriers in the sample.

8-2 Find the magnetic field **B** required to produce a Hall voltage of 14 [μV] across a copper strip that is 0.5 [cm] wide and 0.1 [mm] thick, and that carries a current of 700 [mA]. Assume that **B** is directed perpendicular to the width of the strip that the mobile electron density is 8.4×10^{28} [m^{-3}]. $\times 1.6 \times 10^{-19}$

8-3 Prove that

$$\int_{\text{Vol.}} \nabla \times \mathbf{F} dv = -\oint_S \mathbf{F} \times d\mathbf{s},$$

where S is the surface that bounds the volume of integration. (*Hint*: Start by applying the divergence theorem to the product $\mathbf{F} \times \mathbf{C}$, where **C** is a constant vector.)

8-4 If the relative permeability of cobalt is μ_r = 600 and the density of atoms is 9.02×10^{22} [cm^{-3}], calculate the average dipole moment per atom when a sample is placed in a uniform magnetizing field of value B_m = 0.1 [T].

8-5 A wire loop of radius 1 [cm] that carries a current of 10 [mA] is placed in a homogeneous medium that has a relative permeability of μ_r = 100. Calculate the B-field at a) the center of the loop and b) 10 [cm] along the axis of the loop.

8-6 The B-field in the air just above an iron plate has a magnitude of 0.1 [T] and makes an angle of 1° with respect to the surface normal. Calculate the magnitude of **B** just inside the plate and the angle it makes with respect to the surface normal.

8-7 Figure P8-7 shows a magnetic slab with μ_r = 50. A thin conducting film (with μ_r = 1) lies on top of the slab and carries a surface current of 1.0 [A/mm], directed out of the page. If $|\mathbf{B}_1|$ = 0.01 [T] and θ_1 = 10°, find $|\mathbf{B}_2|$ and θ_2.

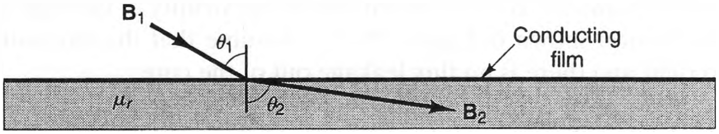


Figure P8-7

8-8 In Figure P8-8, magnetic flux enters the first interface of a three-layer geometry at an angle θ_i. If all three media are nonconducting and have permeabilities μ_1, μ_2, and μ_3, respectively,
(a) show that the angle θ_o is independent of the value of μ_2.
(b) show that $\theta_o = \theta_i$ when $\mu_1 = \mu_3$.

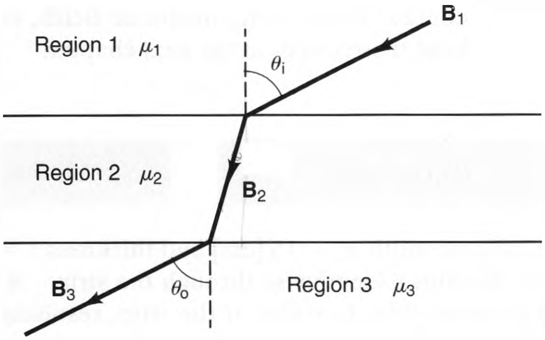


Figure P8-8

8-9 A long solenoid is constructed by wrapping 200 turns/cm of wire around the outer surface of a hollow, circular cylinder that has a relative permeability of μ_r, an inner radius a, and an outer radius b. If the wire carries a current I, calculate $|\mathbf{B}|$ in the solid and hollow portions of the cylinder. Also, calculate the total flux $\int_S \mathbf{B} \cdot d\mathbf{s}$ that passes through the solenoid.

8-10 Draw a flux plot that shows the B-field streamlines in the high-permability "T" shown in Figure P8-10. Assume that flux enters from the bottom and splits equally between the right and left halves.

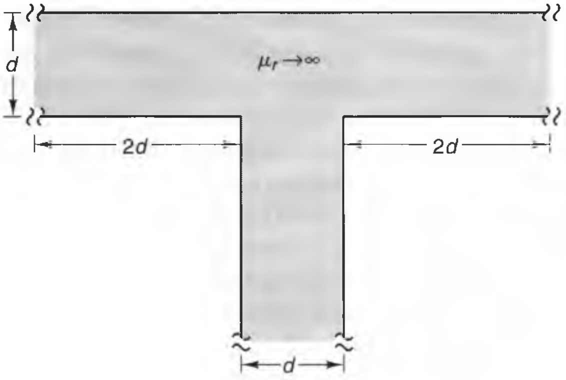


Figure P8-10

8-11 Draw a flux plot that shows the B-field streamlines in the vicinity of the high-permeability core restriction shown in Figure P8-11. Assume that the streamlines flow from left to right and there is no flux leakage out of the core.

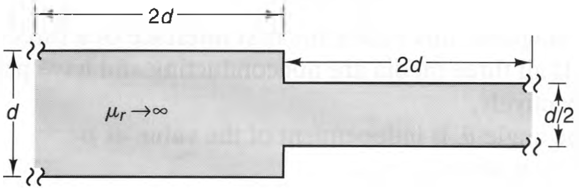


Figure P8-11

8-12 Write a numerical program (using languages such as FORTRAN or C++, or mathematical software programs such as Matlab™ or Mathcad™) that solves Laplace's equation for the magnetic scalar potential inside the restricted core shown in Figure P8-11. Assume that the left- and right-hand edges are constant-potential surfaces. Since there is zero leakage out of the core, the constant-V_m surfaces are perpendicular to the top and bottom walls. This means that the boundary condition along these walls is $\partial V_m / \partial n = 0$, where n is the direction normal to the wall. Plot several constant-V_m surfaces and B-field streamlines.

8-13 Draw a flux plot that shows the B-field streamlines in the air gap shown in Figure P8-13. Assume that the permeability of the core is large and the streamlines flow from left to right.

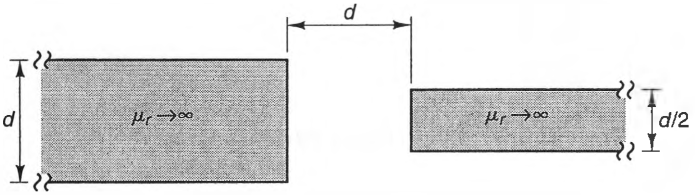


Figure P8-13

8-14 The two-dimensional bend shown in Figure 8-18 has a width of 1 [m]. Estimate the reluctance/meter between the first and last constant magnetic potential lines. Assume that the material has relative permeability μ_r. Compare this value with the reluctance/meter of an unbent section with the same cross section and mean length? (*Hint*: Reluctance can be calculated from B-field flux plots, analogously to calculating capacitance from E-field flux plots.)

8-15 A long cylinder has been uniformly magnetized so that it forms a permanent magnet. If the B-field at the ends of the cylinder is 12 [T], calculate the magnetization **M** inside the cylinder and the magnetization surface current density $\mathbf{J}_{sm}$ that flows around the outer surface of the cylinder. (*Hint*: Remember that **B** at the ends of a long solenoid equals roughly half its value at its center.)

8-16 Calculate the flux passing through the left- and right-hand gaps in the magnetic circuit shown in Figure P8-16. Assume that the core has a relative permeability of 1,000 and a square cross section throughout. Neglect any fringing.

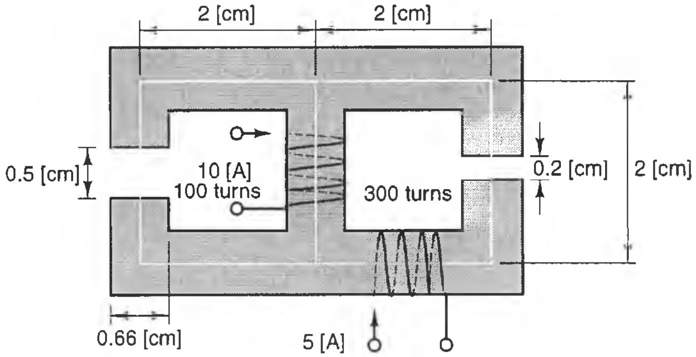


Figure P8-16

8-17 Plot the flux Φ passing through the magnetic circuit shown in Figure P8-17 as a function of x for $0 < x < 1$ [cm]. Assume that the core and bar have relative permeabilities of 1,000 and 500, respectively, and the same square cross sections. Neglect all fringing.

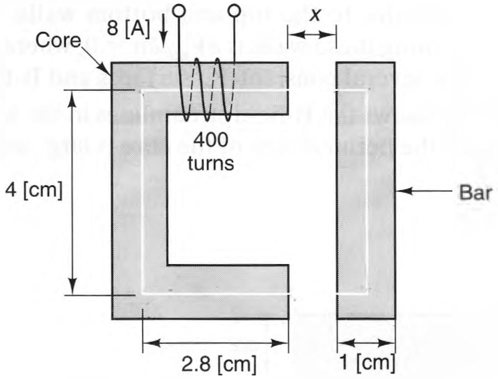


Figure P8-17

8-18 Calculate the flux passing through the magnetic circuit in problem 8-17 when $x = 0.1$ [cm], the bar has a permeability of 500, and the core is Fe–Si alloy (*Hint*: Use the Fe–Si alloy magnetization curve shown in Figure 8-25.)

9

Magnetic Inductance, Energy, and Forces

9-1 Introduction

Throughout Chapters 4–8, electric and magnetic fields were discussed as if they were independent entities. They are indeed independent at low frequencies whenever the time-derivative terms in Maxwell's equations are negligible. In this chapter, we will start a discussion of the more general case where electric and magnetic fields not only are present simultaneously, but also affect each other.

Faraday's law of induction will be the starting point for our discussion. From this law, we will show how electric fields and voltages are generated by either time-varying magnetic fields or the movement of material media through a magnetic field. Applications presented during the discussion will include transformers and generators.

Just as the charges in an electrostatic system interact through their mutual capacitances, we will find that systems of currents also interact through their self and mutual inductances. We will first define the inductance of an element in terms of the magnetic field it generates and will later show how inductance is related to the energy it stores in its magnetic field. We will also derive a number of formulas for the inductances of geometries commonly encountered in engineering systems.

We will conclude this chapter with a discussion of the forces exerted by current carrying circuits on other circuits and magnetic materials. These forces are important, because most practical transducers that convert electrical energy into a mechanical output make use of magnetic forces.

9-2 Faraday's Law of Induction

Whereas only static charge distributions can produce a static E-field, the situation is more complicated when time-varying source distributions are present. Michael Faraday conducted a classic experiment in 1831 that showed that E-fields are produced by time-varying B-fields. A schematic of the experimental setup he used is shown in Figure 9-1. Here, a time-varying current $i(t)$ is established in the primary conducting loop when the switch is closed at $t = 0$. The secondary conducting loop is open circuited and contains no lumped voltage source. Faraday found that a time-varying voltage is generated between the open-circuit terminals of the secondary loop under any of the following circumstances:

1. The current in the primary loop is time varying.
2. Either loop is moving with respect to the other, and a steady current is flowing in the primary loop.
3. A permanent magnet is moved near the secondary loop.

This voltage is an indication that an E-field is induced in the open-circuit gap of the secondary loop as a result of a time-varying B-field in the vicinity of the circuit. Faraday deduced that these effects are accounted for by the following integral relationship between **E** and **B**:

$$\oint_C \mathbf{E} \cdot d\boldsymbol{\ell} = -\frac{d}{dt}\int_S \mathbf{B} \cdot \mathbf{ds}. \tag{9.1}$$

Here, S is any open surface that is bounded in a right-handed sense by a closed contour C. This expression is called ***Faraday's law of induction*** (or simply, ***Faraday's law***).

In words, Faraday's law states that there is a net voltage, or electromotive force (emf), around a closed path whenever a time-varying magnetic flux passes through (or

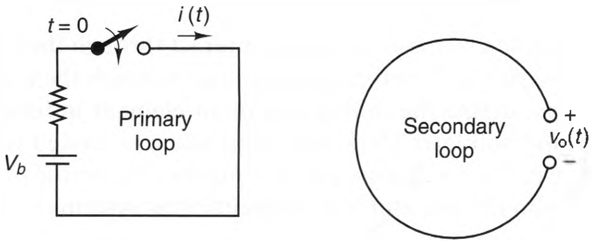


Figure 9-1 A simple network that demonstrates Faraday's law of induction.

links) the path. The sign of the emf is such that it tends to produce currents whose B-fields oppose the time-varying flux linkage. This latter phenomenon is called ***Lenz's law*** and is a convenient method of predicting the direction of the induced currents.

A point form of Faraday's law can be derived by applying Equation (9.1) to a vanishingly small contour. If the contour is stationary, S is also stationary, which means that the order of integration and differentiation can be interchanged in the surface integral, yielding

$$\oint_C \mathbf{E} \cdot \mathbf{d\ell} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot \mathbf{ds}.$$

Next, we can use Stokes's theorem to write the contour integral as a surface integral,

$$\oint_C \mathbf{E} \cdot \mathbf{d\ell} = \int_S \nabla \times \mathbf{E} \cdot \mathbf{ds} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot \mathbf{ds}.$$

When C is made vanishingly small, S becomes a single point, which means that the integrands of the surface integrals must be equal at any point. Thus,

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}. \qquad (9.2)$$

Even though Faraday discovered this law by observing voltages produced in circuits, the ramifications of Faraday's law go far beyond circuits. In fact, Faraday's law is one of the two fundamental laws that specify the relationship between electric and magnetic fields. (The other is Maxwell's curl-H formula for time-varying fields.)

In the discussion that follows, we will examine two important special cases described by Faraday's law. The first case is voltages induced in stationary circuits that are subjected to time-varying magnetic fields. These voltages are called ***transformer emf's*** and are the physical mechanism involved in the operation of transformers. The second case is voltages that are produced in moving circuits. These voltages are called ***motional emf's*** and are encountered in rotating machinery.

Example 9-1

Figure 9-2 shows a high-permeability core that carries a uniform, time-varying B-field, given by $\mathbf{B} = B_o \cos\omega t\, \hat{\mathbf{a}}_z$.

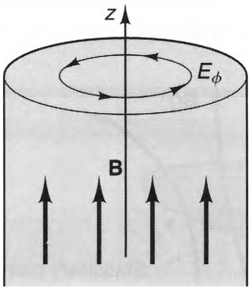


Figure 9-2 A magnetic core carrying a uniform, time-varying B-field.

Calculate the E-field generated inside the core.

Solution:

Using Equation (9.2), we can write

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} = \omega B_o \sin \omega t \, \hat{\mathbf{a}}_z.$$

Expanding the curl operator in cylindrical coordinates (Equation (2.115)), we have

$$\left[\frac{1}{\rho}\frac{\partial E_z}{\partial \phi} - \frac{\partial E_\phi}{\partial z}\right]\hat{\mathbf{a}}_\rho + \left[\frac{\partial E_\rho}{\partial z} - \frac{\partial E_z}{\partial \rho}\right]\hat{\mathbf{a}}_\phi + \frac{1}{\rho}\left[\frac{\partial}{\partial \rho}(\rho E_\phi) - \frac{\partial E_\rho}{\partial \phi}\right]\hat{\mathbf{a}}_z = \omega B_o \sin \omega t \, \hat{\mathbf{a}}_z.$$

Since the right-hand side of this expression has only a z-component, the ρ and ϕ components of $\nabla \times \mathbf{E}$ are zero. Also, we can assume from the cylindrical symmetry of this problem that $\mathbf{E}$ is only a function of ρ. Hence, the preceding vector equation reduces to the scalar equation

$$\frac{1}{\rho}\left[\frac{\partial}{\partial \rho}(\rho E_\phi)\right] = \omega B_o \sin \omega t.$$

The particular solution of this differential equation is

$$E_\phi = \frac{1}{2}\omega \rho B_o \sin \omega t.$$

Thus, a time-varying field in a magnetic core induces an E-field that circulates around the core and is strongest along the outer perimeter. This is true regardless of the cross-sectional shape of the core, although the field expression for a circular cross section is by far the simplest.

9-2-1 TIME-VARYING FIELDS IN STATIONARY CIRCUITS

Figure 9-3 shows a stationary path C in the presence of a time-varying magnetic field. Also shown is the stationary surface S that is bounded by C in a right-handed sense. Since S is time invariant, the order of the differentiation and integration can be reversed in the integral form of Faraday's law, so we can write

$$\oint_C \mathbf{E} \cdot d\boldsymbol{\ell} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot \mathbf{ds}.$$

This equation states that $\mathbf{E}$ is nonconservative whenever a time-varying magnetic field is present, producing a net voltage called ***transformer* emf** around the path. As we saw in Chapter 5, emf's are voltage sources that are capable of driving currents around circuits. What makes transformer emf different from the emf generated by a battery is

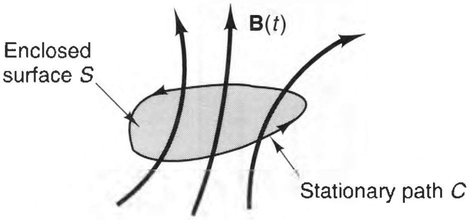


Figure 9-3 A time-varying B-field linking a stationary path.

that a battery's emf is localized to within the battery itself (i.e., the chemical reaction), whereas transformer emf is generated throughout the entire circuit.

Example 9-2

Figure 9-4 shows an open circuited, circular loop of conducting wire that is subjected to a magnetic field $\mathbf{B} = B_0 \sin \omega t\, \hat{\mathbf{a}}_z$. Find the gap voltage V_g if the radius of the loop is a.

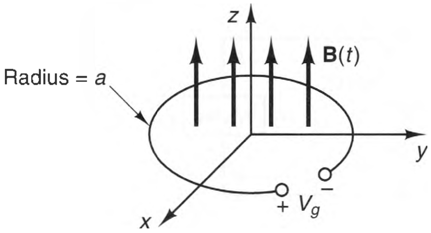


Figure 9-4 A uniform, time-varying B-field linking a circular loop with a gap.

Solution:

For a counterclockwise path around the loop, Faraday's law yields

$$\oint_C \mathbf{E} \cdot \mathbf{d\ell} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot \mathbf{ds} = -\omega \pi a^2 B_o \cos \omega t.$$

This result tells us the net induced voltage around the circuit, but it does not tell us where this voltage will appear in the circuit. For that, we need to consider the circuit parameters of the elements that make up the circuit. In this case, no current can flow in the conductor, since it is open circuited. Hence, the entire voltage appears across the gap, and we have

$$\oint_C \mathbf{E} \cdot \mathbf{d\ell} = V_g = -\omega \pi a^2 B_0 \cos \omega t.$$

If a resistor were placed across the gap, a current would flow in the circuit. As tempting as it is to calculate this current, we are not quite ready to do so, since we would have to account for the B-field that the current would produce. This effect is called self-inductance and is discussed later in the chapter.

Time-varying magnetic fields can cause problems when one uses test equipment to measure voltages. These problems occur when time-varying magnetic flux links the wiring path of the test equipment, so the voltage that is displayed may not be the voltage that the user intended to measure. The following example demonstrates this.

Example 9-3

A uniform magnetic field of value $\mathbf{B} = 2 \sin \omega t\, \hat{\mathbf{a}}_z$ [Wb/m^2] links the circuit shown in Figure 9-5, which consists of two resistors. Here, $\hat{\mathbf{a}}_z$ is out of the page, $\omega = 10^6$ [rad/s], and the surface area covered by the circuit is 1 [cm^2]. If there is no B-field outside this circuit, calculate the voltage V_m measured by a voltage meter for the three configurations shown. Assume that the meter resistance is infinite and that all the transformer emf generated in the circuit is dropped across the resistors (which means that the self-inductance of the loop is negligible).

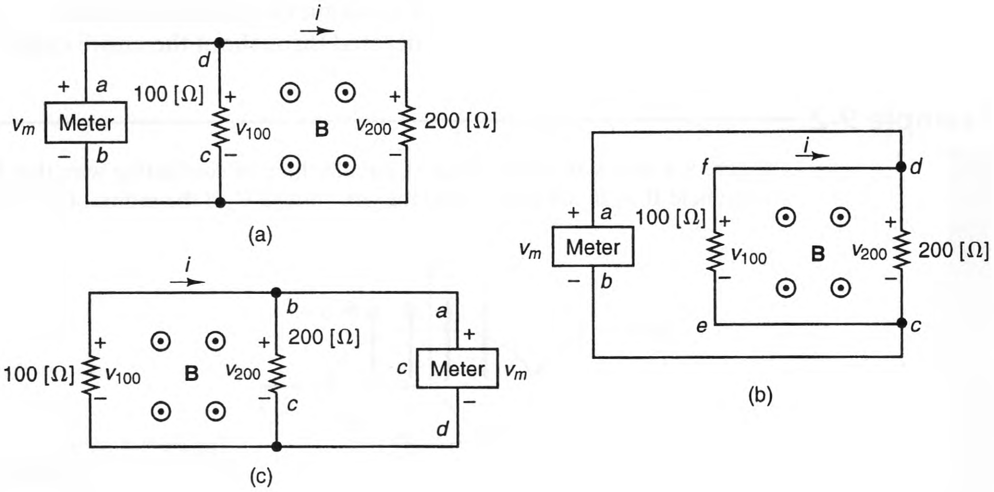


Figure 9-5 Three different voltage meter configurations for a circuit that is linked by a time-varying flux.

Solution:

Since the meter resistance is infinite, the same current *i* flows around the circuit in all three configurations. Evaluating Faraday's law clockwise around the circuit path and noting that **B** and **ds** are antiparallel, we obtain

$$\oint_C \mathbf{E} \cdot \mathbf{d\ell} = (100 + 200)i = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot \mathbf{ds} = 2 \times 10^6 \cos \omega t \times (0.01)^2.$$

Solving for *i*, we have

$$i = 0.666 \cos \omega t \qquad [\text{A}].$$

With this value for *i*, the resistor voltages are

$$v_{100} = -100i = -66.66 \cos \omega t \qquad [\text{V}]$$

$$v_{200} = +200i = 133.3 \cos \omega t \qquad [\text{V}].$$

a) For this configuration, we can evaluate Faraday's law over the path *abcd*. Since no flux is enclosed by this path, we obtain

$$\oint_C \mathbf{E} \cdot \mathbf{d\ell} = v_m - v_{100} = 0,$$

which yields

$$v_m = v_{100} = -66.66 \cos \omega t$$

b) The path *adcb* encloses the same flux as the circuit, so Faraday's law reads

$$\oint_C \mathbf{E} \cdot \mathbf{d\ell} = v_{200} - v_m = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot ds = 200 \cos \omega t.$$

Solving for v_m, we obtain

$$v_m = v_{200} - 200 \cos \omega t = -66.66 \cos \omega t,$$

which shows that in this case the meter is reading v_{100}, *not* v_{200}. We can obtain the same result by integrating along the path *abcefd*. Since no flux is linked by this path, we obtain

$$\oint_C \mathbf{E}\cdot\mathbf{d}\ell = v_m - v_{100} = 0.$$

c) For this configuration, we can evaluate Faraday's law around the path *adcb*. Since no flux is enclosed by this path, we obtain

$$\oint_C \mathbf{E}\cdot\mathbf{d}\ell = v_m - v_{200} = 0,$$

which yields

$$v_m = v_{200} = 133.3\cos\omega t.$$

In this case, we see that the voltage read by the meter is not necessarily the same as what might be expected from a quick look at the circuit. The key to knowing what voltage the meter will read is to know whether or not the meter circuit itself contains any time-varying flux.

9-2-1-1 Transformers. Transformers are devices in which two or more electric circuits are linked by a common magnetic flux. Because of this linkage, changes in the current in one circuit strongly affect the currents and voltages in the other circuits. Transformers are used to provide dc isolation between electric circuits and to change the amplitudes of voltages, currents, and impedances.

Figure 9-6a shows an ***ideal transformer***, which consists of two independent windings around a common core that has infinite permeability. The circuit symbol for this device is shown in Figure 9-6b. The dots next to each winding symbol indicate the polarity of the windings; positive currents flowing into the dotted terminals of each winding produce fluxes that circulate in the same direction. Since the core permeability μ is infinite, all the flux is confined to the core and links every turn of both windings. To start our analysis, let us apply Faraday's law to the contour defined by the path of the first winding. Starting from the top terminal and integrating around the windings and then upward from the bottom terminal towards the starting point, we can write

$$\oint_{C_1} \mathbf{E}\cdot\mathbf{d}\ell = -v_1(t) = -\frac{\partial}{\partial t}\int_{S_1} \mathbf{B}\cdot\mathbf{ds},$$

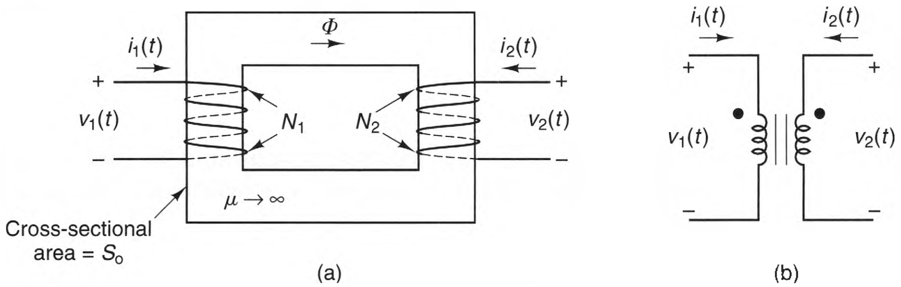


Figure 9-6 An ideal transformer. a) Physical geometry. b) Equivalent circuit.

where S_1 is the surface that is bounded by the path C_1 in a right-handed sense. This surface cannot be drawn on a plane, since the path spirals around the core. However, it can be seen from Figure 9-6a that S_1 intersects the core N_1 times. This means that the integral $\int_{S_1} \mathbf{B} \cdot \mathbf{ds}$ can be written as

$$\int_{S_1} \mathbf{B} \cdot \mathbf{ds} = N_1 \int_{S_o} \mathbf{B} \cdot \mathbf{ds} = N_1 \Phi,$$

where S_o is the cross section of the core and $\Phi = \int_{S_o} \mathbf{B} \cdot \mathbf{ds}$ is the magnetic flux that passes through S_o. Solving for $v_1(t)$, we obtain

$$v_1(t) = N_1 \frac{\partial \Phi}{\partial t}. \tag{9.3}$$

In a similar manner, we can evaluate Faraday's law around the contour of the second winding, S_2, yielding

$$v_2(t) = N_2 \frac{\partial \Phi}{\partial t}. \tag{9.4}$$

Dividing Equation (9.3) by Equation (9.4), we see that the flux terms Φ cancel, resulting in the expression

$$\frac{v_1(t)}{v_2(t)} = \frac{N_1}{N_2} \qquad \text{(Ideal transformer).} \tag{9.5}$$

Hence, for an ideal transformer, the winding voltages have the same waveshapes (i.e., there is no distortion), and their magnitude ratio equals the windings ratio; the winding with the largest number of turns always has the largest voltage. This relationship is responsible for one of the major uses of transformers—transforming voltage levels from one circuit to another. Often, transformers are called step-up or step-down transformers when they are used to increase or decrease voltage levels, respectively.

We can derive a similar relationship between the winding currents by using Ampere's law,

$$\oint_C \mathbf{H} \cdot \mathbf{d\ell} = I.$$

If we choose C to be the clockwise path around the center line of the core, then the current I that passes through this path in a right-handed sense is $N_1 i_1(t) + N_2 i_2(t)$, so we obtain

$$\oint_C \mathbf{H} \cdot \mathbf{d\ell} = N_1 i_1(t) + N_2 i_2(t).$$

However, $H = B/\mu \to 0$ *as* $\mu \to \infty$, so the line integral on the left equals zero, resulting in

$$\frac{i_1(t)}{i_2(t)} = -\frac{N_2}{N_1} \quad \text{(Ideal transformer).} \qquad (9.6)$$

Equation (9.6) states that the currents into each winding of an ideal transformer have opposite signs and their magnitude ratio equals the reciprocal of the windings ratio; the winding with the largest number of turns always has the current with the smallest magnitude. An important consequence of this relationship between the currents is that the total flux Φ produced by both windings is zero. This is because the currents have opposite sigrs and the flux generated by each winding is proportional to the number of turns. This somewhat counterintuitive result occurs because the permeability of an ideal transformer core is infinite.

We can use Equations (9.5) and (9.6) to calculate the energy balance between two terminal ports of an ideal transformer. Remembering that the power entering a terminal pair equals the voltage across the terminals times the current flowing into the positive terminal, we find that the total instantaneous input power is

$$p(t) = v_1(t)i_1(t) + v_2(t)i_2(t) = v_1(t)i_1(t) + \left[\frac{N_2}{N_1}v_1(t)\right] \times \left[-\frac{N_1}{N_2}i_1(t)\right] = 0.$$

Thus, the net instantaneous input power to an ideal transformer is always zero, which means that an ideal transformer does not store or dissipate energy in its core. As we will see later in the chapter, this is consistent with the fact that the magnetic flux inside an ideal transformer is zero.

One important application of transformers is for changing impedance levels. Figure 9-7 shows an ideal transformer with a lumped resistor of value R_L connected to the terminals of a second winding. To find the input resistance at the terminals of the first winding, we can first substitute Equation (9.5) into Equation (9.6) to obtain

$$R_{\text{in}} = \frac{v_1}{i_1} = -\left(\frac{N_1}{N_2}\right)^2 \frac{v_2}{i_2}.$$

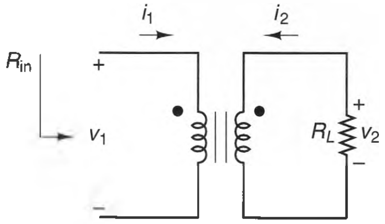


Figure 9-7 An ideal transformer connected to a resistive load.

Then, substituting $v_2/i_2 = -R_L$, we find that this expression becomes

$$R_{\text{in}} = \left(\frac{N_1}{N_2}\right)^2 R_L. \tag{9.7}$$

Thus, the effective impedance of a resistor can be either raised or lowered by attaching it to a transformer with an appropriate turns ratio.

Real transformers deviate from ideal transformers when power losses and flux leakages cannot be ignored. Typically, these deviations are most pronounced at very high frequencies, where core losses become significant, or high powers, where saturation effects become significant.

9-2-1-2 Eddy Currents. Figure 9-8a shows a cross section of a high-permeability core that carries a uniform, time-varying magnetic flux. As we showed in Example 9-1, the time-varying flux induces an emf around the cross section of the core. If the core has a nonzero conductivity, a conduction current $\mathbf{J} = \sigma\mathbf{E}$ is induced that circulates around the axis of the core. These currents are called ***eddy currents*** and are shown in Figure 9-8a. Eddy currents are usually undesirable in transformers and motors, because they produce ohmic (I^2R) losses. These losses are particularly prevalent in ferromagnetic materials, since they have fairly large conductivities.

One way to reduce the eddy current loss in ferromagnetic cores is by fabricating them out of thin strips (called laminae), each separated by a thin layer of insulating material ($\sigma \approx 0$). Such a laminated core is depicted in Figure 9-8b. Because of the insulating layers, eddy currents can flow around each lamina, but not between them. This reduces the dissipated power, since the induced emf in each laminate is proportional to its cross-sectional area, whereas its resistance is proportional to its perimeter. When using such a scheme, the effective permeability of a laminated core is somewhat reduced. This is because a small percentage of its cross section is filled with the nonmagnetic, insulating material. Nevertheless, it is a small price to pay for the significant reduction in the eddy current losses.

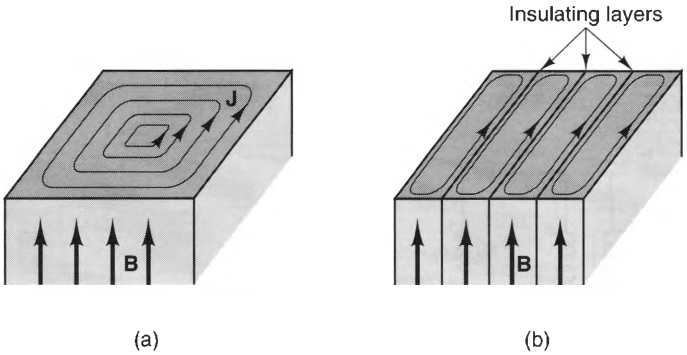


Figure 9-8 Eddy currents a) Solid core. b) Laminated core.

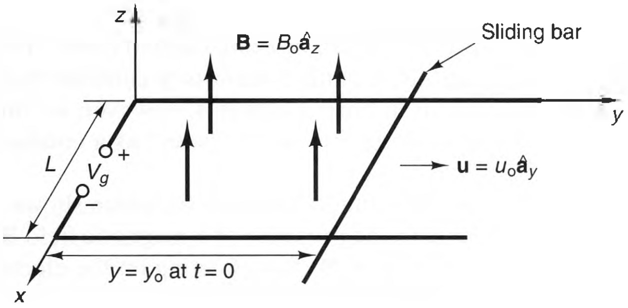

z
$\mathbf{B} = B_o\hat{\mathbf{a}}_z$
Sliding bar
y
$\mathbf{u} = u_o\hat{\mathbf{a}}_y$
L
+
V_g
–
$y = y_o$ at $t = 0$
x

Figure 9-9 A circuit with a sliding bar in a uniform B-field.

9-2-2 Motional emf

A simple experiment that demonstrates motional emf is depicted in Figure 9-9. Here, a metal bar slides over two metal rails in the presence of a uniform, time-invariant magnetic field $\mathbf{B} = B_o\hat{\mathbf{a}}_z$. The velocity of the bar is $\mathbf{u} = u_o\hat{\mathbf{a}}_y$ and its position is $y = y_o$ at $t = 0$. Also, the circuit is open circuited on the left side, which means that no current flows in the circuit. The circuit perimeter changes with time t, so we will denote the counter clockwise path around this circuit by the symbol $C(t)$ and the area that it encloses as $S(t)$. Along this path, Faraday's law reads

$$\oint_{C(t)} \mathbf{E}\cdot\mathbf{d\ell} = -\frac{d}{dt}\int_{S(t)} \mathbf{B}\cdot\mathbf{ds}. \tag{9-8}$$

Since the current is zero, there is no ohmic voltage drop along the conductors, so $\mathbf{E}\cdot\mathbf{d\ell} = 0$ everywhere except in the gap. Thus,

$$\oint_{C(t)} \mathbf{E}\cdot\mathbf{d\ell} = V_g,$$

where V_g is the voltage across the gap.

Turning our attention now to the surface integral, we note that the surface is expanding with time. This means that we must first perform the integration and then perform the time differentiation. This yields

$$-\frac{d}{dt}\int_{S(t)} \mathbf{B}\cdot\mathbf{ds} = -\frac{d}{dt}[B_oL(u_ot + y_o)] = -B_oLu_o,$$

Hence, the voltage in the gap is given by

$$V_g = -B_ou_oL. \tag{9.9}$$

Notice that the sign of this voltage is such that if a resistor were placed across the gap, a clockwise current would be induced whose B-field would counter the increasing flux linking the circuit. This tendency of the induced current to oppose changes in the flux linkage is called ***Lenz's law***.

Because the conducting path in this circuit is in motion, it is natural to choose the same path as the integration contour when applying Faraday's law. However, this requires that the time differentiation be performed after the surface integration, which was no real problem in this case, since the geometry was simple. For more complicated geometries, however, such a procedure can be quite cumbersome. For these situations, it is often attractive to integrate around a *stationary* contour that coincides with the moving contour at some instant in time. To do this, however, we must ask the question; What is the E-field that a stationary observer "sees" as a conductor moves past when a B-field is present?

We can answer this question by referring to Figure 9-10, which shows a conducting wire that is moving with a velocity **u** in the presence of a magnetic field **B**. We will consider the electric ($q\mathbf{E}$) and magnetic ($q\mathbf{u} \times \mathbf{B}$) forces acting on the electrons inside this conductor as measured by two observers: one that moves with the wire and one that is stationary.

To an observer moving with the conductor, the conductor appears to be at rest, just as we appear to be at rest when standing at one spot on the moving earth. This observer perceives no magnetic force and an electric force of value $e\mathbf{E}$, where e is the electron charge. Thus, as far as the moving observer is concerned, the total force acting on each charge is $\mathbf{F} = e\mathbf{E}$.

To find what E-field a stationary observer would measure as the wire passes by, we can use a well-established law from basic physics: *Moving and nonmoving observers always measure the same force on an object, as long as the difference between their velocities is much less than the speed of light.* Hence, a stationary observer will also measure a total force $\mathbf{F} = q\mathbf{E}$ on each charge. However, since such an observer perceives that the charges are indeed moving with velocity **u** in the presence of a magnetic field, a part of this total force must be a magnetic force $q\mathbf{u} \times \mathbf{B}$. This means that the electric force perceived by the stationary observer is different from the one mearured by the moving observer. If the E-field measured by the stationary observer is $\mathbf{E}'$, we must have

$$\mathbf{F} = q\mathbf{E} = q\mathbf{E}' + q\mathbf{u} \times \mathbf{B}.$$

Solving for $\mathbf{E}'$, we obtain the result

$$\underset{\substack{\uparrow \\ \text{Stationary} \\ \text{Observer}}}{\mathbf{E}'} = \underset{\substack{\uparrow \\ \text{Moving} \\ \text{Observer}}}{\mathbf{E}} - \mathbf{u} \times \mathbf{B}, \tag{9.10}$$

which shows that the stationary observer "sees" a different E-field than does the observer who is moving with the conductor.

We can use this result to write Faraday's law in a form that is easier to apply when moving conductors are present. If C is a closed, stationary path that bounds a stationary surface S, we can write

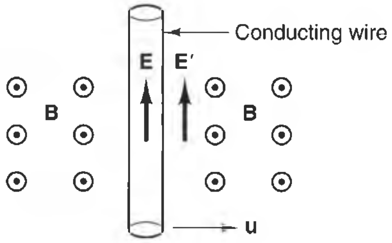


Figure 9-10 A moving conductor in the presence of a B-field. **E** and **E**′ are the electric fields measured in the moving and stationary reference frames, respectively.

$$\oint_C \mathbf{E}' \cdot d\boldsymbol{\ell} = -\frac{d}{dt}\int_S \mathbf{B} \cdot \mathbf{ds} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot \mathbf{ds},$$

where $\mathbf{E}'$ is observed by a stationary observer along the path. If any conductors are moving through the portions of the path at the instant the integrals are evaluated, we can replace $\mathbf{E}'$ with $\mathbf{E} - \mathbf{u} \times \mathbf{B}$, where $\mathbf{E}$ is the E-field "seen" by an observer that is moving with the conductor and $\mathbf{u}$ is the velocity of the conductor relative to the fixed path. This yields the following expression equivalent Faraday's law:

$$\oint_C \mathbf{E} \cdot d\boldsymbol{\ell} = \oint_C \mathbf{u} \times \mathbf{B} \cdot d\boldsymbol{\ell} - \int_S \frac{\partial \mathbf{B}}{\partial t} \cdot \mathbf{ds}. \tag{9.11}$$

Comparing Equations (9.8) and (9.11), we see that their left-hand sides are the same at the instant of time when the paths $C(t)$ and C coincide, but their right-hand sides are different. In essence, the effect of moving the time derivative inside the surface integral of Equation (9.11) is to add the line integral of $\mathbf{u} \times \mathbf{B}$ around the perimeter of the path. When using Equation (9.11), it is important to remember that even though the path C is stationary, $\mathbf{E}$ at each point on the path is the field measured by someone moving through that point with velocity $\mathbf{u}$.

This new form of Faraday's law makes it clear that the electromotive force around a closed path can be considered as the sum of two terms. The first results from the motion of conductors as they "cut" through the lines of a magnetic field. This is the ***motional emf*** contribution. The second contribution, called ***transformer emf***, is the result of a time-varying flux linking the path. Since their left-hand sides are the same, either form of Faraday's law can be used for a given problem, but this new form often simplifies calculations when moving conductors are present.

To see how to apply Equation (9.11), let us return to the moving-bar configuration shown in Figure 9-9. This time, we will choose the integration path to be a stationary, counter clockwise circuit path that coincides with the actual circuit path at some time t_0. The E-field measured in a conductor by someone moving with it is zero when no current is flowing, so $\mathbf{E} = 0$ at all points on the path C except in the gap. Thus, we have

$$\oint_C \mathbf{E} \cdot d\boldsymbol{\ell} = V_g. \tag{9.12}$$

Looking at the right-hand side terms, we first notice that the transformer emf is zero, since $\partial \mathbf{B}/\partial t = 0$. This means that the voltage V_g is generated solely from motional emf. Moreover, the only contribution to the motional emf occurs along the sliding bar, where $\mathbf{u} \times \mathbf{B} = u_0 B_0 \, \hat{\mathbf{a}}_x$. Integrating counter clockwise over the length L of the bar, we obtain

$$V_g = \oint_C \mathbf{u} \times \mathbf{B} \cdot d\boldsymbol{\ell} = \int_L^0 u_0 B_0 dx = -u_0 B_0 L, \tag{9.13}$$

Which agrees with the result given by Equation (9.9).

Example 9-4

Figure 9-11 shows a ***Faraday disk generator***, which consists of a metal disk of radius a, rotating with constant angular velocity ω in the presence of a constant magnetic field $\mathbf{B} = B_0\hat{\mathbf{a}}_z$ The output circuit consists of wires that are connected to the shaft and the disk through low-resistance brushes. Find the open-circuit voltage V_0 if the radius of the shaft is negligible.

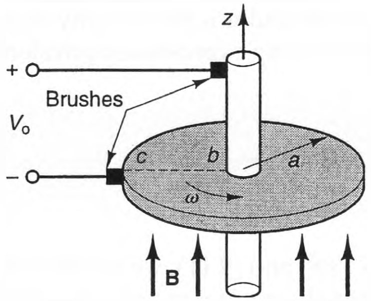


Figure 9-11 A Faraday disk generator.

Solution:

a) Since **B** is time invariant, the transformer emf is zero, so we can write

$$\oint_C \mathbf{E} \cdot d\boldsymbol{\ell} = \oint_C \mathbf{u} \times \mathbf{B} \cdot d\boldsymbol{\ell},$$

where C is the counterclockwise path C along the wires, shaft, and the line segment bc along the disk. We know that $\mathbf{E} = 0$ on metal surfaces when measured by someone moving them, so we have $\mathbf{E} = 0$ and $\mathbf{u} = \omega\rho\,\hat{\mathbf{a}}_\phi$along the line segment cb. This means that $\oint_C \mathbf{E} \cdot d\boldsymbol{\ell} = V_0$.

To evaluate the integral $\oint_C \mathbf{u} \times \mathbf{B} \cdot d\boldsymbol{\ell}$, we note that $\mathbf{u} \times \mathbf{B} = \omega\rho B_0\,\hat{\mathbf{a}}_\rho$along the line segment cb, where $\hat{\mathbf{a}}_\rho$is directed radially outward from the shaft. Since the integration path is directed counter clockwise around the path, $d\boldsymbol{\ell} = d\rho\hat{\mathbf{a}}_\rho$and $\mathbf{u} \times \mathbf{B} \cdot d\boldsymbol{\ell} = \omega\rho B_0 d\rho$. Substituting, we obtain

$$V_0 = \oint_C \mathbf{u} \times \mathbf{B} \cdot d\boldsymbol{\ell} = \int_a^0 \omega\rho B_0 d\rho = -\frac{\omega B_0 a^2}{2}.$$

9-3 Inductance

Faraday's law predicts that a voltage can be induced in a stationary circuit due to the B-field generated by its own time-varying current or the time-varying current in another circuit. We call the former effect *self-inductance* and the latter *mutual inductance*. To model these effects, consider the two thin wires that form the loops shown in Figure 9-

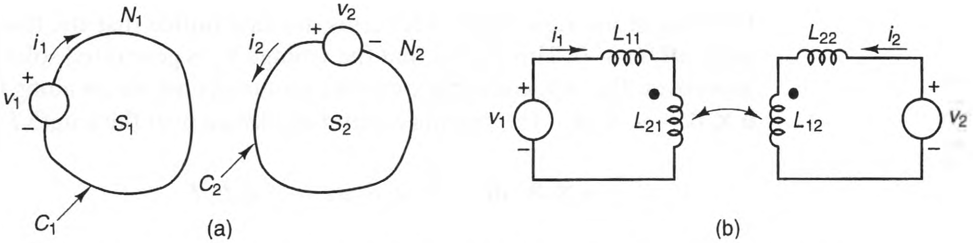


Figure 9-12 Two magnetically coupled loops. a) Geometry. b) Equivalent circuit.

12a, each with N_1 and N_2 turns, respectively. Time-varying voltage sources $v_1(t)$ and $v_2(t)$ are attached to each loop, resulting in the currents $i_1(t)$ and $i_2(t)$, respectively. Applying Faraday's law to a clockwise path C_1 around loop #1, we obtain

$$\oint_{C_1} \mathbf{E} \cdot \mathbf{d\ell} = -\frac{d}{dt}\int_{S_1} \mathbf{B} \cdot \mathbf{ds}, \tag{9.14}$$

where S_1 is the surface that is bounded in a right-handed sense by the path C_1. The integral $\int_{S_1} \mathbf{B} \cdot \mathbf{ds}$ is called a ***flux linkage***, since it represents the flux passing through (i.e., linking) a current path. Flux linkages are typically denoted by the symbol Λ. Because loop #1 has N_1 turns, the flux linkage Λ_1 is simply N_1 times the flux Φ_1 that passes through the cross section of the winding:

$$\Lambda_1 \equiv \int_{S_1} \mathbf{B} \cdot \mathbf{ds} = N_1 \Phi_1 \qquad [\text{Wb}]. \tag{9.15}$$

Substituting Equation (9.15) into Equation (9.14), we find that

$$\oint_{C_1} \mathbf{E} \cdot \mathbf{d\ell} = -\frac{d}{dt}\Lambda_1. \tag{9.16}$$

Since there are two currents, i_1 and i_2, flowing in this system, the flux linkage Λ_1 can be divided into two components:

$$\Lambda_1 = \Lambda_{11} + \Lambda_{21}. \tag{9.17}$$

Here Λ_{11} is the flux linkage in circuit #1 due to i_1 and Λ_{21} is the flux linkage in circuit #1 due to i_2; thus,

$$\Lambda_{11} = \int_{S_1} \mathbf{B}_1 \cdot \mathbf{ds} \tag{9.18}$$

and

$$\Lambda_{21} = \int_{S_1} \mathbf{B}_2 \cdot \mathbf{ds}. \tag{9.19}$$

Substituting Equations (9.17)–(9.19) into Equation (9.16), we obtain

$$\oint_{C_1} \mathbf{E} \cdot \mathbf{d\ell} = -\frac{d}{dt}\Lambda_{11} - \frac{d}{dt}\Lambda_{21}. \tag{9.20}$$

If the wires have zero resistance, the only contribution to the line integral occurs at the voltage source, so

$$\oint_{C_1} \mathbf{E} \cdot \mathbf{d\ell} = -v_1.$$

This means that Equation (9.20) can be written as

$$v_1 = \frac{d\Lambda_{11}}{dt} + \frac{\Lambda_{21}}{dt}. \tag{9.21}$$

If the permeabilities of all the materials are linear, Λ_{11} and Λ_{21} are proportional to i_1 and i_2, respectively, so Equation (9.21) can be written in the form

$$L_{11}\frac{di_1}{dt} + L_{21}\frac{di_2}{dt} = v_1, \tag{9.22}$$

where L_{11} and L_{21} are called ***self-*** and ***mutual inductances*** (respectively), defined by

$$L_{ij} \equiv \frac{\text{flux linking } C_j \text{ due to current in } C_i}{\text{current in } C_i} = \frac{\Lambda_{ij}}{I_i} \quad [\text{H}]. \tag{9.23}$$

The unit of inductance is the henry [H], which is equivalent to webers per ampere [Wb/A]. Notice from this definition that Λ_{ij} is the flux linkage in the j^{th} circuit due to the current I_i in the i^{th} circuit.

A similar application of Faraday's law around the contour C_2 of the second circuit yields

$$L_{12}\frac{di_1}{dt} + L_{22}\frac{di_2}{dt} = v_2. \tag{9.24}$$

Taken as a pair, Equations (9.22) and (9.24) can be represented by the lumped circuit shown in Figure 9-12b. Here, the voltage drops across the self-inductances L_{11} and L_{22} represent emf's caused by time variations of the currents in their own circuits. On the other hand, the mutual inductances L_{12} and L_{21} represent the emf's generated in one circuit due to time variations of the current in the other circuit.

Self-inductances are always positive, regardless of how the currents and voltages are defined. In principal, mutual inductances can be either positive or negative, but it is always possible to define the polarities of the circuit currents so that their mutual inductance has a positive value. Typically, this polarity is indicated using the "dot convention." According to this convention, positive currents flowing into the dotted terminal's dots produce fluxes in each loop that add. Notice that the current directions of i_1 and i_2 in Figure 9-12a follow this convention (assuming that both circuits are in the plane of the paper), so the dots in the equivalent circuit are arranged so that i_1 and i_2 are shown entering the dotted terminals.

The inductances we have discussed so far have been the total self- and mutual-inductances of entire circuits. This is a natural starting point, since Faraday's law of induction specifies induced voltages for complete circuit paths. However, it is often more convenient to divide these inductances into one or more discrete inductances, each associated with a specific portion of the circuit. This makes the most sense when the flux linking a circuit is concentrated in more than one lumped element, such as a tightly wound coil. In these cases, we can still use Equation (9.23) to determine the self- and mutual inductances of the lumped elements, which is demonstrated in the following example.

Example 9-5

Find the self-inductance of the N-turn solenoid shown in Figure 9-13 that consists of N turns of wire around a long, high permeability core. Assume that the core has cross-sectional area S and permeability μ.

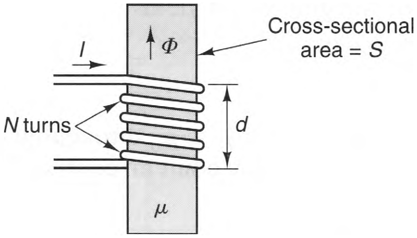


Figure 9-13 A solenoid inductor with a magnetic core.

Solution

If $\mu \gg \mu_0$ and d is much larger than the core diameter, we can assume that the flux leakage out of the core is small, so the field inside the windings is approximately the same as that in an infinite solenoid. Using Equation (7.37), we find that the B-field directed along the axis of the solenoid is given by

$$B \approx \frac{\mu N I}{d},$$

where we have replaced the permeability of free space with the core permeability μ. The flux passing through the core is

$$\Phi = \int_S \mathbf{B} \cdot \mathbf{ds} \approx \frac{\mu N I S}{d}.$$

This flux links the current N times, so the flux linkage is $\Lambda = N\,\Phi$. Using Equation (9.23), we obtain

$$L = \frac{\Lambda}{I} \approx \frac{\mu N^2 S}{d}.$$

9-3-1 MUTUAL INDUCTANCE AND THE NEUMANN FORMULA

An important formula involving the mutual inductance between two circuits can be derived with the help of the magnetic vector potential. For the two circuits shown in Figure 9-14, the mutual inductance L_{21} can be written as

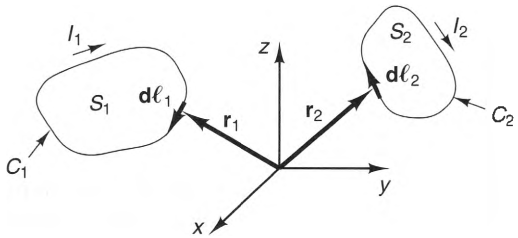


Figure 9-14 Geometry for deriving Neumann's formula for mutual inductance.

$$L_{21} = \frac{\Lambda_{21}}{I_2} = \frac{1}{I_2}\int_{S_1} \mathbf{B}_2 \cdot \mathbf{ds} = \frac{1}{I_2}\int_{S_1} \nabla \times \mathbf{A}_2 \cdot \mathbf{ds},$$

where $\mathbf{B}_2$ and $\mathbf{A}_2$ are the magnetic flux density and the magnetic vector potential generated by the current I_2, respectively. Using Stokes's theorem, we can write the integral on the right as a line integral, which yields

$$L_{21} = \frac{1}{I_2}\oint_{C_1} \mathbf{A}_2 \cdot d\boldsymbol{\ell}_1.$$

If no magnetic materials are present, $\mu = \mu_o$ throughout all space. Thus, we can use Equation (7.44) to represent $\mathbf{A}_2$ at any point on contour C_1. We obtain

$$\mathbf{A}_2 = \frac{\mu_o}{4\pi}\oint_{C_2} \frac{I_2}{|\mathbf{r}_2 - \mathbf{r}_1|}\, d\boldsymbol{\ell}_2,$$

where $\mathbf{r}_2$ is a source point on C_2 and $\mathbf{r}_1$ is a field point on C_1. Substituting this expression into the previous integral yields

$$L_{21} = \frac{\mu_o}{4\pi}\oint_{C_1}\oint_{C_2} \frac{d\boldsymbol{\ell}_1 \cdot d\boldsymbol{\ell}_2}{|\mathbf{r}_2 - \mathbf{r}_1|}. \tag{9.25}$$

Equation (9.25) is called the ***Neumann formula for mutual inductance.*** It can be used directly to calculate the mutual inductance of any filamentary system, although the calculations must almost always be performed numerically because of the difficult double-contour integrals. It also conveys two important properties of the mutual inductance between any pair of circuits. The first is that mutual inductance is a geometrical quantity that is only a function of the contours of the interacting circuits. The second is that mutual inductance is reciprocal, i.e., $L_{21} = L_{12}$. This is easily seen by interchanging the subscripts "1" and "2" in the integrand of Neumann's formula. This result is also true when magnetic materials are present, although it is more difficult to derive.

Example 9-6

Calculate the mutual inductance between the infinite line and the square loop shown in Figure 9-15.

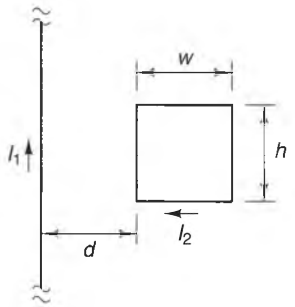


Figure 9-15 An-infinite line current and a square loop of current.

Solution:

Knowing that $L_{12} = L_{21}$, we can proceed by finding L_{12}, since the magnetic field generated by an infinite line is represented by a simple, known formula. Using Equation (7.25), we see that the flux Λ_{12} that links I_2 is given by

$$\Lambda_{12} = \int_{S_2} \mathbf{B}_1 \cdot \mathbf{ds} = \int_0^h \int_d^{d+w} \frac{\mu_o I_1}{2\pi\rho} \hat{\mathbf{a}}_\phi \cdot d\rho dz \hat{\mathbf{a}}_\phi = \frac{\mu_o I_1 h}{2\pi} \ln\left[\frac{d+w}{d}\right].$$

Thus, from Equation (9.23),

$$L_{12} = \frac{\Lambda_{12}}{I_1} = \frac{\mu_o h}{2\pi} \ln\left[\frac{d+w}{d}\right].$$

9-3-2 ENERGY STORAGE IN INDUCTIVE SYSTEMS

The relationship between inductance, current, and energy can be determined by considering the two circuits shown in Figure 9-16. Magnetic materials may or may not be present in these circuits; the only restriction that we will make is that the materials linear.

If we assume that neither circuit has any resistance, the currents i_1 and i_2 and the voltage sources v_1 and v_2 are related by

$$L_{11} \frac{di_1}{dt} + L_{21} \frac{di_2}{dt} = v_1 \tag{9.26}$$

$$L_{12} \frac{di_1}{dt} + L_{22} \frac{di_2}{dt} = v_2. \tag{9.27}$$

If both loops are initially open circuited, let us first establish a current I_1 in the first loop, while maintaining the open circuit in the second loop. The instantaneous power supplied to the first loop during this process is $p_{11}(t) = i_1 v_1 = i_1 L_{11} (di_1/dt)$. Integrating this expression over time, we obtain the energy W_{11} required to establish the steady current I_1:

$$W_{11} = L_{11} \int_o^t i_1 \frac{di_1}{dt} dt = L_{11} \int_o^{I_1} i_1 di_1 = \frac{1}{2} L_{11} I_1^2.$$

During this process, the power supplied to the second loop is zero, since it is open circuited.

Next, to establish a current I_2 in the second loop while maintaining the current I_1 in the first loop, energy must be supplied to both circuits, since current is now flowing in each. The instantaneous power supplied to the second loop is

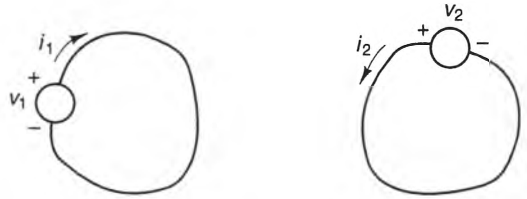


Figure 9-16 Two magnetitically coupled loops with voltage sources.

$p_{22}(t) = i_2 v_2 = i_2 L_{22}\,(di_2/dt)$. Integrating this power over time to find the energy W_{22} supplied to the second circuit, we obtain

$$W_{22} = L_{22}\int_0^t i_2 \frac{di_2}{dt}\,dt = L_{22}\int_0^{I_2} i_2\,di_2 = \frac{1}{2}L_{22}I_2^2.$$

Power must also be supplied to loop #1 to maintain its current I_1 in the presence of the emf generated by the time-varying current in loop #2. The instantaneous power supplied to loop #1 is $p_{12}(t) = I_1 v_1 = I_1 L_{12}\,(di_2/dt)$, which, when integrated, yields

$$W_{12} = L_{12}I_1\int_0^t \frac{di_2}{dt}\,dt = L_{12}I_1\int_0^{I_2} di_2 = L_{12}I_1I_2.$$

The total energy W_m supplied by both sources to establish the currents I_1 and I_2 is the sum of W_{11}, W_{12}, and W_{22}, so we have

$$W_m = \frac{1}{2}L_{11}I_1^2 + L_{12}I_1I_2 + \frac{1}{2}L_{22}I_2^2. \tag{9.28}$$

Since $L_{ij} = L_{ji}$, this result can be written in the form

$$W_m = \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2} L_{ij}I_iI_j. \tag{9.29}$$

For a system of N circuits, Equation (9.29) can be generalized to read

$$W_m = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} I_iI_jL_{ij} \qquad \text{[J]}. \tag{9.30}$$

The energy W_m stored in a system of N loops can also be expressed in terms of the total flux linking each loop. Substituting $\Lambda_{ij} = I_j L_{ij}$ into Equation (9.30), we find that

$$W_m = \frac{1}{2}\sum_{i=1}^{N} I_i \sum_{j=1}^{N} L_{ij}I_j = \frac{1}{2}\sum_{i=1}^{N} I_i \sum_{j=1}^{N} \Lambda_{ji}. \tag{9.31}$$

Since Λ_{ji} is the flux linkage in the i^{th} loop from the j^{th} current, we see that the sum $\sum_{j=1}^{N} \Lambda_{ji}$ is the total flux linking the i^{th} loop, which we denote as Λ_i. Using this, we can write Equation (9.31) as a single sum:

$$W_m = \frac{1}{2}\sum_{i=1}^{N} I_i\Lambda_i \qquad \text{[J]}. \tag{9.32}$$

Finally, we can use Equation (9.28) to determine the maximum value that the mutual inductance L_{12} between two circuits can attain. By completing the square, this expression can be rewritten in the form

$$W_m = \frac{1}{2} L_{22} \left(I_2 + \frac{L_{12}}{L_{22}} I_1 \right)^2 + \frac{1}{2} \left(L_{11} - \frac{L_{12}^2}{L_{22}} \right) I_1^2$$

The first term is always greater than zero for all values of L_{11} and L_{22}, but the second term is greater than zero only when the mutual inductance is less than or equal to the geometric mean of the two self-inductances. Since the energy stored in the magnetic field must be greater than or equal to zero, we must have

$$L_{12} \leqslant \sqrt{L_{11} L_{22}}. \tag{9.33}$$

This is an important result, since it shows the upper bound that the mutual inductance can have in terms of the self-inductances of the circuits. Often, the degree to which two circuits are inductively coupled is described by the coupling coefficient k, which is defined as

$$k = \frac{L_{12}}{\sqrt{L_{11} L_{22}}}, \tag{9.34}$$

where

$$0 \leqslant k \leqslant 1. \tag{9.35}$$

Example 9-7

The circuit shown in Figure 9-17 has $L_{11} = 10$ [μH], $L_{22} = 3$ [μH], and $L_{12} = 1$ [μH]. Calculate the coupling factor between these circuits and the energy stored in the magnetic field at the instant when the currents have values $i_1 = 1$ [mA] and $i_2 = 3$ [mA].

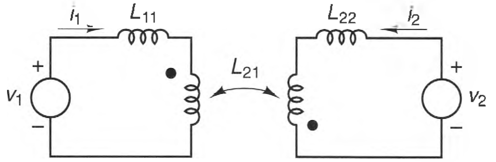


Figure 9-17 Two magnetically coupled circuits.

Solution:

Substituting the self- and mutual inductance values into Equation (9.34), we obtain

$$k = \frac{1 \times 10^{-6}}{\sqrt{10 \times 10^{-6} \times 3 \times 10^{-6}}} = 0.183.$$

Since $k \ll 1$, these circuits are loosely coupled.

We can use Equation (9.28) to determine the magnetic energy W_m stored by these inductors, but we must use the current polarities that correspond to the dot convention. In this case, since the positive direction of i_2 is into the undotted terminal, we must use $-i_2$ in place of i_2. Substituting, we obtain

$$W_m = \frac{1}{2}(1 \times 10^{-3})^2 (10 \times 10^{-6}) + (1 \times 10^{-3})(-3 \times 10^{-3})(1 \times 10^{-6})$$

$$+\frac{1}{2}(-3\times10^{-3})^2(3\times10^{-6})$$

$$= 1.55\times10^{-11} \qquad [\mathrm{J}].$$

9-3-3 ENERGY AND MAGNETIC FIELDS

The energy in an electrostatic system can be expressed in terms of its electric field; the same is true for a magnetostatic system. But whereas we were able to derive the electrostatic expression by constructing a continuous charge distribution charge by charge, the magnetostatic expression cannot be easily derived using this technique. This is because as each current loop is brought in from infinity, work must be expended to keep the currents in the loops constant, since the flux linking them varies while the loops are in motion (even if the motion is exceedingly slow). Instead, we will derive the expression by starting with the energy contained in a system of N loops and extend it to describe the energy contained in a continuous current distribution.

Let us start by considering the volumetric current distribution **J** shown in Figure 9-18 as a collection of N separate loops. Each loop has cross section Δs_i and carries a current $\Delta I_i = J_i\Delta s_i$. From Equation (9.32) the energy contained in the loops is

$$W_m = \frac{1}{2}\sum_{i=1}^{N}\Delta I_i\Lambda_i = \frac{1}{2}\sum_{i=1}^{N}J_i\Delta s_i\Lambda_i,$$

where $\Lambda_i = \Phi_i = \int_{S_i}\mathbf{B}\cdot\mathbf{ds}$. Also, from Equation (7.45), the flux Φ_i passing through the i^{th} loop can be expressed in terms of the line integral of **A** around the contour C_i:

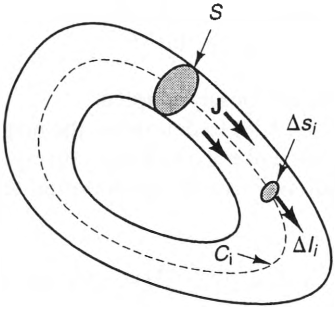
$$\Phi_i = \oint_{S_i}\mathbf{B}\cdot\mathbf{ds} = \oint_{C_i}\mathbf{A}\cdot\mathbf{d\ell}.$$

Hence, W_m can be written in the form

$$W_m = \frac{1}{2}\sum_{i=1}^{N}J_i\Delta s_i\oint_{C_i}\mathbf{A}\cdot\mathbf{d\ell}.$$

As $N\to\infty$, this expression becomes

$$W_m = \frac{1}{2}\int_S J\left[\oint_{C_i}\mathbf{A}\cdot\mathbf{d\ell}\right]ds,$$



where S is the cross-sectional surface of the current distribution that is at all points perpendicular to $\mathbf{J}$. This expression can be further simplified by noting that $\mathbf{J}$ and $\mathbf{d\ell}$ are always parallel. Thus, $Jds\, \mathbf{A} \cdot \mathbf{d\ell} = \mathbf{A} \cdot \mathbf{J}\, dv$, which, when substituted into the integral in the preceding expression, yields

$$W_m = \frac{1}{2}\int_V \mathbf{A} \cdot \mathbf{J}\, dv \qquad [\mathrm{J}] \qquad \text{(Linear media)}, \tag{9.36}$$

where V is the volume filled by the current distribution $\mathbf{J}$. This expression for W_m is the dual of Equation (6.29), which relates the energy stored in an electrostatic system to the charge density ρ_v and the potential distribution V.

We can express W_m in terms of $\mathbf{B}$ and $\mathbf{H}$ alone by first substituting $\nabla \times \mathbf{H} = \mathbf{J}$ into Equation (9.36) to obtain

$$W_m = \frac{1}{2}\int_V \mathbf{A} \cdot \nabla \times \mathbf{H}\, dv.$$

Using Equation (B.5), we can write the integrand

$$\mathbf{A} \cdot \nabla \times \mathbf{H} = -\nabla \cdot (\mathbf{A} \times \mathbf{H}) + \mathbf{H} \cdot \nabla \times \mathbf{A}.$$

However, since $\nabla \times \mathbf{A} = \mathbf{B}$, we have

$$\mathbf{A} \cdot \nabla \times \mathbf{H} = -\nabla \cdot (\mathbf{A} \times \mathbf{H}) + \mathbf{H} \cdot \mathbf{B}.$$

Substituting this into the integral and using the divergence theorem, we obtain

$$W_m = -\frac{1}{2}\oint_S (\mathbf{A} \times \mathbf{H}) \cdot \mathbf{ds} + \frac{1}{2}\int_V \mathbf{B} \cdot \mathbf{H}\, dv, \tag{9.37}$$

where S is the surface that bounds the volume V.

Equation (9.37) is valid for all volumes V that completely enclose the current distribution $\mathbf{J}$. If we let $V \to V_\infty$, S becomes the sphere at infinity, S_∞. However, if $\mathbf{J}$ is contained in a finite volume, $\mathbf{A}$ and $\mathbf{H}$ are proportional to r^{-1} and r^{-2}, respectively, as $r \to \infty$. This means that the product $\mathbf{A} \times \mathbf{H}$ falls off as r^{-3} at large distances from the origin. Thus, the surface integral over S_∞ vanishes, yielding

$$W_m = \frac{1}{2}\int_V \mathbf{B} \cdot \mathbf{H}\, dv \qquad [\mathrm{J}] \qquad \text{(Linear media)}, \tag{9.38}$$

For simplicity, we have dropped the subscript "∞" from the volume V in this expression, but it is to be understood that the integration takes place *everywhere* the product $\mathbf{B} \cdot \mathbf{H}$ is nonzero. The term $1/2\mathbf{B} \cdot \mathbf{H}$ has units of [J/m^3] and is called the ***magnetic energy density***. For isotropic media, $\mathbf{B} = \mu\mathbf{H}$, from which it follows that

$$W_m = \frac{1}{2}\int_V \mu|\mathbf{H}|^2 dv \quad [\mathrm{J}] \quad \textbf{(Linear, isotropic media).} \tag{9.39}$$

The energy W_m contained in a current distribution can be considered to reside either in the current distribution itself (see Equation (9.36)) or in the magnetic fields generated by the currents (see Equation (9.38)). These views are equivalent, because a current distribution and the magnetic field it generates are an inseparable pair; a complete knowledge of one completely defines the other. Also, even though we have used the magnetostatic vector potential in deriving Equations (9.38) and (9.39), these expressions are also valid for time-varying magnetic fields.

9-3-4 INDUCTANCE IN TERMS OF MAGNETIC FIELDS

Earlier in this chapter, we showed that the inductance of a circuit or element can be calculated from a knowledge of the physical layout of the currents. In many cases this is the best way to calculate inductance. But it is also possible to use the relationship between inductance and stored magnetic energy to calculate the inductance of a device in terms of the magnetic fields present in and around the device. To accomplish this, let us first consider a single circuit. Using Equation (9.30) for the case $N = 1$, we find that the magnetic energy stored by the device can be expressed in terms of its self-inductance L_{ii}. Thus,

$$W_m = \frac{1}{2} L_{ii} I_i^2,$$

where I_i is the current flowing in the circuit. Likewise, from Equation (9.38) we know that W_m can also be expressed in terms of the B- and H-fields generated by this circuit. Hence

$$W_m = \frac{1}{2}\int_V \mathbf{B}_i \cdot \mathbf{H}_i dv,$$

where $\mathbf{B}_i$ and $\mathbf{H}_i$ are the magnetic fields generated by this circuit and V is the entire volume over which $\mathbf{B}_i$ and $\mathbf{H}_i$ are nonzero. Equating these two expressions and solving for L_{ii}, we obtain the following expression for the self-inductance of a circuit:

$$L_{ii} = \frac{2W_m}{I_i^2} = \frac{1}{I_i^2}\int_V \mathbf{B}_i \cdot \mathbf{H}_i dv. \tag{9.40}$$

If N circuits are present that carry currents I_i and I_j, respectively, the net magnetic energy stored by these circuits can be expressed in terms of the self-inductances L_{ii} and L_{jj} and the mutual inductance L_{ij}, as well as the magnetic fields generated by both currents. If the B- and H-fields generated by the i^{th} circuit are denoted by $\mathbf{B}_i$ and

$\mathbf{H}_j$, respectively, we can use Equations (9.30) and (9.38) to express the energy stored by the system as

$$W_m = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int_V \mathbf{B}_i \cdot \mathbf{H}_j dv = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} I_i I_j L_{ij}.$$

Collecting terms with the same indices, we obtain the following expression for the mutual inductance L_{ij}:

$$L_{ij} = \frac{2W_{m_{ij}}}{I_i I_j} = \frac{1}{I_i I_j}\int_V \mathbf{B}_i \cdot \mathbf{H}_j dv. \tag{9.41}$$

Note that $\mathbf{B}_i$ and $\mathbf{H}_i$ are generated by I_i and I_j, respectively. The quantity $\int_V \mathbf{B}_i \cdot \mathbf{H}_j dv$ is called the ***mutual energy***.

An advantage of the energy formulations of self- and mutual inductance is that they do not require a knowledge of the flux linking the circuit or element. This is an attractive feature, since it is sometimes difficult to determine the flux linkages when the currents do not flow on thin wires.

Example 9-8

Calculate the self-inductance per meter of the two parallel metal strips shown in Figure 9-19. Assume that the strips are wide enough so that the magnetic field between them is uniform and negligible outside.

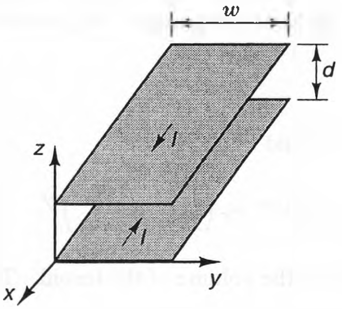


Figure 9-19 Two parallel metal strips, each of width w and separated by a distance d.

Solution:

Since the strips have width v and carry a current I, the current density on each strip is I/w [A/m]. If the ratio w/d is large, we can approximate the strips as infinite sheets and use Equation (7.30) to determine the B-field generated by these currents. Using the superposition principle and noting that the currents are oppositely directed, we find that

$$\mathbf{B} \approx \begin{cases} \frac{\mu_0 I}{w}\hat{\mathbf{a}}_y & 0 < z < d \\ 0 & \text{otherwise} \end{cases}.$$

Substituting this into Equation (9.40), we see that the inductance per unit length (along the x-direction) is given by

$$L = \frac{1}{I^2}\int_V \mathbf{B}\cdot\mathbf{H}dv = \frac{1}{I^2}\int_0^d\int_0^w\int_0^1 \mathbf{B}\cdot\mathbf{H}dxdydz$$

$$= \frac{1}{I^2}\int_0^d\int_0^w\int_0^1 \mu_o\left(\frac{I}{w}\right)^2 dxdydz = \frac{\mu_o d}{w} \qquad [\mathrm{H/m}].$$

We can obtain the same result by using the definition of inductance given by Equation (9.23). The flux per meter passing between the currents is $\Phi = BS$, where $S = 1\ [\mathrm{m}^2]$. This flux links the current once (i.e., $N = 1$), so $\Lambda = \Phi$, and we again obtain

$$L = \frac{\Lambda}{I} = \frac{\mu_o d}{w} \qquad [\mathrm{H/m}].$$

Example 9-9

Calculate the mutual inductance between the two windings on the toroidal core with mean radius ρ_o, cross-sectional radius a, and permeability μ_c, shown in Figure 9-20.

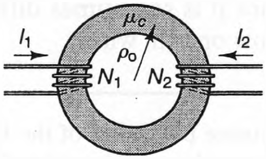


Figure 9-20 A toroidal inductor with two windings.

Solution:

Assuming that $\rho_o \gg a$ and the flux leakage is minimal, the B-field inside the core due to the first winding alone is given by Equation (7.40) when we replace μ_o with μ_c. Hence,

$$B_1 \approx \frac{\mu_c N_1 I_1}{2\pi\rho_o}.$$

Using Equation (9.41), we find that

$$L_{12} = \frac{1}{I_1 I_2}\int_{V_t} \mathbf{B}_1\cdot\mathbf{H}_2 dv \approx \mu_c\left(\frac{N_1}{2\pi\rho_o}\right)\left(\frac{N_2}{2\pi\rho_o}\right)V_t \qquad (\rho_o \gg a),$$

where $V_t \approx (\pi a^2)(2\pi\rho_o)$ is the volume of the toroid. Thus,

$$L_{12} = L_{21} = \frac{\mu_c N_1 N_2 a^2}{2\rho_o} \qquad (\rho_o \gg a),$$

9-3-5 INTERNAL AND EXTERNAL INDUCTANCE

So far in our discussion of inductance, we have considered only circuits and devices where the current flows either in very thin wires or in sheets. This allowed us to consider only the flux linkages that occur outside the current distributions. However, when current flows within conductors with finite cross sections, magnetic fields exist within the wires, and these fields contribute to the total magnetic flux that links the current. This is depicted in Figure 9-21, which shows a current loop with a finite cross section.

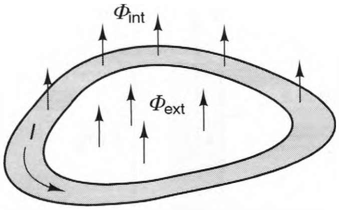


Figure 9-21 A volumetric current with internal and external flux.

The flux that passes through the current-carrying conductors is called the internal flux Φ_{int}, and the flux that passes through the remainder of the circuit is called the external flux Φ_{ext}. We can consider the total self-inductance of this circuit as the sum of two components,

$$L = L_{\text{ext}} + L_{\text{int}}, \tag{9.42}$$

where the external inductance L_{ext} is due to flux linkages outside the current and the internal inductance L_{int} is due to flux linkages inside the current.

Although it is possible in principle to calculate the internal inductance by using the flux linkage definition of inductance (Equation (9.23)), this is not the easiest way. The reason is that a volumetric current can be thought of as of a collection of filamentary loops, each linked by differing amounts of flux. Thus, the internal flux linkage is difficult to calculate directly. On the other hand, the internal inductance is relatively easy to calculate using the energy formulation. This can be accomplished by dividing the volume integral in Equation (9.40) into internal and external parts. We obtain

$$L = \frac{1}{I^2}\int_V \mathbf{B}\cdot\mathbf{H}dv = \frac{1}{I^2}\int_{V_{\text{ext}}} \mathbf{B}\cdot\mathbf{H}dv + \frac{1}{I^2}\int_{V_{\text{int}}} \mathbf{B}\cdot\mathbf{H}dv,$$

where V_{int} and V_{ext} are the volumes inside and outside the current distribution. The first integral on the right-hand side of this expression is the result of flux linkages outside the current, so the external inductance is

$$L_{\text{ext}} = \frac{1}{I^2}\int_{V_{\text{ext}}} \mathbf{B}\cdot\mathbf{H}dv. \tag{9.43}$$

The second integral accounts for internal flux linkages, so it represents the internal inductance

$$L_{\text{int}} = \frac{1}{I^2}\int_{V_{\text{int}}} \mathbf{B}\cdot\mathbf{H}dv. \tag{9.44}$$

To demonstrate how internal inductance can be calculated using the energy formulation of inductance, let us evaluate the internal inductance per meter of straight wire with radius a, shown in Figure 9-22. If we assume that the current flow is uniform throughout the wire's cross section, the B-field inside the wire is given by Equation (7.33),

$$\mathbf{B} = \frac{\mu\rho I}{2\pi a^2}\hat{\mathbf{a}}_\phi \qquad \rho < a,$$

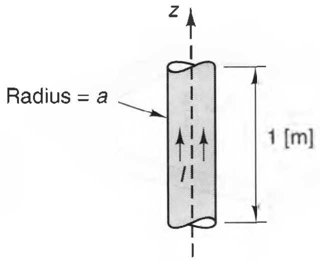


Figure 9-22 A section of a solid wire carrying a uniform current.

where μ is the permeability of the wire. Substituting this into Equation (9.40) and integrating over a 1-meter length of the wire, we obtain

$$L_{\text{int}} = \frac{1}{I^2}\int_{V_{\text{int}}} \mathbf{B}\cdot\mathbf{H}dv = \frac{\mu}{(2\pi a^2)^2}\int_0^1\int_0^{2\pi}\int_0^a \rho^3 d\rho d\phi dz.$$

Evaluating this integral, we find that

$$L_{\text{int}} = \frac{\mu}{8\pi} \qquad [\text{H/m}] \qquad \text{(Solid wire)}, \qquad (9.45)$$

Thus, the internal inductance per unit length of a straight wire is independent of its radius, as long as the current density is uniform throughout the cross section of the wire (as it is at low frequencies). At higher frequencies, however, the majority of the current flows within a thin layer of the wire surface of the wire, reducing the B-field inside the wire. As a result, the internal inductance of a wire becomes progressively smaller at higher frequencies. This subject is discussed further in Chapter 12.

In most cases the internal inductance of a circuit or device is negligible compared to the external inductance. This is particularly true when magnetic materials are present or when the surface bounded by the circuit is much larger than the cross-sectional dimensions of the wires.

9-3-6 DISTRIBUTED INDUCTANCE ON TRANSMISSION LINES

In Chapter 6, we calculated the capacitance characteristics of coaxial and two-wire transmission lines. We will now calculate their inductance characteristics. As we will find in Chapter 11, the capacitance and inductance of a transmission line define its operation over a broad range of frequencies.

9-3-6-1 Coaxial Lines (Cables). Consider the infinite coaxial line shown in Figure 9-23, which consists of a solid inner conductor of radius a and an outer conducting cylinder of inner radius b and outer radius c. Here, the currents carried by the inner and outer conductors are I and $-I$, respectively. If the currents on both conductors are axially symmetric, the B-field in the region $a < \rho < b$ is given by Equation (7.35),

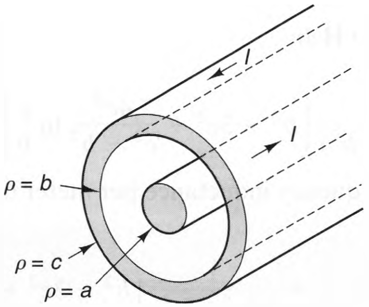


Figure 9-23 A coaxial cable with a solid inner conductor and a solid, finite-width outer conductor.

$$\mathbf{B} = \frac{\mu_o I}{2\pi\rho}\hat{\mathbf{a}}_\phi \qquad a < \rho < b,$$

where we have assumed that both the conductors and dielectrics are nonmagnetic, which is true for most coaxial cables. The B-field outside the outer conductor is zero (since the currents are balanced), so we can calculate the external inductance by integrating $\mathbf{B}\cdot\mathbf{H}$ in region $a < \rho < b$. Using Equation (9.40) and integrating over a unit length, we find that

$$L_{\text{ext}} = \frac{1}{I^2}\int_{V_{\text{ext}}} \mathbf{B}\cdot\mathbf{H}\,dv = \frac{\mu_o}{(2\pi)^2}\int_0^1\int_0^{2\pi}\int_a^b \frac{1}{\rho^2}\rho d\rho d\phi dz,$$

which yields

$$L_{\text{ext}} = \frac{\mu_o}{2\pi}\ln\frac{b}{a} \qquad [\text{H/m}]. \tag{9.46}$$

This expression is valid for all frequencies, since the B-field between the conductors is unaffected by whether or not the current flows uniformly in the conductors, as long as it is rotationally symmetric.

The internal inductance is a function of the magnetic energy contained within the inner and outer conductors. When the current flow is uniform (as it is at low frequencies), the contribution from the inner conductor is given by Equation (9.45),

$$L_{\text{int}_a} = \frac{\mu_o}{8\pi} \qquad [\text{H/m}], \tag{9.47}$$

where we have used the fact that $\mu \approx \mu_o$ for most metals in coaxial cables. To find the internal inductance contribution from the outer conductor, we first note that the B-field inside the outer conductor is given by Equation (7.35),

$$\mathbf{B} = \frac{\mu_o I_o}{2\pi\rho}\frac{(c^2-\rho^2)}{(c^2-b^2)}\hat{\mathbf{a}}_\phi \qquad b < \rho < c.$$

Substituting this into Equation (9.40) and integrating over a 1-meter length yields (after much work!)

$$L_{\text{int}_{bc}} = \frac{1}{I^2}\int_{V_{\text{int}_{bc}}} \mathbf{B}\cdot\mathbf{H}\,dv$$

$$= \frac{\mu_o}{8\pi(c^2-b^2)}\left[b^2 - 3c^2 + \frac{4c^4}{c^2-b^2}\ln\frac{c}{b}\right] \qquad [\text{H/m}]. \qquad (9.48)$$

Thus, the total low-frequency inductance per meter of a coaxial cable is

$$L_{\substack{\text{low}\\\text{freq}}} = \frac{\mu_o}{2\pi}\left[\ln\frac{b}{a} + \frac{1}{4} + \frac{1}{4(c^2-b^2)}\left(b^2 - 3c^2 + \frac{4c^4}{c^2-b^2}\ln\frac{c}{b}\right)\right] \qquad (9.49)$$

At radio frequencies and above, the B-field inside both conductors is negligible, leaving only the external inductance (i.e., the first logarithmic term).

Example 9-10

Calculate the low- and high-frequency inductance per meter of RG-58U coaxial cable, which has a solid inner conductor with radius 0.406 [mm] and a thin, braided outer conductor with mean radius 1.553 [mm]. Assume that the braided outer conductor can be approximated by a solid cylinder of negligible thickness.

Solution:

Since the outer conductor has negligible thickness, Equation (9.48) yields $L_{\text{int}_{bc}} \approx 0$. Hence, the total internal inductance of the cable is given by Equation (9.47),

$$L_{\text{int}} = \frac{\mu_o}{8\pi} = .050 \qquad [\mu\text{H/m}].$$

From Equation (9.46), the external inductance is

$$L_{\text{ext}} = \frac{\mu_o}{2\pi}\ln\left[\frac{1.553}{0.406}\right] = 0.268 \qquad [\mu\text{H/m}].$$

Thus, the total dc inductance per meter is

$$L_{\substack{\text{low}\\\text{freq}}} = L_{\text{int}} + L_{\text{ext}} = 0.318 \qquad [\mu\text{H/m}].$$

At frequencies above a few hundred hertz or so, the internal inductance becomes negligible. Hence,

$$L_{\substack{\text{high}\\\text{freq}}} = 0.268 \qquad [\mu\text{H/m}].$$

9-3-6-2 Two-wire Transmission Lines. A two-wire transmission line is shown in Figure 9-24, which consists of two infinite wires, each of radius a and separated by a distance d. The wires carry oppositely directed, rotationally symmetric currents,[1] each with magnitude I.

[1] This is a good approximation when $d \gg a$. When d is small, however, the current density is higher on the inside surfaces than it is on the outside surfaces. This is analogous to the charge imbalance that occurs when a voltage is impressed between closely spaced wires. (See Section 6-2.3).

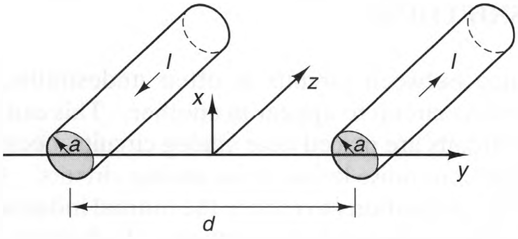


Figure 9-24 A two-wire transmission line.

The total B-field between the wires is the sum of the fields generated by each infinite wire alone. Using Equation (7.33), we find that the B-field along the yz-plane between the wires is

$$\mathbf{B} = \frac{\mu_o I}{2\pi(y + d/2)}\hat{\mathbf{a}}_x + \frac{\mu_o I}{2\pi(d/2 - y)}\hat{\mathbf{a}}_x.$$

The current circulates around the flux between the conductors once, so the external flux-linkage per meter, Λ_{ext}, is simply the net flux per meter between the conductors:

$$\Lambda_{\text{ext}} = \int_{-d/2+a}^{d/2-a} B_x dy = \frac{\mu_o I}{2\pi}\int_{-d/2+a}^{d/2-a}\left(\frac{1}{y + (d/2)} + \frac{1}{(d/2) - y}\right)dy \qquad [\text{Wb/m}]$$

$$= \frac{\mu_o I}{\pi}\ln\left(\frac{d-a}{a}\right) \approx \frac{\mu_o I}{\pi}\ln\left(\frac{d}{a}\right) \qquad [\text{H/m}] \qquad (d \gg a).$$

Substituting this expression for Λ_{ext} into Equation (9.23), we obtain

$$L_{\text{ext}} = \frac{\Lambda_{\text{ext}}}{I} = \frac{\mu_o}{\pi}\ln\left(\frac{d}{a}\right) \qquad [\text{H/m}] \qquad (d \gg a). \qquad (9.50)$$

When $d \gg a$, the B-field inside each conductor is nearly the same as that generated by an isolated conductor. Thus, the total internal inductance per meter at low frequencies is twice that of a single conductor, so that

$$L_{\text{int}} = 2 \times \frac{\mu_o}{8\pi} = \frac{\mu_o}{4\pi} \qquad [\text{H/m}] \qquad (d \gg a),$$

and the total low-frequency inductance of a two-wire line is

$$L_{\substack{\text{low}\\\text{freq}}} = \frac{\mu_o}{\pi}\left(\frac{1}{4} + \ln\frac{d}{a}\right) \qquad [\text{H/m}]. \qquad (9.51)$$

At radio frequencies and above, the flux in the conductors is negligible, so only the logarithmic term remains.

9-3-7 MAGNETIC SHIELDING

The mutual inductance between circuits is often undesirable, since it can cause unwanted signals from one circuit to appear in another. This can be particularly troublesome when digital circuits are placed near analog circuits, because high-level digital currents can cause significant noise levels in the analog circuits. As a result, it is often necessary to take special precautions to reduce the mutual inductance between circuits by reducing the mutual flux linkages between them. Techniques that accomplish this are examples of ***magnetic shielding***.

One magnetic-shielding technique involves placing a circuit or component inside an enclosure made of high-permeability material that diverts the B-fields of adjacent circuits around the shielded circuit. Figure 9-25 shows how the B-field streamlines are diverted around the interior of a spherical shell of high-permeability material (μ_r = 1,000) when it is placed in a uniform B-field.[2] Nearly all of magnetic field lines are diverted around the shielded region, thus reducing the field in that region by a factor of roughly $(\mu_r)^{-1}$. This shielding technique can be very effective, but is often difficult to implement in microelectronic circuits, as magnetic materials tend to be bulkier and less malleable than conductors. Also, since the conductivities of magnetic materials are at best poor, they are not well suited for situations where they must conduct current (as in the case of the outer conductor of a coaxial cable).

Although it is tempting to think that the Faraday shield technique used for electric shielding is also effective for magnetic shielding, such a line of reasoning often leads to poor designs. This is because magnetic fields are *not* directly affected by good conductors such as copper, since their relative permeabilities are nearly unity. A grounded Faraday shield is not a surface of constant magnetic potential. Hence, it cannot, by itself, provide any magnetic shielding. If this be true, the reader may well ask: "Why, then, are coaxial cables often used to provide both electric *and* magnetic shielding?" The answer is that the outer conductor of a coaxial cable shields the outside region from magnetic fields when it passes a current that is exactly equal and opposite to the current on the inner conductor. We saw this in Chapter 7, where we showed that the B-field produced by balanced coaxial currents is zero outside the outer current. From that result, we can state the following principle:

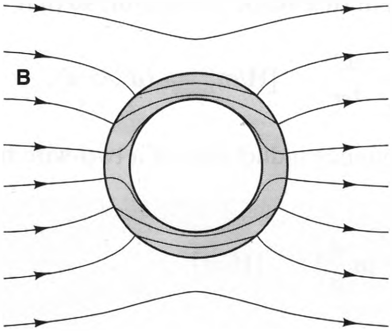


Figure 9-25 An example of magnetic shielding, showing a cross-sectional view of the B-field streamlines when a hollow sphere with μ_r = 1000 is immersed in a uniform magnetic field.

[2] This plot was obtained using exact expressions for the scalar magnetic potentials. (see J. D. Jackson, *Classical Electrodynamics*, 2d ed. (1975), pp. 199–201.) John Wiley & Sons, New York.

A circuit can be shielded from the magnetic field generated by an offending current element by canceling its field with the field of an opposing current element, placed close to the offending current, or, better yet, surrounding it.

The outer conductor of a coaxial cable can provide magnetic shielding, not because it is a constant electric potential surface, but rather because it can pass a current whose field cancels the field generated by the inner conductor. As a result, the magnetic shielding effectiveness of a coaxial cable has *nothing* to do with whether or not it is grounded, but rather on whether it carries an equal and opposite current. Thus, it is very important that the currents on coaxial cables be balanced, which can be harder to accomplish than one might think. Figure 9-26a shows a simple configuration consisting of two chassis connected by a coaxial cable. Both chassis are connected to a common ground.

Even though this circuit consists of only one source and one load, there are two return paths for the current—the cable shield and the ground connection. The amount of current ΔI that flows on the ground loop is a function of the net impedance of this path (including the wire inductances). Because of the ground loop, a net current $\Delta I = I_2 - I_1$, called a ***common-mode current*** flows on the cable. This current produces a nonzero B-field outside the cable that can couple with other circuits or devices. One way to fix the problem is removing one of the chassis grounds, which removes the ground path (often called a ***ground loop***). This fix is attractive from a shielding effectiveness standpoint, but may pose a safety hazard or create unacceptable signal voltages between the chassis. Another technique is to pass the coaxial cable through a ferrite core, as shown in Figure 9-26b. The core presents a large inductive impedance to the common-mode current, resulting in a balanced current (often called a ***differential-mode current***). Such a device is called a ***common-mode choke***.

There are many situations in which coaxial cables are not practical, such as the wiring between lumped elements on printed circuit boards. Here, it is much less expensive to use standard printed circuit board traces. Even with these, good levels of magnetic shielding can still be attained with careful design. One of the most obvious methods is to keep the length-to-area ratio of the circuits as high as possible. This means keeping the current return paths (i.e., ground currents) as close to the feed paths

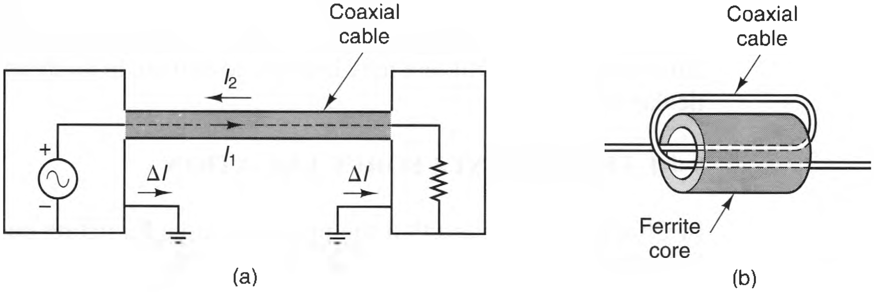


Figure 9-26 a) Two chassis connected by a coaxial cable. b) A common-mode choke.

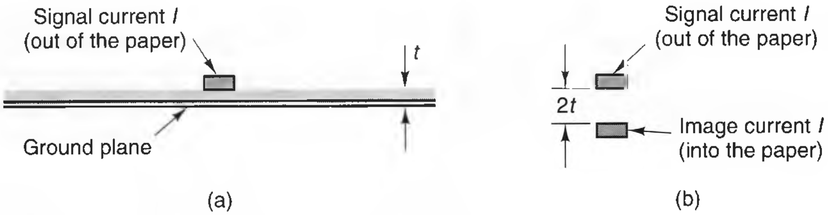


Figure 9-27 A method of reducing magnetic field couplings on printed circuit boards. a) Printed circuit configuration. b) Equivalent geometry.

as possible, thus keeping the area covered by the circuit to a minimum. This is best accomplished by distributing ground traces liberally throughout the board.

Another common technique for reducing magnetic coupling on printed circuit boards is to use multilayer boards. This allows ground layers to be placed directly underneath signal layers. Figure 9-27 depicts a trace on a printed circuit board of thickness t that carries a current I which flows out of the paper. For reasons that will become clear in Chapter 10, the trace and the ground plane form a transmission line, and a current is induced in the ground plane that flows in the opposite direction. Above the board, this current can be modeled as an image current I directed into the paper along an image trace, located a distance $2t$ below the signal trace. This image current generates a B-field that nearly cancels the B-field generated by the signal trace, since the signal and image currents are opposite and very close together.

9-4 Magnetic Forces and Torques

Magnetic forces are used in a large number of electromechanical devices, including motors, relays, electromagnets, and loudspeakers. Although electrical forces can produce similar effects, they are usually orders of magnitude smaller. There are two reasons for this. First, magnetic materials (particularly ferromagnetic materials) can greatly enhance the magnetic fields produced by free currents. Second, low-frequency magnetic fields are produced by currents, whereas low frequency electric fields are produced by charge distributions. It is usually easier (and safer) to create large currents, as opposed to large charges, so magnetic force effects are usually preferred.

We will discuss magnetic forces from three perspectives: the Lorentz force equation, the magnetic energy stored by the elements, and the mutual inductance between the elements. Each perspective is capable of predicting the magnetic force on a given circuit or element, but one may hold an advantage in analysis over the others for a particular system.

9-4-1 THE LORENTZ FORCE EQUATION

In Chapter 3, we found that the magnetic force $\mathbf{F}_m$ acting on a line-current segment C is given by

$$\mathbf{F}_m = \int_C I d\boldsymbol{\ell} \times \mathbf{B}. \tag{9.52}$$

If a current-carrying circuit is constrained to rotate, the net torque $\mathbf{T}$ at an arbitrary point O is given by

$$\mathbf{T}_m = \int_C \mathbf{R} \times \mathbf{dF}_m, \tag{9.53}$$

where $\mathbf{R}$ is a vector that points from O to each point at which the magnetic force $\mathbf{dF}_m$ is applied. If we substitute $I\mathbf{d\ell} \times \mathbf{B}$ for $\mathbf{dF}_m$, Equation (9.53) becomes

$$\mathbf{T}_m = \int_C \mathbf{R} \times (I\mathbf{d\ell} \times \mathbf{B}), \tag{9.54}$$

Also, we can find the component of $\mathbf{T}_m$ along a particular axis of rotation simply by taking the dot product of $\mathbf{T}$ with the unit vector $\hat{\mathbf{a}}_n$ that points along this axis:

$$(\mathbf{T})_n = \hat{\mathbf{a}}_n \cdot \mathbf{T}_m. \tag{9.55}$$

Example 9-11

Figure 9-28 shows a simple dc motor, consisting of a single rectangular loop that rotates between two ferromagnetic pole pieces. The sliding contacts change the current direction every half revolution, so that the current in the rightmost side of the loop is always in the $+z$-direction. If the field current I_f generates a magnetic field $\mathbf{B} = B_0\hat{\mathbf{a}}_y$ in the gap, find the torque T_z acting on the loop along the rotational axis.

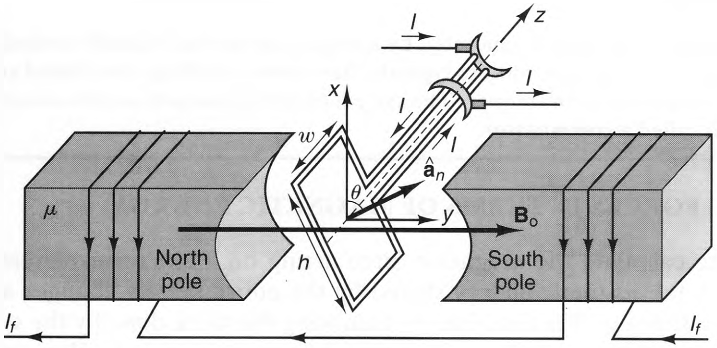


Figure 9-28 A simple electric motor.

Solution:

The contributions to the torque T_z from the front and back portions of the loop are zero. This is because the force $I\mathbf{d\ell} \times \mathbf{B}$ acting on each differential segment of these conductors is in the same direction as the axis of rotation (the z-axis). Thus, we need only calculate the torque contributions on the portions of the loop where the current flows parallel to the z-axis. Using Equation (9.44), and for the moment restricting θ to the range $-90° < \theta < 90°$ (so that the current direction is constant), we find that the contributions to the torque at the point (0,0,0) due to each differential current element on the left and right portions of the loop are given by

$$\mathbf{dT}_{\substack{\text{left}\\\text{right}}} = \mathbf{R} \times I\mathbf{d\ell} \times \mathbf{B},$$

where

$$\mathbf{R} = \pm\frac{h}{2}\sin\theta\,\hat{\mathbf{a}}_x \mp \frac{h}{2}\cos\theta\,\hat{\mathbf{a}}_y + z\,\hat{\mathbf{a}}_z,$$

$$I\,\mathbf{d\ell} = \mp I\,dz\,\hat{\mathbf{a}}_z.$$

Here, the upper and lower signs correspond to the left and right portions of the loop, respectively. Substituting, we have

$$\mathbf{dT}_{\substack{\text{left}\\\text{right}}} = \left(\pm\frac{h}{2}\sin\theta\,\hat{\mathbf{a}}_x \mp \frac{h}{2}\cos\theta\,\hat{\mathbf{a}}_y + z\hat{\mathbf{a}}_z\right)\mathbf{x}\;[\mp I\,dz\,\hat{\mathbf{a}}_z\,\mathbf{x}\,B_o\hat{\mathbf{a}}_y]$$

$$= \frac{h}{2}IB_o\cos\theta\,\hat{\mathbf{a}}_z\,dz \pm IB_o z dz\,\hat{\mathbf{a}}_y.$$

Integrating these contributions from $z = -(w/2)$to $z = w/2$, we obtain

$$\mathbf{T}_{\substack{\text{left}\\\text{right}}} = \int_{-\frac{w}{2}}^{\frac{w}{2}} \mathbf{dT}_{\substack{\text{left}\\\text{right}}} = \frac{h}{2}IwB_o\cos\theta\,\hat{\mathbf{a}}_z.$$

Summing the left and right torque contributions, we find that

$$T_z = \hat{\mathbf{a}}_x \bullet (\mathbf{T}_{\text{left}} + \mathbf{T}_{\text{right}})$$

$$= hwIB_o\cos\theta \qquad -90^\circ < \theta < 90^\circ.$$

This torque tends to make the loop rotate around the z-axis in a right-handed sense. When the loop angle θ exceeds 90°, the current reverses, keeping the torque positive, so the torque can be expressed for all angles as

$$T_z = hwIB_o|\cos\theta|.$$

The $|\cos\theta|$ dependence of the torque generated by a single-loop motor is usually undesirable, since it is not uniform. To avoid this, motors typically have many windings distributed around an ***armature*** to produce a more even torque. The system of sliding contacts used to reverse the current in each loop is called a ***commutator***.

9-4-2 MAGNETIC FORCES IN TERMS OF MAGNETIC ENERGY

It is also possible to calculate the magnetic force acting on a circuit or element by observing how the total magnetic energy stored by the entire system changes as the component is moved slightly. This is done by balancing the work done by the system during this movement with the net change in the stored magnetic energy. We next will derive several equivalent formulas for determining these forces.

9-4-2-1 The Constant-Current Method. Consider the system shown in Figure 9-29, which consists of $N = 3$ rigid circuits and a magnetic core that can slide in and out of its windings. In general, each of these elements will experience a force due to the interactions of their fields. To determine the force $\mathbf{F}_m$ acting on any of the elements, say, circuit #3, we will allow it to move through a directed distance $\mathbf{d\ell}$ in response to the force while the other elements are held in fixed positions.

The work done by the system on the element is $dW = \mathbf{F}_m \bullet \mathbf{d\ell}$ To maintain all the currents constant during this motion, voltage sources must be present in each cir-

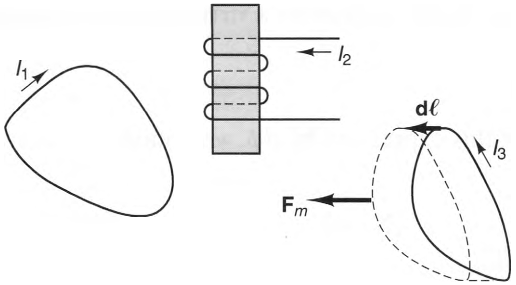


Figure 9-29 Geometry used to find the forces between an arbitrary collection of currents and magnetic materials.

cuit to counter the emf's induced by the time-varying fluxes. Thus, the work dW expended by the system on the component equals the difference between the work dW_s supplied by the sources to maintain the currents and the net change dW_m in the magnetic energy stored by the system. That is,

$$dW = \mathbf{F}_m \cdot \mathbf{d}\ell = dW_s - dW_m. \tag{9.56}$$

The voltage that must be applied to the i^{th} circuit to counter the *emf* generated by the time-varying flux linkage is $d\Lambda_i/dt$, where the positive direction of Λ_i is determined by the positive direction of I_i and the right-hand rule. Thus, the additional energy that must be supplied to maintain the current in the i^{th} circuit constant is

$$\begin{aligned} dW_{s_i} &= \int I_i \frac{d\Lambda_i}{dt}\,dt = I_i \int \frac{d\Lambda_i}{dt}\,dt \\ &= I_i d\Lambda_i = d(I_i\Lambda_i). \end{aligned} \tag{9.57}$$

Summing the contributions from each circuit yields

$$dW_s = d\left(\sum_{i=1}^{N} I_i\Lambda_i\right).$$

But from Equation (9.32), the sum in the parentheses is simply two times the total stored energy W_m. Thus, we can write

$$dW_s = 2dW_m. \tag{9.58}$$

Substituting Equation (9.58) into Equation (9.56), we obtain

$$\mathbf{F}_m \cdot \mathbf{d}\ell = dW_m.$$

Finally, remembering from the properties of the gradient operation that $dW_m = \nabla W_m \cdot \mathbf{d}\ell$, we can conclude that

$$\mathbf{F}_m = \nabla W_m \Big|_{\substack{\text{All } I\text{'s} \\ \text{constant}}} \qquad [\text{N}]. \tag{9.59}$$

For an element that is constrained to rotate about an axis, we can calculate the net torque by noting the changes in the energy stored by the system as it rotates. For instance, the work done by the system for a virtual angular displacement around the z-axis is

$$dW = (\mathbf{T}_m)_z d\phi.$$

Dividing both sides of this expression by $d\phi$, we obtain

$$(\mathbf{T}_m)_z = \left.\frac{\partial W_m}{\partial \phi}\right|_{\substack{\text{All } I\text{'s} \\ \text{constant}}} \qquad [\mathrm{N \cdot m}]. \tag{9.60}$$

Example 9-12

Figure 9-30 shows a U-shaped electromagnet that is lifting an iron bar. The electromagnet consists of N turns of wire around a core of cross section S and constant permeability μ_c. Calculate the lifting force exerted on the bar, assuming that the bar has permeability μ_b and cross section S_b. Assume also that gap has the same cross-sectional area as the core.

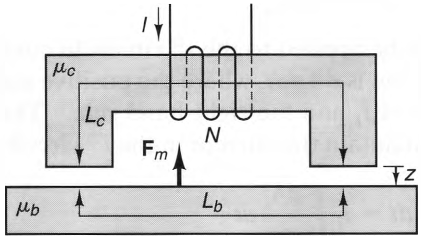


Figure 9-30 An electromagnet lifting an iron bar.

Solution:

If we assume that the bar is slightly separated from the core by air gaps of length z, the flux Φ passing through the core and gaps is obtained by requiring that the magnetic voltage drops around the entire path equal the magnetomotive force of the windings. (See Equations (8.50) and (8.53)). This yields the expression

$$\Phi = \frac{NI}{\mathcal{R}_c + \mathcal{R}_b + 2\mathcal{R}_g}, \tag{9.61}$$

where $\mathcal{R}_c = L_c/(\mu_c S)$, $\mathcal{R}_b = L_b/(\mu_b S_b)$, and $\mathcal{R}_g = z/(\mu_o S)$ are the reluctances of the core, bar, and air gaps, respectively. Also, $B = \Phi/S$ in the core and air gaps, whereas $B = \Phi/S_b$ in the bar. Hence, we can express the magnetic energy W_m of the system as

$$W_m = \frac{1}{2}\int_V \mathbf{B} \cdot \mathbf{H} dv = \frac{\Phi^2}{2}\left[\frac{L_c S}{\mu_c S^2} + \frac{L_b S_b}{\mu_b S_b^2} + 2\frac{zS}{\mu_o S^2}\right]$$

$$= \frac{\Phi^2}{2}[\mathcal{R}_c + \mathcal{R}_b + 2\mathcal{R}_g]. \tag{9.62}$$

Substituting Equation (9.61) into Equation (9.62) yields

$$W_m = \frac{1}{2}\frac{(NI)^2}{(\mathcal{R}_c + \mathcal{R}_b + 2\mathcal{R}_g)} = \frac{1}{2}\frac{(NI)^2}{\left(\mathcal{R}_c + \mathcal{R}_b + 2\dfrac{z}{\mu_o S}\right)}. \tag{9.63}$$

The lifting force $\mathbf{F}_m$ acting on the bar can now be found by substituting Equation (9.63) into Equation (9.59):

$$\mathbf{F}_m = \nabla W_m \Big|_{\substack{\text{All } I\text{'s} \\ \text{constant}}} = \frac{\partial W_m}{\partial z}\hat{\mathbf{a}}_z \Big|_{\substack{\text{All } I\text{'s} \\ \text{constant}}} = -\frac{(NI)^2}{\left(\mathcal{R}_c + \mathcal{R}_b + 2\dfrac{z}{\mu_0 S}\right)^2}\frac{1}{\mu_0 S}\hat{\mathbf{a}}_z.$$

Substituting Equation (9.61) into this expression, we can also express $\mathbf{F}_m$ in terms of the flux Φ:

$$\mathbf{F}_m = -\frac{\Phi^2}{\mu_0 S}\hat{\mathbf{a}}_z. \tag{9.64}$$

Example 9-13

Figure 9-31 shows a solenoid with a partially inserted core. The solenoid has length L, cross-sectional area S, and N turns of wire that are equally spaced and carry a current I. If the rod has relative permeability μ_r and can slide inside the solenoid with negligible friction, find the force on the core.

Solution:

This problem may at first seem difficult, because **B** fringes both near the ends of the solenoid and the core. However, if the ends of the core are deep within the solenoid and air regions, respectively, the shapes of the B-field streamlines will stay the same as the core moves, except that they will shift to the left or right within the core. Thus, the energy stored in the regions where **B** fringes will not change with the depth of penetration.

What will change as the core moves, however, are the percentages of the solenoid that are filled with air and the core. Deep within the air and core regions, **B** is given by

$$\mathbf{B}_{\text{air}} = \frac{\mu_0 NI}{L}\hat{\mathbf{a}}_z,$$

and

$$\mathbf{B}_{\text{core}} = \frac{\mu_0\mu_r NI}{L}\hat{\mathbf{a}}_z,$$

respectively. The volume inside the solenoid that is filled by the core increases linearly with the core penetration distance Δz. Conversely, the volume filled by air decreases by the same amount. Remembering that the magnetic energy density is $1/2\,\mathbf{B}\bullet\mathbf{H}[\text{J/m}^3]$, the net change in the magnetic energy ΔW_m as a result of the displacement of the rod is

$$\Delta W_m = (\mathbf{B}_{\text{core}}\bullet\mathbf{H}_{\text{core}} - \mathbf{B}_{\text{air}}\bullet\mathbf{H}_{\text{air}})\Delta z\, S$$

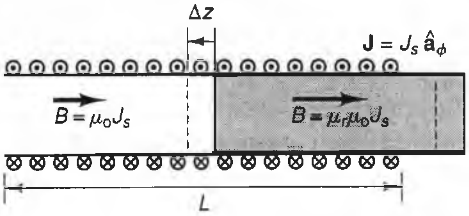


Figure 9-31 A solenoid with a movable magnetic core.

$$= \left(\frac{NI}{L}\right)^2 (\mu_0\mu_r - \mu_0)\Delta z S.$$

Dividing both sides by Δz, we find that

$$\frac{\Delta W_m}{\Delta z} = \frac{dW_m}{dz} = \left(\frac{NI}{L}\right)^2 (\mu_0\mu_r - \mu_0)S.$$

The force acting on the core is found by substituting this expression into Equation (9.59), giving

$$\mathbf{F}_m = \nabla W_m \Big|_{\substack{\text{All } I\text{'s} \\ \text{constant}}} = \frac{dW_m}{dz}\hat{\mathbf{a}}_z = \mu_0\left(\frac{NI}{L}\right)^2 (\mu_r - 1)S\,\hat{\mathbf{a}}_z.$$

As can be seen, the force on the rod is attractive when $\mu_r > 1$ and repulsive when $\mu_r < 1$ (as is the case in superconducting materials). Notice also that this force does not change direction when the current direction is reversed.

9-4-2-2 The Constant-Flux Method. Returning to the system shown in Figure 9-29, let us once again allow the force $\mathbf{F}_m$ to move the element shown a differential distance $\mathbf{d\ell}$ this time under the condition that the flux linking each circuit is held constant while the element is in motion. To do so, the current in each circuit must be adjusted during the displacement in order to keep the flux linkages in each circuit constant. However, no emf's are induced in any of the circuits (since the flux linkages are all constant), so no additional energy is needed to adjust these currents. Thus, the work done on the displaced element equals the net decrease in the magnetic energy:

$$\mathbf{F}_m \cdot \mathbf{d\ell} = -dW_m. \tag{9.65}$$

From the properties of the gradient operation, $dW_m = \nabla W_m \cdot \mathbf{d\ell}$. This means that we can write $\mathbf{F}_m$ in the form

$$\mathbf{F}_m = -\nabla W_m \Big|_{\substack{\text{All } \Phi\text{'s} \\ \text{constant}}} \quad [\text{N}]. \tag{9.66}$$

If the element (or circuit) is constrained to rotate about the z-axis, the torque about this axis is

$$(\mathbf{T}_m)_z = -\frac{\partial W_m}{\partial \phi}\Big|_{\substack{\text{All } \Phi\text{'s} \\ \text{constant}}} \quad [\text{N}\cdot\text{m}]. \tag{9.67}$$

An attractive characteristic of the constant-flux method is that it is often well suited to systems that contain nonlinear ferromagnetic components, since it is not necessary to know how the flux in these elements changes with the impressed current. This is demonstrated in the following example.

Example 9-14

Calculate the lifting force of the electromagnet in Example 9-12 using the constant-flux method. Assume that the core and bar are ferromagnetic materials with nonlinear permeabilities.

Solution:

In Example 9-12, we showed that the magnetic energy W_m stored in the system is given by

$$W_m = \frac{\Phi^2}{2}[\mathcal{R}_c + \mathcal{R}_b + 2\mathcal{R}_g],$$

where $\mathcal{R}_c$, $\mathcal{R}_b$, and $\mathcal{R}_g$ are the reluctances of the core, bar, and gaps, respectively. Using Equation (9.66), we can write

$$\mathbf{F}_m = -\nabla W_m\Big|_{\substack{\text{All } \Phi\text{'s}\\ \text{constant}}} = -\frac{\Phi^2}{2}\nabla[\mathcal{R}_c + \mathcal{R}_b + 2\mathcal{R}_g].$$

However, $\mathcal{R}_c$ and $\mathcal{R}_b$ are both constants, and $\mathcal{R}_g = \frac{z}{\mu_o S}$, so we obtain

$$\mathbf{F}_m = -\frac{\Phi^2}{\mu_o S}\nabla(z) = -\frac{\Phi^2}{\mu_o S}\frac{\partial z}{\partial z}\hat{\mathbf{a}}_z = -\frac{\Phi^2}{\mu_o S}\hat{\mathbf{a}}_z.$$

This is the same expression that was derived using the constant-current method and is valid for both linear and nonlinear materials, as long as the flux Φ in the core is known. If all the materials are linear, Φ can be determined by using Equation (9.61) in Example 9-11. However, if any of the materials are nonlinear, their magnetization curves (such as the one shown in Figure 8-21) must be specified in order to determine Φ.

9-4-2-3 The Inductance Method. Another formulation of the forces in magnetic systems can be obtained by starting with the constant-current formulation, given by Equation (9.59):

$$\mathbf{F}_m = \nabla W_m\Big|_{\substack{\text{All } I\text{'s}\\ \text{constant}}}$$

If the magnetic materials in the system are all linear, we can use Equation (9.30) to express the magnetic energy in terms of the self- and mutual inductances of the system:

$$\mathbf{F}_m = \nabla W_m\Big|_{\substack{\text{All } I\text{'s}\\ \text{constant}}} = \frac{1}{2}\nabla\left[\sum_{i=1}^{N}\sum_{j=1}^{N} I_i I_j L_{ij}\right].$$

Since all the currents of the system are maintained constant during the virtual displacement, this yields

$$\mathbf{F}_m = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} I_i I_j \nabla L_{ji}. \qquad (9.68)$$

If the component is constrained to rotate about an axis (say, the z-axis), the torque about that axis is given by

$$(\mathbf{T}_m)_z = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} I_j I_i \frac{\partial L_{ji}}{\partial \phi}. \qquad (9.69)$$

An advantage of formulating the magnetic forces and torques in terms of self- and mutual inductances is that it is not necessary to calculate the self-inductances of the circuits if they do not change during the virtual displacement. When that is the case, the force or torque is due only to the mutual inductances. This is demonstrated in the following example.

Example 9-15

Calculate the torque exerted on the DC motor loop discussed in Example 9-11, this time using the inductance method.

Solution:

This system consists of two circuits: the two field windings and the rotating loop, which carry the currents I_f and I, respectively. Since the self-inductances of these circuits do not change when the loop is allowed to rotate, we need to consider only the mutual inductances.

The B-field in the gap can be expressed as $B_0 = kI_f$, where the value of the constant k is a function of the number of turns and the reluctance of the flux path. The flux Λ_{21} linking the loop due to the windings is $\Lambda_{21} = kI_fS\ |\sin\theta|$, where $S = wh$ is the area of the loop. From Equation (9.23), the mutual inductances are given by

$$L_{21} = L_{12} = \frac{\Lambda_{21}}{I_f} = k\,w\,h|\sin\theta|.$$

Substituting these values into Equation (9.69), we find that the torque acting on the loop is

$$\begin{aligned}(\mathbf{T}_m)_z &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} I_i I_j \frac{\partial L_{ji}}{\partial \phi} = \frac{1}{2} II_f \frac{\partial L_{12}}{\partial \phi} + \frac{1}{2} II_f \frac{\partial L_{12}}{\partial \phi}\\ &= I\,I_f k\,w\,h\ |\cos\theta|.\end{aligned}$$

Finally, if we substitute $B_0 = kI_f$ into this expression, we obtain the same result as was obtained in Example 9-11.

9-5 Summation

Most of the discussion in this chapter has been an outgrowth of Faraday's law of induction. This law is one of Maxwell's equations and describes one of the ways in which electric and magnetic fields interact. In addition, it is useful in its own right for describing a variety of systems and applications of engineering significance that depend on the electric fields produced by magnetic fields.

Faraday's law has also enabled us to "turn the corner" in our discussion of static electric and magnetic systems into a more general discussion of time-varying systems. We will maintain this direction in Chapter 10, where we will discuss Maxwell's final postulate electromagnetic theory and the time-varying form of Maxwell's equations.

Problems

9-1 Calculate $i(t)$ if the loop shown in Figure P9-1 is immersed in a uniform time-varying magnetic field $\mathbf{B} = 2t$ [T] $\hat{\mathbf{a}}_x$, where t is measured in seconds and $\hat{\mathbf{a}}_x$ is directed out of the page. Assume that the self-inductance of the loop is negligible.

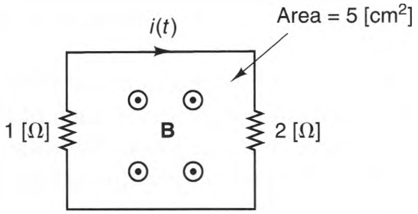


Figure P9-1

9-2 Find the voltage V measured by the meter shown in Figure P9-2, where the B-field is uniform throughout the inner circuit, is zero outside, and varies with time according to

$$\mathbf{B} = B_0 \sin \omega t\, \hat{\mathbf{a}}_n$$

where $\hat{\mathbf{a}}_n$ is directed out of the paper, $B_0 = 0.5$ [T], $\omega = 10$ [rad/s], and the area enclosed by the circuit is 1 [m^2]. Also, assume that the voltage source is $V_s = 10 \cos \omega t$ [V], the meter impedance is infinite, and the self-inductance of the circuit is negligible.

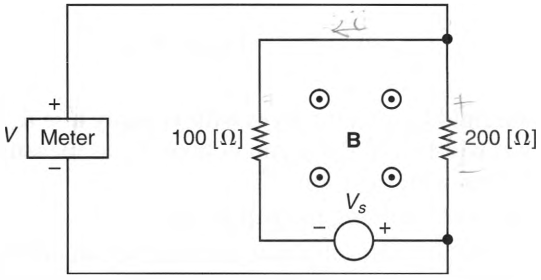


Figure P9-2

9-3 Figure P9-3 shows a two-resistor circuit that is close to a long wire carrying a current $i(t) = 4 \cos(100t)$ [mA]. Calculate $V_{100}(t)$ and $V_{50}(t)$ if the self-inductance of the loop is negligible.

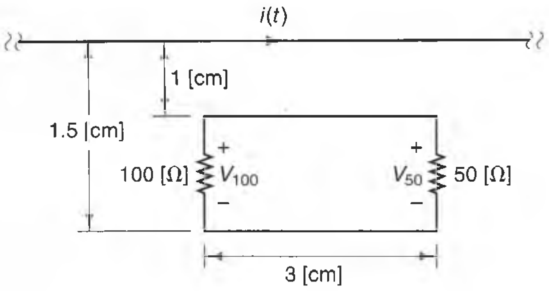


Figure P9-3

9-4 Figures P9-4 a and b show two types of magnetic cores that lie along the z-axis: one solid and the other laminated. The solid core has length L, radius a, and conductivity σ. The laminated core is a collection of N small filamentary cylinders, each with length L and conductivity σ. If $\mathbf{B} = B_o \cos \omega t\ \hat{\mathbf{a}}_z$ is directed upward and the filamentary cylinders fill 90% of the volume of the solid core, find

(a) the average eddy-current power loss in the solid core.

(b) the average eddy-current power loss in each of the filamentary cores.

(c) the average eddy-current power loss in the total laminated core.

Assume that the B-field is uniform throughout both the solid and laminated cores.

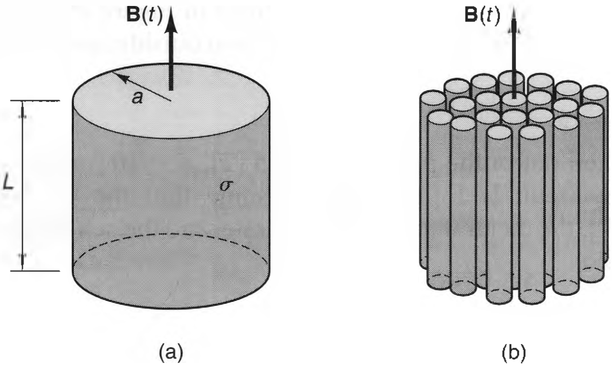


Figure P9-4

9-5 Figure P9-5 shows a rigid loop that moves with velocity $\mathbf{u} = u_o\, \hat{\mathbf{a}}_y$ in a uniform B-field. If the gap has width d, find the gap voltage V_g as measured by

(a) an observer moving with the loop.

(b) a stationary observer watching the gap go by.

9-6 Figure P9-6 shows a rectangular loop that is rotated at an angular frequency of ω [rad/s] in the presence of a uniform B-field of magnitude B. The slip rings are arranged so that the terminals of the rotating loop are always attached to the same points in the stationary circuit. If the loop has self-inductance L, show that the average power dissipated in R is equal to the mechanical work required to rotate the loop.

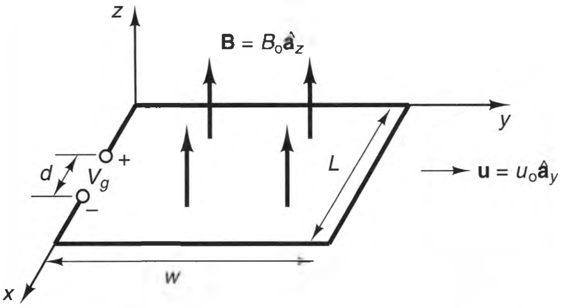

z
$\mathbf{B} = B_o\hat{\mathbf{a}}_z$
y
d
V_g
+
−
L
$\mathbf{u} = u_o\hat{\mathbf{a}}_y$
w
x

Figure P9-5

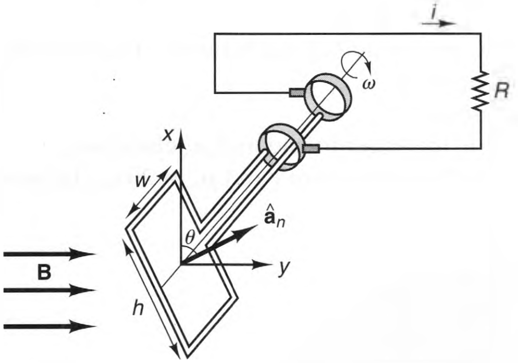


Figure P9-6

9-7 Figure P9-7 shows a square, rigid loop that moves away from an infinite line with current $I = 1{,}000$ [A] at a constant velocity $u = 20$ [cm/s]. If $x = 0$ at $t = 0$ and the loop is 40 [cm] on a side, find $v(t)$ as measured by someone moving with the loop.

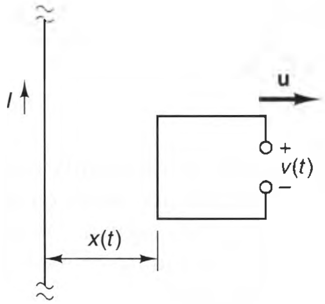


Figure P9-7

9-8 Figure P9-8 shows a loop that rotates about the z-axis at $\omega = 10$ [rad/s]. If $\mathbf{B} = 0.2\ \hat{\mathbf{a}}_y$ [T] and $\phi = 0$ at $t = 0$, find the current i. Assume that the loop inductance is negligible.

9-9 Two concentric current loops lie in the $z = 0$ plane, with radii a and b, respectively. If $a \gg b$, estimate the mutual inductance between these loops if the positive direction of current is the same for both loops.

9-10 Figure P9-10 shows two toroids, one inside the other. The inner solenoid has

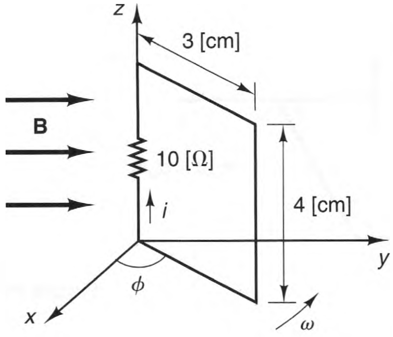


Figure P9-8

2,000 turns, and the outer one has 4,000 turns. Estimate the mutual inductance between these solenoids if

(a) $\mu = \mu_o$ everywhere.

(b) $\mu = 500\mu_o$ inside the inner toroid and μ_o elsewhere.

(c) $\mu = 500\mu_o$ inside the inner toroid and $\mu_o = 200\mu_o$ between the toroids.

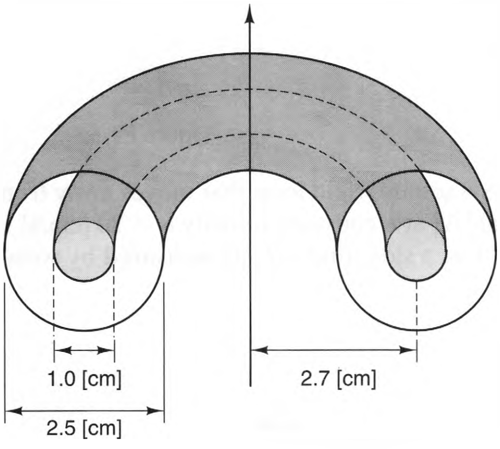


Figure P9-10

9-11 Figure P9-11 shows a high-permeability core with an air gap and two windings. Calculate the mutual inductance between the terminals of the two windings for the positive current directions shown. Assume that the core has a relative permeability $\mu_r = 1{,}000$ and a square cross section with 0.8 [cm] on a side. Also, assume that fringing is negligible.

9-12 Calculate the mutual inductance between the infinite line and the circular loop shown in Figure P9-12. Assume that $d \gg r$.

9-13 Figure P9-13 shows an infinite line and a 1,000-turn square loop that encloses an area of .04 [cm^2]. Both the line and the loop carry currents of 10 amperes, and their polarities are such that their fluxes add inside the loop. Calculate the force acting on the loop, and state whether it is attractive or repulsive, using

(a) the Lorentz force equation.

(b) the inductance method.

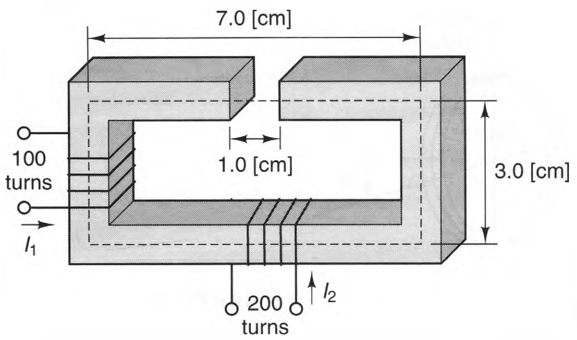


Figure P9-11

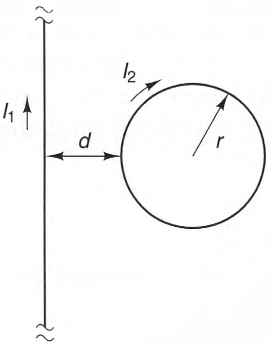


Figure P9-12

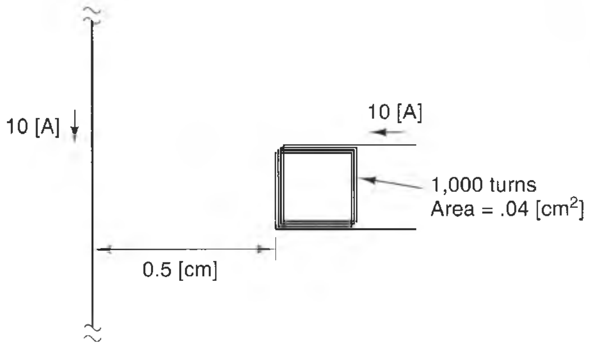


Figure P9-13

9-14 For the magnetic circuit discussed in Problem 9-11, find the attractive force between the pole pieces across the gap when $I_1 = 20$ [A] and $I_2 = 0$.

9-15 Figure P9-15 shows two coils. Coil #1 is stationary and has self inductance $L_{11} = 0.4$ [mH], whereas coil #2 is free to rotate about an axis and has self- inductance $L_{22} = 0.3$ [mH]. If the mutual inductance between the coils is $L_{12} = 0.1 \cos \phi$ [mH],

(a) Find the torque that each coil exerts on the other if $I_1 = 2$ [A] and $I_2 = 1$ [A] when $\phi = 50°$.

(b) If $I_1 = 2 \sin 100t$ [A], find the time-averaged torque acting on coil #2 if the coil is short circuited and rotates at a constant angular velocity that is much less than 100 [rad/s].

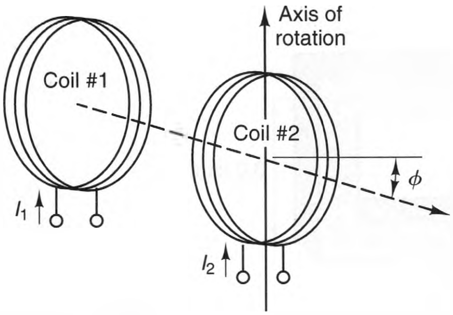


Figure P9-15

9-16 Figure P9-16 shows the cross section of an electromagnet. Here, a coil is wound many times around the center core and has an inductance of 0.6 henry when the plunger is at $x = 0.7$ [cm]. If the reluctance of the core and plunger are negligible compared to the gap reluctance, find the attractive force on the plunger when $x = 0.7$ [cm] and $x = 1.0$ [cm] when the current flowing through the windings is 0.8 ampere.

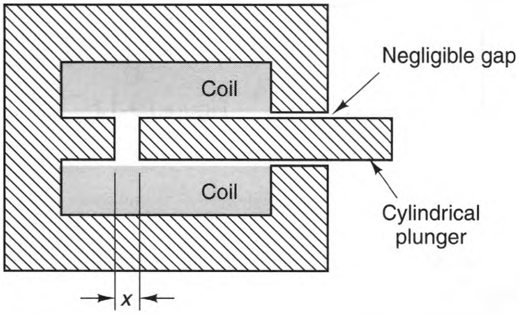


Figure P9-16

10

Time-Varying Electromagnetic Fields

10-1 Introduction

This chapter marks the beginning of a discussion that will last through the remainder of the text—the behavior of high-frequency electromagnetic fields and devices that produce them. This subject is of great importance in electrical engineering, since a large percentage of the systems that electrical engineers design and operate cannot be described by simple circuit analysis alone, or even by the low-frequency field analysis we have presented up to now. For applications like these, a complete description of the time-varying nature of electromagnetic fields is required. This is true even for systems that still "look" like ordinary, lumped-element circuits.

We will start the chapter by discussing the last postulate of electromagnetic theory—displacement current. This will complete our development of Maxwell's equations, which are capable of describing all electromagnetic effects. We will show how Maxwell's equations can be written in a number of forms so that they can be easily applied to different types of situations. Finally, we will end the chapter by showing how standard lumped-component ac circuit analysis follows directly from Maxwell's equations.

10-2 Displacement Current

In Chapter 3, we stated that Maxwell completed the equations that describe electromagnetism simply by adding a new quantity to Ampère's circuital law. Initially, this addition was not based on direct experimental evidence, but rather on Maxwell's belief that light and electromagnetism are, in fact, the same. Maxwell invented the concept of displacement current so that the equations of electromagnetism would also describe the behavior of light waves. The addition did not violate the experimental evidence that was already known about electromagnetism (in particular, Ampère's circuital law) and was later verified by the experiments of Heinrich Hertz in 1888.

Maxwell postulated that Ampère's circuital law, while valid for time-invariant fields, was incomplete for time-varying fields. For fields in free space, Ampère's circuital law reads

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J}_T,$$

where $\mathbf{J}_T$ accounts for the *total* flow of charged particles at a point. One way to show that this expression is incomplete is to take the divergence of both sides, yielding

$$\nabla \cdot \nabla \times \mathbf{B} = \mu_0 \nabla \cdot \mathbf{J}_T.$$

The left-hand side of this expression is identically zero (since the divergence of the curl of any vector is always zero), which leaves

$$\nabla \cdot \mathbf{J}_T = 0. \tag{10.1}$$

On the other hand, the law of charge conservation states that the current density at a point can have a net divergence whenever the charge density at that point is changing with respect to time, i.e.,

$$\nabla \cdot \mathbf{J}_T = -\frac{\partial \rho_{vT}}{\partial t}. \tag{10.2}$$

Comparing Equations (10.1) and (10.2), it is obvious that Ampère's law violates the law of charge conservation at points where the charge density is time varying. Hence, Ampère's circuital law gives an incorrect (or, at best, an incomplete) description of time-varying effects.

Maxwell postulated that, in free space, Ampère's law should be modified to read

$$\nabla \times \mathbf{B} = \mu_0 \left(\mathbf{J}_T + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right). \tag{10.3}$$

This expression is called the time-varying form of Ampère's law, or, simply ***Ampère's law***. The added term, $\epsilon_0 (\partial \mathbf{E}/\partial t)$, is measured in units of amperes per meter squared and is called the ***displacement current density***. In spite of its name, however, this new quantity does *not* represent a flow of charge. Rather, it is present whenever a time-varying charge displacement (or separation) is present.

Even though Maxwell did not introduce the concept of displacement current to force Ampère's circuital law to agree with the law of charge conservation,[1] it is simple to show that it does. Taking the divergence of both sides of Equation (10.3), we obtain

$$\nabla \cdot \mathbf{J}_T + \frac{\partial}{\partial t}\epsilon_o \nabla \cdot \mathbf{E} = 0. \tag{10.4}$$

However, from Gauss' law, we know that in free space, $\nabla \cdot \mathbf{E} = \rho_{vT}/\epsilon_o$, where ρ_{vT} is the total (i.e., free plus bound) charge density. Hence, Equation (10.4) can be written as

$$\nabla \cdot \mathbf{J}_T = -\frac{\partial \rho_{vT}}{\partial t}, \tag{10.5}$$

which is the law of charge conservation.

When materials are present, it is more convenient to express Ampère's law in terms of the magnetic field intensity **H** and the electric flux density **D**, which are related to **B** and **E** by

$$\mathbf{B} = \mu \mathbf{H} = \mu_o(\mathbf{H} + \mathbf{M})$$

and

$$\mathbf{D} = \epsilon \mathbf{E} = \epsilon_o \mathbf{E} + \mathbf{P},$$

where **P** and **M** are the polarization and magnetization, respectively. Substituting these definitions into Equation (10.3), we obtain

$$\nabla \times \mathbf{H} = \mathbf{J}_T + \frac{\partial \mathbf{D}}{\partial t} - \frac{\partial \mathbf{P}}{\partial t} - \nabla \times \mathbf{M}. \tag{10.6}$$

However, according to Equation (8.11), we know that

$$\nabla \times \mathbf{M} = \mathbf{J}_m, \tag{10.7}$$

where $\mathbf{J}_m$ is the magnetization current (defined in Section 8-3-2). Also, from Equation (5.25), we know that the polarization charge density is specified by the polarization **P**:

$$\rho_{vp} = -\nabla \cdot \mathbf{P} \qquad [\mathrm{C/m^3}].$$

Differentiating both sides of this expression with respect to time and using the law of charge conservation, we find that

$$\frac{\partial \mathbf{P}}{\partial t} = \mathbf{J}_p, \tag{10.8}$$

where $\mathbf{J}_p$ is the polarization current, due the movement of polarization charge. Substituting Equations (10.7) and (10.8) into Equation (10.6), we obtain

$$\nabla \times \mathbf{H} = \mathbf{J}_T - \mathbf{J}_m - \mathbf{J}_p + \frac{\partial \mathbf{D}}{\partial t}.$$

Finally, noting that $\mathbf{J}_T = \mathbf{J} + \mathbf{J}_m + \mathbf{J}_p$, we can write this equation in the form

[1] See Paul J. Nahin. *Oliver Heaviside: Sage in Solitude*, (IEEE Press, 1988), New York, Chapter 6, pp. 90–91.

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}, \tag{10.9}$$

where **J** is the free current density and $\partial\mathbf{D}/\partial t$ is the displacement current density.

An integral form of Equation (10.9) can be obtained by taking the dot product of both sides by the differential displacement vector **ds** and integrating over an arbitrary, open surface S, yielding

$$\int_S \nabla \times \mathbf{H} \cdot \mathbf{ds} = \int_S \mathbf{J} \cdot \mathbf{ds} + \int_S \frac{\partial \mathbf{D}}{\partial t} \cdot \mathbf{ds}. \tag{10.10}$$

If S is time invariant, the order of integration and differentiation in the second term on the right-hand side can be interchanged. Also, Stokes's theorem can be used to write the surface integral on the left-hand side as a line integral over the contour C that bounds S in a right-handed sense. In addition, $\int_S \mathbf{J} \cdot \mathbf{ds}$ is the free current I that passes through S. Thus, Equation (10.10) can be written as

$$\oint_C \mathbf{H} \cdot d\boldsymbol{\ell} = I + \frac{\partial}{\partial t} \int_S \mathbf{D} \cdot \mathbf{ds}, \tag{10.11}$$

where C is the closed contour that bounds the open surface S. In this expression, $\frac{\partial}{\partial t} \int_S \mathbf{D} \cdot \mathbf{ds}$ is the displacement current that "flows" out of S in a right-handed sense.

A simple device that illustrates the role of displacement current is the capacitor shown in Figure 10-1. Applying Equation (10.11) to the closed path C' shown in the figure, we note that the surface integration can be performed on any surface that is bounded by this path. Two such surfaces are shown. Here, S_1 is a small disk that intersects the wire, whereas S_2 is shaped so that it passes between the capacitor plates.

If the wire is perfectly conducting, the tangential D-field near it is negligible, so $\mathbf{D} \cdot \mathbf{ds} \approx 0$ everywhere on S_1. Hence, no displacement current flows through S_1, and Ampère's law reads

$$\oint_{C'} \mathbf{H} \cdot d\boldsymbol{\ell} = I. \tag{10.12}$$

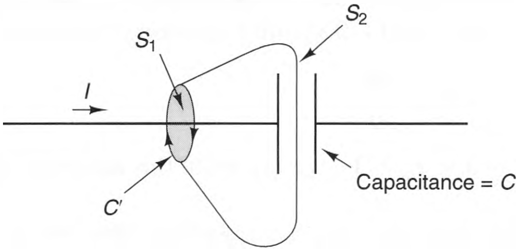


Figure 10-1 Geometry for finding the displacement current between the plates of a capacitor.

Since S_2 passes between the capacitor plates, no conduction current flows through this surface. But a displacement current "flows" through S_2, since $\mathbf{D} \cdot \mathbf{ds} \neq 0$ between the plates. For this surface, Ampère's law reads

$$\oint_{C'} \mathbf{H} \cdot \mathbf{d\ell} = \frac{\partial}{\partial t} \int_{S_2} \mathbf{D} \cdot \mathbf{ds}. \tag{10.13}$$

Comparing Equations (10.12) and (10.13), we conclude that

$$I = \frac{\partial}{\partial t} \int_{S_2} \mathbf{D} \cdot \mathbf{ds}, \tag{10.14}$$

which, in words, states that the conduction current flowing into the capacitor equals the displacement current "flowing" between the plates.

We can derive the voltage–current relationship for a capacitor directly from Equation (10.14) by noting that $\int_{S_2} \mathbf{D} \cdot \mathbf{ds}$ is simply the charge Q on the left-hand plate. If C is the capacitance of the capacitor, then $Q = CV$, and we can write

$$\frac{\partial}{\partial t} \int_{S_2} \mathbf{D} \cdot \mathbf{ds} = \frac{dQ}{dt} = C\frac{dV}{dt}.$$

Substituting this expression into Equation (10.14), we obtain

$$I = C\frac{dV}{dt}, \tag{10.15}$$

which is the familiar V–I characteristic of a capacitor. Integrating this equation with respect to time, we obtain the other familiar capacitor expression,

$$V = \frac{1}{C} \int_{-\infty}^{t} I(t')\,dt'. \tag{10.16}$$

We can generalize Equation (10.15) so that it applies to structures that don't "look" like capacitors, such as the structure shown in Figure 10-2. Here, N wires carry current through a surface S. For this case the total current entering the surface is the sum of the individual wire currents, which yields

$$\sum_{i=1}^{N} I_i = C\frac{dV}{dt}, \tag{10.17}$$

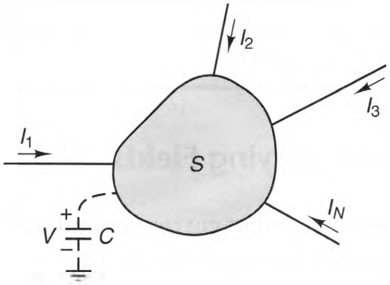


Figure 10-2 Conductor currents flowing into a surface S, showing that their algebraic sum is not zero when displacement current is present.

where I_i is the current flowing toward the surface on the i^{th} wire and C is the capacitance of the surface to ground. This expression is called the ***generalized Kirchhoff's current law*** and states that the net current into a surface (or point) is zero only when the surface-to-ground capacitance is zero.

Example 10-1

Figure 10-3 shows a resistive load enclosed by a metal chassis and fed by a coaxial cable. The cable shield isattached to the chassis, and the load is connected between the inner conductor of the cable and the chassis. If the chassis-to-ground capacitance is 50 [pF] and a sinusoidal, 10 [mv] voltage is present between the chassis and ground, calculate the difference ΔI between the currents on the inner and outer conductors of the transmission line when $f = 60$ [Hz] and $f = 30$ [MHz].

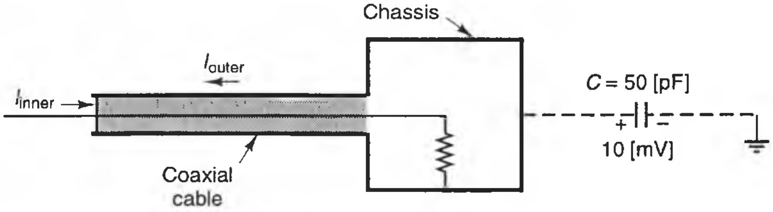


Figure 10-3 A coaxial cable connected to a chassis with a resistive load and a chassis-to-ground capacitance.

Solution:

Using Equation (10.17) and $d(\sin \omega t)/dt = \omega \cos \omega t$, we can write

$$\Delta I = I_{\text{inner}} - I_{\text{outer}} = \left| C\frac{dV}{dt} \right| = \omega C V,$$

where V is the peak chassis-to-ground voltage. At $f = 60$ [Hz], we have

$$\Delta I = (2\pi) \times (60) \times (50 \times 10^{-12}) \times 10^{-2} = 0.188 \text{ [nA]}.$$

At $f = 30$ [MHz], we have

$$\Delta I = (2\pi) \times (30 \times 10^{6}) \times (50 \times 10^{-12}) \times 10^{-2} = 94.20 \text{ [}\mu\text{A]}.$$

Shielded cables and metal chassis are often used to keep circuits from radiating unwanted fields. When the currents on the inner and outer conductors of a coaxial cable are balanced (i.e., $\Delta I = 0$), no net magnetic field is generated outside the cable. (See Section 7-4-1.) However, this example shows that the cable currents will not be balanced if the chassis potential is nonzero and there is a chassis-to-ground capacitance.

10-3 Maxwell's Equations for Time-Varying Fields

Maxwell's invention of displacement current completed the list of postulates necessary to model all time-varying electromagnetic effects:

1. Coulomb's law of force (for static charge distributions)
2. Ampère's law of force (for steady current distributions)
3. The law of charge conservation (or Maxwell's displacement current)
4. Faraday's law of induction
5. The Lorentz force law

The mathematical representations of postulates 1–4 are called Maxwell's equations, which can be written in either differential (point) or integral form. In point form, Maxwell's equations are

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{10.18}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \tag{10.19}$$

$$\nabla \cdot \mathbf{D} = \rho_v \tag{10.20}$$

$$\nabla \cdot \mathbf{B} = 0. \tag{10.21}$$

Although it is not needed to describe the fields themselves, the Lorentz force law (postulate #5) provides the link between electromagnetic fields and the forces that make them known to us. As we discussed in Chapter 3, this law states that the electromagnetic force acting on a point charge is given by

$$\mathbf{F} = Q(\mathbf{E} + \mathbf{u} \times \mathbf{B}), \tag{10.22}$$

where **u**, **E**, and **B** are measured in the same inertial frame of reference.

Each of Maxwell's four equations was determined from different experimental evidence (or conjecture, in the case of displacement current), but the two divergence equations can actually be derived directly from the two curl equations. To show this, let us take the divergence of both sides of Equation (10.18) to obtain

$$\nabla \cdot \nabla \times \mathbf{E} = -\nabla \cdot \frac{\partial \mathbf{B}}{\partial t}.$$

The left side of this expression is identically zero (see Equation (B.8)), and the order of the differentiation on the right-hand side can be interchanged to yield

$$\frac{\partial}{\partial t} \nabla \cdot \mathbf{B} = 0,$$

which states that the divergence of **B** at any point is constant in time. Since a nonzero divergence of **B** has *never* been observed,[2] we conclude that this constant is zero, which agrees with Equation (10.21).

In a similar manner, taking the divergence of both sides of Equation (10.19) yields

[2] While there is some evidence that may support the existence of subatomic magnetic charges, it appears unlikely that they could be of any engineering significance.

$$\nabla \cdot \nabla \times \mathbf{H} = \nabla \cdot \mathbf{J} + \nabla \cdot \frac{\partial \mathbf{D}}{\partial t}.$$

Noting that the left-hand side is identically zero and $\nabla \cdot \mathbf{J} = -(\partial \rho_v / \partial t)$, we can write this expression as

$$\frac{\partial}{\partial t}(\nabla \cdot \mathbf{D} - \rho_v) = 0,$$

which states that the term $(\nabla \cdot \mathbf{D} - \rho_v)$ is constant in time at any point. Experimental evidence shows that this constant is always zero, yielding Equation (10.20).

The integral form of Maxwell's curl equations can be obtained by integrating them over open surfaces and applying Stokes's theorem to the left-hand sides. Similarly, integral forms of the divergence equations can be obtained by integrating them over arbitrary volumes and applying the divergence theorem to the left-hand sides. The resulting equations are

$$\oint_C \mathbf{E} \cdot \mathbf{d\ell} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot \mathbf{ds} \tag{10.23}$$

$$\oint_C \mathbf{H} \cdot \mathbf{d\ell} = I + \int_S \frac{\partial \mathbf{D}}{\partial t} \cdot \mathbf{ds} \tag{10.24}$$

$$\oint_S \mathbf{D} \cdot \mathbf{ds} = Q \tag{10.25}$$

$$\oint_S \mathbf{B} \cdot \mathbf{ds} = 0. \tag{10.26}$$

Maxwell's equations provide a complete description of the behavior of **E**, **D**, **B**, and **H** in terms of the sources **J** and Q. However, the equations do not contain any material-dependent information. This information is provided by the ***constitutive relations***, which, for simple media, are

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P} = \epsilon \mathbf{E} \tag{10.27}$$

$$\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M}) = \mu \mathbf{H} \tag{10.28}$$

$$\mathbf{J} = \sigma \mathbf{E} + \mathbf{J}_i. \tag{10.29}$$

In these expressions, **P** the polarization, **M** the magnetization, and $\mathbf{J}_i$ the impressed (or source) current density. Also, the constitutive parameters ϵ, μ, and σ are the permittivity, permeability, and conductivity of the medium, respectively. For most materials, ϵ, μ, and σ are scalars. However, some materials have directional properties that cannot be modeled with scalar constitutive parameters. In such cases, one or more of the constitutive parameters becomes a 3 × 3 matrix, called a tensor. These materials are important for certain devices (such as microwave circulators), but they are beyond the scope of this text.

One of the distinguishing characteristics of time-varying phenomena is that electric and magnetic fields are always present simultaneously. This is very different from electrostatics and magnetostatics, where one type of field can exist without the other. The following example demonstrates this characteristic.

Example 10-2

Suppose that the E-field in a source-free (i.e., $\mathbf{J}_t = 0$) region of free space is given by

$$\mathbf{E} = E_o \sin(\omega t - \beta z)\hat{\mathbf{a}}_x,$$

where ω and β are constants. Find the H-field that is also present. What value must the constant β be in order for both fields to satisfy all of Maxwell's equations?

Solution:

Substituting **E** into Equation (10.18), we find that

$$\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E} = -\frac{\partial}{\partial z} E_o \sin(\omega t - \beta z)\hat{\mathbf{a}}_y = \beta E_o \cos(\omega t - \beta z)\hat{\mathbf{a}}_y.$$

Integrating with respect to time yields

$$\mathbf{B} = \int \beta E_o \cos(\omega t - \beta z)\,dt = \frac{\beta E_o}{\omega} \sin(\omega t - \beta z)\hat{\mathbf{a}}_y + C\,\hat{\mathbf{a}}_y,$$

where C is some constant. Since the term $C\,\hat{\mathbf{a}}_y$ is a magnetostatic field, it is unrelated to the time-varying E-field and can be ignored, yielding

$$\mathbf{H} = \frac{\mathbf{B}}{\mu_o} = \frac{\beta E_o}{\omega \mu_o} \sin(\omega t - \beta z)\hat{\mathbf{a}}_y.$$

To find the acceptable value of β, let us substitute this expression for **H** into Ampère's law (Equation (10.19)) and see if it yields the E-field we started with. Remembering that $\mathbf{J}_t = 0$, we obtain

$$\epsilon_o \frac{\partial \mathbf{E}}{\partial t} = \nabla \times \mathbf{H} = -\frac{\partial}{\partial z}\left(\frac{\beta E_o}{\omega \mu_o} \sin(\omega t - \beta z)\right)\hat{\mathbf{a}}_x$$

$$= \frac{\beta^2 E_o}{\omega \mu_o} \cos(\omega t - \beta z)\hat{\mathbf{a}}_x.$$

Integrating this expression with respect to time, we get

$$\mathbf{E} = \frac{\beta^2 E_o}{\omega^2 \mu_o \epsilon_o} \sin(\omega t - \beta z)\hat{\mathbf{a}}_x.$$

Comparing this expression with the E-field we started with, we conclude that β must satisfy $\beta^2/(\omega^2 \mu_o \epsilon_o) = 1$. Thus, the necessary value of β is given by

$$\beta = \omega\sqrt{\mu_o \epsilon_o}.$$

It is left as an exercise for the student to show that these expressions for **E** and **B** satisfy the divergence equations $\nabla \cdot \mathbf{D} = 0$ and $\nabla \cdot \mathbf{B} = 0$.

The E- and H-fields we have just examined are unlike any of the fields discussed in the earlier chapters because they *move* (or propagate) with time t. We can see this by noting that the sinusoid terms in both **E** and **H** have the same argument, $\omega t - \beta z$. This means that the zero-crossings of **E** and **H** shift towards increasing values of z as t increases. For instance, one zero-crossing occurs when $\omega t - \beta z = 0$. When the time is advanced by Δt, the position of this zero-crossing shifts by a value

$$\Delta z = \frac{\omega \Delta t}{\beta}.$$

Dividing both sides of the preceding expression by Δt, we obtain

$$\frac{\Delta z}{\Delta t} = \frac{\omega}{\beta} = \frac{1}{\sqrt{\mu_o \epsilon_o}} = 2.9979 \times 10^8 \ [\text{m/s}] \approx 3 \times 10^8 \ [\text{m/s}].$$

Hence, these fields represent disturbances that are moving towards increasing values of z at a rate equal to the speed of light in a vacuum.

There are many types of propagating fields. The specific fields presented in this example are called plane waves and will be discussed at length in Chapter 12.

10-4 Time-Harmonic Fields

Very often we are interested in the fields generated by sources that vary sinusoidally in time. In these cases, it is convenient to express the resulting electric and magnetic fields as phasors. Fortunately, transforming the field quantities encountered in electromagnetics from the time domain to the phasor (i.e., frequency) domain is accomplished in basically the same way as for circuit quantities. We will start our discussion by reviewing the rules of phasors for discrete-time functions (such as the voltages and currents in a circuit) and then generalize these rules for scalar and vector fields.

10-4-1 REVIEW OF PHASORS FOR DISCRETE SCALARS

Phasor notation makes use of the fact that there are only two parameters that are needed to completely specify a time-harmonic waveform: its amplitude and its phase. For instance, consider the sinusoidal current waveform

$$i(t) = I_m \cos(\omega t + \phi) \qquad [\text{A}]. \tag{10.30}$$

Using Euler's identity, we can write $i(t)$ as the real part of a complex exponential function, i.e.,

$$i(t) = \text{Re}[Ie^{j\omega t}] \qquad [\text{A}],$$

where I is the phasor representation of $i(t)$, given by

$$I = I_m e^{j\phi} = I_m \angle \phi \qquad [\text{A}].$$

The steps required to transform a time-harmonic function into its phasor form are summarized as follows:

1. Write the given time function as a cosine function with a phase angle.

2. Express the cosine function as the real part of a complex exponential using Euler's identity.
3. Drop the Re and $e^{j\omega t}$ from the expression.

Conversely, to transform a phasor into its time-domain equivalent, one need only reverse the steps:

1. Reinsert the Re and $e^{j\omega t}$.
2. Use Euler's identity to express the complex exponential in terms of the sine and cosine functions.
3. Perform the Re operation.

It is tempting to think of a phasor as simply a shorthand notation for a time-harmonic function, but it is actually the transformation of the time-harmonic function into the frequency domain. This transformation is often indicated using the notation

$$i(t) \leftrightarrow I. \tag{10.31}$$

One of the most useful features of phasors is that differentiation and integration operators in the time domain transform to algebraic operations in the frequency domain. For instance, consider again the $i(t)$ given by Equation (10.30). Taking the derivative of both sides with respect to t, we have

$$\frac{d}{dt} i(t) = -\omega I_m \sin(\omega t + \phi) = \omega I_m \cos(\omega t + \phi + 90°).$$

Transforming $d(i(t))/dt$ into the frequency domain, we obtain

$$\frac{d}{dt} i(t) \leftrightarrow \omega I_m e^{j90°} e^{j\phi} = j\omega I_m e^{j\phi}. \tag{10.32}$$

Remembering that the phasor representation of $i(t)$ is $I = I_m e^{j\phi}$, we can write Equation (10.32) as

$$\frac{d}{dt} i(t) \leftrightarrow j\omega I, \tag{10.33}$$

which shows that the phasor representation of the time derivative of any time-harmonic function is simply $j\omega$ times the phasor of the function itself. Similarly, if we integrate both sides of Equation (10.30) with respect to time, we obtain

$$\int i dt = \frac{I_m}{\omega} \sin(\omega t + \phi) = \frac{I_m}{\omega} \cos(\omega t + \phi - 90°),$$

which transforms to

$$\int i dt \leftrightarrow \frac{I_m}{\omega} e^{-j90°} e^{j\phi} = \frac{I_m}{j\omega} e^{j\phi}. \tag{10.34}$$

Knowing that $I = I_m\, e^{j\phi}$, we can write Equation (10.34) as

$$\int idt \leftrightarrow \frac{1}{j\omega} I \tag{10.35}$$

Thus, the phasor representation of the integral of a time-harmonic function is simply the phasor of the function divided by $j\omega$.

A simple way to show how phasors can simplify the analysis of linear systems is to consider the circuit shown in Figure 10-4, which consists of a resistor R and an inductor L in series with a time-harmonic voltage source with magnitude V_m and phase ϕ.

The differential equation satisfied by the current $i(t)$ is

$$L\frac{d}{dt}i(t) + Ri(t) = V_m \cos(\omega t + \phi). \tag{10.36}$$

Since this is a linear differential equation, the steady-state response $i(t)$ is time harmonic, with the same frequency ω as the source. Transforming each term of Equation (10.36) into the frequency domain, we obtain

$$j\omega\, LI + RI = V_m\, e^{j\phi}, \tag{10.37}$$

where I is the phasor representation of $i(t)$. Unlike Equation (10.36), this equation can be solved algebraically, yielding

$$I = \frac{V_m e^{j\phi}}{R + j\omega L} = \frac{V_m}{\sqrt{R^2 + (\omega L)^2}} e^{j\psi}, \tag{10.38}$$

where

$$\psi = \phi - \tan^{-1}\frac{\omega L}{R}. \tag{10.39}$$

Transforming I back to the time domain, we finally obtain

$$i(t) = \frac{V_m}{\sqrt{R^2 + (\omega L)^2}} \cos(\omega t - \psi). \tag{10.40}$$

As this example shows, we can use phasors to solve a differential equation by means of simple algebra. The only cost of using phasor analysis is that we must use complex arithmetic.

Most circuit analysis texts use lowercase variables (such as $i(t)$) to represent time-domain waveforms and uppercase variables (such as I) to represent frequency-domain phasors. This convention is not usually followed in electromagnetics, because of the large number of variables that are needed. Instead, uppercase variables are usually used for both time-domain and frequency-domain quantities. This poses no real prob-

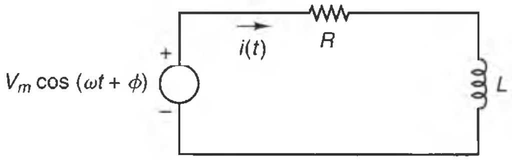


Figure 10-4 An RL circuit with a time-harmonic voltage source.

lem, since it is usually obvious which domain an equation is written in. As a rule, the presence of complex numbers indicates the frequency domain, and the presence of the variable t indicates the time domain.

10-4-2 PHASOR REPRESENTATIONS OF SCALAR AND VECTOR FIELDS

The vector and scalar fields encountered in electromagnetics are functions of position, as well as time. Nevertheless, these quantities can be transformed into the frequency domain whenever they are time harmonic. For instance, consider the volume charge density distribution

$$\rho_v(r, t) = e^{-r^2} \cos(\omega t + \beta r) \qquad [\mathrm{C/m^3}],$$

which is time harmonic at all radial positions r. Using Euler's identity, we can write ρ_v as

$$\rho_v = \mathrm{Re}[e^{-r^2} e^{j\omega t} e^{j\beta r}] \qquad [\mathrm{C/m^3}].$$

If we now drop the Re and the $e^{j\omega t}$, we are left with $e^{-r^2} e^{j\beta r}$, which is the phasor representation of $\rho_v(r, t)$. Thus, we can write

$$\rho_v(r,t) \leftrightarrow e^{-r^2} e^{j\beta r} \qquad [\mathrm{C/m}].$$

As this example shows, the rules for transforming scalar fields into the frequency domain are the same as for discrete quantities. The only difference is that the phasors of scalar fields are functions of position.

The rules for transforming vector fields into the frequency domain are basically the same. This is accomplished most easily by writing these vectors in terms of their components and then transforming each component. For instance, consider the vector

$$\mathbf{E}(t) = E_x(t)\hat{\mathbf{a}}_x + E_y(t)\hat{\mathbf{a}}_y + E_z(t)\hat{\mathbf{a}}_z. \tag{10.41}$$

If $\mathbf{E}(t)$ is time harmonic, each of its components is also time harmonic. Also, since the unit vectors are not functions of time, they are not affected by the transformation. Hence, we can transform $\mathbf{E}(t)$ into the frequency domain as

$$\mathbf{E}(t) \leftrightarrow \mathbf{E} = E_x \hat{\mathbf{a}}_x + E_y \hat{\mathbf{a}}_y + E_z \hat{\mathbf{a}}_z, \tag{10.42}$$

where $\mathbf{E}$ is the phasor (i.e., frequency domain) representation of $\mathbf{E}(t)$, and E_x, E_y, and E_z are the x-, y-, and z-component phasors, defined by

$$E_x(t) \leftrightarrow E_x \tag{10.43a}$$

$$E_y(t) \leftrightarrow E_y \tag{10.43b}$$

$$E_z(t) \leftrightarrow E_z. \tag{10.43c}$$

We can also express the relationship between $\mathbf{E}(t)$ and the phasor $\mathbf{E}$ by the expression

$$\mathbf{E}(t) = \mathrm{Re}[\mathbf{E} e^{j\omega t}]. \tag{10.44}$$

Example 10-3

For the time-harmonic vector

$$\mathbf{B}(t) = 3\cos(\omega t - \beta z + 30°)\hat{\mathbf{a}}_x - 4\sin(\omega t - \beta z)\hat{\mathbf{a}}_y,$$

find its phasor representation **B**. Also, transform **B** back to the time domain, and verify that it equals **B**(t).

Solution:

First, we write both components of **B** in terms of cosine functions:

$$\mathbf{B}(t) = 3\cos(\omega t - \beta z + 30°)\hat{\mathbf{a}}_x + 4\cos(\omega t - \beta z + 90°)\hat{\mathbf{a}}_y.$$

Transforming both components into the frequency domain, we obtain

$$\mathbf{B} = 3e^{j30°}e^{-j\beta z}\hat{\mathbf{a}}_x + 4e^{j90°}e^{-j\beta z}\hat{\mathbf{a}}_y.$$

Also, since $e^{j30°} = 0.866 + j0.5$ and $e^{j90°} = j$, **B** can also be written as

$$\mathbf{B} = (2.598 + j1.5)e^{-j\beta z}\hat{\mathbf{a}}_x + j4e^{-j\beta z}\hat{\mathbf{a}}_y.$$

To obtain **B**(t) from **B**, we simply reinsert the Re and $e^{j\omega t}$, yielding

$$\begin{aligned}\mathbf{B}(t) &= \text{Re}[3e^{j30°}e^{-j\beta z}e^{j\omega t}\hat{\mathbf{a}}_x + 4e^{j90°}e^{-j\beta z}e^{j\omega t}\hat{\mathbf{a}}_y] \\ &= 3\cos(\omega t - \beta z + 30°)\hat{\mathbf{a}}_x + 4\cos(\omega t - \beta z + 90°)\hat{\mathbf{a}}_y \\ &= 3\cos(\omega t - \beta z + 30°)\hat{\mathbf{a}}_x - 4\sin(\omega t - \beta z)\hat{\mathbf{a}}_y,\end{aligned}$$

which is same as the expression we started with.

10-4-3 MAXWELL'S EQUATIONS FOR TIME-HARMONIC FIELDS

Maxwell's equations are particularly simple when written in the frequency domain. For instance, consider the curl-**E** equation (Equation (10.18)),

$$\nabla \times \mathbf{E}(t) = -\frac{\partial \mathbf{B}(t)}{\partial t}.$$

The left-hand side contains spatial derivatives, but no time derivatives, so it retains the same form in the frequency domain:

$$\nabla \times \mathbf{E}(t) \leftrightarrow \nabla \times \mathbf{E}.$$

The right-hand side contains a time derivative of **B**, which transforms to

$$-\frac{\partial \mathbf{B}(t)}{\partial t} \leftrightarrow -j\omega\mathbf{B}.$$

Thus, the frequency-domain version of the curl-**E** equation is

$$\nabla \times \mathbf{E} = -j\omega\mathbf{B}. \qquad (10.45)$$

Similarly, the curl-**H** equation transforms to

$$\nabla \times \mathbf{H} = \mathbf{J} + j\omega \mathbf{D}, \tag{10.46}$$

where **J** is the phasor representation of the volume current density $\mathbf{J}(t)$.

As can be seen from Equations (10.45) and (10.46), only the time-derivative terms in Maxwell's equations change their form when transformed to the frequency domain. As a result, Maxwell's two divergence equations (Equations (10.20) and (10.21)) do not change in the frequency domain, since they contain no time derivatives.

Summarizing, the frequency domain form of Maxwell's equations are

$$\nabla \times \mathbf{E} = -j\omega \mathbf{B} \tag{10.47}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + j\omega \mathbf{D} \tag{10.48}$$

$$\nabla \cdot \mathbf{D} = \rho_v \tag{10.49}$$

$$\nabla \cdot \mathbf{B} = 0. \tag{10.50}$$

Similarly, the integral form of Maxwell's equations in the frequency domain are

$$\oint_C \mathbf{E} \cdot d\boldsymbol{\ell} = -j\omega \int_S \mathbf{B} \cdot d\mathbf{s} \tag{10.51}$$

$$\oint_C \mathbf{H} \cdot d\boldsymbol{\ell} = I + j\omega \int_S \mathbf{D} \cdot d\mathbf{s} \tag{10.52}$$

$$\oint_S \mathbf{D} \cdot d\mathbf{s} = Q \tag{10.53}$$

$$\oint_S \mathbf{B} \cdot d\mathbf{s} = 0. \tag{10.54}$$

The constitutive relations that account for the material properties of a medium retain the same forms in the frequency domain as they do in the time domain, namely,

$$\mathbf{D} = \epsilon_o \mathbf{E} + \mathbf{P} = \epsilon \mathbf{E} \tag{10.55}$$

$$\mathbf{B} = \mu_o(\mathbf{H} + \mathbf{M}) = \mu \mathbf{H} \tag{10.56}$$

$$\mathbf{J} = \sigma \mathbf{E} + \mathbf{J}_i, \tag{10.57}$$

where the constitutive parameters ϵ, μ, and σ are, respectively, the permittivity, permeability, and conductivity of the medium. The constitutive parameters of most materials are independent of frequency and are real, at least at low frequencies. At higher frequencies, however, these parameters often vary with frequency and become complex, indicating that the vectors they relate differ in phase. In Chapter 12, we will show that complex values of ϵ and μ indicate that the medium is lossy, even if its conductivity σ is zero. When the complex nature of these parameters is important, we will use the notation

$$\epsilon = \epsilon' - j\epsilon'' \tag{10.58}$$

$$\mu = \mu' - j\mu'' \tag{10.59}$$

$$\sigma = \sigma' - j\sigma''. \tag{10.60}$$

10-4-4 MAXWELL'S EQUATIONS IN SIMPLE MEDIA

Most of the materials used in electrical devices are relatively elementary. We will call a ***simple medium*** a medium that is 1) homogeneous, 2) isotropic, and 3) linear. This means that the constitutive parameters (ϵ, μ, and σ) of simple media are 1) independent of position, 2) scalar valued, and 3) independent of the field strengths.

Since **D** and **E**, and **B** and **H**, are related by constants in simple media, it is customary to express Maxwell's equations in terms of **E** and **H** only. Hence, for simple media, Equations (10.47)–(10.50) become

$$\left.\begin{aligned} \nabla \times \mathbf{E} &= -j\omega\mu\mathbf{H} & (10.61)\\ \nabla \times \mathbf{H} &= \mathbf{J}_i + \sigma\mathbf{E} + j\omega\epsilon\mathbf{E} & (10.62)\\ \nabla \cdot \mathbf{E} &= \frac{\rho_v}{\epsilon} & (10.63)\\ \nabla \cdot \mathbf{H} &= 0. & (10.64) \end{aligned}\right\} \text{(Simple media).}$$

If a region is source free, the impressed current density $\mathbf{J}_i$ is zero. It then follows that the charge density ρ_v is also zero in a simple, source-free region. To see why, we note that when $\mathbf{J}_i = 0$, Equation (10.62) becomes

$$\nabla \times \mathbf{H} = (\sigma + j\omega\epsilon)\mathbf{E}.$$

If we take the divergence of both sides of this expression and remember that the divergence of the curl of any vector is identically zero, we obtain

$$\nabla \cdot \nabla \times \mathbf{H} = (\sigma + j\omega\epsilon)\,\nabla \cdot \mathbf{E} = 0.$$

Notice in this expression that since σ and ϵ are constants, they can both be taken out of the divergence operation. Then, because the term $(\sigma + j\omega\epsilon)$ is nonzero, we can conclude that $\nabla \cdot \mathbf{E} = 0$, and Equation (10.63) yields[3]

$$\rho_v = 0 \qquad \text{(Simple, source-free media).} \tag{10.65}$$

Thus, at source-free points in simple media, Maxwell's equations become

$$\left.\begin{aligned} \nabla \times \mathbf{E} &= -j\omega\mu\mathbf{H} & (10.66)\\ \nabla \times \mathbf{H} &= (\sigma + j\omega\epsilon)\mathbf{E} & (10.67)\\ \nabla \cdot \mathbf{E} &= 0 & (10.68)\\ \nabla \cdot \mathbf{H} &= 0 & (10.69) \end{aligned}\right\} \text{(Simple, source-free media)}$$

Of all the forms of Maxwell's equations, we will find Equations (10.66)–(10.69) the most useful, since most materials are simple, and we are usually interested in the E- and H-fields away from the sources that cause them.

[3] When fields are not time harmonic, ρ_v can be nonzero while the medium relaxes from a transient event. (See Section 5-3-5).

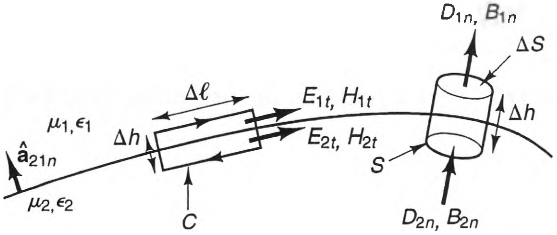


Figure 10-5 Contours and surfaces used to determine the boundary conditions for electric fields, magnetic fields, and currents at the interface between two dissimilar media.

10-5 Boundary Conditions for Time-varying Fields

The properties of time-varying electric and magnetic fields at the interfaces between different media are the same as for time-invariant fields. These properties can be derived directly from Maxwell's equations.

Figure 10-5 shows an interface between two homogeneous media. The upper medium is characterized by μ_1 and ϵ_1, and the lower medium is characterized by μ_2 and ϵ_2. Also shown is a contour C, which is a small path with depth Δh and length $\Delta \ell$ that straddles the interface. We can determine the behavior of the tangential E- and H-fields at the interface by evaluating Equations (10.23) and (10.24) around this contour, obtaining

$$\oint_C \mathbf{E} \cdot \mathbf{d\ell} = -\frac{\partial}{\partial t}\int_S \mathbf{B} \cdot \mathbf{ds}$$

and

$$\oint_C \mathbf{H} \cdot \mathbf{d\ell} = I + \frac{\partial}{\partial t}\int_S \mathbf{D} \cdot \mathbf{ds},$$

where S is the flat surface surrounded by the path and I is the current that passes through S in a right-handed sense. In the limit as $\Delta h \to 0$, the contributions from the contour that are perpendicular to the interface become negligible. Also, since $\Delta S \to 0$ as $\Delta h \to 0$, both surface integrals also become negligible, leaving

$$\oint_C \mathbf{E} \cdot \mathbf{d\ell} \approx E_{1t}\Delta \ell - E_{2t}\Delta \ell \approx 0$$

and

$$\oint_C \mathbf{H} \cdot \mathbf{d\ell} \approx H_{1t}\Delta \ell - H_{2t}\Delta \ell = I \approx J_{sn}\Delta \ell,$$

where J_{sn} is the component of the surface current density that is perpendicular to the loop. Dividing both expressions by $\Delta \ell$, we obtain

$$E_{1t} - E_{2t} = 0 \tag{10.70}$$

and

$$H_{1t} - H_{2t} = J_{sn}. \tag{10.71}$$

Using $\mathbf{B} = \mu\,\mathbf{H}$ and $\mathbf{D} = \epsilon\mathbf{E}$, we find that the preceding conditions become

$$\frac{1}{\epsilon_1} D_{1t} - \frac{1}{\epsilon_2} D_{2t} = 0 \tag{10.72}$$

and

$$\frac{1}{\mu_1} B_{1t} - \frac{1}{\mu_2} B_{2t} = J_{sn}. \tag{10.73}$$

Recognizing that there can be two tangential components for both **D** and **B**, we can generalize Equations (10.70) and (10.71) by using the cross product, yielding

$$\hat{\mathbf{a}}_{21n} \times (\mathbf{E}_1 - \mathbf{E}_2) = 0 \tag{10.74}$$

$$\hat{\mathbf{a}}_{21n} \times (\mathbf{H}_1 - \mathbf{H}_2) = \mathbf{J}_s, \tag{10.75}$$

where $\hat{\mathbf{a}}_{21n}$ is the unit normal to the interface that is directed from region 2 to region 1.

The boundary conditions satisfied by the normal components of the fields are obtained by evaluating Equations (10.25) and (10.26) over the "pillbox" surface shown in Figure 10-5, with height Δh and end-cap area ΔS. If we let $\Delta h \to 0$, the area of the "barrel" portion of the surface becomes zero, and hence, the only contributions to the integral come from the bottom and top surfaces; thus

$$\oint_S \mathbf{D} \cdot \mathbf{ds} \approx D_{1n}\Delta S - D_{2n}\Delta S = Q \approx \rho_s \Delta S,$$

and

$$\oint_S \mathbf{B} \cdot \mathbf{ds} \approx B_{1n}\Delta S - B_{2n}\Delta S = 0,$$

where ρ_s is the charge density along the interface. Dividing both equations by ΔS, we obtain

$$D_{1n} - D_{2n} = \rho_s \tag{10.76}$$

$$B_{1n} - B_{2n} = 0. \tag{10.77}$$

Using $\mathbf{B} = \mu\,\mathbf{H}$ and $\mathbf{D} = \epsilon\mathbf{E}$, we can also write these expressions

$$\epsilon_1 E_{1n} - \epsilon_2 E_{2n} = \rho_s \tag{10.78}$$

$$\mu_1 H_{1n} - \mu_2 H_{2n} = 0. \tag{10.79}$$

Finally, the boundary conditions can be written in more compact form using the unit normal $\mathbf{a}_{21n}$:

$$(\mathbf{D}_1 - \mathbf{D}_2) \cdot \hat{\mathbf{a}}_{21n} = \rho_s \tag{10.80}$$

$$(\mathbf{B}_1 - \mathbf{B}_2) \cdot \hat{\mathbf{a}}_{21n} = 0. \tag{10.81}$$

There are two important special cases of these boundary conditions. The first is when both media are nonconducting. For this case, $\rho_s = 0$ at the interface, so the normal component of **D** is continuous across the interface and the normal component of **E** is discontinuous when $\epsilon_1 \neq \epsilon_2$. The second special case occurs when one region is a perfect dielectric ($\sigma = 0$) and the other is a perfect conductor ($\sigma \to \infty$). For this case, the electric and magnetic fields inside the perfect conductor are zero (except for possibly a magnetostatic field), and Equations (10.70) and (10.79) show that the tangential electric field and the normal magnetic field on the dielectric side of the interface are both zero, yielding

$$E_t = H_n = 0. \tag{10.82}$$

Also, the surface current density along the surface is related to the tangential component of the magnetic field and is given by

$$\mathbf{J}_s = \hat{\mathbf{a}}_n \times \mathbf{H}, \tag{10.83}$$

where $\hat{\mathbf{a}}_n$ is perpendicular to the conductor and points from the conductor to the dielectric.

The most general case occurs when both media are conducting. For this case, the constraints on the tangential components of **J** follow directly from Equation (10.70), which is valid for both conducting and nonconducting media. Using $\mathbf{J} = \sigma \mathbf{E}$, we obtain

$$\frac{J_{1t}}{\sigma_1} = \frac{J_{2t}}{\sigma_2}. \tag{10.84}$$

On the other hand, the constraints on the normal components of the current density **J** can be determined from the integral form of the continuity equation,

$$\oint_S \mathbf{J} \cdot \mathbf{ds} = -\frac{\partial}{\partial t} Q_{\text{enc}}.$$

Here, S is the pillbox surface shown in Figure 10-5, and Q_{enc} is the charge contained within this surface. In the limit as the pillbox height goes to zero, the only contribution to the surface integral comes from the end caps, so $\oint_S \mathbf{J} \cdot \mathbf{ds} \approx (J_{2n} - J_{1n})\Delta s$. Also, on the right side, we have $Q_{\text{enc}} \approx \rho_s \Delta s$, where ρ_s is the surface charge density along the interface. Equating these two expressions and noting that the Δs terms cancel, we obtain

$$J_{2n} - J_{1n} = -\frac{\partial \rho_s}{\partial t}. \tag{10.85}$$

Example 10-4

Figure 10-6 shows a flat, perfectly conducting surfacc. If it is known that the magnetic field is givcn by

$$\mathbf{H} = \begin{cases} 3\cos x\,\hat{\mathbf{a}}_x + 2\cos x\,\hat{\mathbf{a}}_y \quad [\text{A/m}] & z \geqslant 0 \\ 0 & z < 0\text{'} \end{cases}$$

find the current density on the conductor surface.

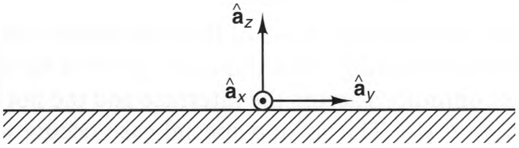


Figure 10-6 A perfectly conducting surface.

Solution:

If we call the regions $z \geqslant 0$ and $z < 0$ regions 1 and 2, respectively, the unit normal from region 2 to region 1 is $\hat{\mathbf{a}}_{21n} = \hat{\mathbf{a}}_z$. Hence, using Equation (10.83), we have

$$\begin{aligned}\mathbf{J}_s &= \hat{\mathbf{a}}_z \times (3\cos(x)\,\hat{\mathbf{a}}_x + 2\cos(x)\,\hat{\mathbf{a}}_y) \\ &= 3\cos(x)\,\hat{\mathbf{a}}_y - 2\cos(x)\,\hat{\mathbf{a}}_x \qquad [\text{A/m}].\end{aligned}$$

10-6 ac Circuit Analysis

Strange as it may seem, a common goal of electromagnetic analysis is to analyze as many parts of a system as possible using ordinary ac circuit analysis. The reason for this is simple: Circuit analysis is much simpler than field analysis. In this section, we will derive the equation for a simple circuit directly from Maxwell's equations. This will demonstrate not only that ac circuit theory is based firmly on electromagnetic theory, but also what assumptions are necessary in order for ac circuit theory to be valid.

Figure 10-7 shows a series circuit consisting of a voltage source, a resistor, an inductor, and a capacitor. These elements are connected with perfectly conducting, filamentary wires that are short enough so that the inductance and capacitance of the wires can be neglected. We will also assume that the only places where an electric field exists along the circuit are between the terminals of the voltage source, between the terminals of the resistor, and between the terminals of the capacitor. Also, we will assume that the magnetic field is negligible everywhere except between the windings of the inductor.

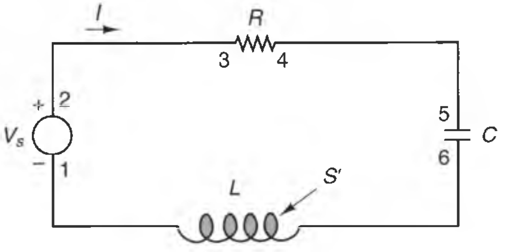


Figure 10-7 An ac circuit in which the magnetic field is significant only near the inductor and the electric field is significant only between the plates of the capacitor.

We can apply Faraday's law (Equation (10.23)) to this circuit, as long as the integration path passes around the voltage source, rather than through it, since a nonconservative force field is present inside a voltage source. Integrating along the circuit perimeter, and assuming that the tangential electric field is zero along the conductors, we can write

$$\int_1^2 \mathbf{E} \cdot \mathbf{d\ell} + \int_3^4 \mathbf{E} \cdot \mathbf{d\ell} + \int_5^6 \mathbf{E} \cdot \mathbf{d\ell} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot \mathbf{ds}, \tag{10.86}$$

where the numbered points in the integrals correspond to the indicated points in Figure 10-7 and S is the surface bounded by the circuit contour. We will now show that each of these integrals represents a familiar circuit expression.

The integral $\int_1^2 \mathbf{E} \cdot \mathbf{d\ell}$ is simply the electromotive force (emf) of the source (see Section 5-2-3); that is,

$$\int_1^2 \mathbf{E} \cdot \mathbf{d\ell} = -\int_2^1 \mathbf{E} \cdot \mathbf{d\ell} = -V_s, \tag{10.87}$$

where V_s is the source voltage from terminal 1 to terminal 2. Next, we note that inside the resistor, $E = J/(\sigma S_R)$ (where S_R is the cross-sectional area of the resistor), so the integral $\int_3^4 \mathbf{E} \cdot \mathbf{d\ell}$ can be written as

$$\int_3^4 \mathbf{E} \cdot \mathbf{d\ell} = I \frac{d_R}{\sigma S_R} = IR \tag{10.88}$$

where d_R is the length of the resistor, $R = d_R/\sigma S_R$ is its resistance, and I is the current flowing clockwise through the circuit. The third line integral, $\int_5^6 \mathbf{E} \cdot \mathbf{d\ell}$, is the voltage across the capacitor. Using Equation (10.16), we can write this voltage in terms of the current flowing into the positive plate, i.e.,

$$\int_5^6 \mathbf{E} \cdot \mathbf{d\ell} = \frac{1}{C} \int_{-\infty}^{t} I(t)\, dt, \tag{10.89}$$

where C is the capacitance of the capacitor. Finally, since $\mathbf{B}$ is negligible everywhere except inside the inductor windings, the surface integral $\int_S (\partial \mathbf{B}/\partial t) \cdot \mathbf{ds}$ takes place over the surface S' linking the inductor windings. Using the definition of inductance, $L = \Phi/I$, we can write the surface integral as

$$\int_{S'} \frac{\partial \mathbf{B}}{\partial t} \cdot \mathbf{ds} = \frac{\partial}{\partial t} \int_{S'} \mathbf{B} \cdot \mathbf{ds} = \frac{\partial}{\partial t}(LI) = L\frac{dI}{dt}. \tag{10.90}$$

Substituting Equations (10.87)–(10.90) into Equation (10.86), we obtain

$$IR + \frac{1}{C}\int_{-\infty}^{t} I dt + L\frac{dI}{dt} = V_s, \tag{10.91}$$

which is the familiar mesh equation used for circuit analysis.

Obviously, one does not need electromagnetic analysis to evaluate Equation (10.91). Electromagnetic analysis is needed, however, to establish under what circumstances a circuit can be modeled using this simple equation. These circumstances are:

1. A thin conducting wire defines the closed contour of the circuit.
2. All wire resistance is incorporated in lumped resistances.
3. The wire inductance is either neglibible or incorporated into lumped inductance elements.
4. Stray capacitance of the wires to ground (or other circuits) is negligible, so each component in the loop has the same current.

As we shall see at various points throughout the remainder of this text that most of these conditions are met when all the wires and components of a circuit are short, compared with the wavelength of the operating frequency.

10-7 Summation

The concept of displacement current was the last postulate of electromagnetic theory to be discovered. Without it, the equations of electromagnetics would not be capable of correctly modeling many effects, particularly at high frequencies. However, it is rarely necessary to think about the meaning of the term "displacement current." In fact, this can be said about *all* the physical postulates that lead up to Maxwell's equations. As a result, most of our effort in the remainder of the text will be devoted to simply using Maxwell's equations to solve various practical problems. Each of these chapters will center on a specific class of devices or phenomena: transmission lines, plane waves, waveguides, and radiation.

PROBLEMS

10-1 Suppose that an electric field $\mathbf{E} = E_0 x^2 e^{-\alpha t}\hat{\mathbf{a}}_y$ exists in a nonmagnetic ($\mu = \mu_0$), nonconducting ($\sigma = 0$) region. Using Maxwell's curl equations, show that this field can exist only if the permittivity ϵ is a specific function of position. Find this function.

10-2 The addition of Maxwell's displacment current term $\partial\mathbf{D}/\partial t$ to Ampère's law make this law consistent with the law of charge conservation. There are, however, an infinite number of terms that would accomplish the same thing. Prove that any term of the form

$$\frac{\partial \mathbf{D}}{\partial t} + \nabla \times \mathbf{G},$$

where $\mathbf{G}$ is any vector, also accomplishes this. Of course, only the choice $\mathbf{G} = 0$ makes Ampère's law agree with all the other experimental evidence pertaining to time-varying fields.

10-3 When observed at large distances from their sources, the electric and magnetic fields generated by many current and charge distributions can be appoximated as spherically symmetric waves whose amplitudes vary inversely with distance. Consider the following spherical wave in source-free (i.e., $\mathbf{J}_i = 0$) free space:

$$\mathbf{E} = \frac{E_o}{r} \sin(\omega t - \beta r)\,\hat{\mathbf{a}}_\theta,$$

(a) Use Maxwell's curl-E equation to find the magnetic field intensity **H** that is associated with the preceding **E**.

(b) By substituting the **H** found in part a) into Maxwell's curl-**H** equation, find the value of β that allows **E** to satisfy Maxwell's curl equations for large values of r.

(c) What other component of **E** must also be present at small values of r in order to satisfy Maxwell's curl equations?

10-4 Displacement current $\mathbf{J}_d = \partial\mathbf{D}/\partial t$ cannot be neglected at high frequencies in insulators (of which free space is an example). In conductors, however, displacement current can often be ignored when the conduction current $\mathbf{J}_c = \sigma\mathbf{E}$ is high. Find the ratio J_c/J_d in seawater at

(a) $f = 30$ [Hz], where $\epsilon_r = 80$ and $\sigma = 4$ [S/m].

(b) $f = 10$ [GHz], where $\epsilon_r = 80$ and $\sigma = 25$ [S/m].

10-5 A homogeneous dielectric with dielectric constant $\epsilon_r = 8$ and conductivity σ is placed between the plates of a parallel-plate capacitor. Determine the value σ at which the conduction and displacement currents in the dielectric are equal at a) $f = 1$ [kHz] and b) 100 [MHz].

10-6 Suppose that the following H-field exists in a source-free vacuum region:

$$\mathbf{H} = -\frac{\beta}{\mu_o\omega} E_o \sin\alpha x \cos(\omega t - \beta z)\,\hat{\mathbf{a}}_x - \frac{\alpha}{\mu_o\omega} E_o \cos\alpha x \sin(\omega t - \beta z)\,\hat{\mathbf{a}}_z.$$

(a) Use Ampère's law to find the **E** associated with the H-field.

(b) By substituting the **E** found in part a) into Maxwell's curl-**E** equation, show that these E- and H-fields are valid only when $\alpha^2 + \beta^2 = \mu_o\epsilon_o\omega^2$

(c) Prove that these E- and H-fields also satisfy Maxwell's divergence equations.

10-7 Let $\mathbf{H} = H_o \cos(k_x x)\cos(\omega t - \beta z)\,\hat{\mathbf{a}}_z$, where k_x and β are real valued. Find **H** in the frequency domain.

10-8 If $\mathbf{E} = E_o e^{-\alpha z} e^{-j\beta z}\,\hat{\mathbf{a}}_x$, find **E** in the time domain if α and β are real valued.

10-9 Show that for sinusoidally varying fields, the conduction and displacement currents are always 90° out of phase in time when ϵ and σ are real valued.

10-10 Find the displacement current density associated with the following magnetic field in a source-free region of free space:

$$\mathbf{H} = H_o \cos k_x x\, e^{-j\beta z}\,\hat{\mathbf{a}}_x.$$

10-11 Consider a three-dimensional space that is divided into two regions, $z > 0$ and $z < 0$, that have permittivites ϵ_1 and ϵ_2, respectively. If both media are nonconducting ($\sigma_1 = \sigma_2 = 0$) and $\mathbf{E}_1 = \alpha\hat{\mathbf{a}}_x + \beta\hat{\mathbf{a}}_y + \zeta\hat{\mathbf{a}}_z$ at $z = 0^+$, find $\mathbf{E}_2$ at $z = 0^-$.

10-12 Consider a three-dimensional space that is divided into two regions, $z > 0$ and $z < 0$, that have permeabilities μ_1 and μ_2, respectively. If

$$\mathbf{B}_1 = \alpha\hat{\mathbf{a}}_x + \beta\hat{\mathbf{a}}_y + \zeta\hat{\mathbf{a}}_z \quad \text{at} \quad z = 0^+,$$

find $\mathbf{B}_2$ at $z = 0^-$ when a surface current $\mathbf{J}_s = J_x\hat{\mathbf{a}}_x + J_y\hat{\mathbf{a}}_y$ flows along the boundary.

10-13 Figure P10-13 shows an interface between two nonconducting media. If the E- and H-fields in region 1 at the interface are

$$\mathbf{E}_1 = 2\hat{\mathbf{a}}_x + \hat{\mathbf{a}}_y - 3\hat{\mathbf{a}}_z \qquad [\text{V/m}]$$

$$\mathbf{H}_1 = -\hat{\mathbf{a}}_x + 2\hat{\mathbf{a}}_y - 4\hat{\mathbf{a}}_z \qquad [\text{A/m}],$$

find $\mathbf{E}_2$ and $\mathbf{H}_2$ at the interface.

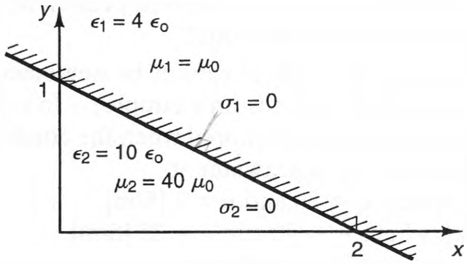


Figure P10-13

10-14 Figure P10-14 shows a rectangular waveguide, which consists of a conducting, rectangular cylinder with cross-sectional dimensions a and b along the x- and y-axis, respectively. In Chapter 13, we will find that a possible magnetic field distribution inside such a waveguide at frequency ω is given by

$$\mathbf{H} = H_o\left[j\beta\left(\frac{a}{\pi}\right)\sin\left(\frac{\pi}{a}x\right)\hat{\mathbf{a}}_x + \cos\left(\frac{\pi}{a}x\right)\hat{\mathbf{a}}_z\right]e^{-j\beta z},$$

where H_o and β are constants. Use the boundary conditions at a perfect conductor to find expressions for the surface current $\mathbf{J}_s$ on the four inside walls of the waveguide.

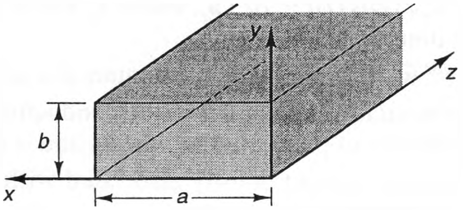


Figure P10-14

11

Transmission Lines

11-1 Introduction

Any configuration of wires and conductors that carries opposing currents is a transmission line. However, when the term "transmission line" is used, it usually refers to a ***uniform transmission line***–two or more conductors that maintain the same cross-sectional dimensions throughout their lengths. Transmission lines are essential components in nearly all electrical systems and devices. Often, they take the form of cables, such as the coaxial cable shown in Figure 11-1a, which consists of a solid center conductor, surrounded by a dielectric core and an outer conductor. The outer conductor can be either solid or braided and is usually grounded at one or both ends. Another common transmission line is a two-wire (or twin-lead) line, shown in Figure 11-1b. Here, two wires are separated by a dielectric that provides mechanical support. When the wires have identical cross sections and the same relationship to ground, the resulting transmission line is called a ***balanced transmission line***.

Transmission lines found on electronic circuit boards are usually planar types, where the conductors lie on flat dielectric sheets. Figures 11-2a–c show examples of microstrip, slot-line, and fin-line transmission lines, respectively. Planar transmission lines are popular because they can be manufactured using the same technology that is

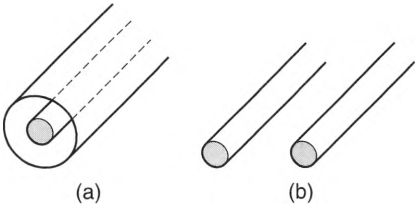


Figure 11-1 Transmission line cables: a) A coaxial cable. b) A two-wire transmission line.

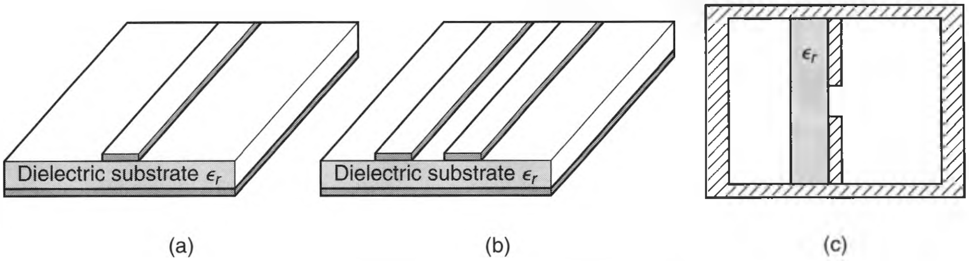

Dielectric substrate ϵ_r
Dielectric substrate ϵ_r
ϵ_r
(a)
(b)
(c)

Figure 11-2 Planar transmission lines: a) A microstrip line. b) A slot line. c) A fin line.

used for printed circuit boards. We will show in this chapter that all conductor traces on ordinary printed circuit boards are really transmission lines, which often do not behave at all like the idealized "wires" used in ordinary circuit analysis, particularly at high frequencies. We will also discuss the major aspects of uniform transmission lines, using both time-domain and frequency-domain analysis.

11-2 TEM Modes on Transmission Lines

When a transmission line is connected to a source (such as an ideal voltage or current source), electric and magnetic fields are induced throughout the line. The way in which these fields distribute themselves is a function of the cross-sectional dimensions of the line, the materials used, the frequency of operation of the line, and the nature of the source. Under the right conditions, any one of an infinite number of distinct electric and magnetic field patterns can be induced on a transmission line. Each of these patterns is called a ***mode***. Because the electric and magnetic fields of each mode are different, each mode has different electrical characteristics.

For a transmission line that lies along the z-axis, we can classify all the possible modes into three basic classes:

TEM modes Transverse-electromagnetic modes, often called transmission-line modes, are characterized by $E_z = 0$ and $H_z = 0$ at all points. Transmission lines that have at least two separate conductors and a homogeneous dielectric can support one TEM mode. This mode is capable of transporting energy and information over a wide band of frequencies, including dc.

Quasi-TEM modes For these modes, E_z and H_z approach zero in the limit as the frequency of operation approaches zero. A single quasi-TEM mode can exist on transmission lines that have at least two conductors and an inhomogeneous dielectric (such as a microstrip transmission line). These modes have nearly the same characteristics as TEM modes and can be analyzed using the same techniques.

Waveguide modes These are modes for which E_z, H_z, or both, are nonzero. Waveguide modes can transport energy or information only when operated above distinct cutoff frequencies. They modes are usually considered to be undesirable on transmission lines and can generally be avoided by operating the line well below their cutoff frequencies.

Any of these three classes of modes can exist on a transmission line. However, as long as the operational frequency is kept below the cutoff frequencies of the waveguide modes, only the TEM or quasi-TEM mode will be transported over large distances. The ***dominant mode*** of a transmission line is its TEM or quasi-TEM mode. For the remainder of this chapter, we will deal exclusively with TEM or quasi-TEM modes. (We will discuss waveguide modes in detail in Chapter 13.) Also, since the quasi-TEM modes are nearly the same as TEM modes, we will call them both TEM modes, unless there is a particular need for a distinction.[1]

11-2-1 CIRCUIT EQUATIONS FOR TEM MODES

As we said in the preceding section, the TEM mode of a transmission line is by far the most desirable mode to use for nearly all practical applications. It is also the easiest mode to model. This is because the electric and magnetic fields of TEM modes produce uniquely determined modal voltage and currents. Since they are scalars, these voltages and currents are relatively easy to measure and model. In this section, we will derive two fundamental equations that characterize these voltages and currents along the line.

Figure 11-3 shows the cross section of a uniform transmission line, aligned so that it runs parallel to the z-axis.

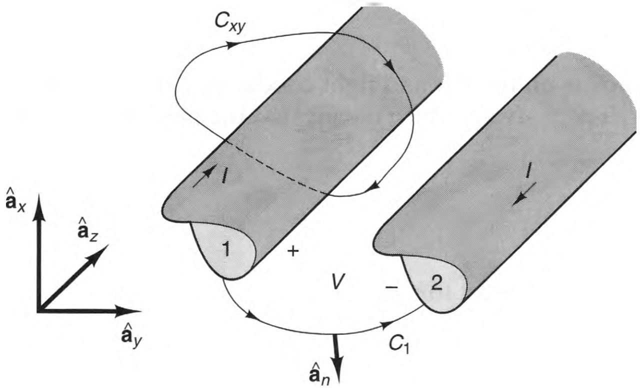


Figure 11-3 Geometry for deriving the equations for voltages and currents on a uniform transmission line. Path C_{xy} is closed and lies in a constant z-plane. Path C_1 extends from conductor 1 to conductor 2 in a constant z-plane.

[1] For a more complete discussion of how quasi-TEM modes differ from TEM modes, see R.E. Collin, *Foundations of Microwave Engineering*, 2d ed. (New York, McGraw Hill: 1992).

If C_{xy} is an arbitrary contour that lies in any constant z-plane, Maxwell's line integral equations read

$$\oint_{C_{xy}} \mathbf{E} \cdot \mathbf{d\ell} = -\frac{\partial}{\partial t}\int_{S_z} \mathbf{B} \cdot \mathbf{ds} \tag{11.1}$$

$$\oint_{C_{xy}} \mathbf{H} \cdot \mathbf{d\ell} = I_{\text{enc}} + \frac{\partial}{\partial t}\int_{S_z} \mathbf{D} \cdot \mathbf{ds}, \tag{11.2}$$

where the surface S_z lies in a constant z-plane and is bounded by the closed path C_{xy}, and I_{enc} is the current passing through S_z in a right-handed sense. Since $E_z = H_z = 0$ for a TEM mode, $\mathbf{B} \cdot \mathbf{ds} = 0$ and $\mathbf{D} \cdot \mathbf{ds} = 0$ everywhere on S_z. Thus, when only a TEM mode is present, Equations (11.1) and (11.2) become

$$\oint_{C_{xy}} \mathbf{E} \cdot \mathbf{d\ell} = 0 \tag{11.3}$$

$$\oint_{C_{xy}} \mathbf{H} \cdot \mathbf{d\ell} = I_{\text{enc}}. \tag{11.4}$$

These equations should look familiar; they are the very same equations satisfied by electrostatic and magnetostatic fields. Hence, we can conclude that the electric and magnetic fields associated with TEM modes distribute themselves throughout constant z-planes just like electrostatic and magnetostatic fields, regardless of the frequency of operation. This is our first clue that we can model TEM modes using ordinary circuit analysis.

Next, let us suppose that conductors #1 and #2 in Figure 11-3 are perfectly conducting and carry currents I and $-I$, respectively. We will also assume that the dielectric material is homogeneous and lossless. We can express the voltage V between the conductors in a constant z-plane by integrating along the path C_1 shown in Figure 11-3; that is,

$$V = -\int_2^1 \mathbf{E} \cdot \mathbf{d\ell} = \int_1^2 \mathbf{E} \cdot \mathbf{d\ell},$$

where "1" and "2" are points on the left and right conductors, respectively. Since C_1 is in the xy-plane, $\mathbf{d\ell} = dx\,\hat{\mathbf{a}}_x + dy\,\hat{\mathbf{a}}_y$, which means that the preceding expression can be written as

$$V = \int_1^2 (E_x dx + E_x dy).$$

Differentiating both sides with respect to z yields

$$\frac{\partial V}{\partial z} = \int_1^2 \left(\frac{\partial E_x}{\partial z}\,dx + \frac{\partial E_y}{\partial z}\,dy\right). \tag{11.5}$$

We can write the integrand of this integral in a more useful form by remembering that $\nabla \times \mathbf{E} = -(\partial \mathbf{B}/\partial t)$ (for all fields) and $E_z = 0$ (for TEM modes). Using these, we can write

$$(\nabla \times \mathbf{E})_x = -\frac{\partial E_y}{\partial z} = -\frac{\partial B_x}{\partial t} \tag{11.6}$$

and

$$(\nabla \times \mathbf{E})_y = \frac{\partial E_x}{\partial z} = -\frac{\partial B_y}{\partial t}. \tag{11.7}$$

Substituting Equations (11.6) and (11.7) into Equation (11.5), we obtain

$$\frac{\partial V}{\partial z} = -\frac{\partial}{\partial t}\int_1^2 (B_y dx - B_x dy), \tag{11.8}$$

noting that the differentiation with respect to time has been brought outside the integral, since the integration path C is stationary. Next, the integrand of Equation (11.8) can be written as a dot product, i.e.,

$$B_y dx - B_x dy = \mathbf{B} \cdot \hat{\mathbf{a}}_n d\ell, \tag{11.9}$$

where

$$\hat{\mathbf{a}}_n = \frac{-dy\,\hat{\mathbf{a}}_x + dx\,\hat{\mathbf{a}}_y}{\sqrt{dx^2 + dy^2}} = \frac{-dy\,\hat{\mathbf{a}}_x + dx\,\hat{\mathbf{a}}_y}{d\ell}. \tag{11.10}$$

is a unit vector that is perpendicular to the path C_1. Substituting Equation (11.9) into Equation (11.8), we obtain

$$\frac{\partial V}{\partial z} = -\frac{\partial}{\partial t}\int_1^2 \mathbf{B} \cdot \hat{\mathbf{a}}_n d\ell. \tag{11.11}$$

The line integral on the right-hand side of Equation (11.11) is the magnetic flux per unit length Φ that passes between the two conductors in a right-handed sense. Remembering that $L = \Phi/I$, where I is the transmission line current and L is the inductance per unit length in [H/m], we can write Equation (11.11) as

$$\frac{\partial V}{\partial z} = -L\frac{\partial I}{\partial t}. \tag{11.12}$$

This expression is the first of two fundamental equations that describe the voltages and currents associated with TEM modes on lossless transmission lines. A similar sequence of steps yields the companion equation

$$\frac{\partial I}{\partial z} = -C\frac{\partial V}{\partial t}, \tag{11.13}$$

where C is the capacitance per unit length of the transmission line in [F/m].

Equations (11.12) and (11.13) were both derived under the assumption that the conductors are perfectly conducting and the dielectric is uniform and lossless. In the real world, neither of these conditions are met. When losses are taken into account, these equations become

$$\frac{\partial I}{\partial z} = -GV - C\frac{\partial V}{\partial t} \tag{11.14}$$

$$\frac{\partial V}{\partial z} = -RI - L\frac{\partial I}{\partial t}, \tag{11.15}$$

where R is the ***resistance per unit length*** of the conductors in [Ω/m] and G is the ***conductivity per unit length*** of the dielectric in [S/m]. The additional terms in these generalized equations are more difficult to derive formally, but are easily justified using simple physical reasoning. For instance, Equation (11.14) states that the rate at which I varies along the line is proportional to $C(\partial V/\partial t)$, which is the capacitive (i.e., displacement) current flowing between the conductors. When the dielectric is lossy, an additional conduction current GV flows in parallel with the capacitive current. Similarly, if the conductors have finite conductivity, an additional term RI is needed in Equation (11.15) to account for the ohmic voltage drop along the line.

The values of the circuit parameters L, C, R, and G depend upon the cross-sectional shape of the transmission line and the materials used. Since the electric and magnetic fields of TEM and quasi-TEM modes distribute themselves throughout the cross section of a transmission line exactly like their electrostatic and magnetostatic counterparts, the values of C, L, and G can be calculated using the techniques discussed earlier in Chapters 6 and 9, respectively. Expressions for R are somewhat harder to come by, since this parameter is related to the penetration of the fields into the conductors. Formulas for the circuit parameters of a number of common transmission lines are presented in Appendix D.

11-2-2 THE UNIT CELL

Since the TEM fields on a transmission line can be described in terms of voltages and currents, it follows that any length of transmission line can be represented as an equivalent network of lumped components. The simplest equivalent circuit is obtained when one considers a very short section. This equivalent circuit is called the ***unit cell*** and is shown in Figure 11-4.

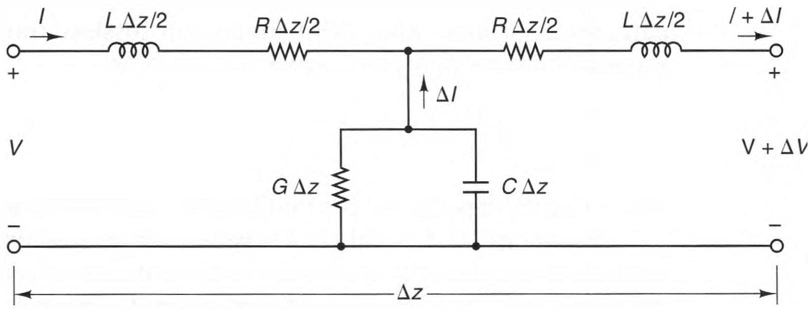


Figure 11-4 The unit cell of an arbitrary transmission line.

In this "T" configuration, the values of the inductance and resistance on either side of the shunt elements are $L\Delta z/2$ and $R\Delta z/2$, respectively, where L and R are the inductance and resistance per unit length, respectively. Similarly, the values of the shunt capacitance and conductance[2] are $C\Delta z$ and $G\Delta z$, where C and G are the capacitance and conductance per unit length of the transmission line.

To show that the voltages and currents on this equivalent circuit are consistent with the transmission-line circuit equations (Equations (11.14) and (11.15)), let us first apply Kirchhoff's voltage law around the outer perimeter of the circuit. A clockwise KVL path around the circuit yields

$$-V + \Delta z\left(\frac{L}{2}\frac{\partial}{\partial t} + \frac{R}{2}\right) I + \Delta z\left(\frac{L}{2}\frac{\partial}{\partial t} + \frac{R}{2}\right)(I + \Delta I) + V + \Delta V = 0,$$

where V and I are the voltage and current at the left-hand terminals, respectively, and $V + \Delta V$ and $I + \Delta I$ are the voltage and current at the right-hand terminals, respectively. As $\Delta z \to 0$, $I + \Delta I \to I$, so

$$RI\,\Delta z + L\Delta z\frac{\partial I}{\partial t} + \Delta V = 0.$$

Dividing both sides by Δz, we find that

$$\lim_{\Delta z\to 0}\frac{\Delta V}{\Delta z} = \frac{\partial V}{\partial z} = -\left(RI + L\frac{\partial I}{\partial t}\right),$$

which is the same as Equation (11.15).

The voltage across the shunt elements $G\Delta z$ and $C\Delta z$ approaches V when Δz is small, so we can express the current ΔI flowing through these elements as

$$\Delta I = -G\Delta z V - C\Delta z\frac{dV}{dt}.$$

Dividing both sides of this expression by Δz, we obtain

$$\lim_{\Delta z\to 0}\frac{\Delta I}{\Delta z} = \frac{\partial I}{\partial z} = -\left(GV + C\frac{\partial V}{\partial t}\right),$$

which is the same as Equation (11.14).

The equivalent circuit shown in Figure 11-4 is exact only in the limit as $\Delta z \to 0$. Nevertheless, it is an excellent approximation when Δz is small. Hence, a transmission line of any length can be modeled as a cascaded chain of unit cells. (See Figure 11-5).

11-3 Transient Voltages and Currents on Transmission Lines

Now that we have developed the differential equations and the circuit parameters that control the voltages and currents on transmission lines, we are ready to investigate how transmission lines respond when driven by sources. Just as in circuit analysis, it is convenient to divide this discussion into two parts: time-domain (i.e., transient) analysis

[2] The shunt resistance is typically called a conductance to distinguish it from the series resistance R.

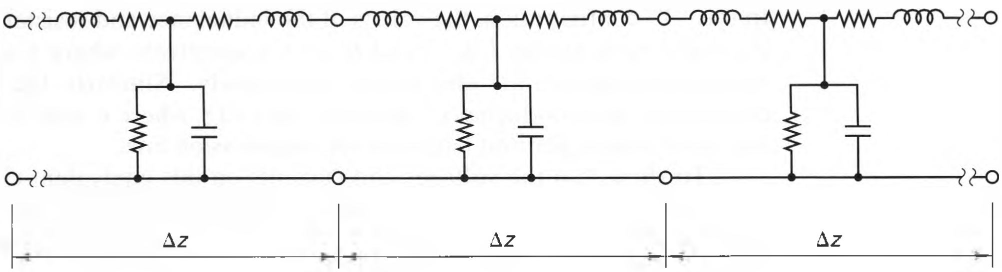


Figure 11-5 An equivalent circuit of a finite length of transmission line, consisting of a cascaded chain of unit cells.

and frequency-domain analysis. First, we will use time-domain analysis to characterize the voltage and current waves that are launched on transmission lines by general, time-domain sources. Later, we will develop frequency-domain techniques for modeling transmission-line responses due to steady-state sources.

11-3-1 TRANSIENT WAVES ON LOSSLESS TRANSMISSION LINES

When no losses are present, $R = G = 0$. For this case, the transmission-line equations become

$$\frac{\partial V}{\partial z} = -L\frac{\partial I}{\partial t} \tag{11.16}$$

$$\frac{\partial I}{\partial z} = -C\frac{\partial V}{\partial t}. \tag{11.17}$$

These are simple differential equations, but they are coupled, since they both contain V and I. To obtain an equation that contains only V, let us first differentiate Equation (11.16) with respect to z, obtaining

$$\frac{\partial^2 V}{\partial z^2} = -L\frac{\partial^2 I}{\partial z\partial t} = -L\frac{\partial^2 I}{\partial t\partial z}. \tag{11.18}$$

Here we have assumed that I is a "well-behaved" function, so the order of differentiation with respect to z and t can be interchanged. Next, if we differentiate Equation (11.17) with respect to t, we have

$$\frac{\partial^2 I}{\partial t\partial z} = -C\frac{\partial^2 V}{\partial t^2}. \tag{11.19}$$

Substituting Equation (11.19) into Equation (11.18), we obtain a differential equation in terms of only V:

$$\frac{\partial^2 V}{\partial z^2} = LC\frac{\partial^2 V}{\partial t^2}. \tag{11.20}$$

We can derive a similar equation for the current I by a similar sequence of steps:

$$\frac{\partial^2 I}{\partial z^2} = LC\frac{\partial^2 I}{\partial t^2}. \tag{11.21}$$

Equations (11.20) and (11.21) are called ***one-dimensional wave equations***.

Propagating Voltage Waves. To understand the nature of the voltages that can exist on lossless transmission lines, let us start by stating the general solution of the voltage wave equation (Equation (11.20)):

$$V(t, z) = V^+(t - z/u) + V^-(t + z/u). \tag{11.22}$$

Here,

$$u = \frac{1}{\sqrt{LC}} \quad \text{[m/s]}, \tag{11.23}$$

and $V^+(t)$ and $V^-(t)$ are arbitrary functions of a single variable, called ***waveform functions***. To verify that Equation (11.22) satisfies Equation (11.20), we note that the second derivatives of V with respect to t and z are

$$\frac{\partial^2 V}{\partial t^2} = (V^+)''\,(t - z/u) + (V^-)''\,(t + z/u) \tag{11.24}$$

and

$$\frac{\partial^2 V}{\partial z^2} = \frac{1}{u^2}(V^+)''\,(t - z/u) + \frac{1}{u^2}(V^-)''\,(t + z/u), \tag{11.25}$$

where $(V^+)''$ and $(V^-)''$ are the second derivatives of V^+ and V^-, respectively. Substituting Equations (11.24) and (11.25) into Equation (11.20), we find that Equation (11.22) is indeed the general solution of the wave equation for all waveform functions $V^+(t)$ and $V^-(t)$, provided that the parameter u is given by Equation (11.23).

The voltage expression given by Equation (11.22) consists of waves that travel along the transmission line. To show this, let us for the moment assume that $V^-(t) = 0$. For this case, the voltage expression becomes

$$V(t, z) = V^+(t - z/u).$$

Figure 11-6 shows $V(t, z)$ as a function of time t for three values of z when $V^+(t)$ is a "pulselike" function. As can be seen, the same waveform is observed at each position, with a time delay that increases linearly with z. Since the waveform shape is the same for all values of z, we call this ***distortionless*** (or ***dispersionless***) ***propagation***. This is a characteristic of all lossless transmission lines. To calculate the propagation velocity, let us observe how fast the value of z must change in order for an observer to "ride" on the same point of the pulse as it moves. This occurs when the argument of V^+ remains constant as time progresses:

$$t - z/u = \text{constant}.$$

Differentiating both sides of the preceding expression with respect to t, we obtain

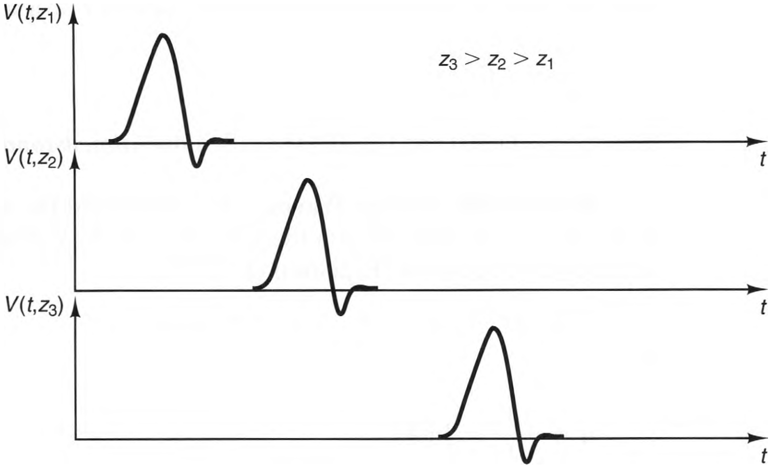


Figure 11-6 A forward-propagating voltage pulse, measured at three points along a lossless transmission line.

$$\frac{dz}{dt} = \frac{1}{\sqrt{LC}} = u.$$

Thus, we can conclude that the waveform $V^+(t - z/u)$ travels (i.e., propagates) towards increasing values of z at a rate of

$$u = \frac{1}{\sqrt{LC}} \quad [\text{m/s}], \tag{11.26}$$

where u is called the ***velocity of propagation***. Waves propagating towards increasing values of z are called ***forward-propagating waves***.

Returning to the general voltage expression given by Equation (11.22), let us now consider the case where $V^+ = 0$ and $V^- \neq 0$. For this case, we have

$$V(t, z) = V^-(t + z/u).$$

To "ride" on the same point of this waveform, we must maintain

$$t + z/u = \text{constant}.$$

Differentiating both sides of this expression with respect to t, we obtain

$$\frac{dz}{dt} = -u = -\frac{1}{\sqrt{LC}},$$

which means that the term $V^-(t + z/u)$ represents a wave traveling in the $-z$ direction at the rate $|u| = 1/\sqrt{LC}$. We will call waves propagating in this direction ***backward-propagating waves***.

Forward-propagating and backward-propagating waves can exist on a transmission line simultaneously. When that happens, both V^+ and V^- are nonzero. This often occurs when an incident wave reflects off a lumped load at the end of a transmission line, as we will show later. When oppositely directed waveforms pass by each other, however, they do not affect each other.

Propagating Current Waves. Associated with each traveling-wave voltage is a traveling-wave current. To show this, we first remember from Equation (11.21) that the current $I(t, z)$ satisfies exactly the same one-dimensional wave equation that $V(t, z)$ does:

$$\frac{\partial^2 I}{\partial z^2} = LC \frac{\partial^2 I}{\partial t^2}. \tag{11.27}$$

Hence, just as with $V(t, z)$, solutions for $I(t, z)$ are always of the form

$$I(t, z) = I^+(t - z/u) + I^-(t + z/u), \tag{11.28}$$

where u is given by Equation (11.26). Although it may appear from Equation (11.28) that the waveform functions $I^+(t)$ and $I^-(t)$ are arbitrary, they have the same shapes as the forward-propagating and backward-propagating voltage waveform functions, $V^+(t)$ and $V^-(t)$, respectively. To show this, substitute Equations (11.28) and (11.22) into Equation (11.16), obtaining

$$-\frac{1}{u}V^+(t - z/u) + \frac{1}{u}V^-(t + z/u) = -LI^+(t - z/u) - LI^-(t + z/u).$$

Both sides of this equation will be equal for all values of t and z only when

$$\frac{V^+(t)}{I^+(t)} = R_o \tag{11.29}$$

and

$$\frac{V^-(t)}{I^-(t)} = -R_o, \tag{11.30}$$

where

$$R_o = \sqrt{\frac{L}{C}} \quad [\Omega]. \tag{11.31}$$

As a result, the current $I(t, z)$ can be written as

$$I(t, z) = \frac{1}{R_o}V^+(t - z/u) - \frac{1}{R_o}V^-(t + z/u). \tag{11.32}$$

The parameter R_o is called the ***characteristic resistance*** of the transmission line. This name is a logical one, since it is measured in ohms and is the ratio of a voltage and a current. However, this resistance is *not* like lumped resistors, which dissipate electrical energy. Rather, the characteristic resistance of a transmission line is an indication of its ability to transport energy via the propagation of voltage and current waves.

Later, when we use phasor analysis to describe time-harmonic waveforms on lossy transmission lines, we will find that the voltage and current waves are related by the characteristic impedance Z_0, which is complex for lossy lines. However, for lossless lines, $Z_o = R_o$, so the variables R_o and Z_o and the terms "characteristic resistance" and "characteristic impedance" can be used interchangeably.

Example 11-1

Calculate the characteristic resistance R_o of RG–58/U coaxial cable which has a solid inner conductor with radius a = 0.406 [mm] and a braided outer conductor with radius b = 1.553 [mm]. Assume that the dielectric is polyethylene, which has a dielectric constant of 2.26.

Solution:

We have already found the distributed capacitance and inductance for this cable in Examples 6-1 and 9-10, respectively:

$L = 0.268$ [μH/m]

$C = 93.73$ [pF/m].

Substituting these values into Equation (11.31) yields

$$R_o = \sqrt{\frac{0.268 \times 10^{-6}}{93.73 \times 10^{-12}}} = 53.47\ [\Omega],$$

which agrees well with the nominal value of 53.5 [Ω] for this type of cable. We can also obtain this result by using Equation (D.12) in Appendix D. Remembering that $R_o = Z_o$ for lossless transmission lines, we find that

$$R_o = \frac{1}{2\pi}\sqrt{\frac{\mu_o}{\epsilon'}}\ \ln\left(\frac{b}{a}\right) = \frac{1}{2\pi}\sqrt{\frac{4\pi \times 10^{-7}}{2.26 \times 8.854 \times 10^{-12}}}\ \ln\left(\frac{1.553}{0.406}\right)$$

$$= 53.47\ [\Omega]$$

Example 11-2

Calculate the characteristic resistance R_o of a microstrip transmission line shown in Figure 11-7, where w = 3 [mm], h = 2 [mm] and ϵ_r = 2.3.

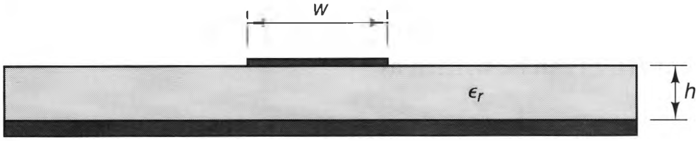


Figure 11-7 A microstrip transmission line.

Solution:

Since $W/h = 1.5 > 1$, we can use equations (D.20) and (D.21). Substituting the values of w, h, and ϵ_r, we find that

$$\epsilon_{\text{eff}} = \frac{1}{2}(2.3 + 1) + \frac{1}{2}(2.3 - 1)\left(1 + \frac{12}{1.5}\right)^{-1/2}$$
$$= 1.87$$

and

$$R_o = Z_o = \frac{120\pi}{\sqrt{1.87}}\{1.5 + 1.393 + 0.667 \ln[1.5 + 1.444]\}^{-1}$$
$$= 76.3 \; [\Omega].$$

Power Transport. As might be expected, the presence of propagating waves on a transmission line is an indication that power is being transported. To show this, consider the situation depicted in Figure 11-8, which shows a section of a uniform transmission line that is carrying both forward-propagating and backward-propagating waves. Viewing this section of transmission line as a one-port network with its terminals at the plane $z = z'$, we see that the power P^+ entering the terminals equals the product of the voltage across the terminals and the current into the positive terminal; that is,

$$P_{\text{in}} = VI,$$

where V and I are the total voltage and current at $z = z'$, respectively, and the positive direction for I is directed towards the right. Substituting $V = V^+ + V^-$ and $I = I^+ + I^-$ into this expression, we obtain

$$P_{\text{in}} = [V^+ + V^-][I^+ + I^-] = V^+I^+ + V^+I^- + V^-I^+ + V^-I^-.$$

Since $V^+ = R_oI^+$ and $V^- = -R_oI^-$, the second and third terms on the right-hand side cancel, and we can write

$$P_{\text{in}} = P^+ - P^-, \tag{11.33}$$

where

$$P^+ = V^+I^+ = \frac{(V^+)^2}{R_o} = R_o(I^+)^2 \quad [\text{W}] \tag{11.34}$$

and

$$P^- = V^-I^- = \frac{(V^-)^2}{R_o} = R_o(I^-)^2 \quad [\text{W}]. \tag{11.35}$$

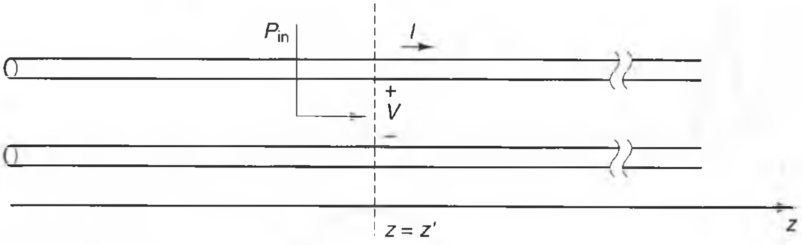


Figure 11-8 Power transport on a transmission line. P_{in} is the power entering the $z = z'$ plane from the left.

Since P^+ and P^- are always positive, we can interpret them as the power entering and leaving the terminal pair, respectively, at $z = z'$. Thus, P^+ is the forward-propagating power transported by the forward-propagating voltage and current waves. Conversely, P^- is the backward-propagating power transported by the backward-propagating voltage and current waves.

11-3-2 WAVE PROPAGATION ON LOSSY TRANSMISSION LINES

When conductor or dielectric losses are present, V and I satisfy Equations (11.14) and (11.15); hence,

$$\frac{\partial I}{\partial z} = -GV - C\frac{\partial V}{\partial t} \tag{11.36}$$

$$\frac{\partial V}{\partial z} = -RI - L\frac{\partial I}{\partial t}. \tag{11.37}$$

We can obtain an equation involving only V by differentiating Equations (11.36) and (11.37) with respect to z and t, respectively, and substituting one into the other:

$$\frac{\partial^2 V}{\partial z^2} = LC\frac{\partial^2 V}{\partial t^2} + (RC + LG)\frac{\partial V}{\partial t} + RGV. \tag{11.38}$$

Conversely, if we differentiate Equations (11.36) and (11.37) with respect to t and z, respectively, and substitute one into the other, we obtain an equation in terms of I alone:

$$\frac{\partial^2 I}{\partial z^2} = LC\frac{\partial^2 I}{\partial t^2} + (RC + LG)\frac{\partial I}{\partial t} + RGI. \tag{11.39}$$

These two equations are similar to those obtained for the lossless case (Equations (11.20) and (11.21)), except that they have some terms which change the nature of the solutions.

General solutions of Equations (11-38) and (11-39) cannot be expressed easily in the time domain. This is because waves launched on lossy transmission lines normally do not maintain the same waveshape as they propagate. Typically, waveforms spread out in time and diminish in amplitude as they propagate on lossy lines. Figure 11-9 depicts such a case. Here, a forward-propagating pulse has a waveshape at $z = 0$ of $V^+(z = 0)$. At $z > 0$, however, the waveshape $V^+(z > 0)$ is not only delayed, but also

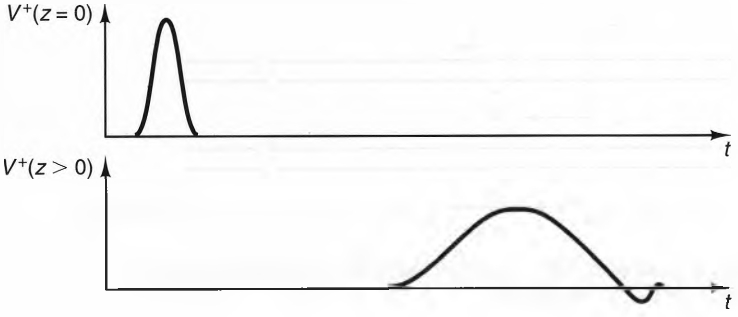


Figure 11-9 A pulse waveform at two points on a lossy transmission line.

attenuated and distorted. This waveform distortion is called ***dispersion*** and is usually an undesirable characteristic of lossy lines.

Lossy lines are easily modeled using frequency-domain analysis, which we will discuss later in this chapter. However, there is one special case where the waves on lossy lines can be easily represented in the time domain. This is called the ***dispersionless*** (or ***nondistorting***) ***case*** and occurs when R, C, L, and G satisfy the relation

$$RC = GL. \tag{11.40}$$

For this case, the general solutions of Equations (11.38) and (11.39) are of the form

$$V(t, z) = e^{-\alpha z} V^{+}(t - z/u) + e^{+\alpha z} V^{-}(t + z/u) \tag{11.41}$$

and

$$I(t, z) = \frac{1}{R_o} e^{-\alpha z} V^{+}(t - z/u) - \frac{1}{R_o} e^{+\alpha z} V^{-}(t + z/u), \tag{11.42}$$

where

$$u = \frac{1}{\sqrt{LC}} \qquad [\text{m/s}], \tag{11.43}$$

$$\alpha = R\sqrt{\frac{C}{L}} = G\sqrt{\frac{L}{C}} \qquad [\text{m}^{-1}], \tag{11.44}$$

$$R_o = \sqrt{\frac{L}{C}} \qquad [\Omega], \tag{11.45}$$

and $V^{+}(t)$ and $V^{-}(t)$ are arbitrary functions of t. Equations (11.41) and (11.42) can be verified by direct substitution into Equations (11.38) and (11.39). The terms in Equations (11.41) and (11.42) that contain $e^{-\alpha z} V^{+}(t - z/u)$ represent forward-propagating waves that grow exponentially weaker when they are observed at increasing values of z. The constant α has units of inverse meters and is called the ***attenuation constant*** of the wave. Similarly, the terms that contain $e^{\alpha z} V^{-}(t + z/u)$ represent backward-propagating waves that become exponentially weaker when observed at increasingly negative values of z. Although both waves decay exponentially as they propagate, their waveshapes do not become distorted.

Oliver Heaviside discovered this nondistorting case in 1887 while studying the performance of long-distance telegraph circuits. At that time, it was known that the maximum rate at which telegraph signals could be transmitted on a line varied inversely with the length of the line, but the reason for this phenomenon was a subject of fierce debate. Heaviside was the first to discover why distortion occurs on lossy lines and theorized that it could be reduced or eliminated on practical transmission lines by adjusting the line parameters so that Equation (11.40) is satisfied. The initial engineering application of Heaviside's idea to telegraph lines was done by George Campbell and, to a lesser extent, Micheal Pupin[3].

[3] See Paul Nahin, *Oliver Heaviside: Sage in Solitude* (New York: IEEE Press, 1988).

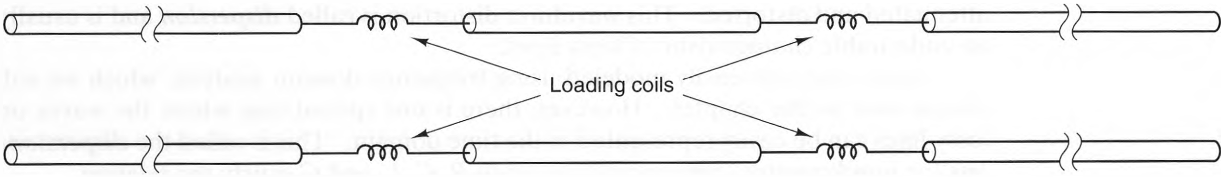


Figure 11-10 Loading coils placed periodically on a transmission line to reduce dispersion.

On practical transmission lines, the values of R, C, L, and G are usually such that $RC > GL$. To obtain a distortionless line, either RC must be decreased, or GL must be increased. This could be accomplished by increasing the conductor spacing, which increases L and decreases C. However, that world result in unacceptable conductor spacings. Another unacceptable solution is to add shunt conductance to the line to increase G, which also increases the attenuation constant α. Heaviside proposed raising L by placing series inductors (called *loading coils*) periodically along the line, as depicted in Figure 11-10. This simple procedure was an immediate success and is still used on analog telephone links to allow distortionless transmission throughout the voice band (0–3 [KHz]).

Example 11-3

A twisted-pair telephone cable transmission line has the following parameters:

$R = 0.107$ [Ω/m]

$L = 543$ [nH/m]

$C = 51.3$ [pF/m]

$G = 51.0$ [pΩ/m].

Find the loading-coil inductance that must be added at each kilometer of the line in order to obtain distortionless propagation.

Solution:

Using Equation (11.40) and the specified values of R, G, and C, we find that the required value of the inductance is

$$L' = \frac{RC}{G} = \frac{0.107 \times 51.3 \times 10^{-12}}{51.0 \times 10^{-12}} = 0.108 \text{ [H/m]}.$$

Using $L' = L + L_C$, where L_C is the loading-coil inductance per meter, we find that

$$L_C = 0.108 \text{ [H/m]} - 543 \text{ [nH/m]} \approx 108 \text{ [H/km]}.$$

11-3-3 LAUNCHING WAVES ON TRANSMISSION LINES

A wave can be launched on a transmission line simply by attaching a voltage across its terminals. Figure 11-11a depicts such a situation. Here, an independent voltage generator $V_g(t)$ and a resistor R_g are connected to the end of an infinite, lossless transmis-

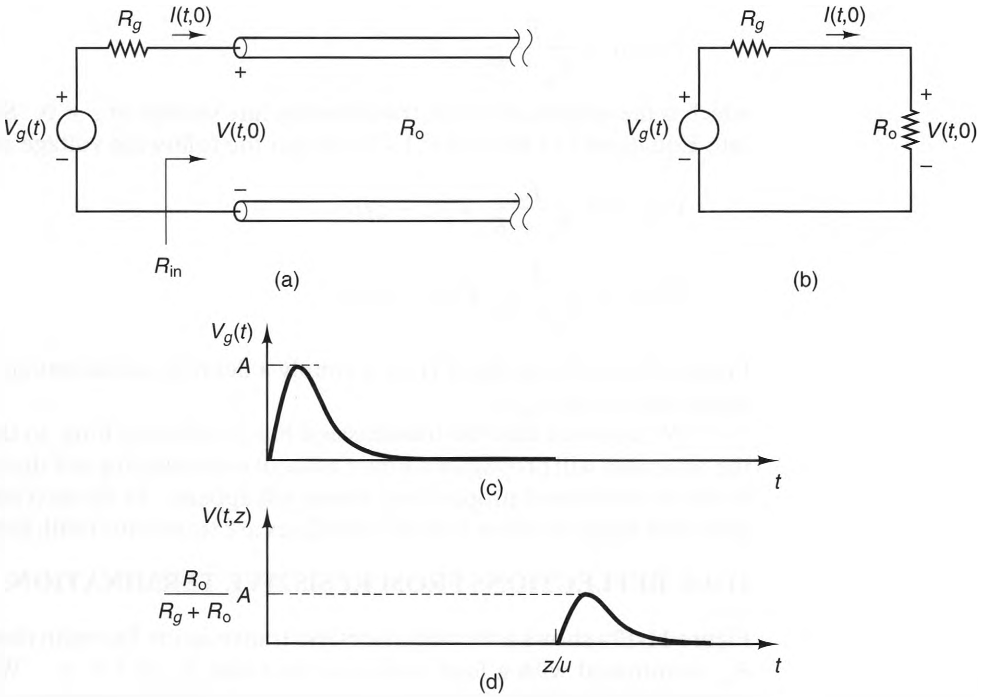


Figure 11-11 Launching waves on a transmission line: a) A voltage generator connected to an infinite transmission line. b) The equivalent circuit as seen by the generator circuit. c) The voltage generator waveform. d) The voltage waveform observed a distance z along the transmission line.

sion line. The waveshape of $V_g(t)$ is shown in Figure 11-11c; it has a peak amplitude of A. Because the line is infinitely long, the total voltage and current on the line consist only of forward-propagating waves; that is,

$$V(t, z) = V^+(t - z/u) \tag{11.46}$$

and

$$I(t, z) = I^+(t - z/u) = \frac{1}{R_o} V^+(t - z/u). \tag{11.47}$$

Since only a forward-propagating wave exists on the line, the resistance R_{in} looking into the line at $z = 0$ is the same for all time t:

$$R_{in} = \frac{V(t, 0)}{I(t, 0)} = \frac{V^+(t - 0/u)}{\frac{1}{R_o} V^+(t - 0/u)} = R_o.$$

Because of this, the input circuit can be redrawn as shown in Figure 11-11b, where the infinite transmission line has been replaced by a resistor of value R_o. Using the voltage divider relation, we obtain

$$V(t, 0) = \frac{R_o}{R_g + R_o} V_g(t), \tag{11.48}$$

which is the amplitude of the transmission line voltage at $z = 0$. Substituting this result into Equations (11.46) and (11.47), we get the following voltage and current waves:

$$V(t, z) = \frac{R_o}{R_g + R_o} V_g(t - z/u) \tag{11.49}$$

$$I(t, z) = \frac{1}{R_g + R_o} V_g(t - z/u). \tag{11.50}$$

Figure 11-11d shows that $V(t, z)$ is simply a delayed and attenuated version of the generator waveform $V_g(t)$.

We assumed that the transmission line is infinitely long, so the waves launched by the generator will propagate forever without encountering any discontinuities. Because of this, no backward-propagating waves will appear. In the next section, we will investigate what happens when transmission lines are terminated with lumped resistors.

11-3-4 REFLECTIONS FROM RESISTIVE TERMINATIONS

Figure 11-12a shows a section of lossless transmission line with characteristic resistance R_o, terminated with a load resistance of value R_L at $z = z'$. We will assume that a source far off to the left of the figure has launched forward-propagating (or incident) voltage and current waveforms that are described by

$$V_{\text{inc}}(t, z) = V^+(t - z/u) \tag{11.51}$$

and

$$I_{\text{inc}}(t, z) = I^+(t - z/u) = \frac{1}{R_o} V^+(t - z/u), \tag{11.52}$$

where $V^+(t - z/u)$ has a peak amplitude of A. If we assume that the waveform $V^+(t)$ is zero for $\tau < 0$, the leading edges of the incident waves will not reach the load until $t = \ell/u$. Thus, $V^+(t - z/u)$ and $I^+(t - z/u)$ are the only waves on the line for $t < \ell/u$.

When the incident waves reach the load, backward-propagating waves will be initiated at the load if $R_L \neq R_o$. To see why, let us suppose that only the forward-propagating waves are present on the line for all values of t. If this were the case, the load voltage $V_L(t)$ and current $I_L(t)$ would simply be the incident waves, evaluated at $z = \ell$; that is,

$$V_L(t) = V(t, \ell) = V^+(t - \ell/u)$$

and

$$I_L(t) = I(t, \ell) = \frac{1}{R_o} V^+(t - \ell/u).$$

However, at the load, the ratio of the voltage and current must equal the load resistance R_L:

$$\frac{V_L(t)}{I_L(t)} = R_L. \tag{11.53}$$

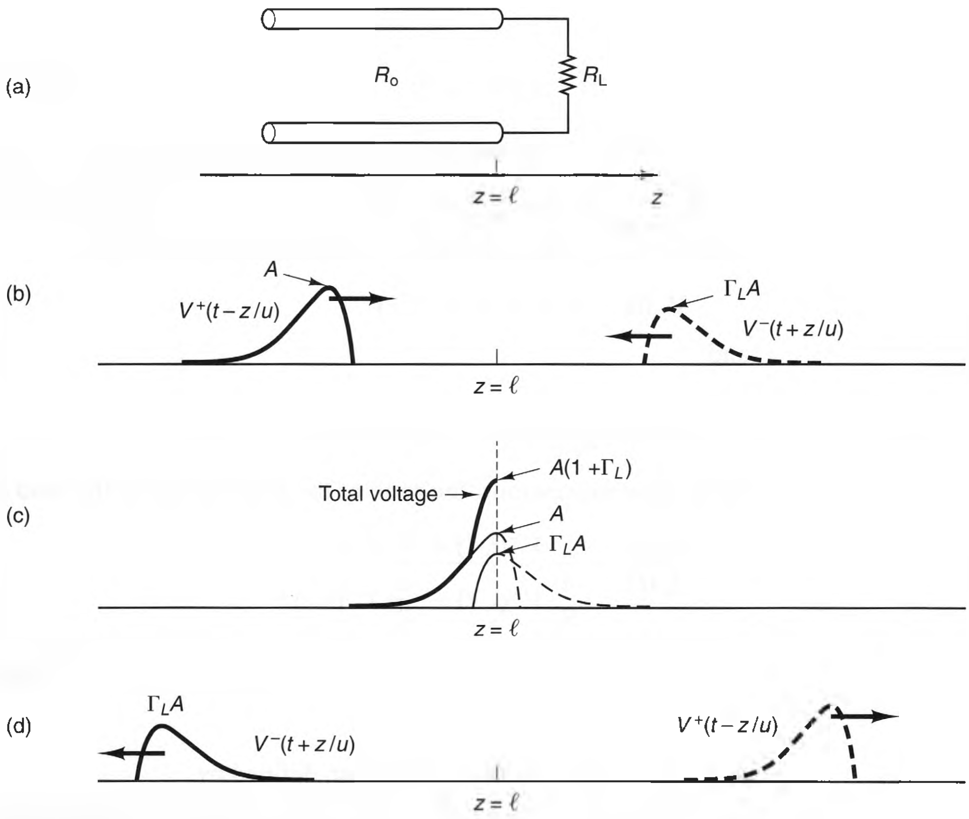


Figure 11-12 The process of reflection at a resistive load: a) A transmission line terminated by a resistor. b)–d) Line voltage along the line before, during, and after the incident pulse reaches the resistor, respectively.

Substituting the expressions for $V_L(t)$ and $I_L(t)$ into this equation, we find that the load resistance must be

$$R_L = R_o.$$

A load that is equal to the characteristic resistance produces no reflections and is called a ***matched load.*** When $R_L \neq R_o$, Equation (11.53) is not satisfied, which means that the incident waves alone cannot satisfy the conditions of both the transmission line and the load.

To model the case where $R_L \neq R_o$, let us again assume that the same forward-propagating waves $V^+(t - z/u)$ and $I^+(t - z/u)$ are incident from the left in Figure 11-12a, but this time let us also speculate that reflected, backward-propagating waves are also present. Hence, the total voltage and current on the line are given by

$$V(t, z) = V^+(t - z/u) + V^-(t + z/u) \tag{11.54}$$

and

$$I(t,z) = I^+(t - z/u) + I^-(t + z/u) = \frac{1}{R_o} V^+(t - z/u) - \frac{1}{R_o} V^-(t + z/u), \qquad (11.55)$$

where $V^-(t)$ is a yet-to-be-determined reflected waveform. Also, note that the negative polarity of the reflected current $I^-(t + z/u)$ occurs because this wave is backward propagating. (See Equation (11.32)). Evaluating these expressions at $z = \ell$, we find that the voltage and current at the load are, respectively,

$$V_L(t) = V^+(t, \ell) + V^-(t, \ell)$$

and

$$I_L(t) = \frac{1}{R_o} V^+(t, \ell) - \frac{1}{R_o} V^-(t, \ell).$$

From these expressions, the ratio of the load voltage to the load current is

$$\frac{V_L(t)}{I_L(t)} = \frac{V^+(t,\ell) + V^-(t,\ell)}{\frac{1}{R_o} V^+(t,\ell) - \frac{1}{R_o} V^-(t,\ell)}.$$

Setting this expression equal to the load resistance R_L and solving for $V^-(t)$, we obtain

$$V^-(t, \ell) = \Gamma_L V^+(t, \ell)\,, \qquad (11.56)$$

where Γ_L is the ***reflection coefficient***, defined by

$$\Gamma_L \equiv \left.\frac{V^-(t)}{V^+(t)}\right|_{\substack{\text{at the}\\ \text{load}}} = \frac{R_L - R_o}{R_L + R_o}. \qquad (11.57)$$

Substituting Equation (11.56) into Equations (11.54) and (11.55), we obtain the complete expressions for the voltage waves on a terminated line:

$$V(t,z) = V^+(t - z/u) + \Gamma_L V^+(t + (z - 2\ell)/u) \qquad (11.58)$$

$$I(t,z) = \frac{1}{R_o}\Big(V^+(t - z/u) - \Gamma_L V^+(t + (z - 2\ell)/u)\Big). \qquad (11.59)$$

These voltage and current waves satisfy the requirements of both the transmission line and the load resistance.

From Equation (11.58), we see that the reflected voltage waveform has the same shape as the incident waveform, with an amplitude that is governed by the reflection coefficient Γ_L. For passive load resistances ($R_L \geqslant 0$), Γ_L has a magnitude that is always less than or equal to unity:

$$-1 \leqslant \Gamma_L \leqslant 1 \quad (R_L \geqslant 0). \qquad (11.60)$$

Notice that $\Gamma_L = 0$ when $R_L = R_o$, which means that no reflection is generated by a matched load. For this case, all the power in the incident voltage and current waves is dissipated by the load.

Figures 11-12b–d shows the incident, reflected, and total voltages on the terminated transmission line for three values of t. When $t_1 < \ell/u$ (Figure 11-12b), the leading edge of the incident wave has yet to reach the load, so only the incident wave appears on the line. Even so, it is convenient to show the yet-to-appear reflected wave as a dotted curve to the right of the load position ($z = \ell$) that propagates towards the left. The peak amplitude of this reflected wave is $\Gamma_L A$, where A is the peak amplitude of the incident wave. Figures 11-12c and 11-12d show the voltages at two instants in time after the incident waveform has reached the load. In these plots, the incident waveform is drawn as a dotted line in the region $z > L$ to remind us that this region of the graph does not represent actual points on the transmission line. In Figure 11-12c, the incident and reflected waveforms appear simultaneously across the load, since the reflected wave is generated at the load the instant the incident wave appears. Figure 11-12d shows that once the incident wave has encountered the load, only the reflected wave is left on the line (assuming that there is no mismatch at the generator).

Finally, we can also describe reflections in terms of the current waves. To do this, we substitute $I^+(t, 0) = V^+(t, 0)/R_o$ and $I^-(t, 0) = - V^-(t, 0)/R_o$ into Equation (11.56) to obtain

$$\left.\frac{I^-(t)}{I^+(t)}\right|_{\substack{\text{at the}\\ \text{load}}} = -\Gamma_L.$$

Hence, the current reflection coefficient is simply the negative of the voltage reflection coefficient Γ_L.

11-3-5 STEP RESPONSE OF TRANSMISSION LINES

We are now ready to discuss the full transient response of transmission lines that are terminated at both ends. To introduce this topic, consider the setup shown in Figure 11-13a. Here, a transmission line with characteristic resistance $R_o = 50$ [Ω] and length $\ell = 3$ [m] is connected to a load resistor $R_L = 100$ [Ω]. The source consists of a 12 [V] battery, a resistor $R_g = 10$ [Ω], and a switch that closes at $t = 0$. Also, the velocity of propagation is $u = 3 \times 10^8$ [m/s], so the one-way propagation delay from end to end is 10 [ns].

When the switch closes at $t = 0$, a step waveform is launched towards the load with an amplitude V_1 given by Equation (11.49):

$$V_1 = \frac{50}{50 + 10} 12 = 10 \text{ [V]}.$$

For $0 < t < 10$ [ns], this is the only voltage wave on the line. Figure 11-13b shows the line voltages at $t = 7$ [ns].

At $t = 10$ [ns], the leading edge of the incident waveform reaches the load, where a reflected wave is produced. The reflection coefficient at the load end is

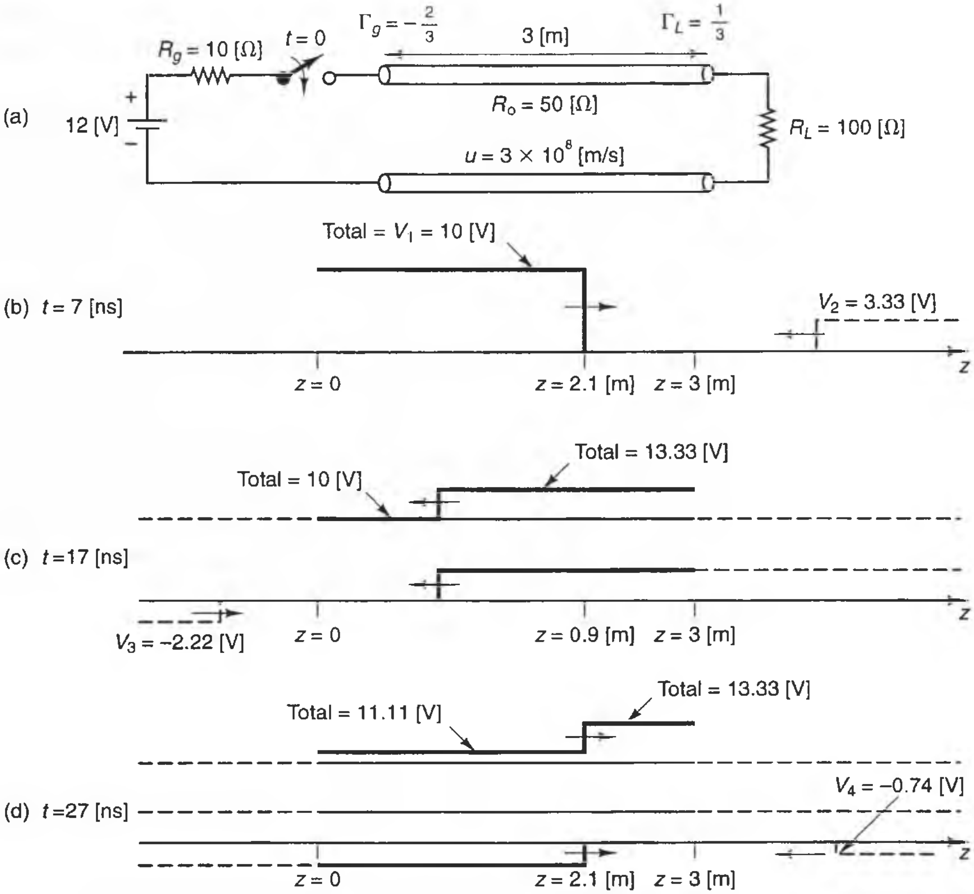


Figure 11-13 Transient response of a transmission line, switched at $t = 0$: a) The circuit. b)–d) Voltage waveforms on the line at three points in time. The arrows show the propagation directions of the leading edges of the waveforms.

$$\Gamma_L = \frac{100 - 50}{100 + 50} = \frac{1}{3},$$

so the first reflected wave has amplitude

$$V_2 = V_1 \Gamma_L = 10 \times \frac{1}{3} = 3.3333\ [\text{V}].$$

Figure 11-13c shows the line voltages at $t = 17$ [ns].

The first reflection from the load reaches the generator terminals at $t = 20$ [ns]. Since the generator resistance is not matched to the transmission line, a reflected wave will be produced that propagates towards the load. The amplitude of this reflected wave is not affected by the battery, because, according to the superposition principle, the battery voltage has already been accounted for in the first forward-propagating wave (launched at $t = 0$). The reflection coefficient at the generator end is

$$\Gamma_g = \frac{10 - 50}{10 + 50} = -\frac{2}{3},$$

so the reflection of V_2 off the generator resistance is

$$V_3 = \Gamma_g V_2 = -\frac{2}{3} \times 3.3333 = -2.222 \text{ [V]}.$$

Figure 11-13d shows the voltages on the line at $t = 27$ [ns].

By now, the method for determining the subsequent reflections on the line should be obvious. To determine the N^{th} reflection at either the generator or load, all that must be known is the amplitude of the approaching $(N - 1)^{\text{th}}$ wave and the reflection coefficient. In this way, the total voltages on the line can be considered as an infinite sum of reflections. Since the reflection coefficients of passive loads have magnitudes less than or equal to the previous one, the higher order reflections eventually have negligible amplitudes. As a result, the step response of a transmission line approaches a constant value along the entire line as $t \rightarrow \infty$.

Figures 11-13b through 11-13d show "snapshots" of the voltages on the line at various points in time. Plots like these give a global picture of how the waves reflect and rereflect off the terminations. Another useful way to determine the step response of a transmission line is by using a ***bounce diagram***, such as the one shown in Figure 11-14. In a bounce diagram, the progression of the leading edges of the incident and reflected voltage waves are displayed as functions of both time and position. In Figure 11-14, the line marked V_1^+ indicates the progress of the leading edge of the wave launched by the generator as it propagates towards the load. This line starts at ($t = 0$,

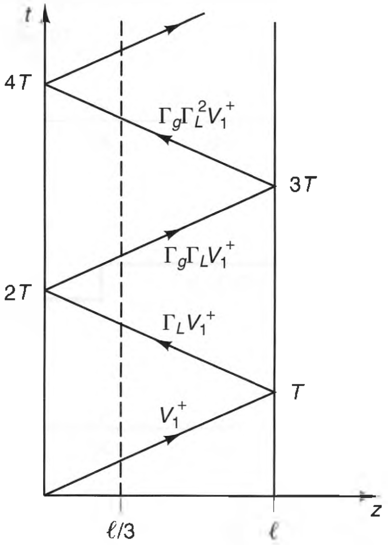


Figure 11-14 A transmission line bounce diagram.

$z = 0$) and ends at ($t = T, z = \ell$), where $T = \ell/u$ is the one-way transit time. The line marked $\Gamma_L V_1^+$ represents the first reflection off the load. This line starts at $t = T$ and has a negative slope, since it represents a backward-propagating wave. In like manner, each of the subsequent reflections are represented by lines that have alternating positive and negative slopes and begin at progressively later times.

To obtain the voltage waveform $V(t, z')$ at a point $z = z'$ on a transmission line, we first draw a vertical line at $z = z'$ on the bounce diagram. Next, starting at $t = 0$ and $z = z'$, we progress vertically on this line, noting the times t_n at which this line intersects the lines representing each wave. At each value t_n, the waveform $V(t, z')$ will exhibit a step discontinuity equal to the value of the newest wave arriving at that point. Figures 11-15a and b show $V(t, z')$ at $z' = 1$ [m] and $z' = 3$ [m], respectively, for the transmission-line network of Figure 11-13a when the line has length $\ell = 3$ [m]. In particular, the waveform at $z' = 3.0$ [m] has fewer jumps in this time interval than the waveform at $z' = 1.5$ [m]. This is because an observer at the load ($z' = 3.0$ [m]) sees the leading edges of the incident wave and its reflection simultaneously, whereas an observer in the center of the transmission line sees them at different times.

Finally, bounce diagrams for the transmission-line currents are the same as voltage bounce diagrams, except that the voltage reflection coefficients, Γ, are replaced with current reflection coefficients, $-\Gamma$.

Rise Time. As we said earlier, every trace on a printed circuit board (PCB) is, in fact, a transmission line. For digital circuits, transmission-line behavior becomes important when the rise time of the signal is less than or comparable to the propagation delay between the signal source and the load.

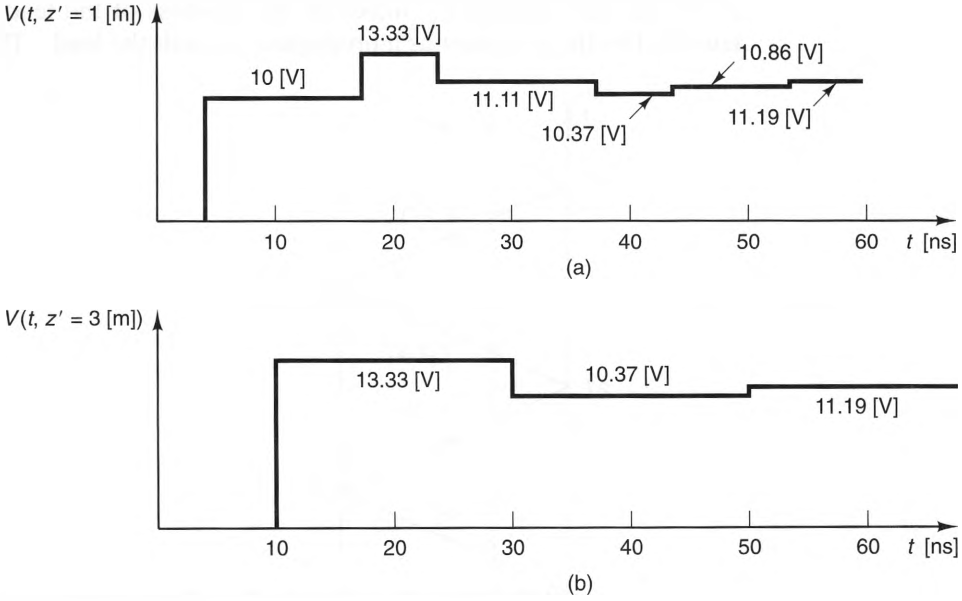


Figure 11-15 Voltage waveforms for the circuit in Figure 11-13a: a) $z' = 1$ [m]. b) $z' = 3$ [m].

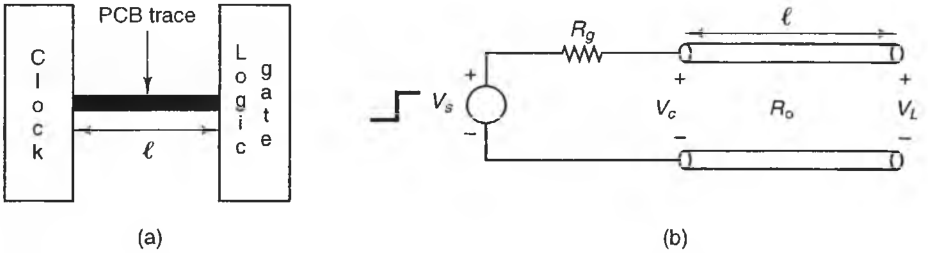


Figure 11-16 Rise time in digital circuits: a) A digital clock and a CMOS logic gate, connected by a PCB trace. b) The equivalent circuit.

Figure 11-16a shows a digital clock generator, connected to the input pin of a CMOS (Complementary Metal-Oxide Semiconductor) logic gate via a PCB trace of length ℓ. Since the input impedance of CMOS gates is extremely high, we can model this circuit with the simplified equivalent circuit shown in Figure 11-16b. In this equivalent circuit, the clock is represented by a unit-step generator in series with a resistance R_g, the PCB trace is replaced by a transmission line with characteristic resistance R_o, and the input terminals of the logic gate are modeled by an open circuit.

Using the analysis techniques we developed in this section, we can calculate the transient voltages at the terminals of the clock and the gate when the step waveform in the clock switches at $t = t'$. Figures 11-17a through 11-17c show these waveforms for three different cases: $R_g \gg R_o$, $R_g = R_o$, and $R_g \ll R_o$, respectively. In each case, the

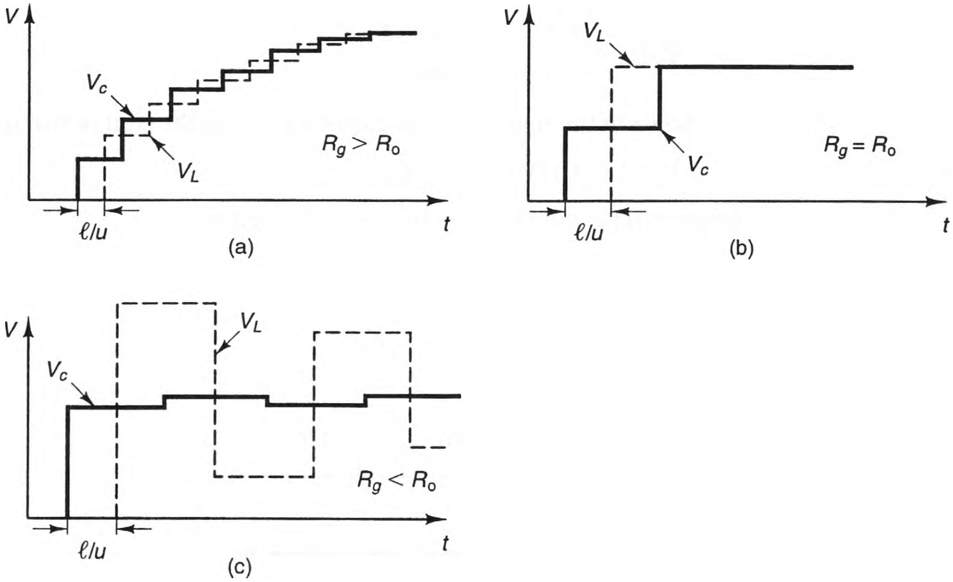


Figure 11-17 Step-response waveforms for the clock circuit shown in Figure 11-16 a) $R_g > R_o$, b) $R_g = R_o$, c) $R_g < R_o$.

logic gate voltage V_L is shown as a dotted curve, and the clock output voltage V_c is shown as a solid curve. As can be seen, these three cases produce very different step responses. Note that the time scale in the first plot is different than in the other two in order to show the entire rise time.

From Figure 11-17a we see that the clock and gate waveforms resemble "staircased" exponentials when $R_g \gg R_o$. For this case, the transmission line can be approximated as a shunt capacitance of value $C\ell$, where C is the capacitance per meter of the transmission line and the 10% to 90% rise time of the gate voltage is approximately

$$t_r \approx 2.2 R_g C\ell. \tag{11.61}$$

As a result, when $R_g \gg R_o$, the rise time is proportional to the length of the PCB trace.

Faster rise times can be attained when the value of R_g approaches R_o. As can be seen from Figure 11-17b, the case where $R_g = R_o$ produces responses with zero rise time. If, however, $R_g < R_o$ (such as in Figure 11-17c), the waveforms exhibit overshoot and ringing, since the reflection coefficients at the opposite ends of the transmission line have opposite signs. This situation is clearly undesirable for digital circuit applications, as ringing requires additional delay in synchronous signal lines and can cause logic errors in asynchronous signal lines.

11-3-6 PULSE RESPONSE OF TRANSMISSION LINES

The pulse response of a transmission-line network is easily found from its step response. To see why, consider the ideal pulse function $\mathrm{P}_T(t)$, shown in Figure 11-18 and defined by

$$\mathrm{P}_T(t) = \begin{cases} 1 & 0 < t < T \\ 0 & \text{otherwise} \end{cases}. \tag{11.62}$$

As shown in the figure, we can express $\mathrm{P}_T(t)$ as the sum of two unit-step functions, i.e.,

$$\mathrm{P}_T(t) = \mathrm{U}(t) - \mathrm{U}(t - T), \tag{11.63}$$

where $\mathrm{U}(t)$ is the unit-step function, defined by

$$\mathrm{U}(t) = \begin{cases} 0 & t < 0 \\ 1 & t \geqslant 0 \end{cases}. \tag{11.64}$$

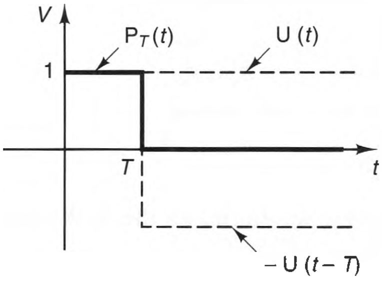


Figure 11-18 An ideal pulse function.

Using Equation (11.63) and the superposition principle, we can express the pulse response of a transmission line as the difference of two unit-step responses. We can also use bounce diagrams to calculate pulse responses. This is accomplished simply by plotting lines on the bounce diagram that indicate the position of both the leading and trailing edges of the pulse, as demonstrated by the following example.

Example 11-4

The air-dielectric transmission line shown in Figure 11-19 is $\ell = 0.09$ [m] long and has characteristic resistance $R_o = 50$ [Ω]. The generator and load resistances are $R_g = 25$ [Ω] and $R_L = 100$ [Ω], respectively. The generator voltage has an amplitude of 12 [V]. Plot the load voltage waveforms when the pulse width is

a) $T = 0.1$ [ns]
b) $T = 0.8$ [ns]

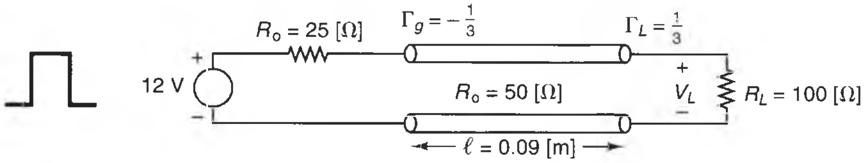


Figure 11-19 A loaded transmission line with a pulse input.

Solution:

The generator and load reflection coefficients are

$$\Gamma_g = \frac{25 - 50}{25 + 50} = -\frac{1}{3}$$

$$\Gamma_L = \frac{100 - 50}{100 + 50} = \frac{1}{3}.$$

Since the dielectric is air, $u \approx 3 \times 10^8$ [m/s], and the one-way propagation delay is

$$\Delta t = \frac{0.09}{3 \times 10^8} = 0.3 \text{ [ns]}.$$

a) The bounce diagram corresponding to the $T = 0.1$ [ns] generator pulse width is shown in Figure 11-20a. Here, the solid lines represent the leading edges of the pulses, whereas the dotted lines represent the trailing edges of the pulses. The voltage $V_L(t)$ across the load is shown in Figure 11-20b and is obtained by summing the contributions along the $z = 0.09$ line in Figure 11-20a. Notice that the incident pulse and its reflections are distinct, since the pulse width is shorter than the propagation delay.

b) Figures 11-21 a and b show the bounce diagram and load voltage waveform, respectively, when the pulse width is $T = 0.8$ [ns]. In this case, the first-reflected pulse arrives at the load before the incident pulse is finished. Notice that since the pulse width is longer than the propagation delay, the output waveform V_L has roughly the same shape as the input waveform.

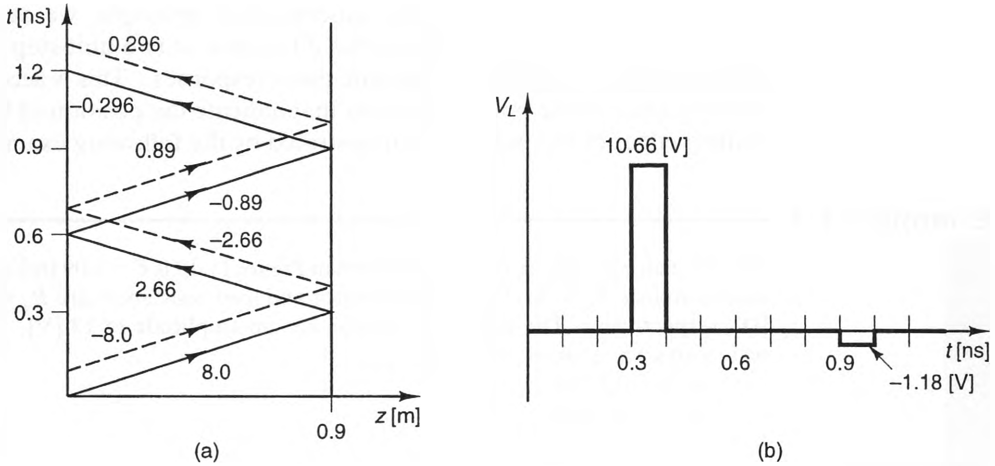


Figure 11-20 Pulse response of circuit in Figure 11-19 when $T = 0.1$ [ns]: a) Bounce diagram. b) Load voltage.

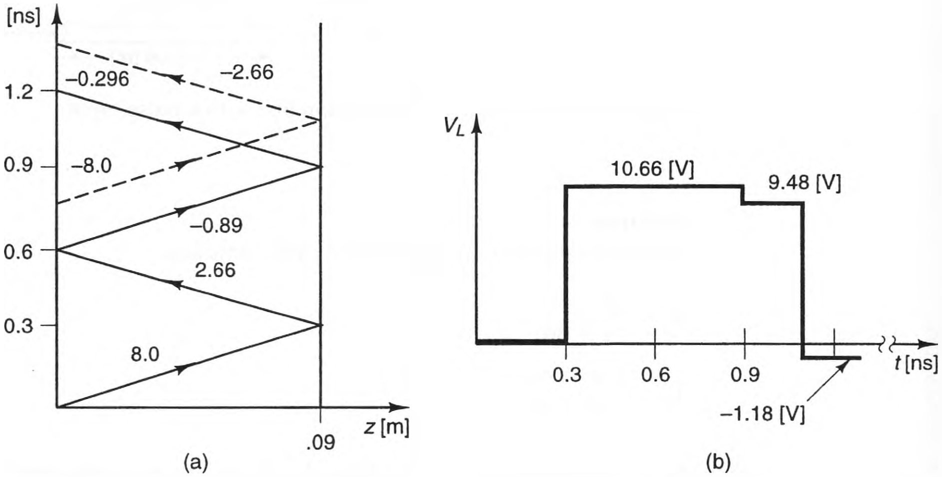
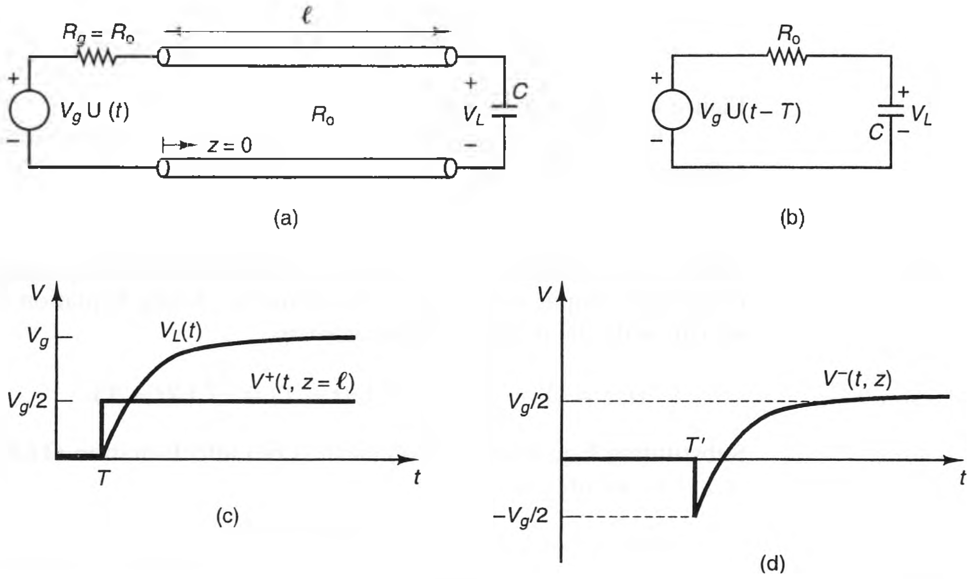


Figure 11-21 Pulse response of circuit in Figure 11-19 when $T = 0.8$ [ns] : a) Bounce diagram. b) Load voltage.

11-3-7 REFLECTIONS FROM REACTIVE LOADS

When a transmission line is terminated with a load that contains either inductance or capacitance, the reflected waveforms have different shapes than those of the incident waves. This happens because a reactive load presents a time-varying impedance to the line, resulting in reflection coefficients that vary with time. In this section, we will develop a procedure for calculating the step responses of transmission lines that are terminated in reactive loads and driven by sources that are matched to the line.



(a) (b) (c) (d)

Figure 11-22 Reflections from a capacitive load: a) The circuit. b) The equivalent circuit at the capacitor terminals. c) The incident and load voltages at the capacitor terminals. d) The reflected waveform.

Let us start by considering the network shown in Figure 11-22a, which consists of a lossless transmission line of length ℓ and characteristic resistance R_o, excited by a step voltage generator with impedance R_o, and terminated with a capacitor of value C. The easiest way to analyze this network is to replace the transmission line and generator with its Thévenin equivalent circuit, as shown in Figure 11-22b. We can find the Thévenin resistance by noting that when the voltage source is turned off, the resistance seen looking into the terminals is R_o, since a matched load simulates an infinite transmission line. Also, we can find the Thévenin voltage $V_{th}(t)$ by calculating the open-circuit voltage $V_{oc}(t)$ at the output terminals. Knowing that the voltage reflection coefficient at an open circuit is +1, we find

$$V_{th}(t) = V_{oc}(t) = V_g \mathrm{U}(t - T).$$

Using the equivalent circuit shown in Figure 11-22b, we can calculate the load voltage $V_L(t)$ at the capacitor by means of standard circuit analysis. The general form of $V_L(t)$ is

$$V_L(t) = \{V_L(\infty) + [V_L(T^+) - V_L(\infty)]e^{-(t-T)/\tau}\}\mathrm{U}(t - T), \qquad (11.65)$$

where τ is the time constant and $V_L(\infty)$ and $V_L(T^+)$ are the capacitor voltages at $t = \infty$ and $t = T^+$, respectively. For this circuit,

$$\tau = R_o C$$

$$V_L(T^+) = V_L(T^-) = 0$$

$$V_L(cc) = V_g,$$

which yields a time response of

$$V_L(t) = V_g[1 - e^{-(t-T)/\tau}]U(t - T). \tag{11.66}$$

Figure 11-22c shows a plot of $V(t)$.

The voltage $V_L(t)$ can also be considered to be the sum of the incident and reflected waves; that is,

$$V_L(t) = V^+(t, \ell) + V^-(t, \ell), \tag{11.67}$$

where $V^+(t, z)$ and $V^-(t, z)$ are the incident (forward-propagating) and reflected (backward-propagating) waves, respectively. Using Equation (11.49) with $R_g = R_o$, we can write the incident voltage wave as

$$V^+(t, z = \ell) = \frac{R_o}{R_o + R_o} V_g U(t - T) = \frac{V_g}{2} U(t - T). \tag{11.68}$$

Substituting Equations (11.68) and (11.66) into Equation (11.67) and solving for V^- $(t, z = \ell)$, we obtain

$$V^-(t, z = \ell) = V_g \left[\frac{1}{2} - e^{-(t-T)/\tau}\right] U(t - T). \tag{11.69}$$

The reflected wave propagates towards decreasing values of z, so its leading edge arrives at a point $z = z'$ at time $t = T'$, where

$$T' = \ell/u + (\ell - z')/u = 2T - z'/u. \tag{11.70}$$

Thus, the reflected wave can be written as

$$V^-(t, z) = V_g \left[\frac{1}{2} - e^{-(t-T')/\tau}\right] U(t - T'). \tag{11.71}$$

This waveform is plotted in Figure 11-22d. As can be seen, it does not have the same shape as the incident waveform.

The foregoing procedure can be used to determine the waves reflected from any reactive load. The example that follows considers the response of an inductive load.

Example 11-5

Calculate the waveform reflected from the inductive load shown in Figure 11-23a.

Solution:

The Thévenin equivalent circuit seen by the load is shown in Figure 11-23b. The general form of the load voltage V_L is given by Equation (11.65), where

$$V_L(T^+) = V_g$$

$$V_L(\infty) = 0$$

$$\tau = \frac{L}{R_o}.$$

Substituting these values into Equation (11.65), we obtain

$$V_L(t) = V_g e^{-(t-T)/\tau} U(t - T).$$

Figure 11-23c shows $V_L(t)$ and the incident voltage wave $V^+(t, \ell)$.

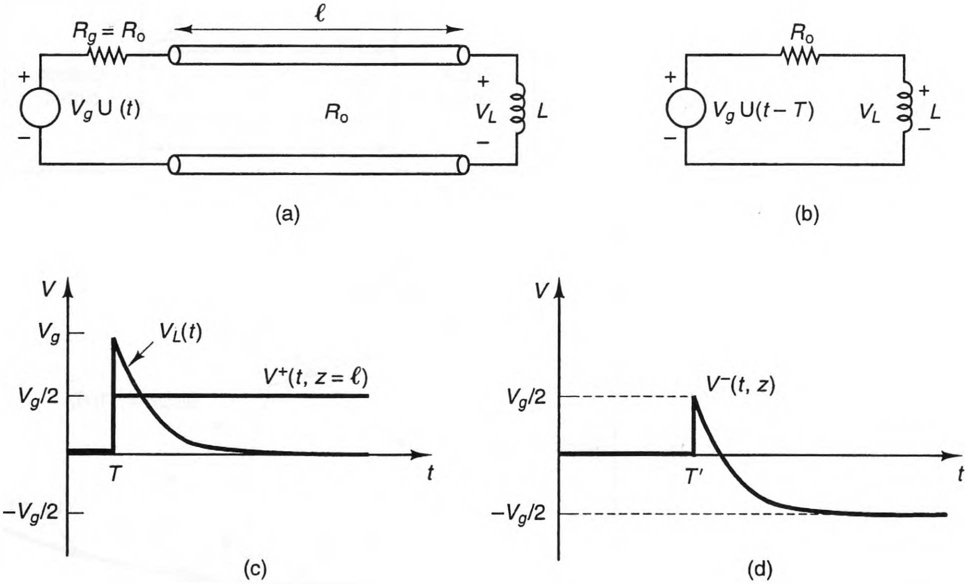


Figure 11-23 Reflection from a inductive load: a) The circuit. b) The equivalent circuit at the inductor terminals. c) The incident and load voltages at the inductor terminals. d) The reflected waveform.

Remembering that the load voltage $V_L(t)$ is the sum of the incident and reflected waves (Equation (11.67)), we find that the reflected voltage at the load is

$$V^-(t, L) = V_g\left[-\frac{1}{2} + e^{-(t-T)/\tau}\right]\mathrm{U}(t-T).$$

Finally, the reflected waveform $V^-(t, z')$ at an arbitrary position z' on the line is obtained by replacing T with

$$T' = \ell/u + (\ell - z)/u = 2T - z'/u.$$

This waveform is plotted in Figure 11.23d.

The transient response of a transmission line contains information about all the discontinuities present along the line. This fact is used in a procedure called ***time-domain reflectometry***, where step waveforms (or pulses) are launched down a transmission line that has unknown loads or discontinuities. By monitoring the voltage at the input terminals of the transmission line, the location and characteristics of the loads and discontinuities along the line can be determined. This is accomplished by noting the position and wave shape of the reflected waveforms. The technique is particularly effective for finding faults in buried lines, since the cable itself does not need to be disturbed during the measurement.

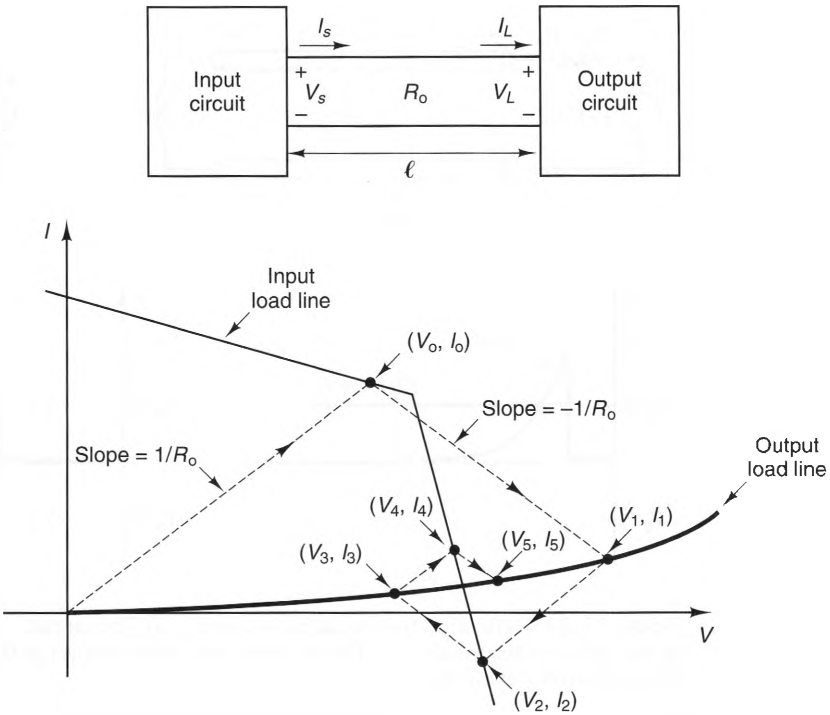


Figure 11-24 Transmission-line response for nonlinear loads: a) Nonlinear input and output circuits, connected by a nondistorting transmission line. b) Bergeron diagram.

11-3-8 TRANSMISSION LINES WITH NONLINEAR LOADS

There are many instances where transmission lines are connected to nonlinear loads. This often occurs in digital circuits, where PCB traces connect chips whose input and output circuits contain nonlinear elements, such as diodes. Transmission lines with nonlinear loads are difficult to analyze when the sources have arbitrary pulse shapes. However, a simple graphical procedure can be used to determine the step response when the nonlinear loads do not contain inductance or capacitance. This method was developed originally by L. J. B. Bergeron,[4] and the resulting graphs are called ***Bergeron diagrams***.

Figure 11-24a shows nonlinear source and load circuits, connected by a lossless transmission line. The $V - I$ plots (called load lines) of the input and output circuits are indicated by the solid curves in Figure 11-24b. These load lines can have any shape, but we will restrict our analysis to the case where the output circuit load line must pass through the origin (0,0), which means that waves can be launched only by the input circuit.

[4] Begeron published his work in French in 1949. A good tutorial summary can be found in P.J. Langlois, "Graphical Analysis of Delay Line Waveforms: A Tutorial," *IEEE Transactions on Education*, vol. 38, no. 1, February 1995, pp. 27–32.

The transient response at the input and output terminals can be found using the following procedure:

1. For $t < 2T$ (where T is the one-way transit time), reflections from the load have not yet reached the input terminals, so the transmission line presents a constant resistance of value R_o to the source circuit. This resistance is represented by a load line with slope $1/R_o$ that passes through the origin (shown as a dotted line in Figure 11-24b). The voltage and current (V_o, I_o) at the source terminals during this time interval occur at the intersection of the input-circuit load line and the dotted line, as shown in the figure.
2. During the time interval $T < t < 3T$, the initial incident wave propagating towards the output terminals appears to come from a Thévenin equivalent circuit that consists of a voltage source $2V_o$ in series with a resistance R_o. This is represented by a load-line with a slope $-1/R_o$ that passes through the point (V_o, I_o), shown as a dotted line in Figure 11-24b. The voltage and current (V_1, I_1) at the output terminals during this time interval occur at the intersection of this line with the output-circuit load line.
3. During the time interval $2T < t < 4T$, the equivalent circuit seen looking towards the load from the source terminals is a voltage source $2(V_1 - V_0)$ in series with a resistor R_o. This corresponds to a load line with slope $1/R_o$ that passes through the point (V_1, I_1), shown as a dotted line in Figure 11-24b. The voltage and current (V_2, I_2) at the output terminals during this time interval occur at the intersection of this line with the input-circuit load line.
4. The procedure for finding subsequent values of (V_n, I_n) is the same. Using this procedure, we find that input terminal voltages and currents have even-numbered subscripts (e.g., $V_o, V_2, \ldots$) and occur at intersections with the input-circuit load line. Output terminal voltages and currents have odd-numbered subscripts and occur at intersections with the output-circuit load line. Eventually, the source and load voltages (and currents) approach the same values, which occur at the intersection of the input and output load lines.

The logic behind these steps may seem difficult to follow, but the procedure itself is quite easy, as is demonstrated by the following example.

Example 11-6

Plot $V_s(t)$ and $V_L(t)$ for the circuit shown in Figure 11-25. Assume that the diode is ideal and that the transmission line is uncharged at $t = 0^-$. Also, assume that the one-way propagation delay on the transmission line is 10 [ns].

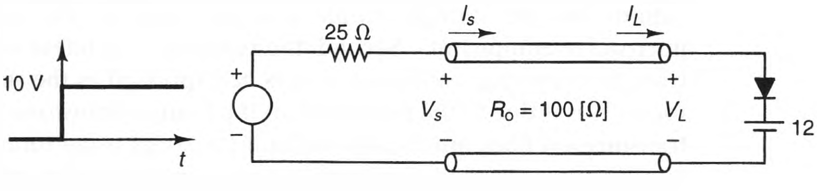


Figure 11-25 A transmission line with a diode output circuit.

Solution:

If we apply Kirchhoff's voltage law around the around the input circuit, we obtain the input-circuit load-line expression

$$V_s = 10 - 25I_s,$$

where I_s is measured left to right. The straight line representing this input load line is shown in Figure 11-26. Similarly, we can obtain the output-circuit load line by noticing that the diode is reverse biased when $V < 12$ [V], so $I_L = 0$ for all V_L less than 12 [V]. On the other hand, the diode is forward biased when $I_L > 0$, so the voltage at the terminals of the output circuit equals the battery voltage (12 [V]) for all $I_L > 0$. Hence, the output-circuit load line consists of horizontal and vertical lines that intersect at the point (0,12[V]).

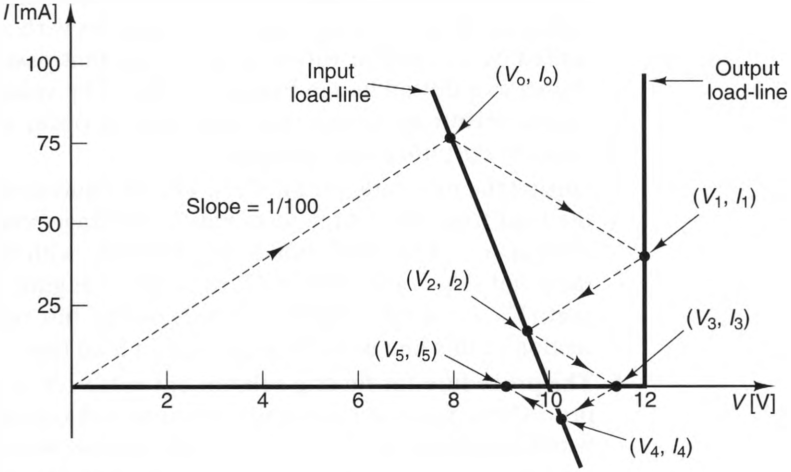


Figure 11-26 Bergeron diagram for the circuit in Figure 11-25.

Using the graphical procedure described, we obtain the $V - I$ plot shown in Figure 11-26. The waveforms of $V_s(t)$ and $V_L(t)$ are shown in Figures 11-27a and b, respectively.

11-4 Time-Harmonic Waves on Transmission Lines

The time-harmonic response of transmission lines is an important special case of the general time-varying case for three reasons. First, many practical engineering applications involve time-harmonic sources, such as the local oscillators in RF and microwave equipment. Second, the response of a linear network line due to an arbitrary time-varying source can always be expressed as the sum of time-harmonic waveforms; this is due to the properties of the Fourier transform. And third, whereas lossy transmission lines are usually difficult to model using time-domain analysis, they are relatively easy to model when the sources are time harmonic.

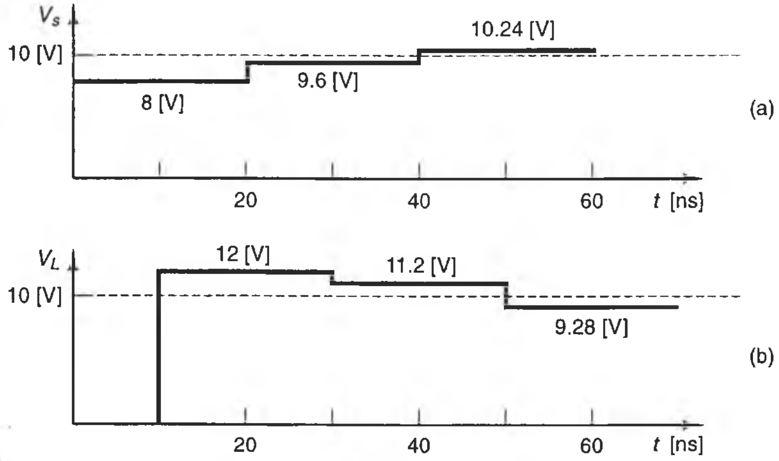


Figure 11-27 Response of the nonlinear circuit shown in Figure 11-25: a) Voltage waveform at the source terminals. b) Voltage waveform at the load terminals.

As in the time-domain analysis of the previous sections, our analysis of transmission lines with time-harmonic sources starts with the transmission-line equations:

$$\frac{\partial V}{\partial z} = -RI - L\frac{\partial I}{\partial t}$$

$$\frac{\partial I}{\partial z} = -GV - C\frac{\partial V}{\partial t}.$$

When all the sources have exactly the same frequency ω, the frequency-domain form of these equations can be derived using the transform rules outlined in Section 10-4-1. Thus, the preceding equations become

$$\frac{\partial V}{\partial z} = -(R + j\omega L)I \tag{11.72}$$

$$\frac{\partial I}{\partial z} = -(G + j\omega C)V, \tag{11.73}$$

where V and I are the phasor representations of the transmission-line voltage and current, respectively. These equations can be decoupled by differentiating one with respect to z and substituting it into the other, resulting in the wave equations

$$\frac{\partial^2 V}{\partial z^2} = \gamma^2 V \tag{11.74}$$

$$\frac{\partial^2 I}{\partial z^2} = \gamma^2 I, \tag{11.75}$$

where

$$\gamma = \sqrt{(R + j\omega L)(G + j\omega C)} \quad [\mathrm{m}^{-1}]. \tag{11.76}$$

The constant γ is called the ***propagation constant*** (for reasons which will soon be obvious) and is a function of the transmission-line parameters R, L, G, and C, as well as the frequency ω. In general, γ is a complex number. Hence, we can write

$$\gamma = \alpha + j\beta, \tag{11.77}$$

where α and β, called the ***attenuation*** and ***phase constants***, respectively, are given by

$$\alpha = \mathrm{Re}[\gamma] = \mathrm{Re}[\sqrt{(R + j\omega L)(G + j\omega C)}] \tag{11.78}$$

$$\beta = \mathrm{Im}[\gamma] = \mathrm{Im}[\sqrt{(R + j\omega L)(G + j\omega C)}]. \tag{11.79}$$

By convention, the principle values of the square roots are implied in Equations (11.78) and (11.79), so $\alpha, \beta \geqslant 0$.

Solutions of Equations (11.74) and (11.75) are easy to find, since they are both second-order, linear, homogeneous differential equations. Their general solutions are

$$V = V^{+}e^{-\gamma z} + V^{-}e^{+\gamma z} \tag{11.80}$$

and

$$I = I^{+}e^{-\gamma z} + I^{-}e^{+\gamma z}, \tag{11.81}$$

where V^{+}, V^{-}, I^{+}, and I^{-} are all constants (possibly complex). We will prove shortly that the terms with $e^{-\gamma z}$ represent forward-propagating waves, whereas the terms with $e^{+\gamma z}$ represent backward-propagating waves.

Just as for the general time-domain case, V and I are not independent quantities. To show this, we can substitute Equations (11.80) and (11.81) into Equations (11.72) and (11.73). Comparing like terms, we obtain

$$\frac{V^{+}}{I^{+}} = Z_{\mathrm{o}} \tag{11.82}$$

and

$$\frac{V^{-}}{I^{-}} = -Z_{\mathrm{o}}, \tag{11.83}$$

where Z_{o} is the ***characteristic impedance*** of the line, given by

$$Z_{\mathrm{o}} = \sqrt{\frac{R + j\omega L}{G + j\omega C}} = \frac{\gamma}{G + j\omega C} = \frac{R + j\omega L}{\gamma} \quad [\Omega]. \tag{11.84}$$

The current I can now be written as

$$I = \frac{V^{+}}{Z_{\mathrm{o}}}e^{-\gamma z} - \frac{V^{-}}{Z_{\mathrm{o}}}e^{+\gamma z}. \tag{11.85}$$

In the sections that follow, we will determine the behavior of these time-harmonic waves on both lossless and lossy transmission lines.

11-4-1 TIME-HARMONIC WAVES ON LOSSLESS TRANSMISSION LINES

On lossless transmission lines, the distributed conductor resistance R and dielectric conductance G are both zero. For this case, the characteristic impedance Z_o is real valued and given by

$$Z_o = \sqrt{\frac{L}{C}} \quad [\Omega] \quad (R = G = 0). \tag{11.86}$$

Comparing Equation (11.86) with Equation (11.31), we see that the characteristic impedance is exactly the same as the characteristic resistance seen by time-domain waveforms on lossless lines. Evaluating Equation (11.76) for the case of zero loss, we find that the propagation constant γ is imaginary; thus,

$$\gamma = \alpha + j\beta \xrightarrow[R=G=0]{} j\omega\sqrt{LC}\,.$$

Therefore, for lossless transmission lines,

$$\alpha = 0 \quad (R = G = 0) \tag{11.87}$$

$$\beta = \omega\sqrt{LC} \quad [\mathrm{m}^{-1}] \quad (R = G = 0). \tag{11.88}$$

Using these values, the voltage and current expressions given by Equations (11.80) and (11.81) become

$$V = V^{+}e^{-j\beta z} + V^{-}e^{+j\beta z} \tag{11.89}$$

$$I = \frac{V^{+}}{Z_o}e^{-j\beta z} - \frac{V^{-}}{Z_o}e^{+j\beta z}. \tag{11.90}$$

Even though the preceding expressions for V and I are written in the frequency domain, they still represent waves that propagate in time. To see this more clearly, let us for the moment transform these expressions into the time domain. To accomplish the transformation, it is necessary to first represent the complex amplitudes V^{+} and I^{-} in exponential form:

$$V^{+} = |V^{+}| \angle \theta_{+} = |V^{+}|\, e^{j\theta_{+}}$$

$$V^{-} = |V^{-}| \angle \theta_{-} = |V^{-}|\, e^{j\theta_{-}}.$$

The time-domain expressions $V(t, z)$ and $I(t, z)$ are found by multiplying Equations (11.89) and (11.90) by $e^{j\omega t}$ and taking the real parts, yielding

$$V(t, z) = \mathrm{Re}\left[|V^{+}|e^{j(\omega t - \beta z + \theta_{+})} + V^{-}e^{j(\omega t + \beta z + \theta_{-})}\right]$$

$$I(t, z) = \mathrm{Re}\left[\frac{|V^{+}|}{Z_o}e^{j(\omega t - \beta z + \theta_{+})} - \frac{|V^{-}|}{Z_o}e^{j(\omega t + \beta z + \theta_{-})}\right].$$

Remembering that Z_o is real for lossless transmission lines, we obtain the following time-domain expressions:

$$V(t, z) = |V^+|\cos(\omega t - \beta z + \theta_+) + |V^-|\cos(\omega t + \beta z + \theta_-) \tag{11.91}$$

$$I(t, z) = \frac{|V^+|}{Z_o}\cos(\omega t - \beta z + \theta_+) - \frac{|V^-|}{Z_o}\cos(\omega t + \beta z + \theta_-). \tag{11.92}$$

The preceding voltage and current expressions vary sinusoidally with both time t and position z. To see that they are indeed propagating waves, let us first consider the terms that contain $\cos(\omega t - \beta z + \theta_+)$. In order to "ride" on a constant phase point on these waves, the argument of the cosine term must remain constant with time. This occurs when

$$\frac{d}{dt}(\omega t - \beta z + \theta_+) = 0.$$

Since ω, β, and θ_+ are constants, this expression becomes

$$\frac{dz}{dt} = \frac{\omega}{\beta}.$$

Noting that dz/dt is a velocity, we conclude that these waves propagate towards increasing values of z at a rate of

$$u_p = \frac{\omega}{\beta} \quad \text{(All transmission lines)}, \tag{11.93}$$

where u_p is called the ***phase velocity*** of the wave, since it is the rate at which the constant phase fronts move. Equation (11.93) is a general expression that is applicable to both lossless and lossy transmission lines. For lossless transmission lines, however, $\beta = \omega\sqrt{LC}$, which yields

$$u_p = \frac{1}{\sqrt{LC}} \quad \text{(Lossless transmission lines)}. \tag{11.94}$$

It is also possible to express u_p in terms of the permittivity and permeability of the dielectric when the dielectric of the transmission line is homogeneous. In Section 13-2 we will show that the phase velocity of *any* TEM wave that passes through a lossless, homogeneous medium equals $1/\sqrt{\mu\epsilon}$, where μ and ϵ are the permeability and permittivity of the medium, respectively. Hence, comparing $1/\sqrt{\mu\epsilon}$ with $1/\sqrt{LC}$ in Equation (11.94), we conclude that the product LC on a lossless, homogeneous-dielectric transmission line (i.e., a TEM line) equals the product $\mu\epsilon$:

$$LC = \mu\epsilon \quad \text{(Lossless, TEM transmission lines)}. \tag{11.95}$$

Substituting Equation (11.95) into Equation (11.94) and assuming that the dielectric is nonmagnetic (as is nearly always the case), we obtain

$$u_p = \frac{1}{\sqrt{\mu_o \epsilon}} = \frac{c}{\sqrt{\epsilon_r}} \quad \text{(Lossless, TEM transmission lines),} \tag{11.96}$$

where c is the speed of light in a vacuum and ϵ_r is the dielectric constant of the medium. Equation (11.96) can also be used with Quasi-TEM transmission lines (such as microstrip lines) if ϵ_r is replaced by an effective dielectric constant ϵ_{eff}. (See Appendix D).

We can also show that the terms which contain $\cos(\omega t + \beta z + \theta_-)$ in Equations (11.91) and (11.92) correspond to waves that propagate towards decreasing values of z. To "ride" on a constant phase point of these waves, an observer must move such that

$$\frac{d}{dt}(\omega t + \beta z + \theta_-) = 0,$$

from which we obtain

$$\frac{dz}{dt} = -u_p = -\frac{\omega}{\beta}.$$

Thus, the phase fronts of these voltage and current waves propagate towards decreasing values of z at a rate u_p.

From the preceding comments, we can conclude that frequency-domain terms that vary with z as $e^{-j\beta z}$ represent forward-propagating waves, whereas terms with $e^{+j\beta z}$ represent backward-propagating waves. Knowing this allows us to determine the direction of propagation of a frequency-domain expression simply by looking at the sign of the exponent. Hence, we rarely find a need to express time-harmonic voltages and currents in the time domain; all the required information is easily determined from the frequency-domain expressions.

An important parameter that is used to describe time-harmonic waves on transmission lines is the ***wavelength***, which is defined as the distance over which the waves repeat themselves and is signified by the symbol λ. To derive an expression for λ, let us assume that a forward-propagating wave is present on a transmission line. Setting $V^- = 0$ in Equation (11.89), we have

$$V(z) = V^+ e^{-j\beta z}.$$

If we require that $V(z + \lambda) = V(z)$, we obtain

$$V^+ e^{-j\beta(z+\lambda)} = V^+ e^{-j\beta z},$$

which simplifies to

$$e^{-j\beta z} = 1.$$

The smallest nonzero value of λ that satisfies this expression is $2\pi/\beta$, so

$$\lambda = \frac{2\pi}{\beta} = \frac{u_p}{f}, \tag{11.97}$$

where we have used $\beta = 2\pi f/u_p$. As might be expected, the wavelength of backward-propagating waves is also given by Equation (11.97.)

Example 11-7

Common "TV twin lead" transmission lines consist of two stranded wires encased in a thin dielectric ribbon. Stranded wires are used to reduce the high-frequency resistance, and the dielectric is needed for mechanical support. Consider a case where 20 A.W.G. wires are used, the wire spacing is 7 [mm], and the dielectric gives rise to an effective dielectric constant of $\epsilon_{\text{eff}} = 1.29$. Calculate Z_o, u_p, β, and λ at 100 [MHz].

Solution:

The diameter of 20 A.W.G. wire is $d = 0.812$ [mm]. Using Equation (D.14), we find that

$$Z_o = \frac{1}{\pi}\sqrt{\frac{\mu_o}{\epsilon'}}\cosh^{-1}\left(\frac{D}{d}\right) = \frac{1}{\pi}\sqrt{\frac{4\pi \times 10^{-7}}{1.29 \times 8.854 \times 10^{-12}}}\cosh^{-1}\left(\frac{7.0}{0.812}\right) = 300\ [\Omega].$$

Also, from Equation (D.7), we have

$$u_p = \frac{c}{\sqrt{\epsilon_{\text{eff}}}} = \frac{3 \times 10^8}{\sqrt{1.29}} = 2.64 \times 10^8\ [\text{m/s}]$$

At $f = 100$ [MHz], Equations (11.88) and (11.97) give us

$$\beta = \frac{\omega}{u_p} = \frac{2\pi \times 100 \times 10^6}{2.64 \times 10^8} = 2.38\ [\text{m}^{-1}]$$

and

$$\lambda = \frac{2\pi}{\beta} = 2.64\ [\text{m}].$$

Example 11-8

Find the conductor width w necessary to obtain a 50 [Ω] characteristic impedance from a microstrip transmission line if the height h is 2 [mm] and the substrate material is RT/Duroid[5] 5880, which has a dielectric constant of 2.26. Also, find the wavelength at $f = 2$ [GHz].

Solution:

The width-to-height ratio is given by either Equations (D.22–23) (valid for $w/h \leqslant 2$) or Equations (D.24–25) (valid for $w/h \geqslant 2$). Let us try the first set and see if consistent values are obtained. From Equation (D.23), we find that

[5] RT/Duroid is a registered trademark of the Rogers Corporation.

$$A = \frac{\pi \times 50}{377}\sqrt{2(2.26 + 1)} + \frac{2.26 - 1}{2.26 + 1}\left(0.23 + \frac{0.11}{2.26}\right) = 1.172.$$

Substituting this value into Equation (D.22) yields

$$\frac{w}{h} \approx 4\left[\frac{1}{2}\exp(1.172) - \exp(-1.172)\right]^{-1} = 3.066.$$

Since 3.066 is >2, we have used the wrong formula, which means that we must instead use Equations (D.24–25). When we do, we get

$$B = \frac{\pi}{2\sqrt{2.26}}\,\frac{377\ [\Omega]}{50\ [\Omega]} = 7.878$$

$$\frac{w}{h} \approx \frac{2.26 - 1}{\pi \times 2.26}\left(\ln(B - 1) + 0.39 - 0.61/2.26\right) + \frac{2}{\pi}\left(B - 1 - \ln(2B - 1)\right) = 3.029.$$

Since $w/h = 3.029 > 2$, this is a valid result. Hence, the conductor strip width necessary to achieve $Z_0 = 50\ [\Omega]$ is

$$w = 3.029 \times h = 6.058\ [\text{mm}].$$

We can now use Equation (D.20) to find the effective dielectric constant:

$$\epsilon_{\text{eff}} = \frac{1}{2}(2.26 + 1) + \frac{1}{2}\frac{(2.26 - 1)}{\sqrt{1 + 12 \times 3.029}} = 1.733.$$

Finally, substituting Equation (11.96) into Equation (11.97) and replacing ϵ_r with ϵ_{eff}, we obtain the wavelength at 300 [MHz],

$$\lambda = \frac{c}{f\sqrt{\epsilon_{\text{eff}}}} = \frac{3 \times 10^8}{\sqrt{1.733} \times 2 \times 10^9} = 113.93\ [\text{mm}].$$

11-4-2 TIME-HARMONIC WAVES ON LOSSY TRANSMISSION LINES

Earlier in this chapter it was stated that waveforms tend to distort (or disperse) as they propagate down lossy transmission lines (except for the case when $RC = GL$). Because of this, lossy transmission lines are difficult to describe using standard time-domain analysis. There is no such problem in the frequency domain, however, since the waves remain sinusoidal no matter how much loss is present. To show this, let us return to Equation (11.80), which is the general phasor expression for the voltage waves that can exist on a transmission line:

$$V = V^+e^{-\gamma z} + V^-e^{+\gamma z}. \tag{11.98}$$

Unlike the lossless case, where the propagation constant γ is imaginary, the presence of loss (either $R \neq 0$ or $G \neq 0$) makes γ complex, so we can express γ in the form

$$\gamma = \alpha + j\beta, \tag{11.99}$$

where

$$\alpha = \text{Re}[\sqrt{(R + j\omega L)(G + j\omega C)}]\quad [\text{Np}\cdot\text{m}^{-1}] \tag{11.100}$$

$$\beta = \text{Im}\,[\sqrt{(R + j\omega L)(G + j\omega C)}] \quad [\text{m}^{-1}]. \tag{11.101}$$

Using these expressions, we can write Equation (11.98) in the form

$$V = V^{+}e^{-\alpha z}e^{-j\beta z} + V^{-}e^{+\alpha z}e^{+j\beta z} \tag{11.102}$$

Since the exponential terms $e^{-\alpha z}$ and $e^{+\alpha z}$ in Equation (11.102) are real valued, the time-domain representation of V is given by

$$V(t, z) = |V^{+}|e^{-\alpha z}\cos(\omega t - \beta z + \theta_{+}) + |V^{-}|e^{+\alpha z}\cos(\omega t + \beta z + \theta_{-}), \tag{11.103}$$

where θ_{+} and θ_{-} are the phases of the phasors V^{+} and V^{-}, respectively. Comparing this expression with Equation (11.91), we see that they are nearly the same, since they both have identical sinusoidal terms. This means that the first and second terms in the preceding expression represent forward and backward-propagating waves, respectively, with a phase velocity and wavelength respectively given by

$$u_p = \frac{\omega}{\beta} \tag{11.104}$$

and

$$\lambda = \frac{2\pi}{\beta}, \tag{11.105}$$

where, for lossy transmission lines, β is given by Equation (11.101). Unlike the lossless case, however, the forward and backward-propagating waves on lossy transmission lines contain the exponential terms $e^{-\alpha z}$ and $e^{\alpha z}$, respectively, which cause the wave amplitudes to decay along their propagation directions. The rate of decay is determined by the attenuation constant α.

There are two ways to specify α. The first is in units of nepers per meter [Np/m], where the neper is a dimensionless unit. This way of specifying α follows directly from Equation (11.100). The other way is to specify α in terms the decibels of loss per meter. To do this, we note that the amplitude of a forward-propagating voltage decays at a rate proportional to $e^{-\alpha z}$, so the decibel loss per meter is

$$\text{dB loss per meter} = -20\log_{10}\frac{|V(z = 0)|}{|V(z = 1)|} = -20\log_{10}[e^{-\alpha}].$$

Simplifying this expression and using $\log_{1o}[e] = 0.434$, we obtain the following conversion relations:

$$\alpha\,[\text{dB/m}] = 8.686 \times \alpha\,[\text{Np/m}] \tag{11.106a}$$

$$\alpha\,[\text{Np/m}] = 0.1151 \times \alpha\,[\text{dB/m}]. \tag{11.106b}$$

Although α can be specified in either [Np/m] or [dB/m], the reader should beware; ***α must be specified in [Np/m] when it is used in formulas that contain the terms $e^{-\alpha z}$ and $e^{\alpha z}$.***

Example 11-9

A forward-propagating voltage wave has an amplitude of 7 [V] at $z = 0$. Calculate the amplitude at $z = 10$ [m] if it is known that $\alpha = 0.5$ [dB/m].

Solution:

Using Equation (11.106b), we find that

$$\alpha = 0.1151 \times 0.5 = 0.0576 \text{ [Np/m]}.$$

Thus, the voltage amplitude at $z = 10$ [m] is

$$V(z = 10) = V(z = 0)e^{-10\alpha} = 7e^{-.576} = 3.937 \text{ [V]}.$$

An alternative method of finding this result is to note that if the transmission line exhibits 0.5 dB/m, then there will be 0.5 × 10 = 5 [dB] loss in 10 meters. Hence,

$$20 \log_{10} \frac{|V(z = 10)|}{|V(z = 0)|} = -5 \text{ [dB]}.$$

Solving for $V(z = 10)$, we again obtain

$$V(z = 10) = 7 \times 10^{-.25} = 3.936 \text{ [V]}.$$

Another important effect of loss on a transmission line is that the characteristic impedance Z_o becomes complex. This can be seen from Equation (11.84):

$$Z_o = \sqrt{\frac{R + j\omega L}{G + j\omega C}} = \frac{\gamma}{G + j\omega C} = \frac{R + j\omega L}{\gamma} \quad [\Omega]. \tag{11.107}$$

Whenever $R \neq 0$ or $G \neq 0$, Z_o is complex and can be expressed in either rectangular or polar form:

$$Z_o = R_o + jX_o = |Z_o| \angle \theta_z. \tag{11.108}$$

When Z_o is complex, the voltage and current waves are out of phase. To see why, let us return to the frequency-domain expression for I (Equation (11.90)):

$$\begin{aligned} I &= \frac{V^+}{Z_o} e^{-\gamma z} - \frac{V^-}{Z_o} e^{+\gamma z} = \frac{V^+}{Z_o} e^{-\alpha z} e^{-j\beta z} - \frac{V^-}{Z_o} e^{+\alpha z} e^{+j\beta z} \\ &= \frac{V^+}{|Z_o|} e^{-\theta_z} e^{-\alpha z} e^{-j\beta z} - \frac{V^-}{|Z_o|} e^{-\theta_z} e^{+\alpha z} e^{+j\beta z}. \end{aligned}$$

Transforming the preceding expression into the time domain, we obtain

$$I(t, z) = \frac{|V^+|}{|Z_o|} e^{-\alpha z} \cos(\omega t - \beta z + \theta_+ - \theta_z) - \frac{|V^-|}{|Z_o|} e^{\alpha z} \cos(\omega t + \beta z + \theta_- - \theta_z). \tag{11.109}$$

Comparing this expression for I with the voltage expression (Equation (11.102)), we see that the forward and backward-propagating voltage and current wave amplitudes are proportional by the factor $|Z_o|$, but the currents lag the voltages by the phase angle θ_z of the characteristic impedance.

Approximate expressions for low-loss transmission lines. Because of the complex square-root operations necessary to calculate both γ and Z_o on lossy lines (see Equations (11.76) and (11.84)), it is often desirable to use approximations of these formulas for low-loss conditions. For most practical transmission lines, $R \ll \omega L$ and $G \ll \omega C$ when operated above a megahertz or so. For this case, we can use the binomial theorem to derive simple expressions for Z_o, R_o, and X_o as follows:

$$Z_o = R_o + jX_o = \sqrt{\frac{R + j\omega L}{G + j\omega C}} = \sqrt{\frac{L}{C}}\left[1 + \frac{R}{j\omega L}\right]^{1/2}\left[1 + \frac{G}{j\omega C}\right]^{-1/2}$$

$$= \sqrt{\frac{L}{C}}\left[1 + \frac{R}{j2\omega L} + \frac{R^2}{8\omega^2 L^2} + \dots\right]\left[1 - \frac{G}{j2\omega C} - \frac{3G^2}{8\omega^2 C^2} + \dots\right]$$

$$\approx \sqrt{\frac{L}{C}}\left[1 + \frac{1}{8\omega^2}\left(\frac{R}{L} - \frac{G}{C}\right)\left(\frac{R}{L} + \frac{3G}{C}\right)\right] - \frac{j}{2\omega}\sqrt{\frac{L}{C}}\left[\frac{R}{L} - \frac{G}{C}\right].$$

Thus,

$$R_o \approx \sqrt{\frac{L}{C}}\left[1 + \frac{1}{8\omega^2}\left(\frac{R}{L} - \frac{G}{C}\right)\left(\frac{R}{L} + \frac{3G}{C}\right)\right]. \tag{11.110}$$

$$X_o \approx -\frac{1}{2\omega}\sqrt{\frac{L}{C}}\left[\frac{R}{L} - \frac{G}{C}\right]. \tag{11.111}$$

If we retain only the first-order terms, these relations can be further simplified to

$$R_o \approx \sqrt{\frac{L}{C}} \tag{11.112}$$

$$X_o \approx 0 \tag{11.113}$$

In a similar manner, we can find approximate expressions for β and α by applying the binomial theorem to Equation (11.76),

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

$$= j\omega\sqrt{LC}\left[1 + \frac{R}{j\omega L}\right]^{1/2}\left[1 + \frac{G}{j\omega C}\right]^{1/2}$$

$$= j\omega\sqrt{LC}\left[1 + \frac{R}{j2\omega L} + \frac{R^2}{8\omega^2 L^2} + \dots\right]\left[1 + \frac{G}{j2\omega C} + \frac{G^2}{8\omega^2 C^2} + \dots\right]$$

$$\approx \frac{1}{2}\left[R\sqrt{\frac{C}{L}} + G\sqrt{\frac{L}{C}}\right]\left[1 + \frac{1}{8\omega^2}\left(\frac{R}{L} - \frac{G}{C}\right)^2\right]$$

$$+ j\omega\sqrt{LC}\left[1 + \frac{1}{8\omega^2}\left(\frac{R}{L} - \frac{G}{C}\right)^2\right].$$

Since $\gamma = \alpha + j\beta$, we have

$$\beta \approx \omega\sqrt{LC}\left[1 + \frac{1}{8\omega^2}\left(\frac{R}{L} - \frac{G}{C}\right)^2\right]. \tag{11.114}$$

$$\alpha \approx \frac{1}{2}\left[R\sqrt{\frac{C}{L}} + G\sqrt{\frac{L}{C}}\right]\left[1 + \frac{1}{8\omega^2}\left(\frac{R}{L} - \frac{G}{C}\right)^2\right]. \tag{11.115}$$

When the loss is small, these expressions can be further simplified by retaining only the first-order terms, in which case we obtain

$$\beta \approx \omega\sqrt{LC} \tag{11.116}$$

and

$$\alpha \approx \alpha_c + \alpha_d, \tag{11.117}$$

where α_c and α_d are the ***conductor*** and ***dielectric attenuation constants***, respectively, given by

$$\alpha_c = \frac{R}{2}\sqrt{\frac{C}{L}} = \frac{R}{2Z_o} \tag{11.118}$$

$$\alpha_d = \frac{G}{2}\sqrt{\frac{L}{C}} = \frac{GZ_o}{2} = \frac{G}{2C}\sqrt{LC}. \tag{11.119}$$

For most transmission lines, the conductor losses are much greater than the dielectric losses, so α_d can usually be neglected.

The characteristic impedance Z_o of a transmission line is always a function of both its material properties and its cross sectional dimensions. However, when losses are low, several other parameters are controlled solely by the properties of the dielectrics. For instance, when losses are low, the phase velocity u_p can be expressed as:

$$u_p \approx \frac{1}{\sqrt{\mu_o \epsilon'}} \qquad \text{(Low loss TEM transmission lines)}, \tag{11.120}$$

where ϵ' is real part of the dielectric permittivity and the dielectric is assumed to be non magnetic (i.e., $\mu = \mu_o$). This formula is derived in the next chapter for plane waves (see Equation 12.82) and is applicable for transmission lines since plane waves are also TEM waves. Remembering that $\beta = \omega/u_p$, we also have

$$\beta = \frac{2\pi}{\lambda} \approx \omega\sqrt{\mu_o \epsilon'}, \qquad \text{(Low loss TEM transmission lines)}, \tag{11.121}$$

The preceding formulas can also be applied to non TEM transmission lines when ϵ' is replaced with the effective dielectric constant ϵ_{eff} of the line (see Appendix D). The dielectric loss constant α_d is also a function of the dielectric properties alone. When losses are small, we have:

$$\alpha_d = \frac{\omega\epsilon''}{2\epsilon'}\sqrt{\mu_o \epsilon'} \qquad \text{(Low loss TEM transmission lines)}, \tag{11.122}$$

where $-\epsilon''$ is the imaginary part of the dielectric's complex permittivity.[1] This expression is also derived in the next chapter (see Equations 12.79) for plane waves and is applicable for transmission lines with uniform dielectrics (i.e., TEM lines).

[6] In Chapter 12, it is shown that the loss of a dielectric can be specified in terms of either σ or ϵ''.

Example 11-10

Calculate the attenuation constant for a polyethylene-filled RG-58/U coaxial cable at $f = 100$ [MHz]. Assume that the complex permittivity of polyethylene is $\epsilon = \epsilon_o\,(2.26 - j\,0.0002)$ [F/m] and the conductivity of copper is 5.8×10^7 [S/m].

Solution:

From Equation (D.13), we find that the attenuation constant for a coaxial cable is

$$\alpha = \frac{\omega\epsilon''}{2\epsilon'}\sqrt{\mu\epsilon'} + \frac{1}{4\pi\sigma\delta Z_o}\left(\frac{1}{a} + \frac{1}{b}\right) \quad \text{[Np/m]},$$

where a and b are the radii of the inner and outer conductors, respectively; $\epsilon = \epsilon' - j\epsilon''$ is the complex permittivity of the dielectric (polyethylene); σ is the conductivity of the conductors (for copper, $\sigma = 5.8 \times 10$ [S/m]); Z_o is the characteristic impedance when losses are neglected; and δ is the skin depth, which is given by Equation (D.11) as

$$\delta = \frac{1}{\sqrt{\pi f \mu \sigma}}.$$

The values of a, b, and Z_o for RG-58/U cable are given in Example 11-1; $a = 0.406$ [mm], $b = 1.553$ [mm], and $Z_o = 53.47$ [Ω]. Using these values, we obtain

$$\delta = \frac{1}{\sqrt{\pi \times 100 \times 10^6 \times 4\pi \times 10^{-7} \times 5.8 \times 10^7}} = 6.61 \times 10^{-6} \quad \text{[m]}$$

and

$$\begin{aligned}\alpha &= \frac{2\pi \times 100 \times 10^6 \times 0.0002\epsilon_o}{2 \times 2.26\epsilon_o}\sqrt{2.26\mu_o\epsilon_o} \\ &\quad + \frac{1}{4\pi \times 5.8 \times 10^7 \times 6.61 \times 10^{-6} \times 53.47}\left(\frac{10^3}{0.406} + \frac{10^3}{1.553}\right) \\ &= 1.39 \times 10^{-4} + 1.205 \times 10^{-2} = 1.22 \times 10^{-2} \quad \text{[Np/m]}.\end{aligned}$$

Comparing the magnitudes of the two components of α, we see that $1.205 \times 10^{-2} \gg 1.39 \times 10^{-4}$, which means that the conductor losses dominate the dielectric losses by more than an order of magnitude.

11-4-3 GROUP VELOCITY AND DISPERSION

We have already shown that the constant-phase fronts of the voltage and current waves travel at the phase velocity

$$u_p = \frac{\omega}{\beta}.$$

On lossless lines, $\beta = \omega\sqrt{LC}$, which means that u_p is the same for all frequencies. This is not the case, however, when losses are present, since β no longer varies linearly with ω. In this section, we will show that waveform distortion occurs when the phase velocity varies with frequency.

Let us consider the propagation of the simple, narrow-band signal

$$V(t) = V_o[1 + m\cos\omega_s t]\cos\omega_c t.$$

This waveform is called an ***amplitude-modulated signal***, where ω_s and ω_c are the ***signal*** and ***carrier frequencies***, respectively, and m is the ***modulation index*** (usually with a value between zero and unity). Figure 11-28 shows a plot of $V(t)$ for the case $w_s \ll w_c$. The carrier amplitude-vs.-time trace is called the ***envelope***, which contains the information carried by the signal. To see how this waveform propagates down a transmission line, let us first use the cosine product identity to write $V(t)$ in the form

$$V(t) = V_o\left[\cos\omega_c t + \frac{m}{2}\cos\omega_U t + \frac{m}{2}\cos\omega_L t\right]. \tag{11.123}$$

Here, ω_U and ω_L are called the ***upper*** and ***lower sideband*** frequencies, respectively, and are given by

$$\omega_U = \omega_c + \omega_s$$

and

$$\omega_L = \omega_c - \omega_s.$$

From Equation (11.123), we see that $V(t)$ is the sum of three distinct sinusoids with frequencies ω_c, ω_U, and ω_L.

By introducing the appropriate propagation delay for each frequency component, we can write the voltage at any point on the line in the form

$$V(t, z) = V_o\left[\cos(\omega_c t - \beta_c z) + \frac{m}{2}\cos(\omega_U t - \beta_U z) + \frac{m}{2}\cos(\omega_L t - \beta_L z)\right], \tag{11.124}$$

where β_c, β_U, and β_L are the phase constants at the frequencies ω_c, ω_U, and ω_L, respectively. Since $\omega_s \ll \omega_c$, we can use Taylor's theorem to calculate approximate values of β_U and β_L. Retaining only the first two terms, we find that

$$\beta_U \approx \beta_c + \omega_s\frac{\partial\beta}{\partial\omega} \tag{11.125}$$

$$\beta_L \approx \beta_c - \omega_s\frac{\partial\beta}{\partial\omega}. \tag{11.126}$$

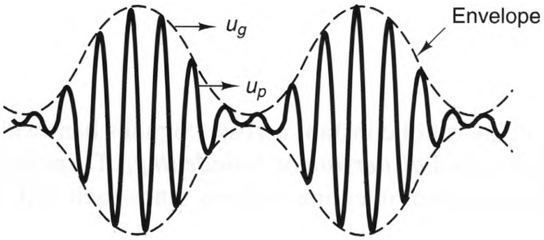


Figure 11-28 An amplitude-modulated signal.

Substituting Equations (11.125) and (11.126) into Equation (11.124) and rearranging, we obtain

$$V(t,z) = V_o \cos(\omega_c t - \beta_c z) + \frac{mV_o}{2}\cos\left[(\omega_c t - \beta_c z) + \left(\omega_s t - \omega_s \frac{\partial \beta}{\partial \omega} z\right)\right]$$

$$+ \frac{mV_o}{2}\cos\left[(\omega_c t - \beta_c z) - \left(\omega_s t - \omega_s \frac{\partial \beta}{\partial \omega} z\right)\right].$$

Using the cosine sum formula, we can write this equation as

$$V(t,z) = V_o\left[1 + m\cos\left(\omega_s t - \omega_s \frac{\partial \beta}{\partial \omega} z\right)\right]\cos(\omega_c t - \beta_c z). \tag{11.127}$$

Comparing Equation (11.127) with the initial waveform, we see that $V(t, z)$ is still an amplitude-modulated signal for all values of z. However, the phases of the carrier and the envelope propagate at different rates. The phase constant of the carrier is β_c, so its phase fronts travel at a velocity

$$u_{\text{carrier}} = \frac{\omega_c}{\beta_c} = u_p\,,$$

which is the same as the phase velocity of a single time-harmonic waveform. On the other hand, the envelope propagates at a velocity

$$u_{\text{envelope}} = \frac{\omega_s}{\left[\omega_s \dfrac{\partial \beta}{\partial \omega}\right]} = \left(\frac{\partial \beta}{\partial \omega}\right)^{-1} = \frac{\partial \omega}{\partial \beta}.$$

Since the envelope is composed of a narrow band of frequencies, this velocity is called the ***group velocity*** and is denoted by the symbol u_g, where

$$u_g \equiv \left(\frac{\partial \beta}{\partial \omega}\right)^{-1} = \frac{\partial \omega}{\partial \beta}. \tag{11.128}$$

From the foregoing definition, we see that $u_g = u_p$ only when β is a linear function of ω, such as when a transmission line has zero loss and the inductance L and capacitance C are independent of frequency. Another important parameter is the ***group delay*** τ_g, which is the inverse of the group velocity:

$$\tau_g \equiv \frac{1}{u_g} \qquad \text{[s/m]}. \tag{11.129}$$

When the group delay is not constant across a signal's bandwidth, distortion will occur. This distortion is called ***dispersion*** or ***broadening***. Dispersion can be particularly troublesome in digital communication systems, since each "bit" is assigned a spe-

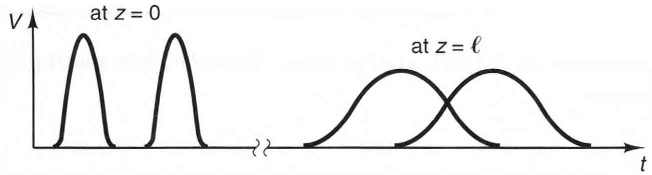


Figure 11-29 Adjacent pulses on a dispersive transmission line at two locations.

cific time slot. As these pulses travel down a dispersive line, the "tail" of one pulse spreads into the leading edge of another, causing ambiguities and errors. This is depicted in Figure 11-29, which shows two pulses that are sent down a dispersive transmission line. At $z = 0$, the pulses show no overlap, but at $z = \ell$, both pulses have broadened, so that they now overlap.

The amount of pulse broadening $\Delta\tau$ incurred by a pulse while it propagates depends upon how much the group delay varies within the pulse bandwidth. We can estimate $\Delta\tau$ for a pulse by calculating the difference between the group delays of the highest and lowest frequency components. We obtain

$$\Delta\tau = \tau_{\max} - \tau_{\min}, \tag{11.130}$$

where $\tau_{\max}$ and $\tau_{\min}$ are, respectively the maximum and minimum group delays within the pulse bandwidth.

It is difficult to derive simple formulas for u_g and τ directly from Equations (11.128) and (11.100), since these formulas involve both square roots and derivatives of complex-valued functions. Worse yet, the values of R, L, G, and C often vary with frequency (particularly R, due to the skin effect, which is discussed in Chapter 12), further complicating the calculations. Because of this, it is usually easier to *measure* the group velocity than it is to calculate it. Figure 11-30 shows the relative group delay τ on a sample of RG-58U coaxial cable. As can be seen, τ varies most rapidly at low frequencies. This is because the wire resistance R is greater than the inductive reactance ωL at low frequencies.

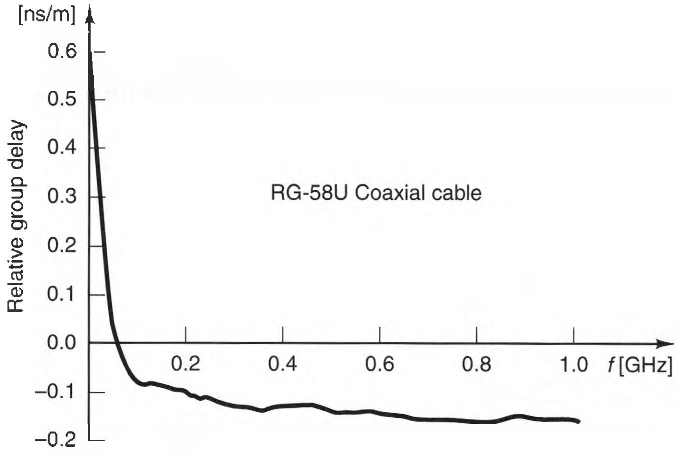


Figure 11-30 Relative group delay vs. frequency on a typical section of RG-58U coaxial cable.

Example 11-11

A 1 [ns] pulse is transmitted down an RG-58U coaxial cable. Calculate how much pulse spreading occurs in a 1.0 [cm] length.

Solution:

From Fourier analysis, the bandwidth of a rectangular pulse is approximately $\Delta f \approx 1/T$, where T is the pulse width. When $T = 1$ [ns], we find that $\Delta f \approx 1$ [GHz]. Thus, the pulse contains frequency components from dc through 1 [GHz].

From Figure 11-30, the maximum group velocity within this bandwidth occurs at dc and is

$$\tau_{\max} = 0.52 \text{ [ns/m]}.$$

The minimum value occurs at 1 [GHz] and is

$$\tau_{\min} = -0.15 \text{ [ns/m]}.$$

Substituting these values into Equation (11.130), we obtain

$$\Delta\tau = 0.52 - (-0.15) = 0.67 \text{ [ns/m]}.$$

Thus, along a 1.0 [cm] length, the pulse spreading is

$$\Delta\tau = 0.67 \text{ [ns/m]} \times .01 \text{ [m]} = 6.7 \text{ [ps]}.$$

11-4-4 REFLECTIONS OF TIME-HARMONIC WAVES

Figure 11-31 shows a section of transmission line with characteristic impedance Z_o, terminated with a load impedance Z_L.

Regardless of what sources are attached to the left-end terminals of the transmission line, the expressions for the total voltage and current on the line at a frequency ω are of the form

$$V = V^+e^{-\gamma z} + V^-e^{+\gamma z}$$

$$I = \frac{1}{Z_o}(V^+e^{-\gamma z} - V^-e^{\gamma z}),$$

where $\gamma = \alpha + j\beta$ and z is measured towards the right. At the load ($z = 0$), $V/I = Z_L$. Substituting this into the preceding expressions for V and I and solving for V^-, we obtain

$$V^- = \frac{Z_L - Z_o}{Z_L + Z_o} V^+ .$$

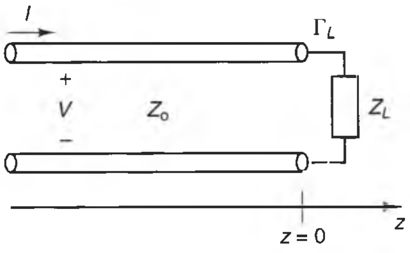


Figure 11-31 A transmission line terminated by an arbitrary impedance.

This expression shows that the backward-propagating voltage phasor V^- is the product of the forward-propagating phasor V^+ and a factor that depends upon the mismatch between Z_o and Z_L. This factor is the ***load reflection coefficient*** Γ_L and is given by

$$\Gamma_L \equiv \left.\frac{V^-}{V^+}\right|_{z=0} = \frac{Z_L - Z_o}{Z_L + Z_o}. \tag{11.131}$$

Comparing Equation (11.131) with the reflection coefficient derived earlier for resistive loads attached to lossless lines (see Equation (11.57)), we see that they are the same, except that Equation (11.131) is also valid for lossy lines and reactive loads. Further, this time-harmonic reflection coefficient can be complex when either Z_o or Z_L is complex, which means that the reflected wave can differ from the incident wave in both amplitude and phase. Finally, we can solve Equation (11.131) for Z_L in terms of Γ_L, yielding

$$Z_L = Z_o\frac{1+\Gamma_L}{1-\Gamma_L}. \tag{11.132}$$

This formula can be used to determine the load impedance when the reflection coefficient is known.

Example 11-12

Calculate the load reflection coefficient on a transmission line if $Z_o = 50\,[\Omega]$ and $Z_L = 100 - j30\,[\Omega]$.

Solution:

Using Equation (11.131), we have

$$\Gamma_L = \frac{Z_L - Z_o}{Z_L + Z_o} = \frac{100 - j\,30 - 50}{100 - j\,30 + 50} = 0.359 - j\,0.128$$
$$= 0.381 \angle -19.65^\circ.$$

11-4-5 INPUT IMPEDANCE AND THE IMPEDANCE TRANSFORMATION

One of the most important effects associated with transmission lines is the way in which they can transform the impedance of a load into a different value when it is viewed through a length of the line. Figure 11-32 depicts such a situation. Here, a

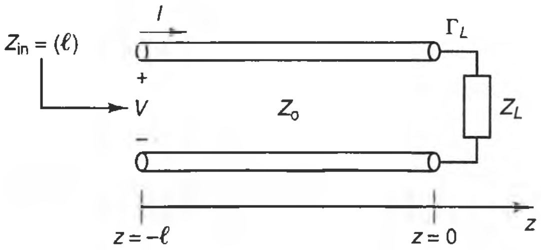


Figure 11-32 Geometry for determining the input impedance a distance ℓ from a load impedance.

section of transmission line of length ℓ and characteristic impedance Z_o is terminated with a load with impedance Z_L. If z is measured from the load, with increasing values towards the right, the general expressions for the voltage and current on the line are

$$V = V^+ e^{-\gamma z} + \Gamma_L V^+ e^{+\gamma z}$$

$$I = \frac{1}{Z_o}(V^+ e^{-\gamma z} - \Gamma_L V^+ e^{+\gamma z}),$$

where we note that the magnitudes of the backward-propagating waves are proportional to the load-reflection coefficient Γ_L. The impedance Z_{in}, looking into the terminals a distance ℓ to the left of the load, is the ratio of the voltage to the current at the input terminals ($z = -\ell$); thus,

$$Z_{\text{in}}(\ell) \equiv \left.\frac{V}{I}\right|_{z=-\ell}.$$

Substituting V and I into this expression, we obtain

$$Z_{\text{in}}(\ell) = Z_o \frac{V^+ e^{\gamma\ell} + \Gamma_L V^+ e^{-\gamma\ell}}{V^+ e^{\gamma\ell} - \Gamma_L V^+ e^{-\gamma\ell}} = Z_o \frac{e^{\gamma\ell} + \Gamma_L e^{-\gamma\ell}}{e^{\gamma\ell} - \Gamma_L e^{-\gamma\ell}}.$$

Substituting $\Gamma_L = (Z_L - Z_o)/(Z_L + Z_o)$ into this expression and multiplying both top and bottom by $(Z_L + Z_o)$ yields

$$Z_{\text{in}}(\ell) = Z_o \frac{Z_L(e^{\gamma\ell} + e^{-\gamma\ell}) + Z_o(e^{\gamma\ell} - e^{-\gamma\ell})}{(Z_L + Z_o)e^{\gamma\ell} - (Z_L - Z_o)e^{-\gamma\ell}}.$$

Finally, using the hyperbolic tangent function,

$$\tanh(\gamma\ell) = \frac{(e^{\gamma\ell} - e^{-\gamma\ell})}{(e^{\gamma\ell} + e^{-\gamma\ell})},$$

$Z_{\text{in}}(\ell)$ we can write as

$$Z_{\text{in}}(\ell) = Z_o \frac{Z_L + Z_o \tanh(\gamma\ell)}{Z_o + Z_L \tanh(\gamma\ell)}. \qquad (11.133)$$

If the line is lossless, $\gamma = j\beta$, and

$$\tanh(\gamma\ell) = \tanh(j\beta\ell) = j\tan(\beta\ell).$$

This means that for lossless transmission lines,

$$Z_{\text{in}}(\ell) = Z_o \frac{Z_L + jZ_o \tan(\beta\ell)}{Z_o + jZ_L \tan(\beta\ell)} \quad \text{(Lossless transmission lines).} \qquad (11.134)$$

Equations (11.133) and (11.134) are called ***impedance transformation formulas***, because they predict how a transmission line transforms the value of a load impedance when viewed from the input terminals of a the line. Both formulas are important, but Equation (11.134) is used most often, since most transmission lines have losses that can be ignored, at least over short lengths.

Example 11-13

Calculate the input impedance of a 1 [m] length of transmission line that is terminated in a load impedance of $Z_L = 20\ [\Omega]$. Assume that the characteristic impedance of the transmission line is 50 [Ω], its effective dielectric constant is $\epsilon_{\text{eff}} = 1.5$, and the frequency of operation is 50 [MHz].

Solution:

From Equations (11.93) and (11.95), the phase constant β is

$$\beta = \frac{2\pi f}{u_p} = \frac{2\pi f}{c}\sqrt{\epsilon_{\text{eff}}} = \frac{2\pi \times 50 \times 10^6 \times \sqrt{1.5}}{3 \times 10^8} = 1.28.$$

Since $\ell = 1$ [m], we also have

$$\tan \beta\ell = \tan 1.28 = 3.37.$$

(Note that the argument of the tangent function is in *radians.*) Finally, using Equation (11.134), the input impedance is

$$Z_{\text{in}}(\ell) = 50\frac{20 + j\,50 \times 3.37}{50 + j\,20 \times 3.37} = 87.7 + j\,50.2\ [\Omega].$$

The following are some special cases that demonstrate important characteristics of the impedance transformation:

(1) When $Z_L = Z_o$, the numerator and denominator of Equation (11.134) are equal, yielding $Z_{\text{in}} = Z_o$ for all values of ℓ.

(2) When $\ell \to 0$, $\tan(\beta\ell) \to 0$ and $Z_{\text{in}} = Z_L$, regardless of the value of Z_o.

(3) When $\ell = n\,(\lambda/2) = (n\pi/\beta)$ (where n is an integer), $\tan(\beta\ell) = 0$ and $Z_{\text{in}} = Z_L$. Thus, the input impedance equals the load impedance when viewed at multiples of a half-wavelength in back of the load, regardless of the value of Z_o.

(4) When $\beta\ell = 2\pi(\ell/\lambda) \ll 1$, $\tan(\beta\ell) \approx \beta\ell$. Using $\beta = (2\pi/\lambda)$, we find that Equation (11.134) becomes

$$Z_{\text{in}} \approx Z_o\frac{Z_L + j2\pi(\ell/\lambda)Z_o}{Z_o + j2\pi(\ell/\lambda)Z_L} \tag{11.135}$$

(5) When $Z_L = 0$ (i.e., a short circuit), Equation (11.134) yields

$$Z_{\text{in}} = jZ_o\tan(\beta\ell) = jZ_o\tan(2\pi\ell/\lambda). \tag{11.136}$$

Hence, Z_{in} is of the form $Z_{\text{in}} = jX$ for all values of ℓ. Figure 11-33 shows a plot of X vs. ℓ. Notice that when $\ell = \lambda/4$, $Z_{\text{in}} = j\infty$, which is an open circuit.

(6) When $Z_L \to \infty$ (i.e., we have an open circuit), Equation (11.134) yields

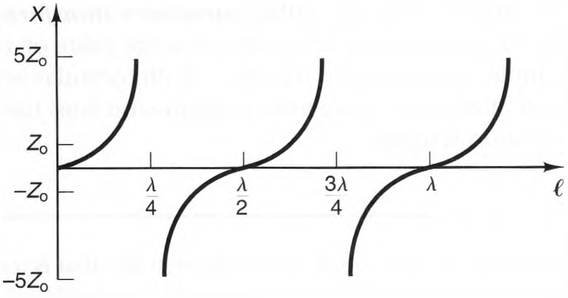


Figure 11-33 Input reactance vs. length for a short-circuited section of transmission line.

$$Z_{\text{in}} = -jZ_o \cot(\beta\ell) = -jZ_o \cot(2\pi\ell/\lambda). \qquad (11.137)$$

Like the short-circuit case, Z_{in} is of the form $Z_{\text{in}} = jX$ for all values of ℓ. Figure 11-34 shows a plot of X vs. ℓ. Notice that when $\ell = \lambda/4$, $Z_{\text{in}} = 0$, which means that an open-circuit load always appears as a short circuit when viewed $\lambda/4$ away.

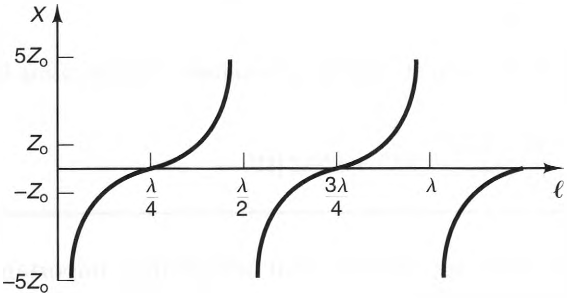


Figure 11-34 Input reactance vs. length for an open-circuited section of transmission line.

11-4-6 TRANSMISSION-LINE EQUIVALENT CIRCUITS

Whenever a transmission line is used to connect components in a circuit, the operation of the circuit is affected. One way to account for this is to model the transmission as a lumped, equivalent circuit. This allows us to use ordinary circuit analysis to model the overall circuit's performance. In this section we will show how any section of transmission line can be modeled as a lumped, two-port network.

Figure 11-35a shows a uniform section of transmission line of length ℓ and characteristic impedance Z_o. Like any linear system of components, we can describe it in terms of its impedance (i.e., Z) parameters. These parameters satisfy the usual two-port network equations,

$$Z_{11}I_1 + Z_{12}I_2 = V_1 \qquad (11.138)$$

$$Z_{21}I_1 + Z_{22}I_2 = V_2, \qquad (11.139)$$

where Z_{11}, Z_{12}, Z_{21}, and Z_{22} are the Z parameters of the network, and the port voltages, V_1 and V_2, and the port currents, I_1 and I_2, are shown in Figure 11-35a. Equations (11.138) and (11.139) can be written in matrix form as

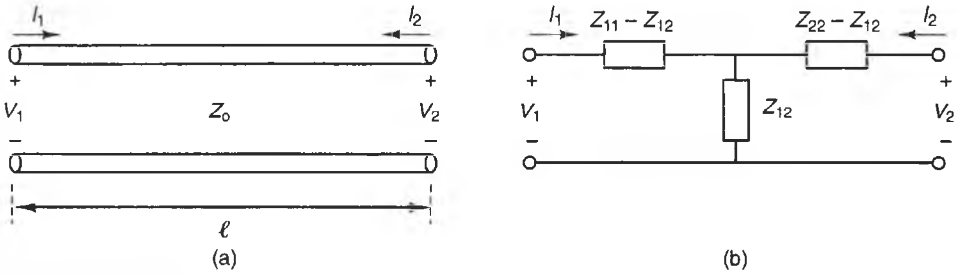


Figure 11-35 a) A uniform section of transmission line. b) An equivalent "T" network in terms of Z parameters.

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \end{bmatrix}. \tag{11.140}$$

If the line section contains no nonlinear or anisotropic materials, it behaves as a ***reciprocal*** network. Reciprocal two-port networks have $Z_{12} = Z_{21}$ and can be represented as a "T" configuration of three passive impedance elements, as shown in Figure 11-35b. Once the values of the Z matrix are known, the equivalent circuit is also known.

The diagonal elements, Z_{11} and Z_{22}, are the easiest Z matrix elements to find. For instance, from Equation (11.138), we find that

$$Z_{11} = \left.\frac{V_1}{I_1}\right|_{I_2=0}. \tag{11.141}$$

Hence, Z_{11} is simply the input impedance looking into port 1 when port 2 is open circuited. Using Equation (11.137), we have

$$Z_{11} = -jZ_o \cot(\beta\ell). \tag{11.142}$$

Similarly, since the transmission line is symmetric,

$$Z_{22} = \left.\frac{V_2}{I_2}\right|_{I_1=0},$$

so

$$Z_{22} = -jZ_o \cot(\beta\ell). \tag{11.143}$$

Knowing the values of Z_{11} and Z_{22}, we can now proceed to find Z_{12}. From Figure 11-35b, we notice that when port 2 is short circuited, the impedance Z_{in}, seen looking into port 1, can be written as

$$Z_{in} = Z_{11} - Z_{12} + Z_{12} \| (Z_{22} - Z_{12}) = Z_{11} - Z_{12} + \frac{Z_{12}(Z_{22} - Z_{12})}{Z_{22}}$$

Solving for Z_{12}, we obtain

$$Z_{12} = \sqrt{Z_{22}(Z_{11} - Z_{in})}. \tag{11.144}$$

However, since we know that the input impedance of a short-circuited transmission line is given by

$$Z_{\text{in}} = jZ_o \tan(\beta\ell), \tag{11.145}$$

we can substitute Equations (11.142), (11.143), and (11.145) into Equation (11.144), obtaining

$$Z_{12} = Z_{21} = -jZ_o \csc(\beta\ell), \tag{11.146}$$

where we note that the negative sign was chosen from the square-root operation to ensure that Z_{12} is capacitive when $\ell \approx 0$. (This is consistent with the unit-cell equivalent circuit shown in Figure 11-4.)

Using Equations (11.142), (11.143), and (11.146), we find that the impedance matrix of a section of lossless, uniform transmission line is

$$[Z] = -jZ_o \begin{bmatrix} \cot(\beta\ell) & \csc(\beta\ell) \\ \csc(\beta\ell) & \cot(\beta\ell) \end{bmatrix}. \tag{11.147}$$

From these values, the lumped loads in the equivalent circuit (Figure 11-35b) are

$$Z_{11} - Z_{12} = -jZ_o \frac{[\cos(\beta\ell) - 1]}{\sin(\beta\ell)} \tag{11.148}$$

$$Z_{22} - Z_{12} = -jZ_o \frac{[\cos(\beta\ell) - 1]}{\sin(\beta\ell)} \tag{11.149}$$

$$Z_{12} = -jZ_o \csc(\beta\ell). \tag{11.150}$$

The following example demonstrates how these impedance values can affect the performance of networks that use transmission lines to connect circuit components.

Example 11-14

Figure 11-36a shows two analog integrated circuits (ICs), connected by a short section of microstrip transmission line of length ℓ and characteristic impedance Z_o. If the input impedance of IC #2 is Z_2 and $\ell = 0.05\lambda$, find a simplified, approximate model for the impedance seen by IC #1 when

(a) $|Z_2| \ll |Z_o|$
(b) $|Z_2| \gg |Z_o|$.

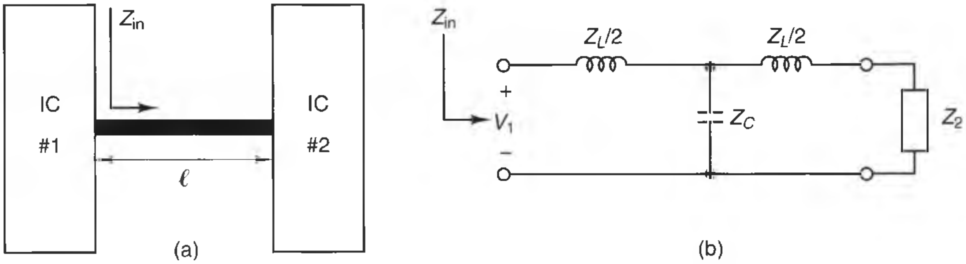


Figure 11-36 (a) Two analog ICs, connected by a length ℓ of transmission line. b) The equivalent circuit at the output port of IC #1.

Solution:

To find the values of the lumped T network of the transmission line, we note that since $\ell = .05\,\lambda$, we have

$$\beta\ell = 2\pi(.05) = 0.314,$$
$$\cos(\beta\ell) = 0.951,$$
$$\sin(\beta\ell) = 0.309.$$

Substituting these values into Equations (11.148), (11.149), and (11.150), we obtain

$$Z_{11} - Z_{12} = Z_{22} - Z_{12} = j(0.158)\,Z_o \equiv Z_L/2$$
$$Z_{12} = -j(3.236)\,Z_o \equiv Z_C\,.$$

Figure 11-36b shows the lumped equivalent circuit of the load "seen" by IC #1.

(a) If $|Z_2| \ll |Z_o|$, we find that $Z_C \gg Z_2 + Z_L/2$. As a result, Z_C can be approximated as an open circuit, yielding the simplified-equivalent circuit shown in Figure 11-37a. Here we see that a short length of high-impedance transmission line effectively adds series inductance to a low-impedance load.

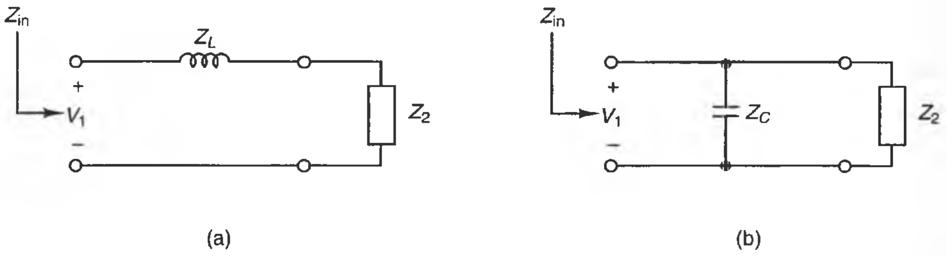


Figure 11-37 Simplified equivalent circuits of the circuit in Figure 11-36b: a) $|Z_2| \ll |Z_o|$. b) $|Z_2| \gg |Z_o|$.

(b) If $|Z_2| \gg |Z_o|$, then $Z_2 \gg Z_L/2$, which means that the impedances of both inductors are negligible compared with Z_2 and Z_C. This yields the circuit shown in Figure 11-37b, which shows that a short length of low-impedance transmission line effectively adds shunt capacitance to a high-impedance load.

In addition to altering the impedance of loads, transmission lines can alter the equivalent circuits of generators. Since lossless transmission lines are linear elements, Thévenin's theorem can be used to analyze networks that contain transmission lines and time-harmonic sources.

Example 11-15

Figure 11-38a shows a sinusoidal generator attached to a $\lambda/8$ section of lossless, 50 [Ω] transmission line. The generator consists of a 100 [mV] voltage source in series with a 75 [Ω] resistor. Find the Thévenin equivalent circuit of this network at the output terminals *ab*.

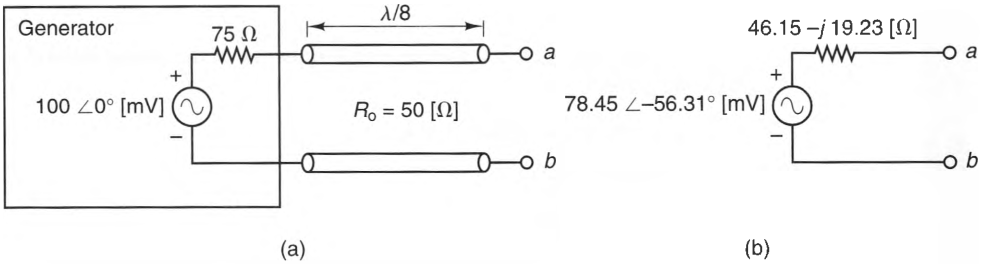


Figure 11-38 a) A sinusoidal generator attached to a section of transmission line. b) The Thévenin equivalent circuit as seen from the transmission line output terminals.

Solution:

To find Z_{th}, we simply set the generator voltage to zero and find the impedance at the output terminals. Using the impedance transformation formula (Equation (11.134)) and $\beta\ell = 2\pi(\ell/\lambda) = \pi/4$, we obtain:

$$Z_{\text{in}} = Z_{\text{th}} = 50\,\frac{75 + j\,50\tan(\pi/4)}{50 + j\,75\tan(\pi/4)} = 46.15 - j\,19.23\ [\Omega].$$

We could find the Thévenin voltage V_{th} by calculating the open-circuit voltage of the network directly, but this method requires a fair amount of work, since both forward- and backward-propagating voltage and current waves will be excited on the line. A simpler method is to attach the same matched (i.e., 50 [Ω]) load to both the original network and the Thévenin equivalent circuit and find the Thévenin voltage source that makes the load voltage in the equivalent circuit equal to that in the original network. This greatly simplifies the transmission line analysis, since only a forward-propagating wave will be present.

When a 50 [Ω] load is attached to the output terminals of the original network, the generator "sees" a 50 [Ω] load. Thus, the voltage V_{in} at the input terminals of the transmission line is determined by the voltage divider relation:

$$V_{\text{in}} = \frac{50}{50 + 75} \times 100 \angle 0^\circ\ [\text{mV}] = 40 \angle 0^\circ\ [\text{mV}].$$

Since only a forward-propagating wave exists on the line for a matched load, the voltage at the output terminals will have the same magnitude, but will be delayed by $\pi/4 = 45^\circ$; hence,

$$V_{\text{out}} = 40 \angle -45^\circ\ [\text{mV}].$$

To find the Thévenin voltage V_{th}, we now apply this same 50 [Ω] load to the Thévenin equivalent circuit of the network, shown in Figure 11-38b. Using the voltage divider, we find that the voltage across a 50 [Ω] load is given by

$$V_{\text{out}} = \frac{50}{50 + 46.15 - j\,19.23} \times V_{\text{th}}.$$

However, this voltage must equal $40 \angle -45^\circ$ [mV], which means that

$$V_{\text{th}} = \frac{50 + 46.15 - j\,19.23}{50} \times 40 \angle -45^\circ\ [\text{mV}] = 78.45 \angle -56.31^\circ\ [\text{mV}].$$

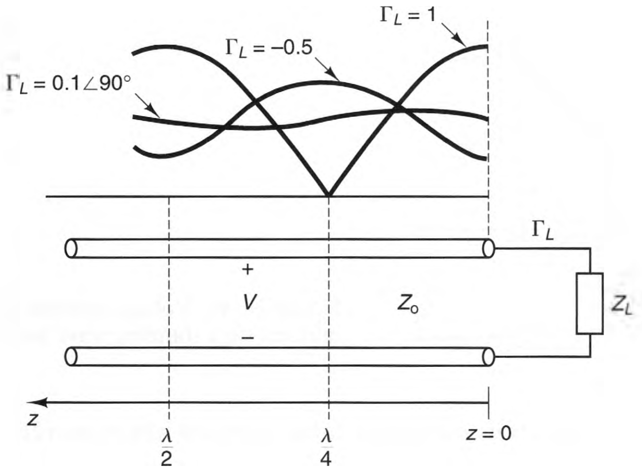


Figure 11-39 Voltage standing wave patterns on a terminated transmission line for three different loads.

11-4-7 STANDING WAVES AND VSWR

Figure 11-39 shows a lossless transmission line with characteristic impedance Z_o, terminated with a load impedance Z_L. If positive values of z lie to the left of the load, the total voltage at any position along the line is given by

$$V(z) = V_{\text{inc}} e^{j\beta z} + \Gamma_L V_{\text{inc}} e^{-j\beta z},$$

where V_{inc} is the complex amplitude of the wave that is incident upon the load and Γ_L is the load reflection coefficient. This expression can be rewritten in the form

$$V(z) = V_{\text{inc}}[e^{j\beta z} + |\Gamma_L| e^{-j(\beta z - \phi_\Gamma)}],$$

or

$$V(z) = V_{\text{inc}}\left[\cos\beta z + j\sin\beta z + |\Gamma_L|\cos(\beta z - \phi_\Gamma) - j|\Gamma_L|\sin(\beta z - \phi_\Gamma)\right],$$

where $|\Gamma_L|$ and ϕ_Γ are the magnitude and phase of Γ_L, respectively. We can find $|V(z)|$ by taking the square root of the sum of the squares of its real and imaginary parts. Remembering that $\sin^2\theta + \cos^2\theta = 1$, this yields

$$|V(z)| = |V_{\text{inc}}|\sqrt{1 + |\Gamma_L|^2 + 2|\Gamma_L|\cos(2\beta z - \phi_\Gamma)}. \tag{11.151}$$

Figure 11-40 shows plots of $|V(z)|$ for three different values of Γ_L. As can be seen, $|V(z)|$ oscillates between maximum and minimum values, with a period of $\lambda/2$. These magnitude distributions are called ***standing wave patterns*** because they do not change with time.

The ratio of the maximum and minimum voltages of a standing wave are directly related to the magnitude of the reflection coefficient. To show this, consider the standing wave pattern depicted in Figure 11-40. Here, $|V|_{\text{max}}$ and $|V|_{\text{min}}$ are the maximum and minimum steady-state voltage magnitudes, respectively. According to

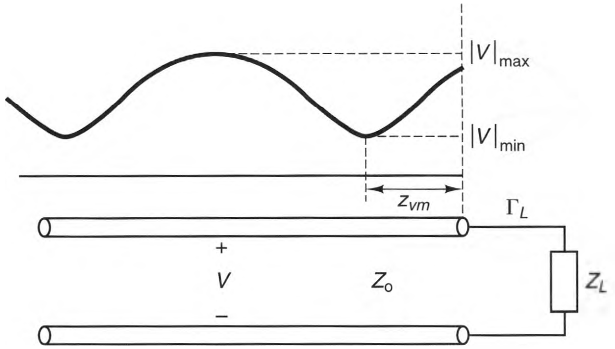


Figure 11-40 Voltage minima and maxima on a standing wave pattern.

Equation (11.151), $|V(z)|$ attains its maximum value at points where $\cos(2\beta z - \phi_\Gamma) = 1$. Thus,

$$|V|_{\max} = |V_{\text{inc}}|\sqrt{1 + |\Gamma_L|^2 + 2|\Gamma_L|} = |V_{\text{inc}}|(1 + |\Gamma_L|). \tag{11.152}$$

Similarly, points of minimum voltage amplitude (called nodes) occur when $\cos(2\beta z - \phi_\Gamma) = -1$, and

$$|V|_{\min} = |V_{\text{inc}}|\sqrt{1 + |\Gamma_L|^2 - 2|\Gamma_L|} = |V_{\text{inc}}|(1 - |\Gamma_L|). \tag{11.153}$$

We define the ***voltage standing wave ratio*** (VSWR) as the ratio of the maximum and minimum voltage magnitudes:

$$\text{VSWR} \equiv \frac{|V|_{\max}}{|V|_{\min}}. \tag{11.154}$$

Substituting Equations (11.152) and (11.153) into this expression, we obtain

$$\text{VSWR} = \frac{1 + |\Gamma_L|}{1 - |\Gamma_L|}. \tag{11.155}$$

Since $|\Gamma_L| \leq 1$, $1 \leq \text{VSWR} \leq \infty$. Solving Equation (11.155) for $|\Gamma_L|$, we can also write

$$|\Gamma_L| = \frac{\text{VSWR} - 1}{\text{VSWR} + 1}. \tag{11.156}$$

Equation (11.156) shows that the magnitude of the reflection coefficient caused by an unknown load can be found by measuring the VSWR. This is important, because it is often easier to measure the VSWR than it is to measure $|\Gamma_L|$ directly, since measuring $|\Gamma_L|$ directly requires a device that can distinguish between waves propagating in opposite directions. By substituting Equation (11.156) into Equations (11.152) and (11.153), expressions for $|V|_{\max}$ and $|V|_{\min}$ in terms of the VSWR can be obtained. We have

$$|V|_{\max} = 2|V_{\text{inc}}| \frac{\text{VSWR}}{\text{VSWR} + 1} \tag{11.157}$$

and

$$|V|_{\min} = 2|V_{\text{inc}}| \frac{1}{\text{VSWR} + 1}, \tag{11.158}$$

where $|V_{\text{inc}}|$ is the magnitude of the wave that is incident upon the load. Equation (11.157) is particularly important, since it shows that the VSWR dictates the maximum voltage on the line for a given incident wave.

Example 11-16

Figure 11-41 shows a transmission line attached to a matched generator and an arbitrary load. If $V_g = 1{,}000$ [V rms], and the breakdown voltage of the transmission line is 700 [V rms], find the range of acceptable VSWRs which guarantee that no breakdown will occur anywhere on the line.

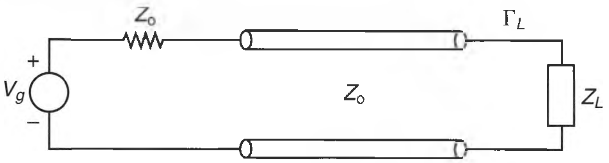


Figure 11-41 A transmission line with a matched generator and unmatched load.

Solution:

It is usually good engineering practice to make sure that $|V|_{\max}$ is less than the specified breakdown value, with a 10% margin of safety. Since the specified value for this cable is 700 [V], we will use 630 [V rms] as the maximum allowable voltage. The generator is matched to the line, so the voltage wave launched towards the load has a magnitude $|V_{\text{inc}}| = V_g/2 = 500$. Solving Equation (11.157) for the maximum VSWR, we find that

$$\text{VSWR} = \frac{|V|_{\max}}{2|V_{\text{inc}}| - |V|_{\max}} = \frac{630}{1000 - 630} = 1.7.$$

Thus, breakdown will be avoided as long as $1 \leq \text{VSWR} \leq 1.7$.

Measuring the VSWR is sufficient to determine the magnitude of Γ_L, but not its phase. Fortunately, the phase of Γ_L can be determined simply by measuring the distance z_{vm} between the first voltage minimum (node) and the load. (See Figure 11-40.)

To see how this is accomplished, we remember that voltage minima occur whenever $\cos(2\beta z - \phi_\Gamma) = -1$, which in turn occurs when

$$2\beta z_{vm} - \phi_\Gamma = \pi.$$

Using $\beta = 2\pi/\lambda$, we can solve this expression for ϕ_Γ, yielding

$$\phi_\Gamma = 720^\circ \left[\frac{z_{vm}}{\lambda} - \frac{1}{4}\right]. \tag{11.159}$$

Even when z_{vm} cannot be measured directly, it can be deduced from the VSWR pattern if the load can be temporarily replaced with a short circuit and we measure the shift in the nodes. This "trick" works because the nodes caused by a short-circuit load occur at integer multiples of $\lambda/2$ behind the short. When using this technique, we observe that z_{vm} is the distance from a load minimum to the closest short-circuit minimum that lies towards the load.

Example 11-17

Figure 11-42 shows an unknown load attached to a 50 [Ω] transmission line. Also shown are the VSWR patterns along a section of the transmission line with the unknown load in place and with the load replaced by a short circuit. Find the impedance of the load.

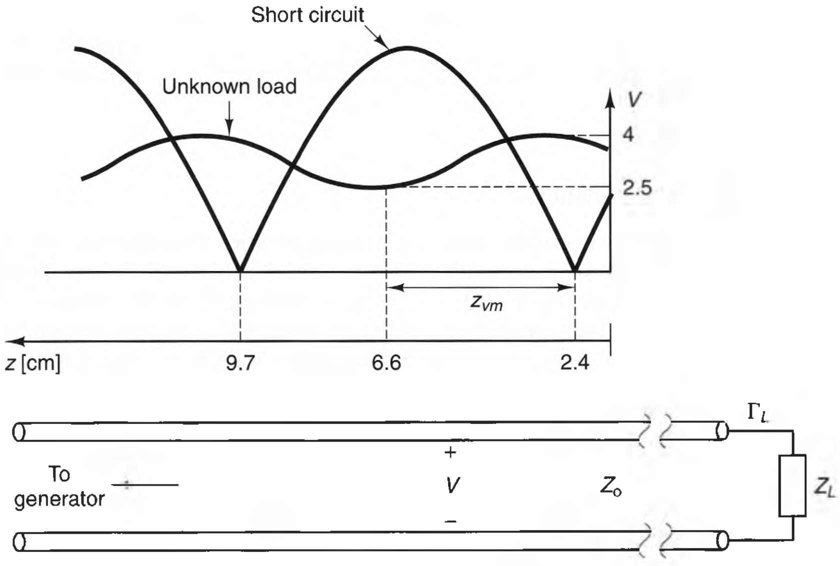


Figure 11-42 Standing wave patterns on a transmission line for an unknown load and a short-circuit load.

Solution:

With the unknown load in place, $V_{\max} = 4$ and $V_{\min} = 2.5$, so

$$\text{VSWR} = \frac{4}{2.5} = 1.6.$$

Substituting this value into Equation (11.156), we find that

$$|\Gamma_L| = \frac{1.6 - 1}{1.6 + 1} = 0.231.$$

To find ϕ_Γ, we must find both λ and z_{vm}. The wavelength λ equals twice the distance between adjacent nodes when either the unknown load or the short circuit is present, but the nodal points are sharper when the short circuit is present. From Figure 11-42, the positions of two successive nodal points are $z_1 = 2.4$ [cm] and $z_2 = 9.7$ [cm]. Thus,

$$\frac{\lambda}{2} = 9.7 - 2.4 = 7.3 \text{ [cm]},$$

or

$$\lambda = 14.6 \text{ [cm]}.$$

Also, z_{vm} is the distance that the nodes shift towards the load when the short circuit is in place, and

$$z_{vm} = 6.6 - 2.4 = 4.2 \text{ [cm]}.$$

Using Equation (11.159), we have

$$\phi_\Gamma = 720° \left[\frac{4.2}{14.6} - \frac{1}{4}\right] = 27.12°.$$

Thus,

$$\Gamma_L = 0.231 \angle 27.12° = 0.205 + j\,0.105.$$

Finally, from Equation (11.132), the load impedance is

$$Z_L = 50\,\frac{1 + 0.205 + j\,0.105}{1 - 0.205 - j\,0.105} = 73.68 + j\,16.38 \text{ [}\Omega\text{]}.$$

11-4-8 EFFECTIVE REFLECTION COEFFICIENTS

We have already seen that the load reflection coefficient Γ_L is the ratio of the reflected and incident voltages, evaluated at the load. In many cases it is helpful to keep track of the relationship between incident and reflected waves at arbitrary points on the line. Such a situation is depicted in Figure 11-43. Here, a transmission line with characteristic impedance Z_0 is terminated at $z = 0$ with a load impedance Z_L. If positive values of z are defined to the left of the load, we can define the ***effective reflection coefficient*** at $z = \ell$ as

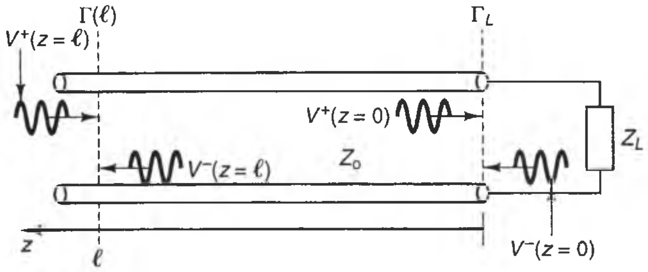


Figure 11-43 The effective reflection coefficient $\Gamma(\ell)$ at an arbitrary point on a transmission line.

$$\Gamma(\ell) = \frac{V^-(z = \ell)}{V^+(z = \ell)}, \tag{11.160}$$

where $V^+(z = \ell)$ and $V^-(z = \ell)$ are the incident and reflected voltages, respectively, at $z = \ell$.

The effective coefficient $\Gamma(\ell)$ is directly related to the load reflection coefficient Γ_L. To see how, we note that because V^+ and V^- propagate in the positive and negative directions, respectively, we can write

$$V^+(z = \ell) = V^+(z = 0)\, e^{j\gamma\ell}$$

and

$$V^-(z = \ell) = V^-(z = 0)\, e^{-j\gamma\ell}.$$

Substituting these into Equation (11.160), we find that

$$\Gamma(\ell) = \Gamma_L e^{-2\gamma\ell}. \tag{11.161}$$

Remembering that $\gamma = \alpha + j\beta$, we can also write

$$\Gamma(\ell) = \Gamma_L e^{-2\alpha\ell}\, e^{-j2\beta\ell}. \tag{11.162}$$

For lossless transmission lines, $\alpha = 0$, which yields

$$\Gamma(\ell) = \Gamma_L e^{-j2\beta\ell} \quad \text{(Lossless transmission lines).} \tag{11.163}$$

From Equation (11.163), we see that the magnitude of $\Gamma(\ell)$ is independent of ℓ on lossless lines. This occurs because the magnitudes of both the incident and reflected fields do not vary with position when there is no loss. On the other hand, the phase of $\Gamma(\ell)$ becomes more and more negative (i.e., delayed) as ℓ increases. This delay occurs because a wave launched towards the load must experience a propagation delay of $\beta\ell$ as it propagates towards the load, and the reflected wave experiences the same delay while propagating back.

On lossy lines, $|\Gamma(\ell)|$ gets smaller as ℓ increases. This occurs because the incident wave diminishes as it approaches the load, and the reflected wave is further diminished as it propagates back. As a result, a highly reflecting load appears less reflecting when viewed through a section of lossy transmission line. This effect can be used to reduce

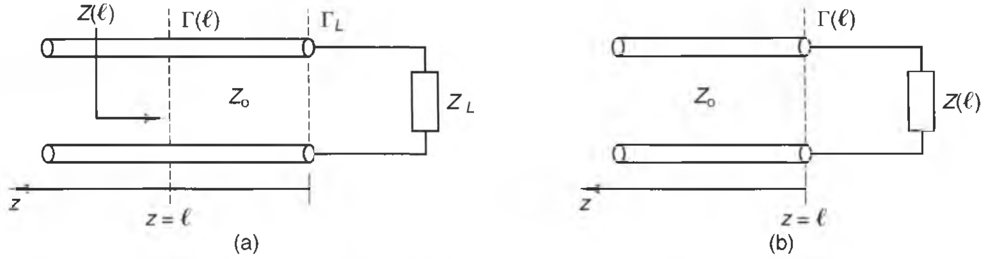


Figure 11-44 a) A transmission line terminated in a lumped load. b) An equivalent load $Z(\ell)$ at $z = \ell$ that yields the same effective reflection coefficient.

the reflections of mismatched loads. Using this technique, very low reflection coefficients can be attained, but at the cost of power dissipated in the lossy matching section.

We saw earlier that the load reflection coefficient Γ_L is related to the load impedance by the expression

$$\Gamma_L = \frac{Z_L - Z_o}{Z_L + Z_o}.$$

A similar relationship exists between the effective reflection coefficient $\Gamma(\ell)$ and the input impedance $Z(\ell)$ at an arbitrary point $z = \ell$. To derive this relationship, consider the circuit shown in Figure 11-44a. Here, a load with impedance Z_L is connected to a transmission line with characteristic impedance Z_o.

Figure 11-44b shows an equivalent circuit, where the transmission line to the right of $z = \ell$ and the load have been replaced by the input impedance $Z(\ell)$, given by Equation (11.134) (or Equation (11.133), for lossy lines). As far as an observer to the left of $z = \ell$ is concerned, the circuits shown in Figures 11-44a & b are identical. Since a load $Z(\ell)$ appears at $z = \ell$ in Figure 11-43b, we can write the reflection coefficient at this point as

$$\Gamma(\ell) = \frac{Z(\ell) - Z_o}{Z(\ell) + Z_o}. \tag{11.164}$$

Here, we see that the effective reflection coefficient $\Gamma(\ell)$ at any point on a transmission line can be determined from the input impedance at that point. Conversely, we can solve Equation (11.164) for $Z(\ell)$ to obtain

$$Z(\ell) = Z_o \frac{1 + \Gamma(\ell)}{1 - \Gamma(\ell)}, \tag{11.165}$$

which shows that the input impedance can be found from the effective reflection coefficient.

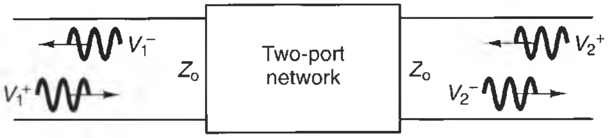


Figure 11-45 Ingoing and outgoing voltagc waves at the ports of a two-port network.

11-4-9 SCATTERING PARAMETERS AND NETWORK ANALYZERS

A useful way to represent the port characteristics of RF and microwave circuits is by using scattering parameters, which are often called S parameters. We will introduce these parameters using Figure 11-45, which shows a linear two-port network. Here, both ports are transmission lines with characteristic impedance Z_o. As is always the case on transmission lines, voltage waves can propagate in two directions. Let us denote the incoming and outgoing voltage phasors at port 1 as V_1^+ and V_1^-, respectively. Similarly, V_2^+ and V_2^- are the phasors of the incoming and outgoing voltage waves at port 2. The S parameters relate the waves according to following equations:

$$V_1^- = S_{11}V_1^+ + S_{12}V_2^+ \tag{11.166a}$$

$$V_2^- = S_{21}V_1^+ + S_{22}V_2^+ . \tag{11.166b}$$

These equations can also be written in matrix form as

$$\begin{bmatrix} V_1^- \\ V_2^- \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} V_1^+ \\ V_2^+ \end{bmatrix} . \tag{11.167}$$

From the preceding equations, we can express each S parameter in terms of the ratio of an outgoing and incoming voltage phasor:

$$S_{11} = \frac{V_1^-}{V_1^+} \quad \text{when } V_2^+ = 0 \text{ (Port 2 matched)} \tag{11.168a}$$

$$S_{12} = \frac{V_1^-}{V_2^+} \quad \text{when } V_1^+ = 0 \text{ (Port 1 matched)} \tag{11.168b}$$

$$S_{21} = \frac{V_2^-}{V_1^+} \quad \text{when } V_2^+ = 0 \text{ (Port 2 matched)} \tag{11.168c}$$

$$S_{22} = \frac{V_2^-}{V_2^+} \quad \text{when } V_1^+ = 0 \text{ (Port 1 matched).} \tag{11.168d}$$

Here, we see that each S parameter is the ratio of an outgoing wave to an incoming wave, under the restriction that one of the ports is terminated with a nonreflecting (i.e., matched) load.

Comparing the definition of S_{11} (Equation (11.168a)) with Equation (11.131), we see that S_{11} is simply the reflection coefficient seen at port 1 when port 2 is terminated by a matched (i.e., nonreflecting) load. Similarly, S_{22} is the reflection coefficient seen at port 2 when port 1 is terminated with a matched load.

The parameters S_{12} and S_{21} are also ratios of outgoing and incoming waves, but they are not reflection coefficients, since the waves are not at the same ports. Rather, these parameters represent the coupling of waves from one port to the other. For instance, S_{21} represents the wave that exits port 2 due to a wave incident upon port 1. Similarly, S_{12} represents the wave that exits port 1 due to a wave incident upon port 2.

The S parameters of a two-port network are directly related to its Z parameters. We can show this by noting that the total voltage and current at each port are related to the incoming and outgoing voltages. Using Equations (11.80) and (11.81), we have

$$V_1 = V_1^+ + V_1^- \tag{11.169a}$$

$$I_1 = \frac{1}{Z_o}[V_1^+ - V_1^-], \tag{11.169b}$$

where V_1 and I_1 are the total voltage and current phasors at port 1. Similarly, at port 2 we have

$$V_2 = V_2^+ + V_2^- \tag{11.170a}$$

$$I_2 = \frac{1}{Z_o}[V_2^+ - V_2^-]. \tag{11.170b}$$

Substituting Equations 11.169 and 11.170 into Equations 11.138, 11.139, and 11.166, and, using matrix algebra, we can derive the following relationships between the Z and S parameters:

$$[\boldsymbol{S}] = [[\boldsymbol{Z}] - Z_o[\boldsymbol{I}]][[\boldsymbol{Z}] + Z_o[\boldsymbol{I}]]^{-1} \tag{11.171}$$

$$[\boldsymbol{Z}] = -Z_o[[\boldsymbol{S}] - [\boldsymbol{I}]]^{-1}[[\boldsymbol{S}] + [\boldsymbol{I}]]. \tag{11.172}$$

Here, $[\boldsymbol{Z}]$ and $[\boldsymbol{S}]$ are the impedance and scattering matrices, respectively, $[\boldsymbol{I}]$ is the identity matrix, and the superscript "-1" denotes "matrix inverse." Thus, one can determine the Z parameters from the S parameters, and vice versa.

Given that the Z and S parameters of a network can each be found one from another, the reader might be wondering why the S parameters are ever used. The following are three reasons why S parameters are useful for describing networks at high frequencies:

(1) Measuring the Z parameters of a network requires placing open circuits at the ports. But many active devices (such as transistors and FETs) will spring into parasitic oscillations when a port is open circuited. This does not happen as often when matched loads are used, so S parameters are easier to measure.

(2) At high frequencies, it is often easier to visualize and measure traveling voltage waves than total voltages. Hence, the S parameters provide a more natural representation at these frequencies.

(3) Scattering matrices have several useful mathematical properties that make them easier to manipulate.[7]

[7] See Samuel Liao, *Microwave Devices and Circuits*, 3d ed., (Upper Saddle River, New Jersey: Prentice-Hall, 1990).

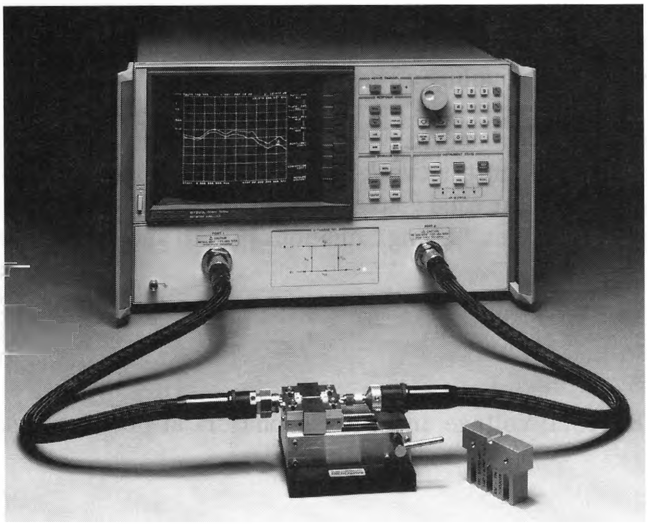


Figure 11-46 A commercial network analyzer, displaying the S-parameters of a 2-port microwave network. Courtesy of the Hewlett-Packard Company.

A ***network analyzer*** is a piece of test equipment that measures the S parameters of a network. A typical network analyzer is shown in Figure 11-46. Network analyzers usually have two ports, each with standard 50 [Ω] coaxial connectors that allow quick connections to the networks being tested. Most modern network analyzers are computer controlled and allow the automatic characterization of networks over wide frequency bands. In addition to S parameters, most analyzers also provide network Z and Y parameters (using relations like Equation (11.172)). Also, many analyzers utilize Fourier transform techniques to simulate the transient responses.

11-4-10 THE SMITH CHART

The Smith chart is a graphical representation of transmission-line parameters that is used both for numerical calculations and for presenting design parameters in a visual setting. Although its use in numerical calculations has diminished with the advent of electronic computing, it remains an important visual design tool for RF and microwave circuits.

Figure 11-47 shows a section of lossless transmission line with characteristic impedance Z_o and length ℓ, and terminated with a load impedance Z_L. The impedance $Z(\ell)$ seen looking into the end of the transmission line is given by the impedance transformation formula

$$Z(\ell) = Z_o \frac{Z_L + jZ_o \tan(\beta\ell)}{Z_o + jZ_L \tan(\beta\ell)}.$$

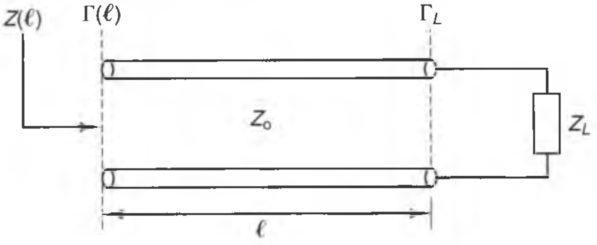


Figure 11-47 The input impedance and effective reflection coefficient a distance ℓ from a load on a transmission line.

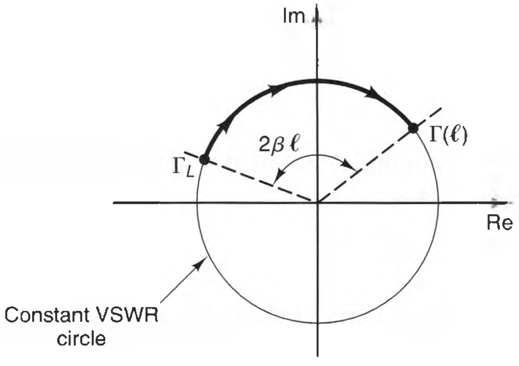


Figure 11-48 The variation of the reflection coefficient $\Gamma(\ell)$ vs. position on a lossless transmission line.

This is a periodic function, since the values of $Z(\ell)$ change in both amplitude and phase as ℓ changes. As a result, it is not easy to tell how the numerical values of $Z(\ell)$ change from position to position simply by looking at the formula. On the other hand, the effective reflection coefficient Γ changes quite simply as a function of the distance ℓ from the load, according to the formula

$$\Gamma(\ell) = \Gamma_L e^{-j2\beta\ell},$$

where Γ_L is the load reflection coefficient. From this expression, we see that only the phase of Γ changes with ℓ. This means that $\Gamma(\ell)$ traces out a circle on the complex plane each time that ℓ increases by $\lambda/2$, such as the circle shown in Figure 11-48. Thus, if the value of $\Gamma(\ell)$ is known for one value of ℓ, its value at any other position can be quickly determined simply by rotating through the appropriate angle—clockwise when moving away from the load and counterclockwise when moving towards the load. Constant $|\Gamma|$ circles are called ***constant-VSWR circles***, since the VSWR is a function of only the magnitude of $\Gamma(\ell)$ and not its phase. (See Equation 11.155.)

To see what kind of graphical relationship the impedance and reflection coefficient have, let us first write $\Gamma(\ell)$ as the sum of a real part Γ_r and an imaginary part Γ_i. We have

$$\Gamma = \Gamma_r + j\Gamma_i,$$

where we have dropped the "(ℓ)" from these quantities to simplify the notation. Next, we define the ***normalized impedance*** z as the input impedance Z divided by the characteristic impedance Z_o. The definition is

$$z \equiv \frac{Z}{Z_o} = r + jx, \tag{11.173}$$

where r and x are the ***normalized resistance*** and ***reactance***, respectively. Notice here that z, r, and x are all dimensionless parameters. Substituting Equation (11.173) into Equation (11.164), we obtain

$$\Gamma = \frac{z-1}{z+1}, \tag{11.174}$$

which can also be written as

$$z = \frac{1+\Gamma}{1-\Gamma}. \tag{11.175}$$

If we express both z and Γ in terms of their real and imaginary parts, Equation (11.165) becomes

$$r + jx = \frac{1 + \Gamma_r + j\Gamma_i}{1 - \Gamma_r - j\Gamma_i}.$$

Multiplying both the numerator and denominator of the right side by the complex conjugate of the denominator yields

$$r + jx = \frac{1 - \Gamma_r^2 - \Gamma_i^2 + j\,2\Gamma_i}{(1 - \Gamma_r)^2 + \Gamma_i^2}.$$

Equating the real and imaginary parts of the right- and left-hand sides of this expression, we obtain

$$r = \frac{1 - \Gamma_r^2 - \Gamma_i^2}{(1 - \Gamma_r)^2 + \Gamma_i^2} \quad \text{and} \quad x = \frac{2\Gamma_i}{(1 - \Gamma_r)^2 + \Gamma_i^2}.$$

Finally, these expressions can be rearranged to obtain the equations that define the Smith chart, namely,

$$\left(\Gamma_r - \frac{r}{r+1}\right)^2 + \Gamma_i^2 = \left(\frac{1}{1+r}\right)^2 \tag{11.176}$$

and

$$(\Gamma_r - 1)^2 + \left(\Gamma_i - \frac{1}{x}\right)^2 = \left(\frac{1}{x}\right)^2. \tag{11.177}$$

Equation (11.176) defines a family of circles in the (Γ_r, Γ_i) plane called ***constant-r circles***, since each circle corresponds to a particular value of r. Each circle has a radius of $1/(1 + r)$ and is centered at the point $(r/(1 + r), 0)$. Figure 11-49a shows several constant-r circles. Each circle represents all the values of Γ that correspond to a particular value of r. If Γ is known at a particular location on the transmission line,

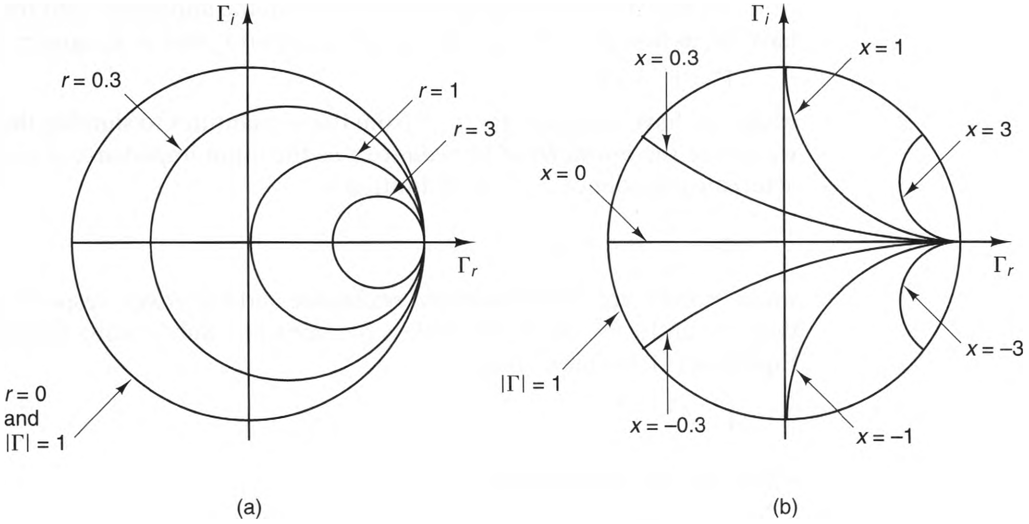


Figure 11-49 a) Constant-r circles and b) Constant-x circles, plotted in Cartesian coordinates.

the value of r at that location can be found simply by plotting Γ in the complex plane and locating the constant-r circle that intersects this point.

In a similar manner, Equation (11.177) defines a family of circles in the (Γ_r, Γ_i) plane called ***constant-x circles***. Each circle has a radius of $1/x$ and is centered at the point $(1, 1/x)$. Several constant-x circles are plotted in Figure 11-49b. Just as with the constant-r circles, the normalized reactance x that corresponds to a particular value of Γ can be determined by noting which constant-x circle intersects Γ on the (Γ_r, Γ_i) plane.

When the constant-r and constant-x circles are plotted together, as shown in Figure 11-50, we obtain the ***Smith chart***, which was devised in 1939 by P.H. Smith.[8] Each point on the Smith chart is a simultaneous plot of the reflection coefficient Γ and the

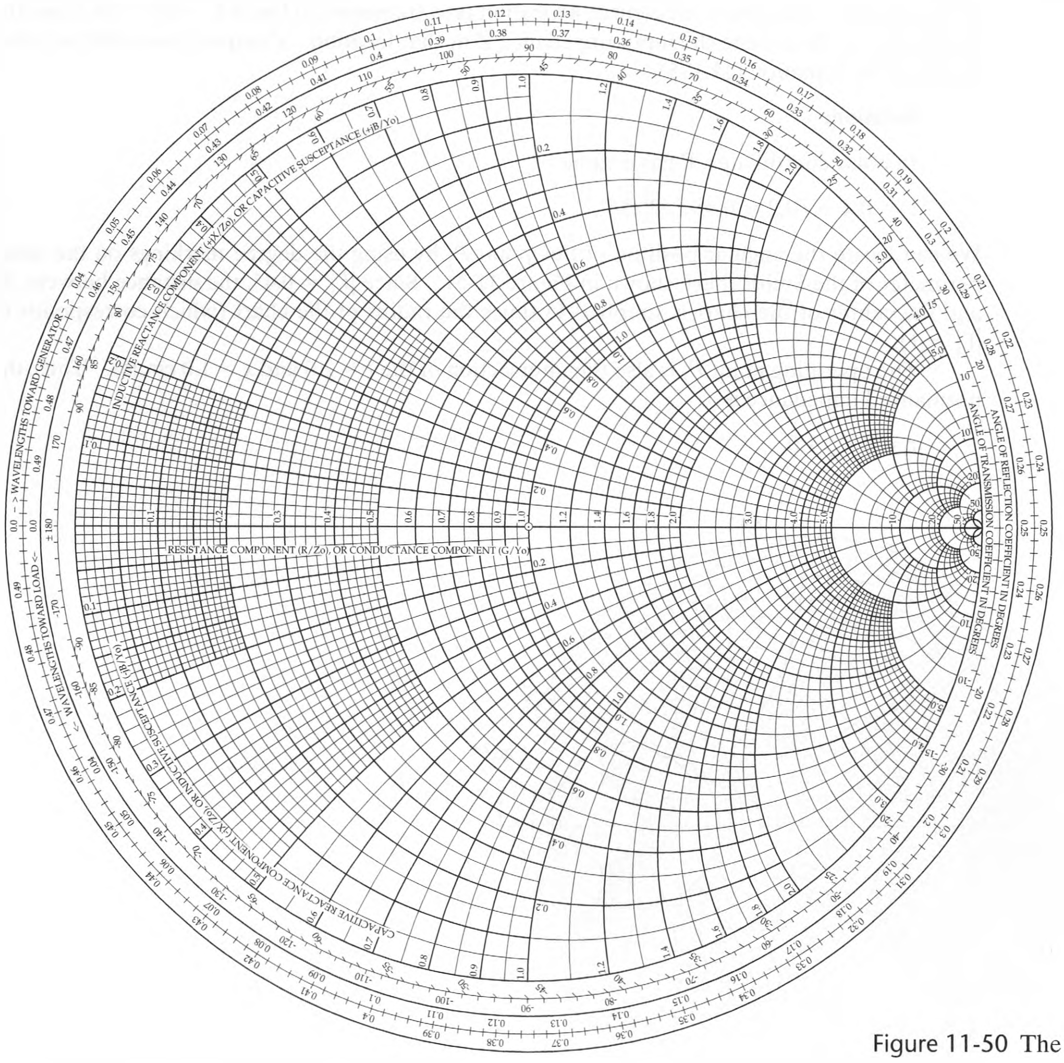

WAVELENGTHS TOWARD GENERATOR —>
<— WAVELENGTHS TOWARD LOAD <—
INDUCTIVE REACTANCE COMPONENT (+jX/Zo), OR CAPACITIVE SUSCEPTANCE (+jB/Yo)
CAPACITIVE REACTANCE COMPONENT (-jX/Zo), OR INDUCTIVE SUSCEPTANCE (-jB/Yo)
RESISTANCE COMPONENT (R/Zo), OR CONDUCTANCE COMPONENT (G/Yo)
ANGLE OF REFLECTION COEFFICIENT IN DEGREES
ANGLE OF TRANSMISSION COEFFICIENT IN DEGREES

Figure 11-50 The Smith chart.

[8] P.H. Smith, "Transmission-line calculator," *Electronics*, vol. 12, p. 29, January 1939. Also, "An improved transmission-line calculator," *Electronics*, vol. 17, p. 130, January 1944.

normalized impedance z at a particular location on a transmission line. Both Γ and z are complex numbers, but they are plotted differently on the Smith chart. The real and imaginary parts of the reflection coefficient, Γ_r and Γ_i, respectively, are plotted in rectangular coordinates, with the horizontal axis representing Γ_r and the vertical axis representing Γ_i. On the other hand, the values of r and x are located on the Smith chart using the constant-r and constant-x circles, respectively. In this way, the value of z that corresponds to a particular reflection coefficient Γ can be read directly off the Smith chart, and vice versa. The process is demonstrated in the example that follows.

Example 11-18

If the effective reflection coefficient at a location on a transmission line is $\Gamma = 0.4 + j\,0.2$, use the Smith chart to determine the input impedance Z at that location. Compare this with the value predicted by Equation (11.175).

Solution:

In polar coordinates, Γ has a value of

$$\Gamma = 0.4 + j\,0.2 = 0.45 \angle 26.56°.$$

We can locate the angular position of this point P_1 by using the degree markings on the outer perimeter of the Smith chart, shown in Figure 11-51. Since $|\Gamma| = 0.45$, the distance between P_1 and the center of the chart is simply 0.45 times the radius of the chart (which corresponds to $|\Gamma| = 1$).

The operating point P_1 is also the intersection of the $r = 2.0$ and $x = 1.0$ circles. Thus, the normalized impedance is

$$z = 2 + j\,1.0.$$

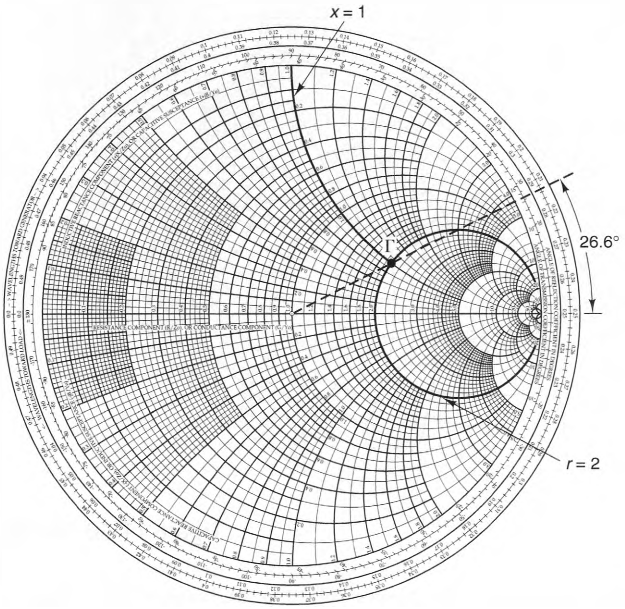


Figure 11-51 Smith chart for Example 11-18.

Since the characteristic impedance of the line is 50 [Ω], the input impedance seen looking towards the load is

$$Z_{\text{in}} = 50z = 100 + j\,50\ [\Omega].$$

Finally, using Equation (11.175), we have

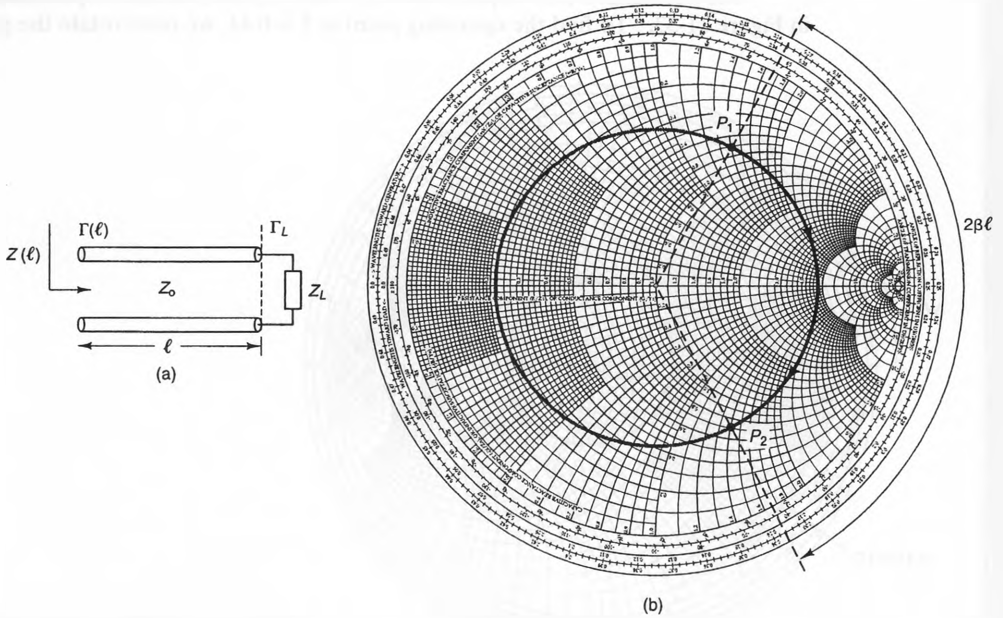
$$z = \frac{1+\Gamma}{1-\Gamma} = \frac{1 + 0.4 + j\,0.2}{1 - 0.4 - j\,0.2} = 2 + j\,1.0,$$

which agrees with the result obtained from the Smith chart.

The previous example shows how the Smith chart can be used to determine the normalized input impedance z of a transmission line from the effective reflection coefficient Γ. Conversely, the Smith chart can also be used to determine the effective reflection coefficient Γ that corresponds to a particular normalized input impedance z. As important as these functions are, however, the real utility of the Smith chart is that it allows for quick graphical impedance transformations from one position to another on a transmission line.

Consider the situation depicted in Figure 11-52a. Here, a load is connected to a lossless transmission line with characteristic impedance Z_o. We will denote the normalized impedance and reflection coefficient at the load as $z_L = r_L + jx_L$ and $\Gamma_L = |\Gamma_L| \angle \phi_L$, respectively. These values are represented by the point P_1 in Figure 11-52b. Using Equation (11.163), we find that the effective reflection coefficient seen a distance ℓ from the load is given by

$$\Gamma = \Gamma_L e^{-j2\beta\ell} = |\Gamma_L| \angle (\phi_L - 2\beta\ell).$$



(a)

(b)

Figure 11-52 Using the Smith chart to compute the normalized impedance vs. position on a lossless transmission line.

In words, this expression says that the new value of Γ is found simply by rotating the point P_1 clockwise through the angle $2\beta\ell = 720° \times (\ell/\lambda)$. This value is represented by the point P_2 in Figure 11-52b. As can be seen, Γ_L and Γ lie on the same constant-VSWR circle. The value of z can be read directly off the Smith chart from the intersection of P_2 with the constant-r and -x circles.

A convenient way to determine how far an operating point on a constant-VSWR circle moves between two positions on a transmission line is to use the wavelength scales along the perimeter of the Smith chart. When an observer moves a distance $\Delta\ell$ towards the load, the phase of the reflection coefficient changes by an amount

$$\Delta\phi = 2\beta\Delta\ell = 720° \times \left(\frac{\Delta\ell}{\lambda}\right), \tag{11.178}$$

This relationship allows us to relate angular positions on the Smith chart using the wavelengths-toward-the-generator (WTG) and wavelengths-toward-the-load (WTL) scales on the outer perimeter of the Smith chart. Movement towards the generator (i.e., away from the load) results in a negative $\Delta\phi$, whereas movement towards the load results in a positive $\Delta\phi$. Notice that one revolution around a VSWR circle corresponds to a change in position of $\lambda/2$.

Example 11-19

A load of value $Z_L = 50 - j\,25$ [Ω] is attached to a lossless, 100 [Ω] transmission line. Use the Smith chart to find Z a distance $\ell = 0.4\lambda$ from the load.

Solution:

The normalized load impedance is $z_L = 0.5 - j\,0.25$, which is represented by the point P_1 in Figure 11-53. To find the operating point at $\ell = 0.4\lambda$, we must rotate the point P_1 clockwise

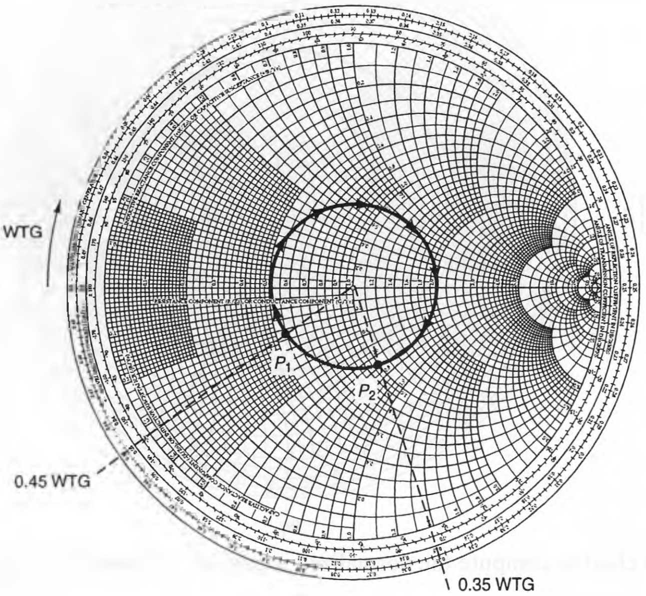


Figure 11-53 Smith chart for Example 11-19.

the appropriate number of degrees. This is most easily accomplished by using the WTG scale at the outer edge of the Smith chart. The point P_1 occurs at 0.45 WTG. Thus, the operating point P_2 in back of the load is found by rotating an additional 0.4λ on the WTG scale. Remembering that a full revolution on the Smith chart is 0.5λ, we find that P_2 is located at $0.4 - (0.5 - 0.45) = 0.35$ WTG. Using the constant-r and -x circles, we obtain

$$z = 0.952 - j\,0.77$$

and

$$Z = 100z = 95.2 - j\,77.0\ [\Omega].$$

Another useful property of the Smith chart is that the VSWR can be read directly off the chart simply by noting where the constant-VSWR circle intersects the real axis. To see how, note that every constant-VSWR circle intersects the real axis twice. Intersections to the right and left of the coordinate center yield $z = r_{\text{max}} \geqslant 1$ and $z = r_{\text{min}} \leqslant 1$, respectively. These are also points of maximum and minimum voltages, respectively, since the incident and reflected waves have phase differences of 0° or 180° when Γ is real. At a point where $z = r_{\text{max}}$,

$$|\Gamma| = \left|\frac{r_{\text{max}} - 1}{r_{\text{max}} + 1}\right| = \frac{r_{\text{max}} - 1}{r_{\text{max}} + 1}.$$

Substituting this into Equation (11.155), we obtain

$$\text{VSWR} = \frac{1 + |\Gamma|}{1 - |\Gamma|} = r_{\text{max}}. \tag{11.179}$$

Using a similar sequence of steps, we can also show that

$$\text{VSWR} = \frac{1}{r_{\text{min}}}. \tag{11.180}$$

The Smith chart can be used to calculate the input admittances as well. At every operating point, the normalized impedance z and the reflection coefficient Γ are related by

$$z = \frac{1 + \Gamma}{1 - \Gamma}. \tag{11.181}$$

If we define the ***normalized admittance*** y as

$$y \equiv \frac{1}{z}, \tag{11.182}$$

it follows from Equation (11.181) that

$$y = \frac{1 - \Gamma}{1 + \Gamma},$$

which can also be written as

$$y = \frac{1 + (-\Gamma)}{1 - (-\Gamma)}. \tag{11.183}$$

Comparing Equations (11.181) and 11.183, we see that normalized impedance and admittance values interchange when the sign of Γ is changed. Since half revolutions on the Smith chart correspond to multiplying Γ by −1, normalized impedance-to-admittance conversions can be obtained graphically on the Smith chart by using the following procedure:

(1) Identify the impedance operating point on the Smith chart by locating either the reflection coefficient Γ or the normalized impedance z. Draw the corresponding constant-VSWR circle.

(2) Locate the admittance operating point by rotating the impedance operating point through 180°.

(3) The normalized admittance can be read directly by interpreting the constant resistance (r) and -reactance (x) circles as constant admittance (g) and -susceptance (b) circles, respectively.

(4) Admittance values at any other position on the line can be obtained by rotating the admittance operating point around the constant-VSWR circle the appropriate number of degrees, using the WTG and WTL scales.

Example 11-20

Figure 11-54 shows a lossless, 50 [Ω] transmission line that is terminated with an unknown impedance. Using the VSWR pattern plotted in this figure, calculate the load impedance and admittance.

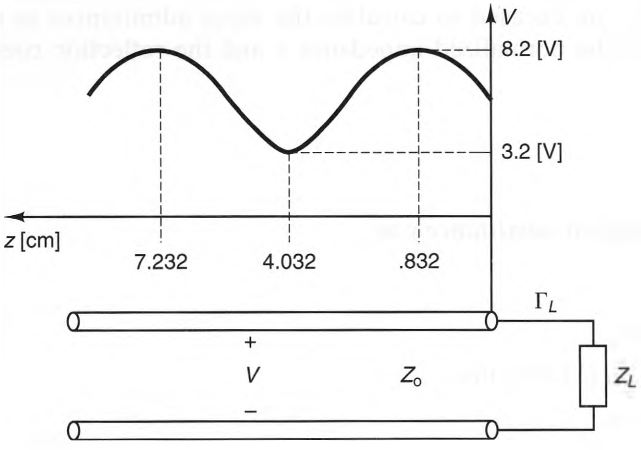


Figure 11-54 VSWR plot for a transmission line with an unknown load.

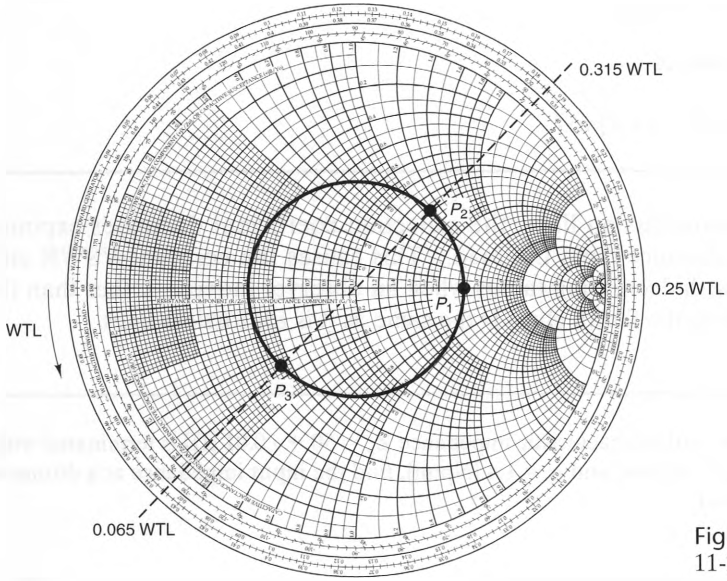


Figure 11-55 Smith chart for Example 11-20.

Solution:

First, we can calculate the wavelength using the distance between successive maxima in Figure 11-54:

$$\lambda = 2(7.232 - 0.832) = 12.8 \text{ [cm]}.$$

Next, from the voltage plot,

$$\text{VSWR} = \frac{V_{\max}}{V_{\min}} = \frac{8.2}{3.2} = 2.56.$$

The constant-VSWR circle, shown in Figure 11-55, passes through the intersection of the $r = 2.56$ circle and the real axis.

To find the load impedance, we note that the impedance at the $z = 0.832$ [cm] voltage maximum is found on the Smith chart at the intersection of the constant-VSWR circle and the positive real axis. This point is indicated as P_1 in Figure 11-55. In terms of wavelengths, the distance from the first maximum to the load is

$$d = \frac{0.832}{12.8} = 0.065\ \lambda.$$

The impedance at the load is obtained by starting at P_1 and rotating counterclockwise 0.065λ on the constant-VSWR circle. Since P_1 occurs at 0.25 on the WTL scale, P_2, is located at $(0.25 + 0.065) = 0.315\lambda$ on the WTL scale. At P_2, we read the normalized impedance

$$z = 1.3854 + j\,1.081,$$

and a load impedance

$$Z_L = 50z = 69.27 + j\,54.05\ [\Omega].$$

The normalized admittance of the load can be found by rotating an additional 1/2 revolution from P_2 (on the constant-VSWR circle) to obtain the point P_3. At P_3, we find

$$y_L = 0.4487 - j\,0.3501$$

and a load admittance of

$$Y_L = \frac{y_L}{50} = 8.97 - j\,7.0\ [\text{mS}].$$

On lossy transmission lines, the reflection coefficient decreases exponentially with increasing distance from the load. This causes the constant-VSWR circles to become spirals with decreasing radius when an observer moves. Other than that, the technique for using the Smith chart is the same as it is for lossless lines.

Example 11-21

A transmission line with characteristic impedance $Z_o = 50 + j\,0.01$ [Ω] is terminated with a load $Z_L = 10$ [Ω]. If $\lambda = 40$ [cm] and $\alpha = 1.4$ [Np/m], find the input impedance at a distance $\ell = 15$ [cm] behind the load.

Solution:

The normalized load impedance is

$$z = \frac{10}{50 + j\,0.01} \approx 0.2 + j\,0,$$

which is represented in Figure 11-56 as the point P_1 at 0.0 on the WTG scale.

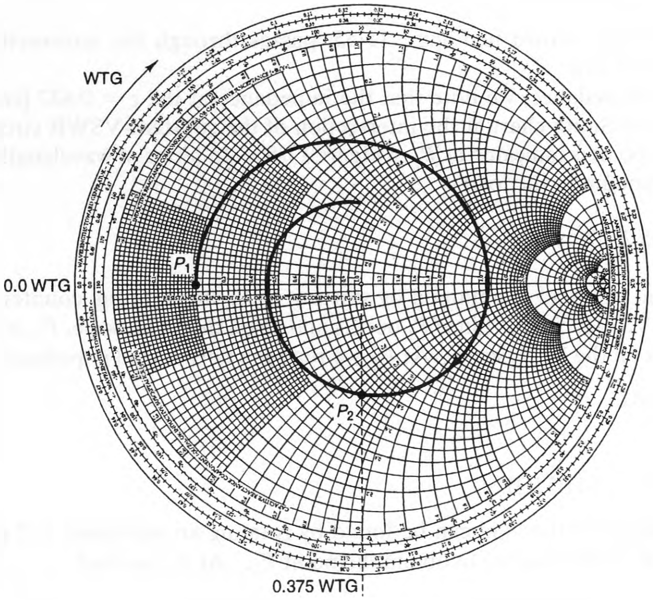


Figure 11-56 Smith chart for Example 11-21.

Remembering that the distance from the origin to the outer perimeter of the Smith chart corresponds to $|\Gamma| = 1$, we see from this figure that $|\Gamma| = 0.666$ at P_1. Using $\alpha = 0.014$ [Np/cm], we find that the value of $|\Gamma|$ a distance $\ell = 15$ [cm] behind the load is

$$|\Gamma| = 0.666 \times e^{-2(.014)(15)} = 0.442.$$

To find the phase of Γ at $\ell = 15$ [cm], we must rotate $15/40 = 0.375$ of a wavelength towards the generator, which is 270° clockwise from P_1. The operating point at $\ell = 15$ [cm] is shown as P_2. Using the constant-r and -x circles, we obtain

$$z_2 = 0.678 - j\,0.735,$$

and

$$Z_2 = (0.678 - j\,0.735) \cdot (50 + j\,0.01) = 33.9 - j\,36.7\ [\Omega].$$

11-4-11 IMPEDANCE MATCHING

Throughout this chapter, we have discussed several characteristics of transmission lines with mismatched loads. Most of these characteristics are undesirable, the worst of which are:

1. Input impedances that vary with line length.
2. Non–optimum power transfer between the source and the load.
3. Waveform distortion.
4. Voltage "hot spots" that can cause dielectric breakdown.

Because of these undesirable effects, measures are often taken to ensure that transmission lines and loads are matched as closely as possible. Sometimes this is as simple as choosing transmission and load impedances that are closely matched. Often, however, one has to make do with an existing load or source. In these cases, it is still possible to improve the match, either by using an impedance transformer or by adding lossless, lumped impedances.

Quarter-Wave Transformer. A quarter-wave transformer is simply a quarter-wavelength-long section of transmission line. When a resistive load is attached to one end of a quarter-wave transformer, the input impedance at the other end is also resistive. Figure 11-57 depicts such a situation, where a resistor of value R_L is attached to a quarter-wavelength section of lossless transmission line with characteristic impedance $\hat{Z}_0$. Noting that $\beta\ell = \pi/2$ when $\ell = \lambda/4$, the impedance trans-

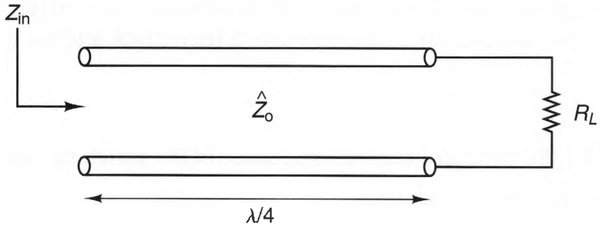


Figure 11-57 A quarter-wave section of transmission line, terminated with a resistive load.

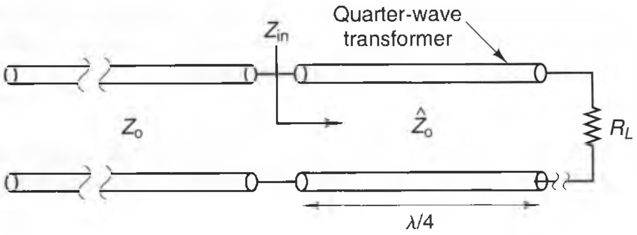


Figure 11-58 A quarter-wave transformer, used to match a resistive load to a transmission line.

formation (Equation (11.134)) predicts that the input impedance is real and given by

$$Z_{\text{in}} = \frac{\hat{Z}_o^2}{R_L}. \qquad (11.184)$$

Since $\hat{Z}_o$ is real for lossless transmission lines, Z_{in} is also real. Hence, R_L can be transformed to any real value simply by choosing a quarter-wavelength transformer with the appropriate characteristic impedance $\hat{Z}_o$.

Figure 11-58 shows how quarter-wave transformers are used to match loads to transmission lines. Here, a resistive load of R_L is to be matched to a transmission line with characteristic impedance Z_o. Between the load and the transmission line is placed a quarter-wave transformer with characteristic impedance $\hat{Z}_o$. In order for the transmission line to "see" a matched impedance, we simply require that the transformed impedance Z_{in} (given by equation 11.184) equal

$$Z_o = \frac{\hat{Z}_o^2}{R_L}.$$

Solving for $\hat{Z}_{o,}$ we find that the quarter-wave matching section must have a characteristic impedance of value

$$\hat{Z}_o = \sqrt{Z_o R_L}. \qquad (11.185)$$

When $\hat{Z}_o$ has this value, the reflection coefficient at the input terminals of the quarter-wave transformer is zero, so all the power incident from the transformer is delivered to the load.

Example 11-22

Design a quarter-wavelength section that matches a 20 [Ω] resistance to a 50 [Ω] microstrip line at $f = 4$ [GHz]. Assume that the dielectric substrate is 0.75 [mm] thick and has a dielectric constant of $\epsilon_r = 2.5$.

Solution:

From Equation (11.185), the characteristic impedance of the matching section must be

$\hat{Z}_o = \sqrt{Z_o R_L} = \sqrt{50 \times 20} = 31.62$ [Ω].

We can use Equations (D.24) and (D.25) to determine the width-to-height ratios of the 50 [Ω] and 31.62 [Ω] PCB traces. For the 50 [Ω] section, we have

$$B = \frac{\pi}{2\sqrt{2.5}}\frac{377\ [\Omega]}{50\ [\Omega]} = 7.49$$

$$\frac{w}{h} = \frac{(2.5-1)}{2.5\pi}\left(\ln(7.49-1) + 0.39 - \frac{0.61}{2.5}\right) + \frac{2}{\pi}\left(7.49 - 1 - \ln(2\times 7.49 - 1)\right)$$

$$= 2.84.$$

Since $h = 0.75$ [mm], this means that $w = 2.13$ [mm].

Similarly, for the 31.62 [Ω] line, we have

$$B = \frac{\pi}{2\sqrt{2.5}}\frac{377\ [\Omega]}{31.62\ [\Omega]} = 11.8$$

$$\frac{w}{h} = \frac{(2.5-1)}{2.5\pi}\left(\ln(11.8-1) + 0.39 - \frac{0.61}{2.5}\right) + \frac{2}{\pi}\left(11.8 - 1 - \ln(2\times 11.8 - 1)\right)$$

$$= 5.37.$$

This means that $w = 5.37 \times 0.75$ [m] $= 4.03$ [mm].

Finally, to determine the physical length of the $\lambda/4$ section, we need to find the wavelength on the 31.62 [Ω] line. Using Equation (D.20), we find that the effective dielectric constant on this line is

$$\epsilon_{\text{eff}} \approx \frac{1}{2}(\epsilon_r + 1) + \frac{1}{2}(\epsilon_r - 1)\left(1 + \frac{12}{5.37}\right)^{-1/2} = 2.17.$$

Using this value, we see that the wavelength on this line is

$$\lambda = \frac{c}{f\sqrt{\epsilon_{\text{eff}}}} = \frac{3\times 10^8}{4\times 10^9 \times \sqrt{2.17}} = 50.9\ [\text{mm}].$$

Hence, the length of the $\lambda/4$ section is

$$\ell = \frac{50.9}{4} = 12.73\ [\text{mm}].$$

Figure 11-59 shows a top view of the input transmission line, the quarter-wave section, and the resistive load.

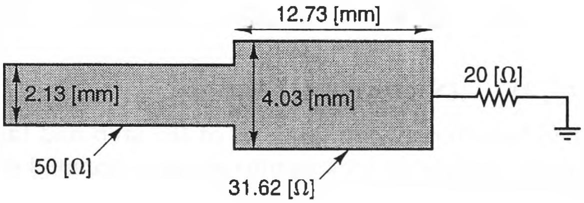


Figure 11-59 A microstrip quarter-wave transformer that matches a 50 [Ω] microstrip line to a 20 [Ω] load.

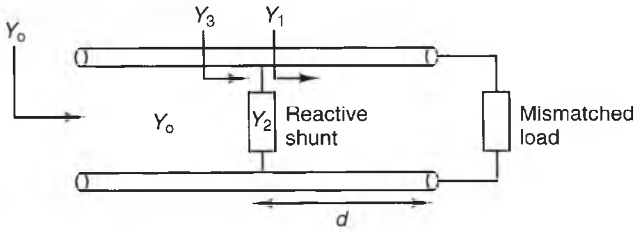


Figure 11-60 Impedance matching using a lumped, reactive shunt.

Stub Tuners. Quarter-wave transformers work well in many situations, but they can only provide perfect matches for resistive loads; reactive loads can't be matched well by quarter-wave transformers. A technique that avoids this problem is depicted in Figure 11-60. Here, a transmission line is terminated with a mismatched load. The essence of this technique is to place a reactive shunt element a distance d away from the load such that the net impedance of the shunt and the transformed impedance of the load exactly equals the characteristic impedance of the transmission line Z_o. When this occurs, the transmission line "sees" a matched load, and all the power from the transmission line is absorbed by the load (since the reactive shunt cannot dissipate power).

The key to this technique lies in choosing the distance d such that the admittance Y_1 seen looking towards the load is of the form

$$Y_1 = Y_o(1 + jb),$$

where $Y_o = 1/Z_o$ is the ***characteristic admittance*** of the transmission line. Next, we choose the admittance Y_2 of the shunt element so that it is given by

$$Y_2 = -jbY_o.$$

For this choice of Y_2, the parallel combination of Y_1 and Y_2 results in a net admittance

$$Y_3 = Y_1 + Y_2 = Y_o(1 + jb) - jbY_o = Y_o,$$

which is a matched admittance. Thus, by choosing the distance d and the reactive admittance Y_2 appropriately, it is possible to match any load (even reactive ones) to a transmission line.

Although any reactive shunt element can be used in this technique, lumped capacitors and inductors are usually too lossy to be of practical use at RF frequencies and above. An attractive alternative is to use sections of short- and open-circuited transmission lines, called stub tuners. Figures 11-61a and b show short-circuited and open-circuited stub tuners, respectively. By choosing the stub lengths appropriately, the necessary shunt susceptances can be obtained. For instance, the admittance Y_s seen looking into a short-circuited stub of length ℓ, shown in Figure 11-61a, is given by Equation (11.136),

$$Y_s = -jY_o\cot(2\pi\ell/\lambda) \qquad \text{(Short-circuited stub)}, \qquad (11.186)$$

where Y_o and λ are the characteristic admittances of the stub and the wavelength, respectively. Similarly, the admittance looking into an open-circuited stub of length ℓ, shown in Figure 11-61b, is given by Equation (11.137),

$$Y_s = jY_o\tan(2\pi\ell/\lambda) \qquad \text{(Open-circuited stub)}. \qquad (11.187)$$

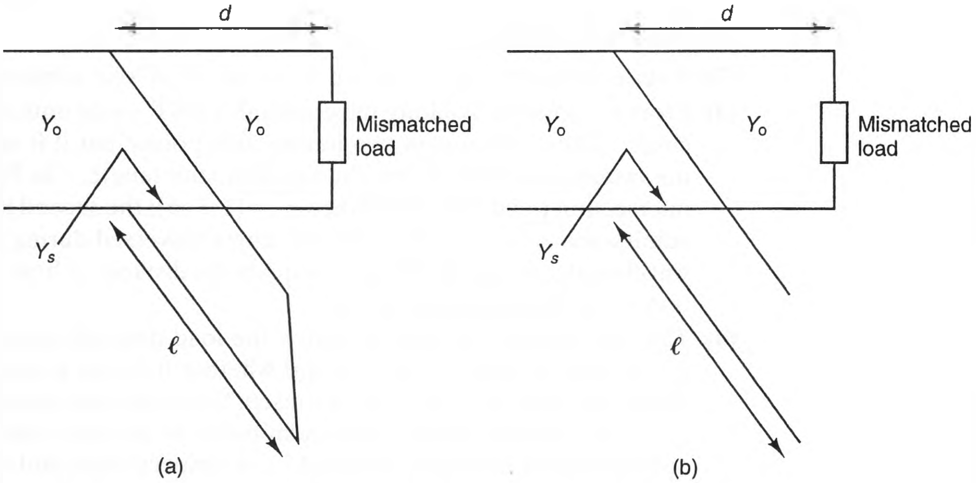


Figure 11-61 a) A short-circuited single-stub tuner. b) An open-circuited single-stub tuner.

Equations (11.186) and (11.187) show that any reactive admittance Y_s can be obtained from an open- or short-circuited stub simply by choosing an appropriate length ℓ.

The appropriate dimensions of a single-stub tuner are easily accomplished by using the Smith chart. Assuming that the characteristic admittances of the transmission line and the stub are equal (Y_o), the single-stub matching technique proceeds as follows (see Figure 11-62):

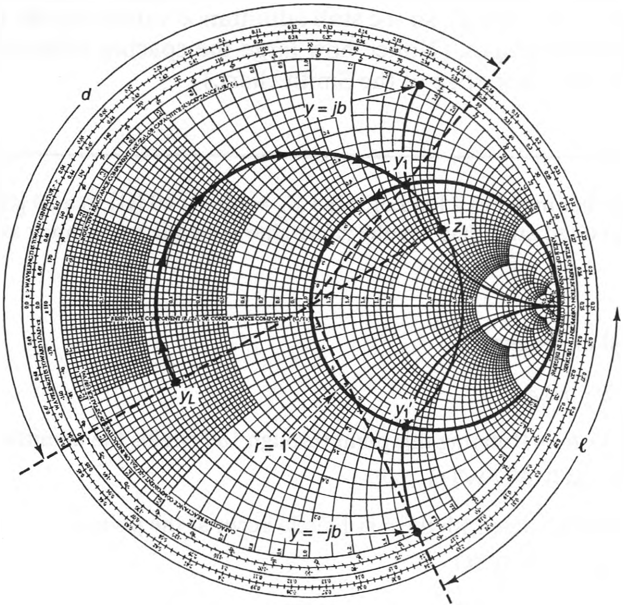


Figure 11-62 Using the Smith chart to determine the position and length of a stub tuner.

(1) Locate the normalized load impedance z_L on the Smith chart, and draw the constant-VSWR circle.

(2) Rotate this point 180° to obtain the normalized load admittance y_L.

(3) From y_L, rotate CW along the constant-VSWR circle until it intersects the $g = 1$ circle. There are always two intersection points, but it is usually best to choose the intersection that yields the smallest stub length. In Figure 11-62, the first intersection point has admittance $y_1 = (1 + jb)$; the second intersection point has admittance $y_1' = (1 - jb)$. The arc length traversed during this rotation (read in wavelengths on the WTG scale) equals the distance d from the load to the location of the lumped reactive load.

(4) The stub length ℓ needed to match the load depends upon which point on the $g = 1$ circle is used (y_1 or y_1'), and whether the stub is open or short circuited. When the point $y_1 = (1 + jb)$ is chosen, the input admittance to the stub must be $y_s = -jb$. To find ℓ, start at the admittance of the stub's end ($y = 0$ or $y = \infty$ for open-circuited and short-circuited stubs, respectively), and rotate clockwise until the point $y_s = -jb$ is reached. For either case, the stub length ℓ equals the arc length traversed on the WTG scale. Figure 11-62 shows the length for a short-circuited stub. The procedure is similar when the point y_1' is chosen, except that the susceptance values of the stubs are positive.

There are several reasons why the shortest stub length is usually the best choice. Among them are the following:

(a) Less space is needed.

(b) All transmission lines have finite losses, so minimizing the stub length also minimizes the power lost in the stub.

(c) The bandwidth over which a good match is obtained is maximized when the short stubs length are used. This is because the VSWR along the stub is infinite (or nearly so when the stub is lossy), so the stub admittance varies rapidly with slight changes in its electrical length. (Notice how large the spacing between the constant b circles is at the outer edge of the Smith chart.)

Example 11-23

Design a single-stub tuner that matches a load of value $Z_L = 60 - j\,40$ [Ω] to a 50 [Ω] transmission line. Use a short-circuited stub that has the same characteristic impedance as the transmission line.

Solution:

The normalized load impedance is

$$z_L = \frac{60 - j\,20}{50} = 1.2 - j\,0.8,$$

which is plotted in Figure 11-63. Rotating this point through 180° yields the admittance:

$$y_L = 0.577 + j\,0.385 \quad \text{at } 0.0777 \text{ WTG.}$$

The intersections of the VSWR circle with the $g = 1$ circle occur at the points

$$y_1 = (1 + j\,0.753) \quad \text{at } 0.1536 \text{ WTG}$$

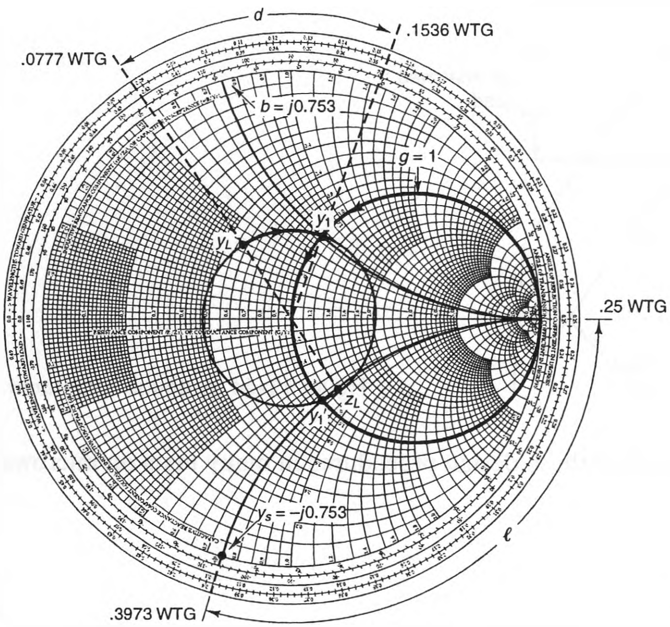


Figure 11-63 Smith chart for Example 11-23.

and

$$y_1' = (1 - j\,0.753) \quad \text{at } 0.3463 \text{ WTG},$$

respectively. Because the matching reactance is a short-circuited stub, we choose the value y_1, since short, short-circuited stubs have negative susceptance. Thus, the optimal distance from the load to the stub is

$$d = 0.1536 - 0.0777 = 0.0759\ \lambda.$$

To find the stub length ℓ, we first locate the stub admittance $y_s = -j\,0.7534$ at 0.3973 on the WTG scale. Rotating counterclockwise towards the short-circuit admittance $y_{sc} = \infty$ (located at 0.25 WTG), we find that

$$\ell = 0.3973 - .25 = 0.1473\ \lambda.$$

Another stub-matching technique that is common in microwave circuits is the ***double-stub technique***. This technique uses two stubs that are located at fixed distances from the load. Figure 11-64 shows a typical double-stub tuner. Double-stub tuners are attractive when it is necessary to quickly change the tuner characteristics as different loads are attached to the transmission line. Unlike the lone stub of a single-stub tuner, which changes its position and length as the load changes, the stubs of a double-stub tuner are at fixed positions. For this network, stub #1 is adjusted so that the normalized admittance at stub #2 is of the form $y = 1 + jb$. Then, stub #2 is adjusted so that it adds a susceptance $-jb$ at that point, yielding a matched impedance to the transmission line. The Smith chart can also be used to design double-stub tuners, but the procedure is more involved than for single-stub tuners.

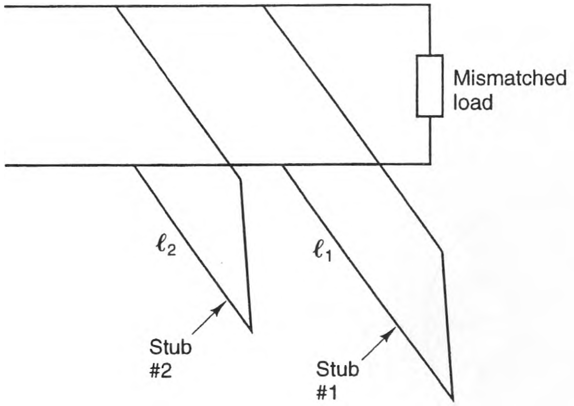


Figure 11-64 A double-stub tuner.

Interested readers can find these design procedures in many texts on microwave theory and techniques.[9]

11-5 Summation

In this chapter we have found that the voltages and currents on transmission lines can be described as waves that are related by a unique characteristic impedance and that propagate at a fixed velocity. These waves can propagate in either direction, and the behavior of a transmission line system is usually a strong function of the reflections that these waves encounter at junctions, loads, and switches.

Most of the wires, traces, or cables that occur in RF, microwave, and digital circuits can be analyzed using the techniques developed in this chapter. This is because the voltages and currents on these lines are usually TEM waves, to which transmission-line analysis applies. When the operating frequencies (or bit rates) are extremely high, however, it is sometimes necessary to use waveguide analysis to model the voltages and currents properly. These methods will be discussed in Chapter 13.

PROBLEMS

11-1 Derive the relation $\partial I/\partial z = -C\,(\partial V/\partial t)$ for lossless transmission lines by applying Ampère's law (Equation (11.4)) to the contour C_{xy} shown in Figure 11-3. (*Hint*: This derivation is similar to that of Equation (11.12).)

11-2 RG-8U coaxial cable has an inner conductor of diameter 1.83 [mm] and an outer conductor of diameter 7.24 [mm]. If the dielectric is solid polyethylene, and the losses of both the conductors and the dielectric are negligible, find

[9] For instance, see R.E. Collin, *Foundations for Microwave Engineering*, 2d ed. (New York: McGraw-Hill Book Company, 1992.)

(a) the capacitance per meter

(b) the inductance per meter at high frequencies (ignore the internal inductance)

(c) the characteristic resistance R_o

(d) the velocity of propagation u.

11-3 Figure P11-3 shows the equivalent circuit of an infinite, lossless transmission line when a single unit cell is added to the line. Since the line is still infinitely long, the addition of the cell does not affect the input impedance Z_{in}; it is still Z_o. Use this circuit to derive an expression for Z_o in terms of the capacitance and inductance elements in the cell. What must be assumed about the length of the cell in order for this expression in order to obtain $Z_o = \sqrt{L/C}$? Use phasor analysis.

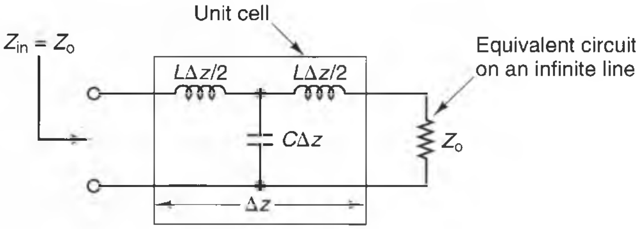


Figure P11-3

11-4 Prove that for the dispersionless case ($RC = GL$), voltage and current waves on transmission lines can be expressed by

$$V(t, z) = e^{-\alpha z}\, V^{+}(t - z/u) + e^{+\alpha z}\, V^{-}(t + z/u)$$

$$I(t, z) = \frac{1}{R_o} e^{-\alpha z}\, V^{+}(t - z/u) - \frac{1}{R_o} e^{+\alpha z}\, V^{-}(t + z/u),$$

where u, α, and R_o are given by Equations (11.43)–(11.45), respectively.

11-5 A source consisting of a 50 [V] dc battery in series with a 25 [Ω] resistor is applied to the input terminals of a $Z_o = 50$ [Ω] transmission line at $t = 0$. Assuming that the transmission line is long enough so that reflections can be ignored, what are the voltage and current at the input terminals for $t > 0$?

11-6 Use the appropriate formulas in Appendix D to calculate the characteristic impedance Z_o of a microstrip transmission line that consists of a 2.5 [mm] wide copper trace on a conductor-backed PCB board that has a dielectric constant of 5.0 and a thickness of 1 [mm].

11-7 Suppose that it is necessary to fabricate a microstrip transmission line with a 50 [Ω] characteristic impedance on a conductor-backed PCB board with a dielectric constant of 3.5 and a dielectric thickness of 1.5 [mm]. Using the appropriate formulas in Appendix D, what trace width is required to attain this impedance?

11-8 Use the appropriate formulas in Appendix D to calculate the characteristic impedance of the air dielectric, strip-line transmission line shown in Figure P11-8 when $b = 1.5$ [cm], $t = 0.1$ [cm], and

(a) $w = 0.3$ [cm]

(b) $w = 0.6$ [cm]

(c) $w = 1.2$ [cm].

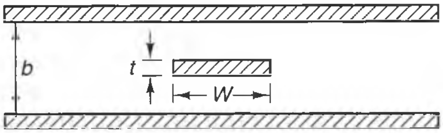


Figure P11-8

11-9 A 10 [V] step-function voltage source with a source resistance of 25 [Ω] is connected to a 1.5 [m] length of 75 [Ω], air-dielectric transmission line ($u = 3 \times 10^8$ [m/s]). If the load resistance is 100 [Ω], plot the voltage, current, and input resistance seen at the input terminals for $0 < t < 40$ [ns].

11-10 A transmission line of length ℓ can be approximated by a single unit cell (with $\Delta z = \ell$) when the rise and fall times of the signals are much longer than the propagation delay T from one end to the other. For the circuit shown in Figure 11-16, prove that the 10% to 90% rise time is given by the approximation $t_r \approx 2.2R_g C\ell$ for the case when $R_g \gg R_o$.

11-11 The transmission-line network shown in Figure P11-11 has a linear input circuit and a nonlinear load. Use a Bergeron graph to calculate and graph the load voltage for $0 < t < 600$ [ps]. Is this waveform approaching a steady-state value? If so, what value?

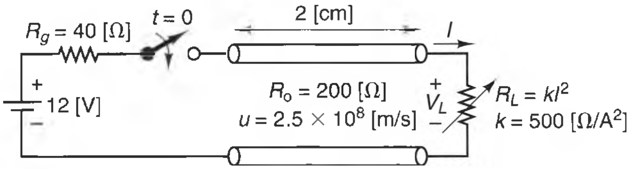


Figure P11-11

11-12 Sketch $V_L(t)$ for $0 < t < 20$ [ns] for the circuit shown in Figure P11-12 if the one-way propagation time on the transmission line is 4 [ns]. Also, what is the steady-state value? Assume that the diode is ideal.

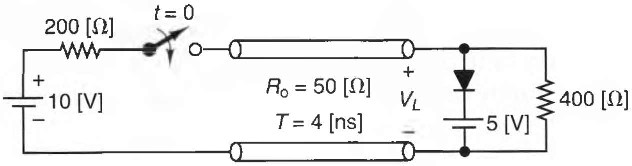


Figure P11-12

11-13 A dc source with an open-circuit voltage of 10 [mV] dc and a series impedance of 50 [Ω] is attached to the input terminals of a lossless transmission line at $t = 0$. The transmission line has a characteristic impedance of 50 [Ω], is 2 [m] long, has a propagation velocity of $u = 0.6455\,c$, and is terminated by a 150 [Ω] resistor in series with a 20 [pF] capacitor. Sketch the reflected waveform at the input terminals of the transmission line.

11-14 Repeat Problem 11-13 for the case where the capacitor is replaced by a 1.2 [μH] inductor.

11-15 Figure P11-15a shows two networks connected by a lossless transmission line characteristic impedance Z_o and one-way propagation delay T. Prove that the output voltages and currents of each network are unchanged when the transmission line is replaced with the Thévenin equivalent circuits shown in Figure P11-15b, where

$$V_a(t - T) = V(\ell, t - T) - Z_o I(\ell, t - T)$$

$$V_b(t - T) = V(0, t - T) + Z_o I(0, t - T).$$

These equivalent circuits are used by the circuit analysis software PSPICE™ to model transmission lines. (*Hint*: Write KVL expressions around the equivalent circuits and use Equations (11.22) and (11.32) to express total voltages and currents in terms of traveling components.)

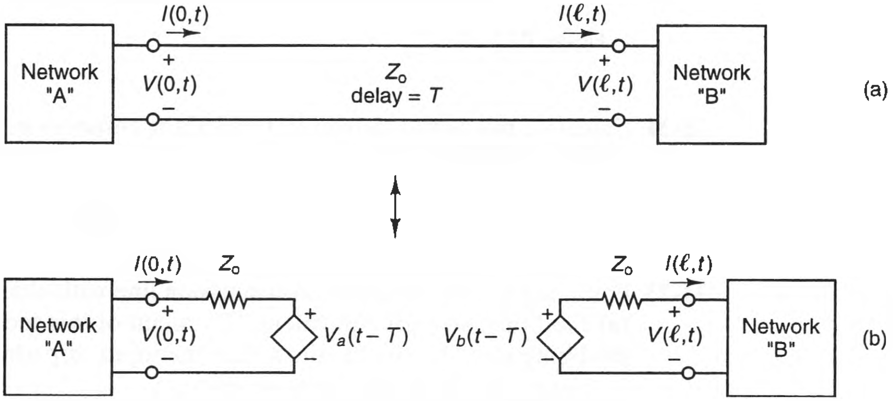


Figure P11-15 a) Original network. b) Equivalent circuits.

11-16 Suppose that the dielectric constant of a coaxial transmission line over a band of frequencies varies as $\sqrt{\epsilon_r} = a + b\omega$, where a and b are constants. Find the expressions for the phase and group velocities on this line as a function of frequency.

11-17 For the twin-lead transmission line discussed in Example 11-7, calculate the attenuation constant α at 100 [MHz] if the wires are made of copper ($\sigma = 5.8 \times 10^7$ [S/m]) and the dielectric is lossless.

11-18 A sinusoidal generator with a 20 [V] open-circuit voltage and an impedance of 75 [Ω] is attached to a lossless 75 [Ω] transmission line that is terminated in an unknown load. If measurements show that the load reflection coefficient Γ_L has a value of 0.45 ∠ 23°, find (a) the load impedance and (b) the voltage amplitude at the load.

11-19 What is the shortest length of a lossless, open-circuited transmission line that can be used to simulate a 20 [pF] capacitor at 400 [MHz]. Assume that the characteristic impedance of the line is 50 [Ω] and the dielectric constant is $\epsilon_r =$ 2.5.

11-20 A lossy transmission line with $R_o = 50$ [Ω] and $X_o \approx 0$ has a loss of 0.5 [dB/m] at 200 [MHz]. If the dielectric constant is $\epsilon_r = 2.0$, find the input impedance of a quarter-wavelength section when the load is a short circuit. What would the input impedance be if the transmission line was lossless?

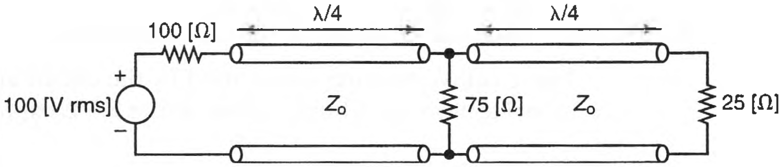


Figure P11-21

11-21 Calculate the power delivered to all three resistors in Figure P11-21. Assume that $Z_o = 50$ [Ω].

11-22 Prove that the frequency-domain equivalent "T" circuit of a finite-length, lossless transmission line (shown in Figure 11-35b) is equivalent to the unit cell (shown in Figure 11-4) when the length of the line is small.

11-23 For a λ/4 section of lossless transmission line with characteristic impedance Z_o,

(a) Find the two-port equivalent "T" circuit of this section.

(b) Using the "T" circuit, show that the input impedance Z_{in} is zero when the output terminals are open circuited.

(c) Using the "T" circuit, show that $Z_{in} = \infty$ when the output terminals are short circuited.

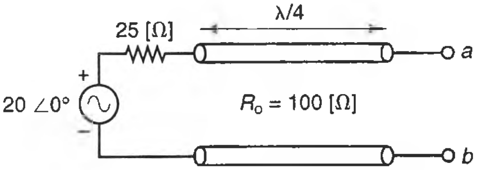


Figure P11-24

11-24 Find the Thévenin equivalent circuit of the network shown in Figure P11-24 with respect to the output terminal pair *ab*.

11-25 Suppose that measurements on a 50 [Ω] transmission line with an unknown load show that VSWR = 1.85 and that two successive voltage minima occur at z = 2.43 [cm] and z = 17.47 [cm] on a scale that increases toward the load. If a voltage minimum occurs at z = 6.28 [cm] when the load is replaced by a short circuit, find the impedance of the load.

11-26 Supose that when using a network analyzer, a person finds the S-parameters of a two-port network to be

$$S_{11} = 0.1 \angle -30°$$
$$S_{12} = S_{21} = 0.6 \angle -15°$$
$$S_{22} = 0.3 \angle 70°.$$

Find the impedance matrix of this network and draw its equivalent "T" circuit if the characteristic impedance of each port is 50 [Ω].

11-27 Using the definitions of the scattering matrix elements (Equation (11.168)), determine the S-matrix of a section of uniform, lossless transmission line with length ℓ and characteristic impedance Z_o. From this matrix, use the S- to Z-matrix conversion (Equation (11.172)) to derive the Z-matrix elements. Does this result agree with the Z-matrix elements given by Equation (11.147)?

11-28 Using the Smith chart, determine the impedance of the load attached to a $Z_o = 75$ [Ω] transmission line if it is known that the input impedance 2 [cm] away from the load is $Z_{in} = 75 + j20$ [Ω] at a frequency where the wavelength on the line is 6 [cm].

11-29 A load of value $Z_L = 125 - j60$ [Ω] is attached to a transmission line with $Z_o = 50$ [Ω]. If the wavelength on the line is 12.5 [cm], use the Smith chart to find

(a) the input impedance 4.5 [cm] away from the load

(b) the VSWR

(c) the closest location from the load where the input impedance seen looking towards the load is real and greater than Z_o.

11-30 A lossy transmission line with $Z_o \approx 50$ [Ω] and $\alpha = 1.5$ [dB/m] is terminated with a load impedance of $Z_L = 100$ [Ω]. If $\lambda = 2.0$ [m], use the Smith chart to find the input impedance 1.0 [m] away from the load.

11-31 Design a single-stub tuner that matches a $Z_L = 40 - j\,60$ [Ω] load to a 50 [Ω] uniform-dielectric transmission line using the shortest possible open-circuited stub that has the same characteristic impedance as the transmission line. Specify all the critical dimensions if the dielectric constant is $\epsilon_r = 2.8$ on both the stub and the transmission line and $f = 4$ [GHz].

11-32 Design a microstrip, quarter-wave transformer that matches a 10 [Ω] load to a 50 [Ω] microstrip transmission line. Assume that the substrate is 1 [mm] thick, the dielectric constant is 3.5, and the frequency is 12.5 [GHz]. Show all the critical dimensions of this design.

11-33 Design a microstrip, single-stub tuner that matches a 50 [Ω] microstrip transmission line to load impedance of $Z_L = 70 + j\,10$ [Ω]. Assume that the frequency is 10 [GHz], the dielectric constant is 2.5, and the substrate thickness is 1.2 [mm]. Assume that the stub is open circuited and has the same characteristic impedance as the transmission line. Show all the critical dimensions of this design (including the 50 [Ω] line).

11-34 Prove that the characteristic impedance of a section of uniform transmission line (lossy or lossless) can be written as

$$Z_o = \sqrt{Z_{sc} Z_{oc}},$$

where Z_{sc} and Z_{oc} are the input impedances when the output terminals are short circuited and open circuited, respectively.

12

Plane Waves

12-1 Introduction

One of the most useful features of electromagnetics is that electromagnetic waves can travel through space, without the need for a guiding structure. These waves are called ***space waves***, because they can propagate through empty space (i.e., a vacuum), which we commonly call free space. Space waves can also propagate in nearly any kind of medium. A large number of applications make use of these waves, including wireless communication systems and radars.

The simplest kind of space waves that can be produced by a source are called ***plane waves***, so named because their constant-amplitude and constant-phase surfaces are flat sheets (planes). Not only are these waves the simplest space waves; they are also excellent approximations of the waves most commonly encountered in engineering practice. This is because nearly all space waves behave like plane waves after they propagate just a few wavelengths from their source.

Plane waves share many common characteristics with the TEM waves found on transmission lines. This should come as no surprise, since both are propagating waves. However, since plane waves are not confined to a guiding structure, there is more variety in the kinds of plane waves that can be produced and the way in which they interact with materials. In this chapter, we will discuss the various properties of plane waves, including wavelength, attenuation, polarization, reflection, and refraction. These topics will be discussed under the assumption that the waves have already been launched by unspecified sources. The specific nature of the sources of the waves will be dealt with in Chapter 14.

12-2 Wave Equations in Simple, Source-Free Media

In Section 10-4-4 we showed that in simple, source-free media, Maxwell's equations become

$$\nabla \times \mathbf{E} = -j\omega\mu\,\mathbf{H} \tag{12.1}$$

$$\nabla \times \mathbf{H} = (\sigma + j\omega\epsilon)\mathbf{E} \tag{12.2}$$

$$\nabla \cdot \mathbf{E} = 0 \tag{12.3}$$

$$\nabla \cdot \mathbf{H} = 0. \tag{12.4}$$

Any field distribution that can exist in a simple, source-free medium must satisfy these equations. Nevertheless, the equations are not the best starting point for our development of plane waves, since they are coupled differential equations that each contain both **E** and **H**. In this section, we will derive a single equation that either **E** or **H** alone must satisfy at every point in the medium.

If we take the curl of both sides of Equation (12.1), we obtain

$$\nabla \times \nabla \times \mathbf{E} = -j\omega\mu\nabla \times \mathbf{H}.$$

Substituting Equation (12.2) into this expression yields

$$\nabla \times \nabla \times \mathbf{E} = -j\omega\mu(\sigma + j\omega\epsilon)\mathbf{E}.$$

Using Equation (B.10), we can write $\nabla \times \nabla \times \mathbf{E}$ in terms of the divergence and Laplacian of **E**:

$$\nabla \times \nabla \times \mathbf{E} = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2\mathbf{E} = -j\omega\mu(\sigma + j\omega\epsilon)\mathbf{E}.$$

Finally, remembering that $\nabla \cdot \mathbf{E} = 0$ in simple, source-free media (i.e., Equation (12.3)), we obtain

$$\nabla^2\mathbf{E} + k^2\mathbf{E} = 0, \tag{12.5}$$

where

$$k^2 = -j\omega\mu\,(\sigma + j\omega\epsilon). \tag{12.6}$$

Equation (12.5) is called the ***vector wave equation***, or the ***vector Helmholtz equation***, and the constant k is called the ***wave number*** of the medium.

When **E** is expressed in terms of its Cartesian components, we can use Equation (2.126) to split the vector wave into three scalar equations,

$$\nabla^2 E_i + k^2 E_i = 0 \quad i = x, y, z, \tag{12.7}$$

where E_x, E_y, and E_z are the x-, y-, and z-components of **E**. Equation (12.7) is called the ***scalar wave equation*** or the ***scalar Helmholtz equation***.

The magnetic field **H** within a simple, source-free region also satisfies the vector and scalar wave equations at each point. This can be shown by taking the curl of Equation (12.2) and proceeding with a similar sequence of steps, yielding

$$\nabla^2 \mathbf{H} + k^2 \mathbf{H} = 0 \tag{12.8}$$

and

$$\nabla^2 H_i + k^2 H_i = 0 \quad \mathrm{i} = x, y, z. \tag{12.9}$$

In the sections that follow, we will use these wave equations to find expressions for the simplest (and most important) space waves: plane waves.

12-3 Plane Waves in Lossless Media

In Cartesian coordinates, the scalar wave equation can be written as

$$\frac{\partial^2 E_i}{\partial x^2} + \frac{\partial^2 E_i}{\partial y^2} + \frac{\partial^2 E_i}{\partial z^2} + k^2 E_i = 0 \quad i = x, y, z. \tag{12.10}$$

When the medium is lossless, σ is zero, and μ and ϵ are both real. According to Equation (12.6), this means that the wave number k is a positive, real number:

$$k = \omega\sqrt{\mu\epsilon} \qquad \text{(Lossless media).} \tag{12.11}$$

Rather than considering the most general case, we will ease ourselves into the subject by first restricting ourselves to plane waves with one E-field component that propagates along the z-axis only. To accomplish this, we will search for E-field solutions of the wave equation that are of the form

$$\mathbf{E} = E_x(z)\hat{\mathbf{a}}_x. \tag{12.12}$$

Substituting this expression into Equation (12.10), we obtain

$$\frac{d^2 E_x}{dz^2} + k^2 E_x = 0.$$

This is a second-order, linear, homogeneous differential equation, so it has two independent solutions, which can be written in the form

$$E_x(z) = E_{xo}^{+} e^{-\gamma z} + E_{xo}^{-} e^{+\gamma z}, \tag{12.13}$$

where E_{xo}^{+} and E_{xo}^{-} are constants (possibly complex) and

$$\gamma = jk = j\omega\sqrt{\mu\epsilon}. \tag{12.14}$$

is the ***propagation constant***.

It is customary to express the propagation constant γ as the sum of a real and imaginary part; that is,

$$\gamma = \alpha + j\beta, \tag{12.15}$$

where α and β are the ***attenuation*** and ***phase constants***, respectively. However, since k is real in lossless media, we have

$$\left.\begin{array}{l} \alpha = 0 \\ \beta = k = \omega\sqrt{\mu\epsilon} \end{array}\right\} \quad \text{(Lossless media)}. \qquad (12.16)\ (12.17)$$

Since α is zero in lossless media, we can write $E_x(z)$ in the form

$$E_x(z) = E_{xo}^{+}e^{-j\beta z} + E_{xo}^{-}e^{+j\beta z}. \tag{12.18}$$

Finally, substituting Equation (12.18) into Equation (12.12), we obtain the complete frequency-domain expression for **E**:

$$\mathbf{E} = E_{xo}^{+}e^{-j\beta z}\,\hat{\mathbf{a}}_x + E_{xo}^{-}e^{+j\beta z}\,\hat{\mathbf{a}}_x. \tag{12.19}$$

Equation (12.19) should look familiar to the reader, since, except for the unit vectors, this expression is the same as the voltage-wave expressions on transmission lines. (See Equation (11.89)). The propagating nature of these waves can be seen more clearly by transforming this expression into the time domain. If we let $E_{xo}^{+} = |E_{xo}^{+}|\angle\theta^{+}$ and $E_{xo}^{-} = |E_{xo}^{-}|\angle\theta^{-}$, the corresponding time-domain expression for **E** is

$$\mathbf{E} = |E_{xo}^{+}|\cos(\omega t - \beta z + \theta^{+})\hat{\mathbf{a}}_x + |E_{xo}^{-}|\cos(\omega t + \beta z + \theta^{-})\hat{\mathbf{a}}_x. \tag{12.20}$$

Here, it is clear that this E-field is the sum of two waves, one with peak amplitude $|E_{xo}^{+}|$ that propagates towards increasing values of z and another with peak amplitude $|E_{xo}^{-}|$ that propagates towards decreasing values of z. Also, using Equations (12.17), (11.96), and (11.97), the phase velocity u_p and wavelength λ of these waves are given by

$$u_p = \frac{\omega}{\beta} = \frac{1}{\sqrt{\mu\epsilon}} \qquad \text{[m/sec]} \tag{12.21}$$

$$\lambda = \frac{2\pi}{\beta} = \frac{\omega}{u_{\rm p}} \qquad \text{[m]}. \tag{12.22}$$

For the case where the medium is a vacuum (free space), $\epsilon = \epsilon_{\rm o}$, $\mu = \mu_{\rm o}$, and the phase velocity equals the speed of light in a vacuum:

$$u_p = 3.0 \times 10^8\,\text{[m/s]} \equiv c \qquad \text{(Vacuum speed of light)}. \tag{12.23}$$

Just as every voltage wave on a transmission line is accompanied by a current wave, every electric field of a plane wave is accompanied by a magnetic field. We can find the H-field associated with the preceding E-field by first solving Equation (12.1) for **H**:

$$\mathbf{H} = \frac{\nabla \times \mathbf{E}}{-j\omega\mu}.$$

Substituting Equation (12.19) into this expression and performing the curl operation, we obtain

$$\mathbf{H} = \frac{-1}{j\omega\mu}\frac{\partial}{\partial z}[E_{xo}^{+}e^{-j\beta z} + E_{xo}^{-}e^{+j\beta z}]\,\hat{\mathbf{a}}_y$$

$$= \frac{\beta}{\omega\mu}[E_{xo}^{+}e^{-j\beta z} - E_{xo}^{-}e^{+j\beta z}]\,\hat{\mathbf{a}}_y.$$

This can be written in the form

$$\mathbf{H} = \frac{E_{xo}^{+}}{\eta}e^{-j\beta z}\,\hat{\mathbf{a}}_y - \frac{E_{xo}^{-}}{\eta}e^{+j\beta z}\,\hat{\mathbf{a}}_y, \tag{12.24}$$

where η is the ***intrinsic*** (or ***wave***) ***impedance*** of the medium and is given by

$$\eta = \frac{\omega\mu}{\beta} = \sqrt{\frac{\mu}{\epsilon}} \quad [\Omega] \qquad \text{(Lossless media)}. \tag{12.25}$$

In a vacuum (free space), the value of η is

$$\eta_o \equiv \sqrt{\frac{\mu_o}{\epsilon_o}} \approx 377\ [\Omega] \approx 120\pi \qquad \text{(Free space)}. \tag{12.26}$$

Finally, since η is real in lossless media, **H** has the following form in the time domain:

$$\mathbf{H} = \frac{|E_{xo}^{+}|}{\eta}\cos(\omega t - \beta z + \theta^{+})\hat{\mathbf{a}}_y - \frac{|E_{xo}^{-}|}{\eta}\cos(\omega t + \beta z + \theta^{-})\hat{\mathbf{a}}_y. \tag{12.27}$$

Comparing Equations (12.19) and (12.27), we see that the forward-propagating E- and H-waves in lossless media have identical phases and a fixed amplitude ratio. The same is true for the backward-propagating waves, except for the 180° phase shift of the H-field. For both the forward- and backward-propagating waves, a right-handed coordinate system is formed by the E- and H-field vectors and the direction of propagation. Figure 12-1 depicts these vectors for a forward-propagating wave. Here, we see that the E-field vector, crossed into the H-field vector, points in the direction of propagation.

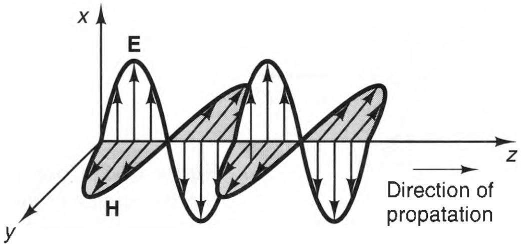


Figure 12-1 E- and H-field vectors at an instant in time for a plane wave propagating in the $+z$ direction.

12-3-1 PLANE WAVE PROPAGATION IN ARBITRARY DIRECTIONS

So far we have only considered simple plane waves that propagate along the z-axis. This has been a convenient starting point, but there is nothing magic about the z-direction; plane waves can be launched in any direction through a medium. In this section, we will use what we have already discovered about waves that propagate along the z-axis to describe plane waves propagating in any direction.

Let us start by noting that when we developed the $+z$ propagating waves in the previous section, the orientations of the x- and y-axes were arbitrary, except that they formed a right-handed coordinate system with the z-axis. Remembering that **E**, **H**, and the direction of propagation form a right-handed coordinate system, it follows that the electric and magnetic field vectors of a $+z$ propagating wave can have any orientation in the xy-plane, as long as the cross product **E X H** points in the $+z$ direction. Thus, the general expressions for $+z$ propagating plane waves can be written in the form

$$\mathbf{E} = [E_{xo}\hat{\mathbf{a}}_x + E_{yo}\hat{\mathbf{a}}_y]e^{-j\beta z} \tag{12.28}$$

$$\mathbf{H} = \left[\frac{E_{xo}}{\eta}\hat{\mathbf{a}}_y - \frac{E_{yo}}{\eta}\hat{\mathbf{a}}_x\right]e^{-j\beta z}, \tag{12.29}$$

where E_{xo} and E_{yo} are the complex amplitudes of the x and y electric field components, respectively, and β equals the wave number k when the medium is lossless. Although we have derived these two expressions in a somewhat nonrigorous way, it is simple to prove that they satisfy the vector wave equations and Maxwell's equations.

We can use Equations (12.28) and (12.29) to write expressions for plane waves propagating in any direction. To accomplish this, consider the situation shown in Figure 12-2, which depicts the E- and H-field vectors of a plane wave that propagates in the direction of the unit vector $\hat{\mathbf{a}}_k$. Two orthogonal coordinate systems are shown in this figure: an unprimed system in which the $+z$ direction coincides with the direction of propagation of the wave and a primed system whose axis directions are arbitrary. Using the figure, we first note that the product βz that appears in Equations (12.28) and (12.29) can be expressed at any point as the dot product of two vectors,

$$\beta z = kz = k\hat{\mathbf{a}}_z \cdot (x\hat{\mathbf{a}}_x + y\hat{\mathbf{a}}_y + z\hat{\mathbf{a}}_z) = \mathbf{k}\cdot\mathbf{r}, \tag{12.30}$$

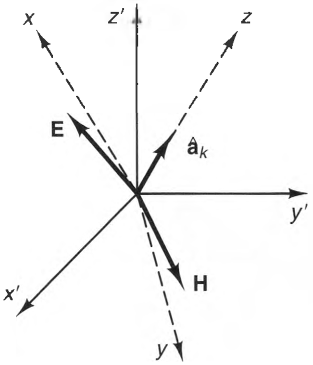


Figure 12-2 Coordinate systems for representing a plane wave propagating in an arbitrary direction.

where $\mathbf{r} = x\hat{\mathbf{a}}_x + y\hat{\mathbf{a}}_y + z\hat{\mathbf{a}}_z$ is the position vector of an arbitrary point (see Equation (2.64a)) and $\mathbf{k} = k\hat{\mathbf{a}}_z$ is a vector, called the ***wave-number vector***, that points in the direction of propagation and whose magnitude equals the wave-number k.

Substituting Equation (12.30) into Equation (12.28), we obtain

$$\mathbf{E} = [E_{xo}\hat{\mathbf{a}}_x + E_{yo}\hat{\mathbf{a}}_y]e^{-j\mathbf{k}\cdot\mathbf{r}},$$

or

$$\mathbf{E} = \mathbf{E}_o e^{-j\mathbf{k}\cdot\mathbf{r}}, \tag{12.31}$$

where $\mathbf{E}_o$, called the ***polarization vector***, can be any vector that is perpendicular to $\mathbf{k}$; that is

$$\mathbf{E}_o \cdot \mathbf{k} = 0. \tag{12.32}$$

We can derive a companion expression for the H-field of this plane wave by noting that Equation (12.29) can be written in terms of the cross product between $\mathbf{E}_o$ and $\mathbf{k}$;

$$\mathbf{H} = \frac{E_{xo}}{\eta}e^{-j\beta z}\hat{\mathbf{a}}_y - \frac{E_{yo}}{\eta}e^{-j\beta z}\hat{\mathbf{a}}_x = \frac{e^{-j\beta z}}{\eta}\hat{\mathbf{a}}_z \times (E_{xo}\hat{\mathbf{a}}_x + E_{yo}\hat{\mathbf{a}}_y)$$

$$= \frac{e^{-j\beta z}}{\eta}\hat{\mathbf{a}}_z \times \mathbf{E}_o = \frac{e^{-j\beta z}}{\eta}\frac{\mathbf{k}}{k} \times \mathbf{E}_o.$$

Remembering that $\beta z = \mathbf{k}\cdot\mathbf{r}$ and $k = \omega\sqrt{\mu\epsilon}$, we can write $\mathbf{H}$ in the form

$$\mathbf{H} = \frac{1}{\omega\mu}(\mathbf{k} \times \mathbf{E}_o)e^{-j\mathbf{k}\cdot\mathbf{r}}. \tag{12.33}$$

Even though Equations (12.32) and (12.33) were derived for the case of a plane wave propagating in the $+z$ direction, they are the same in any coordinate system, since they contain only dot and cross products of the vectors $\mathbf{k}$, $\mathbf{E}_o$, and $\mathbf{r}$. As a result, these equations can describe a plane wave propagating in *any* direction simply by choosing the direction of the wave-number vector $\mathbf{k}$ such that it points in the desired propagation direction; that is,

$$\mathbf{k} = k\hat{\mathbf{a}}_k, \tag{12.34}$$

where $\hat{\mathbf{a}}_k$ is the direction of propagation.

Taken as a set, Equations (12.32), (12.33), and (12.34) can describe any plane wave that propagates in any direction; thus,

$$\mathbf{E} = \mathbf{E}_o e^{-j\mathbf{k}\cdot\mathbf{r}}, \tag{12.35}$$

$$\mathbf{H} = \frac{1}{\omega\mu}(\mathbf{k} \times \mathbf{E}_o)e^{-j\mathbf{k}\cdot\mathbf{r}}, \tag{12.36}$$

$$\mathbf{E}_o \cdot \mathbf{k} = \mathbf{E}_o \cdot k\hat{\mathbf{a}}_k = 0. \tag{12.37}$$

These expressions are valid for any direction of propagation and in any coordinate system. All that is necessary to evaluate these expressions is to find appropriate expressions for the direction of propagation $\hat{\mathbf{a}}_k$, the polarization vector $\mathbf{E}_o$, and the position vector $\mathbf{r}$ in the chosen coordinate system. From Equations (12.35)–(12.37), we can list the following general characteristics of all plane waves:

1. Both **E** and **H** propagate in a direction parallel to the wave-number vector **k**.
2. **E** and **H** are perpendicular to each other.
3. The direction of $\mathbf{E} \times \mathbf{H}$ points in the same direction as the wave-number vector **k**.
4. The ratio of the magnitudes of **E** and **H** equals $(\omega\mu)/k = \eta$, the intrinsic impedance of the medium.

Example 12-1

Find the expression for the plane wave that propagates parallel to the *xy*-plane in the direction indicted in Figure 12-3. Assume that **E** has only a *z*-component.

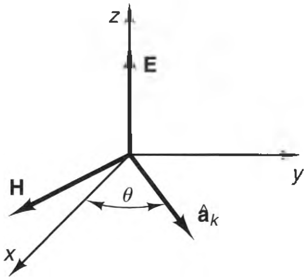


Figure 12-3 A plane wave propagating parallel to the *xy*-plane at an angle θ with respect to the *x*-axis.

Solution:

Since this wave propagates at an angle θ with respect to the *x*-axis, we can express the wave-number vector **k** as

$$\mathbf{k} = k(\cos\theta\,\hat{\mathbf{a}}_x + \sin\theta\,\hat{\mathbf{a}}_y).$$

As a result,

$$\mathbf{k}\cdot\mathbf{r} = k(x\cos\theta + y\sin\theta).$$

Since **E** has only a *z*-component, we also can write

$$\mathbf{E}_o = E_o\hat{\mathbf{a}}_z.$$

Substituting these expressions into Equation (12.35), we obtain

$$\mathbf{E} = E_o\hat{\mathbf{a}}_z e^{-jk(x\cos\theta + y\sin\theta)}.$$

To find **H**, we first evaluate $\mathbf{k} \times \mathbf{E}_o$:

$$\mathbf{k} \times \mathbf{E}_o = k(\cos\theta\,\hat{\mathbf{a}}_x + \sin\theta\,\hat{\mathbf{a}}_y) \times E_o\hat{\mathbf{a}}_z$$
$$= kE_o(\sin\theta\,\hat{\mathbf{a}}_x - \cos\theta\,\hat{\mathbf{a}}_y).$$

Finally, using $(\omega\mu)/k = \eta$, we have, from Equation (12.36),

$$\mathbf{H} = \frac{E_o}{\eta}(\sin\theta\,\hat{\mathbf{a}}_x - \cos\theta\,\hat{\mathbf{a}}_y)e^{-jk(x\cos\theta + y\sin\theta)}.$$

12-3-2 POLARIZATION

The polarization of a plane wave is a measure of how its E-field vector varies with time. To simplify our discussion (without loss of generality), we will consider plane waves that propagate in the $+z$ direction. A general wave of this type can be represented as

$$\mathbf{E} = (|E_{xo}|e^{j\theta_x}\hat{\mathbf{a}}_x + |E_{yo}|e^{j\theta_y}\hat{\mathbf{a}}_y)e^{-j\beta z}, \tag{12.38}$$

where $|E_{xo}|$ and $|E_{yo}|$ are the peak amplitudes of the x- and y-components of $\mathbf{E}$, respectively, and θ_x and θ_y are the phases of these components at $z = 0$, respectively. Transforming this expression to the time domain, we obtain

$$\mathbf{E} = E_{xo}\cos(\omega t - \beta z + \theta_x)\hat{\mathbf{a}}_x + E_{yo}\cos(\omega t - \beta z + \theta_y)\hat{\mathbf{a}}_y. \tag{12.39}$$

In the paragraphs that follow, we will use this expression to describe three classes of polarization: linear polarization, circular polarization, and elliptical polarization.

Linear Polarization. Linear polarization occurs when $\theta_x = \theta_y \equiv \theta$. For this case, Equation (12.39) becomes

$$\mathbf{E} = (|E_{xo}|\hat{\mathbf{a}}_x + |E_{yo}|\hat{\mathbf{a}}_y)\cos(\omega t - \beta z + \theta). \tag{12.40}$$

As can be seen from this expression, the x- and y-components of the field maintain the same ratio for all values of t, which means that $\mathbf{E}$ always lies along a straight line in any constant-z plane. The ***tilt angle*** between $\mathbf{E}$ and the x-axis is

$$\tau = \tan^{-1}\frac{|E_{yo}|}{|E_{xo}|}. \tag{12.41}$$

Figure 12-4 shows $\mathbf{E}$ at several points in time at $z = 0$.

Many practical sources generate linearly polarized plane waves. Lasers, for instance, are often constructed so that their outputs are linearly polarized. Many simple antennas also generate waves that, when viewed at large distances, behave as linearly polarized plane waves. The most common example is the dipole antenna, which is discussed in depth in Chapter 14.

Circular Polarization. Circular polarization occurs when the orthogonal components of $\mathbf{E}$ have equal magnitudes, but differ in phase by ±90°. For simplicity, we will assume that $|E_{xo}| = |E_{yo}| \equiv E_o$, $\theta_x = 0$, and $\theta_y = \pm 90°$. For this case, the y-compo-

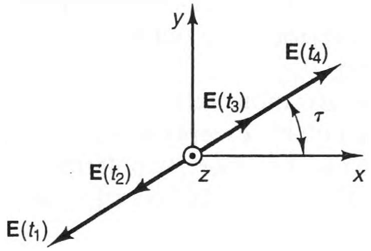


Figure 12-4 The E-field vector of a linearly polarized, $+z$ propagating plane wave as a function of time at a fixed position.

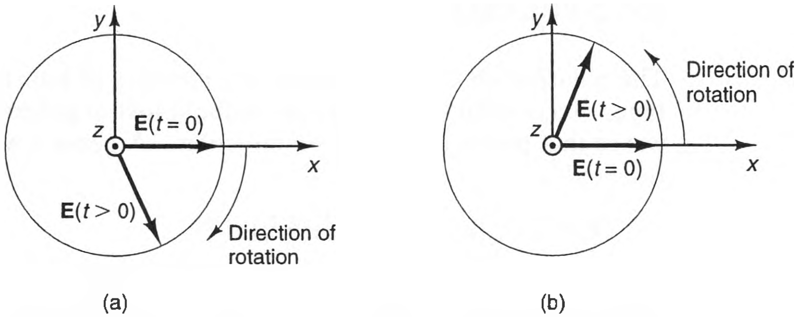


Figure 12-5 E-field rotation of $+z$ propagating, circularly polarized plane waves: a) Left hand polarization. b) Right-hand polarization.

nent of **E** either leads or lags the x-component by 90°, so **E** can be represented in the time and frequency domains as

$$\mathbf{E} = E_o(\hat{\mathbf{a}}_x \pm j\hat{\mathbf{a}}_y)e^{-j\beta z}, \tag{12.42}$$

and

$$\mathbf{E} = E_o(\cos(\omega t - \beta z)\hat{\mathbf{a}}_x \mp \sin(\omega t - \beta z)\hat{\mathbf{a}}_y), \tag{12.43}$$

respectively.

Figure 12-5a shows a plot of **E** for several values of t at $z = 0$ when E_y leads E_x by 90° (i.e., the upper sign in Equation (12.43)). As can be seen, the magnitude of **E** remains constant, but its direction rotates around the z axis once every $2\pi/\omega$ seconds. Since the tip of **E** rotates around the direction of propagation (the z-axis) in a left-handed sense, this is called ***left-hand polarization*** (LHP).

Figure 12-5b shows how **E** varies with time for the case when $\theta_y = -90°$. This case corresponds to the lower sign in Equation (12.43). Here, the behavior is basically the same, except that **E** rotates around the direction of propagation in a right-handed sense. Hence, this is called ***right-hand polarization*** (RHP).

Not only do the E-fields of a circularly polarized wave rotate around the axis of propagation as a function of time; they do the same thing in space. Figure 12-6a shows a "snapshot" of a left-hand, circularly polarized wave. Here, we see that **E** follows a right-handed helix along the direction of propagation. This means that observers at different points along the direction of propagation detect different directions of **E** at the same time t. Similarly, Figure 12-6b shows **E** for a right-hand, circularly polarized wave. In this case, the vector traces out a left-handed helix along the axis of propagation when time is frozen.

Unlike linear polarization, relatively few simple sources generate circularly polarized waves. One notable exception is the helical antenna. Circularly polarized waves can be launched by two linearly polarized sources (such as dipole antennas) when they are oriented perpendicular to each other and fed with currents that are out of phase by 90°.

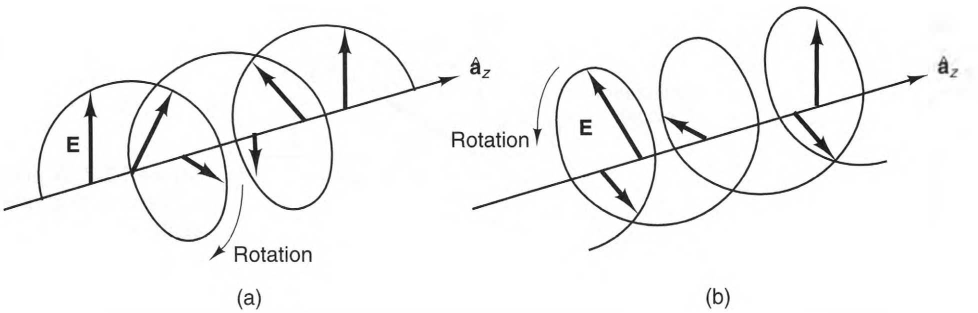


Figure 12-6 The helical paths traced by circularly polarized waves at fixed instants in time: a) Left-hand polarization. b) Right-hand polarization.

Elliptical Polarization. Linear and circular polarization occur when the two orthogonal components of a plane wave have very specific magnitude and phase relationships. We will now discuss a more general case where these relationships are independent of time, but are otherwise arbitrary.[1] Returning to Equation (12.39), let us assume that the x- and y-components of the wave have peak magnitudes $|E_{xo}|$ and $|E_{yo}|$, respectively, and the y component leads the x-component by an angle $\Delta\theta$. Thus, at $z = 0$,

$$\mathbf{E} = E_x \hat{\mathbf{a}}_x + E_y \hat{\mathbf{a}}_y, \tag{12.44}$$

where the x- and y-components of $\mathbf{E}$ are given by

$$E_x = |E_{xo}| \cos(\omega t) \tag{12.45}$$

and

$$E_y = |E_{yo}| \cos(\omega t + \Delta\theta), \tag{12.46}$$

respectively. Since E_x and E_y maintain a fixed relationship in time, it is possible to write one component in terms of the other. To accomplish this, we first solve Equation (12.45) for $\cos(\omega t)$:

$$\cos(\omega t) = \frac{E_x}{|E_{xo}|}. \tag{12.47}$$

Similarly, if we solve Equation (12.46) for $E_y/|E_{yo}|$ and use the cosine reduction formula, we find that

$$\begin{aligned}\frac{E_y}{|E_{yo}|} &= \cos(\omega t + \Delta\theta) = \cos(\omega t)\cos(\Delta\theta) - \sin(\omega t)\sin(\Delta\theta) \\ &= \frac{E_x}{|E_{xo}|}\cos(\Delta\theta) \pm \sqrt{1 - \left(\frac{E_x}{|E_{xo}|}\right)^2}\sin(\Delta\theta),\end{aligned} \tag{12.48}$$

where we have used Equation (12.47) to write $\sin(\omega t)$ in terms of E_x and $|E_{xo}|$. Finally, solving for the square-root term, squaring both sides of the resulting expression, and simplifying yields

[1] A wave is said to be ***unpolarized*** when these amplitude and phase relationships between the orthogonal components vary with time.

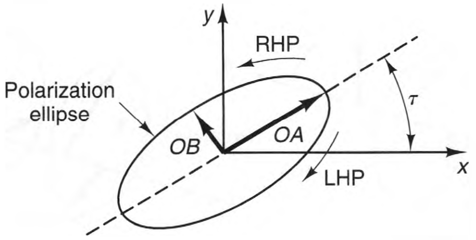


Figure 12-7 A polarization ellipse with major axis OA, minor axis OB, and tilt angle τ.

$$\left(\frac{E_x}{|E_{xo}|}\right)^2 - \frac{2E_xE_y\cos(\Delta\theta)}{|E_{xo}|\,|E_{yo}|} + \left(\frac{E_y}{|E_{yo}|}\right)^2 = \sin^2(\Delta\theta), \tag{12.49}$$

which is the equation of an ellipse, called the ***polarization ellipse***.

Figure 12-7 shows the polarization ellipse of an elliptically polarized plane wave. Right-handed polarization (RHP) occurs when $\Delta\theta < 0$, and left-handed polarization (LHP) occurs when $\Delta\theta > 0$. Two parameters that are commonly used to describe the polarization ellipse are the ***axial ratio*** AR and the ***tilt angle*** τ. The axial ratio is defined as

$$AR = \frac{\text{major axis}}{\text{minor axis}} = \frac{OA}{OB} \qquad 1 \leqslant AR \leqslant \infty, \tag{12.50}$$

where, after a great deal of algebraic manipulation, it can be shown from Equation (12.49) that

$$OA = \left[\frac{1}{2}\left\{|E_{xo}|^2 + |E_{yo}|^2 + [|E_{xo}|^4 + |E_{yo}|^4 + 2|E_{xo}|^2|E_{yo}|^2\cos(2\Delta\theta)]^{1/2}\right\}\right]^{1/2}, \tag{12.51}$$

$$OB = \left[\frac{1}{2}\left\{|E_{xo}|^2 + |E_{yo}|^2 - [|E_{xo}|^4 + |E_{yo}|^4 + 2|E_{xo}|^2|E_{yo}|^2\cos(2\Delta\theta)]^{1/2}\right\}\right]^{1/2}. \tag{12.52}$$

The tilt angle is the angle between the major axis of the ellipse and the x-axis, and is given by

$$\tau = \frac{1}{2}\tan^{-1}\left[\frac{2|E_{xo}|\,|E_{yo}|}{|E_{xo}|^2 - |E_{yo}|^2}\cos(\Delta\theta)\right]. \tag{12.53}$$

Figure 12-8 shows the polarization states for a $+z$ propagating plane wave as the relative amplitudes and phases of the x and y components of **E** are varied. In this diagram, the horizontal and vertical axes are real and imaginary parts of the ratio E_y/E_x, respectively, where E_y and E_x are the complex amplitudes of the x- and y-components of **E**, respectively. As can be seen, linear polarization (AR $= \infty$) occurs when $\text{Im}(E_y/E_x) = 0$, and circular polarization (AR $= 1$) occurs when $\text{Re}(E_y/E_x) = 0$ and $\text{Im}(E_y/E_x) = \pm 1$.

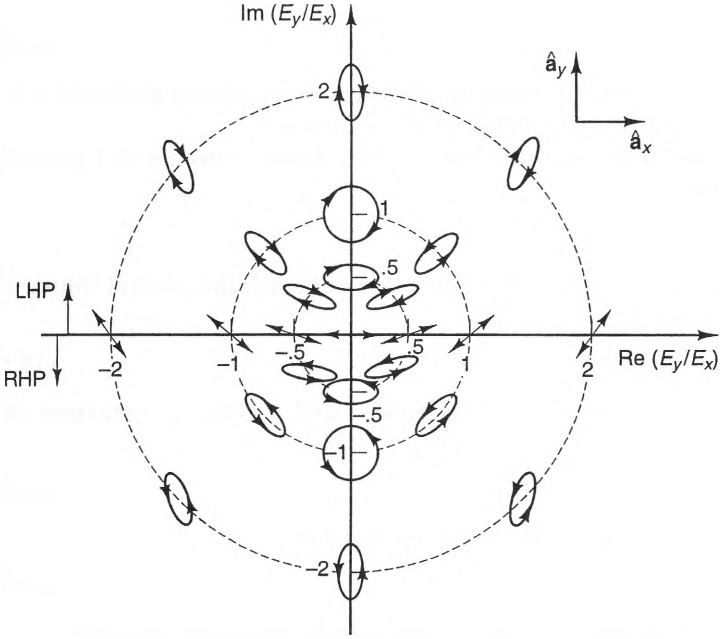


Figure 12-8 Polarization diagram for a $+z$ propagating plane wave.

Example 12-2

Find the polarization ellipse for a plane wave described by

$$\mathbf{E} = 4\cos(\omega t - \beta z)\hat{\mathbf{a}}_x + 2\cos(\omega t + 30° - \beta z)\hat{\mathbf{a}}_y.$$

Solution:

For this wave, we have $E_{x\text{o}} = 4.0$, $E_{y\text{o}} = 2.0$, and $\Delta\theta = 30°$. Using Equations (12.51) and (12.52), we find that

$$OA = \left[\frac{1}{2}\{4^2 + 2^2 + [4^4 + 2^4 + 2 \times 4^2 \times 2^2 \cos(60°)]^{1/2}\}\right]^{1/2} = 4.38$$

$$OB = \left[\frac{1}{2}\{4^2 + 2^2 - [4^4 + 2^4 + 2 \times 4^2 \times 2^2 \cos(60°)]^{1/2}\}\right]^{1/2} = 0.914.$$

From Equation (12.50), the axial ratio is

$$AR = \frac{OA}{OB} = \frac{4.38}{0.914} = 4.79.$$

Finally, using Equation (12.53), we see that the tilt angle is

$$\tau = \frac{1}{2}\tan^{-1}\left[\frac{2 \times 4 \times 2}{4^2 - 2^2}\cos(30°)\right] = 24.55°.$$

12-4 Plane Waves in Lossy Media

Even when a medium is lossy, the electric and magnetic fields in it must still satisfy the wave equation,

$$\nabla^2 E_i + k^2 E_i = 0 \qquad i = x, y, z, \tag{12.54}$$

where the wave number is

$$k = \sqrt{-j\omega\mu(\sigma + j\omega\epsilon)} \qquad [\mathrm{m}^{-1}]. \tag{12.55}$$

However, whereas k is real when the medium is lossless, it is complex when loss is present, either because $\sigma \neq 0$ or because either ϵ or μ is complex.

To demonstrate how plane waves behave in lossy media, consider an x polarized electric field of the form

$$\mathbf{E} = E_x(z)\hat{\mathbf{a}}_x.$$

Since $\mathbf{E}$ has only an x-component, which is a function only of z, the general solution of Equation (12.54) is

$$E_x(z) = E_{xo}^{+}e^{-\gamma z} + E_{xo}^{-}e^{+\gamma z}, \tag{12.56}$$

where E_{xo}^{+} and E_{xo}^{-} are constants (possibly complex) and γ is the propagation constant, given by

$$\gamma = jk = \sqrt{j\omega\mu(\sigma + j\omega\epsilon)} \qquad [\mathrm{m}^{-1}]. \tag{12.57}$$

When loss is present, γ is complex and can be expressed as

$$\gamma = \alpha + j\beta, \tag{12.58}$$

where α and β are the ***attenuation*** and ***phase constants***, respectively, given by

$$\alpha = \mathrm{Re}[\sqrt{j\omega\mu(\sigma + j\omega\epsilon)}] \qquad [\mathrm{Np/m}], \tag{12.59}$$

and

$$\beta = \mathrm{Im}[\sqrt{j\omega\mu(\sigma + j\omega\epsilon)}] \qquad [\mathrm{m}^{-1}]. \tag{12.60}$$

In both Equations (12.59) and (12.60), the square roots are taken so that both α and β are positive when the medium is passive.[2] Also, since $\gamma = jk$, we can write

$$k = \beta - j\alpha. \tag{12.61}$$

Using these definitions, we can write the frequency-domain expression for $\mathbf{E}$ in the form

$$\mathbf{E} = E_{xo}^{+}e^{-\alpha z}e^{-j\beta z}\hat{\mathbf{a}}_x + E_{xo}^{-}e^{\alpha z}e^{j\beta z}\hat{\mathbf{a}}_x. \tag{12.62}$$

The only difference between this expression and the corresponding expression for the lossless case (Equation (12.19)) is that these waves decay exponentially along their respective directions of propagation. This can be seen more clearly by transforming the expression to the time domain. Noting that we can write the complex amplitudes of the waves in the form $E_{xo}^{+} = |E_{xo}^{+}|\angle\theta^{+}$ and $E_{xo}^{-} = |E_{xo}^{+}|\angle\theta^{-}$, we obtain the time-domain expression

$$\mathbf{E} = |E_{xo}^{+}|e^{-\alpha z}\cos(\omega t - \beta z + \theta^{+})\hat{\mathbf{a}}_x + |E_{xo}^{-}|e^{\alpha z}\cos(\omega t + \beta z + \theta^{-})\hat{\mathbf{a}}_x. \tag{12.63}$$

Figure 12-9 shows a "snapshot" of a forward-propagating plane wave in a lossy medium as a function of position.

[2] Passive media can dissipate, but not amplify, waves. Active media, such as are used in lasers and active fiber-optic amplifcrs, are capable of amplifying waves.

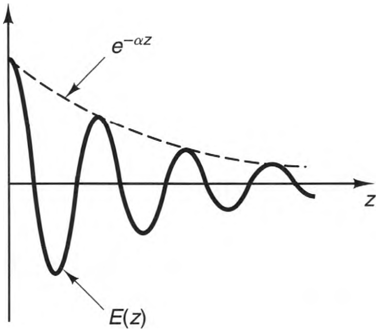


Figure 12-9 E-field amplitude at a fixed time of a plane wave in a lossy medium.

As can be seen, the phase of this wave is dictated by the cosine function, whereas its amplitude is controlled by the exponential function. Just as in the case of transmission lines, it is often convenient to specify the attenuation constant α in terms of the dB attenuation per meter [dB/m], rather than the "natural" units of nepers per meter [Np/m]. The conversion between these two specifications is accomplished by using Equation (11.106),

$$\alpha[\text{Np/m}] = 0.1151 \times \alpha[\text{dB/m}]. \tag{12.64}$$

Since the phase terms in Equations (12.62) and (12.63) have the same form as for the lossless case, the formulas that relate the wavelength λ and phase velocity u_p to the phase constant β are also the same:

$$\lambda = \frac{2\pi}{\beta} \tag{12.65}$$

$$u_p = \frac{\omega}{\beta}. \tag{12.66}$$

However, it is important to remember that $\beta \approx \omega\sqrt{\mu\epsilon}$ in lossy media. Similarly, the group velocity u_g is defined for plane waves just as it was for transmission lines:

$$u_g = \frac{\partial\omega}{\partial\beta} = \left(\frac{\partial\beta}{\partial\omega}\right)^{-1}. \tag{12.67}$$

Since β is not a linear function of ω when loss is present, $u_g \neq u_p$ in lossy media.

The magnetic field that accompanies the E-field given by Equation (12.62) can be determined by substituting **E** into Maxwell's curl-E equation (Equation (12.1)) and solving for **H**, yielding

$$\mathbf{H} = \frac{\boldsymbol{\nabla}\times\mathbf{E}}{-j\omega\mu} = \frac{-1}{j\omega\mu}\frac{\partial}{\partial z}[E_{xo}^{+}e^{-\gamma z} + E_{xo}^{-}e^{\gamma z}]\hat{\mathbf{a}}_v$$

$$= \frac{\gamma}{j\omega\mu}[E_{xo}^{+}e^{-\gamma z} - E_{xo}^{-}e^{+\gamma z}]\hat{\mathbf{a}}_y.$$

This can be written in the form

$$\mathbf{H} = \frac{E_{xo}^{+}}{\eta} e^{-\gamma z} \hat{\mathbf{a}}_y - \frac{E_{xo}^{-}}{\eta} e^{\gamma z} \hat{\mathbf{a}}_y, \tag{12.68}$$

where the intrinsic impedance is

$$\eta = \frac{j\omega\mu}{\gamma} = \sqrt{\frac{j\omega\mu}{\sigma + j\omega\epsilon}} \qquad [\Omega]. \tag{12.69}$$

Since γ is complex in lossy media, η is also complex. If we denote $\eta = |\eta| \angle \theta_\eta$, we can express **H** in the time domain as

$$\mathbf{H} = \frac{|E_{xo}^{+}|}{|\eta|} e^{-\alpha z} \cos(\omega t - \beta z + \theta^{+} - \theta_\eta)\hat{\mathbf{a}}_y$$

$$- \frac{|E_{xo}^{-}|}{|\eta|} e^{-\alpha z} \cos(\omega t + \beta z + \theta^{-} - \theta_\eta)\hat{\mathbf{a}}_y.$$

Comparing this expression with the one obtained for the lossless case (see Equation (12.27)), we see that they are nearly the same. One difference, as we would expect, is that the forward- and backward-propagating H-fields decay in lossy media along their propagation directions. Another difference is that there is an additional phase delay θ_η in both the forward- and backward-propagating H-fields in lossy media that occurs because the intrinsic impedance of lossy materials is complex. As a result, **H** lags behind **E** in phase by the angle θ_η in lossy media.

Finally, just as we can have linearly, circularly, and elliptically polarized plane waves in lossless media that propagate in any direction, the same is true in lossy media. For instance, Equations (12.35)–(12.37) can still be used to describe the E- and H-fields propagating in any direction $\hat{\mathbf{a}}_k$ simply by using the appropriate values of the wave number k and the intrinsic impedance η in these expressions.

12-5 Medium Characterization

When a medium is lossy, both the propagation constant γ and the intrinsic impedance η are complex, either[3] because of ***conduction loss***, where σ is nonzero and finite, or because of ***polarization loss***, where the permittivity ϵ is complex. The physical mechanisms responsible for these losses are different, but they have the same effect on the behaviors of **E** and **H**. Because of this, it is customary to lump all the loss parameters of a material into a single parameter that yields the correct propagation constant γ and intrinsic impedance η.

The key to understanding how to lump both the conductivity and polarization loss into a single parameter is to notice how σ and ϵ appear in the formulas for γ and η:

$$\gamma = \sqrt{j\omega\mu(\sigma + j\omega\epsilon)} \qquad [\mathrm{m}^{-1}]. \tag{12.70}$$

$$\eta = \frac{j\omega\mu}{\gamma} \qquad [\Omega]. \tag{12.71}$$

[3] Magnetization loss (where μ is complex) is also possible in the magnetic materials. They are sometimes used in microwave devices, but are not usually suitable for propagating plane waves.

When μ is real, both γ and η are correctly specified as long as the sum $\sigma + j\omega\epsilon$ has the correct value.

For good conductors, it is customary to specify the total loss in terms of an effective conductivity $\sigma_{\text{eff}} = \sigma + \omega\epsilon''$ and a real permittivity $\epsilon_{\text{eff}} = \epsilon'$, where ϵ' is the real part of the complex permittivity. Thus, for good conductors, we have

$$\epsilon_{\text{eff}} = \epsilon' = \epsilon_r\epsilon_o \tag{12.72a}$$

(Good conductors),

$$\sigma_{\text{eff}} = \sigma + \omega\epsilon'' \tag{12.72b}$$

where σ and $\epsilon = \epsilon' - j\epsilon''$ are the actual conductivity and permittivity of the material and $\epsilon_r = \epsilon'/\epsilon_o$ is the relative permittivity. For good dielectrics, it is customary to specify the total loss in terms of only an effective complex permittivity ϵ_{eff}, so

$$\epsilon_{\text{eff}} = \epsilon' - j\left(\epsilon'' + \frac{\sigma}{\omega}\right) \tag{12.73a}$$

(Good dielectrics),

$$\sigma_{\text{eff}} = 0 \tag{12.74b}$$

where σ and $\epsilon = \epsilon' - j\epsilon''$ are the actual conductivity and permittivity of the material, respectively. Typically, the subscripts "eff" are not used in practice, since it is usually understood that all of the loss has been specified using effective values.

Example 12-3

Suppose that measurements have shown that a certain material has the following values of conductivity σ and permittivity ϵ at 100 [MHz]:

$$\sigma = 500\ [\mu\text{S/m}]$$

$$\epsilon = \epsilon_o(2.5 - j0.4).$$

Find an equivalent set of parameters by assuming that the loss is 1) conductive loss and 2) polarization loss.

Solution:

1) To treat all the loss as conduction loss, we solve Equation (12.72b) for σ_{eff}, obtaining

$$\sigma_{\text{eff}} = \sigma + \omega\epsilon'' = 500 \times 10^{-6} + (2\pi \times 100 \times 10^6) \times (0.4 \times 8.854 \times 10^{-12})$$
$$= 2.73\ [\mu\text{S/m}].$$

Thus, this medium can be described by the parameters

$$\epsilon = 2.5\epsilon_o$$

$$\sigma = 2.72 \quad [\mu\text{S/m}].$$

2) To treat all the loss as polarization loss, we solve Equation (12.73a) for ϵ_{eff}, obtaining

$$\epsilon_{\text{eff}} = \epsilon' - j\left(\epsilon'' + \frac{\sigma}{\omega}\right)$$

$$= 2.5\epsilon_o - j\epsilon_o\left(0.4 + \frac{500 \times 10^{-6}}{2\pi\epsilon_o 100 \times 10^6}\right)$$

$$= \epsilon_o(2.5 - j0.49)$$

Hence, this medium can also be described by the parameters

$$\epsilon = \epsilon_o(2.5 - j0.49)$$

$$\sigma = 0.$$

Another way to specify the loss characteristics of a material is write the complex effective permittivity in the form

$$\epsilon_{\text{eff}} = \epsilon' \,(1 - j\tan\phi) = \epsilon_r\epsilon_o\,(1 - j\tan\phi), \tag{12.74}$$

where ϕ and $\tan\phi$ are called the ***loss angle*** and the ***loss tangent*** of the medium, respectively and $\epsilon_r = \epsilon'/\epsilon_o$ is the relative permittivity. Substituting Equation (12.74) into Equation (12.73a) and solving for $\tan\phi$ yields

$$\tan\phi = \frac{\epsilon''}{\epsilon'} + \frac{\sigma}{\omega\epsilon'}$$

For dielectrics, where σ is assumed to be zero, the loss tangent is given by

$$\tan\phi = \frac{\epsilon''}{\epsilon'} \qquad \text{(Dielectrics; } \sigma = 0\text{)}. \tag{12.75}$$

For conductors, where ϵ'' is assumed to be zero, the loss tangent is given by

$$\tan\phi = \frac{\sigma}{\omega\epsilon'} \qquad \text{(Conductors; } \epsilon'' = 0\text{)}. \tag{12.76}$$

It is possible to specify the loss characteristics of a material by using either an equivalent conductivity, a complex permittivity, or a loss tangent, but it is important to note that, for most materials, these parameters are functions of frequency. As a result, one must always check to see if a tabulated value is valid at the frequency of interest. Generally, tabulated values of complex permittivity tend to be more accurate over larger bandwidths than are tabulated values of conductivity—particularly for low-loss dielectrics. The loss tangents of a number of materials used in engineering practice are shown in Table C-3.

Example 12-4

Suppose that a sample of mica has a relative permittivity of $\epsilon_r = 5.4$ and a loss tangent of $\tan\phi = 0.0006$ at a frequency of 100 [MHz]. Calculate the effective conductivity σ at this frequency.

Solution:

Using Equation (12.76), we find that

$$\sigma = \omega\epsilon\tan\phi = (2\pi \times 100 \times 10^6) \times (5.4 \times 8.54 \times 10^{-12}) \times (0.006)$$

$$= 1.8 \times 10^{-5} \qquad [\text{S/m}].$$

The formulas for γ and η are cumbersome to evaluate when loss is present, because they involve the square roots of complex numbers. However, when the loss is either very small or very large, approximate formulas are available that are more straightforward to evaluate. We will now discuss these two special cases: low-loss dielectrics and good conductors.

Low-Loss Dielectrics. A material is considered to be low loss when its loss tangent is much less than unity; that is,

$$\tan\phi = \frac{\epsilon''}{\epsilon'} \ll 1 \qquad \text{(Low-loss dielectrics)}. \tag{12.77}$$

We can derive approximate formulas for the phase constant β and the attenuation constant α by substituting Equation (12.77) into Equation (12.70) and retaining the three lowest terms of the binomial expansion

$$\gamma = \alpha + j\beta \approx j\omega\sqrt{\mu\epsilon'}\left[1 - j\frac{\epsilon''}{2\epsilon'} + \frac{1}{8}\left(\frac{\epsilon''}{\epsilon'}\right)^2\right]. \tag{12.78}$$

Hence,

$$\alpha \approx \frac{\omega\epsilon''}{2\epsilon'}\sqrt{\mu\epsilon'} = \frac{\omega\sqrt{\epsilon_r}\tan\phi}{2c} \tag{12.79}$$

$$\beta \approx \omega\sqrt{\mu\epsilon'}\left[1 + \frac{1}{8}\left(\frac{\epsilon''}{\epsilon'}\right)^2\right] = \frac{\omega\sqrt{\epsilon_r}}{c}\left[1 + \frac{1}{8}\tan^2\phi\right], \tag{12.80}$$

where $\epsilon_r = \epsilon'/\epsilon_o$ and c is the speed of light in a vacuum. Using a similar procedure, we find that approximate formulas for the intrinsic impedance η and the phase velocity u_p are

$$\eta \approx \sqrt{\frac{\mu}{\epsilon'}}\left[1 + j\frac{\epsilon''}{2\epsilon'}\right] = \frac{\eta_o}{\sqrt{\epsilon_r}}\left[1 + j\frac{1}{2}\tan\phi\right], \tag{12.81}$$

and

$$u_p = \frac{\omega}{\beta} \approx \frac{1}{\sqrt{\mu\epsilon'}}\left[1 - \frac{1}{8}\left(\frac{\epsilon''}{\epsilon'}\right)^2\right] = \frac{c}{\sqrt{\epsilon_r}}\left[1 - \frac{1}{8}\tan^2\phi\right], \tag{12.82}$$

where $\eta_o \approx 377$ [Ω] is the intrinsic impedance of free space.

When the material loss is specified in terms of its conductivity, we can replace $\tan\phi$ in the preceding formulas with $\sigma/(\omega\epsilon)$ and obtain

$$\alpha \approx \frac{\sigma}{2}\sqrt{\frac{\mu}{\epsilon}} \tag{12.83}$$

$$\beta \approx \omega\sqrt{\mu\epsilon}\left[1 + \frac{1}{8}\left(\frac{\sigma}{\omega\epsilon}\right)^2\right] \tag{12.84}$$

$$\eta \approx \sqrt{\frac{\mu}{\epsilon}}\left[1 + j\frac{\sigma}{2\omega\epsilon}\right] \tag{12.85}$$

$$u_p = \frac{\omega}{\beta} \approx \frac{1}{\sqrt{\mu\epsilon}}\left[1 - \frac{1}{8}\left(\frac{\sigma}{\omega\epsilon}\right)^2\right], \tag{12.86}$$

where it is assumed that ϵ is the real part of the complex permittivity.

Good Conductors. A material is considered to be a good conductor when its loss tangent is much greater than unity; that is,

$$\tan\phi = \frac{\sigma}{\omega\epsilon} \gg 1 \qquad \text{(Good conductors)}. \tag{12.87}$$

For this case, the propagation constant γ can be approximated as

$$\gamma = \alpha + j\beta = \sqrt{j\omega\mu(\sigma + j\omega\epsilon)} = j\omega\sqrt{\mu\epsilon}\sqrt{\left(1 + \frac{\sigma}{j\omega\epsilon}\right)} \tag{12.88}$$

$$\approx j\omega\sqrt{\mu\epsilon}\sqrt{\frac{\sigma}{j\omega\epsilon}} = \sqrt{j\omega\mu\sigma} = \frac{(1+j)}{\sqrt{2}}\sqrt{\omega\mu\sigma}.$$

Thus,

$$\alpha \approx \beta \approx \sqrt{\pi f \mu\sigma}. \tag{12.89}$$

Also, substituting Equation (12.89) into Equation (12.71), we obtain

$$\eta \approx \frac{(1+j)}{\sqrt{2}}\sqrt{\frac{\omega\mu}{\sigma}} = \sqrt{\frac{\omega\mu}{\sigma}}\angle 45^\circ. \tag{12.90}$$

Since the attenuation constant for a good conductor is large, the depth of penetration of the fields is small. The distance into the conductor at which the field strengths are diminished by the factor $e^{-1} = 0.368$ is called the ***skin depth*** δ. Using Equation (12.89), we can write

$$\delta = \frac{1}{\alpha} = \frac{1}{\sqrt{\pi f \mu \sigma}}. \tag{12.91}$$

Also, since $\alpha \approx \beta$ and $\beta = 2\pi/\lambda$, we can also write

$$\delta = \frac{\lambda}{2\pi}, \tag{12.92}$$

which shows that the wavelength in a good conductor is of the same order of magnitude as the skin depth.

Example 12-5

A plane wave is incident from free space into seawater. Calculate the distance below the surface where the E-field is 10% of its value at the surface when a) $f = 30$ [Hz] and b) $f = 10$ [GHz]. At these frequencies, seawater has the following constitutive parameters: [4]

$$\mu = \mu_o, \ \epsilon' = 80\,\epsilon_o, \ \sigma = 4 \text{ [S/m] } @ f = 30 \text{ [Hz]}$$

$$\mu = \mu_o, \ \epsilon' = 80\,\epsilon_o, \ \epsilon'' = 45\,\epsilon_o \quad @ f = 10 \text{ [GHz]}.$$

Solution:

a) At 30 [Hz], the loss tangent is

$$\tan\phi = \frac{\sigma}{\omega\epsilon} = \frac{4}{2\pi \times 30 \times 80 \times 8.854 \times 10^{-12}} = 3 \times 10^7 \gg 1,$$

which means that seawater is a good conductor at this frequency. Using Equation (12.89), we have

$$\alpha \approx \sqrt{\pi \times 30 \times 4\pi \times 10^{-7} \times 4} = 2.18 \times 10^{-2} \text{ [Np/m]}.$$

Since **E** decays proportionally to $e^{-\alpha z}$, the depth d at which the electric field is 10% of its value at the surface satisfies the expression

$$e^{-\alpha d} = 0.1 .$$

Solving for d, we obtain

$$d = -\frac{\ln(0.1)}{2.18 \times 10^{-2}} = 106 \text{ [m]}.$$

b) At 10 [GHz], the loss tangent is

$$\tan\phi = \frac{\epsilon''}{\epsilon'} = \frac{45\,\epsilon_o}{80\,\epsilon_o} = 0.56,$$

which means that the seawater is neither a good conductor nor a good dielectric, so we must use Equation (12.70) to find the attenuation constant:

$$\alpha = \text{Re}[\sqrt{j\omega\mu(\sigma + j\omega\epsilon)}] = \text{Re}[\sqrt{-\omega^2\mu_o\epsilon_o(80 - j45)}]$$

$$= \frac{\omega}{c}\text{Re}[\sqrt{-(80 - j45)}] = \frac{(2\pi \times 10 \times 10^9)}{3 \times 10^8}\text{Re}[2.43 + j9.27]$$

$$= 508.8 \text{ [Np/m]}.$$

Hence, the depth at which **E** is 10% of its value at the surface is

$$d = -\frac{\ln(0.1)}{\alpha} = \frac{2.3}{508.8} = 0.005 \text{ [m]}$$

From the results of parts a) and b), we see that the transmission characteristics of seawater at low frequencies are far superior to those at microwave frequencies. Because of this, submarine radio communications are conducted using either very low-frequency electromagnetic waves (usually 40 [Hz] to 10 [KHz]) or sound waves (used in sonar), which decay at a much slower rate.

[4] Taken from Moore, Fung, and Ulaby, *Microwave Remote Sensing*, Volume 3, Artech House, Deadham, MA, 1981.

12-6 Power Transmission

We found in the previous chapter that propagating voltage and current waves on transmission lines transmit power in the direction of propagation. The same is true for plane waves. We will show this by first deriving an important theorem of electromagnetic theory—Poynting's theorem. Poynting's theorem describes the electromagnetic power balance over an arbitrary volume. Using this theorem, we will develop formulas for the power transported by a plane wave.

12-6-1 POYNTING'S THEOREM

We will start our discussion by considering fields in the time domain. If we take the dot product of both sides of Maxwell's curl-H equation (Equation 10.19) with **E**, we obtain

$$\mathbf{E}\cdot\nabla\times\mathbf{H} = \mathbf{E}\cdot\mathbf{J} + \mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t}.$$

Using the vector identity given by Equation (B.1), we can expand the left-hand side of the preceding expression so that it reads

$$\mathbf{H}\cdot\nabla\times\mathbf{E} - \nabla\cdot(\mathbf{E}\times\mathbf{H}) = \mathbf{E}\cdot\mathbf{J} + \mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t}.$$

Rearranging, we have

$$-\nabla\cdot(\mathbf{E}\times\mathbf{H}) = \mathbf{E}\cdot\mathbf{J} + \mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t} - \mathbf{H}\cdot\nabla\times\mathbf{E}.$$

Also, from Maxwell's curl-**E** equation, we have

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

Substituting this into the former expression, we obtain

$$-\nabla\cdot(\mathbf{E}\times\mathbf{H}) = \mathbf{E}\cdot\mathbf{J} + \mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t} + \mathbf{H}\cdot\frac{\partial\mathbf{B}}{\partial t}. \tag{12.93}$$

If the medium is simple, we can also write

$$\mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t} = \epsilon\mathbf{E}\cdot\frac{\partial\mathbf{E}}{\partial t} = \frac{\epsilon}{2}\frac{\partial}{\partial t}(\mathbf{E}\cdot\mathbf{E}) = \frac{\partial}{\partial t}\left(\frac{\epsilon E^2}{2}\right) \tag{12.94}$$

and

$$\mathbf{H}\cdot\frac{\partial\mathbf{B}}{\partial t} = \mu\mathbf{H}\cdot\frac{\partial\mathbf{H}}{\partial t} = \frac{\mu}{2}\frac{\partial}{\partial t}(\mathbf{H}\cdot\mathbf{H}) = \frac{\partial}{\partial t}\left(\frac{\mu H^2}{2}.\right) \tag{12.95}$$

Substituting Equations (12.94) and (12.95) into Equation (12.93), we obtain

$$-\nabla\cdot(\mathbf{E}\times\mathbf{H}) = \mathbf{E}\cdot\mathbf{J} + \frac{\partial}{\partial t}\left(\frac{\epsilon E^2}{2} + \frac{\mu H^2}{2}\right).$$

Integrating this expression over an arbitrary volume V, we find that

$$-\int_V \nabla \cdot (\mathbf{E} \times \mathbf{H}) dv = \int_V \mathbf{E} \cdot \mathbf{J} dv + \frac{\partial}{\partial t} \int_V \left(\frac{\epsilon E^2}{2} + \frac{\mu H^2}{2} \right) dv.$$

Finally, we can use the divergence theorem to express the left-hand side as a surface integral over the surface S that bounds V:

$$-\oint_S \mathbf{E} \times \mathbf{H} \cdot d\mathbf{s} = \int_V \mathbf{E} \cdot \mathbf{J} dv + \frac{\partial}{\partial t} \int_V \left(\frac{\epsilon E^2}{2} + \frac{\mu H^2}{2} \right) dv. \qquad (12.96)$$

Equation (12.96) is called ***Poynting's theorem***. To see what this theorem says physically, note initially that the first term on the right-hand side is an expression of Joule's law (Equation (5.17)), which represents the instantaneous power dissipated inside V:

$$P_{\text{diss}} = \int_V \mathbf{E} \cdot \mathbf{J} dv. \qquad (12.97)$$

If the volume contains only passive media and no sources, $P_{\text{diss}} \geqslant 0$.

Next, the integrals involving E^2 and H^2 represent the electric and magnetic energies,[5] W_e and W_m, respectively, stored within V (see Equations (6.34) and (9.39)), where the total stored energy is

$$W = W_e + W_m = \int_V \left(\frac{\epsilon E^2}{2} + \frac{\mu H^2}{2} \right) dv. \qquad (12.98)$$

Thus, the time derivative of this volume integral represents the time rate of change in the energy stored in V. When $\partial W/\partial t > 0$, the net power stored is increasing, whereas when $\partial W/\partial t < 0$, the net power stored is decreasing.

From the foregoing comments, we can conclude that the right-hand side of Equation (12.96) is the sum of the dissipated and stored power within the volume V. This power must come from somewhere, so it must equal the net power transferred through the bounding surface S into the volume. Thus, the surface integral on the left-hand side of Equation (12.96) must be equal to the net power flowing into V through its bounding surface S; that is,

$$P_{\text{in}} = -\oint_S \mathbf{E} \times \mathbf{H} \cdot d\mathbf{s} \qquad [\text{W}]. \qquad (12.99)$$

Conversely, the negative of the integral must be the net power flowing *out* of V through its bounding surface S:

$$P_{\text{out}} = \oint_S \mathbf{E} \times \mathbf{H} \cdot d\mathbf{s} \qquad [\text{W}]. \qquad (12.100)$$

[5] Although these expressions were derived rigorously for static fields, they have been found to be true for time-varying fields as well.

Equation (12.100) is an important corollary of Poynting's theorem and is valid for any distribution of **E** and **H**, not just plane waves. When **E** and **H** are represented in the time domain, the cross product **E** × **H** is denoted by the symbol $\mathscr{S}$ and called the ***instantaneous Poynting vector*.**[6]

$$\mathscr{S} = \mathbf{E} \times \mathbf{H} \qquad [\mathrm{W/m^2}]. \tag{12.101}$$

Since $\mathscr{S}$ is measured in units of watts per square meter, it is natural to assume that $\mathscr{S}$ represents the direction and density of the power flow at a point. As it turns out, this interpretation is always correct for time-varying fields, but can occasionally yield strange results for problems involving static fields.[7] Fortunately, since we are interested only in time-varying fields for the remainder of this text, this interpretation will pose no such problems.

Example 12-6

Use Poynting's theorem to calculate the power transported by a forward-propagating TEM mode in the coaxial transmission line shown in Figure 12-10.

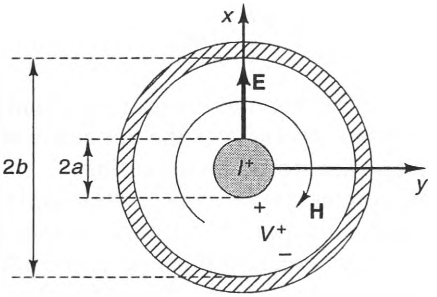


Figure 12-10 A TEM wave propagating into the paper on a coaxial cable.

Solution:

If the instantaneous voltage at a point on the transmission line is V^+, the instantaneous current I^+ (into the page) on the inner conductor is

$$I^+ = \frac{V^+}{Z_0},$$

where Z_0 is the characteristic impedance of the line. The electric and magnetic fields of a coaxial TEM wave are given by Equations (5.65) and (7.35), respectively;

[6] Named after John Henry Poynting (1852–1914), a British physicist.

[7] For an interesting discussion of these problems, see Daniel R. Frankl, *Electromagnetic Theory* (Englewood Cliffs, NJ: Prentice Hall, 1986), pp. 205–206.

$$\mathbf{E} = \frac{V^+}{\rho \ln\left(\dfrac{b}{a}\right)} \hat{\mathbf{a}}_\rho$$

$$\mathbf{H} = \frac{I^+}{2\pi\rho} \hat{\mathbf{a}}_\phi = \frac{V^+}{2\pi Z_o \rho} \hat{\mathbf{a}}_\phi .$$

Using these fields, we find that the instantaneous Poynting vector is

$$\mathscr{S} = \frac{V^+}{\rho \ln b/a} \hat{\mathbf{a}}_\rho \times \frac{V^+}{2\pi Z_o \rho} \hat{\mathbf{a}}_\phi = \frac{(V^+)^2}{2\pi Z_o \rho^2 \ln b/a} \hat{\mathbf{a}}_z .$$

The power transmitted through any constant-z surface is

$$P_{\text{trans}} = \int_S \mathscr{S} \cdot \mathbf{ds} = \frac{(V^+)^2}{2\pi Z_o \ln b/a} \int_0^{2\pi} \int_a^b \frac{1}{\rho^2} \rho \, d\rho \, d\phi$$

$$= \frac{(V^+)^2}{Z_o} = V^+ I^+ ,$$

which is the familiar result predicted by ordinary circuit theory.

12-6-2 AVERAGE POWER FLOW AND THE COMPLEX POYNTING VECTOR

When the fields are time varying, it is often desirable to keep track of the average power, rather than the instantaneous power. We define the ***average Poynting vector*** for time-periodic fields as

$$\mathscr{S}_{\text{ave}} \equiv \frac{1}{T} \int_0^T \mathscr{S} \, dt = \frac{1}{T} \int_0^T \mathbf{E} \times \mathbf{H} \, dt, \tag{12.102}$$

where T is the waveform period. If **E** and **H** are time harmonic, we can write them at any point as

$$\mathbf{E} = |E| \cos(\omega t + \theta_E) \hat{\mathbf{a}}_E$$

$$\mathbf{H} = |H| \cos(\omega t + \theta_H) \hat{\mathbf{a}}_H ,$$

where the unit vectors $\hat{\mathbf{a}}_E$ and $\hat{\mathbf{a}}_H$ point in the direction of **E** and **H**, respectively. Substituting these expressions into Equation (12.102) and using the trigonometric identity

$$\cos a \cos b = \frac{1}{2} \cos(a + b) + \frac{1}{2} \cos(a - b),$$

we obtain

$$\mathscr{S}_{\text{ave}} = \frac{1}{2} |E| \hat{\mathbf{a}}_E \times |H| \hat{\mathbf{a}}_H \cdot \frac{1}{T} \int_0^T [\cos(2\omega t + \theta_E + \theta_H) + \cos(\theta_E - \theta_H)] \, dt,$$

where $T = 2\pi/\omega$. The average value of the first term inside the integral is zero, and the second term inside the integral is a constant, so it is equal to its own average. Hence, we can write

$$\mathscr{S}_{\text{ave}} = \frac{1}{2}|E|\,\hat{\mathbf{a}}_E \times |H|\,\hat{\mathbf{a}}_H \cdot \cos(\theta_E - \theta_H).$$

This can be rewritten using complex notation as

$$\mathscr{S}_{\text{ave}} = \frac{1}{2}\operatorname{Re}\{(|E|e^{j\theta_E}\hat{\mathbf{a}}_E) \times (|H|e^{-j\theta_H}\hat{\mathbf{a}}_H)\}$$

where "Re" denotes "the real part of." Noting that $|H|e^{-j\theta_H}\ \hat{\mathbf{a}}_H$ is simply the complex conjugate of **H**, we can write $\mathscr{S}_{\text{ave}}$ in the form

$$\mathscr{S}_{\text{ave}} = \frac{1}{2}\operatorname{Re}(\mathbf{S}) \qquad [\text{W/m}^2], \tag{12.103}$$

where

$$\mathbf{S} = \mathbf{E} \times \mathbf{H}^* \qquad [\text{W/m}^2] \tag{12.104}$$

and "*" denotes "the complex conjugate of." The vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}^*$ is called the ***complex Poynting vector*** and is the vector analog of complex power used in ac circuit analysis. Comparing Equations (12.102) and (12.103), we see that the average Poynting vector $\mathscr{S}$ for time-harmonic fields can be determined either by averaging the instantaneous Poynting vector $\mathscr{S}$ or by taking one-half of the real part of the complex Poynting vector **S**. Using these definitions, we see that the average power that passes through a surface S is given by

$$P_{\text{ave}} = \int_S \mathscr{S}_{\text{ave}} \cdot \mathbf{ds}. \tag{12.105}$$

12-6-3 THE POWER TRANSMISSION OF PLANE WAVES

We can use the Poynting vector to determine the power that is transported by a plane wave. Let us start by considering the following time-domain description of a linearly polarized plane wave in a lossless medium:

$$\mathbf{E} = E_{xo}^{+}\cos(\omega t - \beta z)\,\hat{\mathbf{a}}_x$$

$$\mathbf{H} = \frac{E_{xo}^{+}}{\eta}\cos(\omega t - \beta z)\,\hat{\mathbf{a}}_y.$$

The instantaneous Poynting vector associated with this wave is

$$\mathscr{P}(t) = \mathbf{E} \times \mathbf{H} = \left(E_{xo}^{+} \cos(\omega t - \beta z)\hat{\mathbf{a}}_x\right) \times \left(\frac{E_{xo}^{+}}{\eta} \cos(\omega t - \beta z)\hat{\mathbf{a}}_y\right)$$

$$= \frac{|E_{xo}^{+}|^2}{\eta} \cos^2(\omega t - \beta z)\hat{\mathbf{a}}_z \qquad [\text{W/m}^2].$$

Thus, $\mathscr{P}(t)$ always points in the direction of propagation and varies in amplitude between zero and $|E_{xo}^{+}|^2/\eta$. Given that the time average of the term $\cos^2(\omega t - \beta z)$ is 1/2, we can express the average Poynting vector of the plane wave as

$$\mathscr{S}_{\text{ave}} = \frac{1}{2}\frac{|E_{xo}^{+}|^2}{\eta}\hat{\mathbf{a}}_z \qquad [\text{W/m}^2]. \qquad (12.106)$$

Since our choice of $+z$ propagation and x-polarization was arbitrary, we can generalize Equation (12.106) to read

$$\mathscr{S}_{\text{ave}} = \frac{1}{2}\frac{|E_o|^2}{\eta}\hat{\mathbf{a}}_k \qquad [\text{W/m}^2] \qquad \text{(Linearly polarized plane waves in lossless medium),} \qquad (12.107)$$

where $\hat{\mathbf{a}}_k$ is the direction of propagation and E_o is the peak amplitude (in time) of the electric-field vector. From this expression, we see that the average power density of a linearly polarized plane wave in a lossless medium is independent of position.

Example 12-7

A linearly polarized plane wave propagates through free space at an angle θ with respect to the $z = 0$ plane, as shown in Figure 12-11. If the peak amplitude of **E** is 10 [mV/m], calculate the average power that passes through the 1 [m^2] surface shown in the figure.

Solution:

Using Equation (12.107), we can represent the average Poynting vector as

$$\mathscr{S}_{\text{ave}} = \frac{1}{2}\frac{|10 \times 10^{-3}|^2}{377}\hat{\mathbf{a}}_k = 132.6\hat{\mathbf{a}}_k \qquad [\text{nW/m}^2],$$

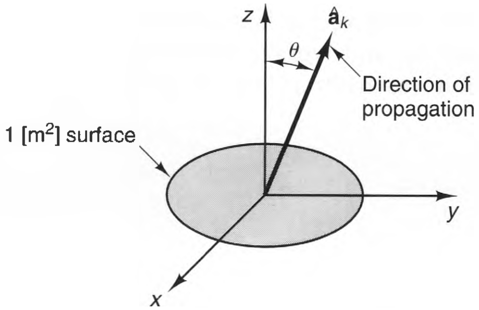


Figure 12-11 A plane wave propagating at an angle θ through a 1 [m^2] surface.

where $\hat{\mathbf{a}}_k$ points in the direction of propagation.

The average power that passes through the surface is

$$P_{\text{ave}} = \int_S \mathscr{S}_{\text{ave}} \cdot \mathbf{ds} = 132.6 \times \int_S \hat{\mathbf{a}}_k \cdot \hat{\mathbf{a}}_z dz.$$

From Figure 12-11, we see that $\hat{\mathbf{a}}_k \cdot \hat{\mathbf{a}}_z = \cos\theta$. Since the total surface area is 1 [m^2], we finally obtain

$$P_{\text{ave}} = 132.6 \cos\theta \qquad [\text{nW}].$$

Thus, the power that passes through the surface is maximized when the direction of propagation is parallel to the surface normal.

In the preceding paragraphs, we saw that the instantaneous power density of a linearly polarized wave is proportional to $\cos^2(\omega t - \beta z)$. This means that the power density of a linearly polarized plane wave is small for a significant percentage of its period. Let us now consider the electric and magnetic fields of a circularly polarized wave, given by

$$\mathbf{E} = E_o^+ \cos(\omega t - \beta z)\hat{\mathbf{a}}_x \pm E_o^+ \sin(\omega t - \beta z)\hat{\mathbf{a}}_y$$

and

$$\mathbf{H} = \frac{E_o^+}{\eta}\cos(\omega t - \beta z)\hat{\mathbf{a}}_y \mp \frac{E_o^+}{\eta}\sin(\omega t - \beta z)\hat{\mathbf{a}}_x.$$

The instantaneous Poynting vector for such a wave is

$$\mathscr{S}(t) = \frac{(E_o^+)^2}{\eta}\Big(\cos(\omega t - \beta z)\hat{\mathbf{a}}_x \pm \sin(\omega t - \beta z)\hat{\mathbf{a}}_y\Big) \times$$
$$\Big(\cos(\omega t - \beta z)\hat{\mathbf{a}}_y \mp \sin(\omega t - \beta z)\hat{\mathbf{a}}_x\Big)$$
$$= \frac{(E_o^+)^2}{\eta}(\cos^2(\omega t - \beta z) + \sin^2(\omega t - \beta z))\hat{\mathbf{a}}_z.$$

Simplifying this expression, we obtain

$$\mathscr{S}(t) = \frac{(E_o^+)^2}{\eta}\hat{\mathbf{a}}_z. \tag{12.108}$$

Here, we see that the Poynting vector is a constant in both time and position, which means that the instantaneous and average Poynting vectors are equal. Given that our choice of $+z$ propagation was arbitrary, we can generalize Equation (12.106) to read

$$\mathscr{S}(t) = \mathscr{S}_{\text{ave}} = \frac{(E_o)^2}{\eta}\hat{\mathbf{a}}_k \qquad [\text{W/m}^2] \qquad \begin{array}{c}\text{(Circularly polarized plane}\\ \text{wave in lossless medium),}\end{array} \tag{12.109}$$

where $\hat{\mathbf{a}}_k$ is the direction of propagation and E_o is the amplitude of the electric-field vector.

Comparing Equations (12.109) and (12.107), we see that the average power density of a circularly polarized wave is exactly twice that of a linearly polarized wave that has the same peak amplitude. There are two ways to explain this. The first is that a circularly polarized wave consists of two orthogonal waves that have equal amplitudes. Hence, the total power is twice the value of either one. The second explanation is that the amplitudes of **E** and **H** remain constant for a circularly polarized wave, so such a wave is "on" twice as much as a linearly polarized wave.

Finally, to see how loss affects the power density of plane waves, let us write the frequency-domain representations of the electric and magnetic field of a simple plane wave in a lossy medium. We have

$$\mathbf{E} = E_x^+ e^{-\alpha z} e^{-j\beta z}\hat{\mathbf{a}}_x$$

and

$$\mathbf{H} = \frac{E_{xo}^+}{\eta} e^{-\alpha z} e^{-j\beta z}\hat{\mathbf{a}}_y,$$

where η is complex valued. The complex Poynting vector associated with this wave is

$$\mathbf{S} = \mathbf{E} \times \mathbf{H}^* = (E_{xo}^+ e^{-\alpha z} e^{-j\beta z}\hat{\mathbf{a}}_x) \times \left(\frac{E_{xo}^+}{\eta} e^{-\alpha z} e^{-j\beta z}\hat{\mathbf{a}}_y\right)^*$$

$$= \frac{|E_{xo}^+|^2}{\eta^*} e^{-2\alpha z}\hat{\mathbf{a}}_z.$$

Using $\mathscr{S}_{\text{ave}} = \dfrac{1}{2}\text{Re}(\mathbf{E} \times \mathbf{H}^*)$, we find that the average Poynting vector is

$$\mathscr{S}_{\text{ave}} = \frac{1}{2}\text{Re}\left[(E_{xo}^+ e^{-\alpha z} e^{-j\beta z}\hat{\mathbf{a}}_x) \times \left(\frac{E_{xo}^+}{\eta} e^{-\alpha z} e^{j\beta z}\hat{\mathbf{a}}_y\right)\right],$$

which can be simplified to read

$$\mathscr{S}_{\text{ave}} = \frac{1}{2}\frac{|E_{xo}^+|^2}{|\eta|} e^{-2\alpha z}\cos\theta_\eta \hat{\mathbf{a}}_z \quad [\text{W/m}^2] \quad \text{(Linearly polarized plane wave in lossy medium)}, \qquad (12.110)$$

where $|\eta|$ and θ_η are the magnitude and phase of the intrinsic impedance η, respectively.

Example 12-8

A 1-GHz, linearly polarized plane wave propagates through a lossy medium in the $+x$ direction and has a magnitude of 1 [V/m] at $x = 0$. Use Poynting's theorem to calculate how much power is dissipated in the 1 [m^3] volume shown in Figure 12-12. Assume that the medium has a relative permittivity of $\epsilon_r = 2.0$ and a loss tangent of $\tan\phi = 0.001$.

Solution:

Since $\tan\phi = \epsilon''/\epsilon' \ll 1$, the medium is a low-loss dielectric. Using Equations (12.79) and (12.81), we find that

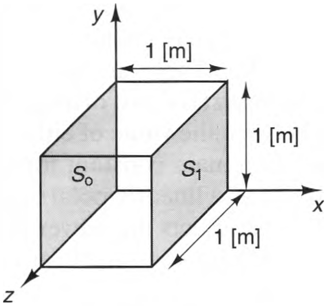


Figure 12-12 A square volume in which power is dissipated by a propagating plane wave.

$$\alpha \approx \frac{\omega\epsilon''}{2\epsilon'}\sqrt{\mu\epsilon'} = \frac{\omega}{2} \times \frac{\epsilon''}{\epsilon'} \times \frac{\sqrt{\epsilon_r}}{c} = 0.0148 \qquad \text{[Np/m]}$$

and

$$\eta \approx \sqrt{\frac{\mu}{\epsilon'}}\left[1 + j\frac{\epsilon''}{2\epsilon'}\right] = 266.38\angle 0.029° .$$

From Equation (12.110), the average Poynting vector is

$$\begin{aligned}\mathscr{S}_{\text{ave}} &= \frac{1}{2}\frac{|1|^2}{266.38} e^{-2(0.0148)x} \cos(0.029°)\,\hat{\mathbf{a}}_x \\ &= 1.88e^{-0.0296x}\hat{\mathbf{a}}_x \qquad [\text{mW/m}^2].\end{aligned}$$

The average power dissipated P_{diss} in the volume equals the net average power P_{in} flowing into the bounding surface. Since $\mathscr{S}_{\text{ave}}$ has only an x component, the sole contributions to this integral come from the faces S_0 and S_1, which have areas of 1 [m^2] each and are located at $x = 0$ and $x = 1$, respectively. Thus,

$$\begin{aligned}P_{\text{diss}} &= -\oint_S \mathscr{S}_{\text{ave}} \cdot \mathbf{ds} = \int_{S_0} \mathscr{S}_{\text{ave}} \cdot \hat{\mathbf{a}}_x ds - \int_{S_1} \mathscr{S}_{\text{ave}} \cdot \hat{\mathbf{a}}_x ds \\ &= 1.88 \times 10^{-3}[\text{W/m}^2][1 - e^{-0.0296}] \times 1\,[\text{m}^2] \\ &= 54.8\,[\mu\text{W}].\end{aligned}$$

12-7 Plane-Wave Reflection and Transmission at Planar Boundaries, Normal Incidence

Up to this point, we have discussed the characteristics of plane waves in source-free, homogeneous regions of space. A sizable percentage of space does indeed fit this description, but the reflections of waves off of material interfaces must often be considered in order to determine the actual performance of a system or device. For example, the waves generated by broadcast TV stations can reflect off buildings and hills,[8] giving rise to one or more delayed signals that cause undesirable "ghosts". Wave reflections can also be beneficial, as they are in lasers. In this case, waves are reflected

[8] Yes, there *are* hills in Kansas!

back and forth through an amplifying medium to produce the laser's high-amplitude output beam.

We will begin this section by considering the reflection and transmission experienced by plane waves that are normally incident upon a planar dielectric interface. After this, we will discuss the reflections and transmissions of normally incident plane waves from stratified interfaces.

12-7-1 REFLECTION AND TRANSMISSION COEFFICIENTS

Figure 12-13 shows the interface between two dielectric regions. In region 1 ($z < 0$) the constitutive parameters are ϵ_1, μ_1, and σ_1, and in region 2 ($z > 0$) the constitutive parameters are ϵ_2, μ_2, and σ_2. We will assume that an incident plane wave launched by sources somewhere to the left of the interface propagates in the $+z$ direction and has electric and magnetic fields described by

$$\mathbf{E}^i = E^i e^{-\gamma_1 z}\hat{\mathbf{a}}_x \ (z \leqslant 0) \tag{12.111}$$

$$\mathbf{H}^i = \frac{E^i}{\eta_1} e^{-\gamma_1 z}\hat{\mathbf{a}}_y \quad (z \leqslant 0), \tag{12.112}$$

where γ_1 and η_1 are the propagation constant and intrinsic impedance of region 1, respectively. These waves satisfy the wave equation in region 1, but reflected and transmitted fields will also be present because of the interface. We will proceed by assuming that the reflected and transmitted fields have the same polarization as the incident field and then prove this assumption to be true by showing that the necessary boundary conditions at the interface are satisfied when these reflected and transmitted fields have specific amplitudes.

We expect that reflected fields will exist in region 1 and propagate towards the left, which can be expressed as

$$\mathbf{E}^r = E^r e^{\gamma_1 z}\hat{\mathbf{a}}_x \ (z \leqslant 0) \tag{12.113}$$

$$\mathbf{H}^r = -\frac{E^r}{\eta_1} e^{\gamma_1 z}\hat{\mathbf{a}}_y \quad (z \leqslant 0), \tag{12.114}$$

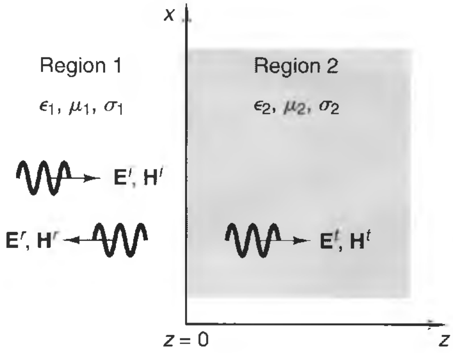


Figure 12-13 A plane wave normally incident upon a planar interface between two dissimilar media.

where the minus sign in the H-field expression occurs because this field is propagating towards the left. The transmitted fields in region 2 propagate towards the right, so we can write them as

$$\mathbf{E}^t = E^t e^{-\gamma_2 z}\hat{\mathbf{a}}_x \quad (z \geqslant 0) \tag{12.115}$$

$$\mathbf{H}^t = \frac{E^t}{\eta_2} e^{-\gamma_2 z}\hat{\mathbf{a}}_y \quad (z \geqslant 0), \tag{12.116}$$

where γ_2 and η_2 are, respectively, the propagation constant and intrinsic impedance of region 2.

The incident, reflected, and transmitted fields each satisfy the wave equation in their respective regions. However, they will not satisfy the necessary boundary conditions at the interface unless E^r and E^t assume specific values. These boundary conditions require that the tangential electric and magnetic fields both be continuous across the interface. Starting with the tangential electric fields, we note that the incident, reflected, and transmitted fields are all tangent to the interface. Also, the total tangential field in region 1 is the sum of the incident and reflected fields. Equating these fields at the interface, we have

$$E^i e^{-\gamma_1 z}\hat{\mathbf{a}}_x\Big|_{z=0^-} + E^r e^{+\gamma_1 z}\hat{\mathbf{a}}_x\Big|_{z=0^-} = E^t e^{-\gamma_2 z}\hat{\mathbf{a}}_x\Big|_{z=0^+}.$$

Canceling the common terms, this results in the following relationship between the amplitudes of the incident, reflected, and transmitted fields:

$$E^i + E^r = E^t. \tag{12.117}$$

Next, continuity of the tangential H-field at the interface requires that

$$\frac{E^i}{\eta_1} e^{-\gamma_1 z}\hat{\mathbf{a}}_y\Big|_{z=0^-} - \frac{E^r}{\eta_1} e^{\gamma_1 z}\hat{\mathbf{a}}_y\Big|_{z=0^-} = \frac{E^t}{\eta_2} e^{-\gamma_2 z}\hat{\mathbf{a}}_y\Big|_{z=0^+},$$

which results in another expression for the incident, reflected, and transmitted field amplitudes, namely,

$$\frac{E^i}{\eta_1} - \frac{E^r}{\eta_1} = \frac{E^t}{\eta_2}. \tag{12.118}$$

Substituting Equation (12.118) into Equation (12.117) and solving for E^r and E^t in terms of E^i, we obtain

$$E^r = E^i \frac{\eta_2 - \eta_1}{\eta_2 + \eta_1} \tag{12.119}$$

and

$$E^t = E^i \frac{2\eta_2}{\eta_2 + \eta_1}. \tag{12.120}$$

The ratio of E^r to E^i is called the ***reflection coefficient*** and is designated by the symbol Γ. Using Equation (12.119), we find that

$$\Gamma \equiv \frac{E^r}{E^i} = \frac{\eta_2 - \eta_1}{\eta_2 + \eta_1}. \tag{12.121}$$

When both media are passive, Γ is real and has a magnitude less than or equal to unity. This means that the reflected wave cannot have a larger magnitude than the incident wave. Values of Γ close to unity are obtained when there is a big mismatch between the impedances, i.e., which $\eta_2 >> \eta_1$ or $\eta_1 >> \eta_2$. We can also define an H-field reflection coefficient as the ratio of the reflected and incident H-fields at the interface. Just as in the case of the waves on transmission lines, the H-field reflection coefficient is simply the negative of the E-field reflection coefficient Γ.

The amplitude ratio of the transmitted and incident electric fields at the interface is called the ***transmission coefficient*** and is designated by the symbol T. Using Equation (12.120), we can write

$$T \equiv \frac{E^t}{E^i} = \frac{2\eta_2}{\eta_2 + \eta_1}. \tag{12.122}$$

Substituting the definitions of Γ and T into Equations (12.117) and (12.118), we obtain two important relationships between the reflection and transmission coefficients:

$$T = 1 + \Gamma \tag{12.123}$$

$$\Gamma^2 + \frac{\eta_1}{\eta_2} T^2 = 1. \tag{12.124}$$

Unlike the reflection coefficient Γ, the transmission coefficient T can have a magnitude greater than unity. This occurs when the incident medium has a lower characteristic impedance than the transmission medium. In the limit as η_2 approaches infinity, T approaches 2.0, which means that the E-field of the transmitted wave is twice as big as the incident wave. Although this may at first seem like a violation of energy, we will prove shortly that it is not, since the transmitted H-field is nearly zero.

12-7-2 REFLECTED AND TRANSMITTED POWER

When a plane wave transports power towards a planar interface, power is also transported by the reflected and transmitted waves. As might be expected, the amount of power transported by the reflected and transmitted waves is proportional to the reflection and transmission coefficients, Γ and T, respectively. If the incident medium is lossless, these relationships are particularly simple.

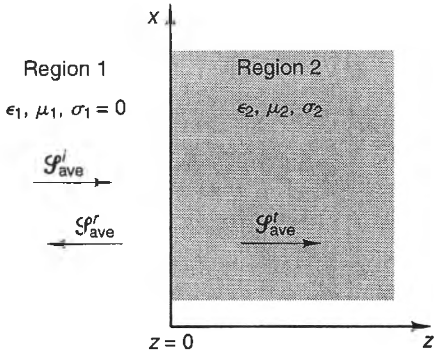

x
Region 1
$\epsilon_1, \mu_1, \sigma_1 = 0$
$\mathscr{S}^i_{\rm ave}$
$\mathscr{S}^r_{\rm ave}$
Region 2
$\epsilon_2, \mu_2, \sigma_2$
$\mathscr{S}^t_{\rm ave}$
z = 0
z

Figure 12-14 Incident, reflected, and transmitted Poynting vectors at the planar interface between two dissimilar media.

Figure 12-14 shows a plane wave propagating towards an interface between two media. We will assume that region 1 is lossless (i.e., $\sigma_1 = 0$), but region 2 may have loss. Knowing that the total fields in the incident region are the sums of the incident and reflected fields, we can express the total E- and H-fields in this region as

$$\mathbf{E}^1 = E^i(e^{-j\beta_1 z} + \Gamma e^{j\beta_1 z})\hat{\mathbf{a}}_x,$$

and

$$\mathbf{H}^1 = \frac{E^i}{\eta_1}(e^{-j\beta_1 z} - \Gamma e^{j\beta_1 z})\hat{\mathbf{a}}_y,$$

where we note that η_1 is a real number, since we have assumed that the incident medium is lossless. Using Equations (12.103) and (12.104), we see that the average Poynting vector $\mathscr{S}^{(1)}_{\rm ave}$ in this region is

$$\begin{aligned}\mathscr{S}^{(1)}_{\rm ave} &= \frac{1}{2}\,\mathrm{Re}[\mathbf{E}^1 \times (\mathbf{H}^1)^*] \\ &= \frac{|E^i|^2}{2}\,\mathrm{Re}\left[(e^{-j\beta_1 z} + \Gamma e^{j\beta_1 z})\left(\frac{e^{j\beta_1 z}}{\eta_1^*} - \frac{\Gamma^* e^{-j\beta_1 z}}{\eta_1^*}\right)\right]\hat{\mathbf{a}}_z \\ &= \frac{|E^i|^2}{2}\,\mathrm{Re}\left[\frac{1}{\eta_1^*} + \frac{\Gamma e^{j2\beta_1 z}}{\eta_1^*} - \frac{\Gamma^* e^{-j2\beta_1 z}}{\eta_1^*} - \frac{|\Gamma|^2}{\eta_1^*}\right]\hat{\mathbf{a}}_z. \end{aligned} \tag{12.125}$$

Evaluating this expression at the interface ($z = 0$), and noting that the cross-coupling term $(\Gamma - \Gamma^*)/\eta_1{}^*$ is imaginary when η_1 is real, we find that

$$\mathscr{S}^{(1)}_{\rm ave}\Big|_{z=0} = \frac{1}{2}\frac{|E^i|^2}{\eta_1}\hat{\mathbf{a}}_z - \frac{1}{2}\frac{|\Gamma E^i|^2}{\eta_1}\hat{\mathbf{a}}_z \qquad [\mathrm{W/m^2}].$$

Taking a closer look at this expression, we see that $\mathscr{S}^{(1)}_{\rm ave}|_{z=0}$ is simply the sum of the average Poynting vectors of the incident and reflected waves; that is,

$$\mathscr{S}^{(1)}_{\rm ave}\Big|_{z=0} = \mathscr{S}^i_{\rm ave} + \mathscr{S}^r_{\rm ave} \qquad (\sigma_1 = 0), \tag{12.126}$$

where

$$\mathscr{P}^i_{\text{ave}} = \frac{1}{2}\frac{|E^i|^2}{\eta_1}\hat{\mathbf{a}}_z \qquad [\text{W/m}^2], \tag{12.127}$$

and

$$\mathscr{P}^r_{\text{ave}} = -\frac{1}{2}\frac{|\Gamma E^i|^2}{\eta_1}\hat{\mathbf{a}}_z = -|\Gamma|^2\mathscr{P}^i_{\text{ave}} \qquad [\text{W/m}^2]. \tag{12.128}$$

The average Poynting vector $\mathscr{P}^{(2)}_{\text{ave}}$ on the transmission side of the interface is much simpler, since only the transmitted fields are present. Using Equations (12.103) and (12.104), we find that

$$\mathscr{P}^{(2)}_{\text{ave}} = \frac{|E^i|^2}{2}|T|^2\text{Re}\left[e^{-\alpha_2 z}e^{-j\beta_2 z}\left(\frac{e^{-\alpha_2 z - j\beta_2 z}}{\eta_2}\right)^*\right]\hat{\mathbf{a}}_z$$

$$= \frac{|E^i|^2|T|^2}{2|\eta_2|}e^{-2\alpha_2 z}\cos\theta_\eta\,\hat{\mathbf{a}}_z,$$

where θ_η is the angle of the intrinsic impedance η_2. As we can see from this expression, power density decays exponentially with increasing values of z inside the transmission medium if this medium is lossy. Evaluating the expression at $z = 0$, we obtain the net average power density $\mathscr{P}^t_{\text{ave}}$ injected into the transmission medium,

$$\mathscr{P}^t_{\text{ave}} \equiv \mathscr{P}^{(2)}_{\text{ave}}\Big|_{z=0} = \frac{|E^i|^2|T|^2}{2|\eta_2|}\cos\theta_\eta\,\hat{\mathbf{a}}_z \qquad [\text{W/m}^2]. \tag{12.129}$$

According to the law of power conservation, the average power injected into the interface from the incident medium must equal the average power coming out of the interface in the transmitted region. In terms of the average Poynting vectors, this means that

$$\mathscr{P}^{(1)}_{\text{ave}}\Big|_{z=0} = \mathscr{P}^{(2)}_{\text{ave}}\Big|_{z=0}.$$

Substituting Equations (12.126) and (12.129) into this expression, we obtain

$$\mathscr{P}^i_{\text{ave}} + \mathscr{P}^r_{\text{ave}} = \mathscr{P}^t_{\text{ave}} \qquad (\sigma_1 = 0), \tag{12.130}$$

$$|\mathscr{P}^i_{\text{ave}}| = |\mathscr{P}^r_{\text{ave}}| + |\mathscr{P}^t_{\text{ave}}| \qquad (\sigma_1 = 0). \tag{12.131}$$

Also, by substituting Equations (12.127), (12.128), and (12.129) into Equation (12.131), we obtain the following relationship between the reflection and transmission coefficients:

$$|\Gamma|^2 + \frac{\eta_1|T|^2}{|\eta_2|}\cos\theta_\eta = 1 \qquad (\sigma_1 = 0). \tag{12.132}$$

In words, Equations (12.130) and (12.131) state that power transported through a lossless medium towards an interface divides itself between the reflected and transmitted waves. This is a very convenient result, since it allows us to calculate the power density transmitted across the interface simply by subtracting the power density of the reflected wave from the power density of the incident wave. It should be noted, however, that the power densities of the incident, reflected, and transmitted waves are not so simply related when the incident medium is lossy, since the real part of the cross-coupling term in Equation (12.125) no longer vanishes.

Taking a closer look at Equations (12.128) and (12.129), we can see why it is not a violation of the conservation of energy for the transmission coefficient T to have a value greater than unity. Noting that $|\Gamma| \approx 1$ and $|T| \approx 2$ when $\eta_2 \to \infty$, we have

$$|\mathscr{S}^t_{\text{ave}}| = \frac{|E^i|^2|T|^2}{2|\eta_2|}\cos\theta_\eta \approx 0 \quad (\eta_2 \to \infty)$$

$$|\mathscr{S}^r_{\text{ave}}| = \frac{1}{2}\frac{|\Gamma E^i|^2}{\eta_1} \approx |\mathscr{S}^{\,i}_{\text{ave}}| \quad (\eta_2 \to \infty).$$

Here we see that even though the transmitted E-field has twice the magnitude of the incident field, very little power is transmitted, since the impedance of the medium is high. As a result, nearly all the incident power is reflected by the surface.

12-7-3 REFLECTION AND ABSORPTION OF WAVES FROM CONDUCTING SURFACES

An important property of metals is that they are good reflectors of waves. This allows metal surfaces to either confine waves to some area of space or send them off in another direction. A wide range of practical devices make use of this property, including transmission lines, waveguides, resonant cavities, and antennas.

The easiest metal surfaces to model are perfect conductors, which have $\sigma \to \infty$. Figure 12-15 shows a plane wave, normally incident upon a perfectly conducting slab. From Equation (12.69), the intrinsic impedance of the conductor is

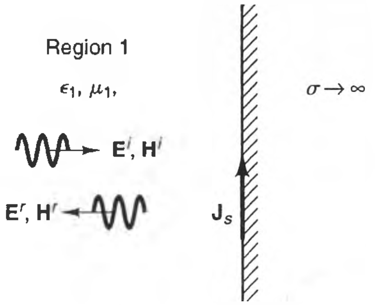


Figure 12-15 Incident and reflected waves at a perfectly conducting slab.

$$\eta = \lim_{\sigma \to \infty} \sqrt{\frac{j\omega\mu}{\sigma + j\omega\epsilon}} = 0.$$

Using this value, we obtain the following reflection and transmission coefficients:

$$\Gamma = -1 \tag{12.133}$$

$$T = 0. \tag{12.134}$$

Since $|\Gamma| = 1$, the incident and reflected fields have the same magnitude, so all the incident power is reflected by the surface, regardless of its thickness.

In addition to the wave reflected from the surface of a perfect conductor, a surface current is induced that flows within an infinitesimal depth of the conductor. This follows from the boundary conditions, since the H-field is zero inside the conductor and nonzero just outside it. Using Equation (10.83), we see that the current density on the conductor surface is given by

$$\mathbf{J}_s = \hat{\mathbf{a}}_n \times \mathbf{H}_s = 2\hat{\mathbf{a}}_n \times \mathbf{H}^i \quad \text{[A/m]} \quad \text{(Perfectly conducting surface)}, \tag{12.135}$$

where $\mathbf{H}^i$ and $\mathbf{H}_s = 2\mathbf{H}^i$ are the incident and total H-fields at the surface, respectively, and the unit vector $\hat{\mathbf{a}}_n$ points outward from the conductor surface.

When the conductor is lossy, the E-field transmitted into the conductor is small, but definitely not zero. This case is depicted in Figure 12-16, where a linearly polarized plane wave is incident from a perfect dielectric towards a lossy slab with conductivity σ. Here, the E-field in the dielectric is a nearly perfect standing wave, since the reflected fields have almost the same magnitudes as the incident fields. At the interface, the transmitted E-field has a magnitude equal to the difference between the incident and reflected E-fields. Because the transmitted fields decay exponentially, we can, for all practical purposes, assume that these fields exist only within a few skin depths δ of the surface, where

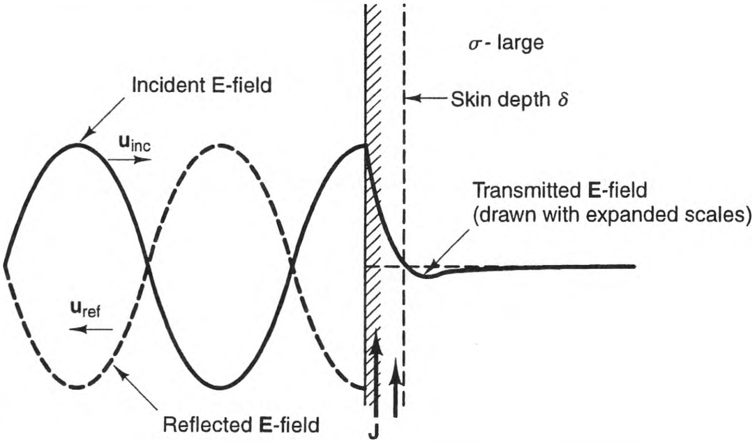


Figure 12-16 The incident, reflected, and transmitted fields near the surface of a lossy, conducting surface.

$$\delta = \frac{1}{\alpha} = \frac{1}{\sqrt{\pi f \mu \sigma}}. \tag{12.136}$$

Also, the wavelength in the slab is much smaller than it is in the dielectric, since β is large in a good conductor.

We can calculate the average power P that is dissipated per square meter of the conductor surface by using Joule's law (Equation (12.97)). The time-averaged power density for a sinusoidal field is $\frac{1}{2}\,\mathrm{Re}[\mathbf{E} \bullet \mathbf{J}^*]$, so we can integrate the power density over the thickness of the slab to find

$$P = \frac{1}{2}\,\mathrm{Re}\left[\int_0^\infty \mathbf{E} \bullet \mathbf{J}^*\, dz\right] \qquad [\mathrm{W/m^2}]. \tag{12.137}$$

In this expression, we have replaced the actual slab thickness with infinity, which is a good approximation as long as the slab is at least several skin depths thick. Most metals are isotropic, so **E** and **J** point in the same direction. This means that the dot product $\mathbf{E} \bullet \mathbf{J}^*$ is equal to EJ^*, where E and J are the components of **E** and **J** along the direction of polarization. Knowing that $J = \sigma E$ and $E = \eta H$ inside the conductor, we can write the current density as

$$J = \sigma E = \sigma \eta H = \sigma \eta H_s e^{-\alpha z} e^{-j\beta z} \qquad [\mathrm{A/m^2}], \tag{12.138}$$

where H_s is the value of the total H-field at the surface, η and σ are the intrinsic impedance and conductivity of the metal, and α and β are the attenuation and phase constants in the metal. Substituting Equation (12.138) into Equation (12.137) and integrating, we obtain

$$P = \frac{1}{2}\int_0^\infty \sigma |\eta|^2 |H_s|^2 e^{-2\alpha z}\, dz = \frac{\sigma |\eta|^2 |H_s|^2}{4\alpha}. \tag{12.139}$$

This can be simplified by using the approximate expressions for η and α (Equations (12.90) and (12.89), respectively), yielding the result

$$P \approx \frac{1}{2} R_s |H_s|^2 \qquad [\mathrm{W/m^2}], \tag{12.140}$$

where R_s is called the ***surface resistance*** of the slab and is given by

$$R_s = \frac{1}{\sigma\delta} = \sqrt{\frac{\omega\mu}{2\sigma}} \qquad [\Omega]. \tag{12.141}$$

Rather than flowing as a surface current, as it does when the conductivity σ is infinite, the current in a lossy conductor flows as a volumetric current that decays exponentially from the surface. If we denote the total current per unit length flowing beneath the surface as I, we have from Equation (12.138) that

$$I = \int_0^\infty J dz = \int_0^\infty \sigma\eta H_s e^{-\alpha z} e^{-j\beta z}\, dz = \frac{\sigma\eta}{\alpha + j\beta} H_s \approx H_s,$$

since $(\sigma\eta)/(\alpha + j\beta) \approx 1$ for a good conductor. (See Equations (12.88) and (12.90)). Also, $H_s \approx 2H^i = J_s$, where J_s is the surface current density that would flow if the conductivity were infinite. Thus, we find that

$$I \approx J_s \qquad [\text{A/m}], \tag{12.142}$$

which shows that the total current flowing beneath the surface of a lossy conductor is the same as that which flows within an infinitesimal depth of a perfect conductor. Because of this, J_s can be considered as the equivalent surface current density for the volumetric current in a lossy conductor.

By substituting Equation (12.142) into (12.140), we can express the dissipated power density at a lossy conductor surface in terms of the product of the surface resistance R_s and the square of the equivalent surface current density J_s:

$$P \approx \frac{1}{2} R_s |J_s|^2 \qquad [\text{W/m}^2]. \tag{12.143}$$

This expression shows that the dissipated power can be thought of as occurring in an equivalent resistance R_s through which the surface current flows. This resistance is a function of the resistivity of the conductor, σ^{-1}, and the depth of penetration of the current into the slab.

Even though the current actually decays exponentially within the conductor, it is instructive to see what power would be dissipated if the same amount of current flowed uniformly within exactly one skin depth. In this case, the volumetric current density would be J_s/δ, so the power dissipated would be

$$P = \frac{1}{2}\int_0^\delta \frac{1}{\sigma}\left[\frac{J_s}{\delta}\right]^2 dz = \frac{1}{2}\frac{J_s^2}{\sigma\delta^2}\delta.$$

But since $R_s = 1/(\sigma\delta)$, we obtain

$$P \approx \frac{1}{2} R_s |J_s|^2,$$

which is the same dissipated power as when the current decays exponentially.

Finally, even though Equations (12.140) and (12.143) were derived for the case of a normally incident plane wave, they are excellent approximations even when the incident waves are incident at oblique angles to the surface. Since the conductor impedance is much smaller than the dielectric impedance, the direction of propagation inside the conductor is nearly always approximately normal to the surface, regardless of the angle of incidence.[9] Because of this, Equations (12.140) and (12.143) can be used to find the power dissipated by metal reflectors and enclosures, regardless of nature of the incident fields.

[9] This is a consequence of Snell's law of refraction, which is discussed later in the chapter.

Example 12-9

Calculate the power dissipated per square meter beneath the surface shown in Figure 12-17, where a time-harmonic plane wave is normally incident upon a copper plate. The frequency of the wave is 1.0 [GHz] and the peak amplitude of the E-field is 700 [V/m].

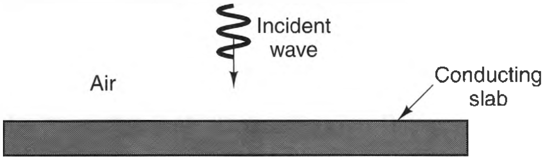


Figure 12-17 A plane wave normally incident from air upon a lossy, conducting slab.

Solution:

From Table C-2, the conductivity of copper is $\sigma = 5.8 \times 10^7$ [S/m]. Using Equations (12.136) and (12.141), we find that the skin depth δ and surface resistance R_s are

$$\delta = \sqrt{\frac{1}{\pi f \mu \sigma}} = \left[\frac{1}{(\pi \times 10^9) \times (4\pi \times 10^{-7}) \times (5.8 \times 10^7)}\right]^{1/2} = 2.1 \qquad [\mu\text{m}].$$

$$R_s = \frac{1}{(2.1 \times 10^{-6}) \times (5.8 \times 10^7)} = 8.25 \qquad [\text{m}\Omega].$$

The peak incident H-field is,

$$H^i = \frac{700\,[\text{V/m}]}{377\,[\Omega]} = 1.86 \qquad [\text{A/m}].$$

Since σ is large, the total H-field at the surface is approximately $2 \times H^i = 3.71$ [A/m]. Thus, from Equation (12.140), we obtain

$$P = \frac{1}{2}(8.25 \times 10^{-3}) \times (3.71)^2 = 56.9 \qquad [\text{mW/m}^2].$$

12-7-4 REFLECTIONS AND TRANSMISSIONS FROM LAYERED MEDIA

There are many situations where a plane wave passes through several layers of different materials. Two common examples are the passage of light through a lens with multiple optical coatings and the waves generated by a ground-probing radar as they pass through various strata in the ground. To model these situations, the reflected and transmitted waves from each interface must be accounted for. There are two ways that this can be accomplished. In the first, each reflection and transmission is dealt with one at a time. The second method uses an impedance transformation that lumps multiple reflections into a single parameter. Both methods produce correct results, but differ in the work they entail and the amount of information they provide.

Multiple Reflections. Figure 12-18 shows a stack of three different lossless, dielectric regions that form planar interfaces at $z = 0$ and $z = \ell$, respectively. All three regions are of infinite extent in the x- and y-directions, and the second region has thick-

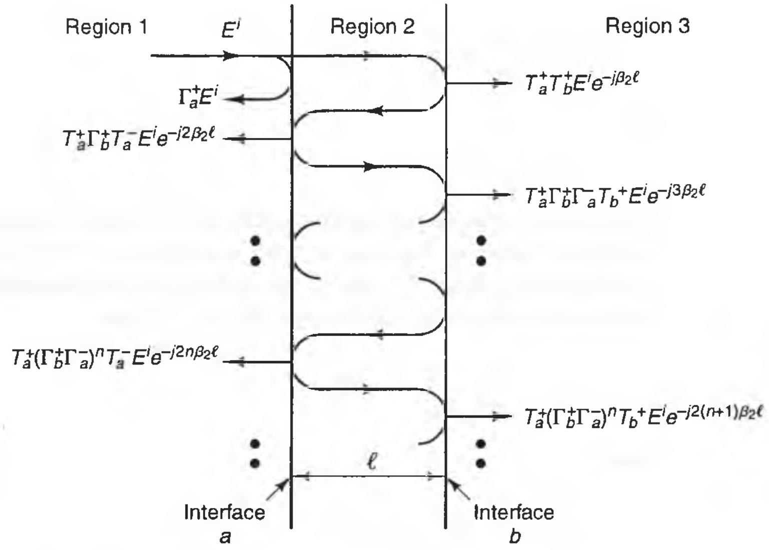

Region 1
E^i
Region 2
Region 3
$T_a^+T_b^+E^ie^{-j\beta_2\ell}$
$\Gamma_a^+E^i$
$T_a^+\Gamma_b^+T_a^-E^ie^{-j2\beta_2\ell}$
$T_a^+\Gamma_b^+\Gamma_a^-T_b^+E^ie^{-j3\beta_2\ell}$
$T_a^+(\Gamma_b^+\Gamma_a^-)^nT_a^-E^ie^{-j2n\beta_2\ell}$
$T_a^+(\Gamma_b^+\Gamma_a^-)^nT_b^+E^ie^{-j2(n+1)\beta_2\ell}$
ℓ
Interface
a
Interface
b

Figure 12-18 Multiple reflections and transmissions of a plane wave incident upon the planar interfaces between three dissimilar media.

ness ℓ along the z-direction. When a plane wave $\mathbf{E}^i$ is normally incident from the left ($z < 0$), an infinite number of transmitted and reflected waves are produced at the interfaces. However, all the waves reflected towards the left in region 1 have the same wavelength, so their sum can be considered to be a single reflected wave $\mathbf{E}^r$. Similarly, all the waves transmitted into region 3 combine to form a single transmitted wave $\mathbf{E}^t$. In the analysis that follows, we will develop expressions for $\mathbf{E}^r$ and $\mathbf{E}^t$ by following the progression of the incident field and its reflections.

Let us start our analysis by assuming that the incident field can be represented as

$$\mathbf{E}^i = E^ie^{-j\beta z}\hat{\mathbf{a}}_x.$$

When this wave arrives at the $z = 0$ interface (interface a), a reflected wave of amplitude $\Gamma_a^+E^i$ and a transmitted wave of amplitude $T_a^+E^i$ are produced, where Γ_a^+ and T_a^+ are reflection and transmission coefficients, given by

$$\Gamma_a^+ = \frac{\eta_2 - \eta_1}{\eta_2 + \eta_1},$$

and

$$T_a^+ = 1 + \Gamma_a^+ = \frac{2\eta_2}{\eta_2 + \eta_1}.$$

In these formulas, the subscript a indicates coefficients that are associated with interface a, and the superscript "+" indicates coefficients that correspond to waves propagating in the $+z$ direction (i.e., from region 1 to region 2).

While propagating through region 2, the first transmitted wave undergoes a phase delay of $e^{-j\beta_2\ell}$. When this wave strikes interface b, transmitted and reflected waves are produced, with amplitudes of $T_a^+T_b^+E^ie^{-j\beta_2\ell}$ and $T_a^+\Gamma_b^+E^ie^{-j\beta_2\ell}$, respectively, where

$$\Gamma_b^+ = \frac{\eta_3 - \eta_2}{\eta_3 + \eta_2},$$

and

$$T_b^+ = 1 + \Gamma_b^+.$$

The reflected wave propagates backward in region 2 and produces new transmitted and reflected waves at interface a, with amplitudes $T_a^+ \Gamma_b^+ T_a^- E^i e^{-j2\beta_2 \ell}$ and $T_a^+ \Gamma_b^+ \Gamma_a^- E^i e^{-j2\beta_2 \ell}$, respectively, where Γ_a^- and T_a^- are reflection and transmission coefficients at interface a for waves propagating towards the left. Thus,

$$\Gamma_a^- = \frac{\eta_1 - \eta_2}{\eta_1 + \eta_2} = -\Gamma_a^+$$

and

$$T_a^- = 1 + \Gamma_a^- = 1 - \Gamma_a^+ = \frac{2\eta_1}{\eta_2 + \eta_1}.$$

By now, the sequence of waves produced by successive reflections from the two interfaces should be obvious. We can determine the net fields in each region by summing the infinite series of rays. For instance, the amplitude E^r of the reflected field in region 1 can be written as

$$E^r = E^i[\Gamma_a^+ + T_a^+ T_a^- \Gamma_b^+ e^{-j2\beta_2 \ell} + T_a^+ T_a^- \Gamma_a^- (\Gamma_b^+)^2 e^{-j4\beta_2 \ell}$$
$$+ T_a^+ T_a^- (\Gamma_a^-)^2 (\Gamma_b^+)^3 e^{-j6\beta_2 \ell} + \cdots],$$

which can also be written as

$$E^r = E^i \Gamma_a^+ + E^i T_a^+ T_a^- \Gamma_b^+ e^{-j2\beta_2 \ell} \sum_{n=1}^{\infty} (\Gamma_a^- \Gamma_b^+ e^{-j2\beta_2 \ell})^{(n-1)}.$$

As complicated as the infinite series in this expression may appear, it is simply a geometric series. From ordinary calculus, it is known that the sum of a geometric series can be written in the form

$$\sum_{n=1}^{\infty} ar^{n-1} = \frac{a}{1-r} \quad |r| < 1.$$

Using this result, we can write the reflected field in region 1 as

$$\mathbf{E}^r(z) = E^r e^{+j\beta_1 z} \hat{\mathbf{a}}_x, \tag{12.144}$$

where

$$E^r = E^i \left[\Gamma_a^+ + \frac{T_a^+ T_a^- \Gamma_b^+ e^{-j2\beta_2 \ell}}{1 - \Gamma_a^- \Gamma_b^+ e^{-j2\beta_2 \ell}} \right]. \tag{12.145}$$

Employing a similar sequence of steps, we can derive an expression for the transmitted field E^t in region 3 ($z > \ell$), which can be written in the form

$$\mathbf{E}^t(z) = E^t e^{-j\beta_1(z-\ell)}\hat{\mathbf{a}}_x, \tag{12.146}$$

where

$$E^t = E^i \frac{T_a^+ T_b^+ e^{-j\beta_2\ell}}{1 - \Gamma_a^- \Gamma_b^+ e^{-j2\beta_2\ell}}. \tag{12.147}$$

If the reflection coefficient Γ is defined as the net reflected wave amplitude to the incident wave amplitude at interface a, we have, from Equations (12.145),

$$\Gamma = \frac{E^r}{E^i} = \Gamma_a^+ + \frac{T_a^+ T_a^- \Gamma_b^+ e^{-j2\beta_2\ell}}{1 - \Gamma_a^- \Gamma_b^+ e^{-j2\beta_2\ell}}, \tag{12.148}$$

Similarly, the transmission coefficient T is defined as the ratio of the transmitted wave amplitude in region 3 to the incident wave amplitude in region 1. Using Equation (12.147) we have

$$T = \frac{E^t}{E^i} = \frac{T_a^+ T_b^+ e^{-j\beta_2\ell}}{1 - \Gamma_a^- \Gamma_b^+ e^{-j2\beta_2\ell}}. \tag{12.149}$$

As can be seen from Equations (12.148) and (12.149), both Γ and T are functions not only of the impedance values of all three media, but also of the thickness of the "sandwiched" medium (medium 2). This is because the phases of the individual reflected and transmitted waves depend on the phase delay across the region.

To see how Γ and T vary with the frequency f of the incident wave, let us consider the case where regions 1 and 3 are both free space. This situation is depicted in Figure 12-19, which shows a dielectric slab of thickness ℓ with free space on both sides. For this case, we have

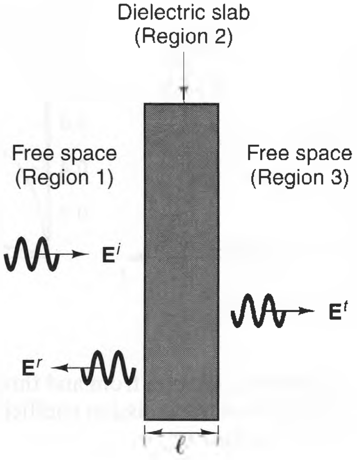


Figure 12-19 Incident, reflected, and transmitted waves at a finite-width, dielectric slab.

$$\begin{aligned}\Gamma_b^+ &= \Gamma_a^- = -\Gamma_a^+,\\ T_a^+T_a^- &= (1+\Gamma_a^+)(1-\Gamma_a^+) = 1-(\Gamma_a^+)^2,\\ T_a^+T_b^+ &= (1+\Gamma_a^+)(1+\Gamma_b^+) = 1-(\Gamma_a^+)^2.\end{aligned}$$

Substituting these into Equations (12.148) and (12.149), we obtain

$$\Gamma = \sqrt{R}\left[1 - \frac{1-R}{1-R\exp(-j4\pi f\tau)}\right], \tag{12.150}$$

and

$$T = \frac{1-R}{1-R\exp(-j4\pi f\tau)}\,e^{-j\beta_2\ell}. \tag{12.151}$$

In both of these expressions, τ is the one-way propagation time across the slab and R is the reflectivity of either side of the slab alone; that is,

$$R = |\Gamma_a^+|^2 = \left|\frac{\sqrt{\epsilon_r}-1}{\sqrt{\epsilon_r}+1}\right|^2, \tag{12.152}$$

where ϵ_r is the dielectric constant of the slab.

Figures 12-20a and b shows plots $|T|$ and $|\Gamma|$ vs. f, respectively, for three different values of R. The curves for $|T|$ and $|\Gamma|$ are complementary, since one parameter is high when the other is low. In particular, note that there are an infinite number of resonant frequencies at which $\Gamma = 0$ and $|T| = 1$. These frequencies are given by

$$f_m = \frac{m}{2\tau} = \frac{mc}{2\ell\sqrt{\epsilon_r}} \qquad m = 1, 2, \ \ldots\ , \tag{12.153}$$

where c is the speed of light in a vacuum and ϵ_r is the dielectric constant of the slab. At each resonant frequency, the wavelength λ_m in the slab is an integral fraction of the slab width:

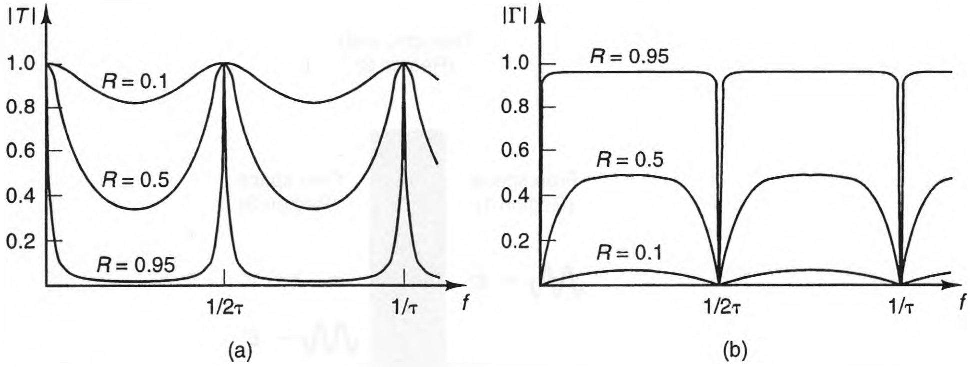


Figure 12-20 Reflection and transmission from and through a finite-width dielectric slab for different reflectivities R: a) Net transmission coefficient $|T|$ vs. frequency. b) Net reflection coefficient $|\Gamma|$ vs. frequency.

$$\lambda_m = \frac{2\ell}{m} \quad m = 1, 2, \ \ldots . \tag{12.154}$$

The slab appears as nonreflecting at these frequencies regardless of the width, since all the waves directed back towards the source sum exactly to zero. When this occurs, the transmitted wave has the same magnitude as the incident wave.

It can also be seen from Figure 12-20 that the off-resonance behavior of the slab is a strong function of the reflectivity R. When R is small, the reflection and transmission coefficients vary only slightly about the values 0 and 1, respectively. However, when R is close to unity (i.e., 100% reflectivity), $\Gamma \approx 1$ and $T \approx 0$ at nearly all frequencies, except the resonant frequencies, where $\Gamma \approx 0$ and $T = 1$. If a wave containing a spread of frequencies is incident upon the slab, the transmitted wave will contain mostly the resonant frequencies if R is relatively large. Because of this characteristic, dielectric slabs are often used as filters, particularly at optical frequencies (which include both visible and infrared frequencies). Slabs used in this way are called ***etalons***.

Figure 12-21 shows a schematic of a tunable wedge etalon filter for use in optical fibers. Here, the narrow optical beam[10] from the fiber is deflected vertically by a voltage controlled deflector. Since the etalon thickness varies along the vertical axis, the frequency passed by the filter varies as the deflector voltage changes. Etalon filters like this are often used to allow optical receivers to "see" only the light emitted at certain frequencies.

Reflection Analysis Using Effective Wave Impedances. It is possible to extend the multiple-reflection technique discussed in the preceding section to model the reflection and transmission through any number of planar interfaces. However, this method becomes quite cumbersome when more than two interfaces are present, since the number of possible reflections grows dramatically in that case. In this section, we will develop a technique that avoids the problem by replacing pairs of regions with a single equivalent medium that has the same reflection properties. With this method, any number of interfaces can be modeled.

To develop the technique, let us return to the three-medium problem discussed in the previous section. This configuration is redrawn in Figure 12-22a. As we saw before, an infinite number of backward-propagating waves are produced in region 1, due to the reflections and transmissions at the interfaces. But since they are all of the same wavelength and frequency, they can be treated as a single wave, with electric and

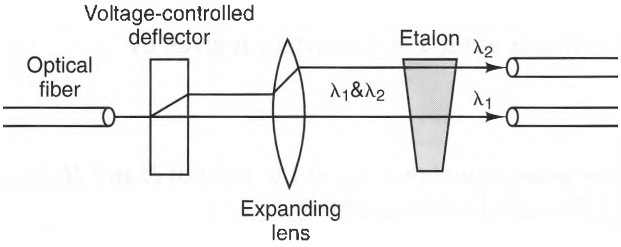


Figure 12-21 An etalon filter system for wavelength-selective routing of signals on optical fibers.

[10] Although the beam emitted by an optical fiber is very narrow, it can be approximated as a plane wave within its beamwidth.

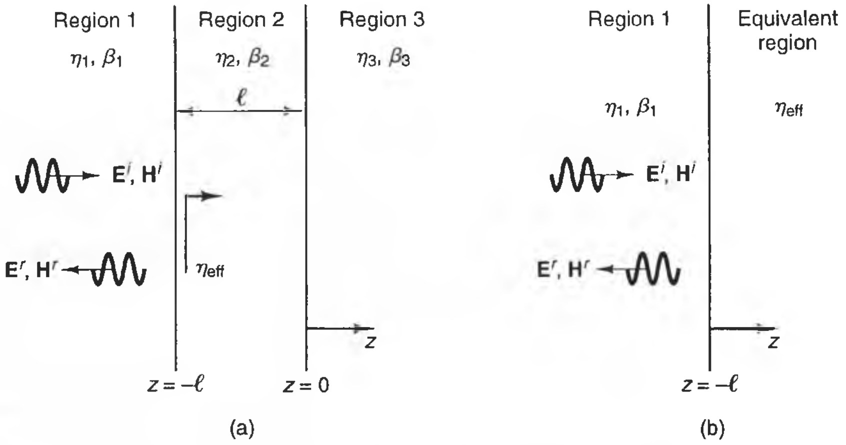


Figure 12-22 Two geometries that have the same net reflection coefficient at $z = -\ell$ for a wave incident from the left: a) A planar, three-medium geometry. b) An equivalent two-medium geometry.

magnetic fields $\mathbf{E}^r$ and $\mathbf{H}^r$, respectively. The same is true of the forward-propagating waves in region 3, which form the waves $\mathbf{E}^t$ and $\mathbf{H}^t$. In the middle region (region 2), where both forward-propagating and backward-propagating waves are present, we can consider them to form the waves $\mathbf{E}^+$ and $\mathbf{H}^+$, and $\mathbf{E}^-$ and $\mathbf{H}^-$, respectively.

To simplify the problem further, let us see if it is possible to produce the same reflected fields $\mathbf{E}^r$ and $\mathbf{H}^r$ from a single interface between region 1 and some yet-to-be-determined medium with intrinsic impedance η_{eff} that fills the entire region $z > -\ell$. This situation is depicted in Figure 12-22b. In order for the substitution to work, we must find a value for η_{eff} that maintains exactly the same ratio of the electric and magnetic fields just to the right of the $z = -\ell$ interface as were there in the original problem (Figure 12-22a). Noting that $\mathbf{E}^-$ and $\mathbf{H}^-$ can be thought of as the reflections of $\mathbf{E}^+$ and $\mathbf{H}^+$ off the $z = 0$ (interface b), we can write

$$E(z) = E^+(e^{-\gamma_2 z} + \Gamma_o e^{+\gamma_2 z}), \tag{12.155}$$

and

$$H(z) = \frac{E^+}{\eta_2}(e^{-\gamma_2 z} - \Gamma_o e^{+\gamma_2 z}), \tag{12.156}$$

where the reflection coefficient at the $z = 0$ interface is given by

$$\Gamma_o = \frac{\eta_3 - \eta_2}{\eta_3 + \eta_2}. \tag{12.157}$$

If we define the ***effective wave impedance*** η_{eff} as the ratio of E and H at $z = -\ell^+$, we can use Equations (12.155) and (12.156) to obtain

$$\eta_{\text{eff}} = \left.\frac{E}{H}\right|_{z=-\ell^+} = \eta_2 \frac{e^{+\gamma_2\ell} + \Gamma_o e^{-\gamma_2\ell}}{e^{+\gamma_2\ell} - \Gamma_o e^{-\gamma_2\ell}}. \tag{12.158}$$

Substituting Equation (12.157) into Equation (12.158) and simplifying the resulting expression, we can write η_{eff} in the form

$$\eta_{\text{eff}} = \eta_2 \frac{\eta_3 + \eta_2 \tanh(\gamma_2 \ell)}{\eta_2 + \eta_3 \tanh(\gamma_2 \ell)} \quad [\Omega]. \tag{12.159}$$

If medium 2 is lossless, $\gamma_2 = j\beta_2$, which means that Equation (12.159) can be simplified to read

$$\eta_{\text{eff}} = \eta_2 \frac{\eta_3 + j\eta_2 \tan(\beta_2 \ell)}{\eta_2 + j\eta_3 \tan(\beta_2 \ell)} \quad [\Omega] \quad (\alpha_2 = 0). \tag{12.160}$$

Comparing Equations (12.159) and (12.160) with the impedance transformation formulas for transmission lines (Equations (11.133) and (11.134), respectively), we see that they are of exactly the same form. Thus, we can interpret η_{eff} as the transformed impedance of region 3 at the point $z = -\ell^+$, just inside region 2.

Returning to Figure 12-22a, if we replace regions 2 and 3 with a homogenous region of intrinsic impedance η_{eff} (given by Equation (12.159) or (12.160)), there will be no change in the fields to the left of $z = -\ell$. This is because the equivalent region maintains exactly the same ratio of the electric and magnetic fields as did the original two regions at $z = -\ell$. The net reflection coefficient at the $z = -\ell$ interface can easily be calculated using the simplified geometry of Figure 12-22b; we obtain

$$\Gamma = \frac{E^r}{E^i} = \frac{\eta_{\text{eff}} - \eta_1}{\eta_{\text{eff}} + \eta_1}. \tag{12.161}$$

It is left as an exercise to the reader to show that Equations (12.161) and (12.148) are equivalent.

In addition to making reflection calculations easier, the effective-wave impedance technique provides additional insight into ways to either reduce or enhance the reflections. For example, suppose it is necessary to transmit a plane wave across an interface between two dissimilar materials. Unless some sort of impedance-matching technique is used, the impedance mismatch between the two materials will give rise to a reflection, which, in turn, reduces the power that can be transported across the interface. To remedy this, consider the quarter-wave transformer shown in Figure 12-23, which consists of a quarter-wavelength layer that is sandwiched between two dissimilar materials.

To find the net reflection coefficient Γ of this geometry, the quarter-wavelength layer and the right-hand medium can be replaced by a single medium of intrinsic impedance η_{eff}, given by Equation (12.160). Noting that $\tan(\beta\ell) \to \infty$ when $\ell = \lambda/4$, we find that

$$\eta_{\text{eff}} = \eta \frac{\eta_2 + j\eta \tan(\beta\ell)}{\eta + j\eta_2 \tan(\beta\ell)} \to \frac{\eta^2}{\eta_2} \text{ (when } \ell = \lambda/4\text{)},$$

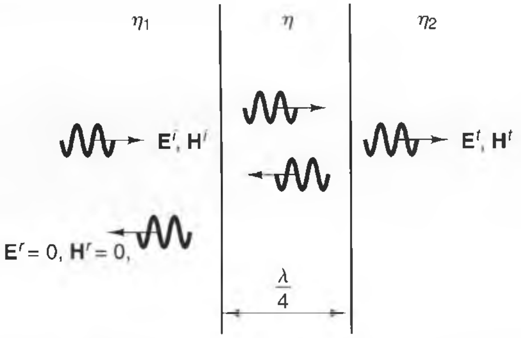

η_1
η
η_2
$\mathbf{E}^i, \mathbf{H}^i$
$\mathbf{E}^t, \mathbf{H}^t$
$\mathbf{E}^r = 0, \mathbf{H}^r = 0,$
$\frac{\lambda}{4}$

Figure 12-23 A quarter-wave transformer.

where η and η_2 are the intrinsic impedances of the quarter-wave section and the right-hand medium, respectively. To determine what value of η achieves a net zero reflection at the leftmost interface, we simply set η_{eff} equal to η_1, yielding

$$\eta = \sqrt{\eta_1 \eta_2} \qquad \text{(Quarter-wave transformer).} \qquad (12.162)$$

This quarter-wave transformer is analogous to the quarter-wave transformers discussed in the previous chapter for use on transmission lines. In both cases, they transform the impedance of a load to the impedance level of the incoming wave. Quarter-wave transformers are often used as antireflection coatings on optical lenses.

It is easy to apply this effective impedance technique to larger numbers of interfaces. Consider the situation depicted in Figure 12-24a, which shows a plane wave incident upon three parallel interfaces. Here, the interfaces are identified by the letters a, b, and c, starting with the rightmost interface. To find the net reflection coefficient for these interfaces, we first calculate the effective impedance η_b, as seen from interface c, looking towards interface b; we get

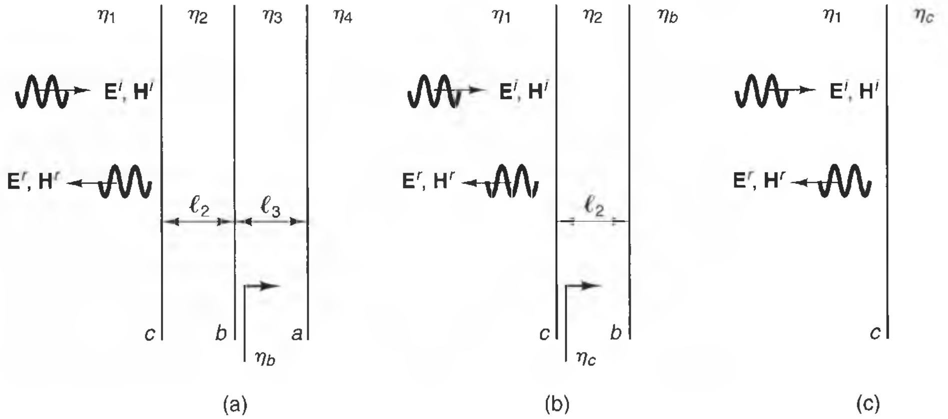

η_1
η_2
η_3
η_4
$\mathbf{E}^i, \mathbf{H}^i$
$\mathbf{E}^r, \mathbf{H}^r$
ℓ_2
ℓ_3
c
b
a
η_b
(a)
η_1
η_2
η_b
$\mathbf{E}^i, \mathbf{H}^i$
$\mathbf{E}^r, \mathbf{H}^r$
ℓ_2
c
b
η_c
(b)
η_1
η_c
$\mathbf{E}^i, \mathbf{H}^i$
$\mathbf{E}^r, \mathbf{H}^r$
c
(c)

Figure 12-24 Reflection from three planar interfaces: a) Original four-medium geometry. b) Equivalent three-medium geometry. c) Equivalent two-medium geometry.

$$\eta_b = \eta_3 \frac{\eta_4 + j\eta_3 \tan(\beta_3 \ell_3)}{\eta_3 + j\eta_4 \tan(\beta_3 \ell_3)},$$

where ℓ_3 is the width of region 3. Replacing regions 3 and 4 with a single region of intrinsic impedance η_b, we obtain the geometry shown in Figure 12-24b. We can then eliminate the next interface by calculating the effective impedance η_c, just to the right of the interface c. Using the impedance transformation formula, we obtain

$$\eta_c = \eta_2 \frac{\eta_b + j\eta_2 \tan(\beta_2 \ell_2)}{\eta_2 + j\eta_b \tan(\beta_2 \ell_2)},$$

where ℓ_2 is the width of region 2. We can now replace the two media at the far right of Figure 12-24b with a single medium of impedance η_c, which results in the geometry shown in Figure 12-24c. Since this final geometry is a simple interface between two media, the net reflection coefficient is

$$\Gamma = \frac{E^r}{E^i} = \frac{\eta_c - \eta_1}{\eta_c + \eta_1}.$$

By now, the procedure for modeling any number of parallel dielectric interfaces should be obvious: Remove the interfaces one by one, starting with the interface farthest from the incident wave. By working sequentially towards the incident wave, any number of interfaces can be modeled. Comparing this technique with the multiple-reflection method discussed in the previous section, we see that this technique is easier, but does not provide information about the fields inside any of the media except the incident medium. This is not a problem, however, if one is interested only in the net reflection coefficient seen by the incident field.

Example 12-10

Calculate the net reflection coefficient Γ for the plane wave incident upon the stratified media shown in Figure 12-25. Assume that the dielectric constants of the media are $\epsilon_{r_1} = 1$, $\epsilon_{r_2} = 6.25$, $\epsilon_{r_3} = 2.25$, and $\epsilon_{r_4} = 1$. Also, assume that $\ell_2 = \lambda_2/8$ and $\ell = \lambda_3/5$.

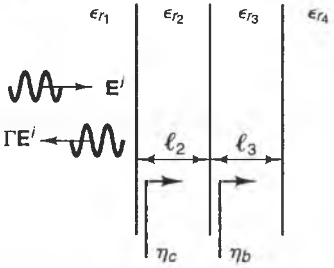


Figure 12-25 Incident and reflected waves at a stack of three interfaces of dissimilar materials.

Solution:

Using the specified dielectric constants, we first calculate the following parameters:

$$\eta_1 = \eta_4 = \frac{377}{\sqrt{1}} = 377 \qquad [\Omega]$$

$$\eta_2 = \frac{377}{\sqrt{6.25}} = 150.8 \qquad [\Omega]$$

$$\eta_3 = \frac{377}{\sqrt{2.25}} = 251.33 \qquad [\Omega]$$

$$\tan(\beta_3\ell_3) = \tan\left[\frac{2\pi}{\lambda_3} \times \frac{\lambda_3}{5}\right] = \tan(2\pi/5) = 3.08$$

$$\tan(\beta_2\ell_2) = \tan(2\pi/8) = 0.785.$$

The effective wave impedance η_b just to the right of the interface between regions 2 and 3 can be found from Equation (12.160):

$$\eta_b = \eta_3 \frac{\eta_4 + j\eta_3 \tan(\beta_3\ell_3)}{\eta_3 + j\eta_4 \tan(\beta_3\ell_3)}$$

$$= 251.33 \frac{377 + j(251.33)(3.08)}{251.33 + j(377)(3.08)} = 176.94 - j43.34.$$

Next, the effective wave impedance η_c just to the right of the interface between regions 1 and 2 is

$$\eta_c = \eta_2 \frac{\eta_b + j\eta_2 \tan(\beta_2\ell_2)}{\eta_2 + j\eta_b \tan(\beta_2\ell_2)}.$$

$$= 150.8 \frac{(176.94 - j43.34) + j(150.8)(0.785)}{150.8 + j(176.94 - j43.34)(0.785)}$$

$$= 121.66 - j30.2 \qquad [\Omega].$$

Finally, the effective reflection coefficient is

$$\Gamma = \frac{\eta_c - \eta_1}{\eta_c + \eta_1} = \frac{121.66 - j30.2 - 377}{121.66 - j30.2 + 377}$$

$$= 0.515\angle -169.8^\circ.$$

This results in a power reflection coefficient of

$$|\Gamma|^2 = 0.265 = 26.5\%.$$

12-8 Plane-Wave Reflection and Transmission at Planar Boundaries, Oblique Incidence

Up to this point, we have discussed the reflection and transmission of plane waves when the direction of propagation is normal to the interfaces. We will now discuss the more general case where the incident wave strikes the interface at a nonperpendicular (i.e., oblique) angle. This analysis is more cumbersome for oblique incidence than for normal incidence, so we will restrict our comments to single interfaces of lossless media. We will start by considering the case where the incident electric-field vector is parallel to the interface. This will be followed by the case where the incident H-field vector is parallel to the interface.

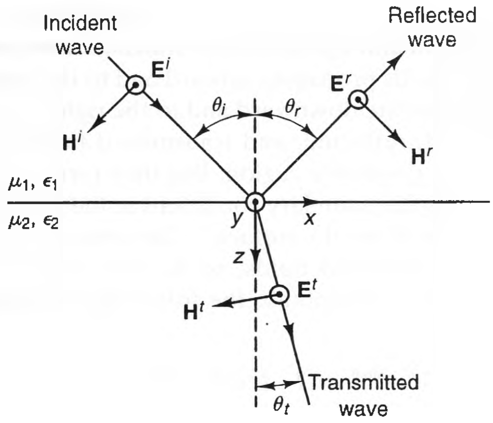


Figure 12-26 Reflection and transmission of an obliquely incident, perpendicularly polarized plane wave at a planar interface.

12-8-1 PERPENDICULAR POLARIZATION

Figure 12-26 shows a plane wave incident upon the interface between two lossless media, characterized by μ_1 and ϵ_1, and μ_2 and ϵ_2, respectively. The incident field $\mathbf{E}^i$ is y polarized and propagates downward and to the right in medium 1. Using Equation (12.35), we can represent $\mathbf{E}^i$ by

$$\mathbf{E}^i = E^i \hat{\mathbf{a}}_y e^{-jk_1(z\cos\theta_i + x\sin\theta_i)}, \tag{12.163}$$

where θ_i is the angle between the surface normal and the direction of propagation. If we define the ***plane of incidence*** as the plane containing the surface normal and the propagation vector of the incident field, we see that $\mathbf{E}^i$ is perpendicular to that plane. Hence, this choice of polarization is called ***perpendicular polarization.***[11] Also, from Equation (12.36), the magnetic field

$$\mathbf{H}^i = \frac{E^i}{\eta_1}(-\cos\theta_i\,\hat{\mathbf{a}}_x + \sin\theta_i\,\hat{\mathbf{a}}_z)e^{-jk_1(z\cos\theta_i + x\sin\theta_i)}. \tag{12.164}$$

Because of the impedance change at the interface, both reflected and transmitted fields must also be present. These fields must have the same polarization as the incident field in order to satisfy the boundary conditions at the interface. We can represent the reflected and transmitted fields using the expressions

$$\mathbf{E}^r = \Gamma_\perp E^i \hat{\mathbf{a}}_y e^{-jk_1(-z\cos\theta_r + x\sin\theta_r)} \tag{12.165}$$

$$\mathbf{H}^r = \frac{\Gamma_\perp E^i}{\eta_1}(\cos\theta_r\hat{\mathbf{a}}_x + \sin\theta_r\hat{\mathbf{a}}_z)e^{-jk_1(-z\cos\theta_r + x\sin\theta_r)}, \tag{12.166}$$

$$\mathbf{E}^t = T_\perp E^{-i}\hat{\mathbf{a}}_y e^{-jk_2(z\cos\theta_t + x\sin\theta_t)} \tag{12.167}$$

$$\mathbf{H}^t = \frac{T_\perp E^i}{\eta_2}(-\cos\theta_t\hat{\mathbf{a}}_x + \sin\theta_t\hat{\mathbf{a}}_z)e^{-jk_2(z\cos\theta_t + x\sin\theta_t)}, \tag{12.168}$$

[11] Other names are horizontal polarization, E polarization, and TE (trasverse electric) polarization.

where the superscripts r and t denote reflected and transmitted fields, respectively, and $\Gamma_\perp$ and $T_\perp$ are the ***perpendicular reflection*** and ***transmission coefficients***, respectively. Notice that the reflected fields propagate upward and to the right at an angle θ_r, while the transmitted fields propagate downward and to the right at an angle θ_t.

Each of the incident, reflected, and transmitted fields defined above satisfies Maxwell's equations in its respective region. But they form a correct solution for this problem only if they satisfy the boundary conditions at the interface, which require that $E_{\tan}$ and $H_{\tan}$ be continuous across the surface. The total electric field in region 1 is the sum of the incident and reflected fields, so we can evaluate Equations (12.163), (12.165), and (12.167) at $z = 0$ to obtain the following expression for the tangential electric fields:

$$E^i e^{-jk_1 x \sin\theta_i} + \Gamma_\perp E^i e^{-jk_1 x \sin\theta_r} = T_\perp E^i e^{-jk_2 x \sin\theta_t}. \tag{12.169}$$

Also, the tangential component of H-field is the x-component, so we must evaluate Equations (12.164), (12.166), and (12.168) at $x = 0$ to obtain

$$-\frac{E^i}{\eta_1}\cos\theta_i e^{-jk_1 x \sin\theta_i} + \frac{\Gamma_\perp E^i}{\eta_1}\cos\theta_r e^{-jk_1 x \sin\theta_r}$$
$$= -\frac{T_\perp E^i}{\eta_2}\cos\theta_t e^{-jk_2 x \sin\theta_t}. \tag{12.170}$$

To satisfy Equations (12.169) and (12.170), we must find the appropriate reflection and transmission coefficients, $\Gamma_\perp$ and $T_\perp$, respectively. However, in order for these values to be independent of z, the three exponential terms in both expressions must all be identical functions of x. Hence, we must require that

$$e^{-jk_1 x \sin\theta_i} = e^{-jk_1 x \sin\theta_r} = e^{-jk_2 x \sin\theta_t},$$

which is satisfied when

$$k_1 \sin\theta_i = k_1 \sin\theta_r = k_2 \sin\theta_t. \tag{12.171}$$

This equation is often called the *surface phase-matching requirement.*

Equation (12.171) specifies a unique relationship between the incident, reflected, and transmitted angles. The equality of the first two terms is called ***Snell's law of reflection*** and states that the angle of incidence equals the angle of reflection:

$$\theta_i = \theta_r. \tag{12.172}$$

The relationship between θ_i and θ_t that is specified in Equation (12.171) is called ***Snell's law of refraction.*** This law states that the angle of transmission θ_t is given by

$$\sin\theta_t = \frac{k_1}{k_2}\sin\theta_i. \tag{12.173}$$

The law is often written in the form

$$\sin\theta_t = \frac{n_1}{n_2}\sin\theta_i, \tag{12.174}$$

where n_1 and n_2 are the ***indices of refraction*** of the two media. The index of refraction of a medium (lossless or lossy) is defined as the ratio of the speed of light in a vacuum to the phase velocity in the medium.

$$n \equiv \frac{c}{u_p}. \tag{12.175}$$

In lossless, nonmagnetic media, however, $u_p = 1/\sqrt{\mu_o \epsilon}$. This means that n equals the square root of the relative permittivity in lossless, nonmagnetic materials:

$$n = \sqrt{\epsilon_r} \qquad \text{(Lossless, nonmagnetic materials).} \tag{12.176}$$

A medium with a large index of refraction is called a dense medium, since the dielectric constant ϵ_r is usually large when the number of atoms per unit volume is high. Conversely, a medium with a small index of refraction is called a rare medium. The rarest medium is a vacuum (free space), where $n = 1$.

We can shed more light on the meaning of Snell's laws by observing the constant-phase planes shown in Figure 12-27. Here, we see that the distance between the incident and reflected phase planes (shown as dotted lines) are the same, since both waves have the same wavelength. This means that θ_i and θ_r must be equal in order for these phase planes to track each other along the interface. On the other hand, the

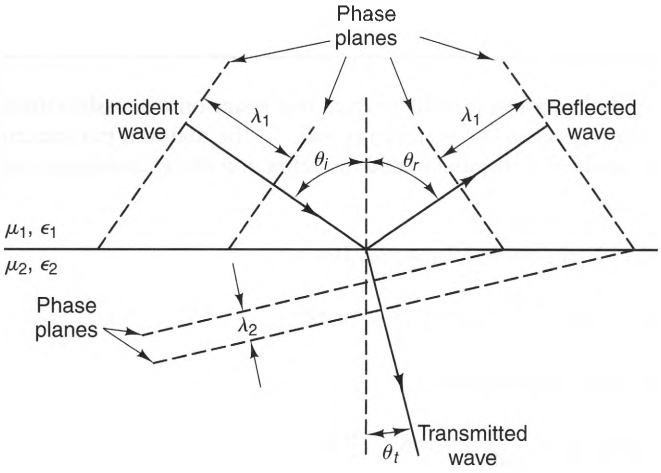


Figure 12-27 Constant-phase planes of the incident, reflected, and transmitted waves at an interface. These show how Snell's laws of reflection and transmission maintain the same distance between the planes along the interface.

spacing between the transmitted phase planes is different, since the wavelength is different in the transmission region. In order for the phase planes below the surface to track those above, the direction of propagation of the transmitted wave must bend towards the surface normal when $\lambda_2 < \lambda_1$ and away from the surface normal when $\lambda_2 > \lambda_1$, respectively.

Now that the correct relationships between the incident, reflected, and transmitted angles have been determined, we can return to the electric and magnetic field boundary conditions (Equations (12.169) and (12.170), respectively). Substituting Equations (12.172) and (12.173) into the boundary condition equations and canceling the common terms, we obtain

$$1 + \Gamma_\perp = T_\perp .$$

and

$$1 - \Gamma_\perp = T_\perp \frac{\eta_1 \cos\theta_t}{\eta_2 \cos\theta_i}.$$

Solving these expressions for $\Gamma_\perp$ and $\mathrm{T}_\perp$, we find that

$$\Gamma_\perp = \frac{\eta_2 \cos\theta_i - \eta_1 \cos\theta_t}{\eta_2 \cos\theta_i + \eta_1 \cos\theta_t} \qquad (12.177)$$

$$T_\perp = \frac{2\eta_2 \cos\theta_i}{\eta_2 \cos\theta_i + \eta_1 \cos\theta_t}. \qquad (12.178)$$

Notice that when $\theta_i = 0$, $\theta_r = \theta_t = 0$, and $\Gamma_\perp$ and $T_\perp$ assume the same values as for the case of normal incidence (Equations (12.121) and (12.122), respectively). Also, when the second medium is a perfect conductor, $\eta_2 = 0$, which yields

$$\Gamma_\perp = -1 \quad (\sigma_2 \to \infty) \qquad (12.179)$$

$$T_\perp = 0 \quad (\sigma_2 \to \infty). \qquad (12.180)$$

Example 12-11

A perpendicularly polarized plane wave is incident from free space onto a lossless dielectric surface at an angle of 30° with respect to the surface normal. If the material parameters are $\epsilon = 4.0\,\epsilon_o$ and $\mu = \mu_o$, find the angle of transmission and the reflection and transmission coefficients.

Solution:

From Equation (12.174), the angle of transmission

$$\theta_t = \sin^{-1}\left[\frac{n_1}{n_2}\sin\theta_i\right] = \sin^{-1}\left[\frac{1}{\sqrt{4}}\sin(30^\circ)\right] = 14.48^\circ.$$

The intrinsic impedances of the two media are

$$\eta_1 = \sqrt{\frac{\mu_o}{\epsilon_o}} = 377 \quad [\Omega], \ \eta_2 = \sqrt{\frac{\mu_o}{4\epsilon_o}} = 188.5 \quad [\Omega].$$

Substituting these values into Equations (12.177) and (12.178), we find that

$$\Gamma_\perp = \frac{188.5\cos(30°) - 377\cos(14.48°)}{188.5\cos(30°) + 377\cos(14.48°)} = -0.382$$

$$T_\perp = \frac{2 \times 188.5\cos(30°)}{188.5\cos(30°) + 377\cos(14.48°)} = 0.618.$$

12-8-2 PARALLEL POLARIZATION

Parallel polarization[12] occurs when the incident electric-field vector lies in the plane of incidence. The situation is depicted in Figure 12-28. We can represent the incident electric and magnetic fields for this case by the expressions

$$\mathbf{E}^i = E^i(\cos\theta_i\hat{\mathbf{a}}_x - \sin\theta_i\hat{\mathbf{a}}_z)e^{-jk_1(z\cos\theta_i + x\sin\theta_i)} \tag{12.181}$$

$$\mathbf{H}^i = \frac{E^i}{\eta_1}\hat{\mathbf{a}}_y e^{-jk_1(z\cos\theta_i + x\sin\theta_i)}. \tag{12.182}$$

Since the incident field is polarized parallel to the plane of incidence, we will now show that the reflected and transmitted fields have the same polarization and can be expressed as

$$\mathbf{E}^r = \Gamma_\parallel E^i(\cos\theta_r\hat{\mathbf{a}}_x + \sin\theta_r\hat{\mathbf{a}}_z)e^{-jk_1(-z\cos\theta_r + x\sin\theta_r)} \tag{12.183}$$

$$\mathbf{H}^r = -\frac{\Gamma_\parallel E^i}{\eta_1}\hat{\mathbf{a}}_y e^{-jk_1(-z\cos\theta_r + x\sin\theta_r)}, \tag{12.184}$$

and

$$\mathbf{E}^t = T_\parallel E^i(\cos\theta_t\hat{\mathbf{a}}_x - \sin\theta_t\hat{\mathbf{a}}_z)e^{-jk_2(z\cos\theta_t + x\sin\theta_t)} \tag{12.185}$$

$$\mathbf{H}^t = \frac{T_\parallel E^i}{\eta_2}\hat{\mathbf{a}}_y e^{-jk_2(z\cos\theta_t + x\sin\theta_t)}. \tag{12.186}$$

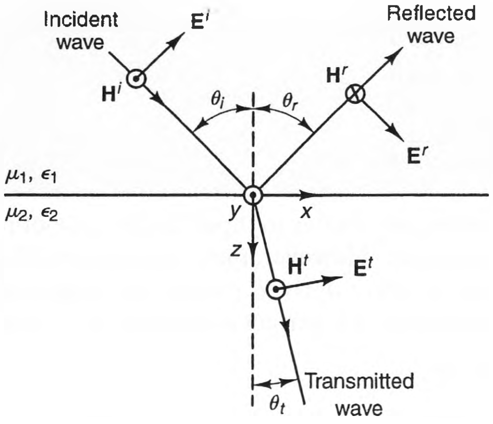


Figure 12-28 Reflection and transmission of an obliquely incident, parallel-polarized plane wave at a planar interface.

[12] Other names for this are vertical polarization, H polarization, and TM (transverse magnetic) polarization.

In these expressions, $\Gamma_{\|}$ and $T_{\|}$ are the parallel reflection and transmission coefficients, respectively.

To find the appropriate values of $\Gamma_{\|}$ and $T_{\|}$, we must require the total tangential electric and magnetic fields to be continuous at $x = 0$. Evaluating Equations (12.181)–(12.186) at $x = 0$, and remembering that the total fields in region 1 are the sum of the incident and reflected fields, we obtain the following expressions for the electric and magnetic fields, respectively:

$$E^i \cos\theta_i e^{-jk_1 x \sin\theta_i} + \Gamma_{\|} E^i \cos\theta_r e^{-jk_1 x \sin\theta_r} = T_{\|} E^i \cos\theta_t e^{-jk_2 x \sin\theta_t} \tag{12.187}$$

$$\frac{E^i}{\eta_1} e^{-jk_1 x \sin\theta_i} - \frac{\Gamma_{\|} E^i}{\eta_1} e^{-jk_1 x \sin\theta_r} = \frac{T_{\|} E^i}{\eta_2} e^{-jk_2 x \sin\theta_t}. \tag{12.188}$$

The exponential terms in these expressions are the same as for the perpendicular case (Equation (12.170)), so Snell's laws of reflection and transmission are the same for parallel polarization as they are for perpendicular polarization:

$$\theta_i = \theta_r \tag{12.189}$$

$$\sin\theta_t = \frac{k_1}{k_2}\sin\theta_i = \frac{n_1}{n_2}\sin\theta_i. \tag{12.190}$$

Substituting Equations (12.189) and (12.190) into Equations (12.187) and (12.188) and canceling the common terms, we obtain

$$1 + \Gamma_{\|} = T_{\|}\frac{\cos\theta_t}{\cos\theta_i}$$

$$1 - \Gamma_{\|} = T_{\|}\frac{\eta_1}{\eta_2}.$$

Finally, solving the preceding expressions for $\Gamma_{\|}$ and $T_{\|}$, we get

$$\Gamma_{\|} = \frac{\eta_2\cos\theta_t - \eta_1\cos\theta_i}{\eta_2\cos\theta_t + \eta_1\cos\theta_i} \tag{12.191}$$

$$T_{\|} = \frac{2\eta_2\cos\theta_i}{\eta_2\cos\theta_t + \eta_1\cos\theta_i}. \tag{12.192}$$

Note that these expressions are similar to those for the case of perpendicular polarization, but they are not identical. However, these formulas yield the same reflection and transmission coefficients as were derived earlier for normal incidence when $\theta_i = 0$. Also, when the second medium is a perfect conductor, $\eta_2 = 0$, which yields

$$\Gamma_{\|} = -1 \quad (\sigma_2 \to \infty) \tag{12.193}$$

$$T_{\|} = 0 \quad (\sigma_2 \to \infty). \tag{12.194}$$

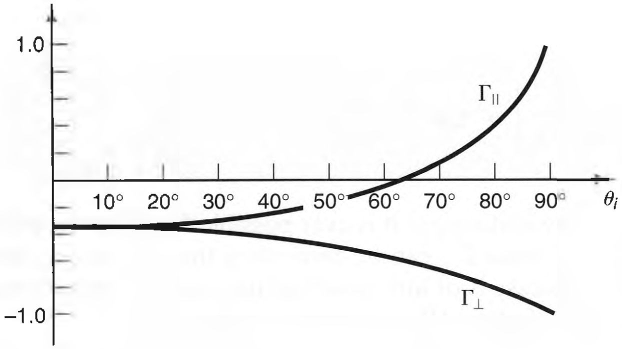


Figure 12-29 Plots of the reflection coefficients for perpendicular- and parallel-polarized waves that are incident from free space onto a medium with $\epsilon = 4\epsilon_0$ and $\mu = \mu_0$.

12-8-3 THE BREWSTER ANGLE

Figure 12-29 shows the values of $\Gamma_\perp$ and $\Gamma_\parallel$ as a function of θ_i when a wave is incident from free space onto a medium characterized by $\epsilon = 4\epsilon_o$ and $\mu = \mu_o$. As can be seen, $\Gamma_\perp$ and $\Gamma_\parallel$ are equal at $\theta_i = 0$, but as $\theta_i \to 90°$, they approach $+1$ and -1, respectively. Notice also that $\Gamma_\perp$ is never zero, whereas $\Gamma_\parallel$ equals zero at one angle, $\theta_i = 63.4°$, for this case. The incident angle that yields $\Gamma_\parallel = 0$ is called the ***Brewster angle*** and is denoted by the symbol θ_B. It is also called the ***polarizing angle***, because at this angle, the reflected fields are always linearly polarized, with the electric field perpendicular to the plane of incidence.

To find a general formula for the Brewster angle, we set the expression for $\Gamma_\parallel$ (Equation (12.191)) equal to zero:

$$\frac{\eta_2 \cos\theta_t - \eta_1 \cos\theta_i}{\eta_2 \cos\theta_t + \eta_1 \cos\theta_i} = 0.$$

This equation is satisfied when the numerator is zero, which occurs when

$$\eta_2 \cos\theta_t = \eta_1 \cos\theta_i .$$

If both media are nonmagnetic, $\mu_1 = \mu_2 = \mu_o$, and the preceding expression becomes

$$\frac{1}{\sqrt{\epsilon_2}} \cos\theta_t = \frac{1}{\sqrt{\epsilon_1}} \cos\theta_i,$$

or

$$\epsilon_1(1 - \sin^2\theta_t) = \epsilon_2(1 - \sin^2\theta_i).$$

But we also know from Snell's law that $\sin^2\theta_t = (\epsilon_1/\epsilon_2)\sin^2\theta_i$. Substituting this into the preceding expression and solving for $\sin\theta_i$, we obtain

$$\sin\theta_i = \sqrt{\frac{1 - \epsilon_1/\epsilon_2}{1 - (\epsilon_1/\epsilon_2)^2}} = \sqrt{\frac{1}{1 + \epsilon_1/\epsilon_2}}.$$

Hence, $\Gamma_{\parallel} = 0$ when $\theta_i = \theta_B$, where

$$\theta_B = \sin^{-1}\left[\sqrt{\frac{1}{1 + \epsilon_1/\epsilon_2}}\right] = \tan^{-1}\left(\sqrt{\frac{\epsilon_2}{\epsilon_1}}\right). \tag{12.195}$$

The reader might be wondering if it is ever possible for $\Gamma_{\perp}$ to equal zero. Theoretically, the answer is yes, since $\Gamma_{\perp}$ can be zero when the two media have different values of μ. However, this case is of little practical use, since magnetic materials usually have high losses, particularly at RF frequencies and above.

The polarizing effect at the Brewster angle is often used in optical devices to either reflect or transmit a specific polarization. One common application that uses the Brewster effect is glare-resistant sunglasses. Sunlight is unpolarized, which means that its the polarization state is random and fluctuates rapidly. Reflections of sunlight off most surfaces contain more power in the perpendicular component than in the parallel component if the angle of incidence is relatively close to the Brewster angle. Such a situation is depicted in Figure 12-30. When the angle of incidence is close to the Brewster angle, most of the power in the reflected wave is polarized parallel to the ground. If these reflections are viewed through a lens that passes only light that is polarized perpendicular to the earth, very little of the reflection will be seen. Lenses that pass only one type of polarization are called ***polarizing*** (or Polaroid™) ***lenses***.

Another important use of the Brewster angle is in lasers. A simplified schematic of a laser is shown in Figure 12-31. A laser is an oscillator that amplifies waves by forcing them to reflect back and forth in an optical cavity that contains a lasing medium. The lasing medium, which can be gas, liquid, or solid, is induced into an amplifying state by a process called ***pumping***, whereby a critical number of its atoms are excited

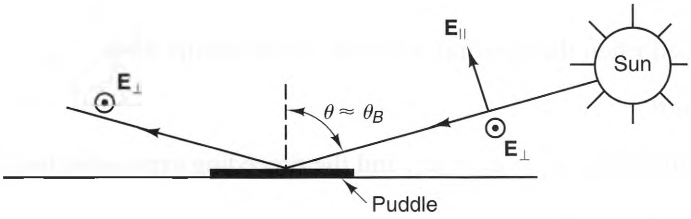


Figure 12-30 Unpolarized sunlight that becomes mostly polarized after reflecting off a water puddle at roughly the Brewster angle.

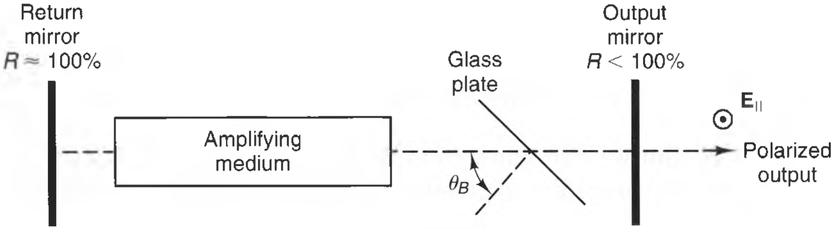


Figure 12-31 A glass plate in a laser cavity at the Brewster angle, yielding a polarized output.

into elevated energy states. Common methods of pumping are optical flashlamps (common in solid lasers), electric discharges (common in gas lasers), and minority carrier injection (used in semiconductor lasers). Sustained oscillations occur when the pumping rate is sufficient to make the net power gain per round trip in the optical cavity greater than zero.

The outputs of many lasers are randomly polarized, since the gain in the cavity is often independent of the wave polarization. To obtain a polarized output, a glass plate can be placed between the mirrors and aligned so that the waves enter and leave it at the Brewster angle.[13] The parallel-polarized waves are unaffected by the plate, but lasing for perpendicular polarization is extinguished because the gain per round trip is pushed below the threshold value.

Example 12-12

Calculate the Brewster angle of the interface between free space and a glass plate that has a dielectric constant of $\epsilon_r = 7.0$.

Solution:

Using Equation (12.195), we obtain

$\theta_B = \tan^{-1}(\sqrt{7}) = 69.3°.$

12-8-4 TOTAL REFLECTION AND THE CRITICAL ANGLE

When a plane wave is incident from a dense medium into a rare medium, such as from water to air, Snell's law of refraction predicts that the transmission angle θ_t must be larger than the angle of incidence θ_i. The angle of incidence that results in $\theta_t = 90°$ is called the ***critical angle*** θ_c. To calculate θ_c, note that $\sin \theta_t = 1$ when $\theta_t = 90°$. Using Snell's law of reflection, we find that

$$\sin \theta_c = \frac{n_2}{n_1} \sin \theta_t = \frac{n_2}{n_1},$$

where n_1 and n_2 are the refraction indices of regions 1 and 2, respectively. Solving for θ_c, we obtain

$$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right). \tag{12.196}$$

When both media are nonmagnetic, θ_c can be written in terms of the permittivities of the media:

[13] It is shown in Problem 12-28 that when the incident field enters a flat plate at the Brewster angle, it also leaves the plate at the Brewster angle.

$$\theta_c = \sin^{-1}\left(\sqrt{\frac{\epsilon_2}{\epsilon_1}}\right) \qquad \text{(Nonmagnetic media)}. \qquad (12.197)$$

To see what happens when θ_i is greater than the critical angle, note that Snell's law of refraction predicts that $\sin\theta_t$ is greater than unity:

$$\sin\theta_t = \frac{n_1}{n_2}\sin\theta_i > 1 \quad (\theta_i > \theta_c). \qquad (12.198)$$

This means that θ_t is an imaginary number and can no longer be interpreted as the angle of transmission of an ordinary plane wave. Also, since

$$\cos\theta_t = \sqrt{1 - \sin^2\theta_t},$$

we find that $\cos\theta_t$ is imaginary when $\theta_i > \theta_c$. Hence, we can write[14]

$$\cos\theta_t = -jA, \qquad (12.199)$$

where A is a positive real number, defined by

$$A = \sqrt{\sin^2\theta_t - 1} = \sqrt{\frac{n_1^2}{n_2^2}\sin^2\theta_i - 1}. \qquad (12.200)$$

Even when $\theta_i > \theta_c$, Snell's law of reflection still predicts that the angle of reflection equals the angle of incidence. But the reflection coefficients do something peculiar when $\theta_i > \theta_c$ for both perpendicular and parallel polarization. To see this, let us start with the perpendicular polarization case. If we substitute Equation (12.199) into Equation (12.177), we obtain

$$\Gamma_\perp = \frac{\eta_2\cos\theta_i + j\eta_1 A}{\eta_2\cos\theta_i - j\eta_1 A}. \qquad (12.201)$$

Since the numerator and denominator of this expression are complex conjugates, $\Gamma_\perp$ has a magnitude of 1.0. This means that the incident field undergoes ***total reflection*** at the interface. Writing $\Gamma_\perp$ in complex form, we obtain

$$\Gamma_\perp = 1\angle\phi_\perp \qquad (\theta_i > \phi_c,) \qquad (12.202)$$

where the phase shift is

$$\phi_\perp = 2\tan^{-1}\left[\frac{\eta_1\sqrt{\frac{n_1^2}{n_2^2}\sin^2\theta_i - 1}}{\eta_2\sqrt{1 - \sin^2\theta_i}}\right]. \qquad (12.203)$$

When both media are nonmagnetic, $\mu_1 = \mu_2 = \mu_o$, $n_1^2 = \epsilon_1/\epsilon_o$, and $n_2^2 = \epsilon_2/\epsilon_o$, so Equation (12.203) can be further simplified to read

[14] The choice $\cos\theta_t = jA$ is also possible, but would result in a transmitted field that increases exponentially with depth, violating the conservation-of-energy principle.

$$\phi_\perp = 2\tan^{-1}\left[\frac{\sqrt{\sin^2\theta_i - \epsilon_2/\epsilon_1}}{\sqrt{1-\sin^2\theta_i}}\right] \qquad \text{(Nonmagnetic media)}. \tag{12.204}$$

Given that the reflected fields have the same magnitudes as the incident fields when $\theta_i > \theta_c$, it may at first seem logical to assume that the transmitted fields are zero. But such is not the case. To see why, let us take a closer look at the expressions for the transmitted electric and magnetic fields when the incident field is perpendicularly polarized. Using Equations (12.167) and (12.168), we can write the transmitted fields in the form

$$\mathbf{E}^t = T_\perp E^i \hat{\mathbf{a}}_y e^{-\alpha_2 z} e^{-j\beta_{2z} x} \tag{12.205}$$

$$\mathbf{H}^t = \frac{T_\perp E^i}{\eta_2}(jA\hat{\mathbf{a}}_x + \sin\theta_t \hat{\mathbf{a}}_z)e^{-\alpha_2 z}e^{-j\beta_{2z}x}, \tag{12.206}$$

where A and $\sin\theta_t$ are given by Equations (12.200) and (12.198), respectively, and

$$T_\perp = \frac{2\eta_2\cos\theta_i}{\eta_2\cos\theta_i - j\eta_1 A} \tag{12.207}$$

$$\alpha_2 = k_2 A = k_2\sqrt{\frac{n_1^2}{n_2^2}\sin^2\theta_i - 1} \qquad \text{[Np/m]} \tag{12.208}$$

$$\beta_{2z} = k_2\sin\theta_t = k_2\frac{n_1}{n_2}\sin\theta_i \qquad [\text{m}^{-1}]. \tag{12.209}$$

As can be seen from Equation (12.207), the transmission coefficient $T_\perp$ does *not* equal zero when $\theta_i > \theta_c$, which means that the transmitted fields are nonzero, even though the incident field is totally reflected. The transmitted fields decay exponentially with increasing values of z and exhibit no phase changes with increasing depth, which is the characteristic of an ***evanescent wave***. This decay has nothing to do with loss, however. Rather, it occurs because the incident wave has been totally reflected. Because it decays quickly with depth, this wave is often called a ***surface wave***. Such a situation is depicted in Figure 12-32.

The existence of nonzero transmitted fields when the reflection coefficient $\Gamma_\perp$ has unity magnitude may appear to violate the principle of energy conservation. To see why this is not a problem, let us calculate the complex Poynting vector **S** in the region where the transmitted wave exists:

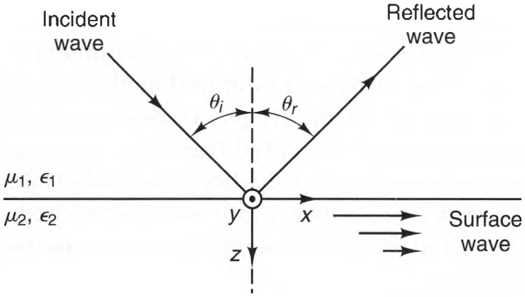


Figure 12-32 A totally reflected wave and the accompanying surface wave.

$$\mathbf{S} = \mathbf{E}^t \times (\mathbf{H}^t)^* = \frac{|T_\perp|^2 |E^i|^2}{\eta_2} e^{-2\alpha_2 z}[jA\hat{\mathbf{a}}_z + \sin\theta_t \hat{\mathbf{a}}_x]. \tag{12.210}$$

This vector has both x- and z-components. However, only the x-component is real, which means that a net average power flows parallel to the interface inside the transmission medium, but not perpendicular to it. This is consistent with the fact that the incident and reflected fields have the same magnitude and, as a result, transport no net power through the interface.

Now that we have identified the evanescent behavior of the transmitted fields when the incident field is perpendicularly polarized, let us next consider the case where the incident field is parallel polarized. The reflection coefficient for this case can be found by substituting Equation (12.199) into Equation (12.191):

$$\Gamma_\parallel = \frac{-j\eta_2 A - \eta_1 \cos\theta_i}{-j\eta_2 A + \eta_1 \cos\theta_i}.$$

Writing this in polar form, we have

$$\Gamma_\parallel = 1\angle\phi_\parallel \qquad (\theta_i > \theta_c), \tag{12.211}$$

where the phase shift is

$$\phi_\parallel = -180^\circ + 2\tan^{-1}\left[\frac{\eta_2\sqrt{\frac{n_1^2}{n_2^2}\sin^2\theta_i - 1}}{\eta_1\sqrt{1 - \sin^2\theta_i}}\right]. \tag{12.212}$$

When both media are nonmagnetic, Equation (12.212) can be further simplified to read

$$\phi_\parallel = -180^\circ + 2\tan^{-1}\left[\frac{\sqrt{\frac{\epsilon_1^2}{\epsilon_2^2}\sin^2\theta_i - \epsilon_1/\epsilon_2}}{\sqrt{1 - \sin^2\theta_i}}\right] \qquad \text{(Nonmagnetic media).} \tag{12.213}$$

The transmitted fields for parallel polarization are evanescent, with exactly the same rates of decay as for the perpendicular polarization case. The formula for the transmission coefficient is slightly different:

$$T_\parallel = \frac{2\eta_2 \cos\theta_i}{-jA\eta_2 + \eta_1 \cos\theta_i}. \tag{12.214}$$

Finally, an interesting question concerning the phenomenon of total reflection is the following: If no power flows into the transmitted medium, how did the transmitted fields get there? The answer is that the frequency-domain analysis we have used to describe this process has neglected the initial transient response of the waves when they are first turned on. Although $|\Gamma| = 1$ when the steady state is reached, $|\Gamma| < 1$ during the initial transient period. Once steady state is reached, energy is no longer passed across the interface, and the stored energy in the transmitted fields simply propagate parallel to the interface.

Example 12-13

A 100-MHz plane wave is incident from fresh water ($\epsilon_r = 80$, $\sigma \approx 0$) into air at an angle of incidence of 60°. If the field is perpendicularly polarized, calculate the reflection coefficient and the rate of decay of the transmitted field in dB/m.

Solution:

From Equation (12.196), the critical angle for this interface is

$$\theta_c = \sin^{-1}\left(\sqrt{\frac{1}{80}}\right) = 6.42°.$$

Thus, an incident angle of 60° is well beyond the critical angle, and we can use Equation (12.208) to find the attenuation constant α_2 in the air. Remembering that $\beta = \omega/c$ in air,

$$\alpha_2 = k_2\sqrt{\frac{n_1^2}{n_2^2}\sin^2\theta_i - 1} = \frac{2\pi \times 100 \times 10^6}{3 \times 10^8}\sqrt{80\sin^2 60° - 1}$$
$$= 16.087 \text{ [Np/m]}.$$

Substituting this result into Equation (12.64) to convert α_2 from [Np/m] to [dB/m], we obtain

$$\alpha_2 = \frac{16.087}{.1151} = 139.7 \qquad \text{[dB/m]}.$$

The magnitude of the reflection coefficient $\Gamma_\perp$ is unity, since $\theta_i > \theta_c$. The phase of the reflection coefficient can be found using Equation (12.204);

$$\phi_\perp = 2\tan^{-1}\left[\frac{\sqrt{\sin^2 60° - 1/\sqrt{80}}}{\sqrt{1 - \sin^2 60°}}\right] = 115.9°.$$

Thus, the reflection coefficient is

$$\Gamma_\perp = 1.0\angle 115.9°.$$

12-9 Summation

In this chapter, we have discussed various aspects of plane waves. Plane waves are the simplest, and most important examples of space waves. Even though true plane waves are nonphysical in that they contain an infinite amount of power, they are excellent approximations of the waves generated by real sources.

We will again see plane waves in Chapter 14 when we discuss antennas and radiation. There, we will nearly always assume that the fields radiated by an antenna behave like plane waves when they are viewed far from their sources.

Problems

12-1 The frequency-domain expression for the electric field intensity of a plane wave in air is

$$\mathbf{E} = 3e^{-j(2.094y + \pi/4)}\hat{\mathbf{a}}_x \qquad \text{[V/m]},$$

where y is measured in meters. Find the frequency of this wave (in [MHz]) and time-domain expressions for both **E** and **H**.

12-2 The time-domain expression for the magnetic field intensity of a plane wave in air is

$$\mathbf{H} = 0.1\cos(\omega t + \pi z + 20°)\hat{\mathbf{a}}_x \qquad \text{[A/m]},$$

where z is measured in meters. Find the frequency (in MHz) of this wave and its wavelength.

12-3 If the frequency-domain representation of the H-field in a source-free region filled with air is given by

$$\mathbf{H} = 3e^{-j(\beta y + \pi/4)}\hat{\mathbf{a}}_x - 4e^{-j\beta z}\hat{\mathbf{a}}_y \qquad \text{[A/m]},$$

find the frequency-domain representation of **E**.

12-4 Find the time-domain expression for the E- and H-fields of a plane wave that propagates in air in the direction $1/(\sqrt{5})\,(\hat{\mathbf{a}}_x - 2\hat{\mathbf{a}}_z)$. Assume that **H** has only a y-component and has a magnitude of 10 [A/m].

12-5 Determine the direction of propagation and the polarization state of the following plane wave:

$$\mathbf{E} = E_o[(1 + j)\hat{\mathbf{a}}_x + (1 - j)\hat{\mathbf{a}}_y]e^{-j\beta z}.$$

12-6 Prove that:

(a) any linearly polarized wave can be resolved into the sum of right- and left-hand circularly polarized waves,

(b) any elliptically polarized wave can be resolved into the sum of right- and left-hand circularly polarized waves.

12-7 Prove that a linearly polarized wave can always be constructed by adding two elliptically polarized waves with opposite rotations.

12-8 Find the tilt angle τ and the axial ratio AR of the plane wave

$$\mathbf{E} = 3e^{-j(\beta z + 20°)}\hat{\mathbf{a}}_x + e^{-j(\beta z + 40°)}\hat{\mathbf{a}}_y.$$

12-9 Prove that the attenuation and phase constants of a medium can always be expressed in the form

$$\alpha = \omega\sqrt{\frac{\mu\epsilon}{2}}\left[\sqrt{1 + \left(\frac{\sigma}{\omega\epsilon}\right)^2} - 1\right]^{1/2} \qquad \text{[Np/m]}$$

$$\beta = \omega\sqrt{\frac{\mu\epsilon}{2}}\left[\sqrt{1 + \left(\frac{\sigma}{\omega\epsilon}\right)^2} + 1\right]^{1/2} \qquad [\text{m}^{-1}].$$

12-10 Certain materials become anisotropic when subjected to a strong magnetic bias field (such as from a magnet) so that counter rotating, circularly polarized waves propagate through them with different velocities. This effect is called *Faraday rotation* and can be used to create devices that rotate the plane of polarization of linearly polarized waves. For the device shown in Figure P12-10, prove that if a plane wave propagation in the $+z$ direction has polarization $\mathbf{E} = E_i\,\hat{\mathbf{a}}_x$ at $z = 0$, then the polarization vector will rotate clockwise as the wave propagates if $\beta_L > \beta_R$ and counterclockwise if $\beta_L < \beta_R$, where β_L and β_R are the phase constants for left- and right-hand propagating waves, respectively. Also if $(\beta_L - \beta_R)\ell = \pi$, show that $\mathbf{E}_o = E_i\,\hat{\mathbf{a}}_y$ at $z = \ell$. (*Hint:* Remember that any linearly polarized wave can be considered as the sum of two circularly polarized waves—in this case with different velocities.)

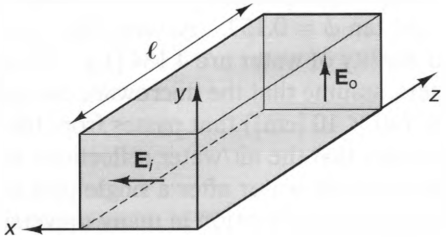


Figure P12-10

12-11 Prove by direct substitution that all possible plane waves represented by Equations (12.35)–(12.37) satisfy Maxwell's frequency-domain equations for all possible directions of propagation.

12-12 Find the complex permittivity of a material at 100 [MHz] that has a dielectric constant of 2.5 and a conductivity of 1.39×10^{-4} [S/m].

12-13 Calculate the attenuation in [dB/m] of a 1 [GHz] plane wave as it passes through a medium with a loss tangent of $\tan\phi = 0.05$ if it is known that the medium is nonmagnetic and the wavelength at that frequency is 17.32 [cm].

12-14 Calculate the intrinsic impedance η, the attenuation constant (in [Np/m]), the skin depth δ, and the wavelength of copper ($\sigma = 5.8 \times 10^7$ [S/m], $\epsilon = \epsilon_o$, $\mu = \mu_o$) at a) $f = 60$ [Hz], b) $f = 100$ [MHz], and c) $f = 10$ [GHz].

12-15 Calculate the effective complex permittivity ϵ of a nonmagnetic material that exhibits an attenuation of 0.1 [dB/km] at a frequency of 10 [GHz] if the wavelength at that frequency is 1.342 [cm].

12-16 A linearly polarized plane wave with a peak electric field amplitude of 3 [V/m] propagates through a lossless, nonmagnetic medium and has a power density of 22 [mW/m²]. Find the relative permittivity ϵ_r of the medium.

12-17 At large distances from an antenna, the radiated electric and magnetic fields can always be written in the form

$$\mathbf{E} = E_\theta \hat{\mathbf{a}}_\theta + E_\phi \hat{\mathbf{a}}_\phi$$

$$\mathbf{H} = \frac{1}{\eta}(E_\phi \hat{\mathbf{a}}_\theta - E_\theta \hat{\mathbf{a}}_\phi),$$

where η is the intrinsic impedance of the medium. Find the expression for the power density $\mathscr{S}_{\text{ave}}$ in terms of E_θ and E_ϕ. Since both E_θ and E_ϕ are always proportional to $1/r$ when r is large, how does $\mathscr{S}_{\text{ave}}$ decay with increasing values of r?

12-18 A uniform current $\mathbf{J} = J_o \hat{\mathbf{a}}_z$ flows in a conducting slab with conductivity σ that lies in the region $|x| < d$.

(a) Calculate the dissipated power per unit surface area of the upper face P_{diss} using Joule's law (Equation (12.97)).

(b) Calculate P_{diss} by integrating the instantaneous Poynting vector $\mathscr{P}$ around the outside surface of the slab. (*Hint:* Use Ampere's law to find **H**.)

12-19 Calculate the time it takes to boil a container of distilled water in a microwave oven, where the container is 2 [cm] high and has a cross section of 100 [cm^2]. Microwave ovens typically operate at 2.4 GHz, where distilled water is characterized by $\epsilon_r = 78$ and $\tan \phi = 0.157$. Assume that water starts at 23° [C], and the specific heat and density of water are 4.184 [J/g · °C] and 1.0 [g/cm^3], respectively. To simplify matters, assume that the microware energy is a 1.5 [kW], square-apertured plane wave (10 × 10 [cm]) that passes from the top of the container to the bottom. Assume also that the air/water reflections are negligible and the energy that is not absorbed by the water after a single pass is absorbed by the grease on the oven floor (not a bad assumption in many university living quarters!).

12-20 A 300 [MHz] plane is normally incident from free space onto a lossless, nonmagnetic slab of infinite thickness. Find the refection and transmission coefficients if it is known that the velocity of propagation in the slab is $0.85c$.

12-21 A 10 [GHz] uniform plane wave is normally incident from free space onto an ocean surface. If the power density of the incident wave is 100 [mW/m^2] and the seawater has $\epsilon_r = 80$ and $\tan \phi = 0.56$, calculate the time average power density

(a) just below the surface

(b) at a depth of 1 [cm].

12-22 A 1 [GHz] plane wave is normally incident upon a thick, nonmagnetic metal slab. If the slab has a conductivity of $\sigma = 1 \times 10^7$ [S/m] and the power density of the incident wave is 10 [W/m^2], find the power (in [W/m^2]) that is dissipated in the slab for each square meter of surface area.

12-23 Suppose that the spacing between adjacent resonant frequencies of a particular etalon is 40 [GHz]. If the etalon is 3 [mm] thick, find its dielectric constant.

12-24 A quarter-wave stack consists of a series of quarter-wavelength-thick layers, usually deposited on a substrate (such as glass). Highly reflecting mirrors can be made from these stacks by alternating values of permittivity. Figure P12-24 shows a quarter-wave stack consisting of a number of cells, where each cell is a pair of high- and low-permittivity layers. Find the minimum number of layers necessary to attain 99.9% *power* reflectivity if $\epsilon_H = 6\,\epsilon_o$, $\epsilon_L = 1.8\,\epsilon_o$, and $\epsilon_s = 4\,\epsilon_o$.

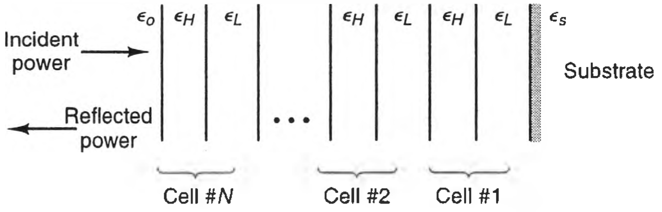


Figure P12-24

12-25 The window of lowest loss in silica glass optical fibers is centered at a free-space wavelength of approximately 1.55 [μm] and has a width of approximately 200 [nm]. The frequency range of semiconductor lasers is much narrower than this window, so many laser frequencies can be used simultaneously to transmit different signals. (This is called wavelength division multiplexing.) One way to filter each wavelength out separately for detection is by using etalons. Calculate the etalon width that is necessary to obtain a resonant wavelength spacing of 10 [nm] if the relative permittivity of the etalon is 6.0.

12-26 Suppose that transparent coating is applied to a glass substrate to eliminate reflections of violet light ($\lambda_o = 0.4$ [μm]) at normal incidence. If the substrate has $\epsilon_r = 6$ and $\mu_r = 1$, find:

(a) the required coating permittivity and thickness.

(b) the power reflectivity for red light ($\lambda_o = 0.7$ [μm]).

12-27 A circularly polarized plane wave is incident from free space onto a flat surface with $\epsilon_r = 2.5$ and $\mu_r = 1$. If the angle of incidence is 40°, calculate the axial ratio *AR* of the reflected wave.

12-28 The dielectric slab in Figure P12-28 is illuminated by an incident plane wave. Show that:

(a) the input and output rays are parallel.

(b) if the angle of incidence θ_1 equals the Brewster angle for the input face, then the angle θ_2 automatically equals the Brewster angle of the output face.

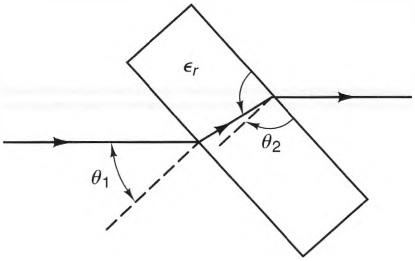


Figure P12-28

12-29 Suppose that a glass plate with $\epsilon_r = 4.5$ is inserted into a helium–neon laser cavity at the Brewster angle to polarize the laser's output, as shown in Figure 12-31. If the cavity wavelength is $\lambda_o = 0.6328$ [μm] and multiple reflections within the plate are negligible, calculate the net reflection loss, in dB per pass through the plate, of the unwanted polarization.

12-30 A 300 [MHz] plane wave with a power density of 100 [mW/m^2] is incident from fresh water ($\epsilon_r = 80$, $\sigma = 0$) into air at an incident angle of 40° and is polarized perpendicular to the plane of incidence. Calculate the height above the water surface at which the E-field has a magnitude of 10^{-4} [V/m].

12-31 Show that the reflection and transmission coefficients for a plane wave incident upon a planar boundary between two nonmagnetic media can be written in the form

$$\Gamma_\perp = \frac{\sin(\theta_t - \theta_i)}{\sin(\theta_t + \theta_i)} \qquad T_\perp = \frac{2\cos\theta_i \sin\theta_t}{\sin(\theta_t + \theta_i)}$$

$$\Gamma_\parallel = \frac{\tan(\theta_t - \theta_i)}{\tan(\theta_t + \theta_i)} \qquad T_\parallel = \frac{2\cos\theta_i \sin\theta_t}{\sin(\theta_t + \theta_i)\cos(\theta_t - \theta_i)}.$$

13

Waveguides

13-1 Introduction

Waveguides are similar to transmission lines in that they, too, are used to transport electromagnetic energy and signals along a fixed path. But whereas transmission lines carry TEM (or quasi-TEM) modes, waveguides carry non-TEM modes, often called ***waveguide modes***. This difference between transmission lines and waveguides may seem subtle, but their mechanical and electrical properties are quite different. For instance, transmission lines must have at least two separate conductors, whereas waveguides can consist of a single conductor or only dielectrics (as in the case of an optical fiber). Typically, waveguides must be operated over smaller bandwidths than transmission lines, but they generally exhibit smaller losses, which makes them attractive for many applications.

There are many different types of waveguides, but they can usually be placed into two broad classes: metal waveguides and dielectric waveguides. Metal waveguides employ conductors to confine and guide the waves and are typically used in the RF, microwave, and millimeter wave ranges. Figure 13-1 shows three types of metal waveguides: rectangular, circular, and ridge.

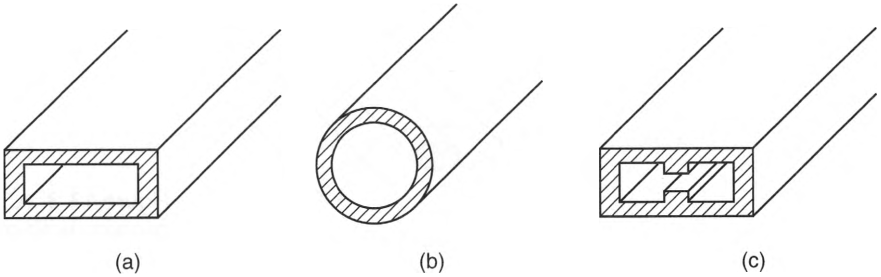


Figure 13-1 Three types of metal waveguides: a) Rectangular. b) Circular. c) Ridge.

Each of these types of metal waveguide has distinct properties, but they are similar enough so that a detailed analysis of one gives insight into most of the properties of the others.

Unlike metal waveguides, dielectric waveguides direct waves using the reflections at the interfaces of dissimilar dielectrics. Dielectric waveguides do not confine waves as tightly as do metal waveguides, but they have many attractive properties, particularly at frequencies where the conductive losses of metal waveguides make them unattractive. Figure 13-2 shows two types of dielectric waveguides: a slab waveguide and an optical fiber.

The advantages of dielectric waveguides are most pronounced at optical frequencies, where the losses of certain dielectric materials are extremely low. This property is directly responsible for the rapid increase in the use of fiber-optic links that have revolutionized computer networks and communication systems.

13-2 Waveguide Modes

Even though the electric and magnetic fields inside different types of waveguides distribute themselves differently, all waveguides share a number of common properties. The most important of these is that they can support an infinite number of field configurations, called ***modes***. In this section, we will identify the basic types of modes present in waveguiding structures.

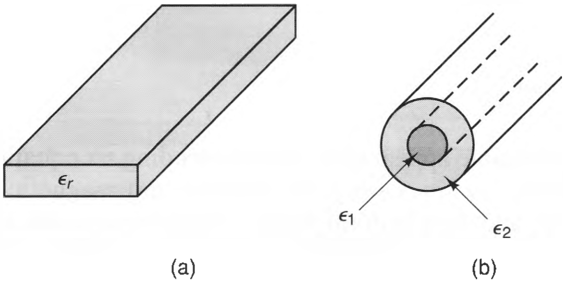


Figure 13-2 Two dielectric waveguides: a) Dielectric slab. b) Optical fiber.

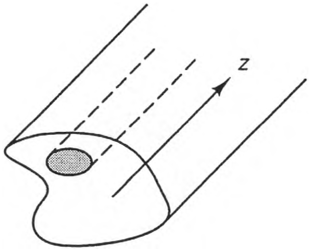


Figure 13-3 A uniform waveguide that contains both conductors and dielectrics.

Figure 13-3 shows a section of a waveguide with an arbitrary cross section that is uniform along the z-axis. The materials used can be either dielectrics or conductors, and more than one kind of material may be present. To model the kinds of fields that exist within and around this waveguide, we start with Maxwell's two curl equations in source-free media:

$$\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H} \tag{13.1}$$

$$\nabla \times \mathbf{H} = j\omega\epsilon\mathbf{E}\ . \tag{13.2}$$

If we assume that the z dependence of each component of **E** and **H** is of the form $E_i, H_i \propto e^{-\gamma z}$ $(i = x, y, \text{or } z)$, where γ is the propagation constant, these two vector equations can be written as the following six scalar equations:

$$\frac{\partial E_z}{\partial y} + \gamma E_y = -j\omega\mu H_x \tag{13.3}$$

$$-\gamma E_x - \frac{\partial E_z}{\partial x} = -j\omega\mu H_y \tag{13.4}$$

$$\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = -j\omega\mu H_z \tag{13.5}$$

$$\frac{\partial H_z}{\partial y} + \gamma H_y = j\omega\epsilon E_x \tag{13.6}$$

$$-\gamma H_x - \frac{\partial H_z}{\partial x} = j\omega\epsilon E_y \tag{13.7}$$

$$\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = j\omega\epsilon E_z. \tag{13.8}$$

In spite of the formidable appearance of the preceding six equations, we can simplify matters significantly by separating the electric and magnetic fields into two classes: ***transverse fields*** and ***longitudinal fields***. Transverse fields are directed perpendicular to the direction of propagation (i.e., E_x, E_y, H_x, and H_y), whereas longitudinal fields are directed parallel to the direction of propagation (i.e., E_z, H_z). Next, we can manipulate the six equations so that each transverse component is specified only in terms of the longitudinal components. For instance, solving Equation (13.4) for H_y, we obtain

$$H_y = \frac{\gamma}{j\omega\mu} E_x + \frac{1}{j\omega\mu}\frac{\partial E_z}{\partial x}.$$

Substituting this into Equation (13.6) and solving for E_x, we get

$$E_x = -\frac{1}{h^2}\left[\gamma\frac{\partial E_z}{\partial x} + j\omega\mu\frac{\partial H_z}{\partial y}\right],$$

where $h^2 = \gamma^2 + \omega^2\mu\epsilon$. Similar expressions can be derived for the other transverse components and are summarized as follows:

$$E_x = -\frac{1}{h^2}\left[\gamma\frac{\partial E_z}{\partial x} + j\omega\mu\frac{\partial H_z}{\partial y}\right] \tag{13.9}$$

$$E_y = -\frac{1}{h^2}\left[\gamma\frac{\partial E_z}{\partial y} - j\omega\mu\frac{\partial H_z}{\partial x}\right] \tag{13.10}$$

$$H_x = -\frac{1}{h^2}\left[-j\omega\epsilon\frac{\partial E_z}{\partial y} + \gamma\frac{\partial H_z}{\partial x}\right] \tag{13.11}$$

$$H_y = -\frac{1}{h^2}\left[j\omega\epsilon\frac{\partial E_z}{\partial x} + \gamma\frac{\partial H_z}{\partial y}\right], \tag{13.12}$$

where

$$h^2 = \gamma^2 + \omega^2\mu\epsilon. \tag{13.13}$$

Equations (13.9)–(13.13) suggest that we can divide the modes in waveguiding structures into the following classes:

TEM modes These modes have $E_z = H_z = 0$ and $h^2 = \gamma^2 + \omega^2\mu\epsilon = 0$. Examples of TEM modes are plane waves and transmission-line modes. Waveguides with finite cross-sectional dimensions can have a TEM mode only if there are at least two separate conductors and a uniform dielectric (such as is found in coaxial cables).

TE modes Transverse-electric modes, sometimes called H 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 uniform dielectrics.

TM modes Transverse-magnetic modes, sometimes called E 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 uniform 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. EH modes are often possible in metal waveguides with inhomogeneous dielectrics and also in optical fibers.

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 are often possible in metal waveguides with inhomogeneous dielectrics and also in optical fibers.

We have already encountered two examples of TEM modes in Chapters 11 and 12—transmission-line modes and plane waves. These modes have many attractive properties, not the least of which is that they can propagate without attenuation at all frequencies when material losses are zero. To see why all TEM modes have this characteristic, we note that when $h^2 = \gamma^2 + \omega^2\mu\epsilon = 0$, we have

$$\gamma = j\beta = j\omega\sqrt{\mu\epsilon} \qquad \text{(TEM modes)}. \tag{13.14}$$

Thus, TEM modes propagate with no attenuation when the medium is lossless, regardless of the frequency of operation. Also, the phase velocity of a TEM mode in a lossless medium is always given by

$$u_p = \frac{\omega}{\beta} = \frac{1}{\sqrt{\mu\epsilon}} \qquad \text{(TEM modes)}. \tag{13.15}$$

Single-conductor waveguides cannot support TEM modes. This is because B-field lines always must close upon themselves (since $\nabla \cdot \mathbf{B} = 0$). According to Ampère's law, however, any closed loop of magnetic flux must be accompanied by either a conduction current or a displacement current flowing through the loop. For a TEM mode, B-field lines can exist only in the transverse plane. But if there is no longitudinal D-field (since $E_z = 0$) and no longitudinal conduction current (since there is no inner conductor) inside the waveguide, then no B-field loops can exist.

Waveguides that have only one conductor cannot support TEM modes, but they can support the other types of modes listed, which are all types of ***waveguide modes***. As we shall see in this chapter, waveguide modes behave quite differently than TEM modes, particularly at lower frequencies. Since most metal waveguides share many basic characteristics, we will carefully analyze the characteristics of a specific type—rectangular waveguides—and then offer general comments about several other types of metal waveguides. Likewise, since most dielectric waveguides have many similarities, we will analyze the modes of the dielectric slab waveguides and then discuss the characteristics of the fiber-optic cables used in optical communication systems.

13-3 Metal Waveguides

Metal waveguides can be made with nearly any cross-sectional shape. Although they are usually hollow metal tubes, they can also be filled with dielectrics, or even other conductors. In spite of their differences, the modes of all metal waveguides share many common features. This is particularly true when the waveguides are filled with a homogeneous dielectric, such as air.

Consider the metal waveguide shown in Figure 13-4. The cross-sectional shape of this waveguide is arbitrary, but we will consider the dielectric to be lossless and homogeneous. We will assume that the z-axis lies along the waveguide axis. At every point in the waveguide, the longitudinal electric and magnetic fields (E_z and H_z, respectively) satisfy the scalar wave equation (Equation (12.7)),

$$\nabla^2 E_z + k^2 E_z = 0 \tag{13.16}$$

and

$$\nabla^2 H_z + k^2 H_z = 0, \tag{13.17}$$

where $k = \omega\sqrt{\mu\epsilon}$ is the wave number of the medium and

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

is the Laplacian. Since we are interested only in fields with a z-dependence of the form $e^{-\gamma z}$,

$$\frac{\partial^2}{\partial z^2} e^{-\gamma z} = \gamma^2 e^{-\gamma z}.$$

This means that Equations (13.16) and (13.17) can be written as

$$\nabla_t^2 E_z + h^2 E_z = 0 \tag{13.18}$$

and

$$\nabla_t^2 H_z + h^2 H_z = 0, \tag{13.19}$$

where

$$h^2 = \gamma^2 + \omega^2 \mu\epsilon \tag{13.20}$$

is the same variable that appears in Equations (13.9)–(13.13) and ∇_t^2 is the ***transverse Laplacian operator***, which, in Cartesian coordinates, can be expressed as

$$\nabla_t^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}. \tag{13.21}$$

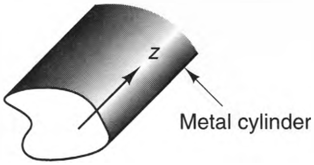


Figure 13-4 A metal waveguide with a lossless, homogeneous dielectric.

Solutions of Equations (13.18) and (13.19) exist for all values of h, but the boundary conditions imposed by the highly conducting metal walls can be satisfied only when h is restricted to certain values, called ***eigenvalues.*** Every waveguide shape has a distinct set of eigenvalues and corresponding modal fields. If the metal walls are perfectly conducting (which is a good approximation for most metal waveguides), each eigenvalue is real and independent of frequency. Solving Equation (13.20) for γ, we find that

$$\gamma = \alpha + j\beta = \sqrt{h^2 - \omega^2\mu\epsilon} = jk\sqrt{1 - (f_c/f)^2} \qquad [\mathrm{m}^{-1}], \tag{13.22}$$

where

$$f_c = \frac{h}{2\pi\sqrt{\mu\epsilon}} \qquad [\mathrm{Hz}] \tag{13.23}$$

is called the mode ***cutoff frequency*** and k is the wave number of the dielectric. When $f > f_c$, the propagation constant γ is imaginary, and thus, the mode is called a ***propagating mode.*** On the other hand, when $f < f_c$, γ is real, which means that the fields decay exponentially with increasing values of z. When waveguide modes are operated below their cutoff frequencies, they are called ***evanescent modes***, or ***nonpropagating modes.***

Even though the eigenvalues of metal waveguides are independent of frequency, most of the other parameters associated with metal waveguides are strong functions of the operating frequency, particularly when operated close to the cutoff frequency. Also, the fields of each mode are distributed differently throughout the waveguide, giving each mode a unique impedance, attenuation, and velocity.

13-3-1 TM MODES IN RECTANGULAR WAVEGUIDES

Figure 13-5 shows a rectangular waveguide with perfectly conducting walls of width a and height b, filled with a lossless dielectric. For TM modes, $H_z = 0$ and $E_z \neq 0$. The scalar wave equation for E_z in Cartesian coordinates can be written as

$$\frac{\partial^2 E_z}{\partial x^2} + \frac{\partial^2 E_z}{\partial y^2} + h^2 E_z = 0. \tag{13.24}$$

Because the conducting walls of a rectangular waveguide lie along the coordinate axes, we can use the separation-of-variables technique (discussed earlier in Chapter 5) to

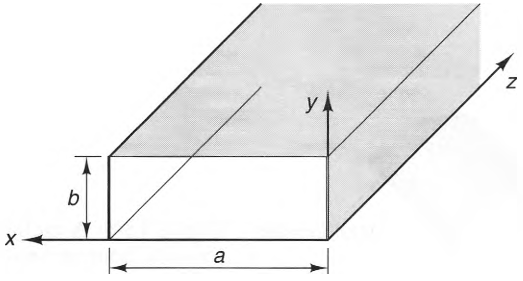


Figure 13-5 Geometry for calculating the modes in rectangular waveguides.

solve this wave equation. This technique greatly simplifies the solution procedure and starts by assuming that the transverse electric field E_z can be written as the product of functions that each depend on a single variable:

$$E_z = X(x)Y(y)e^{-\gamma z}.$$

Substituting this expression into Equation (13.24), we have

$$X''(x)Y(y)e^{-\gamma z} + X(x)Y''(y)e^{-\gamma z} + h^2X(x)Y(y)e^{-\gamma z} = 0,$$

where $X''(x) = d^2[X(x)]/dx^2$ and $Y''(y) = d^2[Y(y)]/dy^2$. Dividing both sides of the foregoing expression by $X(x)Y(y)e^{-\gamma z}$, we obtain

$$\frac{X''(x)}{X(x)} + \frac{Y''(y)}{Y(y)} + h^2 = 0. \tag{13.25}$$

In this expression, the term $[X''(x)]/[X(x)]$ appears to be a function of x. However, since none of the other terms are functions of x, this term must actually be a constant. By the same reasoning, the term $[Y''(y)]/[Y(y)]$ must also be a constant. If we denote these constants as $-k_x^2$ and $-k_y^2$, respectively, Equation (13.25) can be written as

$$h^2 = k_x^2 + k_y^2, \tag{13.26}$$

where

$$\frac{d^2X}{dx^2} + k_x^2X = 0 \tag{13.27}$$

$$\frac{d^2Y}{dy^2} + k_y^2Y = 0. \tag{13.28}$$

Equations (13.27) and (13.28) are both homogeneous, second-order differential equations in a single variable. Their general solutions are

$$X(x) = A\cos k_xx + B\sin k_xx \tag{13.29}$$

$$Y(y) = C\cos k_yy + D\sin k_yy, \tag{13.30}$$

which means that the general solution for E_z is

$$E_z = [A\cos k_xx + B\sin k_xx][C\cos k_yy + D\sin k_yy]e^{-\gamma z}. \tag{13.31}$$

We can find the appropriate values of the unknown constants in this expression by requiring that E_z vanish at the perfectly conducting walls of the waveguide. Starting with the wall at $x = 0$, we require that

$$E_z\Big|_{x=0} = [A\cos(0) + B\sin(0)][C\cos k_yy + D\sin k_yy]e^{-\gamma z} = 0.$$

Since the values of y and z are arbitrary along this wall, we must have $A = 0$. Similarly, at $y = 0$, we require that

$$E_z\Big|_{y=0} = [A\cos k_xx + B\sin k_xx][C\cos(0) + D\sin(0)]e^{-\gamma z} = 0.$$

Since x and z are arbitrary along this wall, B or C must be zero. However, if both A and B are zero, then E_z (and all the transverse fields) would be zero at all points within the waveguide. Instead, we choose $C = 0$, so E_z can be expressed in the form

$$E_z = E_o \sin k_x x \sin k_y y \, e^{-\gamma z}, \tag{13.32}$$

where the new constant E_o is simply the product of the constants B and D.

Turning our attention to the wall at $x = a$, we now require that

$$E_z \Big|_{x=a} = E_o \sin k_x a \sin k_y y \, e^{-\gamma z} = 0.$$

This equation is satisfied whenever

$$k_x = \frac{m\pi}{a} \qquad m = 1, 2, \ldots, \infty, \tag{13.33}$$

where we note that $m = 0$ is not allowed, since this would result in $E_z = 0$ at all points inside the waveguide. Similarly, along the wall at $y = b$, we must also require that

$$E_z \Big|_{y=b} = E_o \sin k_x x \sin k_y b \, e^{-\gamma z} = 0,$$

which is satisfied whenever

$$k_y = \frac{n\pi}{b} \qquad n = 1, 2, \ldots, \infty. \tag{13.34}$$

Again, we cannot have $n = 0$, since this results in null fields. Substituting Equations (13.33) and (13.34) into Equation (13.26), we find that the allowed values of h^2 are given by

$$h_{mn}^2 = \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2. \tag{13.35}$$

Substituting Equations (13.33) and (13.34) into Equation (13.32), we can now write the allowed solution for E_z in the form

$$E_z = E_0 \sin\left(\frac{m\pi}{a} x\right) \sin\left(\frac{n\pi}{b} y\right) e^{-\gamma_{mn} z}. \tag{13.36}$$

Using Equations (13.22) and (13.36), we find that the propagation constant is

$$\gamma_{mn} = \alpha_{mn} + j\beta_{mn} = \sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 - \omega^2 \mu \epsilon}$$

$$= jk\sqrt{1 - \left(\frac{f_{c_{mn}}}{f}\right)^2}, \tag{13.37}$$

where the cutoff frequency for each mode is

$$f_{c_{mn}} = \frac{1}{2\sqrt{\mu\epsilon}} \sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2}. \tag{13.38}$$

Now that we have a complete description of the longitudinal field E_z, the rest of the field components of the TM modes can be found by simply substituting Equation (13.36) into Equations (13.9)–(13.12), yielding

$$E_x = -E_o \frac{\gamma_{mn}}{h_{mn}^2}\left(\frac{m\pi}{a}\right)\cos\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right)e^{-\gamma_{mn}z} \qquad (13.39)$$

$$E_y = -E_o \frac{\gamma_{mn}}{h_{mn}^2}\left(\frac{n\pi}{b}\right)\sin\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right)e^{-\gamma_{mn}z} \qquad (13.40)$$

$$E_z = E_o \sin\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right)e^{-\gamma_{mn}z} \qquad \text{(TM modes)} \qquad (13.41)$$

$$H_x = E_o \frac{j\omega\epsilon}{h_{mn}^2}\left(\frac{n\pi}{b}\right)\sin\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right)e^{-\gamma_{mn}z} \qquad (13.42)$$

$$H_y = -E_o \frac{j\omega\epsilon}{h_{mn}^2}\left(\frac{m\pi}{a}\right)\cos\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right)e^{-\gamma_{mn}z} \qquad (13.43)$$

$$H_z = 0. \qquad (13.44)$$

Figure 13-6 shows the electric and magnetic field patterns of some of the lower order TM modes in rectangular waveguides. Also shown in this figure are the surface currents $\mathbf{J}_s$ on the conducting walls for each mode. These are obtained by using Equation (12.135),

$$\mathbf{J}_s = \hat{\mathbf{a}}_n \times \mathbf{H}_s, \qquad (13.45)$$

where $\mathbf{H}_s$ is the magnetic field at the wall and the unit vector $\hat{\mathbf{a}}_n$ points outward from each wall into the waveguide. In addition to having different modal field patterns, we will show later in this section that each mode has many other distinct operating characteristics.

Example 13-1

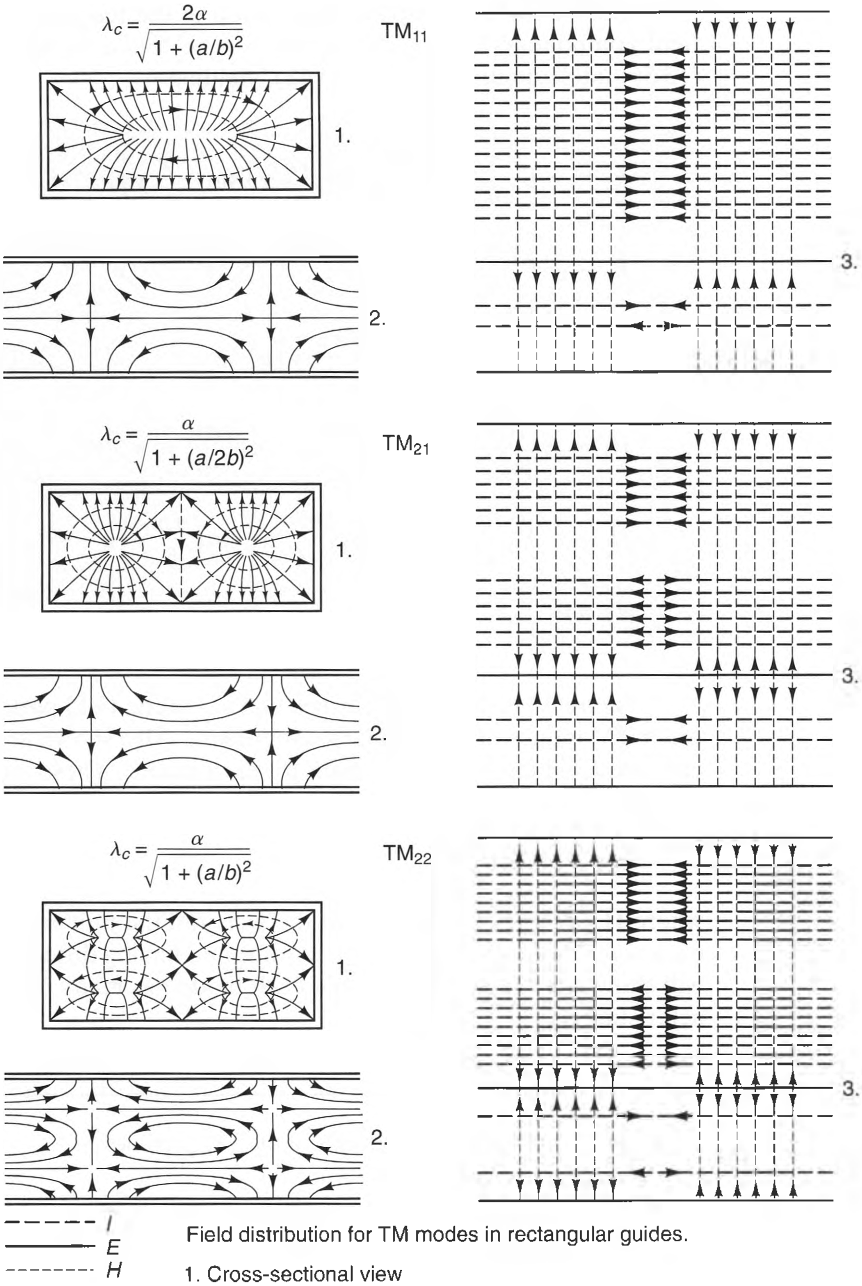
Find expressions for the wall currents of the TM$_{mn}$ modes.

Solution:

According to Equation (13.45), the wall currents are controlled by the magnetic fields that exist at the walls. For the wall at $x = 0$, $\hat{\mathbf{a}}_n = \hat{\mathbf{a}}_x$, so we have

$$\mathbf{J}_s\Big|_{x=0} = \hat{\mathbf{a}}_x \times (H_x\hat{\mathbf{a}}_x + H_y\hat{\mathbf{a}}_y) = H_y\hat{\mathbf{a}}_z\Big|_{x=0} = -\hat{\mathbf{a}}_z E_o \frac{j\omega\epsilon}{h_{mn}^2}\left(\frac{m\pi}{a}\right)\sin\left(\frac{n\pi}{b}y\right)e^{-\gamma_{mn}z}.$$

Since the fields are symmetric, the surface current on the wall $x = b$ is given by a similar expression.


$\lambda_c = \frac{2\alpha}{\sqrt{1+(a/b)^2}}$
TM11
1.
2.
3.
$\lambda_c = \frac{\alpha}{\sqrt{1+(a/2b)^2}}$
TM21
1.
2.
3.
$\lambda_c = \frac{\alpha}{\sqrt{1+(a/b)^2}}$
TM22
1.
2.
3.
I
E
H

Field distribution for TM modes in rectangular guides.

1. Cross-sectional view
2. Longitudinal view
3. Surface view

Figure 13-6 TM mode patterns in rectangular waveguides (adapted from N. Marcuvitz, *Waveguide Handbook*, London: Peter Peregrinus Ltd. on behalf of the Institution of Electrical Engineers, 1986).

At the $y = 0$ wall, $\hat{\mathbf{a}}_n = \hat{\mathbf{a}}_y$. Substituting, we obtain

$$\mathbf{J}_s \bigg|_{y=0} = \hat{\mathbf{a}}_y \times (H_x \hat{\mathbf{a}}_x + H_y \hat{\mathbf{a}}_y) = -H_x \hat{\mathbf{a}}_z \bigg|_{y=0} = -\hat{\mathbf{a}}_z E_o \frac{j\omega\epsilon}{h_{mn}^2}\left(\frac{n\pi}{b}\right)\sin\left(\frac{m\pi}{a}x\right)e^{-\gamma_{mn}z}.$$

A similar expression can be derived along the wall at $y = b$.

From these expressions, we see that the wall currents for all the TM modes arc directed along the waveguide axis. This is true for all the modal currents depicted in Figure 13-6.

13-3-2 TE MODES IN RECTANGULAR WAVEGUIDES

For TE modes, $E_z = 0$ and $H_z \neq 0$, and the wave equation for H_z in Cartesian coordinates can be written as

$$\frac{\partial^2 H_z}{\partial x^2} + \frac{\partial^2 H_z}{\partial y^2} + h^2 H_z = 0. \tag{13.46}$$

Since the boundary conditions for TE modes are also enforced on surfaces that lie along the coordinate axes, we can again use the separation-of-variables technique to find the general solution for H_z. Using the same steps that we used for the TM modes, we can write the general solution for H_z in the form

$$H_z = [A\cos k_x x + B\sin k_x x][C\cos k_y y + D\sin k_y y]e^{-\gamma z}. \tag{13.47}$$

where, again, we have

$$h^2 = k_x^2 + k_y^2. \tag{13.48}$$

Before we can evaluate the unknown constants in Equation (13.47), we must first take a closer look at the boundary conditions that must be imposed on H_z. Unlike the tangential components of the electric field, there is no reason to suppose that H_z vanishes at the perfectly conducting walls. However, we can derive a useful boundary condition for H_z by considering the geometry of Figure 13-7, which shows the interface between a dielectric and a perfect conductor. Remembering that the only nonzero fields at the surface of a perfect conductor are the tangential magnetic and the normal electric fields, we can write Maxwell's curl-**H** equation as

$$\left[\frac{\partial H_{2t}}{\partial(1t)} - \frac{\partial H_{1t}}{\partial(2t)}\right]\hat{\mathbf{a}}_n - \frac{\partial H_{2t}}{\partial n}\hat{\mathbf{a}}_{1t} + \frac{\partial H_{1t}}{\partial n}\hat{\mathbf{a}}_{2t} = j\omega\epsilon E_n \hat{\mathbf{a}}_n$$

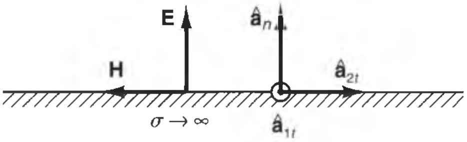


Figure 13-7 Geometry for determining magnetic field boundary conditions at the surface of a perfect conductor.

where $\hat{\mathbf{a}}_n$ is the direction normal to the surface, and $\hat{\mathbf{a}}_{1t}$ and $\hat{\mathbf{a}}_{2t}$ are the two tangential directions. The right-hand side of this expression has only an n component, so the $(1t)$ and $(2t)$ components on the left-hand side must be zero. This means that

$$\frac{\partial H_{1t}}{\partial n} = \frac{\partial H_{2t}}{\partial n} = 0.$$

Thus, we can write the H-field boundary condition at a conducting surface as

$$\frac{\partial \mathbf{H}_{\text{tan}}}{\partial n} = 0 \qquad \text{(Near a perfectly conducting surface)}, \tag{13.49}$$

which, in words, states that the tangential magnetic field has zero slope along the normal direction at the surface of a perfect conductor.

We can now return to the matter of determining the unknown constants in the general solution for H_z. (See Equation (13.47)). Starting with the wall at $x = 0$, the magnetic field boundary condition requires that

$$\left.\frac{\partial H_z}{\partial x}\right|_{x=0} = [-k_x A \sin(0) + k_x B \cos(0)][C \cos k_y y + D \sin k_y y]e^{-\gamma z} = 0,$$

which is satisfied if $B = 0$. Similarly, at the $y = 0$ wall, we must have

$$\left.\frac{\partial H_z}{\partial y}\right|_{y=0} = [A \cos k_x x + B \sin k_x x][-k_y C \sin(0) + k_y D \cos(0)]e^{-\gamma z} = 0,$$

which is satisfied when $D = 0$. Using $B = D = 0$, we can now write H_z in the form

$$H_z = H_o \cos k_x x \cos k_y y \, e^{-\gamma z}, \tag{13.50}$$

where we have replaced the product AC with the constant H_o.

The appropriate values of k_x and k_y are found by requiring that H_z satisfy the boundary condition at the two remaining walls. At $x = a$, we have

$$\left.\frac{\partial H_z}{\partial x}\right|_{x=a} = -k_x H_o \sin k_x a \cos k_y y \, e^{-\gamma z} = 0.$$

This is satisfied when

$$k_x = \frac{m\pi}{a} \qquad m = 0, 1, 2, \ldots, \infty. \tag{13.51}$$

Similarly, at $y = b$, we require that

$$\left.\frac{\partial H_z}{\partial y}\right|_{y=b} = -k_y H_o \cos k_x x \sin k_y b \, e^{-\gamma z} = 0,$$

which is satisfied when

$$k_y = \frac{n\pi}{b} \qquad n = 0, 1, 2, \ldots, \infty. \tag{13.52}$$

Combining Equations (13.50)–(13.52), we can write the allowed solutions for H_z in the form

$$H_z = H_o \cos\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right)e^{-\gamma_{mn}z}. \tag{13.53}$$

Substituting Equations (13.51) and (13.52) into Equation (13.48), we find that the allowed values of h^2 for TE modes are given by

$$h_{mn}^2 = \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2, \tag{13.54}$$

which is the same expression we encountered for the TM modes. (See Equation (13.35)). As a result, the modal propagation constants γ_{mn} and cutoff frequencies $f_{c_{mn}}$ for TE modes are the same as for TM modes; that is,

$$\gamma_{mn} = \alpha_{mn} + j\beta_{mn} = \sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 - \omega^2\mu\epsilon}$$

$$= j\omega\sqrt{\mu\epsilon}\sqrt{1 - \left(\frac{f_{c_{mn}}}{f}\right)^2}, \tag{13.55}$$

where

$$f_{c_{mn}} = \frac{1}{2\sqrt{\mu\epsilon}}\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2}. \tag{13.56}$$

However, unlike the TM modes, where neither m nor n can be zero, H_z for TE modes vanishes only when both m and n are zero. Hence, either m or n (but not both) can be zero for TE modes.

We obtain the complete listing of all the field components for TE$_{mn}$ modes by substituting Equation (13.53) for H_z into Equations (13.9)–(13.12):

$$E_x = H_o\frac{j\omega\mu}{h_{mn}^2}\left(\frac{n\pi}{b}\right)\cos\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right)e^{-\gamma_{mn}z} \tag{13.57}$$

$$E_y = -H_o\frac{j\omega\mu}{h_{mn}^2}\left(\frac{m\pi}{a}\right)\sin\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right)e^{-\gamma_{mn}z} \tag{13.58}$$

$$E_z = 0$$

$$H_x = H_o\frac{\gamma_{mn}}{h_{mn}^2}\left(\frac{m\pi}{a}\right)\sin\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right)e^{-\gamma_{mn}z} \qquad \text{(TE modes)} \tag{13.59}$$

$$H_y = H_o\frac{\gamma_{mn}}{h_{mn}^2}\left(\frac{n\pi}{b}\right)\cos\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right)e^{-\gamma_{mn}z} \tag{13.60}$$

$$H_z = H_o\cos\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right)e^{-\gamma_{mn}z}. \tag{13.61}$$

Figure 13-8 shows the electric and magnetic fields associated with some of the lowest order TE modes in rectangular waveguides. Also shown are the surface currents along the conducting walls. Comparing these modal patterns with the TM_{mn} modal patterns (Figure 13-6), we see that the TE modes behave differently than the TM_{mn} modes, even when the modal parameters m and n are the same.

13-3-3 MODAL HIERARCHY AND THE DOMINANT RANGE

As we have seen in the preceding sections, an infinite number of distinct modes can exist in a rectangular waveguide. Each mode has a cutoff frequency, which means that it can propagate signals and energy over long distances only when it is operated above cutoff. Since each mode has different operating characteristics, it is important to know the order in which the modes come "on-line" as the operating frequency is increased.

The dominant mode of a waveguide is the mode with the lowest cutoff frequency. For rectangular waveguides, the TE_{10} mode is the dominant mode.[1] From Equations (13.57)–(13.61), the field components of the TE_{10} mode are

$$\left.\begin{aligned} E_y &= -j\omega\mu H_o\left(\frac{a}{\pi}\right)\sin\left(\frac{\pi}{a}x\right)e^{-\gamma_{10}z} && (13.62)\\ H_x &= \gamma_{10}H_o\left(\frac{a}{\pi}\right)\sin\left(\frac{\pi}{a}x\right)e^{-\gamma_{10}z} && (13.63)\\ H_z &= H_o\cos\left(\frac{\pi}{a}x\right)e^{-\gamma_{10}z} && (13.64)\end{aligned}\right\}\quad (\mathrm{TE}_{10}\text{ mode}).$$

Also,

$$f_{c_{10}} = \frac{1}{2a\sqrt{\mu\epsilon}} = \frac{u_{\mathrm{TEM}}}{2a}, \tag{13.65}$$

and

$$\gamma_{10} = j\beta_{10} = j\omega\sqrt{\mu\epsilon}\sqrt{1-\left(\frac{f_{c_{10}}}{f}\right)^2}, \tag{13.66}$$

where $u_{\mathrm{TEM}} = 1/\sqrt{\mu\epsilon}$ is the phase velocity of a TEM wave (such as a plane wave) in the same dielectric.

[1] This assumes that the wall dimensions have a ratio $a/b > 1$, which is the standard convention. If, however, $a/b < 1$, the TE_{01} mode is the dominant mode.

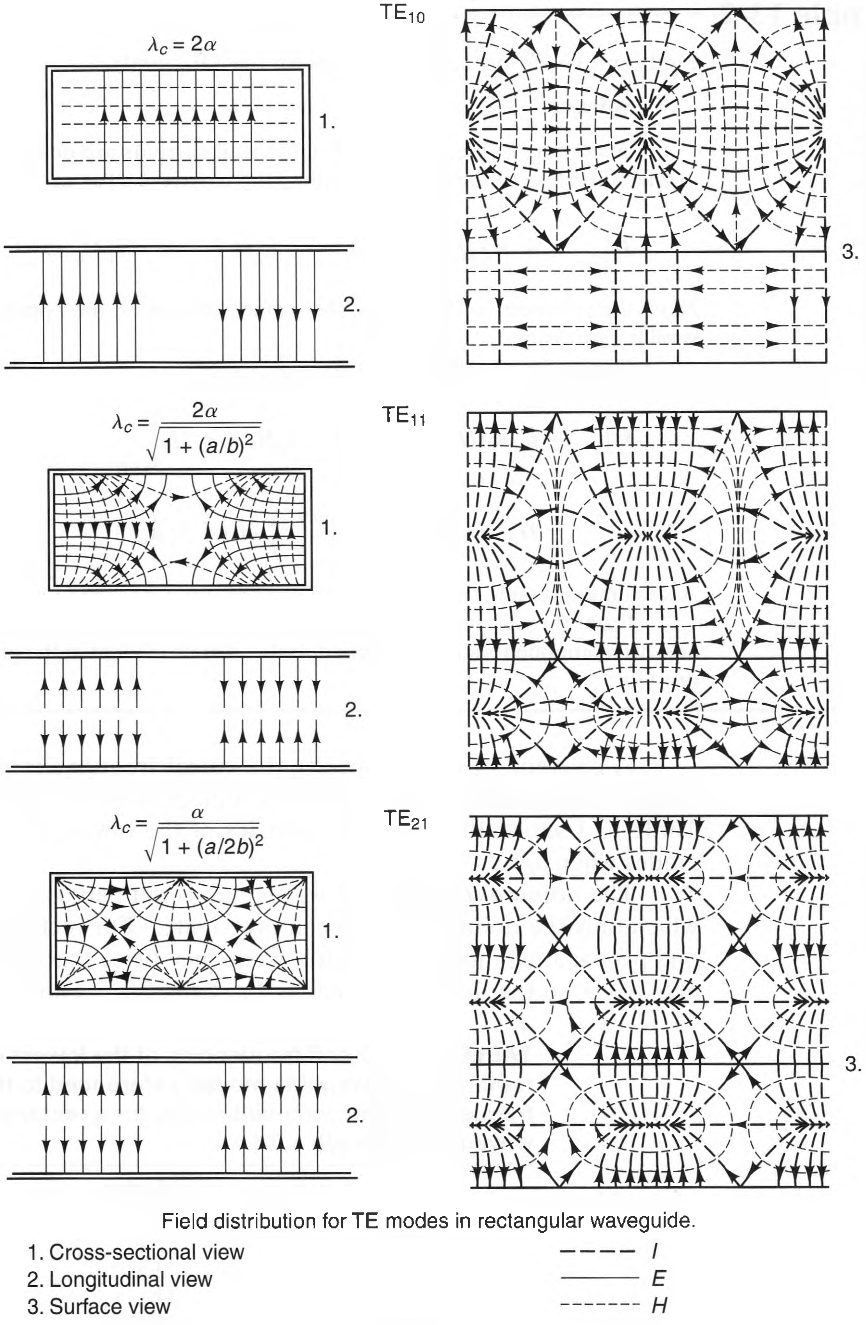

TE10
$\lambda_c = 2\alpha$
1.
2.
3.
TE11
$\lambda_c = \frac{2\alpha}{\sqrt{1 + (a/b)^2}}$
1.
2.
TE21
$\lambda_c = \frac{\alpha}{\sqrt{1 + (a/2b)^2}}$
1.
2.
3.
Field distribution for TE modes in rectangular waveguide.
1. Cross-sectional view
2. Longitudinal view
3. Surface view
---- I
——— E
------- H

Figure 13-8 TE mode patterns in rectangular waveguides (adapted from N. Marcuvitz, *Waveguide Handbook*, London: Peter Peregrinus Ltd. on behalf of the Institution of Electrical Engineers, 1986).

Example 13-2

Find expressions for the wall currents for the TE_{10} mode.

Solution:

According to Equation (13.45), the wall currents are determined by the H-fields and the normal direction to the surfaces. Along the narrow wall at $x = 0$, $\hat{\mathbf{a}}_n = \hat{\mathbf{a}}_x$, so we obtain

$$\mathbf{J}_s\bigg|_{x=0} = \hat{\mathbf{a}}_x \times (H_x\hat{\mathbf{a}}_x + H_z\hat{\mathbf{a}}_z) = -\hat{\mathbf{a}}_y H_z\bigg|_{x=0} = -\hat{\mathbf{a}}_y H_0 e^{-\gamma_{10}z}.$$

From the symmetry of the fields, the surface current on the opposite narrow wall is given by a similar expression.

On the broad wall at $y = 0$, $\hat{\mathbf{a}}_n = \hat{\mathbf{a}}_y$, so we have

$$\mathbf{J}_s\bigg|_{y=0} = \hat{\mathbf{a}}_y \times (H_x\hat{\mathbf{a}}_x + H_z\hat{\mathbf{a}}_z) = \hat{\mathbf{a}}_x H_z - \hat{\mathbf{a}}_z H_x\bigg|_{y=0}$$

$$= H_0\left(\cos\left(\frac{\pi}{a}x\right)\hat{\mathbf{a}}_x - \left(\frac{a\gamma_{10}}{\pi}\right)\sin\left(\frac{\pi}{a}x\right)\hat{\mathbf{a}}_z\right)e^{-\gamma_{10}z}.$$

A similar expression can be derived along the opposite wall at $y = b$.

From these expressions, we see that $\mathbf{J}_s$ has only one component along the narrow walls and two components along the broad walls. This is evident in the graph of the TE_{10} mode in Figure 13-8.

A convenient way to specify the cutoff frequencies of waveguide modes is in terms of the cutoff frequency of the dominant mode ($f_{c_{10}}$ for rectangular waveguides). Table 13.1 shows the cutoff frequencies of the lowest order rectangular waveguide modes when $a/b = 2.1$.

The ***dominant range*** of a waveguide is defined as the range of frequencies for which only the dominant mode can propagate. For rectangular waveguides, the second propagating mode has a cutoff frequency twice that of the dominant mode when $a/b > \sqrt{3}$, so the dominant range is $f_{c_{10}} < f < 2f_{c_{10}}$. However, since the propagation

TABLE 13.1 Cutoff frequencies of the lowest order rectangular waveguide modes (referenced to the cutoff frequency of the dominant mode) for a rectangular waveguide with *a*/*b* = 2.1.

$f_c / f_{c_{10}}$	Modes
1.0	TE_{10}
2.0	TE_{20}
2.1	TE_{01}
2.326	$\mathrm{TE}_{11}, \mathrm{TM}_{11}$
2.9	$\mathrm{TE}_{21}, \mathrm{TM}_{21}$
3.0	TE_{30}
3.662	$\mathrm{TE}_{31}, \mathrm{TM}_{31}$
4.0	TE_{40}

parameters of any mode are rapid functions of frequency near cutoff, it is best to stay away from the modal cutoff frequencies. A good rule of thumb for rectangular waveguides is to operate within the so-called usable range, $1.25 f_{c_{10}} < f < 0.95 f_c$, where f_c is the cutoff frequency of the next highest mode.

Example 13-3

Find the dominant range of type WR-75 rectangular waveguide, which is filled with air and has inside dimensions $a = 1.905$ [cm] and $b = 0.953$ [cm].

Solution:

Using Equation (13.65), we obtain

$$f_{c_{10}} = \frac{3 \times 10^8}{2 \times (.01905)} = 7.87 \qquad \text{[GHz]}.$$

Also, since $a/b = 1.99 > \sqrt{3}$, the dominant range is

$$7.87 < f < (2) \times 7.87 \qquad \text{[GHz]},$$

or

$$7.87 < f < 15.74 \qquad \text{[GHz]}.$$

The usable frequency range is

$$(1.25) \times 7.87 < f < (0.95) \times 15.74 \qquad \text{[GHz]},$$

or

$$9.84 < f < 15.35 \qquad \text{[GHz]}.$$

The modes of other metal waveguides are also specified in terms of mode indices. However, because of differences in the mathematical formulations, their modal hierarchy is often different. For instance, the dominant mode for a circular waveguide is the TE_{11} mode, whose field patterns are shown in Figure 13-9.

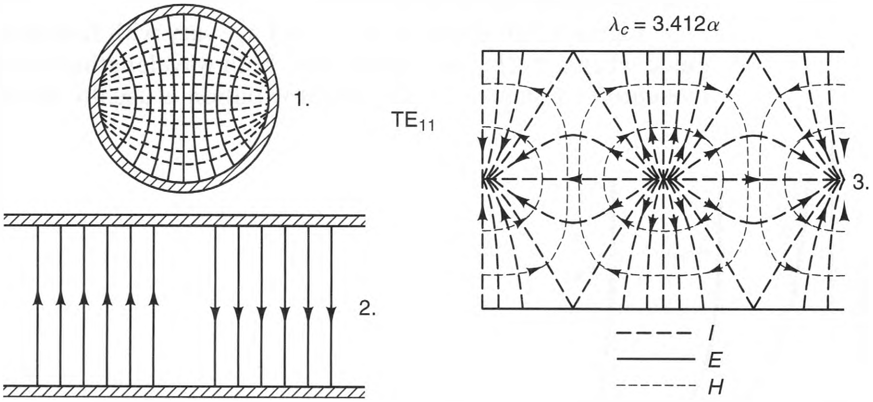


Figure 13-9 TE_{11} mode pattern for circular waveguides (adapted from N. Marcuvitz, *Waveguide Handbook*, London: Peter Peregrinus Ltd. on behalf of the Institution of Electrical Engineers, 1986).

As can be seen, this mode looks much like a "rounded" version of the TE_{10} in a rectangular waveguide.

13-3-4 PROPERTIES OF PROPAGATING WAVEGUIDE MODES

Waveguide modes are propagating fields when they are operated above cutoff. Like the TEM waves we encountered on transmission lines and as plane waves, propagating waveguide modes can be described in terms of their wavelengths, phase and group velocities, and wave impedances, as well as the power they transport.

Wavelength. Above cutoff, each waveguide mode contains the phase term $e^{-j\beta z}$. Using Equations (13.22), we can write β as

$$\beta = k\sqrt{1-\left(\frac{f_c}{f}\right)^2} \quad f > f_c, \tag{13.67}$$

where f_c is the cutoff frequency of the mode and $k = \omega\sqrt{\mu\epsilon}$ is the wave number of the dielectric. Just as for a TEM wave, the wavelength λ_g of a waveguide mode is defined as the distance between points of identical phase along the direction of propagation. Hence, $\beta\lambda_g = 2\pi$, which means that

$$\lambda_g = \frac{2\pi}{\beta}. \tag{13.68}$$

Substituting Equation (13.67) into Equation (13.68), we obtain

$$\lambda_g = \frac{\lambda}{\sqrt{1-\left(\frac{f_c}{f}\right)^2}}. \tag{13.69}$$

where $\lambda = 2\pi/k$ is the wavelength of a TEM wave (such as a plane wave) of the same frequency in the same dielectric.

Figure 13-10 shows how $\lambda_{g_{10}}$ and $\lambda_{g_{20}}$ vary with frequency in a rectangular waveguide when $a = 2b$. As can be seen, both wavelengths vary rapidly near their respective cutoff frequencies. This rapid variation is usually deemed undesirable. On the

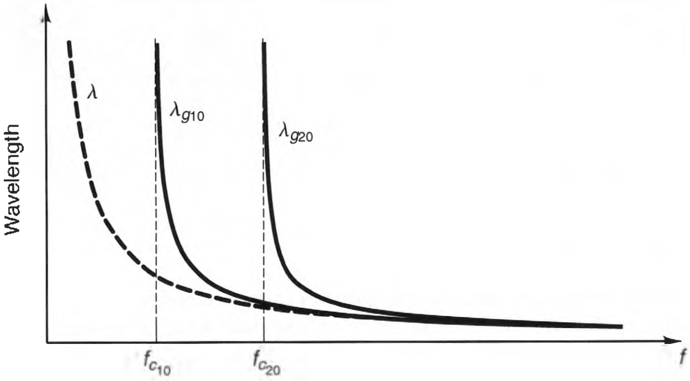


Figure 13-10 Wavelength vs. frequency for a TEM wave, and the TE_{10} and TE_{20} modes in a rectangular waveguide.

other hand, as $f \to \infty$, both $\lambda_{g_{10}}$ and $\lambda_{g_{20}}$ approach λ asymptotically. This is an indication that waveguide modes share many of the characteristics of TEM modes when they are operated far above cutoff.

Another parameter that is often specified for a waveguide mode is its ***cutoff wavelength*** λ_c. This is defined as the free-space wavelength (i.e., wavelength in a vacuum) at the modal cutoff frequency. From this definition, we can write λ_c in the form

$$\lambda_c = \frac{c}{f_c}, \tag{13.70}$$

where c is the speed of light in a vacuum. Since the dominant mode has the lowest cutoff frequency, it also has the largest cutoff wavelength. For an air-filled, rectangular waveguide, the cutoff wavelength of the TE_{10} mode is

$$\lambda_{c_{10}} = 2a, \tag{13.71}$$

where a is the width of the broad wall.

Example 13-4

An air-filled, rectangular waveguide has dimensions $a = 1$ [cm] and $b = 0.6$ [cm]. Calculate the cutoff wavelengths of the TE_{10} and the TE_{20} modes. If the waveguide is operated at a frequency of 18 [GHz], calculate the guide wavelength, and compare it to the free-space wavelength.

Solution:

From Equation (13.71), the cutoff wavelength for the TE_{10} mode is

$$\lambda_{c_{10}} = 2 \times 1.0 \text{ [cm]} = 2 \text{ [cm]}.$$

Given that $f_{c_{20}} = 2f_{c_{10}}$, we can conclude from Equation (13.70) that

$$\lambda_{c_{20}} = 0.5\lambda_{c_{10}} = 1 \text{ [cm]}.$$

At 18 [GHz], the free-space wavelength is

$$\lambda_o = \frac{c}{f} = \frac{3 \times 10^{10} \text{ [cm/s]}}{18 \times 10^{9} \text{ [s}^{-1}\text{]}} = 1.67 \text{ [cm]}.$$

Since $\lambda_{c_{10}} > \lambda_o$ and $\lambda_o > \lambda_{c_{20}}$, only the TE_{10} mode propagates.

Finally, the guide wavelength at this frequency can be found from Equation (13.69). Using $f_{c_{10}} = c/2a = 15$ [GHz] and $\lambda = \lambda_o$ (since the dielectric is air), we obtain

$$\lambda_{g_{10}} = \frac{1.67 \text{ [cm]}}{\sqrt{1 - \left(\frac{15}{18}\right)^2}} = 3.02 \text{ [cm]}.$$

Notice that since the operating frequency is relatively close to the cuttoff frequency, the guide wavelength is significantly longer than the free-space wavelength.

Wave Velocities and Dispersion. The basic formulas that relate the phase and group velocities to the phase constant β are the same for waveguide modes as they are for TEM waves. (See Equations (12.21) and (12.66).) Hence,

$$u_p = \frac{\omega}{\beta}, \tag{13.72a}$$

and

$$u_g = \frac{\partial \omega}{\partial \beta} = \left(\frac{\partial \beta}{\partial \omega}\right)^{-1}, \tag{13.72b}$$

respectively. Because β is not a linear function of frequency for waveguide modes, u_p and u_g are different, even when the dielectric is lossless. This can be seen by first writing the phase constant β as an explicit function of frequency. Using Equation (13.67) and remembering that $k = 2\pi f\sqrt{\mu\epsilon}$, we can write

$$\beta = 2\pi f\sqrt{\mu\epsilon}\sqrt{1 - \left(\frac{f_c}{f}\right)^2} \qquad f > f_c. \tag{13.73}$$

Substituting this into Equations (13.72a) and (13.72b), we obtain

$$u_p = \frac{u_{\text{TEM}}}{\sqrt{1 - \left(\frac{f_c}{f}\right)^2}} \tag{13.74}$$

and

$$u_g = u_{\text{TEM}}\sqrt{1 - \left(\frac{f_c}{f}\right)^2}, \tag{13.75}$$

where $u_{\text{TEM}} = 1/\sqrt{\mu\epsilon}$ is the velocity of a TEM wave in the dielectric.

Figure 13-11 shows how u_p and u_g vary with frequency for a typical waveguide mode. As can be seen, both 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 u_p and u_g behave differently near cutoff: 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 relatively, which states that energy and matter cannot travel faster than the speed of light in a vacuum. But the behavior 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, and which is always less than or equal to c.

The rapid variations in both u_p and u_g near cutoff are nearly always undesirable. The behavior of u_g near cutoff is particularly troublesome when a waveguide is used to

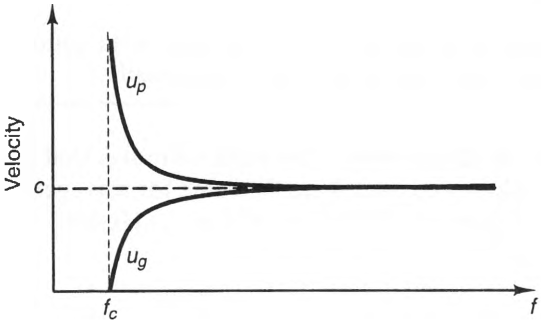


Figure 13-11 Phase and group velocities vs. frequency for waveguide modes.

propagate modulated signals, such as pulses. As we saw earlier in Chapter 11, the differential propagation delay (or pulse spreading) per meter Δt is related to the group velocity given by the expression

$$\Delta t = \left.\frac{1}{u_g}\right|_{\max} - \left.\frac{1}{u_g}\right|_{\min}, \tag{13.76}$$

where $\left.\frac{1}{u_g}\right|_{\max}$ and $\left.\frac{1}{u_g}\right|_{\min}$ are the maximum and minimum inverse group velocities within the bandwidth, respectively. Substituting Equation (13.75) into Equation (13.76), we obtain

$$\Delta t = \frac{1}{u_{\text{TEM}}}\left[\frac{1}{\sqrt{1-\left(\frac{f_c}{f_{\min}}\right)^2}} - \frac{1}{\sqrt{1-\left(\frac{f_c}{f_{\max}}\right)^2}}\right] \quad [\text{s/m}]. \tag{13.77}$$

Unlike the modes on transmission lines, which exhibit differential propagation delays (i.e., dispersion) only when the dielectrics are lossy or frequency dependent, Equation (13.77) shows that waveguide modes are *always* dispersive, even when the dielectric is lossless and walls are perfectly conducting. This dispersion is called ***waveguide dispersion***. Waveguides can also exhibit ***material dispersion***, which occurs when the dielectric parameters are frequency dependent.

Example 13-5

Calculate the differential delay experienced by the 18 [GHz] sinusoidal radar pulse with 5 [ns] duration (shown in Figure 13-12a) as it propagates in a 1 meter length of air-filled type WR-51 rectangular waveguide.

Solution:

The Fourier transform of the pulse, shown in Figure 13-12b, is a sinc function. It has fre-

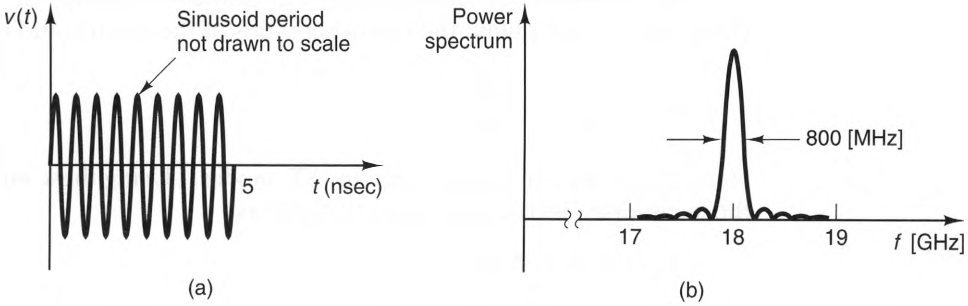


Figure 13-12 A microwave pulse: a) Time-domain representation. b) Frequency-domain power spectrum.

quency components over an infinite range, but most of the signal's energy is contained in a bandwidth

$$\Delta f \approx 4/T = \frac{4}{5 \times 10^{-9}} = 800\,[\text{MHz}].$$

Thus,

$$f_{\max} = f + \Delta f/2 = 18.4\,[\text{GHz}]$$

$$f_{\min} = f - \Delta f/2 = 17.6\,[\text{GHz}].$$

For WR-51 waveguide, $a = 1.295$ [cm] and $b = 0.648$ [cm]. Using Equation (13.65), we obtain

$$f_{c_{10}} = \frac{u_{\text{TEM}}}{2a} = \frac{3 \times 10^8}{2 \times 1.295 \times 10^{-2}} = 11.583\,[\text{GHz}].$$

All other modes are cut off within the waveform bandwidth, so we can assume that all the signal's energy is carried by the TE_{10} mode. Using Equation (13.77), the differential time delay is

$$\Delta t = \frac{1}{3 \times 10^8}\left[\frac{1}{\sqrt{1 - \left(\frac{11.583}{17.6}\right)^2}} - \frac{1}{\sqrt{1 - \left(\frac{11.583}{18.4}\right)^2}}\right] = 0.137\,[\text{ns/m}].$$

Thus, the pulse width grows 0.137 [ns] for every meter that it propagates.

Wave Impedance. The ratio of the transverse electric and magnetic fields of a mode is called its ***wave impedance***, defined by the relation

$$Z \equiv \frac{|\mathbf{E}_t|}{|\mathbf{H}_t|}, \tag{13.78}$$

where $|\mathbf{E}_t|$ and $|\mathbf{H}_t|$ are the magnitudes of the transverse electric and magnetic fields, respectively. For plane waves, the wave impedance is the intrinsic impedance η of the dielectric, which, if it is lossless, is independent of frequency. For waveguide modes, however, we will now show that the wave impedance is related not only to the intrinsic impedance of the dielectric, but also to the type of mode (TE, TM, etc.) and the frequency of operation.

Let us start by considering TE modes. Setting $E_z = 0$ in Equations (13.9)–(13.12), we obtain the following ratios of the electric and magnetic fields:

$$Z_{\text{TE}} = \frac{E_x}{H_y} = -\frac{E_y}{H_x} = \frac{j\omega\mu}{\gamma}, \tag{13.79}$$

Here, Z_{TE} is the wave impedance for TE modes. Using these equalities, we can write the transverse electric and magnetic fields as

$$\mathbf{E}_t = E_x\hat{\mathbf{a}}_x + E_y\hat{\mathbf{a}}_y \tag{13.80}$$

$$\mathbf{H}_t = \frac{1}{Z_{\text{TE}}}[-E_y\hat{\mathbf{a}}_x + E_x\hat{\mathbf{a}}_y]. \tag{13.81}$$

We can easily see from these two expressions that $\mathbf{E}_t \cdot \mathbf{H}_t = 0$, which means that the transverse electric and magnetic fields of TE modes are mutually orthogonal throughout the waveguide. Using Equations (13.79) and (13.81), we can relate $\mathbf{E}_t$ and $\mathbf{H}_t$ to the wave impedance Z_{TE} by the vector formula

$$\mathbf{H}_t = \frac{1}{Z_{\mathrm{TE}}} \hat{\mathbf{a}}_z \times \mathbf{E}_t. \tag{13.82}$$

When a TE mode is operated above cutoff, $\gamma = jk\sqrt{1 - (f_c/f)^2}$. Substituting this into Equation (13.79), we obtain the formula:

$$Z_{\mathrm{TE}} = \frac{\eta}{\sqrt{1 - \left(\frac{f_c}{f}\right)^2}}, \tag{13.83}$$

where $\eta = \sqrt{\mu/\epsilon}$ is the intrinsic impedance of the dielectric. Figure 13-13 shows a plot of Z_{TE} as a function of frequency for a typical TE mode. As can be seen, Z_{TE} approaches infinity near cutoff, which means that the electric field $\mathbf{E}_t$ is much greater than the magnetic field $\mathbf{H}_t$. On the other hand, as $f \to \infty$, $Z_{\mathrm{TE}} \to \eta$, which is yet another indication that waveguide modes behave much like TEM modes when operated far above cutoff.

For TM modes, we can set $H_z = 0$ in Equations (13.9)–(13.12) to obtain the following ratios of the transverse electric and magnetic field components that define the wave impedance Z_{TM}:

$$Z_{\mathrm{TM}} = \frac{E_x}{H_y} = -\frac{E_y}{H_x} = \frac{\gamma}{j\omega\epsilon}. \tag{13.84}$$

From these expressions, it follows that $\mathbf{E}_t$ and $\mathbf{H}_t$ can be related by the vector formula

$$\mathbf{H}_t = \frac{1}{Z_{\mathrm{TM}}} \hat{\mathbf{a}}_z \times \mathbf{E}_t. \tag{13.85}$$

Also, substituting $\gamma = jk\sqrt{1 - (f_c/f)^2}$ into Equation (13.84), we obtain

$$Z_{\mathrm{TM}} = \eta\sqrt{1 - \left(\frac{f_c}{f}\right)^2}. \tag{13.86}$$

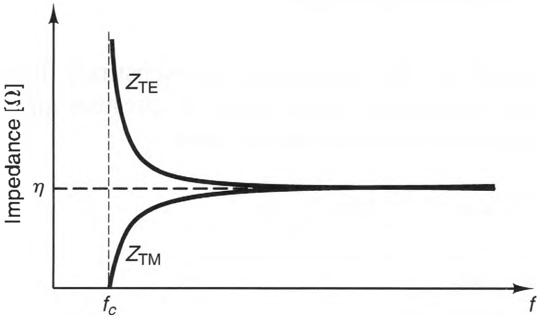


Figure 13-13 TE and TM wave impedances vs. frequency for waveguide modes.

Figure 13-13 shows a plot of Z_{TM} as a function of frequency for a typical TM mode. Like Z_{TE}, $Z_{\text{TM}} \to \eta$ as $f \to \infty$. But near cutoff, $Z_{\text{TM}} \to 0$. Hence, the TM magnetic fields are much stronger than the electric fields at frequencies near cutoff.

Finally, we note from Equations (13.83) and (13.86) that the wave impedances of both TE and TM modes are reactive (i.e., imaginary) when $f < f_c$. This occurs because the propagation constant γ is real below cutoff.

Example 13-6

An air-filled waveguide has dimensions a = 1.25 [cm] and b = 0.2 [cm]. Find the frequency at which the wave impedance of the TE$_{10}$ mode is twice its value in free space.

Solution:

Since the waveguide is filled with air, the intrinsic impedance of the dielectric is $\eta = \eta_o$. From Equation (13.83), the ratio of the wave impedance to the intrinsic impedance to free space is

$$\frac{Z_{\text{TE}}}{\eta_o} = \frac{1}{\sqrt{1 - \left(\frac{f_c}{f}\right)^2}}.$$

This ratio equals 2.0 when $\sqrt{1 - (f_c/f)^2} = 1/2$, which occurs when $f = f_c/\sqrt{0.75}$. The cutoff frequency for the TE$_{10}$ mode is $f_c = c/(2a) = 12$ [GHz]. Hence, $Z_{\text{TE}} = 2\eta_o$ when f = 13.86 [GHz].

Transmitted Power. The power P transported by a waveguide mode can be found by integrating the average Poynting vector $\mathscr{S}_{\text{ave}}$ over the cross section of the waveguide. From Equation (12.105), the power transmitted by a waveguide mode is given by

$$P = \int_S \mathscr{S}_{\text{ave}} \cdot \mathbf{ds} = \frac{1}{2} \text{Re} \int_S (\mathbf{E} \times \mathbf{H}^*) \cdot \mathbf{ds} \qquad [\text{W}], \tag{13.87}$$

where S is the waveguide cross section, and **E** and **H** are the modal electric and magnetic fields, respectively. Using $E_x = ZH_y$ and $E_y = -ZH_x$ (where Z is the wave impedance of the mode), we can write Equation (13.87) as

$$P = \frac{1}{2} \text{Re}\left(\frac{1}{Z^*}\right) \int_S [|E_x|^2 + |E_y|^2]\, ds. \tag{13.88}$$

To evaluate this expression, all that is necessary is to substitute the appropriate expressions for E_x and E_y that correspond to the mode in question and integrate. For the case of the TE$_{10}$ in a rectangular waveguide, we have

$$E_x = 0, \quad E_y = -j\omega\mu H_o\left(\frac{a}{\pi}\right)\sin\left(\frac{\pi}{a}x\right)$$

$$Z = Z_{\text{TE}} = \frac{j\omega\mu}{\gamma} = \frac{\omega\mu}{\beta}.$$

Substituting these into Equation (13.88) yields

$$P_{10} = \frac{1}{2}\,\omega\mu\beta_{mn}\,|H_o|^2\left(\frac{a}{\pi}\right)^2 \int_0^b\int_0^a \sin^2\left(\frac{\pi}{a}x\right)dxdy.$$

Integrating, we obtain

$$P_{10} = \frac{ab}{4}\,\omega\mu\beta_{10}\,|H_o|^2\left(\frac{a}{\pi}\right)^2 \qquad [\mathrm{W}]. \qquad (13.89)$$

This can be rewritten as

$$P_{10} = |E_{\max}|^2\,(ab)/(4\eta)\,\sqrt{1-\left(\frac{f_{c_{10}}}{f}\right)^2} \qquad [\mathrm{W}], \qquad (13.90)$$

where $\eta = \sqrt{\mu/\epsilon}$ is the intrinsic impedance of the dielectric and $E_{\max} = \omega\mu\,(a/\pi)\,H_o$ is the maximum electric field strength inside the waveguide. Similar expressions can be derived for all other rectangular waveguide modes.

Figure 13-14 shows a plot of P_{10} vs. frequency when the maximum electric field strength $E_{\max}$ is held constant. As can be seen, $P_{10} \approx (ab)|E_{\max}|^2/(4\eta)$ when $f \gg f_{c_{10}}$. Except for a factor of 2 (caused by the sinusoidal taper of the E-field across the width of the waveguide), this is the same result as would be obtained for a plane wave of identical amplitude. (See Equation (12.107)). On the other hand, $P_{10} \rightarrow 0$ as $f \rightarrow f_c$, which shows that significant power can be transmitted only when $E_{\max}$ is very large.

13-3-5 PROPERTIES OF MODES BELOW CUTOFF

When a waveguide mode is operated below its cutoff frequency, its propagation constant γ is a real quantity, which means that its fields decay exponentially as a function of position. For metal waveguides with uniform dielectrics, the attenuation and phase constants (α and β, respectively) are given by Equation (13.22),

$$\alpha = k\sqrt{\left(\frac{f_c}{f}\right)^2 - 1} \qquad [\mathrm{Np/m}] \qquad (f<f_c), \qquad (13.91)$$

$$\beta = 0 \qquad (f<f_c), \qquad (13.92)$$

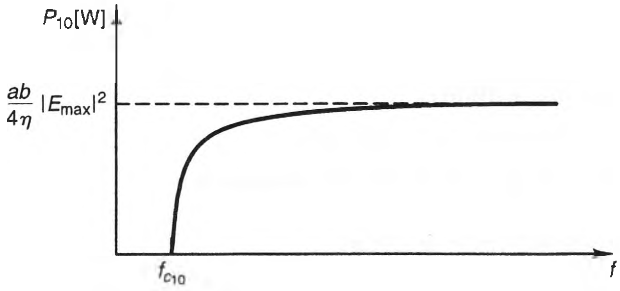


Figure 13-14 Power transmitted in the TE_{10} mode vs. frequency for a fixed maximum electric field $E_{\max}$.

where $k = \omega\sqrt{\mu\epsilon}$ is the wavenumber of the dielectric and f_c is the cutoff frequency. By convention, the positive root of Equation (13.91) is used, yielding $\alpha \geqslant 0$. Since the phase constant β is zero below cutoff, the phases of these modes do not vary with position, which means that they have no wavelength or phase velocity. As a result, these modes are called ***evanescent***, or ***nonpropagating***, ***modes***.

Because evanescent modes decay exponentially, they are not capable of transporting power over large distances. This makes them useful for situations where it is necessary to restrict the amount of power that leaks through holes, seams, and joints in various structures. If the dimensions of these apertures are small enough so that all the waveguide modes are cut off, the leakage power can be controlled. The example that follows shows how this can be accomplished.

Example 13-7

Figure 13-15 shows a microwave furnace, used for drying large objects such as wood. Furnaces of this type operate by converting standard 60-Hz power into microwave power (usually at 2.45 GHz) and directing it into a metal enclosure that contains the objects to be dried. Since the metal walls of the enclosure cannot absorb the energy, all of the microwave energy is absorbed by the object. Also shown in the figure is a viewing port, constructed out of a narrow section of square waveguide.

If the viewing port is 4 [cm] on each side, calculate the minimum length d which guarantees that the field strength at the end is at least 140 [dB] down from the level at the furnace wall.

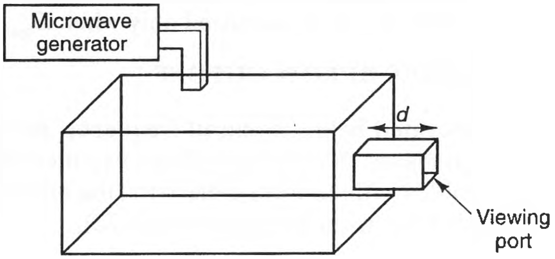


Figure 13-15 A microwave oven with a viewing port.

Solution:

Since the TE$_{10}$ decays most slowly, we will find the minimum length based on this mode. For a 4 [cm] square waveguide, the cutoff frequency of the TE$_{10}$ mode is

$$f_{c_{10}} = \frac{3 \times 10^8}{2 \times 0.04} = 3.75 \qquad [\text{GHz}].$$

From Equation (13.91), the attenuation constant is

$$\alpha = \frac{2\pi \times 2.45 \times 10^9}{3 \times 10^8}\sqrt{\left(\frac{3.75}{2.4}\right)^2 - 1} = 61.6 \qquad [\text{Np/m}].$$

From Equation (12.64), the loss in dB/m is

$$\text{dB loss per meter} = 8.69\alpha = 535.4 \qquad [\text{dB/m}].$$

Hence, the length needed to obtain at least 140 [dB] attenuation is

$$d \geqslant \frac{140\ [\text{dB}]}{524.2\ [\text{dB/m}]} = 0.2627\ [\text{m}] = 26.2\ [\text{cm}].$$

Even though waveguides are most often used in their dominant range, evanescent modes are usually present at discontinuities, such as at bends, junctions between waveguide sections, or holes in waveguide walls. These higher order, evanescent modes appear because no single waveguide mode is capable of satisfying the additional boundary conditions imposed by the discontinuities. But an infinite collection of modes can. Thus, even when a single mode is incident upon a discontinuity, a large number of modes is generated at the discontinuity. Most of these additional modes are higher order modes that are below cutoff and, as a result, are not observed far from the discontinuity. Even so, the presence of such modes changes the wave impedance of the dominant mode in these regions, causing reflections that are similar to those obtained when a lumped load is placed on a transmission line. We will investigate this further when we discuss the equivalent circuits of waveguide discontinuities.

13-3-6 LOSSES IN METAL WAVEGUIDES

Until now, our analysis of metal waveguides has assumed that both the walls and the dielectric are lossless. This assumption has resulted in waveguide modes that propagate with no attenuation when operated above cutoff. As might be expected, that types of behavior changes when losses are present.

Power losses in waveguides are the result of dielectric losses, metal losses, or both. Of these, dielectric losses are easiest to model. This is because the presence of a lossy dielectric does not change the $E_{\text{tan}} = 0$ boundary condition at the walls. Hence, all the expressions derived for the TE and TM modes for the lossless case still apply, except that the permittivity ϵ is now complex, $\epsilon = \epsilon' - j\epsilon''$. Using Equation (13.37), we find that the dielectric attenuation constant above cutoff is given by,

$$\alpha_d = \text{Re}(\gamma) = \text{Re}\left[\sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 - \omega^2\mu(\epsilon' - j\epsilon'')}\right]. \tag{13.93}$$

Figure 13-16 shows a plot of $\alpha_{d_{10}}$ vs. frequency for a waveguide with dimensions $a = 2$ [cm], $b = 1$ [cm], filled with a dielectric with permittivity $\epsilon = (1 - j.01)\,\epsilon_o$. As can be seen from this plot, the attenuation is nonzero at all frequencies, although it is small above 7.5 [GHz], which is the cutoff frequency when $\epsilon'' = 0$. Thus, when ϵ'' is small, the cutoff frequency is essentially the same as in the lossless case. Using Equation (13.93) and the binomial expansion, we can derive the following approximation for α_d above cutoff when the loss is small;

$$\alpha_d \approx \frac{\omega\epsilon''\eta}{2\sqrt{1 - \left(\frac{f_c}{f}\right)^2}} \qquad \text{when} \qquad \frac{\epsilon''}{\epsilon'} \ll 1 - \left(\frac{f_c}{f}\right)^2, \tag{13.94}$$

Here, η is the intrinsic impedance of the dielectric and f_c is the cutoff frequency of the mode when no loss is present.

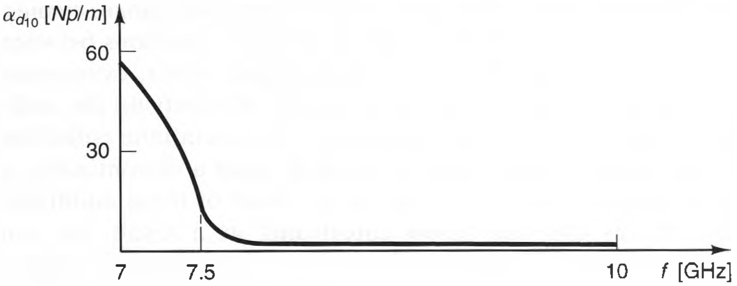


Figure 13-16 The attenuation constant vs. frequency for the dominant mode in a rectangular waveguide with dimensions $a = 2$ [cm], $b = 1$ [cm], and a dielectric with permittivity $\epsilon = (1 - j.01)\,\epsilon_0$.

Whereas the effects of dielectric loss are easy to calculate, the effects of metal losses are less straightforward to model. This is because $E_{\tan}$ is no longer zero at the waveguide walls, which means that the modal fields distribute themselves differently throughout the waveguide cross section. Fortunately, when the wall conductivity is high (such as when the walls are made of copper), we can derive a simple, yet accurate, expression for the attenuation constant α_c due to the conductor losses. When **E** and **H** decay proportional to $e^{-\alpha_c z}$, the power P transmitted through the waveguide will decay as

$$P = P_o e^{-2\alpha_c z} \qquad [\mathrm{W}], \tag{13.95}$$

where P_o is the power at $z = 0$. Differentiating this expression with respect to z, we obtain

$$\frac{dP}{dz} = -2\alpha_c P \qquad [\mathrm{W/m}].$$

Solving for α_c, we get

$$\alpha_c = \frac{-1}{2}\frac{dP/dz}{P}. \tag{13.96}$$

Since $-(dP/dz)$ is the power loss per unit length, we can express the attenuation constant in the form

$$\alpha_c = \frac{1}{2}\frac{P_L}{P} = \frac{1}{2}\frac{\text{Power loss per meter}}{\text{Transmitted power}}. \tag{13.97}$$

When the wall conductivity σ is high, the E- and H-fields distribute themselves throughout the waveguide nearly the same as for the perfectly conducting case. This means that we can obtain excellent estimates of the transmitted power P and the power loss per meter P_L using the E- and H-field expressions derived for the case of perfectly conducting walls. Using Poynting's theorem, we find that the transmitted power P equals the integral of the average Poynting vector over the cross section S of the waveguide. Thus, we can write

$$P = \frac{1}{2}\,\mathrm{Re}\left[\int_S \mathbf{E} \times \mathbf{H}^* \cdot \mathbf{ds}\right] \qquad [\mathrm{W}], \tag{13.98}$$

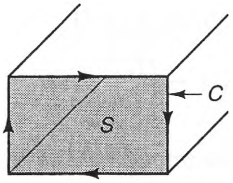


Figure 13-17 Geometry showing the surface S and contour C used to determine the attenuation constant in a rectangular waveguide due to metal losses.

where the waveguide cross section S is shown in Figure 13-17. Similarly, the power lost in the waveguide walls is given by Equation (12.140), which relates the power dissipated per square meter at a conducting surface in terms of the tangential H-field and the surface resistance of the conductor. Integrating this expression around the perimeter of the waveguide cross section, we obtain

$$P_L = \frac{1}{2} R_s \oint_C |\mathbf{H}|^2 d\ell \qquad [\mathrm{W/m}], \tag{13.99}$$

where the integration contour C is the perimeter of the waveguide, shown in Figure 13-17. In Equation (13.99),

$$R_s = \frac{1}{\sigma\delta} = \sqrt{\frac{\omega\mu}{2\sigma}} \qquad [\Omega], \tag{13.100}$$

is the surface resistance, and σ and δ are the conductivity and skin depth of the metal, respectively. Substituting Equations (13.96) and (13.98) into Equation (13.99), we obtain

$$\alpha_c = \frac{R_s}{2} \frac{\oint_C |\mathbf{H}|^2 d\ell}{\mathrm{Re} \int_S \mathbf{E} \times \mathbf{H}^* \cdot \mathbf{ds}}. \tag{13.101}$$

For the TE$_{10}$ mode in a rectangular waveguide, the integral in the denominator of Equation (13.101) has already been evaluated. (See Equation (13.89).) Hence,

$$\mathrm{Re} \int_S \mathbf{E} \times \mathbf{H}^* \cdot \mathbf{ds} = 2P_{10} = \frac{ab}{2} \omega\mu\beta |H_o|^2 \left(\frac{a}{\pi}\right)^2 \qquad [\mathrm{W}]. \tag{13.102}$$

Next, by integrating on all four sides of the contour C, it is straightforward (but tedious) to show that for the TE$_{10}$ mode, we have

$$\oint_C |\mathbf{H}|^2 d\ell = H_o^2 \beta^2 a \left(\frac{a}{\pi}\right)^2 + H_o^2 [a + 2b]. \tag{13.103}$$

Substituting Equations (13.102) and (13.103) into Equation (13.101), and also using Equations (13.65), (13.66), we get

$$\alpha_{c_{10}} = R_s \frac{[1 + (2b/a)(f_c/f)^2]}{\eta b \sqrt{1 - \left(\frac{f_c}{f}\right)^2}},$$

where η is the intrinsic impedance of the dielectric. To see the complete frequency dependence of this expression, we can substitute Equation (13.100) into it and obtain

$$\alpha_{c_{10}} = \frac{1}{\eta b}\left[\frac{\pi f \mu}{\sigma\left[1 - \left(\frac{f_c}{f}\right)^2\right]}\right]^{1/2}\left[1 + \frac{2b}{a}\left(\frac{f_c}{f}\right)^2\right]. \tag{13.104}$$

This can also be written in the form

$$\alpha_{c_{10}} = \frac{\lambda}{b\lambda_g}\left(\frac{\pi}{\lambda\eta\sigma}\right)^{1/2}\left[1 + \left(\frac{\lambda_g}{\lambda_c}\right)^2\left(1 + 2\frac{b}{a}\right)\right], \tag{13.105}$$

where λ, λ_g, and λ_c are the TEM, guide, and cutoff wavelengths, respectively. As can be seen from Equations (13.104) and (13.105), the attenuation $\alpha_{c_{10}}$ decreases as the height b of the waveguide decreases. This, of course, is attractive, but is offset by the fact that the dominant range decreases when a is less than $b\sqrt{3}$. Because of this trade-off, the typical compromise is to choose the width-to-height ratio as $a/b \approx 2$.

Figure 13-18 shows a plot of $\alpha_{c_{10}}$ as a function of frequency for a standard, air-filled WR-75 copper waveguide with dimensions a = 1.905 [cm] and b = 0.953 [cm], and a cutoff frequency of 7.874 [GHz]. In this plot, the shaded region is the dominant range. As can be seen, the lowest attenuation occurs near the high end of that range.

Finally, metal losses are usually much greater than dielectric losses, particularly when the dielectric is air. Thus, in these cases, only the metal losses need to be considered. When both types of loss are present, however, the total loss constant α can usually be approximated as the sum of the conductor constant α_c (Equation (13.101)) and the dielectric loss constant α_d (Equation (13.94)):

$$\alpha = \alpha_d + \alpha_c\,. \tag{13.106}$$

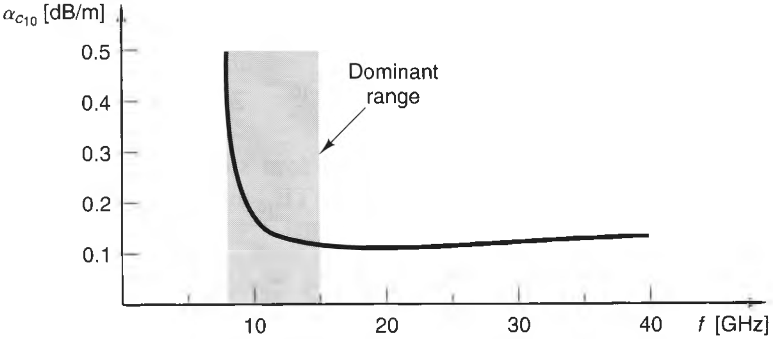


Figure 13-18 Attenuation coefficient vs. frequency for the dominant mode in a rectangular copper waveguide with dimensions a = 1.905 [cm] and b = 0.953 [cm].

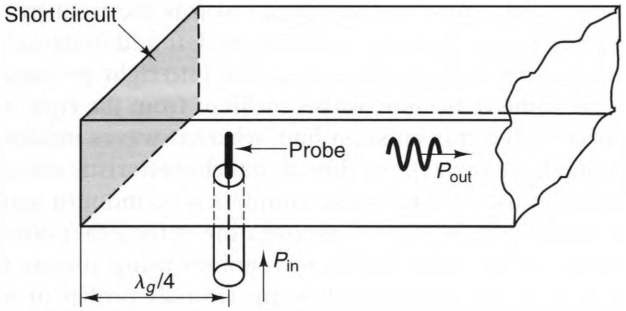


Figure 13-19 A coax-to-waveguide coupler that launches the TE_{10} mode.

13-3-7 WAVEGUIDE COUPLERS

An efficient way to generate waves in waveguides is to use couplers that convert input power from a transmission line into waveguide modes, and vice versa. Figure 13-19 shows a TE_{10}-mode coupler, which consists of a coaxial cable that enters the broad wall of a waveguide. The portion of the center conductor that extends into the waveguide is called the probe.

An exact analysis of this coupler is quite involved, but the general idea is that the probe acts as a small monopole antenna[2] and launches outward-propagating waves that are polarized parallel to the probe and are strongest near the probe. These waves are a close match to the fields of the TE_{10} mode, so they are accepted by the waveguide as forward- and backward-propagating TE_{10} modes. The short-circuit plate, placed $\lambda_g/4$ in back of the probe, allows the backward-propagating mode to reflect and add constructively with the forward-propagating mode. The net effect is that power incident from the transmission line is launched as a TE_{10} mode in the waveguide. If the probe is properly designed, very little power is reflected back into the transmission line from the coupler. Conversely, if a TE_{10} wave is incident from the waveguide upon the coupler, the process reverses itself, and power is delivered to the transmission line.

Figure 13-20 shows a transition that launches a right-propagating TE_{10} wave in a waveguide that is open at both ends. Here, two probes are fed such that their currents are 90° out of phase and are placed $\lambda_g/4$ apart inside the waveguide. The propagation

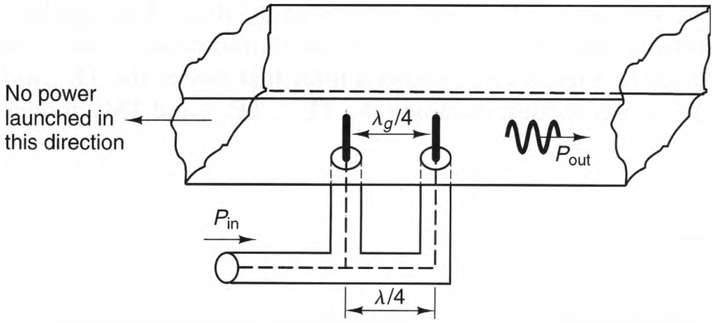


Figure 13-20 A coax-to-waveguide coupler that launches the TE_{10} mode in one direction only.

[2] We will discuss monopole antennas in Chapter 14.

delay between the probes is 90°, so waves launched towards the right by each probe add constructively, whereas waves launched towards the left add destructively. As a result, this transition couples waves from the coaxial line into right-propagating waves in the waveguide. By the same reasoning, waves incident from the right towards the coupler will deliver power to the transmission line, whereas waves incident from the left will not. Couplers that have this type of directional characteristic are called ***directional couplers***. A common use of directional couplers is to monitor signals propagating in one direction, while ignoring signals propagating in the other direction.

Higher order modes can be launched in waveguides using similar techniques. Here again, the idea is to place the radiating elements (usually probes or small loops) at positions inside the waveguide where the fields of the desired modes are strong.

13-3-8 MODE FILTERS

Metal waveguides are almost always operated so that only a single propagating mode is present. This is because the presence of more than one propagating mode causes dispersion, since different modes propagate with different group velocities. In addition, each mode behaves differently at bends, twists, couplers, and other discontinuities in a waveguide, resulting in frequency-dependent reflections and phase shifts.

As might be expected, the easiest way to ensure single-mode operation is to operate a waveguide in its dominant range. However, there are times when it is necessary to operate above the dominant range, a technique called ***overmoded operation***. As an example, waveguide feeds for antennas mounted on tall towers are often operated in this way to take advantage of the lower conductor losses of certain higher order modes. In order for these feeds to operate successfully, however, filters must be placed in the waveguide to ensure that only one mode is present. Even when a waveguide is operated in its dominant range, the nonpropagating, higher order modes produced at discontinuities in the waveguide may also upset its operation. Here again, mode filters can be used to remove energy from these unwanted modes.

There are two popular methods of fabricating mode filters. Both make use of the different modal patterns of each mode. The first technique involves placing thin, resistive cards parallel to E-field "hot spots" of the undesired modes. The unwanted modes drive conduction currents in the cards that dissipate power. If the E-field of the desired mode is perpendicular to a card, it is unaffected if the card is very thin. Obviously, this technique will work only if field patterns of the desired and undesired modes are different enough so that the cards affect the undesired mode, while leaving the desired mode unchanged. Figure 13-21 shows a filter that passes the TE_{10} mode and absorbs several higher order modes, including the TE_{11}, TE_{21}, and TM_{11} modes.

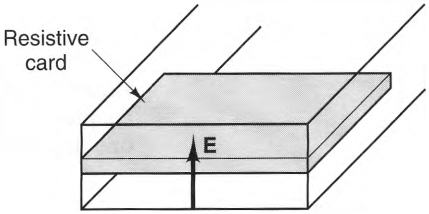


Figure 13-21 A restive-card mode filter that filters out higher order modes.

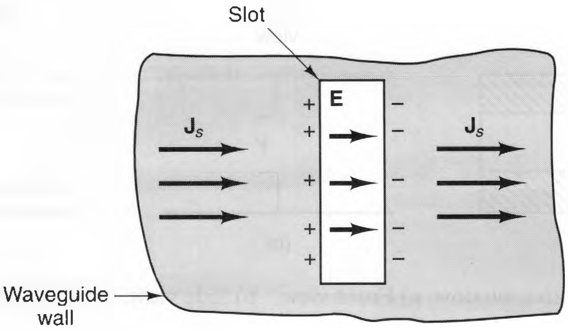


Figure 13-22 A narrow slot in a waveguide wall, showing the charge separation across the slot and the resulting E-field.

Another way to filter undesired modes is to cut narrow slots in the waveguide walls so that they disrupt the surface currents of those modes. As depicted in Figure 13-22, an electric field is excited across a slot when it breaks lines of current, because of the buildup of opposing charges on both sides of the slot. This E-field radiates power from the unwanted mode outside the waveguide, just as power is radiated from a horn antenna (which we will discuss in Chapter 14). On the other hand, the slots have little effect on currents that run parallel to the slots if they are narrow. To filter unwanted modes, the slots are placed so that only the currents of the undesired modes are affected. Figure 13-23 shows a filter slot that passes the TE_{10} and radiates the TE_{01} mode. Because these slots radiate power, they can also be used as antennas. We will discuss this application further in the next chapter.

13-3-9 LUMPED ELEMENTS IN WAVEGUIDES

Just as it is possible to place lumped elements (loads) on transmission lines to make devices such as filters and couplers, the same is true for waveguides. However, the lumped loads placed in waveguides often do not look like simple, two-terminal circuit elements, such as lumped inductors and capacitors. Rather, waveguide lumped elements often look more like obstructions, which alter the nearby electric and magnetic field distributions. Fortunately, even though these field distributions tend to be rather complex, their transmission and reflection characteristics can frequently be modeled using simple equivalent circuits.

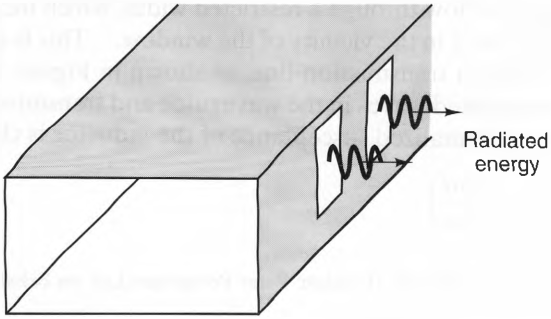


Figure 13-23 Waveguide slots that radiate power from the TE_{01} mode.

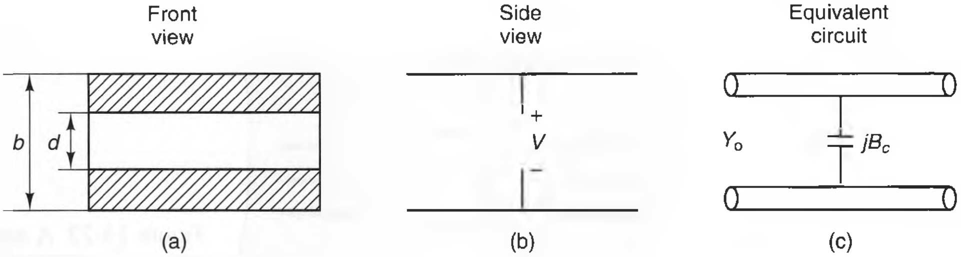


Figure 13-24 A capacitive window: a) Front view. b) Side view. c) Equivalent circuit.

Figure 13-24a shows two thin metal fins attached to the top and bottom walls of a waveguide, forming an opening called a ***capacitive window***. The side view shown in Figure 13-24b shows why the name is appropriate, since the entire voltage between the top and bottom walls of the waveguide is dropped across the narrow gap of the window. This increases the E-field within the window and increases the energy stored in the electric field.

On the other hand, the window has essentially no effect on the magnetic field, since it does not cause a redistribution of the wall currents. Hence, the net effect of the window is analogous to placing a lumped capacitor across the lines of a transmission line, as shown in Figure 13-24c. In fact, when the capacitor susceptance is chosen correctly, there is a one-to-one correspondence between the incident, reflected, and transmitted waves on the transmission line circuit and the incident, reflected, and transmitted waves in the waveguide. Using a more detailed analysis,[3] we find that the appropriate values of the actual susceptance B_c and the normalized susceptance b_c are given by

$$b_c = \frac{B_c}{Y_o} \approx \frac{4b}{\lambda_g} \ln\left(\csc\frac{\pi d}{2b}\right), \tag{13.107}$$

where λ_g is the guide wavelength and Y_o is the characteristic admittance of the transmission line (any convenient value can be used—usually unity). The magnitude and phase of the waves transmitted and reflected by the window are the same as those transmitted and reflected by the capacitively loaded transmission line at the plane of the window.

Figure 13-25a shows an ***inductive window***, so named because its equivalent circuit is a shunt inductor. In Figure 13-25b, we see that the window forces the current on the top and bottom walls to flow through a restricted width, which increases the current density and the magnetic field in the vicinity of the window. This is analogous to placing a shunt inductor across a transmission-line, as shown in Figure 13-25c. The incident, reflected, and transmitted waves in the waveguide and transmission line networks are equivalent when the normalized susceptance of the inductor is chosen as

$$b_i = \frac{B_i}{Y_o} \approx -\frac{\lambda_g}{a}\cot^2\left(\frac{\pi d}{2a}\right). \tag{13.108}$$

[3] See N. Marcuvitz, *Waveguide Handbook*, (London: Peter Peregrinus Ltd. on behalf of the Institution of Electrical Engineers, 1986).

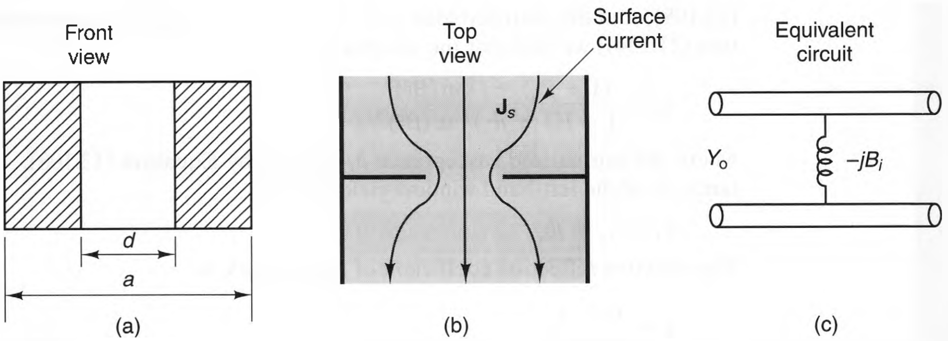


Figure 13-25 An inductive window: a) Front view. b) Top view. c) Equivalent circuit.

The following example shows how lumped elements in a waveguide can be used to alter the frequency characteristics of a waveguide system.

Example 13-8

Consider the waveguide network shown in Figure 13-26a. Here, two identical inductive shunts are placed a distance $\ell = 1.5$ [cm] apart in a waveguide with dimensions $a = 2$ [cm] and $b = 1$ [cm]. Use the transmission-line equivalent circuit for this network to calculate the reflection coefficient at the plane of the first window when $d/a = 1/2$ and $1/4$.

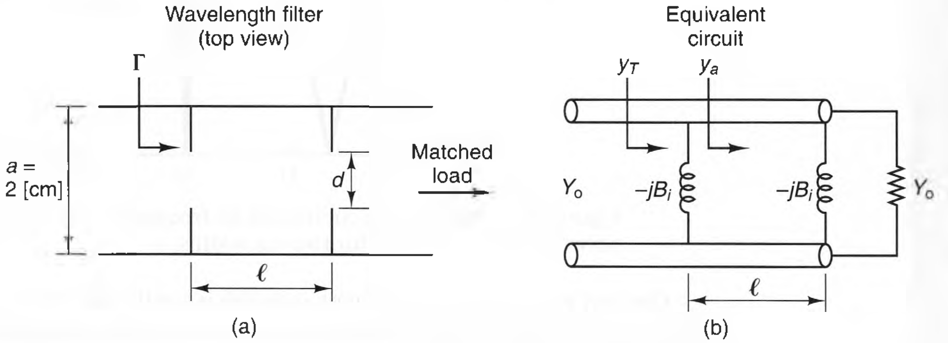


Figure 13-26 A waveguide network consisting of two inductive shunts: a) Side view. b) Equivalent circuit.

Solution:

The equivalent circuit for this network is shown in Figure 13-26b, where the resistive admittance Y_0 is placed in parallel with the right-hand inductor to model the matched load. We will approach problem by first finding the normalized admittance y_a just to the right of the left-most inductor. We will then combine this admittance with the admittance of the left-hand window to find the total admittance y_T that the shunts and the remainder of the waveguide present to the input waveguide.

The normalized admittance at the plane of the right-hand window is the parallel combination of the normalized admittances of the window ($y = jb_i$, where b_i is given by Equation

(13.108)) and the matched load ($y = 1$). Using the admittance transformation formula (Equation (11.134)) we find that the admittance

$$y_a = \frac{(1 + jb_i) + j\tan(\beta\ell)}{1 + j(1 + jb_i)\tan(\beta\ell)},$$

where the normalized susceptance b_i is given by Equation (13.108). Adding y_a to the admittance jb_i of the left-hand window yields

$$y_T = y_a + jb_i\,.$$

The effective reflection coefficient of the network is

$$\Gamma = \frac{y_T - 1}{y_T + 1}.$$

Figure 13-27 shows a plot of $|\Gamma|$ vs. frequency for two different window widths. As can be seen, $|\Gamma|$ exhibits a band-pass characteristic for both widths, but decreasing the window width d increases the resonant frequency and decreases the bandwidth.

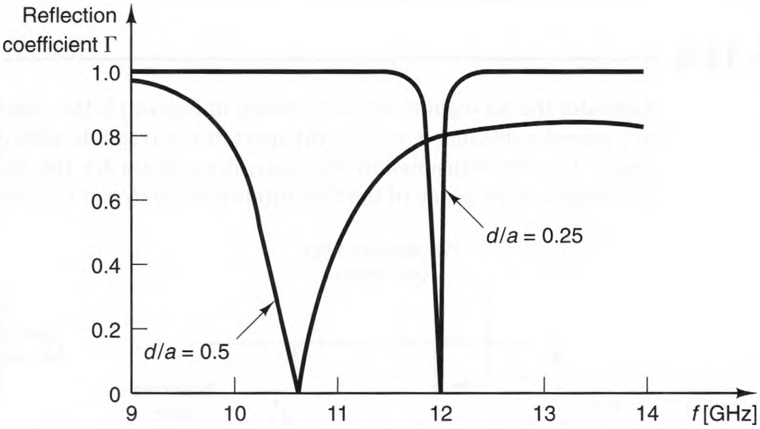


Figure 13-27 Reflection coefficient vs. frequency for the waveguide network in Figure 13-26 for two iris widths.

One way to explain why the filter bandwidth is small when the window width is narrow is to notice the similarity between this microwave filter and the etalon filter described in Chapter 12. (See Figures 12-18 and 12-19.) In both cases, waves reflect back and forth and form standing waves in a cavity, and the transmission bandwidth is smallest when the input and output faces are highly reflecting. Since the inductive windows look more and more like short-circuit plates when the window widths are small, it is reasonable to expect the filter bandwidth to be narrow for small window widths. We will have more to say about the fields in cavities later in this chapter.

13-3-10 SURVEY OF COMMON METAL WAVEGUIDES

In the preceding sections, we have seen that rectangular waveguides are capable of transporting signals and energy using any one of an infinite set of waveguide modes. Each mode has a distinct cutoff frequency, wavelength, impedance, and velocity of propagation. Since the same is true for all metal waveguides, regardless of their shape, one might wonder why different types of waveguides are used in engineering practice. The reason is that each type of waveguide has certain specific electrical or mechanical characteristics that may make it more or less suitable for a specific application. In this section we will briefly summarize and compare the notable features of the most common types: rectangular, circular, elliptical, and ridge waveguides.

Rectangular waveguides are popular because they have a large dominant range and moderate losses. 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 (depicted in Figure 13-28a) 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. Moreover circular waveguides are used for making rotary joints, which are needed when a section of waveguide must be able to rotate, such as the feed of a revolving antenna.

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. For instance, there were serious plans in the early 1970s to use 5-centimeter-diameter circular waveguides to carry long-distance telephone traffic by means of these low-loss modes. The measured performance of these waveguides was 1.5 [dB/km] from 45–110 [GHz], which is very impressive for metal waveguides. As luck would have it, however, the systems were never used commercially, since fiber-optic cables appeared on the scene about that time, with lower loss, greater bandwidth, and much lower cost.

An elliptical waveguide is shown in Figure 13-28b. As might be expected by their shape, these waveguides bear similarities to both circular and rectangular waveguides. Like circular waveguides, they are easy to bend. But unlike circular waveguides, in which the direction of polarization tends to rotate as the waves pass through bends and twists, the polarization is much more stable in elliptical waveguides. This is

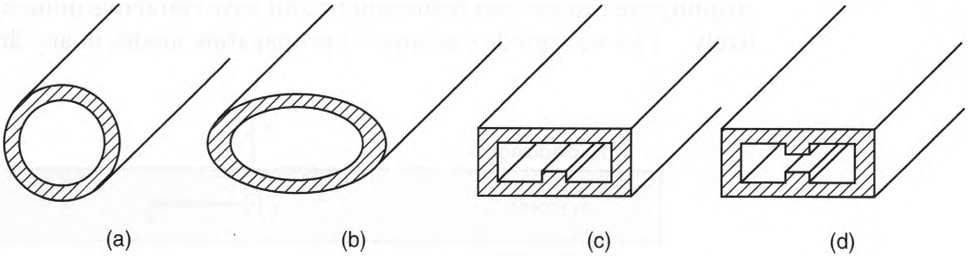


Figure 13-28 Common types of metal waveguide: a) Circular. b) Elliptical. c) Single ridge. d) Double ridge.

because the cutoff frequencies are different for modes directed along the major and minor axes of the elliptical cross section. This property makes elliptical waveguides attractive for feeding antennas, in which the polarization state is very important.

The most popular types of ridge waveguides are the single and double structures shown in Figures 13-28c and d, respectively. In these waveguides, the ridges 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, the increased-frequency bandwidth is obtained at the expense of increased attenuation and decreased power-handling capacity.

13-4 Dielectric Waveguides

Dielectric waveguides have been used for many years in microwave applications, but it is the relatively recent development of low-loss optical fibers that has made dielectric waveguides the important technology they are today. At the present time, optical fibers and cables offer the lowest losses of any guided-wave medium of electromagnetic energy.

Whereas metal waveguides confine waves by reflecting them off metal surfaces, dielectric waveguides utilize the total reflection of waves that occurs at the interface between two dielectrics when the incident angle is greater than the critical angle. Because more than one dielectric is involved, dielectric waveguides are typically harder to analyze than metal waveguides. This is particularly true of optical fibers, which are circular and often have complicated cross sections. Nevertheless, we can understand many of the properties of all dielectric waveguides by studying the characteristics of a relatively simple example—the dielectric slab waveguide.

13-4-1 THE DIELECTRIC SLAB WAVEGUIDE

Figure 13-29 shows a dielectric slab waveguide, which consists of a uniform, dielectric slab, bounded on the top and bottom by an infinite, uniform dielectric (usually free space). Using the nomenclature of optical fibers, we will call the slab the ***core*** and the surrounding medium the ***cladding***. We will assume that both the core and cladding are lossless and nonmagnetic and have refractive indices of n_1 and n_2, respectively. This waveguide can support propagating modes in any direction that is paral-

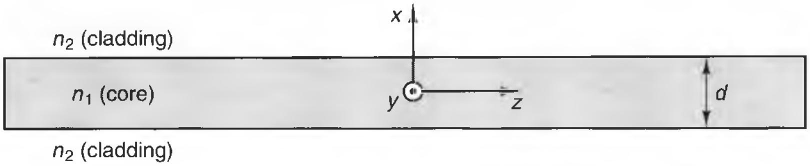


Figure 13-29 Geometry for determining the modes of a slab waveguide.

lel to the yz-plane, but we will restrict our analysis to the modes that propagate along the z-axis.

One way to find the modes of this waveguide is to solve the wave equation in the core and cladding regions and then require the necessary boundary conditions at the top and bottom interfaces. This method is straightforward, but tedious. An easier way to obtain the same result is to postulate that the modes are simply plane waves that reflect back and forth inside the core and then prove that stable modes are obtained when these waves propagate only at certain angles with respect to the waveguide axis. This technique has the added advantage that it gives special insights into the behavior of waveguide modes.

Referring to Figure 13-30, let us start by postulating that some source has launched a y-polarized plane wave inside the slab that propagates upward and to the right, at an angle θ with respect to the upper surface. If θ is greater than the critical angle θ_c, a downward-propagating wave of the same magnitude is produced as the incident wave strikes the top face. This reflected wave, in turn, produces an upward-propagating wave due to the reflection at the bottom face. The process repeats itself an infinite number of times, producing an infinite series of upward- and downward-propagating waves.

For most incident angles θ, the total electric and magnetic fields inside the slab will sum to zero. This occurs because each wave has a different phase, and hence, they add destructively. But there are certain incident angles at which the waves add constructively, producing modes that propagate without attenuation down the slab. To find these allowed values of θ, we need only require that all the upward-propagating waves have identical phase fronts. This is the same as requiring that the first upward-propagating wave turn back into itself after one complete reflection cycle. We can accomplish that with the help of Figure 13-30, which shows rays (dark, solid lines) that represent a plane wave as it progresses through one reflection cycle. The dotted line represents a constant-phase plane of the incident wave just before it strikes the top face at the point P_1.

In order for the twice-reflected wave to have the same constant-phase planes as the incident wave, its phase at the point P_2 must be the same as the phase of the incident wave at P_1 or must differ by a multiple of 2π. Noting that the phase difference

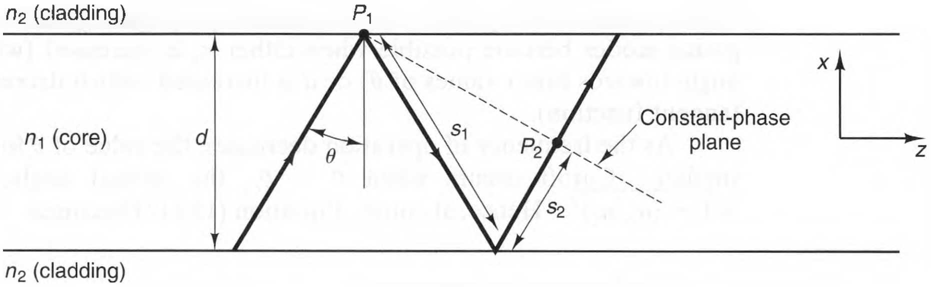


Figure 13-30 Ray diagram showing a complete reflection cycle as a plane wave reflects off of both interfaces of a dielectric slab waveguide. Permitted angles of reflection yield phases at P_1 and P_2 that differ by a multiple of 2π.

between these two points equals the sum of the propagation delay along the distance $s_1 + s_2$ and the phase shift 2ϕ caused by the two reflections, we require that

$$-k_1(s_1 + s_2) + 2\phi = -2\pi m \qquad m = 0, 1, 2, \ldots \tag{13.109}$$

In this expression, k_1 is the TEM phase constant in the slab. For TE modes, **E** is directed out of the page, so ϕ is simply the reflection phase shift for perpendicular polarization, $\phi_\perp$, which is given by Equation (12.204):

$$\phi_\perp = 2\tan^{-1}\left[\frac{\sqrt{\sin^2\theta - (n_2/n_1)^2}}{\cos\theta}\right]. \tag{13.110}$$

Also, using simple trigonometry, we can show that

$$s_1 + s_2 = 2d\cos\theta. \tag{13.111}$$

Substituting Equations (13.110) and (13.111) into Equation (13.109) and using $k_1 = 2\pi n_1/\lambda_o$ (where $\lambda_o = c/\omega$ is the free-space wavelength), we find that

$$4\tan^{-1}\left[\frac{\sqrt{\sin^2\theta - (n_2/n_1)^2}}{\cos\theta}\right] - \frac{4\pi n_1 d}{\lambda_o}\cos\theta = -2\pi m \qquad \text{(TE modes)},$$

which can be rewritten as

$$\tan\left[\frac{\pi n_1 d}{\lambda_o}\cos\theta - m\frac{\pi}{2}\right] = \frac{\sqrt{\sin^2\theta - (n_2/n_1)^2}}{\cos\theta} \text{(TE Modes; } m = 0, 1, 2, \ldots\text{).} \tag{13.112}$$

Equation (13.112) is an exact expression for the allowed propagation angles for each TE mode in a dielectric slab waveguide. This expression cannot be solved explicitly for θ, but a graphical solution can be obtained by plotting both sides of the expression separately and noting the intersection points of the curves that are produced. Such a graphical solution is shown in Figure 13-31 for the case where $n_1 = 2.3$, $n_2 = 1.0$, and $d = 1.4\lambda_o$. There are three points of intersection among these curves, which correspond to three possible propagating modes. Additional propagating modes become possible when either n_1 is increased (which shifts the critical angle towards lower values of θ) or d is increased (which decreases the period of the tangent function).

As the frequency of operation decreases, the value of θ for each mode becomes smaller. Cutoff occurs when $\theta = \theta_c$, the critical angle, at which $\cos\theta_c = \sqrt{1 - (n_2/n_1)^2}$. Hence, at cutoff, Equation (13.112) becomes

$$\frac{\pi n_1 d}{\lambda_c}\sqrt{1 - (n_2/n_1)^2} = m\frac{\pi}{2},$$

where λ_c is the cutoff wavelength in free space. Substituting $f_c = c/\lambda_c$, we find that

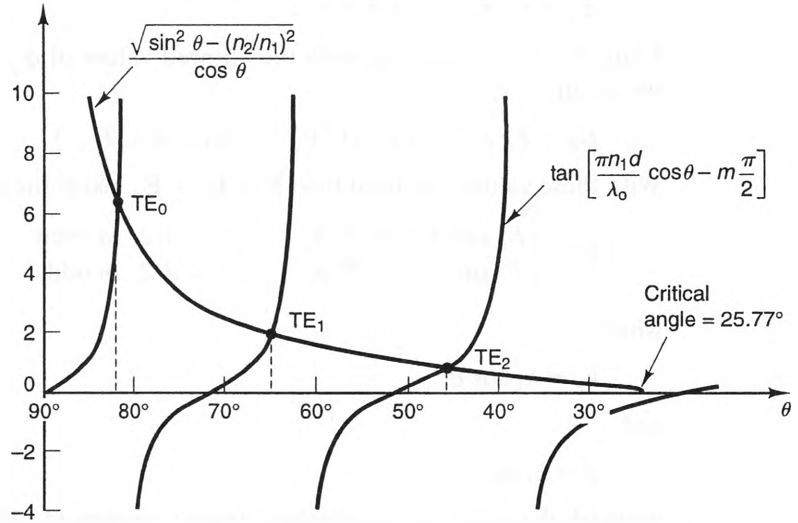


Figure 13-31 Graphical solution for the lowest order modes of a slab waveguide, obtained by graphing the left- and right-hand sides of Equation (13.112) for $n_1 = 2.3$, $n_2 = 1.0$, and $d = 1.4\lambda_o$.

$$f_{c_m} = m\frac{c}{2d\sqrt{n_1^2 - n_2^2}}. \qquad \text{(TE and TM modes; } m = 0, 1, 2, \ldots\text{)}. \qquad (13.113)$$

It can be shown that this formula also applies to the TM modes.

Unlike metal waveguides, in which all modes have nonzero cutoff frequencies, the dominant modes in a dielectric slab waveguide, TE$_o$ and TM$_o$, propagate at all frequencies. Dominant-mode operation occurs when there is only one point of intersection in Equation (13.112), which happens when

$$\frac{d}{\lambda_o} < \frac{1}{2}[(n_1)^2 - (n_2)^2]^{-1/2} \qquad \text{(Single-mode operation)}. \qquad (13.114)$$

Thus, single-mode operation at any frequency can be attained by making the slab thin, decreasing the refractive-index difference between the core and cladding, or both. Conversely, the number of propagating modes is large when the slab is thick or the difference between the indices of refraction is large.

The electric fields in the core can be found by summing the upward- and downward-propagating plane waves, $\mathbf{E}_u$ and $\mathbf{E}_d$, respectively, that make up each mode. Referring to Figure 13-30, we can represent $\mathbf{E}_u$ and $\mathbf{E}_d$ by the expressions

$$\mathbf{E}_u = E_u\hat{\mathbf{a}}_y e^{-j\beta_1(x\cos\theta + z\sin\theta)}$$

$$\mathbf{E}_d = E_d\hat{\mathbf{a}}_y e^{-j\beta_1(-x\cos\theta + z\sin\theta)}.$$

At the upper face, $\mathbf{E}_u$ is the incident wave and $\mathbf{E}_d$ is the reflected wave. Hence,

$$\mathbf{E}_d = \Gamma_\perp \mathbf{E}_u \qquad \text{at } x = d/2.$$

Using $\Gamma_\perp = 1\angle\phi_\perp$, along with the allowed values of $\phi_\perp$ given by Equation (13.110), we obtain

$$E_d = E_u\, e^{-jm\pi} = (-1)^m E_u \qquad (m = 0, 1, 2, \ldots).$$

With these values, the total field $\mathbf{E} = \mathbf{E}_u + \mathbf{E}_d$ inside the slab is

$$\mathbf{E} = \begin{cases} E_o \cos(k_x x) e^{-j\beta z}\,\hat{\mathbf{a}}_y & |x| < d/2,\ m \text{ even} \\ E_o \sin(k_x x) e^{-j\beta z}\,\hat{\mathbf{a}}_y & |x| < d/2,\ m \text{ odd} \end{cases}, \tag{13.115}$$

where

$$k_x = k_1 \cos\theta, \tag{13.116}$$

and

$$\beta = k_1 \sin\theta, \tag{13.117}$$

in which the value of θ is obtained from Equation (13.112). Remembering that the guide wavelength λ_g equals $2\pi/b$, we also have

$$\lambda_g = \frac{2\pi}{k_1 \sin\theta} = \frac{\lambda_o}{n_1 \sin\theta}, \tag{13.118}$$

where λ_o is the free-space wavelength.

The fields in a slab waveguide are not confined to just the core region. The reason for this is that evanescent waves are generated in the cladding as a result of the total reflection in the core. Remembering that the tangential E-field at a dielectric interface is continuous, we can find the amplitude of these evanescent fields by evaluating the fields in the core at the core–cladding interfaces and adding the exponential decay factor $e^{-\alpha_2|x|}$, yielding

$$\mathbf{E} = \begin{cases} E_o \cos(k_x d/2) e^{-\alpha_2(|x|-d/2)} e^{-j\beta z}\,\hat{\mathbf{a}}_y & |x| > d/2,\ m \text{ even} \\ \pm E_o \sin(k_x d/2) e^{-\alpha_2(|x|-d/2)} e^{-j\beta z}\,\hat{\mathbf{a}}_y & |x| > d/2,\ m \text{ odd} \end{cases}, \tag{13.119}$$

where the upper sign is used for $x > d/2$ and the lower sign is used for $x < -d/2$. The cladding attenuation constant is found by using Equation (12.208) and is

$$\alpha_2 = k_2\sqrt{\frac{n_1^2}{n_2^2}\sin^2\theta - 1} = k_1\sqrt{\sin^2\theta - (n_2/n_1)^2}, \tag{13.120}$$

where θ is the appropriate propagation angle for the mode in question.

Figure 13-32 shows how the field strength of the TE$_o$ mode varies throughout the core and cladding regions at two frequencies, one near cutoff and the other far above cutoff. As can be seen, the fields are more confined to the core at frequencies far above cutoff. This is an important point to consider, since real claddings have finite thicknesses, and fields can radiate from the edge of the core. In general, the lowest radiation losses are obtained when the slab is operated far above cutoff.

The TM modes of a slab waveguide are similar to the TE modes and can be derived by a similar procedure. Essentially, the only change necessary is that the

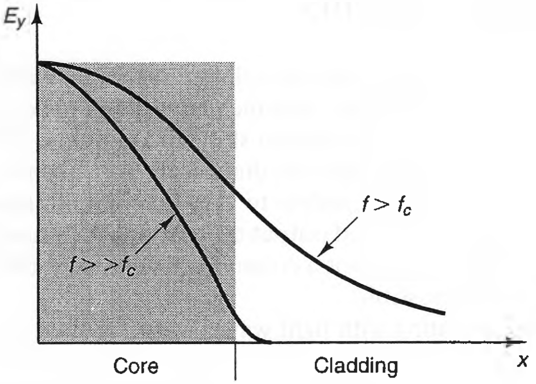


Figure 13-32 TE_0 field-strength in the core and the cladding for high and low frequencies, showing that the fields are most tightly bound to the core at higher frequencies.

upward- and downward-propagating waves in the core are polarized parallel (rather than perpendicular) to the xz-plane. This means that the reflection phase shifts used in Equation (13.109) become $\phi_{\parallel}$, rather than $\phi_{\perp}$, which results in slightly different values of the propagation angles for TM modes. Even so, the TM_m modes have the same cutoff frequencies as the TE_m modes.

Example 13-9

For a slab waveguide with $\epsilon_1 = 1.96\epsilon_0$, $\epsilon_2 = \epsilon_0$, and $d = 1$ [cm], find the dominant range and the operating characteristics at $f = f_{c_1}/2$.

Solution:

The refractive indices of the core and cladding are $n_1 = \sqrt{1.96} = 1.4$ and $n_2 = 1.0$, respectively. Substituting these values into Equation (13.113), we find that

$$f_{c_1} = \frac{c}{2d\sqrt{n_1^2 - n_2^2}} = \frac{3 \times 10^8}{(2)(1.0 \times 10^{-2})\sqrt{1.4 - 1.0}} = 23.7 \qquad [\text{GHz}].$$

When $f = f_{c_1}/2 = 11.85$ [GHz], $\lambda_0 = c/(11.9 \times 10^9) = 2.52$ [cm]. Hence, from Equation (13.112), the propagation angle for the TE_0 mode satisfies the expression

$$\tan[1.47 \cos\theta] = \frac{\sqrt{\sin^2\theta - 0.714}}{\cos\theta}.$$

Solving this expression numerically yields

$$\theta = 64.47^\circ.$$

Substituting this value into Equation (13.117), we have

$$\beta = k_1 \sin\theta = \frac{2\pi f n_1}{c}\sin\theta = \frac{2\pi \times 11.9 \times 10^9 \times \sqrt{1.4}}{3 \times 10^8}\sin(64.47^\circ)$$
$$= 2.66\ [\text{cm}^{-1}].$$

Finally, using Equation (13.118), we obtain

$$\lambda_g = \frac{2\pi}{266.1} = 2.36\ [\text{cm}].$$

13-4-2 FIBER-OPTIC WAVEGUIDES

The idea of communicating using optical wavelengths (or light waves) has been around for a long time. Various "low-tech" schemes have been used throughout history, including smoke signals, lanterns, and ship-to-ship blinker lights. The first modern system was devised in 1880 by Alexander Graham Bell, the inventor of the telephone. This system used a voice-modulated mirror to vary the light intensity directed towards a selenium photocell. The modulated output current was then used to drive a receiver headset. Although the device was not a commercial success, it did show that communication via light waves was possible.

The idea of communicating with light waves was to remain merely a curiosity for nearly a century. There were two reasons for this. The first was that the optical sources available during that time (such as incandescent, fluorescent, and arc lamps) could be modulated only at low speeds. This made their data rates smaller than those of competing technologies. Second, the optical attenuation between the transmitter and receiver sights was high, either because of spreading losses (as in open-air systems) or because of high material losses (as in dielectric waveguides).

Renewed interest in light-wave technology was aroused in the 1960s and 1970s by two inventions. The first was the laser, which provided a nearly single-wavelength optical source that could be modulated at high data rates. During the 1960s, many optical communication systems were developed that utilized unguided-laser propagation in the atmosphere. These systems represented a significant increase in the performance of light-wave communication, but there were many problems associated with the unguided propagation of the light waves. Among them were the need for line-of-sight links and a clear, unturbulent atmosphere.

The second major breakthrough was the development of low-loss silica glasses that could be formed into optical fibers. Throughout the 1970s, the loss per kilometer of optical fibers dropped dramatically, the result of new fabrication techniques that could eliminate most of the impurities that cause loss. Figure 13-33[4]shows the current state of the art in fiber attenuation as a function of the free-space wavelength λ_o. Also shown are the three operating windows that are used by most fiber-optic systems.

The usable bandwidth of a typical optical fiber is approximately 25,000 [GHz], which is roughly a thousand times greater than the entire RF spectrum used for radio communications in free space. This enormous bandwidth is responsible for the significant changes that are occurring in communication and computing systems. In the past, the communication channel between two electronic systems was often the low-speed bottleneck that limited the overall system performance. But with optical fibers, the electronic systems themselves are often the low-speed bottleneck. This means that it is becoming increasingly desirable to use optical technology to perform tasks that in the past were performed electrically, such as switching and computing.

[4] Adapted from G. Keiser, *Optical Fiber Communications*, 2d ed., (New York: McGraw-Hill, 1991), p. 11.

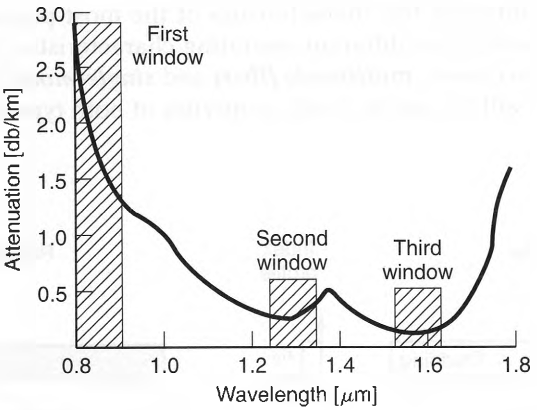


Figure 13-33 Fiber attenuation vs. frequency for state-of the art optical fibers. Also shown are the three most common frequency windows used in optical communication systems.

Figure 13-34 shows the basic components of a typical fiber-optic communication system. In this figure, the three basic system components are the optical source, the optical channel, and the receiver. The optical source is usually a laser or an LED whose output is modulated by the electrical signal that contains the message. Common modulation techniques are amplitude shift keying (ASK), phase shift keying (PSK), and frequency shift keying (FSK). In simple systems, the optical channel consists solely of an optical fiber. In more complicated systems, the optical channel also includes connectors, couplers, optical amplifiers, and repeaters, which both amplify and reconstruct digital signals. Finally, the receiver consists of a photodetector (usually a PIN diode) that converts the optical signal into an electrical signal, followed by amplifiers and demodulators.

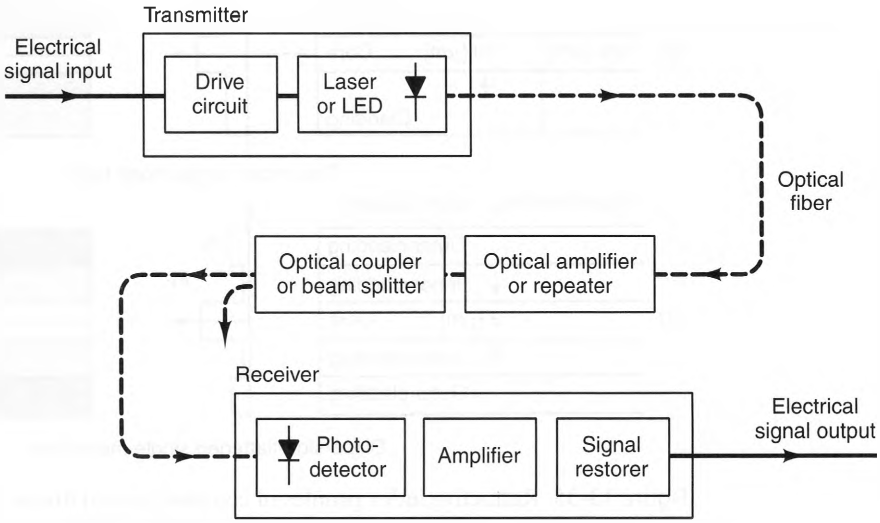


Figure 13-34 Block diagram showing the basic components in an optical communication system.

Figure 13-35 summarizes the characteristics of the most popular optical fibers. Although each construction has different operating characteristics, optical fibers are usually divided into two classes: ***multimode fibers*** and ***single-mode fibers***. In the following paragraphs, we will discuss the basic properties of both types of fiber.

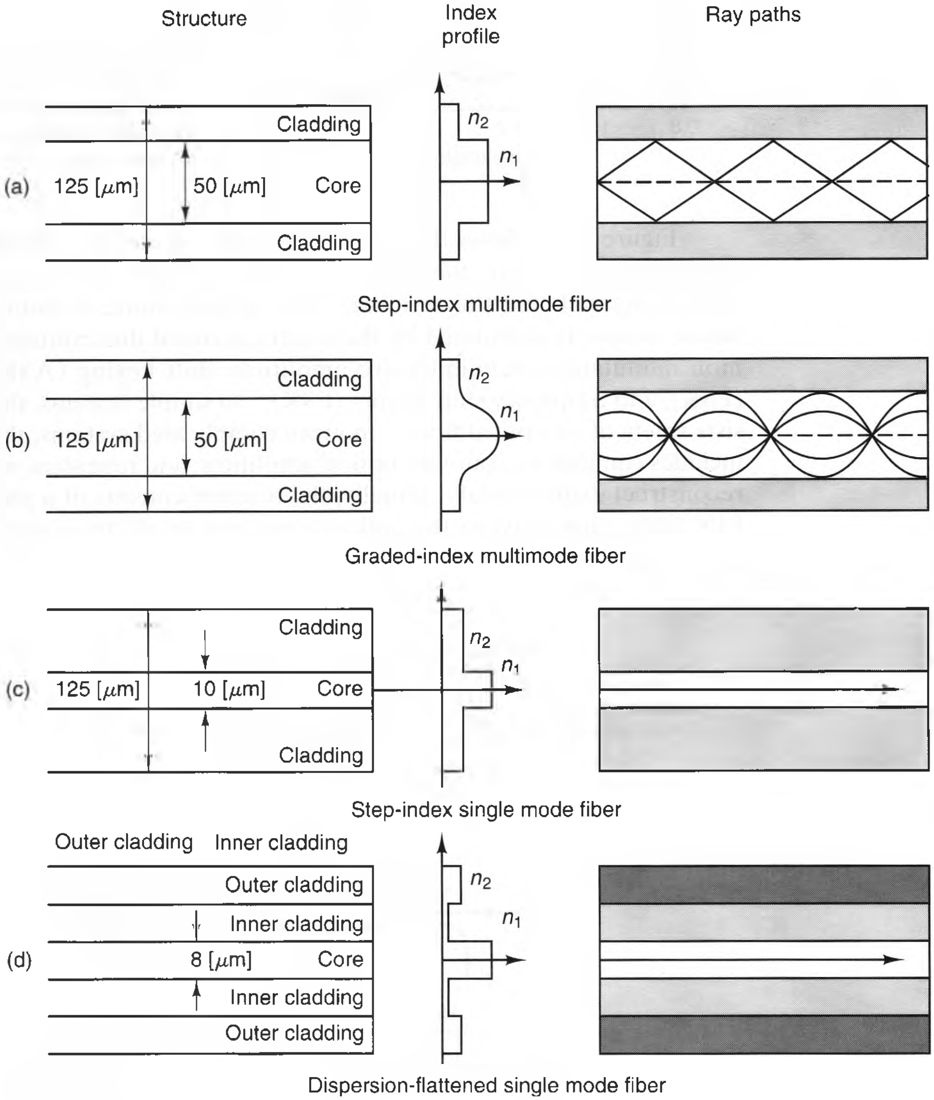


Figure 13-35 Refractive index profiles of common optical fibers: a) Step-index, multimode fiber. b) Graded-index, multimode fiber (GRIN). c) Step-index, single-mode fiber.

Multimode Fibers. There are two principal types of multimode fibers in common use: ***step-index*** and ***graded-index fibers***. The index profiles of these fibers are shown in Figures 13-35a and b, respectively. In step-index fibers, the core has a constant refractive index, yielding modes that are very similar to those of the slab waveguide. When the core diameter is large with respect to the free-space wavelength, the number of modes supported by these fibers is large, often in excess of many thousands of modes.

A major advantage of step-index, multimode fibers is that they accept a relatively large percentage of the light produced by LEDs and lasers. This occurs both because the core diameter is large and because these fibers are able to direct the incident light into a large number of modes. The high-power capability of these fibers makes them ideal for use in short-distance, local area networks. In these networks, a signal emitted by one user is shared by many users simultaneously, which means that the number of users is limited by how much optical power can be injected into the fiber. Step-index, multimode fibers permit large numbers of users, while allowing the use of relatively low-power optical sources.

The major disadvantage of step-index, multimode fibers is that they have high dispersion. This is because the higher order modes reflect off the core–cladding interface more often than lower order modes do, so their path lengths are larger. As a result, the propagation delays of higher order modes are larger than those of lower order modes. The high dispersion of these fibers limits their distance–bandwidth product. Although this makes them unsuitable for long-distance networks, the distances typically encountered in local area networks are usually small enough so that dispersion is not a problem.

Graded-index fibers (GRINs) achieve lower dispersions by using a tapered refractive index in the core, with the highest index at the center of the core. Because there is no abrupt change in the refractive index at the core–cladding interface, the waves that would otherwise to escape the core are gradually bent back towards the core. Graded-index fibers have less dispersion than step-index fibers with comparable core radii. This is because higher order modes spend a large percentage of time in the outer region of the core, where the velocity of propagation is faster. As a result, the longer path lengths of these modes are, in part, compensated by faster propagation speeds. This velocity compensation in graded-index fibers allows for higher signaling rates than can be achieved with step-index fibers.

A figure of merit that is often used to describe multimode fibers is the ***numerical aperture***, which indicates the light-gathering ability of the fiber. Figure 13-36 shows a plane wave approaching the end of a step-index fiber at an angle θ_o with respect to the optical axis.

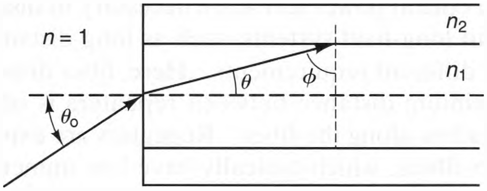


Figure 13-36 Geometry for calculating the numerical aperture NA of a step-index, optical fiber.

As this ray enters the core, it makes an angle θ with the optical axis, and the angle of incidence at the core–cladding interface is ϕ. Using Snell's law of refraction, we find that

$$\sin\theta_o = n_1 \sin\theta = n_1 \sqrt{1 - \sin^2\phi}.$$

The maximum entrance angle $\theta_{o,\max}$ occurs when $\phi = \sin^{-1}(n_2/n_1)$, which is the critical angle for the core–cladding interface. The sine of $\theta_{o,\max}$ is the numerical aperture, NA. For a step-index fiber,

$$\mathrm{NA} = \sin\theta_{o,\max} = \sqrt{n_1^2 - n_2^2}. \qquad (13.121)$$

The larger the NA of a fiber, the more optical power it can capture from a source. A typical value of NA for a step-index multimode fiber is in the range from 0.19 to 0.25.

Example 13-10

Calculate the numerical apertures of the following fibers:

a) a step-index, all-plastic fiber having core and cladding refractive indices of $n_1 = 1.60$ and $n_2 = 1.48$, respectively,

b) a step-index fiber with a silica core ($n_1 = 1.45$) and a silicone resin cladding ($n_2 = 1.4$).

Solution:

For part a),

$$\mathrm{NA} = \sqrt{(1.6)^2 - (1.48)^2} = 0.608.$$

For part b),

$$\mathrm{NA} = \sqrt{(1.45)^2 - (1.4)^2} = 0.377.$$

As these numbers indicate, plastic fibers tend to have higher numerical apertures. Unfortunately, they also have higher losses.

Since the refractive index is not constant inside the core of graded-index fibers, the maximum acceptance angle of these fibers depends upon the radial position at which the ray enters the fiber. The core is most dense at its center, so the acceptance angle is less for rays entering further away from the optical axis. As a result, GRIN fibers tend to accept less light than step-index fibers with the same core diameters. Typical average NAs for GRIN fibers are in the range from 0.16 to 0.21.

Single-Mode Fibers. Multimode fibers are popular for short-haul communication networks, where high optical power levels are necessary to distribute signals to a large number of users. But long-haul systems, such as long-distance telephone and computer networks, have different requirements. Here, fiber dispersion becomes a major factor, since the maximum distance between repeaters is often determined by the amount of pulse spreading along the fiber. Repeaters are expensive, so it is desirable to use low-dispersion fibers, which typically have low numerical apertures. Fortunately, the power requirements of long-haul systems are not as great as those of short-haul systems, since the number of users tied to the ends is usually smaller. For these applications, single-mode fibers are most often used. As their name implies, these fibers support only one propagating mode and thus eliminate intermodal dispersion effects.

There are many kinds of single-mode fibers. The simplest are step-index fibers with very small core diameters. (See Figure 13-35c.) An exact analysis of the modes in these fibers shows that single-mode operation is attained when the core radius satisfies the inequality[5]

$$\frac{a}{\lambda_o} < \frac{2.405}{2\pi\sqrt{n_1^2 - n_2^2}}. \tag{13.122}$$

Example 13-11

Calculate the maximum core radius that guarantees single-mode operation for a fiber with $n_1 = 1.48$ and $n_2 = 1.46$, where the operating wavelength is $\lambda_o = 1.3$ [μm].

Solution:

Using Equation (13.122), we have

$$a < 1.3 \times 10^{-6} \times \frac{2.405}{2\pi\sqrt{(1.48)^2 - (1.46)^2}},$$

which yields

$$a < 2.05\ [\mu\text{m}].$$

When only one mode propagates in a fiber, intermodal dispersion is eliminated. Nevertheless, this still leaves two other dispersion mechanisms that can limit signaling bandwidths. The first is ***material dispersion***, which is due to the changes in the refractive index of the core with the optical wavelength. Very little can be done about this component of the dispersion, since the silica glasses used in fibers are already very pure. The second dispersion mechanism is ***waveguide dispersion***, which occurs for all non-TEM modes in any kind of waveguide.

Single-mode optical fibers can be designed so that the material and waveguide dispersion components cancel. This is possible because these two dispersion mechanisms have opposite frequency characteristics. The index of refraction of silica glass increases slightly with increasing frequency, causing larger phase delays at higher optical frequencies. On the other hand, the group velocity of the dominant waveguide mode increases at higher frequencies, thus decreasing the phase delays. The material dispersion is fixed, but the waveguide dispersion can be varied by simply changing the core diameter. Zero net dispersion at a single optical wavelength can be achieved by choosing a core diameter that is even smaller than that needed to ensure single-mode operation. These fibers are called ***dispersion-shifted fibers***, since they exhibit zero total dispersion at a higher wavelength than the wavelength at which the material dispersion alone is zero.

[5] See G. Keiser, *Optical Fiber Communications*, (New York: McGraw Hill, 1991).

It is also possible to attain low dispersion over a band of frequencies by using more complicated index profiles within the cladding, such as the profile shown in Figure 13-35d. Fibers of this type are called ***dispersion-flattened*** fibers. They are attractive when a number of signals with slightly different optical wavelengths are multiplexed to propagate along the same fiber.

13-5 Cavity Resonators

At low frequencies, circuits with lumped inductors and capacitors often exhibit one or more frequencies at which the input impedance is purely real. These frequencies are called resonant frequencies, or simply resonances, and the circuits are often called resonators. At or near the resonant frequencies, the network appears resistive. Away from the resonances, however, the circuits appear highly reactive. These characteristics make resonators very useful for tuning and filtering in electronic systems.

Theoretically, it would seem possible to construct lumped resonators at any frequency. However, lumped elements become increasingly difficult to fabricate at microwave frequencies and above. For instance, it is difficult to imagine how one would construct a capacitor for use at optical frequencies. At high frequencies, a more attractive way to make a resonator is to allow waves to reflect back and forth within some sort of enclosure. These enclosures, called cavities, can be a dielectric (possibly air) surrounded by metal walls, or simply a block of dielectric. The sharp resonances are a result of the constructive and destructive interference that the waves exhibit as they reflect back and forth within the resonator. Cavity resonators are very important in a wide range of electrical and optical applications, including oscillator circuits, filters, tuned amplifiers, and laser cavities.

Cavity resonators are similar to waveguides in that they both support a large number of distinct modes. But whereas each waveguide mode can exist over a broad range of frequencies, resonator modes are usually restricted to very narrow frequency ranges. To understand why, consider the rectangular cavity shown in Figure 13-37, which has perfectly conducting walls of width a, height b, and length d along the x-, y-, and z-axes, respectively, and is filled with a lossless dielectric. The easiest way to analyze the modes of this resonator is to recognize that the structure is nothing more than a rectangular waveguide that is closed at both ends. Given this, it follows that wave-

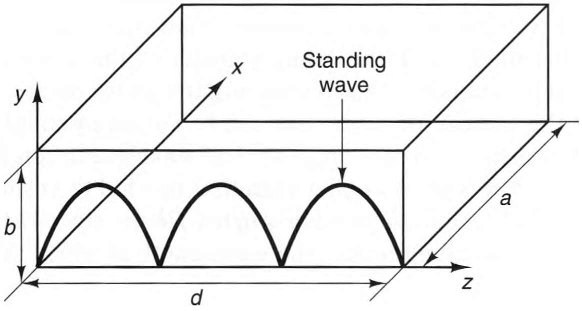


Figure 13-37 A rectangular resinator (cavity), showing E vs. z for one resonant mode.

guide modes propagating along the z-axis will experience constructive interference when the length d is a multiple of $\lambda_g/2$, where λ_g is the guide wavelength. For these lengths, standing waves are formed in the cavity, one of which is depicted in Figure 13-37. Since $\lambda_g = 2\pi/\beta$, the condition for resonance is

$$\beta = p\frac{\pi}{d}, \qquad p = 1, 2, \ldots. \tag{13.123}$$

From Equations (13.55) and (13.56), the propagation constant β_{mn} for the mnth TE or TM mode in a rectangular waveguide is

$$\beta_{mn} = \sqrt{k^2 - \left(\frac{m\pi}{a}\right)^2 - \left(\frac{n\pi}{b}\right)^2}, \tag{13.124}$$

where $k = 2\pi f/u$ and $u = c/\sqrt{\epsilon_r}$ are the wave number and the speed of light in the dielectric, respectively. Comparing Equations (13.123) and (13.124), we find that they are both satisfied only when k takes on discrete values, which occur at the resonant frequencies, given by

$$f_{mnp} = \frac{u}{2}\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2 + \left(\frac{p}{d}\right)^2}. \tag{13.125}$$

In this expression, the indices m, n, and p can take on all positive integer values, including zero, as long as only one index is zero at a time. Thus, there are a triply infinite number of resonant frequencies and corresponding modal field distributions throughout the cavity. If the x-, y-, and z-axes are chosen such that $a > d > b$, then the lowest resonant frequency is f_{101}.

Since the lowest resonance has $m = 1$ and $n = 0$, the E- and H-fields of this mode correspond to the fields of the TE waveguide mode. We can obtain the corresponding cavity fields by summing forward- and backward-propagating modes whose amplitudes add such that $E_y = 0$ at $z = 0$ and $z = d$. Substituting $p = \beta d/\pi = 1$ into Equations (13.62)–(13.64) and noting that $\gamma = j\beta$ changes sign for backward-propagating waves, we obtain

$$E_y = -2\omega\mu H_o\left(\frac{a}{\pi}\right)\sin\left(\frac{\pi}{a}x\right)\sin\left(\frac{\pi}{d}z\right) \tag{13.126}$$

$$H_x = j2H_o\left(\frac{a}{d}\right)\sin\left(\frac{\pi}{a}x\right)\cos\left(\frac{\pi}{d}z\right) \tag{13.127}$$

$$H_z = -j2H_o\cos\left(\frac{\pi}{a}x\right)\sin\left(\frac{\pi}{d}z\right), \tag{13.128}$$

where H_o is an arbitrary constant. We can find the energies stored in these electric and magnetic fields by substituting Equations (13.126)–(13.128) into Equations (6.34) and (9.39), respectively. The time-averaged electric energy is

$$W_e = \frac{\epsilon}{2}\int_{\text{Vol.}}\frac{1}{2}\text{Re}(\mathbf{E}\cdot\mathbf{E}^*)\,dv = \epsilon|H_o|^2\left(\frac{\omega\mu a}{\pi}\right)^2\int_0^a\int_0^b\int_0^d \sin^2\frac{\pi x}{a}\sin^2\frac{\pi z}{d}\,dx\,dy\,dz$$

$$= \frac{ab\,d\epsilon}{4}\,|H_{\text{o}}|^2 \left(\frac{2f\mu a}{\pi}\right)^2 \qquad [\text{J}]. \tag{13.129}$$

Similarly, the time-averaged magnetic energy is

$$W_m = \frac{\mu}{2}\int_{\text{Vol.}} \frac{1}{2}\,\text{Re}(\mathbf{H}\bullet\mathbf{H}^*)\,dv$$

$$= \mu\,|H_{\text{o}}|^2 \int_0^a \int_0^b \int_0^d \left(\left(\frac{a}{d}\right)^2 \sin^2\frac{\pi x}{a}\cos^2\frac{\pi z}{d} + \cos^2\frac{\pi x}{a}\sin^2\frac{\pi z}{d}\right) dx\,dy\,dz$$

$$= \frac{ab\,d\mu}{4}\,|H_{\text{o}}|^2 \left[\left(\frac{a}{d}\right)^2 + 1\right] \qquad [\text{J}]. \tag{13.130}$$

However, since $f_{101} = 1\,(2\sqrt{\mu\,\epsilon})\,(1/a^2 + 1/d^2)^{1/2}$, we find that $W_e = W_m$, which is analogous to what happens in lumped LC networks at resonance.

The resonant frequencies given by Equation (13.125) were derived under the assumption that cavity walls are lossless, with no power transferred out of the cavity through apertures or couplers. This assumption led to infinitely sharp resonances, which become broader when loss is present. Just as with lumped RLC resonant circuits, the more loss that is present, the broader the resonance will be. The bandwidth and power dissipation of a resonator are related by its quality factor, which is defined by

$$Q = 2\pi\,\frac{\text{maximum energy stored}}{\text{energy loss per cycle of oscillation}} = \frac{4\pi f_{\text{o}} W_e}{P_L}, \tag{13.131}$$

where W_e is the time-averaged energy stored in the electric field and P_L is the time-averaged dissipated power at resonance. This is the same definition for the quality factor as is used for lumped-element tuned circuits.[6] Also as in lumped circuits, the quality factor Q and the 3dB bandwidth BW of a cavity resonator are related by

$$\text{BW} = \frac{2\pi f_{\text{o}}}{Q}, \tag{13.132}$$

where f_{o} is the resonant frequency of the cavity. The derivation of this formula for cavity resonators is beyond the scope of this text,[7] but is a direct consequence of the Fourier transform principle that networks with short transient responses (i.e., low Q) have large bandwidths.

When conduction losses in the metal walls of a cavity resinator are dominant, we can use Equation (13.99) to express the dissipated power P_L in terms of the H-field on the cavity walls. Substituting, we obtain

$$Q = 4\pi f_{\text{o}}\,\frac{W_e}{R_s \oint_S H^2 ds}, \tag{13.133}$$

[6] See David Irwin, *Basic Engineering Circuit Analysis*, 4th ed., (New York: Macmillan, 1993).

[7] See R. E. Collin, *Foundations for Microwave Engineering*, 2d ed., (New York: McGraw-Hill, 1992).

where S is the surface of the conducting walls and R_s is the surface resistance of the walls and is given by Equation (13.100). For the case where $a > d > b$, it can be shown by direct substitution of the modal fields that the Q of the lowest order mode is

$$Q_{101} = \frac{(2\pi f_{101}ad)^3 b\eta}{u^3 2\pi^2 R_s(2a^3b + 2d^3b + a^3d + d^3a)}, \tag{13.134}$$

where u and η are the speed of light and the intrinsic impedance the dielectric, respectively, and $f_{101} = (u/2)\sqrt{(1/a)^2 + (1/d)^2}$.

Example 13-12

Calculate the frequency, the Q, and the bandwidth of the lowest order resonance of an air-filled, square metal cavity that is 2 [cm] on a side. Assume that the walls are made of copper.

Solution:

Using Equation (13.125), we find that the lowest resonant frequency is

$$f_{101} = \frac{u}{2}\sqrt{\left(\frac{1}{a}\right)^2 + \left(\frac{1}{d}\right)^2} = \frac{3 \times 10^8}{.02 \times \sqrt{2}} = 10.6 \quad \text{[GHz]}.$$

For copper, $\sigma = 5.8 \times 10^7$ [S/m]. From Equation (13.100), the surface resistance is

$$R_s = \sqrt{\frac{\pi f_{101}\mu_o}{\sigma}} = 0.027 \quad [\Omega].$$

Substituting these values into Equation (13.133), we obtain

$$Q_{101} = \frac{(2\pi f_{101}ad)^3 b\eta_o}{c^3 2\pi^2 R_s(2a^3b + 2d^3b + a^3d + d^3a)} = 10{,}320.$$

Finally, using Equation (13.132), we find that

$$\text{BW} = \frac{2\pi f_o}{Q} = 6.45 \quad \text{[MHz]}.$$

Since the half-power bandwidth is a very small fraction of the resonant frequency, this is a sharp resonance.

The rectangular cavity just described, while interesting, is of no practical value, as there are no input or output ports. In practice, cavities are usually attached to transmission lines or waveguides. Just as with lumped circuits, resonance occurs at those frequencies at which the input impedance Z_{in} is real valued. A series resonance occurs when the imaginary part of Z_{in} has a positive slope at resonance (just like a series RLC circuit). A parallel resonance occurs when the imaginary part of the input admittance $(1/Z_{\text{in}})$ has a positive slope at resonance (just like a parallel RLC circuit).

There are a number of ways that cavities can be coupled to waveguide and transmission-line circuits. One method is to construct the resonator inside a waveguide by placing lumped elements (such as capacitive or inductive windows) inside the waveguide. This kind of resonator was discussed in Example 13-7. Another method is

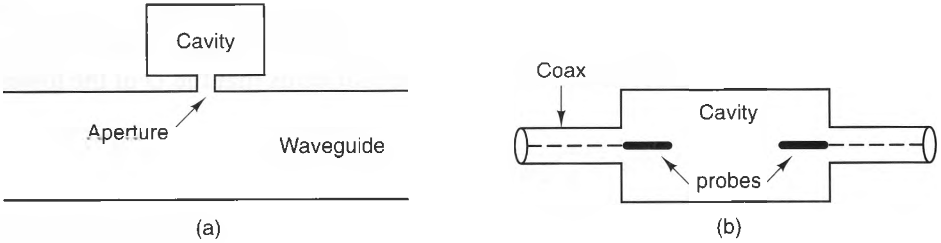


Figure 13-38 Typical cavity configurations: a) A shunt-mounted cavity in a waveguide wall. b) A series configuration with coaxial ports.

shown in Figure 13-38a. Here, a cavity is coupled to a waveguide by an aperture in the side wall of the waveguide. The aperture is positioned so that the waveguide fields excite the desired cavity mode. Figure 13-38b shows another method, in which, probes attached to the center conductors of the coaxial cables extend into the cavity and act as antennas, coupling energy in and out of the cavity.

Figures 13-39a and b show two types of microstrip resonators that use simple microstrip components as the resonating elements. In Figure 13-39a, the resonator is simply a strip of open-circuited, microstrip transmission line, which is capacitively coupled to the input transmission line. The resonant frequencies of the strip can be determined by simple transmission-line analysis. A similar effect can be obtained from the circular-disk resonator shown in Figure 13-39b. Here, the resonant frequencies of the disk are a function of its radius.

A problem often encountered when using microstrip resonators, such as the ones depicted in Figure 13-39, is that they must have dimensions on the order of a free-space wavelength at resonance. At low frequencies, this often makes them unacceptably large. A way around the problem is to use dielectric resonators, such as the one shown in Figure 13-40. These resonators are often called ***dielectric puck resonators***, because they look like miniature hockey pucks. Here, the fields are confined to a region of high dielectric constant, where the wavelength is small. This allows a small resonator to resonate at a relatively low frequency. Power can be coupled to the puck using capacitive coupling from a microstrip transmission line. Because of their small size and sharp frequency characteristics, dielectric puck resonators are often used as feedback elements in microwave oscillators.

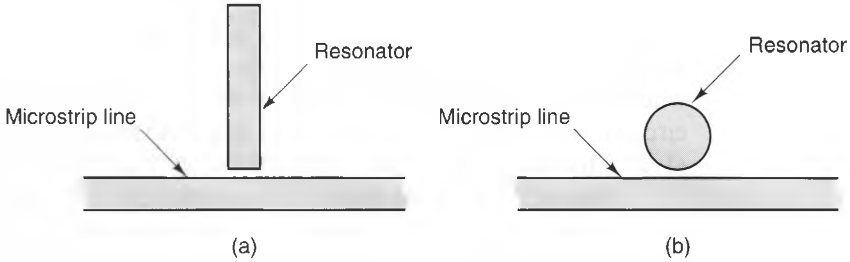


Figure 13-39 Microstrip resonators: a) Rectangular strip. b) Circular patch.

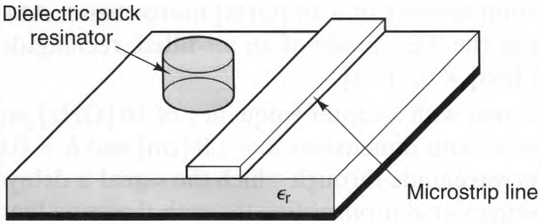


Figure 13-40 A dielectric puck resinator, coupled to a microstrip transmission line.

13-6 Summation

In this chapter, we have seen how electromagnetic waves can be transported by simple metal and dielectric structures using non-TEM modes. Unlike the TEM modes encountered with transmission lines, non-TEM modes have distinct cutoff frequencies, below which they will not propagate. Even when operated above cutoff, these waveguide modes exhibit variations with frequency that must be anticipated in order to obtain acceptable system performance. Nevertheless, waveguides offer low-loss transmission characteristics that make their less-than-desirable dispersion characteristics well worth the trouble.

The optical revolution that started with the invention of low-loss optical fibers is only in its beginning stages. In addition to the low loss, the bandwidth capabilities of a single fiber exceed those of air. Only time will tell what percentage of functions that are at present performed electrically will someday be performed optically.

Problems

13-1 Show that the component E_x, given by Equation (13.9), satisfies the wave equation (Equation (13.16)) as long as both E_z and H_z do also.

13-2 The TM modes for rectangular waveguides were found by finding the appropriate longitudinal component E_z that satisfies the boundary conditions at the conductor walls. Show that the transverse components E_x and E_y for these modes also satisfy the correct boundary conditions.

13-3 Repeat Problem 13-2 for the TE modes in rectangular waveguides.

13-4 An air-filled, rectangular waveguide has dimensions $a = 2.5$ [cm] and $b = 1.25$ [cm]. At an operating frequency of 17 [GHz], what modes are above cutoff?

13-5 An air-filled, rectangular waveguide has dimensions $a = 1.5$ [cm] and $b = 0.8$ [cm]. Calculate the ratio of the guide wavelength to the free-space wavelength for the dominant mode at $f =$ a) 10.2 [GHz], b) 15 [GHz], and c) 30 [GHz].

13-6 A section of WR-75 waveguide ($a = 1.905$ [cm] and $b = 0.953$ [cm]) is operated at 6 [GHz]. Find the attenuation of the dominant mode through this section, in dB/m.

13-7 Find the dimensions of a square waveguide that has a cutoff frequency of 12.5 [GHz].

13-8 Calculate the group velocity of a 14 [GHz] narrowband, AM-modulated signal that propagates in the TE_{10} mode of an air-filled, rectangular waveguide with dimensions 1.15 [cm] × 0.8 [cm].

13-9 A narrowband signal with a center frequency of 10 [GHz] propagates along an air-filled waveguide with dimensions $a = 1.8$ [cm] and $b = 0.6$ [cm]. Calculate the length of this waveguide through which the signal is delayed by 1 [μs], compared with the same signal propagating through the same length in air.

13-10 Calculate the ratio of the wave impedance to the free-space characteristic impedance for the dominant mode in an air-filled, rectangular waveguide with dimensions $a = 2$ [cm] and $b = 1$ [cm] at $f =$ a) 7.7 [GHz], b) 10.0 [GHz], and c) 15 [GHz].

13-11 Show that in the region near the center of a rectangular waveguide, the TE_{10} mode behaves like a plane wave. What is the polarization state of this plane wave? What is the intrinsic impedance? (*Hint:* Expand Equations (13.62)–(13.64) in a Taylor's series about $x = a/2$.)

13-12 A rectangular, air-filled waveguide has dimensions $a = 3.4$ [cm] and $b = 1.8$ [cm] and walls made of copper. Calculate the attenuation (in dB/m) for the dominant mode at $f =$ a) 6 [GHz] and b) 12 [GHz].

13-13 Figure P13-13 shows a horn antenna attached to a lossless rectangular waveguide. Measurements at 7 [GHz] show that the horn is not perfectly matched to the wave impedance of the TE_{10} mode, resulting in a VSWR of 2.5 in the waveguide and a voltage minimum 1.2 [cm] behind the neck of the horn. Using the reactive shunt (or stub) tuner method of impedance matching discussed in Chapter 11, find the location and the dimensions of the inductive iris that achieves a perfect match.

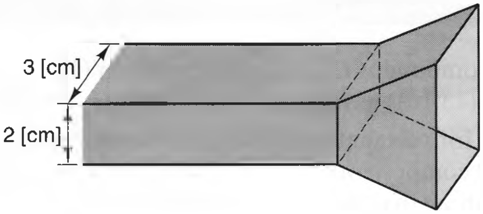


Figure P13-13

13-14 Repeat Problem 13, this time using a capacitive iris.

13-15 For a rectangular waveguide with a lossless dielectric, prove that the normalized frequency at which the TE_{10} mode exhibits its minimum attenuation is

$$\frac{f}{f_c} = \left[g + \sqrt{g^2 - (2b/a)}\right]^{1/2}, \qquad \text{where } g \equiv \frac{3}{2} + \frac{3b}{a}.$$

When $a/b = 2$, where does this frequency lie with respect to the dominant range?

13-16 Calculate the maximum power that can be sent through an $a = 2.0$ [cm] and $b = 1.0$ [cm], air-filled waveguide at 12 [GHz] such that dielectric breakdown does not occur. Assume that the dielectric strength of air at atmospheric pressure is 3×10^6 [V/m].

13-17 A sinusoidal pulse with a center frequency of 12.5 [GHz] is launched down an air-filled, rectangular waveguide with dimensions $a = 1.9$ [cm] and $b = 0.9$ [cm]. If the bandwidth of this pulse is 500 [MHz], estimate the pulse spread per meter.

13-18 For a slab waveguide where the core and cladding have refractive indices of n_1 and n_2, respectively, prove that the phase constant β of any propagating mode satisfies the inequality

$$n_2\beta_o < \beta < n_1\beta_o ,$$

where β_o is the phase constant of a TEM wave in free space at the same frequency. Also, interpret the extreme values $\beta = n_1\beta_o$ and $\beta = n_2\beta_o$.

13-19 For a dielectric slab waveguide, the field intensities decay very slowly away from the core–cladding interface when the core thickness d is small with respect to the TEM wavelength in the core. Show that when $k_1 d \ll 2$ and both the core and cladding are nonmagnetic, the following approximations are valid for the dominant TE mode:

$$\beta \approx k_1 \sin\theta_c = k_2$$

$$\alpha_2 \approx \frac{d}{2}(k_1^2 - k_2^2).$$

The constants k_1 and k_2 are the wave numbers of the core and cladding, respectively.

13-20 For a dielectric slab waveguide with $n_1 = 1.46$ and $n_2 = 1.44$, calculate the maximum slab thickness d that results in single mode propagation at the frequency at which the free-space wavelength is 0.8 [μm].

13-21 A dielectric slab waveguide with thickness 10 [μm] has $n_1 = 1.5$ and $n_2 = 1.49$. Find the number of modes that can propagate at the frequency whose free-space wavelength is 0.94 [μm].

13-22 Find the thickness d of a dielectric slab waveguide with a dielectric constant of $\epsilon_r = 4.0$ whose dominant asymmetrical mode (i.e., $m = 1$) has a cutoff frequency of 18 [GHz]. Assume that the cladding is air.

13-23 Find the numerical aperture NA of a step-index optical fiber that has $n_1 = 1.51$ and $n_2 = 1.49$.

13-24 Figure P13–24 shows a laser source illuminating the end of an optical fiber. The laser can be considered a point source that radiates power uniformly within a 40° angle of rotation about the fiber axis. If the fiber has a core radius of 50 [μm] and a numerical aperture of 0.2, calculate the percentage $P\%$ of the laser's power that is captured by the fiber if the distance from the laser to the fiber is a) 10 [μm], b) 1 [mm], c) 1 [cm].

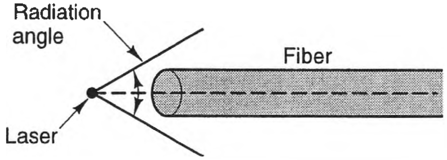


Figure P13-24

13-25 Calculate the maximum core diameter that results in single-mode operation of a step-index optical fiber with $n_1 = 1.51$ and $n_2 = 1.50$ at a free-space wavelength of 1.3 [μm].

13-26 A 40 [km] optical link uses a fiber that has a loss of 0.75 [dB/km] and has connectors spaced every 4 [km] and connectors at the ends. Each connector adds a loss of 0.25 [dB]. Determine the power that must be launched in the fiber if the power required at the detector is 1.5 [μW].

13-27 For a metal, air-filled cavity with dimensions 4 [cm] × 5 [cm] × 2 [cm], find the four lowest order resonant modes and their frequencies.

13-28 An air-filled rectangular cavity has copper walls ($\sigma = 5.8 \times 10^7$ [S/m]) and dimensions 6 [cm] × 5 [cm] × 3 [cm]. Calculate the Q of the dominant resonance and its frequency.

13-29 When measurements are conducted in a shielded room, it is best to operate the equipment at frequencies well below all of the resonant frequencies of the room. For a shielded room with dimensions 3 [m] × 2 [m] × 2.5 [m], calculate the lowest resonant frequency.

13-30 Design a cubic cavity resonator that has a dominant resonant frequency of 9 [GHz] if it is filled with a) air, b) a dielectric with $\epsilon_r = 150$.

14

Radiation and Antennas

14-1 Introduction

One of the most useful properties of electromagnetic fields is that they can transport energy and signals from one place to another. When these waves propagate without a guiding structure, this process is called ***radiation***. We have already encountered some aspects of radiation in our discussion of plane waves, but there we did not consider the sources of these waves. In this chapter, we will look closely at how these waves are actually launched by sources.

Radiation can be viewed as either a desirable or an undesirable phenomenon, depending upon the situation. Situations where electromagnetic radiation is desirable are numerous. The most notable examples are communication systems and radars, both of which use electromagnetic waves to transport information.

A common example of undesirable radiation is the interference caused by personal computers with other devices, such as television receivers or cellular phones. In these cases, the high-speed digital signals inside the computer generate electromagnetic waves that leak out of the chassis and cables. Emissions of this type are called ***electromagnetic interference*** (EMI) and are regulated by governmental agencies such as the

Federal Communications Commission (FCC). The engineering discipline that seeks to minimize these sorts of problems is called ***electromagnetic compatibility*** (EMC).

In what follows, we will introduce the subject of electromagnetic radiation by first discussing electromagnetic potentials, which greatly simplify the analysis of radiating structures. Next, we will discuss the properties of the simplest radiator: a short filament of current. This radiator is the prototype for all radiating elements, since any current distribution can be modeled as a collection of these elemental sources. After this, we will discuss a number of different types of antennas, including both simple elements and arrays of simple elements.

14-2 Electromagnetic Potentials

The fundamental source of all electromagnetic radiation is the acceleration of charges, which occurs whenever time-varying current or charge distributions are present. Because of this, it is often convenient to use formulas for the radiated electric and magnetic fields that are explicit functions of the currents and charges that cause them. In that case, Maxwell's equations are not the most convenient starting point for the analysis. A better approach is to use ***electromagnetic potentials***, which are similar to the electric and magnetic potentials derived earlier in Chapters 4 and 7. We will start our discussion of radiation by defining these potentials and showing how they can be used to determine the electric and magnetic fields generated by known current distributions.

It takes a fair number of steps to derive the electromagnetic potentials from Maxwell's equations, so we will start our discussion by stating the final result. Figure 14-1 shows arbitrary current and charge distributions, **J** and ρ_v, respectively, that lie within a volume V. For an observer at the position indicated by the vector **r**, the fields **E** and **B** can be represented in the frequency domain as

$$\mathbf{E} = -\nabla V - j\omega\mathbf{A} \tag{14.1}$$

$$\mathbf{B} = \nabla \times \mathbf{A}, \tag{14.2}$$

where the ***electric scalar potential*** V and the ***magnetic vector potential*** **A** are defined by

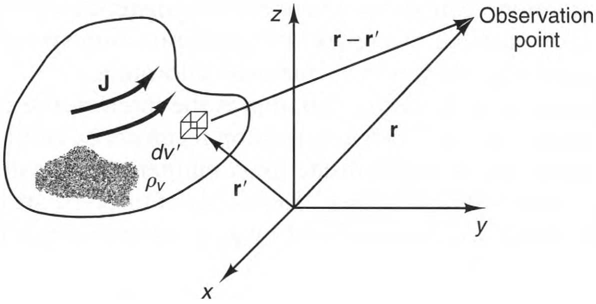


Figure 14-1 Geometry for deriving the electromagnetic potentials.

$$V(\mathbf{r}) = \frac{1}{4\pi\epsilon} \int_{\text{Vol.}} \frac{\rho_v(\mathbf{r}')e^{-jk|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}\,dv' \qquad [\text{V}] \tag{14.3}$$

$$\mathbf{A}(\mathbf{r}) = \frac{\mu}{4\pi} \int_{\text{Vol.}} \frac{\mathbf{J}(\mathbf{r}')e^{-jk|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}\,dv' \qquad [\text{Wb/m}] \text{ or } [\text{T} \cdot \text{m}]. \tag{14.4}$$

In these expressions, $\mathbf{r}'$ is a dummy position vector that sweeps through all the points inside the volume, and k is the wave number of the medium. (See Equation (12.6).)

Equations (14.1) and (14.2) state that the electric and magnetic fields **E** and **B** can be obtained by performing straightforward vector operations on the scalar potential V and the vector potential, respectively, **A**. These potentials are explicit functions of the current and charge distributions present in the system. Hence, if the current and charge distributions on a radiating structure are known, **E** and **B** can be calculated by simply evaluating the integrals given in Equations (14.3) and (14.4) and substituting V and **A** into Equations (14.1) and (14.2). While this may seem like a lot of work, it is often easier than solving Maxwell's equations directly, since the integral form of Maxwell's equations contain the unknowns **E** and **B** inside the integrals.

To derive Equations (14.1)–(14.4), we start by restating Maxwell's equations in differential form:[1]

$$\nabla \times \mathbf{E} = -j\omega\mathbf{B} \tag{14.5}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + j\omega\epsilon\mathbf{E} \tag{14.6}$$

$$\nabla \cdot \epsilon\mathbf{E} = \rho_v \tag{14.7}$$

$$\nabla \cdot \mathbf{B} = 0. \tag{14.8}$$

Because **B** has zero divergence at all points regardless of how the sources are configured, Theorem IV in Section 2-5-6 allows us to express **B** as the curl of a yet-to-be-determined vector potential:

$$\mathbf{B} = \nabla \times \mathbf{A}. \tag{14.9}$$

Next, if we substitute Equation (14.9) into Equation (14.5), we obtain

$$\nabla \times \mathbf{E} = -j\omega\, \nabla \times \mathbf{A},$$

which can be rewritten as

$$\nabla \times (\mathbf{E} + j\omega\mathbf{A}) = 0.$$

Since the vector $(\mathbf{E} + j\omega\mathbf{A})$ has zero curl, we can use Theorem III of Section 2-5-6 to write it as the negative gradient of a yet-to-be-determined scalar potential V:

$$\mathbf{E} + j\omega\mathbf{A} = -\nabla V.$$

[1] In our analysis, it is assumed that any loss in the host medium is treated as polarization loss, i.e., $\epsilon = \epsilon' - j\epsilon''$. This means that **J** is assumed to be a source (or impressed) current.

Solving for **E**, we have

$$\mathbf{E} = -\nabla V - j\omega\mathbf{A}. \tag{14.10}$$

Equations (14.9) and (14.10) show how **E** and **B** can be obtained once **A** and V are known. To derive the equations that define **A** and V, we first substitute Equations (14.9) and (14.10) into Equation (14.6), which, when μ is constant, yields

$$\frac{1}{\mu}\nabla \times \nabla \times \mathbf{A} = \mathbf{J} + j\omega\epsilon(-j\omega\mathbf{A} - \nabla V).$$

Using $\nabla^2\mathbf{A} \equiv \nabla(\nabla \cdot \mathbf{A}) - \nabla \times \nabla \times \mathbf{A}$ and rearranging the resulting expression, we obtain

$$\nabla^2\mathbf{A} + k^2\mathbf{A} - \nabla[\nabla \cdot \mathbf{A} + j\omega\mu\epsilon V] = -\mu\mathbf{J}, \tag{14.11}$$

where $k = \omega\sqrt{\mu\epsilon}$ is the wave number of the medium. Similarly, substituting Equation (14.10) into Equation (14.7), we find that when ϵ is constant,

$$\nabla^2 V + j\omega\nabla \cdot \mathbf{A} = -\frac{\rho_v}{\epsilon}. \tag{14.12}$$

Equations (14.11) and (14.12) are coupled differential equations that relate the potentials **A** and V to the sources **J** and ρ_v. We can uncouple them by noting from Equation (14.9) that the essential property of the vector potential **A** is its curl. Since the curl and divergence of a vector can be specified separately (see Section 2-5-6), we can choose $\nabla \cdot \mathbf{A}$ to be anything that is convenient. The choice that uncouples these differential equations is

$$\nabla \cdot \mathbf{A} = -j\omega\mu\epsilon V. \tag{14.13}$$

Using this relation, which is called the ***Lorentz gauge***,[2] we find that Equations (14.11) and (14.12) reduce to the following uncoupled equations:

$$\nabla^2 V + k^2 V = -\frac{\rho_v}{\epsilon}. \tag{14.14}$$

$$\nabla^2\mathbf{A} + k^2\mathbf{A} = -\mu\mathbf{J}. \tag{14.15}$$

We can further simplify Equation (14.15) by noting that in Cartesian coordinates, the Laplacian of the vector **A** can be written in terms of the Lapacians of its components. (See Equation (2.126).) Hence, Equation (14.15) can be written as

$$\nabla^2 A_i + k^2 A_i = -\mu J_i \qquad i = x, y, \text{or } z. \tag{14.16}$$

Comparing Equations (14.16) and (14.14), we see that they are the same, except for the source terms on their right-hand sides. Hence, they have similar solutions. Also, note that these equations become Poisson's equation (Equation (4.61)) when $k = 0$. Because of this, it should not be surprising that the solutions for V and **A** are

[2] Other gauges can also be used, such as Coulomb's gauge: $\nabla \cdot \mathbf{A} = 0$. When we use these other gauges, the resulting expressions for **A** are more complex, although they yield the same E- and B-fields when substituted into Equations (14.7) and (14.8).

similar to the solution of Poisson's equation (Equation (4.46)). For instance, the particular solution for V is

$$V = \frac{1}{4\pi\epsilon} \int\limits_{\text{Vol.}} \frac{\rho_v(\mathbf{r}')e^{-jk|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}\, dv' \quad [\text{V}]. \tag{14.17}$$

Comparing Equation (14.17) with the electrostatic potential function (given by Equation (4.46)), we see that they are the same when $k = 0$, which occurs when $\omega = 0$. When $\omega \neq 0$, however, the time-harmonic potential V has an additional phase term inside the integral, which, as we will see shortly, is caused by propagation delays.

A rigorous proof that equation 14.17 satisfies equation 14.14 is tedious, but we can outline a heuristic proof by first taking the Laplacian of the scalar potential V that is given by Equation (14.17). Because the ∇^2 operator does not involve the primed coordinates, it can be brought inside the integral, yielding

$$\nabla^2 V = \frac{1}{4\pi\epsilon} \int\limits_{\text{Vol.}} \rho_v(\mathbf{r}')\nabla^2\left[\frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}\right] dv'. \tag{14.18}$$

An exact evaluation of the Laplacian term inside the integral reveals that

$$\nabla^2\left[\frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}\right] = -k^2\frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|} - 4\pi\delta(x-x')\delta(y-y')\delta(z-z'), \tag{14.19}$$

where $\delta(x)$ is the Dirac delta (or impulse) function, with the property $\int_{-\infty}^{\infty} \delta(x - x')\,dx = 1$. The exponential term on the right-hand side of this expression can be found evaluating the Laplacian in Cartesian coordinates, using Equation (2.123). The second term is nonzero only at the point $\mathbf{r} = \mathbf{r}'$ and can be derived by using the divergence theorem (see problem 14-35). Substituting Equation (14.19) into Equation (14.18), we obtain

$$\nabla^2 V = -k^2\frac{1}{4\pi\epsilon} \int\limits_{\text{Vol.}} \frac{\rho_v(\mathbf{r}')e^{-jk|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}\, dv' - \frac{1}{\epsilon}\int\limits_{\text{Vol.}} \rho_v(\mathbf{r}')\delta(x-x')(y-y')\delta(z-z')\,dv'$$

Comparing the first integral in this expression with Equation (14.17), we see this term is $-k^2V$. Using the sampling property of the delta functions, the second integral equals ρ_v/ϵ, evaluated at $\mathbf{r}' = \mathbf{r}$. Thus, the above expression becomes

$$\nabla^2 V + k^2 V = -\frac{\rho_v}{\epsilon},$$

which is the differential equation that defines V. (see Equation (14.14).)

We can write the solution for the vector potential $\mathbf{A}$ almost by inspection. This is because the Cartesian components of $\mathbf{A}$ satisfy the same differential equation as V does (i.e., Equation (14.16)). Thus, we can write

$$A_i = \frac{\mu}{4\pi}\int_{\text{Vol.}} \frac{J_i(\mathbf{r}')e^{-jk|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}\, dv' \quad (i = x, y, \text{ or } z). \tag{14.20}$$

Summing the three components of **A**, we obtain the complete expression,

$$\mathbf{A} = \frac{\mu}{4\pi}\int_{\text{Vol.}} \frac{\mathbf{J}(\mathbf{r}')e^{-jk|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}\,dv' \qquad [\text{Wb/m}] \text{ or } [\text{T}\cdot\text{m}]. \tag{14.21}$$

Interestingly, even though we used the Cartesian coordinate system to develop Equation (14.21), this final expression is not specific to the Cartesian system—any coordinate system can be used.

Example 14-1

Calculate **E** and **H** if it is known that the vector and scalar potentials are

$$\mathbf{A} = A_o e^{-jkz}\hat{\mathbf{a}}_x \qquad [\text{Wb/m}]$$

$$V = 0,$$

where $k = \omega\sqrt{\mu\epsilon}$.

Solution:

First, we note that $\nabla \cdot \mathbf{A} = \partial\,[A_o e^{-jkz}]/\partial x = 0 = V$, so these potentials satisfy the Lorentz condition. Next, using Equation (14.9), we have

$$\mathbf{H} = \frac{1}{\mu}\nabla \times \mathbf{A} = \frac{A_o}{\mu}\frac{\partial}{\partial z}(e^{-jkz})\,\hat{\mathbf{a}}_y = \frac{-jk\,A_o}{\mu}e^{-jkz}\hat{\mathbf{a}}_y.$$

To find **E**, we can substitute the potentials into Equation (14.10), yielding

$$\mathbf{E} = -\nabla V - j\omega\mathbf{A} = -j\omega A_o e^{-jkz}\hat{\mathbf{a}}_x.$$

These E- and H-fields form a linearly polarized plane wave. To see this more clearly, let us select the value of the constant A_o so that it equals $-E_o/j\omega$. Substituting, we find the familiar expressions for a plane wave propagating in the $+z$ direction,

$$\mathbf{E} = E_o e^{-jkz}\hat{\mathbf{a}}_x$$

$$\mathbf{H} = \frac{-kE_o}{\omega\mu}e^{-jkz}\hat{\mathbf{a}}_y = \frac{E_o}{\eta}e^{-jkz}\hat{\mathbf{a}}_y.$$

We can also use the scalar and magnetic potentials to represent electromagnetic fields in the time domain. To do this, we first transform Equations (14.1) and (14.2) into the time domain by replacing the $j\omega$ with $\partial/\partial t$, yielding

$$\mathbf{B} = \nabla \times \mathbf{A} \tag{14.22}$$

$$\mathbf{E} = -\nabla V - \frac{\partial \mathbf{A}}{\partial t}. \tag{14.23}$$

Next, we note that the integrands of both potentials (Equations (14.17) and (14.21)) contain terms of the form $F(\omega, \mathbf{r}')e^{-jk|\mathbf{r}-\mathbf{r}'|}$, where $F(\omega, \mathbf{r}')$ is some function of both frequency and position $\mathbf{r}'$. Since $k = \omega\sqrt{\mu\epsilon}$, the phase of the exponential term is a linear function of ω as long as μ and ϵ are independent of frequency.

This allows us to use the well-known Fourier-transform, time-delay theorem;

$$f(t-\tau) \leftrightarrow F(\omega)e^{-j\omega\tau},$$

where $f(t)$ and $F(\omega)$ form a Fourier transform pair. If we let $\tau = |\mathbf{r}-\mathbf{r}'|/u$, where $u = 1/\sqrt{\mu\epsilon}$ is the speed of light in the medium, we obtain the following Fourier transform pair:

$$f\left[t - \frac{|\mathbf{r}-\mathbf{r}'|}{u}, \mathbf{r}'\right] \leftrightarrow F(\omega, \mathbf{r}')e^{-jk|\mathbf{r}-\mathbf{r}'|}. \tag{14.24}$$

Using Equation (14.24), we find that Equations (14.17) and (14.21) become

$$V = \frac{1}{4\pi\epsilon}\int_{\text{Vol.}} \frac{\rho_v(t', \mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,dv' \quad [\text{V}] \tag{14.25}$$

$$\mathbf{A} = \frac{\mu}{4\pi}\int_{\text{Vol.}} \frac{\mathbf{J}(t', \mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,dv' \quad [\text{Wb/m}] \text{ or } [\text{T}\cdot\text{m}], \tag{14.26}$$

where $t' = t - |\mathbf{r}-\mathbf{r}'|/u$ is the ***delayed time***. According to Equations (14.25) and (14.26), the potentials $\mathbf{A}$ and V resulting from sources at the point $\mathbf{r}'$ at time t' are not observed at the point $\mathbf{r}$ until the delayed time $t' + |\mathbf{r}-\mathbf{r}'|/u$, which is the shortest time that electromagnetic energy can propagate between the points $\mathbf{r}$ and $\mathbf{r}'$ in a homogeneous medium. Because of this time delay, the scalar and vector potentials for time-varying fields are often called ***retarded potentials***.

14-3 The Infinitesimal Dipole

The simplest radiation source is a short segment of current, which, for reasons that will soon be obvious, is called an ***infinitesimal dipole***. This source is also called a ***Hertzian dipole***. Figure 14-2 shows such a current segment, which lies along the z axis and carries a time-harmonic current of amplitude I_0 at all points along its length $\Delta\ell$. The easiest way to calculate the electric and magnetic fields generated by a Hertzian dipole is to first calculate the vector potential $\mathbf{A}$, then $\mathbf{B}$, and then $\mathbf{E}$.

To evaluate the vector potential of a filamentary current, it is best to replace the volumetric current density $\mathbf{J}$ in the expression with the filamentary current I using the substitution

$$\mathbf{J}(\mathbf{r}')dv' = I(\mathbf{r}')\mathbf{d\ell}',$$

With this substitution, the vector potential $\mathbf{A}$ can be written as

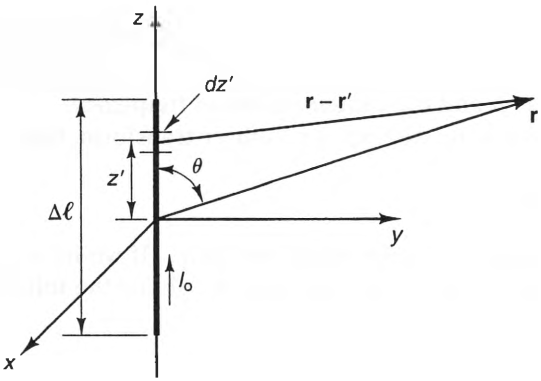


Figure 14-2 Geometry for calculating the fields of a Hertzian dipole.

$$\mathbf{A}(\mathbf{r}) = \frac{\mu}{4\pi}\int_{C'} \frac{I(\mathbf{r}')e^{-jk|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}\,d\boldsymbol{\ell}' \qquad \text{(Filamentary currents),} \tag{14.27}$$

where C' is the contour of the filament.

For the infinitesimal dipole, the terms necessary to evaluate the integral for **A** are

$$I(\mathbf{r}') = I_o, \quad d\boldsymbol{\ell}' = dz'\hat{\mathbf{a}}_z$$
$$\mathbf{r}' = z\hat{\mathbf{a}}_z, \quad \mathbf{r} = r\hat{\mathbf{a}}_r$$
$$|\mathbf{r}-\mathbf{r}'| = |r\hat{\mathbf{a}}_r - z'\hat{\mathbf{a}}_z|.$$

Substituting these expressions into Equation (14.27) and noting that the unit vector $\hat{\mathbf{a}}_z$ is a constant along the filament, we obtain

$$\mathbf{A} = \frac{\mu I_o\hat{\mathbf{a}}_z}{4\pi}\int_{-\Delta\ell/2}^{\Delta\ell/2} \frac{e^{-jk|r\hat{\mathbf{a}}_r - z'\hat{\mathbf{a}}_z|}}{|r\hat{\mathbf{a}}_r - z'\hat{\mathbf{a}}_z|}\,dz', \tag{14.28}$$

When the filament length $\Delta\ell$ is small, we can evaluate this integral as follows. First, we can expand the term $|r\hat{\mathbf{a}}_r - z'\hat{\mathbf{a}}_z|$ using Cartesian components:

$$|r\hat{\mathbf{a}}_r - z'\hat{\mathbf{a}}_z| = |r\sin\theta\cos\phi\hat{\mathbf{a}}_x + r\sin\theta\sin\phi\hat{\mathbf{a}}_y + (r\cos\theta - z')\hat{\mathbf{a}}_z|$$
$$= \sqrt{r^2 - 2rz'\cos\theta + (z')^2}.$$

If $z' \ll r$, we can ignore the $(z')^2$ term and use a two-term binomial expansion of the remaining terms, yielding

$$|r\hat{\mathbf{a}}_r - z'\hat{\mathbf{a}}_z| \approx \sqrt{r^2 - 2rz'\cos\theta} \approx r - z'\cos\theta. \tag{14.29}$$

Substituting Equation (14.29) into Equation (14.28), we have

$$\mathbf{A} \approx \frac{\mu I_o\hat{\mathbf{a}}_z}{4\pi}\int_{-\Delta\ell/2}^{\Delta\ell/2} \frac{e^{-jk(r-z'\cos\theta)}}{r - z'\cos\theta}dz' = \frac{\mu I_o\hat{\mathbf{a}}_z e^{-jkr}}{4\pi}\int_{-\Delta\ell/2}^{\Delta\ell/2} \frac{e^{jkz'\cos\theta}}{r - z'\cos\theta}dz'. \tag{14.30}$$

As long as $\Delta\ell \ll r$, we can safely ignore the term z'-$\cos\theta$ in the denominator. However, in order for us to be able to ignore this same term in the exponential function, we must have $kz' \ll 1$; otherwise, there will be a significant phase error in the integral. Since $k = 2\pi/\lambda$, the product kz' will be small throughout the integration if we restrict the filament length $\Delta\ell$ such that $\Delta\ell \ll \lambda$. Hence, Equation (14.30) can be approximated as

$$\mathbf{A} \approx \frac{\mu I_0 \Delta\ell}{4\pi} \frac{e^{-jkr}}{r} \hat{\mathbf{a}}_z \qquad (\Delta\ell \ll r \text{ and } \Delta\ell \ll \lambda). \tag{14.31}$$

Now that we have an expression for the vector potential **A**, finding the E- and H-fields is straightforward. Substituting Equation (14.31) into Equation (14.2), we have

$$\mathbf{H} = \frac{1}{\mu} \nabla \times \mathbf{A} = \frac{I_0 \Delta\ell}{4\pi} \nabla \times \left(\frac{e^{-jkr}}{r} \hat{\mathbf{a}}_z \right). \tag{14.32}$$

Using Table B-3, we can express $\hat{\mathbf{a}}_z$ in the spherical coordinate system as

$$\hat{\mathbf{a}}_z = \cos\theta \hat{\mathbf{a}}_r - \sin\theta \hat{\mathbf{a}}_\theta.$$

Hence, Equation (14.32) can be written as

$$\mathbf{H} = \frac{I_0 \Delta\ell}{4\pi} \nabla \times \left(\frac{e^{-jkr}}{r} \cos\theta\, \hat{\mathbf{a}}_r - \frac{e^{-jkr}}{r} \sin\theta\, \hat{\mathbf{a}}_\theta \right).$$

Since the vector in the brackets does not vary with ϕ and has no ϕ component, most of the terms in the curl expression vanish, yielding

$$\mathbf{H} = \frac{I_0 \Delta\ell}{4\pi r} \left[\frac{\partial}{\partial r} \left(r \frac{e^{-jkr}}{r} \cos\theta \right) - \frac{\partial}{\partial\theta} \left(-\frac{e^{-jkr}}{r} \sin\theta \right) \right] \hat{\mathbf{a}}_\phi.$$

Evaluating the derivatives, we obtain

$$\mathbf{H} = H_\phi \hat{\mathbf{a}}_\phi = \frac{I_0 \Delta\ell}{4\pi} \sin\theta \left[\frac{jk}{r} + \frac{1}{r^2} \right] e^{-jkr} \hat{\mathbf{a}}_\phi. \tag{14.33}$$

At this point, we have a choice of how to proceed to find **E**. One way is to use the Lorentz gauge to find V from **A** and then use the relation $\mathbf{E} = -\nabla V - j\omega\mathbf{A}$. Another, simpler way is to substitute the H-field expression we have just found into Maxwell's curl-H equation and solve for **E**. Knowing that $\mathbf{J} = 0$ everywhere except on the filament itself, we can write

$$\mathbf{E} = \frac{\nabla \times \mathbf{H}}{j\omega\epsilon},$$

Substituting Equation (14.33) into this expression and evaluating the curl operation, we obtain

$$\mathbf{E} = E_r \hat{\mathbf{a}}_r + E_\theta \hat{\mathbf{a}}_\theta, \tag{14.34}$$

where

$$E_r = \frac{I_o \Delta\ell}{4\pi} 2\eta \cos\theta \left[\frac{1}{r^2} - \frac{j}{kr^3}\right] e^{-jkr} \tag{14.35}$$

$$E_\theta = \frac{I_o \Delta\ell}{4\pi} \eta \sin\theta \left[\frac{jk}{r} + \frac{1}{r^2} - \frac{j}{kr^3}\right] e^{-jkr}, \tag{14.36}$$

and $\eta = \sqrt{\mu/\epsilon}$ is the intrinsic impedance of the medium.

As the reader has no doubt noticed, the E- and H-field expressions given by Equations (14.33), (14.35), and (14.36) are *not* simple, even though a short current filament is about as simple a source as one could imagine. What makes these expressions so formidable is that they contain terms with three different rates of decay: $1/r$, $1/r^2$, and $1/r^3$. In the discussion that follows, we will show that the $1/r^3$ and $1/r^2$ components are similar to those generated by time-invariant currents and charges, whereas the $1/r$ components are radiation fields, found only for time-varying sources.

Starting with the electric field, we note from Equations (14.35) and (14.36) that the $1/r^3$ terms are dominant when $r \ll 1/k = \lambda/2\pi$, where λ is the wavelength of a plane wave in the ambient medium. This region is called the ***near-zone region***, where we find that

$$\mathbf{E} \approx \frac{-jI_o \Delta\ell}{4\pi r^3} \frac{\eta}{k} [2\cos\theta\, \hat{\mathbf{a}}_r + \sin\theta\, \hat{\mathbf{a}}_\theta] \qquad (r \ll \lambda/2\pi). \tag{14.37}$$

We can interpret this result more easily by noting that since the current is uniform along the filament, point charges $\pm Q_o$ collect at the upper and lower endpoints, respectively. In the frequency domain, $I_o = \partial Q_o/\partial t$, so it follows that in the frequency domain that

$$I_o = j\omega Q_o \tag{14.38}$$

Substituting Equation (14.38) into Equation (14.37) and using $\eta/k = (\omega\epsilon)^{-1}$, we find that

$$\mathbf{E} \approx \frac{Q_o \Delta\ell}{4\pi\epsilon r^3} [2\cos\theta\, \hat{\mathbf{a}}_r + \sin\theta\, \hat{\mathbf{a}}_\theta] \qquad (r \ll \lambda/2\pi). \tag{14.39}$$

This expression is the same as the field generated by a static dipole (Equation (4.54)), except that the amplitude of the field varies sinusoidally with time. Hence, at short distances, the field generated by a short current filament is dominated by the effects of the charges collecting at the endpoints. This is why we call this source an infinitesimal dipole.

The near-zone behavior of the H-field can be found by noting that the $1/r^2$ term in Equation (14.33) is dominant in that region. Thus, in the near zone we have

$$\mathbf{H} \approx \frac{I_o \Delta\ell}{4\pi r^2} \sin\theta\, \hat{\mathbf{a}}_\phi \qquad (r \ll \lambda/2\pi). \tag{14.40}$$

Except for the fact that this field varies sinusoidally in time, the H-field component distributes itself exactly as the field of a steady-current filament. (See Equation (7.22).)

From the foregoing comments, we see that the electric and magnetic fields close to an infinitesimal dipole behave like those generated by a static dipole and a steady current, respectively. Except for the sinusoidal variations in these fields, there is nothing really new here. When r is large, however, the fields take on a distinct time-varying behavior. To see this, we note that when $r \gg \lambda$, only the $1/r$ terms in the expressions for **E** and **H** are significant. Hence, in the far zone, we have

$$\mathbf{E} \approx \frac{jkI_o\Delta\ell}{4\pi}\eta \sin\theta \frac{e^{-jkr}}{r}\hat{\mathbf{a}}_\theta \qquad (r \gg \lambda) \tag{14.41}$$

$$\mathbf{H} \approx \frac{jkI_o\Delta\ell}{4\pi}\sin\theta \frac{e^{-jkr}}{r}\hat{\mathbf{a}}_\phi \qquad (r \gg \lambda). \tag{14.42}$$

The $1/r$ decay rate exhibited by these fields is new: Static current and charge distributions with finite dimensions do not generate fields that decay proportional to $1/r$. Also, the fields are outward propagating, as indicated by the e^{-jkr} phase terms. Because of this, we also call the region $r \gg \lambda$ the ***radiation zone***, and the fields in this region are called ***radiation fields***. When viewed from a global perspective, the fields in the radiation zone have spherical phase fronts, as shown in Figure 14-3. However, when viewed over a small range of angles, these phase fronts appear planar, which is one indication that they can be approximated as plane waves. Other indications are that $|\mathbf{E}|/|\mathbf{H}| = \eta$, and **E** and **H** are perpendicular both to each other and to the direction of propagation. As a result, these fields "look" like plane waves when viewed locally by an observer in the far zone. This is useful in analysis and design, since plane waves are simple to model.

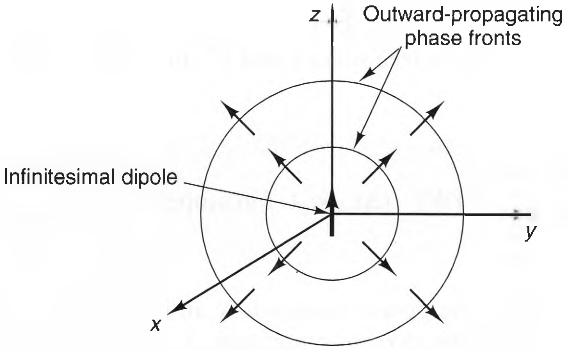


Figure 14-3 Radiation phase fronts of an infinitesimal (Hertzian) dipole.

14-3-1 RADIATED POWER AND RADIATION RESISTANCE

Given that the fields generated by an infinitesimal dipole propagate outward from the source and behave locally like plane waves in the far zone, it is logical to assume that these fields carry power with them. To show that this is so, we first calculate the complex Poynting vector

$$\mathbf{S} = \mathbf{E} \times \mathbf{H}^* \qquad [\mathrm{W/m^2}].$$

associated with these fields. Substituting Equations (14.33)–(14.36) into this expression, we obtain

$$\mathbf{S} = \eta\left(\frac{I_o\Delta\ell}{4\pi}\right)^2\left\{2\sin\theta\cos\theta\left[-\frac{jk}{r^3} - \frac{j}{kr^5}\right]\hat{\mathbf{a}}_\theta + \sin^2\theta\left[\frac{k^2}{r^2} - \frac{j}{kr^5}\right]\hat{\mathbf{a}}_r\right\} \qquad [\mathrm{W/m^2}].$$

Remembering that the average Poynting vector $\mathscr{S}_{\text{ave}}$ is 1/2 times the real part of **S**, we can write

$$\mathscr{S}_{\text{ave}} = \frac{1}{8}\,\eta\sin^2\theta\left(\frac{I_o\Delta\ell}{\lambda r}\right)^2\hat{\mathbf{a}}_r \qquad [\mathrm{W/m^2}], \tag{14.43}$$

where we have used $k = 2\pi/\lambda$. Comparing the preceding expressions for **S** and $\mathscr{S}_{\text{ave}}$, we see that the complex power **S** has components in both the r and θ directions, but the average power $\mathscr{S}_{\text{ave}}$ has only an r component, since the θ component of **S** is imaginary and corresponds to stored, not radiated, energy.

We can find the total power radiated by an infinitesimal dipole simply by integrating the average Poynting vector $\mathscr{S}_{\text{ave}}$ around any closed surface that completely encloses the dipole:

$$P_{\text{rad}} = \oint_S \mathscr{S}_{\text{ave}} \cdot \mathbf{ds}.$$

Since $\mathscr{S}_{\text{ave}}$ has only an r component, a sphere is the easiest surface for integrating. For this case, $\mathbf{ds} = r^2 \sin\theta\, d\theta\, d\phi\, \hat{\mathbf{a}}_r$, which yields

$$P_{\text{rad}} = \frac{1}{8}\,\eta\left(\frac{I_o\Delta\ell}{\lambda}\right)^2\int_0^{2\pi}\int_0^{\pi}\frac{\sin^2\theta}{r^2}\,\hat{\mathbf{a}}_r\cdot\hat{\mathbf{a}}_r r^2\sin\theta\, d\theta\, d\phi.$$

Since $\hat{\mathbf{a}}_r \cdot \hat{\mathbf{a}}_r = 1$, $\eta \approx 120\pi\,[\Omega]$ (in free space), and $\int_0^\pi \sin^3\theta\, d\theta = 4/3$, P_{rad} can be written as

$$P_{\text{rad}} = 40\pi^2 I_o^2\left(\frac{\Delta\ell}{\lambda}\right)^2 \qquad [\mathrm{W}] \quad (\Delta\ell \ll \lambda, I \text{ in amperes}). \tag{14.44}$$

Equation (14.44) shows that the power radiated by an infinitesimal dipole is proportional to the squares of both the current magnitude I_o and the length-to-wavelength ratio $\Delta\ell/\lambda$. This observation has important applications both for elements that are designed to radiate and for those that are not. For instance, in order for a short dipole to radiate large amounts of power, the current levels must be high. On the other hand, the radiation from printed circuit board (PCB) traces can become troublesome whenever the circuit paths or component lengths approach even 1/20th of a wavelength.

Since a short current segment radiates power, it follows that it must present a resistive load to the source that drives this current. To see why, consider the situation shown in Figure 14-4a, which depicts a sinusoidal current source with magnitude I_o, connected to the midpoint of a short, perfectly conducting wire dipole of total length $\Delta\ell$. If the wire is loaded with large capacitors at the ends,[3] the current excited along the wire will be uniform, and it will radiate a power P_{rad}, given by Equation (14.44). This power must be supplied by the current source, so

$$P_{\text{in}} = P_{\text{rad}} = 40\pi^2 I_o^2 \left(\frac{\Delta\ell}{\lambda}\right)^2. \tag{14.45}$$

Also, since the wire is passive, the equivalent circuit looking into the terminals must be an impedance $Z_{\text{in}} = R_{\text{in}} + jX_{\text{in}}$ [Ω], which is depicted in Figure 14.4b. In terms of this impedance, the input power to the dipole can be expressed as

$$P_{\text{in}} = \frac{1}{2} R_{\text{in}} I_o^2. \tag{14.46}$$

Substituting Equation (14.45) into Equation (14.46) and solving for R_{in}, we obtain

$$R_{\text{in}} = \frac{2\,P_{\text{in}}}{I_o^2} = 80\pi^2\left(\frac{\Delta\ell}{\lambda}\right)^2. \tag{14.47}$$

The filaments are perfectly conducting, so the input resistance R_{in} can only be due to radiation effects. For that reason, we call this resistance the ***radiation resistance*** of the wire, denoted by the symbol R_r, and given by

$$R_r = \frac{2\,P_{\text{rad}}}{I_o^2} = 80\pi^2\left(\frac{\Delta\ell}{\lambda}\right)^2 \; [\Omega] \qquad (\Delta\ell \ll \lambda). \tag{14.48}$$

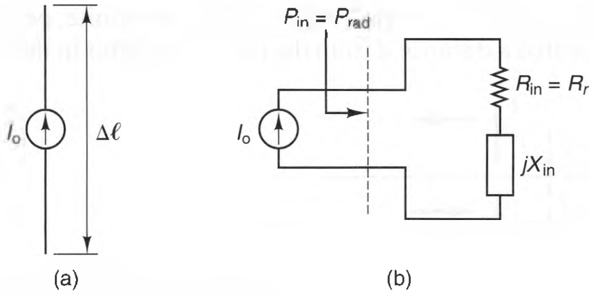


Figure 14-4 An infinitesimal (Hertzian) dipole, driven by a current source. a) Physical geometry. b) Equivalent circuit at the input terminals.

[3] In practice, this can be accomplished by attaching metal discs to the wire ends.

From the standpoint of the source that drives the current, radiation resistance is indistinguishable from ohmic resistance. In both cases, the source must continually supply energy to the element in order to keep the current amplitude constant with time. In the case of ohmic resistance, this energy is transformed into thermal energy. Radiation resistance, on the other hand, converts the energy into propagating electromagnetic waves.

Example 14-2

Calculate the radiation resistance of a 1 [cm] length of uniform current if the frequency is 900 [MHz] and the host medium is air.

Solution:

At 900 [MHz], the wavelength in air is:

$$\lambda = \frac{c}{f} = \frac{3 \times 10^8}{900 \times 10^6} = 33.3 \qquad [\mathrm{cm}].$$

Since $\Delta\ell = 1$ [cm] $\ll \lambda$, the radiation resistance of this current is given by Equation (14.48),

$$R_r = 80\pi^2\left(\frac{1 \times 10^{-2}}{33.3 \times 10^{-2}}\right)^2 = 0.711 \qquad [\Omega].$$

As can be seen from this example, the radiation resistance of a short current segment is only on the order of a fraction of an ohm, making it a relatively inefficient radiator. Later, we will show that currents that flow on structures a half wavelength or larger typically have much larger, and more usable, radiation resistances.

14-3-2 THE FIELDS OF CURRENTS ABOVE A PERFECTLY CONDUCTING GROUND

There are many situations where currents flow above large conducting planes. Typical examples are antennas above the earth and wires mounted near a conducting chassis. In many cases, we can consider these conducting planes to be infinitely large and perfectly conducting. Also, since all current distributions are simply collections of infinitesimal dipoles, we can model these situations by considering the fields generated by infinitesimal dipoles above an infinite, perfectly conducting ground.

Figure 14-5a shows two infinitesimal dipoles above an infinite, perfectly conducting plane. Both are located a distance d from the plane, but differ in that one current is

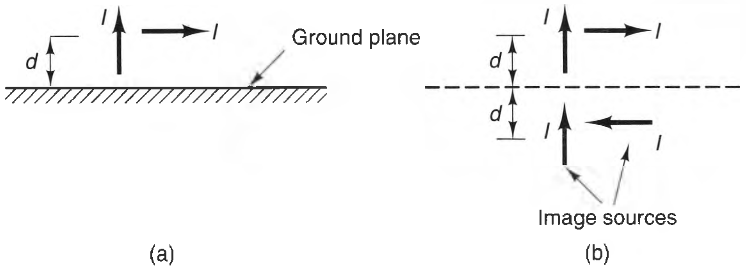


Figure 14-5 Currents above a conducting ground plane. a) Physical geometry. b) Equivalent geometry.

oriented parallel to the ground and the other perpendicular to the ground. Both of these currents radiate fields that, in turn, induce secondary currents in the ground plane. We can model the effect of these secondary currents by replacing the ground plane with image sources that maintain the same boundary conditions imposed by a conductor: $E_{\text{tan}} = 0$ at all points on the plane. This equivalent geometry is shown in Figure 14-5b.

The image of a perpendicular current element above a ground plane is an identical current element the same distance below the ground plane and with the same orientation. In this case, both the source and its image generate E_r and E_θ components (where the z-axis is perpendicular to the ground plane), each having a component parallel to the ground plane. However, since E_r is proportional to cos θ and E_θ is proportional to sin θ, the tangential component of E_r from the source is canceled by the tangential component of E_r from the image. The same occurs for the tangential components arising from the E_θ components of the currents.

The image of a tangentially oriented current above a ground plane is also an identical current element, located the same distance below the ground plane, but with an opposite orientation. An easy way to see why the source and the image must be directed oppositely is to consider a point on the ground plane that is exactly between the two currents. For this case, the E_θ components of both sources are parallel to the ground plane and will cancel when the currents are oppositely directed.

The fields of the image sources can add either constructively or destructively with the fields of the original sources, depending upon the position of the observer and the distance of the source current from the ground plane. One special case occurs when the distance from the source to the ground plane is small. In this case, an observer above the ground plane "sees" the source and its image at essentially the same point, so the image of a perpendicular current has the effect of doubling the field of the source, whereas the image of a tangential current tends to cancel the fields of the source. For that reason, antennas mounted very close (in wavelengths) to a ground are nearly always oriented so that the currents flow perpendicular to the ground. A common example is the vertical dipole antenna, which we will discuss later.

Example 14-3

A current element in free space with length $\Delta\ell = 0.5$ [cm] and peak current $I_o = 0.1$ [A] at 1 [GHz] is positioned $d = 1$ [cm] above a perfectly conducting ground plane that is tangent to the current. Calculate the magnitude of the E-field 1 [km] above the ground plane, directly above the source. Compare this with the E-field magnitude if the ground plane were removed.

Solution:

The wavelength in free space is $\lambda = c/f = 30$ [cm]. Thus, $\Delta\ell \ll \lambda$, and this current can be considered as an infinitesimal dipole. Also, since 1 [km] $\gg \lambda$, the observation point is in the far zone, which means that we can use Equation (14.41) for the E-fields radiated by both the source and its image. Directly above the sources, sin $\theta = 1$, so, using the superposition principle, we can express the field due to the source and its oppositely directed image as

$$E_\theta = \frac{jkI_o\Delta\ell}{4\pi}\eta\left(\frac{e^{-jkr_1}}{r_1} - \frac{e^{-jkr_2}}{r_2}\right),$$

where r_1 is the distance from the observer to the source (10^5 [cm] $-$ 1 [cm]), r_2 is the distance from the observer to the image (10^5 [cm] $+$ 1 [cm]), and $k = 2\pi/\lambda = 0.209$ [cm^{-1}]. Since $1/r_1 \approx 1/r_2 \approx 1/r$, where $r = 1$ [km], we can write this expression as

$$E_\theta \approx \frac{jkI_o\Delta\ell}{4\pi r}\eta(e^{-jkr_1} - e^{-jkr_2}).$$

Also, since $r_1 = r - d$ and $r_2 = r + d$, we can write the preceding expession as

$$E_\theta \approx \frac{jkI_o\Delta\ell}{4\pi r}\eta e^{-jkr}(e^{jkd} - e^{-jkd}) = \frac{jkI_o\Delta\ell}{4\pi r}\eta e^{-jkr}(j\,2\sin kd),$$

where we have used Euler's identity to write $e^{jkd} - e^{-jkd}$ as $j\,2\sin kd$. Using the values $\eta = 377$ [Ω], $|e^{-jkr}| = 1$, and $2\sin kd = 2\sin(1.047) = 0.416$, we find that

$$|E_\theta| \approx \frac{0.209[\text{cm}^{-1}] \times 0.1[\text{A}] \times 0.5[\text{cm}] \times 377[\Omega] \times 0.416}{4\pi \times 10^5[\text{cm}]} = 1.3 \qquad [\mu\text{V/m}].$$

If the ground were not present, $|E_\theta|$ would be given by the same expression, but without the 2 sin $kd = 0.416$ term, so we would have

$$|E_\theta| \approx \frac{1.3[\mu\text{V/m}]}{0.416} = 3.135 \qquad [\mu\text{V/m}].$$

Thus, for this case, we see that the ground plane has the effect of reducing the field of the dipole by (3.135 − 1.3)/3.135 × 100 = 58%.

14-4 Transmitting Antenna Parameters

The infinitesimal dipole is a simple prototype that demonstrates the basic operating characteristics of many antennas. Just as circuit engineers never stop using Ohm's law, antenna engineers often find themselves modeling antenna performance using the properties of infinitesimal dipoles. Before we discuss more complicated antennas, however, it is worthwhile to define the various parameters that are used in engineering practice to specify the performance of an antenna when it is acting as a transmitter. In general, apart from its mechanical characteristics, what distinguishes one type of antenna from another is 1) the range of angles over which its radiated power is directed, 2) the input impedance looking into the feed-point terminals, and 3) its bandwidth.

In this section we will define a number of parameters that are commonly used to quantify the operating characteristics of antennas. Although the discussion will center upon the transmitting characteristics of antennas, we will see later in the chapter that most these parameters can also be used to describe their receiving characteristics.

14-4-1 RADIATION INTENSITY

In the preceding section, we saw that the radiated power density generated by an infinitesimal dipole is directed radially outward from the source and decays as $1/r^2$. This is true for all radiation sources with finite dimensions, which means that we can express the radiated power density of any radiation source as a function of angular position, divided by r^2:

$$\mathcal{S}_{\text{ave}} = \frac{U(\theta, \phi)}{r^2}\hat{\mathbf{a}}_r \qquad [\text{W/m}^2]. \tag{14.49}$$

The function $U(\theta, \phi)$ is called the ***radiation intensity***, which, as we will soon show, is measured in units of watts per steradian:

$$U(\theta, \phi) = r^2\,\mathcal{S}_{\text{ave}} \qquad [\text{W/sr}]. \tag{14.50}$$

Since the fields in the radiation zone always behave locally as plane waves, we can express the power density using Equation (12.107),

$$\mathcal{S}_{\text{ave}} = \frac{|\mathbf{E}|^2}{2\eta}.$$

Substituting this into Equation (14.50), we obtain an expression that specifies the radiation intensity U in terms of the E-field:

$$U(\theta, \phi) = r^2\frac{|\mathbf{E}|^2}{2\eta} \qquad [\text{W/sr}]. \tag{14.51}$$

Whereas the power density $\mathcal{S}_{\text{ave}}$ is a measure of the power passing through a unit area, the radiation intensity U is a measure of the power passing through a unit solid angle. To see this more clearly, let us calculate the power P_{rad} radiated by a source. Using $P_{\text{rad}} = \oint_S \mathcal{S}_{\text{ave}} \bullet \mathbf{ds}$ and integrating around a sphere, we obtain

$$P_{\text{rad}} = \oint_S \frac{U(\theta, \phi)}{r^2}\hat{\mathbf{a}}_r \bullet \hat{\mathbf{a}}_r r^2 \sin\theta\, d\theta\, d\phi.$$

This can be written as

$$P_{\text{rad}} = \oint_S U(\theta, \phi) d\Omega, \tag{14.52}$$

where $d\Omega = \sin\theta\, d\theta\, d\phi$ is the area traced on a unit sphere when θ and ϕ change by the amounts $d\theta$ and $d\phi$, respectively. Since $d\Omega$ is measured in units of steradians [sr], U is measured in units of watts per steradian [W/sr].

Example 14-4

Suppose that an antenna has a radiation intensity given by

$$U(\theta, \phi) = \begin{cases} 1.5\cos\theta & [\text{W/sr}] \quad 0 < \theta < \pi/2,\ 0 < \phi < 2\pi \\ 0 & \qquad\quad\ \text{otherwise} \end{cases}.$$

Find the power radiated by this antenna.

Solution:

Using Equation (14.52), we obtain

$$P_{\text{rad}} = \oint_S U(\theta, \phi) d\Omega = \int_0^{2\pi} \int_0^{\pi/2} U(\theta, \phi) \sin\theta \, d\theta \, d\phi$$

$$= \int_0^{2\pi} \int_0^{\pi/2} 1.5 \cos\theta \sin\theta \, d\theta \, d\phi = 3\pi \int_0^{\pi/2} \cos\theta \sin\theta \, d\theta$$

$$= \frac{3\pi}{2} = 4.71 \qquad [\text{W}].$$

14-4-2 ANTENNA PATTERNS

A ***radiation pattern*** is simply a plot of the radiation characteristics of an antenna. There are two types of radiation patterns. A plot of the radiated power at a constant radius is called a ***power pattern***. Similarly, a plot of the electric (or magnetic) field magnitude at a constant radius is called a ***field pattern***. These patterns can be plotted either in absolute units or in dB. Figure 14-6a shows the radiation pattern of a typical antenna. As can be seen, this pattern consists of a number of lobes. The largest lobe is usually called the ***main lobe***, and the others are called ***side lobes***. The minima between the lobes are called ***nulls***.

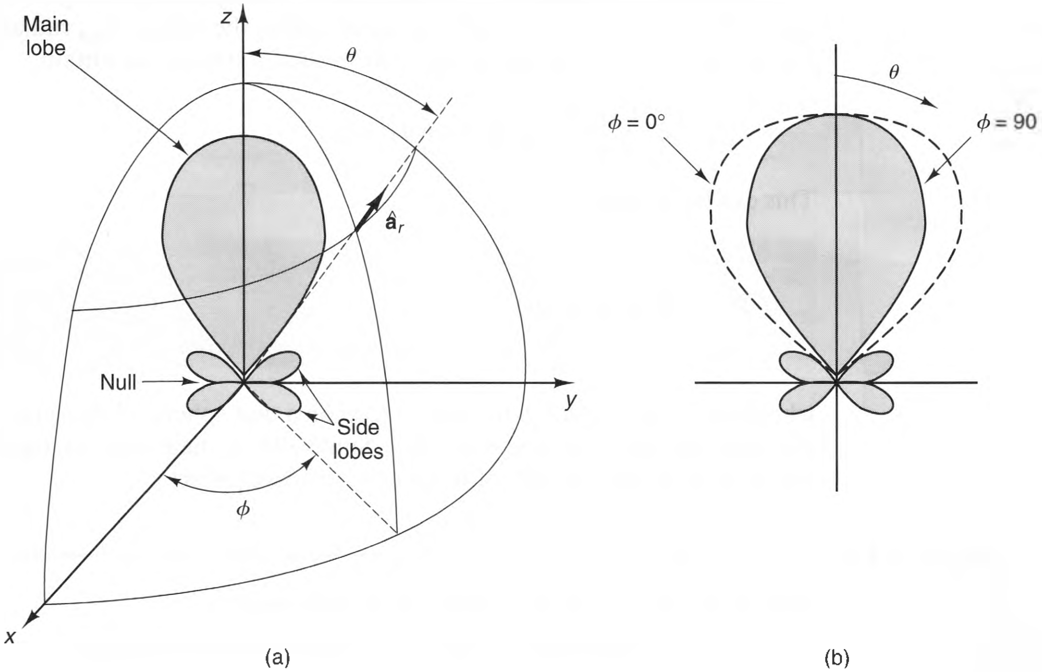


Figure 14-6 Antenna patterns. a) A three-dimensional pattern. b) Two-dimensional cuts.

Radiation patterns are three-dimensional entities, but they are usually measured and displayed as a series of two-dimensional patterns, called ***cuts***. For most antennas, two cuts are sufficient to convey a good idea of their three-dimensional patterns. Figure 14-6b shows two cuts of the three-dimensional radiation pattern in Figure 14-6a for $\phi = 0°$ and $\phi = 90°$, respectively. As can be seen, these cuts are similar, but not identical, indicating that the three-dimensional pattern is not symmetrical.

The radiation patterns of linearly polarized antennas are often specified in terms of their ***E-Plane*** and ***H-Plane*** patterns. By definition, the E-plane contains the direction of maximum radiation and the electric field vector. Similarly, the H-plane contains the direction of maximum radiation and the magnetic field vector. These planes are perpendicular to each other, since **E** and **H** are always perpendicular in the far zone. Figure 14-7 shows the E-plane and H-plane patterns of a horn antenna.

Two kinds of antenna pattern shapes are given special names. The first is an ***isotropic pattern***, which is the same in all directions. No antenna can have a truly isotropic pattern, but it is often a convenient idealization. An ***omnidirectional pattern*** is rotationally symmetric about an axis, often called the azimuthal axis. Figure 14-8 shows an omidirectional antenna pattern. Omnidirectional antennas are often used in broadcast applications, since they provide uniform coverage in all directions around them.

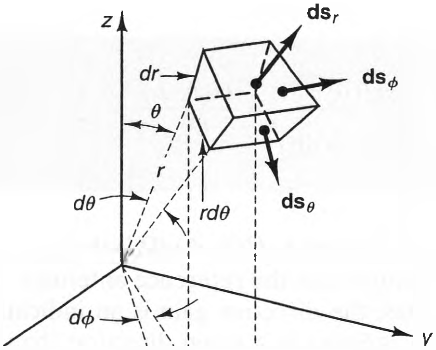


Figure 14-7 *E*- and *H*- plane patterns for a horn antenna.

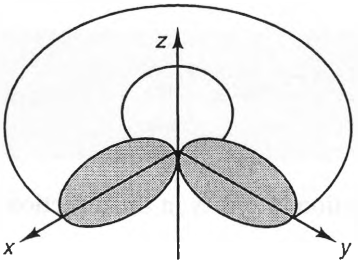


Figure 14-8 An omnidirectional antenna pattern.

14-4-3 DIRECTIVE GAIN AND DIRECTIVITY

The ***directive gain*** D_g of an antenna is the ratio of its radiation intensity $U(\theta, \phi)$ in a given direction to the radiation intensity $U_{\text{ref}}(\theta, \phi)$ of a reference antenna that radiates the same total power P_{rad}:

$$D_g(\theta, \phi) = \frac{U(\theta, \phi)}{U_{\text{ref}}(\theta, \phi)} \quad \text{(Dimensionless)}, \qquad (14.53)$$

where,

$$P_{\text{rad}} = \oint_S U(\theta, \phi) d\Omega = \oint_S U_{\text{ref}}(\theta, \phi) d\Omega.$$

In practice, a lossless, isotropic radiator is usually chosen as the reference antenna. For this case, $U_{\text{ref}}(\theta, \phi)$ is independent of θ and ϕ, so

$$\oint_S U_{\text{ref}} d\Omega = 4\pi U_{\text{ref}}.$$

Substituting the preceding expressions into Equation (14.53) and noting that $U_{\text{ave}} = 1/(4\pi) \oint_S U(\theta, \phi) d\Omega$, we obtain the following equivalent expressions for D_g:

$$D_g(\theta, \phi) = \frac{4\pi U(\theta, \phi)}{P_{\text{rad}}} = \frac{4\pi U(\theta, \phi)}{\oint_S U(\theta, \phi) d\Omega} = \frac{U(\theta, \phi)}{U_{\text{ave}}} \quad \text{(Istropic reference radiator)} \qquad (14.54)$$

In this expression, $U_{\text{ave}} = 1/(4\pi) \oint_S U(\theta, \phi) d\Omega$ is the average radiation intensity of the antenna under consideration (not the reference antenna).

As its name implies, the directive gain is an indication of how much more (or less) power an antenna radiates in a given direction than an isotropic antenna does. The directive gain in the direction of maximum radiation is called the ***directivity*** and is given by

$$D_o = D_g(\theta, \phi)\Big|_{\text{max}} = \frac{4\pi U_{\text{max}}}{P_{\text{rad}}} = \frac{U_{\text{max}}}{U_{\text{ave}}} \quad \text{(Istropic reference radiator)}, \qquad (14.55)$$

where U_{max} is the radiation intensity in the direction of maximum radiation.

Example 14-5

Find the directivity of an infinitesimal dipole.

Solution:

The average Poynting vector for an infinitesimal dipole is given by Equation (14.43),

$$\mathscr{P}_{\text{ave}} = \frac{1}{8}\eta \sin^2\theta \left(\frac{I_{\text{o}}\Delta\ell}{\lambda r}\right)^2 \hat{\mathbf{a}}_r \qquad [\text{W/m}^2].$$

Hence, the radiation intensity $U(\theta, \phi)$ is of the form

$$U(\theta, \phi) = U_{\max} \sin^2\theta,$$

where $U_{\max}$ is the radiation intensity along $\theta = 90°$. Substituting this into Equation (14.55), we have

$$D_{\text{o}} = \frac{4\pi U_{\max}}{\oint_S U_{\max}\sin^2\theta\, d\Omega} = \frac{4\pi}{\int_0^{2\pi}\int_0^{\pi} \sin^2\theta \sin\theta\, d\theta\, d\phi} = \frac{4\pi}{(8\pi/3)}$$

Thus,

$$D_{\text{o}} = 1.5.$$

We can also express this in dB:

$$D_{\text{o}} = 10\log(1.5) = 1.76\ [\text{dB}].$$

14-4-4 POWER GAIN AND RADIATION EFFICIENCY

The directive gain $D_g(\theta, \phi)$ indicates where an antenna radiates its energy, but does not indicate what percentage of the input power is actually radiated. To convey this additional information, we define the ***power gain*** $G_g(\theta, \phi)$ of an antenna as the ratio of the radiation intensity $U(\theta, \phi)$ to the radiation intensity U_{ref} of a *lossless* reference antenna that has the same input power P_{in}. Usually, the reference antenna is an isotropic radiator, so $P_{\text{in}} = 4\pi U_{\text{ref}}$, which yields

$$G_g(\theta, \phi) = \frac{4\pi U(\theta, \phi)}{P_{\text{in}}} \qquad \text{(Isotropic reference radiator)}. \tag{14.56}$$

The maximum power gain of an antenna is denoted by

$$G_{\text{o}} = \frac{4\pi U_{\max}}{P_{\text{in}}} \qquad \text{(Isotropic reference radiator)}. \tag{14.57}$$

Usually, when the term "antenna gain" is used in practice, the maximum power gain is implied.

Comparing Equation (14.57) with Equation (14.55), we see that the definitions of the directivity and gain of an antenna differ only in that the directivity has the radiated power in the denominator, whereas the gain has the input power in the denominator. This difference is subtle, but important. The ratio of these quantities is the ***radiation efficiency*** η_r of the antenna, given by

$$\eta_r = \frac{G_o}{D_o} = \frac{P_{\text{rad}}}{P_{\text{in}}}. \tag{14.58}$$

Most antennas have efficiencies well over 90%. Exceptions to this are electrically small antennas and ones that make extensive use of dielectrics, such as a horn antenna with a lossy dielectric lens.

14-4-5 RADIATION RESISTANCE AND INPUT IMPEDANCE

Figure 14-9a shows an arbitrary antenna with a pair of input terminals a and b. If the antenna is not receiving power from waves generated by other sources, the Thévenin equivalent circuit looking into these terminals will consist only of an impedance

$$Z_{\text{in}} = R_{\text{in}} + jX_{\text{in}} \qquad [\Omega], \tag{14.59}$$

where R_{in} and X_{in} are the input resistance and reactance, respectively. This Thévenin equivalent circuit is shown in Figure 14-9b. In general, the antenna input resistance R_{in} is the sum of two components:

$$R_{\text{in}} = R_{ri} + R_L. \tag{14.60}$$

R_{ri} and R_L are the ***input radiation resistance*** and ***input loss resistance*** of the antenna, respectively. As their names imply, the loss resistance R_L accounts for that portion of the input power that is dissipated in heat, whereas the input radiation resistance R_{ri} accounts for power that is radiated by the antenna. Using

$$P_{\text{rad}} = \frac{1}{2} I_{\text{in}}^2 R_{ri}$$

we can write the input radiation resistance R_{ri} in terms of the terminal current I_{in} and the power P_{rad} radiated by the antenna:

$$R_{ri} = \frac{2P_{\text{rad}}}{I_{\text{in}}^2} \qquad [\Omega]. \tag{14.61}$$

Also, the radiation efficiency η_r can be expressed in terms of R_{ri} and R_L as

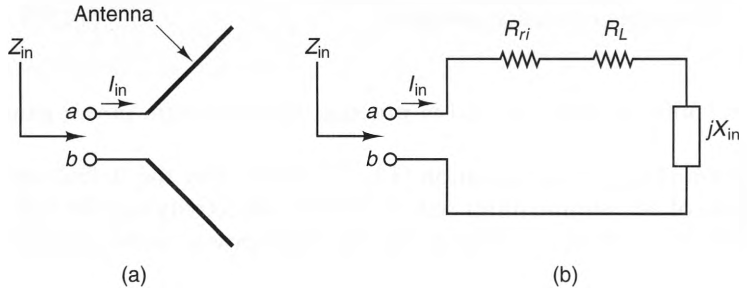


Figure 14-9 The input impedance of an antenna: a) Input terminals of an antenna. b) The equivalent circuit as seen looking into the input terminals.

$$\eta_r = \frac{P_{\text{rad}}}{P_{\text{in}}} = \frac{R_{ri}}{R_{ri} + R_L} \quad \text{(Dimensionless)}. \tag{14.62}$$

Even though the input radiation resistance is a parameter that is measured in the near zone of an antenna, Equation (14.61) shows that it can be calculated by knowing only the far-zone radiation pattern. On the other hand, it is not so easy to calculate the input reactance X_{in}. This is because X_{in} represents energy stored in the fields close to the antenna. Hence, X_{in} can only be calculated by first calculating the fields close to the antenna, which are typically much harder to calculate than the far-zone fields.

Another common resistance parameter used to describe antennas is the radiation resistance

$$R_r = \frac{2P_{\text{rad}}}{I_{\text{max}}^2} \quad [\Omega], \tag{14.63}$$

where I_{max} is the maximum current on the antenna. This resistance is related to the input radiation resistance R_{ri}, but they are equal only when the maximum current on the antenna appears at the input terminals. In general, R_{ri} and R_r are related by

$$R_{ri} = \left[\frac{I_{\text{max}}}{I_{\text{in}}}\right]^2 R_r. \tag{14.64}$$

14-5 Simple Antennas

Simple antennas consist of a single radiating element or structure. The major classes of simple antennas include straight wire antennas, loop antennas, aperture antennas, and reflector antennas. These antennas are often used individually, but they can also be used as the basic building blocks of larger antenna structures called arrays. In this section, we will outline the basic characteristics of the major classes of simple antennas. Later, we will discuss how these elements can be arranged to form arrays.

14-5-1 DIPOLES

We have already discussed one type of dipole, the infinitesimal (or short) dipole. This antenna is simple to model, but is not very practical for a number of reasons, the most important being that its input impedance is undesirable—a small resistance in series with a very large capacitive reactance. Because of this, it is very difficult to design efficient matching networks that allow short dipoles to be driven by amplifying circuits. However, when the dipole length is on the order of a half wavelength or more, its input impedance becomes much more attractive. This, along with its mechanical simplicity, makes the finite-length dipole attractive for a number of applications.

In order to determine the fields generated by finite-length dipoles, we must first determine what kinds of current distributions are excited on these wires by a voltage feed. Calculating these currents directly from Maxwell's equations is a difficult problem, since both the currents and the fields must be found simultaneously. A simpler, alternative procedure is to consider the wire configurations shown in Figure 14-10.

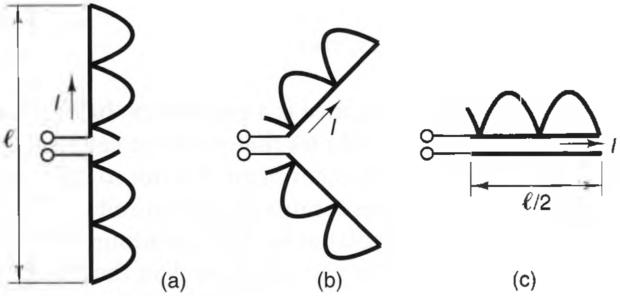


Figure 14-10 Sequence of wire configurations for determining the current distributions on dipoles: a) A straight dipole. b) A slightly bent dipole. c) An open-circuited transmission line.

In Figure 14-10a, a dipole antenna of length ℓ is fed at its center by a voltage source. Figures 14-10b and c show the same geometry, but with the dipole arms bent progressively towards each other until they are parallel. At the end of this progression, the wires form a uniform, open-circuited transmission line. Even though the properties of transmission lines and antennas are quite different, the current on the wires remains remarkably constant throughout the progression. This means that we can use the transmission-line current as an approximation of the antenna current. Using standard transmission-line analysis, we find that the current on the upper wire of Figure 14-10c is of the form

$$I(z) = I_{\mathrm{m}} \sin\left[k\left(\frac{\ell}{2} - z\right)\right],$$

where $k = \omega\sqrt{\mu_0\epsilon_0}$ is the phase constant of a transmission line with an air dielectric and $z = 0$ occurs at the terminals. Since the currents on the upper and lower wires have even symmetry, the preceding expression for $I(z)$ can be used over the entire length of the dipole in Figure 14-10a by replacing z with $|z|$, yielding

$$I(z) = I_m \sin\left[k\left(\frac{\ell}{2} - |z|\right)\right]. \tag{14.65}$$

This approximation for $I(z)$ is most accurate when the dipole is fed at its center, but can also be used for off-center-fed dipoles.

Figures 14-11a–c show the current distributions excited on dipoles of three different lengths.

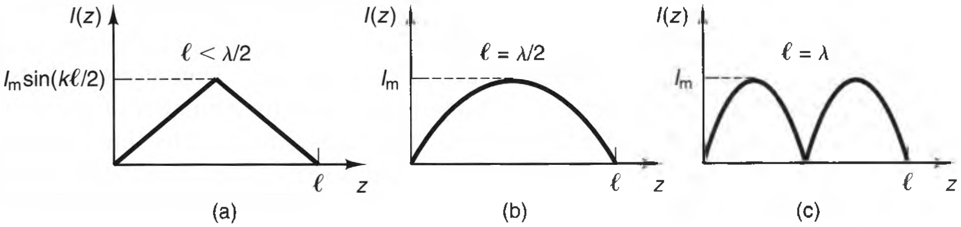


Figure 14-11 Current distributions on dipole antennas: a) $\ell \ll \lambda/2$. b) $\ell = \lambda/2$. c) $\ell = \lambda$.

As can be seen, when $\ell \ll \lambda/2$, the shape of the current is roughly triangular. For longer lengths, $I(z)$ takes on a sinusoidal shape, with more lobes on longer length wires. For all lengths, the current I_o at the center of the wire is given by $I_o = I_m \sin(k\,\ell/2)$. However, when ℓ is an odd multiple of half wavelengths, $I_o = I_m$.

Having found an expression for the dipole current $I(z)$, we can calculate the total radiated field by summing the contributions from each infinitesimal segment along the dipole. Referring to Figure 14-12 and using the field of an infinitesimal current segment (Equation (14.41)), we can write the far-zone contribution from the segment at $z = z'$ as

$$\mathbf{dE} = \frac{jkI(z')\,dz'}{4\pi}\,\eta\sin\theta'\,\frac{e^{-jkR}}{R}\,\hat{\mathbf{a}}_\theta, \tag{14.66}$$

where R is the length of a line from the segment to the observer and θ' is the angle that this line makes with the z-axis. When $r \gg \ell/2$, all lines drawn from the wire to the observer are nearly parallel, so $\theta \approx \theta'$ and

$$R \approx r - z'\cos\theta, \tag{14.67}$$

where r and θ are the position coordinates of the observer. Substituting Equation (14.67) into Equation (14.66) yields

$$\mathbf{dE} = \frac{jkI(z')e^{-jkr}\,dz'}{4\pi r}\,\eta\sin\theta\,\hat{\mathbf{a}}_\theta e^{jkz'\cos\theta}. \tag{14.68}$$

Replacing $I(z')$ with $I_m \sin\{k[(\ell/2) - |z|]\}$ and integrating all the contributions to the field along the wire, we obtain

$$\mathbf{E} = \int_{\text{BOTTOM}}^{\text{TOP}} \mathbf{dE} = \frac{jkI_m e^{-jkr}}{4\pi r}\,\eta\sin\theta\,\hat{\mathbf{a}}_\theta \int_{-\ell/2}^{\ell/2} \sin\left[k\left(\frac{\ell}{2} - |z|\right)\right] e^{jkz'\cos\theta}\,dz'.$$

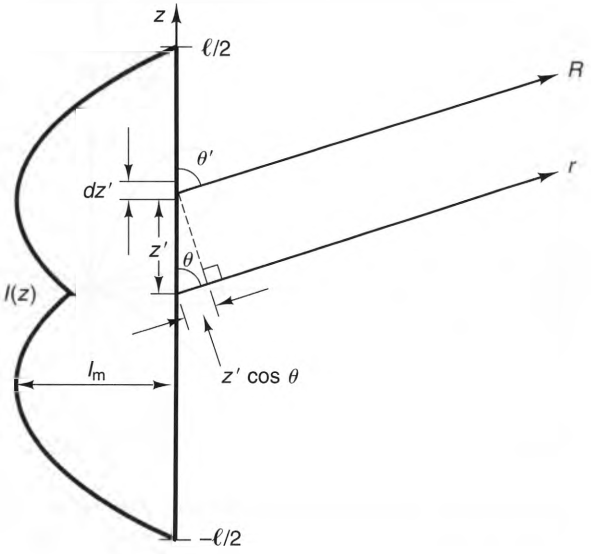


Figure 14-12 Geometry for determining the far-zone radiated field of a finite-length dipole.

Using Euler's identity to expand the exponential term and evaluating the resulting integral, we get

$$\mathbf{E} = \frac{j\eta I_m e^{-jkr}}{2\pi r} F(\theta)\,\hat{\mathbf{a}}_\theta \qquad [\text{V/m}], \tag{14.69}$$

where $F(\theta)$ is called the ***pattern function*** and is given by

$$F(\theta) = \frac{\cos\left(\dfrac{k\ell}{2}\cos\theta\right) - \cos\dfrac{k\ell}{2}}{\sin\theta}. \tag{14.70}$$

Since in the far zone **E** and **H** behave locally as plane waves, we also have $H_\phi = E_\theta/\eta$. Hence,

$$\mathbf{H} = \frac{jI_m e^{-jkr}}{2\pi r} F(\theta)\,\hat{\mathbf{a}}_\phi. \tag{14.71}$$

Figure 14-13 shows the pattern functions for four different-length wire antennas, each normalized to a maximum amplitude of unity. As can be seen, dipole pattern functions become more complex as their lengths increase. The reason for this is that the phase difference between the fields emanating from the endpoints of the dipole becomes more pronounced as the dipole length increases. At $\theta = 90°$, all delays are

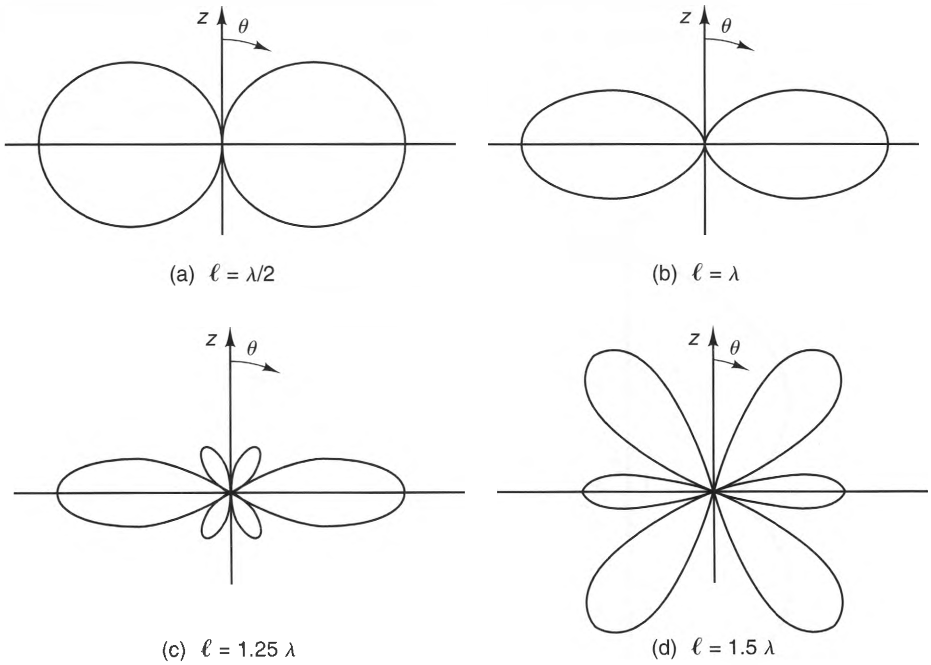


Figure 14-13 Pattern functions for dipole antennas oriented along the z-axis: a) $\ell = \lambda/2$. b) $\ell = \lambda$. c) $\ell = 1.25\lambda$. d) $\ell = 1.5\lambda$.

equal, since the current lies along the z-axis. But as θ approaches 0° or 180°, contributions from the different points on the dipole arrive with significant phase differences.

We can find the power radiated by a wire antenna by integrating the radiation intensity over a large sphere that surrounds the wire. Substituting Equations (14.69) and (14.70) into Equation (14.51), we see that the radiation intensity is

$$U(\theta, \phi) = \frac{\eta I_m^2}{8\pi^2} F^2(\theta).$$

Substituting $U(\theta, \phi)$ into Equation (14.55), we find that the directivity D_o is given by

$$D_o = \frac{U_{\max}}{U_{\text{ave}}} = \frac{F^2(\theta)\,|_{\max}}{\dfrac{1}{4\pi}\displaystyle\int_0^{2\pi}\int_0^{\pi} F^2(\theta)\sin\theta\, d\theta\, d\phi} = \frac{F^2(\theta)\,|_{\max}}{\dfrac{1}{2}\displaystyle\int_0^{\pi} F^2(\theta)\sin\theta\, d\theta}.$$

The integral in this expression cannot be evaluated in closed form, but it can be manipulated into a form that contains well-tabulated functions,[4] or it can be evaluated numerically. Figure 14-14 show D_o as a function of ℓ/λ. As can be seen, $D_o \approx 1.5$ when $\ell \ll \lambda/2$. For $\ell = \lambda/2$, the directivity is

$$D_o = 1.64 = 2.15\ [\text{dB}] \qquad \text{(Half wave dipole)}. \tag{14.72}$$

Remembering that the current at the dipole center is $I_{\text{in}} = I_m \sin(k\,\ell/2)$, we can determine the input radiation resistance of a lossless, center-fed dipole using Equation (14.61):

$$R_{\text{in}} = R_{ri} = \frac{2P_{\text{rad}}}{I_{\text{in}}^2} = \frac{\eta}{2\pi\sin^2(k\,\ell/2)}\int_0^{\pi} F^2(\theta)\sin\theta\, d\theta. \tag{14.73}$$

The dark curve in Figure 14-15 is a plot of R_{in} as a function of dipole length. The values in this curve were obtained using Equation (14.73), except for the lengths in the range 0.8λ to 1.0λ, where Equation (14.73) predicts unreasonably large values. These incorrect values occur because the approximate current distribution goes to zero for this range of dipole lengths, whereas the actual current distribution does not. To

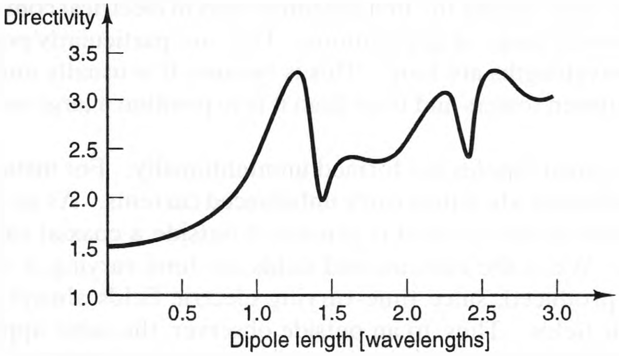


Figure 14-14 Dipole directivity vs. length in wavelengths.

[4] See C. Balanis, *Antenna Theory* (Harper & Row publishers). New York, 1982.

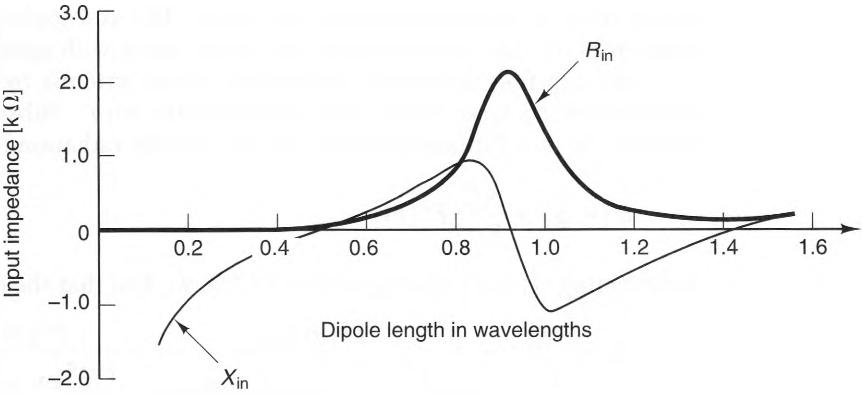


Figure 14-15 Input resistance and reactance of a center-fed dipole vs. length.

account for that behavior, the values of R_{in} plotted in the aforesaid range were obtained using an advanced numerical technique that calculates the exact current distribution on the wire. The figure also shows the values of the input reactance X_{in}, which were obtained using the same numerical technique throughout the entire range of dipole lengths.

As can be seen from Figure 14-15, the input impedance of a center-fed dipole is purely resistive for certain lengths, called ***resonant lengths***. The shortest resonant length is $\ell \approx \lambda/2$ ($\ell = 0.475\lambda$, to be more exact), for which we obtain

$$Z_{\text{in}} \approx 73 + j\,0 \quad [\Omega] \qquad \text{(Half wave dipole)}. \tag{14.74}$$

There are other resonant lengths, but the half-wave dipole is by far the most popular choice, since it has the shortest length, the simplest radiation pattern, and a relatively large bandwidth over which $\text{Im}(Z_{\text{in}})$ is small.

Dipole antennas were among the first antennas used in electrical communications and are still used in a wide range of applications. They are particularly popular at RF frequencies, where wavelengths are long. This is because it is usually much easier to mount a long wire between towers and trees than it is to position a large metal surface, such as a reflector.

There are times when dipoles are formed unintentionally. For instance, coaxial cables act as dipole radiators when they carry unbalanced currents. As we discussed in Section 9-3-7, a nonzero magnetic field is generated outside a coaxial cable that has unbalanced currents. When the currents and fields are time varying, a time-varying electric field is also produced, since time-varying electric fields always accompany time-varying magnetic fields. Thus, to an outside observer, the cable appears to be a thick wire dipole carrying the common-mode current (i.e., the sum of the inner and outer currents).

Radiation from cables is a significant problem in digital equipment, since it causes interference with communication services (such as radio and television) and can also interfere with the operation of the digital systems themselves. A common way of suppressing this radiation is to wrap an offending cable around a ferrite core, as shown in Figure 9-26b, which reduces the common-mode current.

Two types of wire antennas that are closely related to dipoles are monopoles and folded dipoles, which we will discuss in the paragraphs that follow.

Monopole Antennas. Figure 14-16a shows a monopole antenna, consisting of a straight wire, mounted perpendicular to a conducting ground plane and fed with a voltage V_{in} at the base. At first look, this antenna may appear to have little relation to the dipole, since currents flow on both the wire and the ground plane. However, we can determine the fields of a monopole by using the equivalent geometry shown in Figure 14-16b. Here, the ground plane is replaced by an image wire that carries a current which is a mirror image of the monopole current. To ensure that the current on the equivalent dipole has the required even symmetry, a voltage V_{in} is applied symmetrically, just below the $z = 0$ plane. The fields above the $z = 0$ plane are unchanged, since the $E_{\text{tan}} = 0$ boundary condition once imposed by the ground plane is now maintained by the image currents. As a result, the radiation pattern of a monopole is identical to a dipole whose length is exactly twice that of the monopole.

We can find the input impedance of a monopole using the equivalent dipole shown in Figure 14-16b. Starting with $Z_{\text{in}} = V_{\text{in}}/I_{\text{in}}$, we can rewrite this as

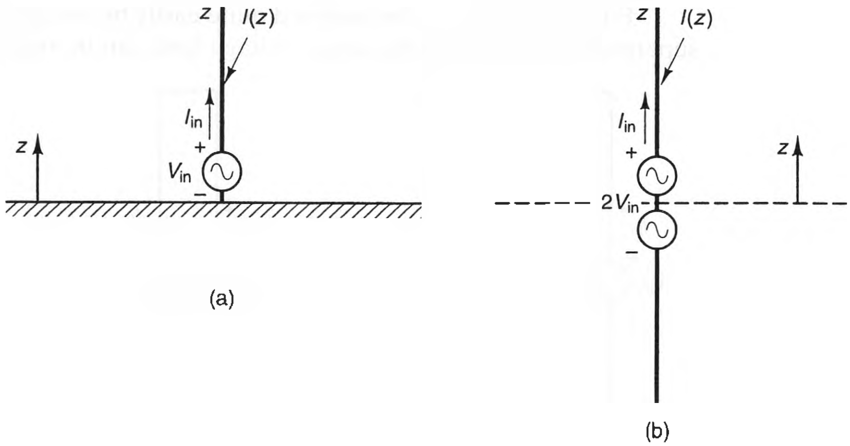


Figure 14-16 A monopole over a conducting plane and its equivalent dipole. a) The monopole and ground plane. b) The equivalent dipole in free space.

$$Z_{\text{monopole}} = \frac{V_{\text{in}}}{I_{\text{in}}} = \frac{1}{2}\frac{2V_{\text{in}}}{I_{\text{in}}}.$$

Since the two series voltage sources in the equivalent dipole constitute a single voltage of value $2V_{\text{in}}$, it follows that $2V_{\text{in}}/I_{\text{in}}$ is the impedance of the equivalent dipole. Thus, the monopole input impedance is given by

$$Z_{\text{monopole}} = \frac{1}{2} Z_{\text{dipole}}. \tag{14.75}$$

For a monopole of length $\ell \approx \lambda/4$, we have

$$Z_{\text{monopole}} = \frac{1}{2} \times 73 = 36.5 \qquad [\Omega] \qquad (\lambda/4 \text{ monopole}). \tag{14.76}$$

Because monopole antennas require only half the wire length of dipole antennas, they are often used in low-frequency systems, where wavelengths are long. They also have radiation patterns that are ideal for ground-to-ground communication systems, since the direction of maximum radiation is parallel to the ground. Also, unlike horizontal dipoles, which must be mounted several wavelengths above the ground to be effective (due to reflections off the ground that tend to cancel the fields), monopoles work best when mounted directly above the ground.

Folded Dipoles. Another common variation of the dipole wire antenna is the folded dipole, shown in Figure 14-17a. As can be seen, this antenna consists of two $\lambda/2$ dipoles connected in parallel, with the feed point at the center of one of the dipoles. Although at first glance it may appear that a folded dipole antenna will act more as a loop antenna than as a dipole, the small area enclosed by the wires prevents it from acting in the loop mode.

Folded dipoles can be analyzed quite easily by recognizing that, according to the superposition principle, the single voltage feed can be represented as the sum of the

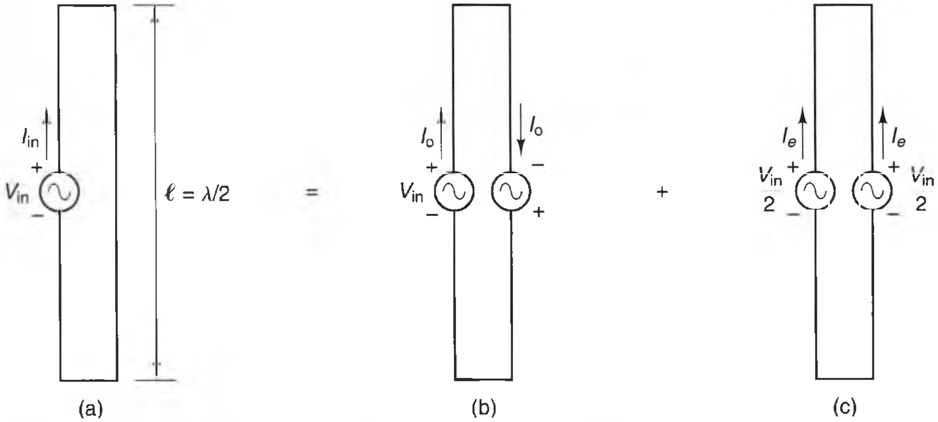


Figure 14-17 A folded dipole and its odd and even components. a) A folded dipole, fed at a single port. b) The odd-mode excitation. c) The even-mode excitation.

odd and even voltage feed configurations shown in Figures 14-17b & c, respectively. Thus, the folded dipole current distribution can be expressed as the sum of the even- and odd-mode current distributions. The input current of the folded dipole can now be expressed as

$$I_{\text{in}} = I_e + I_o, \tag{14.77}$$

where I_e and I_o are the currents at the original feed point due to the even- and odd-mode voltage feeds, respectively. Also, the radiated fields of the folded dipole are the sums of the fields radiated by the odd- and even-mode configurations.

The odd-mode configuration of Figure 14-17b is a transmission-line mode, since the sources drive currents that are oppositely directed on the upper and lower wires. Transmission-line modes do not radiate, so this part of the total current distribution does not contribute to the radiation pattern of the folded dipole. Also, the odd-mode current distribution has no effect on the input impedance of the folded dipole. This is because the impedance "seen" by both voltage generators in the odd-source configuration is infinite, since they are located $\lambda/4$ away from short circuits. Hence, $I_o = 0$ when $\ell \approx \lambda/2$.

For the even-mode configuration, shown in Figure 14-17c, the currents at the wire ends are zero, which means that the current distributions on both wires are essentially the same as the current distribution on a single dipole. Since the wires are closely spaced, their radiated fields add in the far zone, producing a standard dipole pattern. As a result, the even-mode excitation behaves as a thick, center-fed dipole, with voltage feed $V_{\text{in}}/2$ and terminal current $2I_e$. The input impedance of a thick dipole is roughly the same as for a thin one, so the ratio of the terminal voltage $V_{\text{in}}/2$ to the terminal current $2I_e$ must be 73 [Ω], which gives us

$$\frac{V_{\text{in}}/2}{2I_{\text{e}}} = 73 \qquad [\Omega].$$

Multiplying both sides by 4 and using $I_{\text{in}} = I_e$, we obtain the ratio of V_{in} to I_{in}, which is the input impedance of the folded dipole;

$$Z_{\text{F dipole}} = \frac{V_{\text{in}}}{I_{\text{in}}} = 4 \times 73 = 292 \qquad [\Omega] \qquad (\ell \approx \lambda/2). \tag{14.78}$$

Thus, a folded dipole has the same pattern function as a standard dipole and an input impedance that is four times larger than a standard dipole.

Because of their relatively high input impedances, folded dipoles are often used with FM radio and television receivers. They provide a particularly good match to the 300 [Ω] twin-lead transmission lines commonly used in these systems.

14-5-2 LOOP ANTENNAS

Loop antennas usually consist of one or more circular loops of wire. Figure 14-18 shows a single-turn loop antenna, fed by a voltage source. Unlike wire antennas, which are analytically simple for all lengths, only small-circumference loops can be analyzed easily.

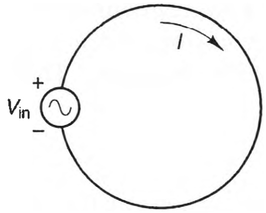


Figure 14-18 A small loop antenna with a uniform current.

We can determine the radiating characteristics of small loop antennas by estimating the current distribution on the loop, calculating the vector potential **A** generated by this current, and then calculating the radiated E- and H-fields.

From a lumped-circuit point of view, the short loop shown in Figure 14-18 is essentially a short circuit driven by a voltage source. Thus, it is reasonable to assume that the current is uniform along the loop. This is verified by experimental measurements. Referring to Figure 14-19, we see that the vector potential generated by a uniform loop of current is given by

$$\mathbf{A} = \frac{\mu I_o}{4\pi}\oint_{C'} \frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}\,\mathbf{d\ell}'. \tag{14.79}$$

In this expression, **r** and **r**′ represent the positions of the observer and the dummy integration points along the loop, respectively, and **dℓ**′ is the differential displacement vector along the loop. Rather than integrating directly, we will rewrite the expression in terms of an integral that we encountered in Chapter 7, while discussing the magnetic dipole. To accomplish this, we first write the exponential term in the integrand as

$$e^{-jk|\mathbf{r}-\mathbf{r}'|} = e^{-jkr}e^{-jk(|\mathbf{r}-\mathbf{r}'|-r)}.$$

As long as $a \ll r$, $|\mathbf{r}-\mathbf{r}'|$ and r are nearly equal, so the second exponential can be approximated by the first two terms of its Maclaurin series:

$$e^{-jk[|\mathbf{r}-\mathbf{r}'|-r]} \approx 1 - jk[|\mathbf{r}-\mathbf{r}'| - r].$$

Substituting into Equation (14.79), we obtain

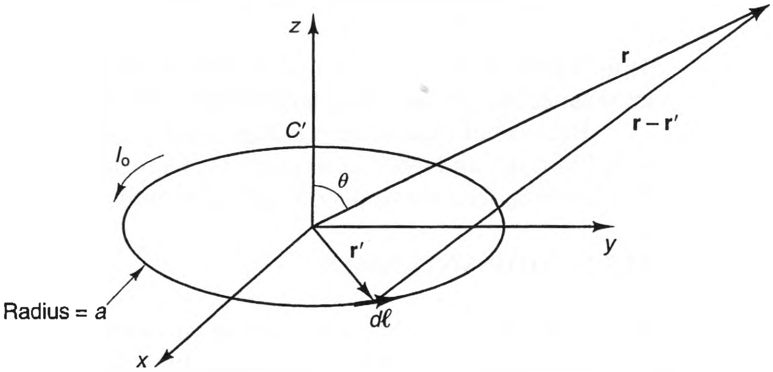


Figure 14-19 Geometry for determining the fields radiated by a small, constant-current loop.

$$\mathbf{A} = \frac{\mu I_o}{4\pi} e^{-jkr}\left[(1 + jkr)\oint_{C'} \frac{\mathbf{d\ell}'}{|\mathbf{r} - \mathbf{r}'|} - jk\oint_{C'} \mathbf{d\ell}'\right]. \tag{14.80}$$

The second integral in this expression is zero (since the path of integration is closed), and the first integral is the same as that found in Equation (7.55), namely,

$$\oint_C \frac{\mathbf{d\ell}'}{|\mathbf{r} - \mathbf{r}|'} \approx \frac{S}{r^2}\sin\theta\,\hat{\mathbf{a}}_\phi \qquad r \gg a, \tag{14.81}$$

where S is the area of the loop. Substituting Equation (14.81) into Equation (14.80), we obtain the final expression for **A**:

$$\mathbf{A} = \frac{\mu S I_o e^{-jkr}}{4\pi r^2}[1 + jkr]\sin\theta\,\hat{\mathbf{a}}_\phi \qquad r \gg a \quad \text{and} \quad a \ll \lambda. \tag{14.82}$$

Having found **A**, it is a simple matter to find **E** and **H**. Using Equation (14.2) and $\mathbf{H} = \mathbf{B}/\mu$, we have

$$\mathbf{H} = \frac{1}{\mu}\nabla \times \mathbf{A} = \frac{1}{\mu}\nabla \times \left[\frac{\mu S I_o e^{-jkr}}{4\pi r^2}[1 + jkr]\sin\theta\right].$$

Evaluating the appropriate derivatives, we find that

$$\mathbf{H} = H_r\hat{\mathbf{a}}_r + H_\theta\hat{\mathbf{a}}_\theta, \tag{14.83}$$

where

$$H_r = \frac{jkSI_o}{2\pi}\cos\theta\left[\frac{1}{r^2} - \frac{j}{kr^3}\right]e^{-jkr} \tag{14.84}$$

$$H_\theta = \frac{jkSI_o}{4\pi}\sin\theta\left[\frac{jk}{r} + \frac{1}{r^2} - \frac{j}{kr^3}\right]e^{-jkr}. \tag{14.85}$$

Finally, we can use Maxwell's curl-H equation to find **E**. Noting that $\mathbf{J} = 0$ except on the loop itself, we have

$$\mathbf{E} = \frac{\nabla \times \mathbf{H}}{j\omega\epsilon}.$$

Substituting Equations (14.84) and (14.85) into this expression and evaluating the curl operation, we obtain

$$\mathbf{E} = E_\phi\hat{\mathbf{a}}_\phi = -\frac{j\eta kSI_o}{4\pi}\sin\theta\left[\frac{jk}{r} + \frac{1}{r^2}\right]e^{-jkr}\,\hat{\mathbf{a}}_\phi. \tag{14.86}$$

Comparing the fields generated by a small loop with those of a small dipole (Equations (14.33), (14.35), and (14.36)), we see that they are the same, except that the roles of **E** and **H** are reversed. Sources that have this kind of reversal in the roles of **E** and **H** are called ***duals*** of each other. The pattern function for a small loop is shown in Figure 14-20. As can be seen, the pattern is omidirectional, with maximum radiation in the plane perpendicular to the axis of the loop.

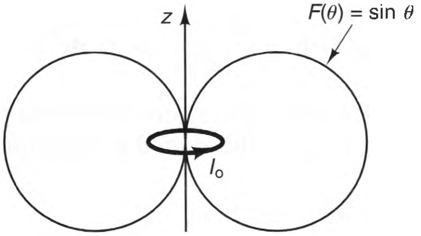


Figure 14-20 The pattern function for a small, constant-current loop antenna.

Because the radiation patterns of small loops and dipoles are duals, the radiation resistance for a small loop antenna can be derived using a similar sequence of steps as was used for the small dipole, resulting in the expression

$$R_r = \frac{2P_{\text{rad}}}{I_o^2} = \eta\frac{2\pi}{3}\left(\frac{kS}{\lambda}\right)^2 = 20\pi^2\left(\frac{C}{\lambda}\right)^4 \quad [\Omega] \qquad (C \ll \lambda), \tag{14.87}$$

where $C = 2\pi a$ is the circumference of the loop and $\eta \approx 377$ [Ω] in free space. As with the small dipole, the radiation resistance of a small loop is also small—too small to be of much practical value in most cases. However, unlike the situation with short dipoles, there are two "tricks" that are often used to dramatically increase the radiation resistance of small loops. The first is to use multiple turns. According the superposition principle, the E- and H-fields generated by an N-turn loop are both proportional to N, so the radiation resistance varies as N^2. The second trick is to wrap the multiple turns around a high-permeability core, which increases the strength of both **E** and **H** in the far zone. When both of these tricks are used, the radiation resistance is given by

$$R_r = 20\pi^2\left(\frac{C}{\lambda}\right)^2\left(\frac{\mu_c}{\mu_o}\right)^2 N^2 \quad [\Omega] \qquad (C \ll \lambda), \tag{14.88}$$

where μ_c is the permeability of the core. Multiturn loops with ferrite cores are called ***loop-stick antennas*** and are used in AM broadcast receivers because of their relatively high radiation resistances and small sizes.

Even though we have not yet discussed the receiving properties of antennas, there is one important receiving property of loop antennas that is worth discussing here: Loop antennas are relatively insensitive to near-zone electric sources, such as the sparks created by electrical machinery. This makes these antennas particularly attractive for use as receiving antennas in electrically noisy environments, since they respond well to far-zone electric and magnetic sources, but not near-zone electric sources. This property follows directly from that fact that the near-zone transmitted E-fields of loop antennas are much weaker than the H-fields. Since the transmitting and receiving properties of antennas are related,[5] loop antennas respond well to incident H-fields, but not E-fields. The near-zone fields of short electric currents are predominately electric (see Equations (14.33)–(14.36)), so loop antennas respond much less to these sources than dipoles do.

[5] We will show this later in the chapter.

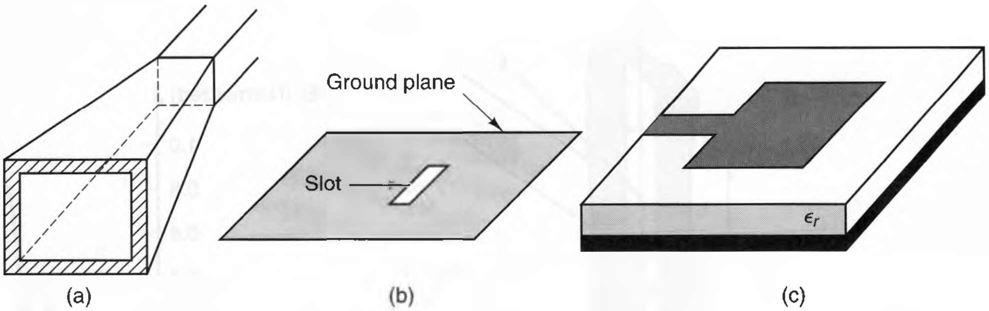


Figure 14-21 Examples of aperture antennas: a) A horn antenna. b) A slot antenna. c) A microstrip patch antenna.

14-5-3 APERTURE ANTENNAS

As the name implies, an aperture antenna is characterized by an aperture, or opening, from which the radiated fields are emitted. Popular types of aperture antennas are horns, slots, and microstrip patches, shown in Figure 14-21.

In a way, aperture antennas are the easiest antennas to understand, since they look like they are capable of launching waves. The operation of aperture antennas is most easily explained using Huygens's principle (Christiaan Huygens, 1629–1695), which, in words, states that each point in an advancing wave front acts as a source of spherical, secondary wavelets that propagate outward. Figure 14-22 demonstrates this by showing a plane wave impinging upon a slit aperture in a screen. Here, we see that the secondary wavelets lead to a spreading of the wave as it emerges from the aperture. This spreading is called ***diffraction*** and occurs whenever a wave is incident upon a sharp discontinuity. Hence, the antenna pattern of an aperture antenna is actually a diffraction pattern.

We can derive an expression for the far-zone pattern of this simple aperture antenna by using Figure 14-23a and a simplified expression of Huygens's principle (which can be derived from Maxwell's equations),[6]

$$E(\mathbf{r}) \approx \frac{jk}{2\pi}\int_S E_a(\mathbf{r}')\frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}ds'. \tag{14.89}$$

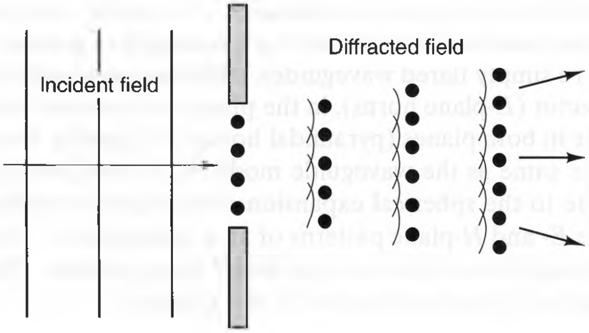


Figure 14-22 Diffraction from a uniformly illuminated aperture.

[6] See Stutzman and Thiele, *Antenna Theory and Design*. New York.

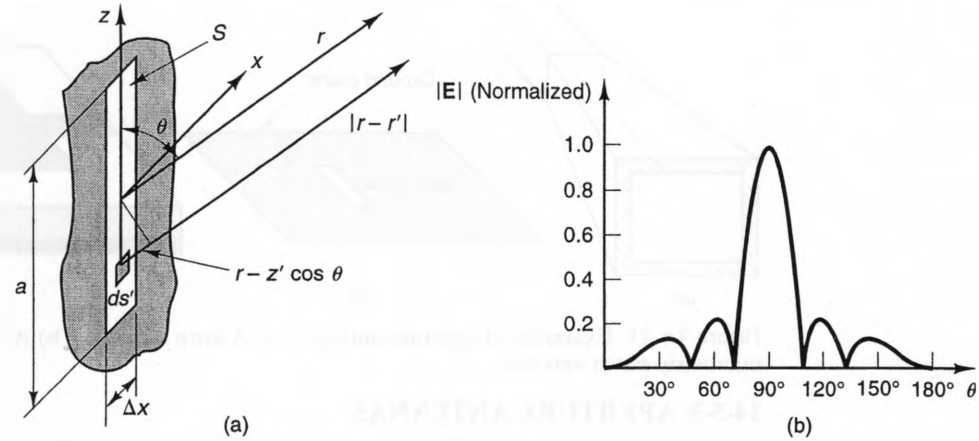


Figure 14-23 Radiation from a narrow aperture: a) The geometry. b) Normalized far-zone E-field vs. observation angle.

In this expression, $E(\mathbf{r})$ is the magnitude of the electric field at the observation point $\mathbf{r} = (r, \theta, \phi)$, S is the aperture surface, $E_a(\mathbf{r}')$ is the magnitude of the electric field in the aperture, and $\mathbf{r}' = (x', y', z')$ is the dummy position vector that sweeps over all the aperture points during the integration. At large distances from the aperture, we can write

$$\left|\mathbf{r} - \mathbf{r}'\right| \approx r - z' \cos \theta.$$

Assuming that the field in the aperture is uniform and that the aperture width Δx is small, Equation (14.89) can be evaluated for the aperture as follows:

$$E(r, \theta, \phi) \approx \frac{jk\Delta x e^{-jkr}}{2\pi r} \int_{-a/2}^{a/2} E_a e^{+jkz' \cos \theta} dz' = jE_a \Delta x \frac{e^{-jkr}}{r} \frac{\sin\left(\pi \dfrac{a}{\lambda} \cos \theta\right)}{\pi \cos \theta}. \tag{14.90}$$

Figure 14-23b shows this radiation pattern as a function of θ for the case where $a = 3\lambda$. As can be seen, the pattern consists of a main lobe at $\theta = 90°$ and a number of side lobes. The size of the main lobe is inversely proportional to the aperture width a. This characteristic is shared by all aperture antennas. To obtain a narrow main lobe, the aperture dimensions must be on the order of a wavelength or greater.

Horn antennas are simply flared waveguides. The flare can be in the plane that contains the E-field vector (*E*-plane horns), in the plane that contains the H-field vector (*H*-plane horns), or in both planes (pyramidal horns). Typically, the aperture distribution is roughly the same as the waveguide mode in the feed, with a phase taper across the aperture due to the spherical expansion of the fields from the waveguide. Figure 14-24 shows the *E*- and *H*-plane patterns of an *E*-plane horn.[7] As can be seen, the *E*-plane pattern is much more directive than the *H*-plane pattern. This is because the aperture is wide in the *E*-plane and narrow in the *H*-plane.

[7] From equations derived in C. Balanis, *Antenna Theory*, pp. 539.

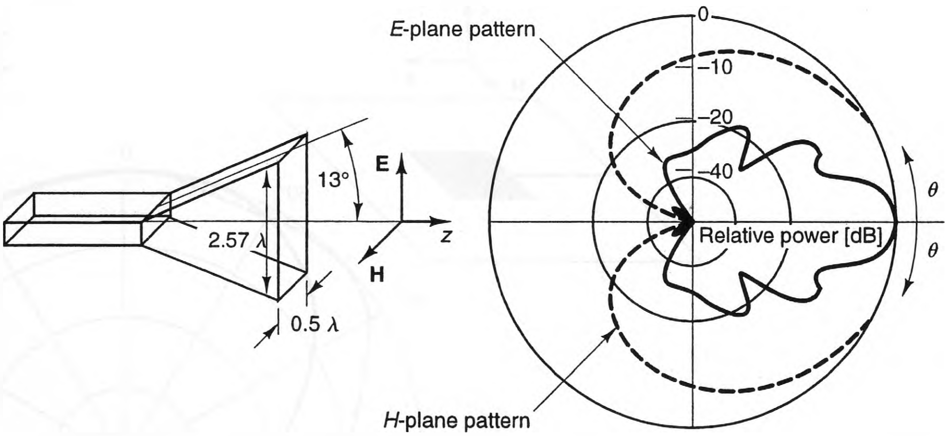


Figure 14-24 Far-zone radiation from a typical horn antenna: a) The geometry. b) The far-zone E- and H-plane patterns in dB.

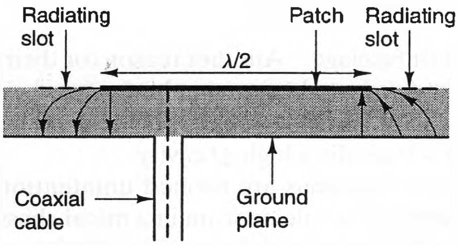


Figure 14-25 Cross-sectional view of a microstrip antenna with a coaxial feed.

Microstrip patch antennas consist of metal patches mounted on a dielectric sheet, with a ground plane underneath. The most popular patch shapes are rectangles and circles. Often, these patches are fed by microstrip transmission lines, as shown in Figure 14-21c. Another common way to feed patch antennas is shown in Figure 14-25. Here, the center conductor of a coaxial cable is brought up from the ground plane and attached to a point on the patch. The apertures of these antennas are not the patches themselves, but rather the regions just beside the patch edges, where the E-field lines fringe. This fringing is shown in the figure; the apertures (or radiating slots) are indicated by the dotted lines.

Real input impedances are obtained when the patch dimensions are roughly $\lambda/2$ on a side. Figure 14-26 shows the radiation pattern of a typical patch antenna. As can be seen, the direction of maximum radiation is perpendicular to the patch. Also, the radiation pattern is quite broad, which makes patch antenna useful for applications in which the position of the receiver (or transmitter) is unknown, such as mobile communication receivers.

Although the gain of single patch antennas is low, their planar construction makes them an ideal choice for the elements in large, planar arrays. As we will see later in this chapter, high-gain antennas can be made out of arrays of relatively low-gain elements. Patch antennas are also attractive in situations where the antenna height must be kept

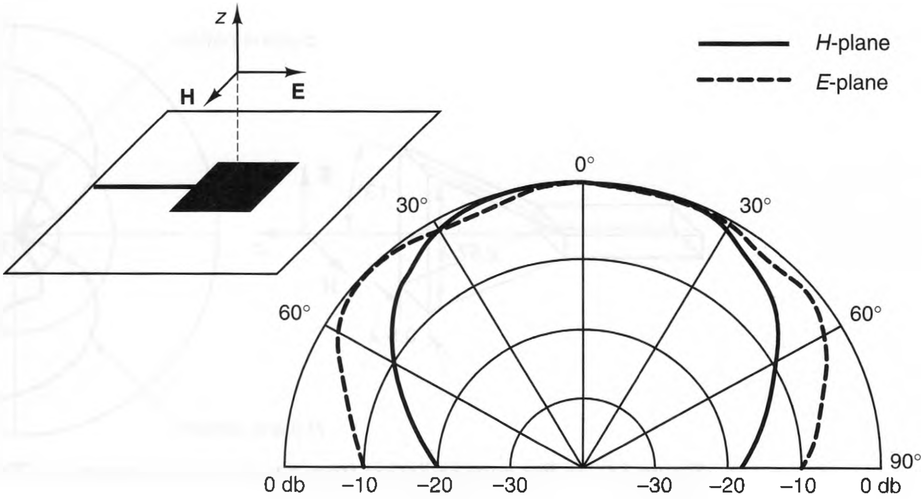


Figure 14-26 E- and H-plane radiation patterns of a rectangular patch antenna, $\lambda/2$ on a side, with a microstrip transmission-line feed.

low, such as on an aircraft fuselage. Another reason for their popularity in large arrays is that they can often be fabricated using standard printed-circuit techniques. On the downside, however, the impedance bandwidths of these antennas are small, since the region beneath the patch is basically a high-Q cavity.

Sometimes aperture antennas are formed unintentionally in electronic equipment. A common example is a hole or seam in a metal chassis. Figure 14-27 shows a seam on a chassis, possibly formed along the edges of a lid or an access panel. Even if the seam is thin, the radiation can be significant when its length is even a fraction of a wavelength. A common method of reducing the radiation is to place conducting tape along the seam. Another common solution is to place flexible metal fingers on each side of the chassis doors that mesh as the door is closed.

14-5-4 REFLECTOR ANTENNAS

Reflector antennas use reflectors (usually metal) to reshape the pattern of a smaller feed antenna into a more directive pattern. Figure 14-28 shows three popular types of reflector antennas: a planar reflector, a corner reflector, and a parabolic reflector. As can be seen from this figure, each of these reflectors redirect rays back towards the feed, but with different angular spreads.

The radiation patterns of planar and corner reflector antennas can usually be obtained using image theory to replace the conducting surfaces with equivalent

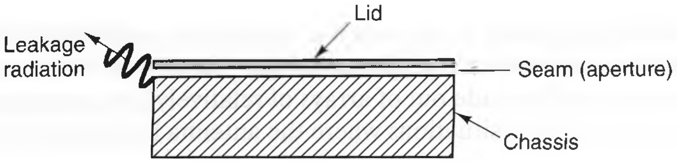


Figure 14-27 Radiation from a seam in a conducting chassis.

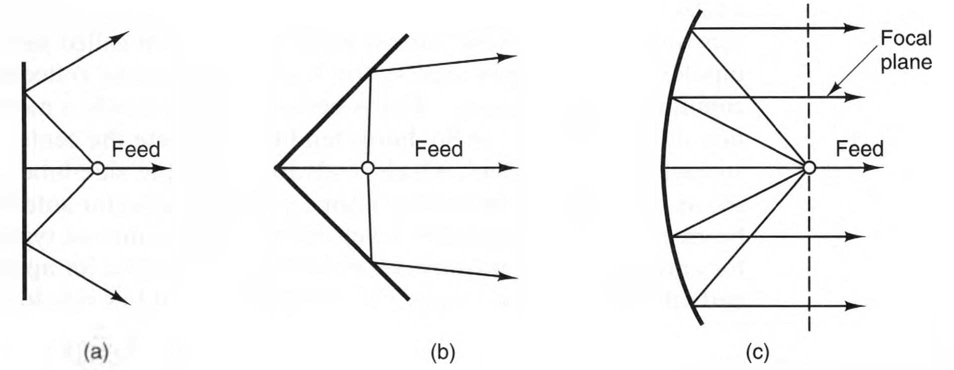


Figure 14-28 Examples of reflector antennas: a) A planar reflector. b) A corner reflector. c) A parabolic reflector.

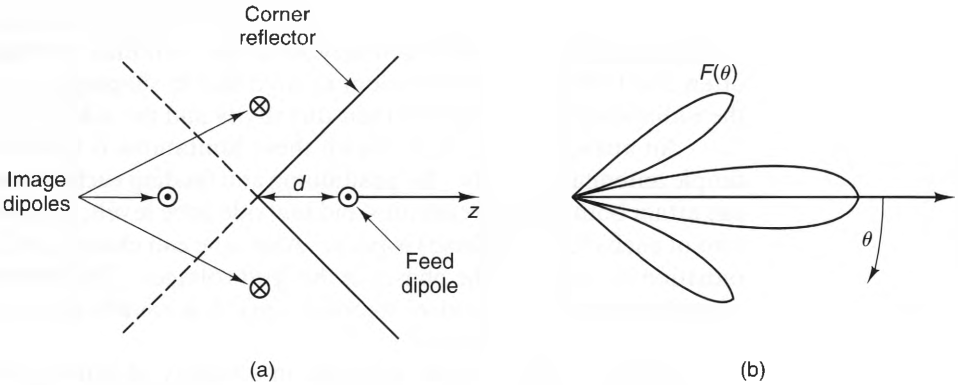


Figure 14-29 A corner reflector with a dipole feed: a) The reflector, the feed dipole, and its images. b) The far-zone radiation pattern.

sources. For instance, Figure 14-29a shows the equivalent sources for a 90° corner reflector, fed by a dipole antenna (directed out of the paper). This geometry can be simplified by replacing the corner reflector with three image dipoles. Figure 14-29a shows the image polarity necessary to maintain the constant potential surfaces of the corner reflector. With the image dipoles in place, the pattern function of this antenna can be obtained simply by summing the fields generated by each of the four dipoles. Using the array techniques discussed in the next section, it can be shown that the pattern function of this antenna in the plane of the paper is

$$F(\theta) = 2\left[\cos\left(2\pi\frac{d}{\lambda}\cos\theta\right) - \cos\left(2\pi\frac{d}{\lambda}\sin\theta\right)\right]. \tag{14.91}$$

Figure 14-29b shows the radiation pattern for a 90° corner reflector when $d = 0.8\lambda$.

Parabolic reflectors are distinctive because they reflect all the rays emanating from a point source parallel to the axis of the reflector. This is an attractive feature, since the reflected fields on the focal plane resemble a plane wave, as shown in Figure

14-28c. This feature makes parabolic reflector antennas highly directive, with pattern functions that have very narrow main beams, often called ***pencil beams***. Although dipoles are sometimes used as the feeds for parabolic reflector antennas, the most common feeds are horns. This is because horns provide a more symmetric illumination of the reflector. Also, horns tend to illuminate the center of the reflector more strongly than the edges, which tends to reduce the side-lobe levels.[8] Figure 14-30 shows the radiation pattern for a large parabolic reflector antenna. The narrow main beam and −30 [dB] side-lobe levels shown in the figure are typical of parabolic reflector antennas. This characteristic makes them attractive for applications such as radar and satellite communications, where high gains and low side-lobe levels are required.

14-6 Antenna Arrays

In our discussion of the basic antennas in the previous section, we saw that these antennas can usually be made more directive by making them large with respect to a wavelength. This often poses problems, however, since such structures can become difficult to fabricate and maneuver when they are large. Another problem is that the antennas often don't offer as much freedom as we'd like in shaping the exact characteristics of the radiation patterns, such as their directivity and the side-lobe characteristics.

An attractive way to deal with these limitations is to construct arrays of small, simple antenna elements. By positioning and feeding each element appropriately, one can attain both large directivities and low side-lobe levels, even when the radiated pattern of each element (alone) is poor. Also, one can change the direction of maximum radiation by changing the phases of the feed voltages. This ***electronic beam steering*** is often better than mechanical steering, since it is usually faster and does not demand heavy positioning equipment.

Antenna arrays can be arranged in a variety of geometrical configurations, but planar arrays are the most popular. Figure 14-31 shows three types of planar arrays: linear, rectangular, and circular.

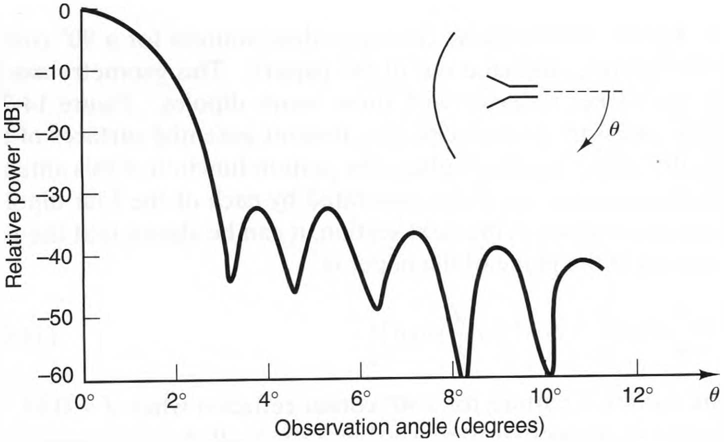


Figure 14-30 Radiation pattern of a typical large parabolic antenna with a horn feed. (Courtesy of M.C. Bailey, NASA Langley Research Center).

[8] This side lobe reduction with feed taper is similar to what occurs in the tapered arrays discussed later in the chapter.

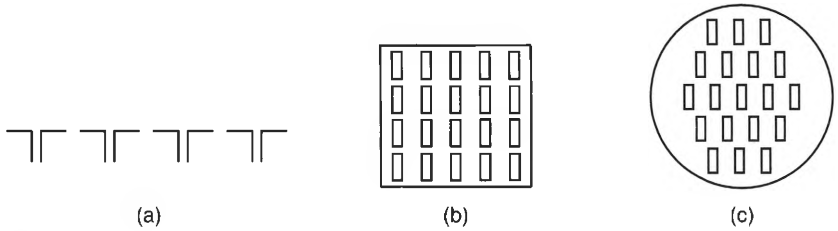


Figure 14-31 Examples of antenna arrays: a) A linear array. b) A rectangular, planar array. c) A circular, planar array.

We will start our discussion of arrays by deriving the characteristics of a simple two-element array; then we will discuss the characteristics of several uniform and nonuniform arrays.

14-6-1 A TWO-ELEMENT ARRAY

Consider the two short dipoles with currents

$$I_1 = |I_o| e^{j\delta/2} \tag{14.92a}$$

$$I_2 = |I_o| e^{-j\delta/2}, \tag{14.92b}$$

shown in Figure 14-32. The dipoles have equal lengths $\Delta\ell$ and lie parallel to the z-axis. They are positioned symmetrically along the y-axis and spaced a distance d apart. The dipole currents have equal magnitudes, but phases that differ by the phase shift δ.

Using the superposition principle, we find that the total electric field **E** radiated by this array is the sum of the fields radiated by each individual dipole. When the observer is far from the origin, only the phase difference between the dipole fields must be accounted for. Using Equation (14.41), the total field can be written as

$$\mathbf{E} = \frac{j\eta k |I_o| \Delta\ell}{4\pi r} \sin\theta\, \hat{\mathbf{a}}_\theta [e^{j\delta/2} e^{-jkr_1} + e^{-j\delta/2} e^{-jkr_2}], \tag{14.93}$$

where r_1 and r_2 are the distances from the observer to the left- and right-hand dipoles, respectively, and r is the distance from the origin to the observer. When r is large, we can write

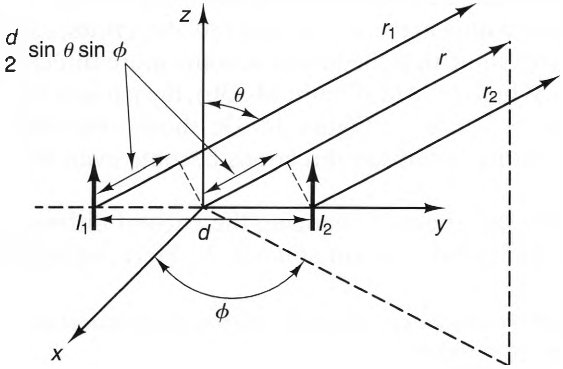


Figure 14-32 Geometry for determining the far-zone radiated fields of a two-element, linear array of Hertzian dipoles.

$$r_1 \approx r + \frac{d}{2}\sin\theta\sin\phi \tag{14.94a}$$

$$r_2 \approx r - \frac{d}{2}\sin\theta\sin\phi . \tag{14.94b}$$

Substituting Equations (14.94)a and b into Equation (14.93), we obtain

$$\begin{aligned}\mathbf{E} &\approx \frac{j\eta k|I_o|\Delta\ell e^{-jkr}}{4\pi r}\sin\theta\,\hat{\mathbf{a}}_\theta[e^{-j\frac{1}{2}(kd\sin\theta\sin\phi-\delta)} + e^{j\frac{1}{2}(kd\sin\theta\sin\phi-\delta)}]\\ &= \frac{j\eta k|I_o|\Delta\ell e^{-jkr}}{4\pi r}\sin\theta\,\hat{\mathbf{a}}_\theta \times 2\cos\left[\frac{1}{2}(kd\sin\theta\sin\phi-\delta)\right].\end{aligned}$$

This expression can be written in the form

$$\mathbf{E} \approx \frac{j\eta k|I_o|\Delta\ell e^{-jkr}}{4\pi r}\sin\theta\,\hat{\mathbf{a}}_\theta\Lambda(\theta,\phi), \tag{14.95}$$

where $\Lambda(\theta, \phi)$ is the ***array factor***, defined by

$$\Lambda(\theta,\phi) = 2\cos\left[\frac{1}{2}(kd\sin\theta\sin\phi-\delta)\right]. \tag{14.96}$$

Equation (14.95) is an example of the ***pattern multiplication theorem***, which states that

> The combined pattern of N identical elements can always be expressed as the element pattern times an array factor $\Lambda(\theta, \phi)$ that accounts for the number of elements, their relative positions, and their feed currents.[9]

If we were to replace the infinitesimal dipoles in the array in Figure 14-32 with, say, horns, we would replace the element pattern in Equation (14.95) with the pattern of a horn, but the array factor would remain the same.

Figure 14-33 shows the radiation patterns in the $\theta = 90°$ plane for three different dipole spacings when the dipole currents are in phase (i.e., $\delta = 0$). As can be seen, all three array factors have maxima in the $\phi = 0°$ and 180° directions, called the ***broadside directions***. These figures show that the arrays become more directive as the element spacing d is increased. When $d = \lambda/2$ (Figure 14-33b), the pattern has deep nulls along the ***end-fire directions***, $\phi = \pm 90°$. Figure 14-33c shows that when d is increased beyond $\lambda/2$, the lobes in the broadside directions become even narrower, but other lobes are formed.

Figure 14-34 shows the effect of changing the current differential phase shift δ when the element spacing is held constant at $d = \lambda/3$. Here, we see that the direction

[9] This theorem assumes that the elements are uncoupled, meaning that a current on one element does not excite appreciable currents on other elements.

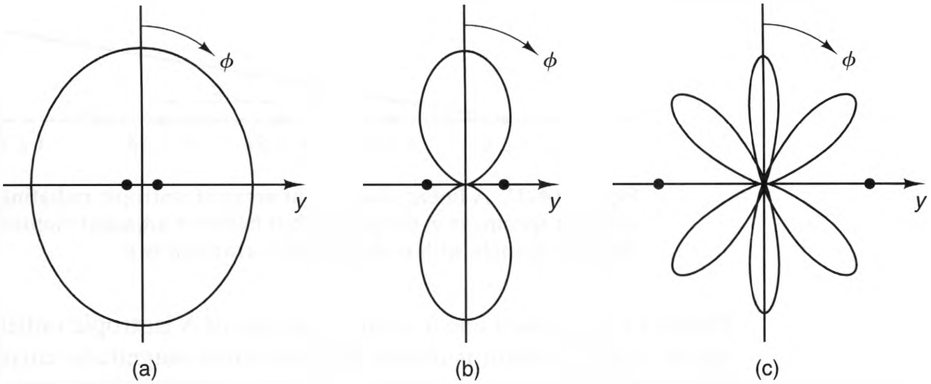


Figure 14-33 Array factor plots in the $\theta = 90°$ plane for three two-element broadside arrays with different spacings between the elements along the y-axis: a) $d = 0.2\lambda$. b) $d = 0.5\lambda$. c) $d = 1.5\lambda$.

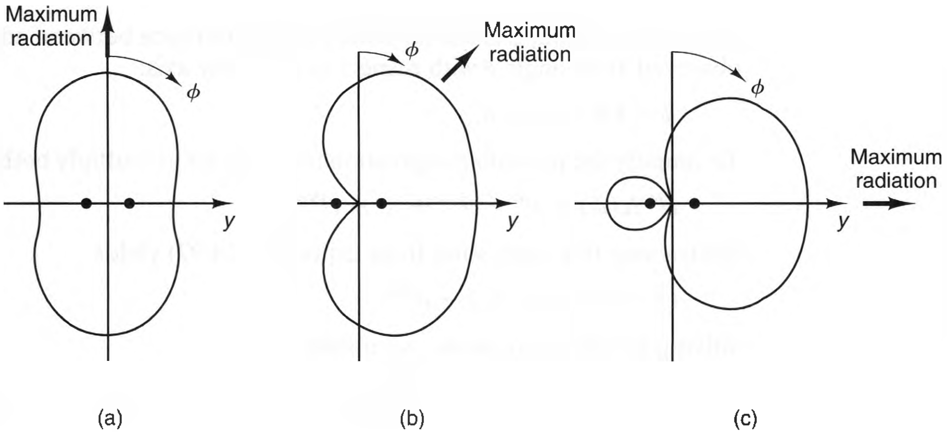


Figure 14-34 Array factor plots in the $\theta = 90°$ plane for three two-element arrays along the y-axis, each with $d = \lambda/3$ and phase shift a) $\delta = 0°$, b) $\delta = 60°$, c) $\delta = 120°$.

of maximum radiation varies with δ. When $\delta = 0$, the lobes are directed broadside to the array axis ($\phi = 0°$ and $\phi = 180°$). As δ is increased, the lobes bend towards the $\phi = 90°$ end-fire direction. When $\delta = -kd$ (120° when $d = \lambda/3$), radiation along the forward end-fire direction is maximized. The shift in the radiation pattern shown in the figure is an example of electronic beam steering.

14-6-2 *N*-ELEMENT LINEAR ARRAYS WITH UNIFORM AMPLITUDE AND SPACING

We can extend our analysis of two-element arrays to include any number of equally spaced elements. Arrays that have their elements along a line are called ***linear arrays***.

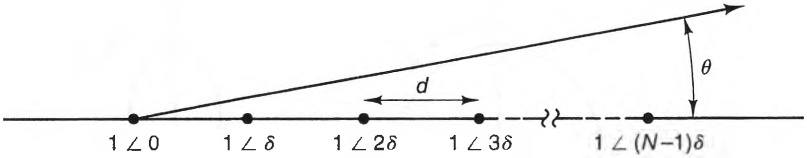


Figure 14-35 A linear, N-element array of isotropic radiating elements. The element spacing is d, the phase shift between adjacent elements is δ, and the observer's angle with respect to the array axis is θ.

Figure 14-35 shows a linear array consisting of N isotropic radiators with element spacing d. Each element is driven with the same-magnitude current, with a progressive phase shift δ between adjacent elements. To an observer far from the array, the field from each element arrives with a phase that is determined by the position of the element, the current phase shift of the feed, and the observer's angular coordinate θ. Adding these contributions, we can write the array factor Λ in the form

$$\Lambda(\psi) = 1 + e^{j\psi} + e^{j2\psi} + \dots + e^{j(N-1)\psi}. \tag{14.97}$$

In this expression, ψ is the far-zone phase difference between adjacent elements when observed at an angle θ with respect to the array axis:

$$\psi = kd\cos\theta + \delta.$$

To simplify the preceding expression for $\Lambda(\psi)$, let us multiply both sides by $e^{j\psi}$, obtaining

$$e^{j\psi}\Lambda(\psi) = e^{j\psi} + e^{j2\psi} + \dots + e^{jN\psi}.$$

Subtracting this expression from Equation (14.97) yields

$$(1 - e^{j\psi})\Lambda(\psi) = 1 - e^{jN\psi}.$$

Solving for the array factor, we obtain

$$\Lambda(\psi) = \frac{1 - e^{jN\psi}}{1 - e^{j\psi}} = \frac{e^{jN\psi/2}[e^{-jN\psi/2} - e^{jN\psi/2}]}{e^{j\psi/2}[e^{-j\psi/2} - e^{j\psi/2}]} = e^{j[\psi(N-1)/2]}\left[\frac{\sin(N\psi/2)}{\sin(\psi/2)}\right].$$

Finally, since we are usually concerned only with the magnitude of the far-zone radiation pattern, we can drop the exponential phase term, yielding

$$\Lambda(\psi) = \left|\frac{\sin(N\psi/2)}{\sin(\psi/2)}\right|, \tag{14.98}$$

where

$$\psi = kd\cos\theta + \delta. \tag{14.99}$$

Notice that $\Lambda(\psi)$ attains a maximum value of $\Lambda_{\max} = N$ when $\psi = 0$.

The array factor $\Lambda(\psi)$ of a uniform, linear array is a function of the number of elements, their spacing, and the phase difference between each element. Even when the number of elements is fixed, there is a great variety of different array factors that can be obtained, depending upon the element spacing d and differential phase shift δ that are chosen. Two common special cases of linear arrays are broadside arrays and ordinary end-fire arrays, which we will discuss next.

Broadside Arrays. Arrays that generate their maximum radiation perpendicular to the array axis are called ***broadside arrays***. Broadside arrays are popular for broadcast applications, since their array factors are omnidirectional. To be most effective, broadside arrays should be constructed using radiating elements that also have omnidirectional patterns.

The array factor $\Lambda(\psi)$ for any linear array is maximized when $\psi = kd\cos\theta + \delta = 0$. For a broadside array, this maximum occurs when $\theta = \pm 90°$, which means that the element-to-element phase shift δ is zero:

$$\delta = 0 \qquad \text{(Broadside arrays)}. \tag{14.100}$$

We could also determine this criterion by noticing that all of the elements are equidistant from a far-zone observer at $\theta = \pm 90°$, which means that the fields of each element will add constructively when the current feeds all have the same phase.

Two things control the directivity of a broadside array: the number of elements N and the distance d between the elements. To illustrate the effect of N, Figure 14-36 shows the normalized array factors when $d = 0.5\lambda$ for $N = 3$ and $N = 5$, respectively. As can be seen, the main lobe becomes narrower as N increases, but at the expense of more side lobes. For a fixed number of elements, increasing d has a similar effect. This latter observation can be seen from Figure 14-33 for the case $N = 2$. Here, we see that when d is large, the peak radiation intensities of some side lobes will be as great as that of the main lobe. These strong side lobes, called ***grating lobes***, are usually undesirable. In practice, the element spacing is typically chosen such that $d \leqslant \lambda/2$ for broadside arrays to avoid grating lobes.

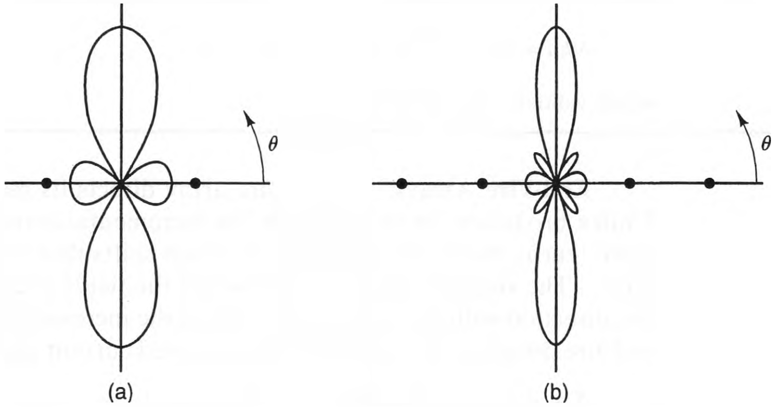


Figure 14-36 Array factor plots for two broadside arrays with the same element spacing, $d = 0.5\lambda$: a) $N = 3$. b) $N = 5$.

We can obtain an expression for the directivity of a broadside array by substituting the array factor $\Lambda(\psi)$ into Equation (14.55), yielding

$$D_{\text{o}} = \frac{4\pi|\Lambda_{\max}|^2}{\oint_S \Lambda(\theta,\phi)|^2 d\Omega} = \frac{4\pi N^2}{\int_0^{2\pi}\int_0^{\pi}\left[\dfrac{\sin\left(\dfrac{N}{2}kd\cos\theta\right)}{\sin\left(\dfrac{1}{2}kd\cos\theta\right)}\right]^2 \sin\theta\, d\theta\, d\phi}.$$

When $Nkd > 6$, the following approximation is accurate to at least 10%:

$$D_{\text{o}} \approx 2N\left(\frac{d}{\lambda}\right) = 2\left(1+\frac{L}{d}\right)\left(\frac{d}{\lambda}\right), \tag{14.101}$$

Here, $L = (N-1)d$ is the total length of the array. This expression clearly shows that the directivity of a broadside array increases as its length L increases.

Example 14-6

A 10-element broadside array of isotropic radiators is to have a directivity of $D_{\text{o}} = 5$ (7 [dB]). What is the minimum element spacing that achieves this directivity?

Solution:

Since the individual elements are isotropic, the directivity of the array equals the directivity of the array factor alone. Using Equation (14.101), we have

$$5 = 2 \times (10) \times \left(\frac{d}{\lambda}\right).$$

Solving for d, we find that the minimum spacing is

$$d = \lambda/4.$$

Finally, to see if our use of Equation (14.101) was justified, we find that

$$Nkd = 10 \times \left(\frac{2\pi}{\lambda}\right) \times \left(\frac{\lambda}{4}\right) = 15.7 > 6,$$

which validates our procedure.

End-Fire Arrays. An end-fire array directs its main lobe along the array axis. Unlike broadside arrays, in which the incremental current phase shift δ is always the same (zero), there are a number of phase shifts that result in end-fire radiation patterns. The simplest scheme occurs when the fields of each element arrive in the end-fire direction with the same phase. Since the incremental propagation delay along the end-fire direction ($\theta = 0°$) is kd, the required current phase shift δ is

$$\delta = -kd \qquad \text{(Ordinary end-fire array)}, \tag{14.102}$$

where the negative sign indicates that the element phases are progressively more delayed in the end-fire direction.

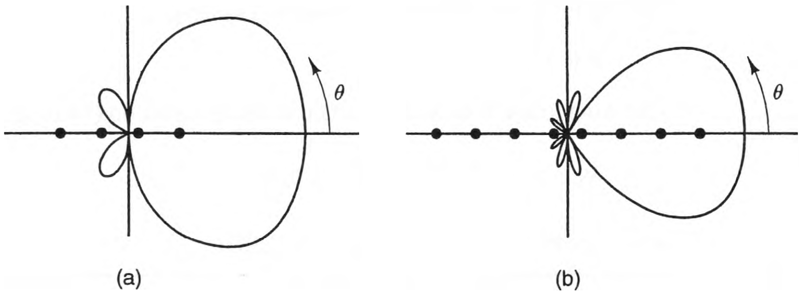


Figure 14-37 Array factor plots for two end-fire arrays with the same element spacing, $d = \lambda/4$, and $\delta = -kd$: a) $N = 4$, b) $N = 8$.

As with broadside arrays, the directivities of end-fire arrays are functions of both the number of elements N and the element spacing d. Figure 14-37 shows the pattern functions for arrays with $N = 4$ and $N = 8$, both with $d = 0.25\lambda$. As can be seen, the pattern becomes more directive as N increases, although the increase in directivity is more modest than occurs for broadside arrays. The directivity of end-fire arrays can also be increased by increasing d, but at the expense of higher side lobe levels. In practice, the array spacing is usually chosen so that $d \leqslant \lambda/4$ in order to avoid high lobes in the backward direction ($\theta = 180°$).

We can obtain an expression for the directivity of an end-fire array by substituting the array factor $\Lambda(\psi)$ into Equation (14.55), yielding

$$D_o = \frac{4\pi|\Lambda_{\text{max}}|^2}{\oint_S |\Lambda(\theta, \phi)|^2 d\Omega} = \frac{4\pi N^2}{\displaystyle\int_0^{2\pi}\int_0^{\pi} \left[\frac{\sin\left[\frac{N}{2} kd(\cos\theta - 1)\right]}{\sin\left[\frac{1}{2} kd(\cos\theta - 1)\right]}\right]^2 \sin\theta \, d\theta \, d\phi}.$$

When $Nkd > 3$, the following approximation is accurate to at least 10%:

$$D_o \approx 4N\left(\frac{d}{\lambda}\right) = 4\left(1 + \frac{L}{d}\right)\left(\frac{d}{\lambda}\right). \tag{14.103}$$

Here, $L = (N - 1)d$ is the total length of the array.

Example 14-7

A 10-element ordinary end fire array of isotropic radiators is to have a directivity of $D_o = 5$ (7 [dB]). What is the minimum element spacing that achieves this directivity?

Solution:

Since the individual elements are isotropic, the directivity of the array equals the directivity of the array factor alone. Using Equation (14.103), we have

$$5 = 4 \times (10) \times \left(\frac{d}{\lambda}\right).$$

Solving for d, we find that the minimum spacing is

$$d = \lambda/8.$$

Checking to see if we were justified in using Equation (14.103), we see that

$$Nkd = 10 \times \left(\frac{2\pi}{\lambda}\right) \times \left(\frac{\lambda}{8}\right) = 7.58 > 3,$$

which justifies our use of Equation (14.103).

14-6-3 NONUNIFORM ARRAYS

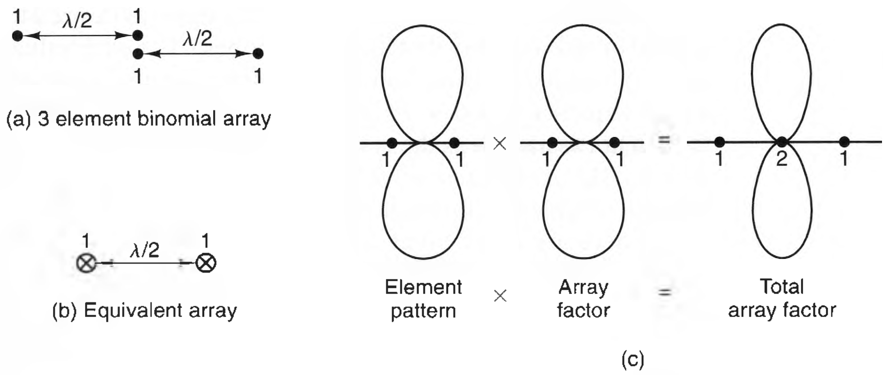
In our discussion of uniform, linear arrays we saw that they are capable of achieving very high directivities when the number of elements is large. Since the current magnitudes at the feeds are all the same, the feed networks for these arrays are relatively simple to build. On the down side, however, the side-lobe levels of uniform arrays tend to be quite high. For instance, the first side lobe of a large array can be reduced by no more than 13.5[dB] from the main beam level. For applications in which the side-lobe levels are critical (such as radars), this kind of performance is unacceptable.

One way to reduce side-lobe levels of an array is to use nonuniform element feeds, since side-lobe levels decrease when the center elements of an array are excited more strongly than the edge elements. These are called ***tapered feeds***. A simple array that uses tapered feeds is the binomial array, which produces a single main lobe with no side lobes. We can develop the characteristics of binomial arrays by repeatedly using the pattern multiplication theorem and the properties of two element uniform arrays.

To start our discussion of binomial arrays, let us review the characteristics of two-element broadside arrays. As shown in Figure 14-33, two-element arrays have a single main lobe with deep nulls in the end-fire directions when the element spacing is $d = \lambda/2$ and the feed currents are identical. Further increases in d will narrow the main beam, but only at the expense of radiation in the end-fire directions.

Let us now consider the radiation pattern of the two, two-element broadside arrays shown in Figure 14-38a. Here, each two-element array is fed with uniform currents ($1 \angle 0°$) and have element spacings of $d = \lambda/2$. The center-to-center offset spacing between the arrays is also $\lambda/2$, so the two overlapping elements in the center can be considered as a single element with a feed current of $2 \angle 0°$. Hence, this configuration can also be considered as a three-element nonuniform array with feeds 1–2–1 (all in phase).

We can determine the radiation pattern of this three-element array by considering it as a two-element array whose the elements are themselves two-element arrays. This interpretation is depicted in Figure 14-38b, where each symbol $\otimes$ denotes a two-element broadside array with $d = \lambda/2$. Using the pattern multiplication theorem, we see that the radiation pattern of the three-element nonuniform array is the product of the element pattern and the array factor of a two-element broadside array. For this case, the element pattern and the array factor are the same, and the total array factor for the three-element binomial array is shown in Figure 14-38c. This pattern function is more directive than the two-element pattern and yet has no side lobes. Also, we



(a) 3 element binomial array

(b) Equivalent array

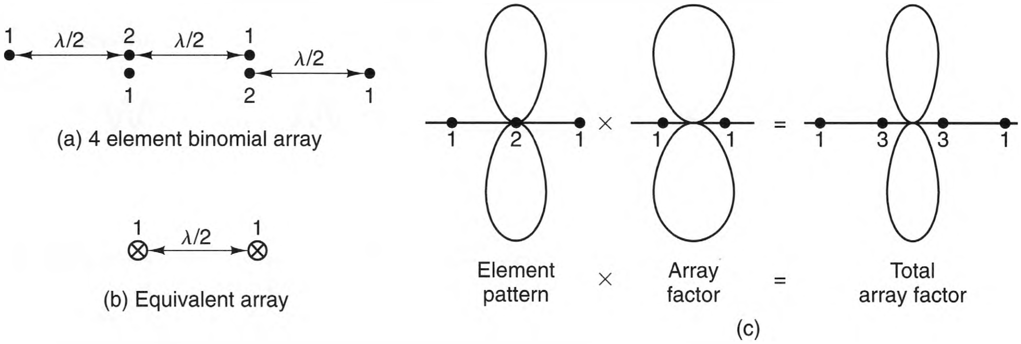
Element pattern × Array factor = Total array factor

(c)

Figure 14-38 Constructing a three-element binomial array out of two, two-element arrays: a) Two overlapping two-element arrays. b) An equivalent two-element array in which each element is a two-element array. c) The binomial array pattern equals the product of the element pattern and the array factor.

notice that the feed sequence 1–2–1 is the third row of Pascal's triangle, which means that these feeds are binomial coefficients for $m = 3$.

We can extend this same idea to construct N-element binomial arrays. For example, to construct a four-element binomial array, we can place two three-element binomial arrays as shown in Figure 14-39a. The overlapping elements in the center can be considered to form single elements, so this entire configuration is a four-element array, with element spacing $d = \lambda/2$ and feed sequence 1–3–3–1 (the fourth row of Pascal's triangle). We can determine the radiation pattern of this array by treating it as a two-element array whose elements are three-element binomial arrays. This interpretation is depicted in Figure 14-39b, where the symbol ⊗ represents a three-element binomial array. From the pattern multiplication theorem, the total radiation pattern is the product of the compound element pattern (from Figure 14-38c) and the array factor of a two-element array.



(a) 4 element binomial array

(b) Equivalent array

Element pattern × Array factor = Total array factor

(c)

Figure 14-39 Constructing a four-element binomial array out of two, three-element binomial arrays: a) Two overlapping three-element binomial arrays. b) An equivalent two-element array in which each element is a three-element binomial array. c) The four-element binomial array pattern equals the product of the element pattern and the array factor.

This process can be repeated as many times as desired to obtain an N-element, binomial broadside array that has no side lobes. For each value of N, the current feeds always follow the binomial distribution. For instance, a five-element binomial array has a feed sequence 1–4–6–4–1. The beamwidths become more narrow as N is increased, though not as narrow as the main beam of a uniform array with the same number of elements. This is typical of all side-lobe reduction schemes; side lobes are reduced at the expense of widening the main beam.

There are many other feed-tapering schemes that produce lowered side lobes. Generally, the amount of side-lobe reduction is proportional to the contrast between the feed amplitudes in the middle of the array and those on the edges. Unfortunately, it is often difficult to fabricate feed networks that drive some elements with large currents and others with small currents. As a result, practical designs usually represent a trade-off between side-lobe levels and the cost of the feed network.

14-7 Properties of Receiving Antennas

So far, we have only considered the transmitting properties of antennas. In this discussion, we have seen that a voltage or current applied to the terminals of an antenna produces radiated fields. It is just as possible, however, for an antenna to capture power from an incident wave and direct it to a load. This property is obvious for an aperture antenna, such as a horn. In that case, the horn simply captures a portion of the incoming wave front and delivers the power to the waveguide feed. The same is true of all antennas, even when the physics of the receiving mechanism is not so obvious.

In this section, we will look closely at the receiving characteristics of antennas and demonstrate how they are directly related to their transmitting properties.

14-7-1 ANTENNA EQUIVALENT CIRCUITS

Let us start by considering a pair of antennas that are capable of both transmitting and receiving energy to and from each other. Such a situation is depicted in Figure 14-40a, which shows two antennas in free space. Each antenna has a pair of terminals (i.e., a port) to which generators or loads can be attached.

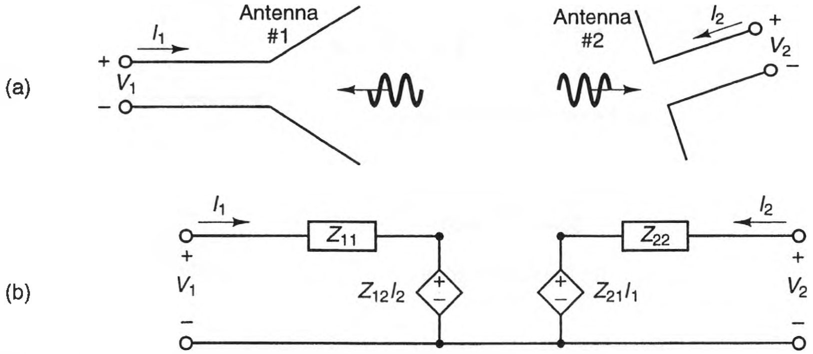


Figure 14-40 Coupled antennas: a) Two arbitrary antennas. b) The equivalent circuit.

Since free space is a linear medium, we can relate the port voltages and currents of these antennas using Z (impedance) parameters, just as we do with any other two-port network:

$$V_1 = Z_{11}I_1 + Z_{12}I_2 \tag{14.104}$$

$$V_2 = Z_{21}I_1 + Z_{22}I_2\,. \tag{14.105}$$

The parameters Z_{11}, Z_{12}, Z_{21}, and Z_{22} make up the Z-matrix, which completely defines the port characteristics of this antenna system. Figure 14-40b shows a lumped equivalent circuit that has the same port characteristics.

When antennas #1 and #2 are far apart, their input impedances are Z_{11} and Z_{22}, respectively. To show this, we note that Z_{11} is the ratio of the voltage and current at antenna #1 when the terminals of antenna #2 are short circuited:

$$Z_{11} = \left.\frac{V_1}{I_1}\right|_{I_2=0}.$$

However, if antenna #2 has no power applied to it, it will have a negligible effect on the currents and voltages on antenna #1 when the distance between them is large. The same is true if a source is placed at the terminals of antenna #2 and the terminals of antenna #1 are short circuited. As a result, we have

$$Z_{11} \approx Z_{A1} \tag{14.106}$$

and

$$Z_{22} \approx Z_{A2}\,, \tag{14.107}$$

where Z_{A1} and Z_{A2} are the input impedances of antennas #1 and #2, respectively, when they are driven as isolated transmitters. We will call these impedances the ***self-impedances*** of the antennas.

Unlike the self-impedance terms Z_{11} and Z_{22}, which are nearly independent of the presence of the other antenna, the mutual impedance terms Z_{12} and Z_{21} are functions of the distance between the antennas and their relative orientations. They are also functions of the polarization states of the antennas, since a receiving antenna responds best to the same polarization it emits when transmitting. Free space is reciprocal,[10] so the reciprocity principle applies, which states that the off-diagonal elements of the impedance matrix are equal. Hence,

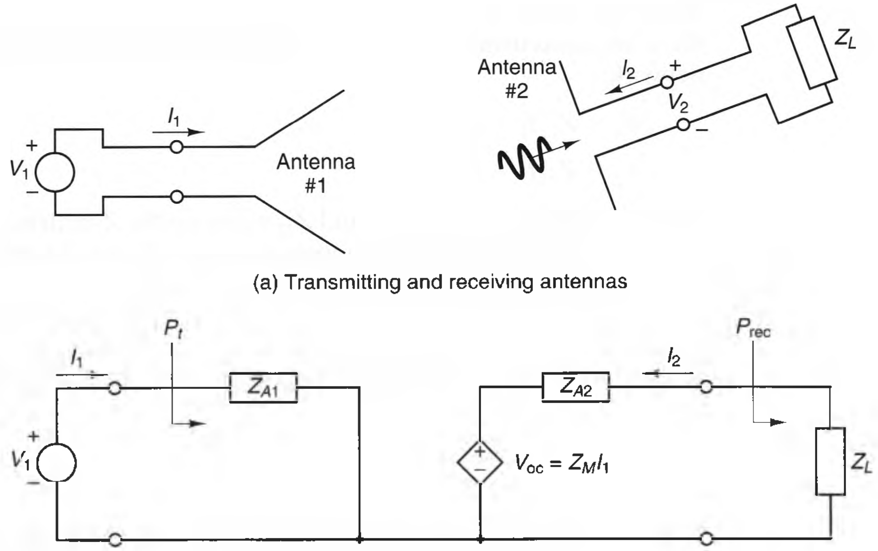
$$Z_{12} = Z_{21} \equiv Z_M\,, \tag{14.108}$$

where Z_M is called the ***mutual impedance*** of the antennas. Using Equations (14.106), (14.107), and (14.108), we can write the port characteristics of our two-antenna network as

$$V_1 \approx Z_{A1}I_1 + Z_MI_2 \tag{14.109}$$

$$V_2 \approx Z_MI_1 + Z_{A2}I_2. \tag{14.110}$$

[10] Most media are reciprocal, but there are some important exceptions, such as the ionosphere under the influence of the earth's magnetic field.



(a) Transmitting and receiving antennas

(b) Equivalent circuit

Figure 14-41 Transmitting and receiving antennas: a) A transmitting antenna connected to a voltage source and a receiving antenna attached to a load. b) The equivalent circuit when the mutual coupling is weak.

Consider now the situation illustrated in Figure 14-41a, which shows a voltage generator connected to antenna #1 and a passive load connected to antenna #2. Figure 14-41b shows the equivalent circuit for this network, where we have assumed that the antennas are spaced far enough apart so that the effect of the receiving antenna on the transmitting antenna is negligible. For this case, the dependent voltage source normally present in the transmitter circuit (see Figure 14.40b) can be neglected. Using standard circuit analysis, we find that the transmitted power is given by

$$P_t = \frac{1}{2}\frac{V_1^2}{|Z_{A1}|^2}\operatorname{Re}(Z_{A1}) \qquad [\mathrm{W}]. \tag{14.111}$$

Similarly, we can express the received power delivered to the load Z_L by the expression

$$P_{\text{rec}} = \frac{1}{2}\frac{|V_{\text{oc}}|^2}{|Z_{A2} + Z_L|^2}\operatorname{Re}(Z_L) \tag{14.112}$$

where V_{oc} is the open-circuit voltage at the receiving-antenna terminals and is given by

$$V_{\text{oc}} = Z_M I_1 = \frac{Z_M V_1}{Z_{A1}}. \tag{14.113}$$

When $Z_L = Z_{A2}^*$ (i.e., a conjugate-matched load), P_{rec} attains its maximum value, which is given by

$$P_{\text{rec}} = \frac{1}{8}\frac{|V_{\text{oc}}|^2}{\operatorname{Re}(Z_{A2})} \qquad [\mathrm{W}]\ (\text{Conjugate-matched load}). \tag{14.114}$$

Substituting Equations (14.112) and (14.113) into Equation (14.114) and rearranging, we obtain the following expression for the ratio of the received power to the transmitted power:

$$\frac{P_r}{P_t} = \frac{1}{4}\frac{|Z_M|^2}{\mathrm{Re}(Z_{A1})\mathrm{Re}(Z_{A2})} \qquad \text{(Conjugate-matched load)}. \qquad (14.115)$$

A noteworthy feature of Equation (14.115) is that it is symmetric in the subscripts "1" and "2." This means that the same transfer function is obtained if power is supplied to antenna #2 and antenna #1 is terminated in a conjugate-matched load. Hence, we see that

The power transferred between two antennas is independent of which antenna is transmitting, as long as the receiving antenna is terminated in a conjugate-matched load.

As a final point, some readers might be wondering what happens to the power that is "lost" in the self-impedance of an antenna that is acting as a receiver. At first glance, it may appear that there should be no lost power, particularly if the antenna is constructed with lossless materials. Closer examination, however, reveals that this "lost" power is actually power that is reradiated by the antenna. This should not be surprising, since a current distribution on an antenna will always radiate power—even when the current is caused by an incoming field.

14-7-2 EFFECTIVE APERTURE

The effective aperture of an antenna is a parameter that represents the electrical surface area that an antenna presents to an incoming wave. We can introduce the concept by considering the horn antenna shown in Figure 14-42. Here, a plane wave impinges upon the horn, which delivers power to a matched load. Because the incident power density is measured in watts per square meter, it is reasonable that the power delivered to the load is proportional to the product of the incident power density and the area of the aperture. This line of reasoning is sensible for antennas that have obvious physical apertures, such as horn and reflector antennas. But even thin wire antennas are capable of "grabbing" power from an incident wave, in spite of their apparent lack of a physical aperture.

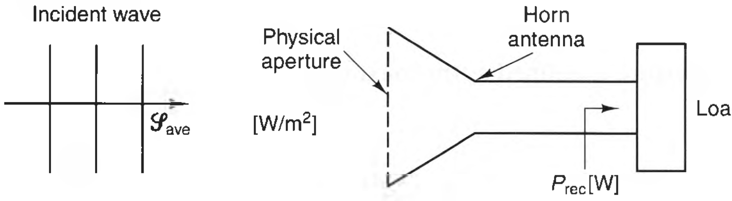


Figure 14-42 A plane wave impinging upon a horn antenna that is terminated with a load.

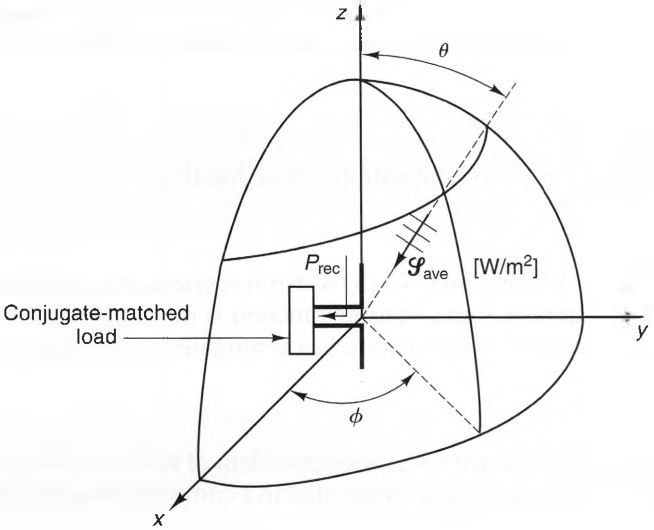


Figure 14-43 An experimental setup for measuring the effective aperture of an antenna.

To probe the matter further, let us consider the situation shown in Figure 14-43, which depicts an arbitrary antenna at the origin and an incident plane wave approaching from the direction (θ, ϕ). If P_{rec} is the power delivered to a conjugate-matched load and $\mathcal{S}_{\text{ave}}$ is the incident power density, we define the effective aperture $A_e(\theta, \phi)$ of the antenna as the ratio of P_{rec} to $\mathcal{S}_{\text{ave}}$ when the polarization state of the incident wave matches the antenna's polarization state. In equation form, this definition can be stated as

$$A_e(\theta, \phi) = \frac{P_{\text{rec}}}{\mathcal{S}_{\text{ave}}} \quad [\text{m}^2] \quad \text{(Conjugate-matched load)}. \tag{14.116}$$

Using Equation (14.114), we can also express $A_e(\theta, \phi)$ in terms of the open-circuit voltage and the self-impedance of the antenna:

$$A_e(\theta, \phi) = \frac{|V_{\text{oc}}|^2}{8\mathcal{S}_{\text{ave}}} \frac{1}{\text{Re}[Z_A]} \quad [\text{m}^2] \quad \text{(Conjugate-matched load)}. \tag{14.117}$$

The maximum effective aperture is defined as the maximum value of $A_e(\theta, \phi)$:

$$A_{em} \equiv A_e(\theta, \phi)\Big|_{\text{max}} . \tag{14.118}$$

Example 14-8

Find the maximum effective area A_{em} of the short dipole shown in Figure 14-44.

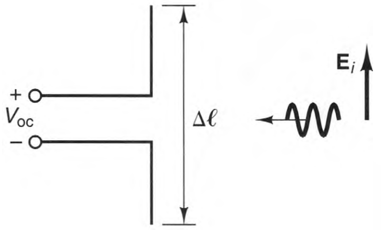


Figure 14-44 A plane wave that is normally incident upon a short, open-circuited dipole.

Solution:

We will assume that the capacitance of the antenna tips is high enough so that the wire current is uniform when the antenna is driven as a transmitter. This means that the open-circuit voltage (when the antenna is receiving) equals the product of the incident electric-field intensity E_i and the dipole length $\Delta\ell$ (see Problem 14.28):

$$V_{oc} = E_i \Delta\ell .$$

The input resistance of this antenna is given by Equation (14.48) and is

$$\mathrm{Re}(Z_A) = R_A = 80\pi^2\left(\frac{\Delta\ell}{\lambda}\right)^2 = \frac{2\pi\eta}{3}\left(\frac{\Delta\ell}{\lambda}\right)^2 .$$

where η and λ are the free-space impedance and wavelength, respectively. Also, the incident power density $\mathcal{S}_{\text{ave}}$ is given by Equation (12.107):

$$\mathcal{S}_{\text{ave}} = \frac{1}{2}\frac{|E_i|^2}{\eta} \qquad [\mathrm{W/m^2}].$$

Substituting the preceding three expressions into Equations (14.117) and (14.118), we obtain

$$A_{em} = \frac{(E_i\Delta\ell)^2}{8(E_i^2/2\eta)(2\pi\eta\Delta\ell^2/3\lambda^2)} = \frac{3\lambda^2}{8\pi} = 0.119\lambda^2 \qquad [\mathrm{m^2}].$$

This result is noteworthy because it is independent of the dipole length $\Delta\ell$, as long as $\Delta\ell \ll \lambda$.

To show how the effective aperture A_e and directivity D of an antenna are related, consider the two antennas shown in Figure 14-45. Here, antenna #1 is fixed at the origin. Antenna #2 is free to move along a sphere of constant radius, but it is always oriented such that it directs its maximum radiation intensity towards the origin. When antenna #2 radiates power P_{t2}, the power density $\mathcal{S}_{\text{ave}}$ arriving at antenna #1 is independent of the angular position (θ, ϕ) and is given by

$$\mathcal{S}_{\text{ave}} = \frac{P_{t2}}{4\pi r^2} D_2 ,$$

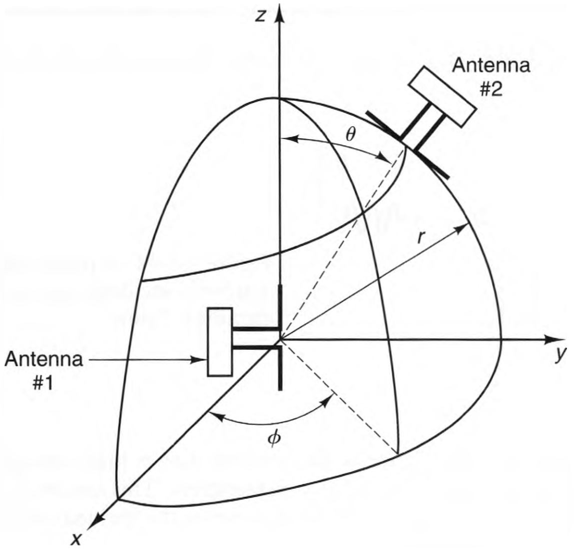


Figure 14-45 An experimental setup for showing that the transmitting and receiving patterns of an antenna are identical.

where D_2 is the directivity of antenna #2. Substituting this into Equation (14.116) and solving for the ratio P_{r1}/P_{t2}, we obtain

$$\frac{P_{r1}}{P_{t2}} = \frac{A_{e1}(\theta, \phi) D_2}{4\pi r^2}, \tag{14.119}$$

where $A_{e1}(\theta, \phi)$ is the effective aperture of antenna #1 along the direction (θ, ϕ).

When the experiment is reversed, so that antenna #1 is the transmitter, the power density arriving at antenna #2 is

$$\mathcal{S}_{\text{ave}} = \frac{P_{t1}}{4\pi r^2} D_1(\theta, \phi),$$

where $D_1(\theta, \phi)$ is the directive gain of antenna #1 along the direction (θ, ϕ). If P_{r2} is the power received by antenna #2, we can substitute the preceding expression into Equation (14.119) to find the ratio P_{r2}/P_{t1}:

$$\frac{P_{r2}}{P_{t1}} = \frac{A_{em2} D_1(\theta, \phi)}{4\pi r^2}. \tag{14.120}$$

Here, A_{em2} is the effective aperture of antenna #2 when the incident waves arrive along its direction of maximum effective aperture.

Since the power transferred between two antennas is independent of which one is transmitting, the ratios P_{r1}/P_{t2} and P_{r2}/P_{t1} must be identical. Setting Equations (14.119) and (14.120) equal to each other, we find that

$$\frac{A_{e1}(\theta, \phi)}{D_1(\theta, \phi)} = \frac{A_{em2}}{D_2}.$$

However, since both antennas were arbitrary, we can conclude that both sides of this expression are constants, regardless of the types of antennas used. Hence, we can drop the subscripts from the expression and write

$$\frac{A_e(\theta, \phi)}{D_g(\theta, \phi)} = \text{constant}, \tag{14.121}$$

which states that the ratio of an antenna's directive gain in a direction (θ, ϕ) to its effective aperture in that direction is a universal constant. *This means that the transmit and receive patterns of an antenna are always identical.*

We can find the universal constant in Equation (14.121) by using the results of Example 14-7. In that example, we showed that the maximum effective aperture of a Hertzian dipole is given by

$$A_{em} = \frac{3\lambda^2}{8\pi} \qquad \text{(Hertzian dipole)}.$$

But we also know that the directivity of a Hertzian dipole is

$$D_o = \frac{3}{2} \qquad \text{(Hertzian dipole)}.$$

Taking the ratio of these expressions, we find that

$$\frac{A_{em}}{D_o} = \frac{3\lambda^2/8\pi}{3/2} = \frac{\lambda^2}{4\pi}.$$

Thus, the universal constant is $\lambda^2/(4\pi)$, and Equation (14.121) can be rewritten as

$$A_e(\theta, \phi) = \frac{\lambda^2}{4\pi} D_g(\theta, \phi) \qquad [\text{m}^2]. \tag{14.122}$$

Since $A_e(\theta, \phi)$ and $D_g(\theta, \phi)$ have their maximum values in the same direction (θ, ϕ), we also have

$$A_{em} = \frac{\lambda^2}{4\pi} D_o \qquad [\text{m}^2]. \tag{14.123}$$

Example 14-9

Find the power delivered to a conjugate-matched load from an antenna that is illuminated by a 2 [GHz] plane wave. Assume that the antenna has a directivity of 5 [dB] and the incoming wave is incident along the antenna's direction of maximum sensitivity, is polarization matched to the antenna, and has a peak amplitude of 500 [mV/m].

Solution:

The directivity in linear units is

$$D_o = 10^{\frac{5}{10}} = 3.16.$$

Also, the free-space wavelength is

$$\lambda = \frac{3 \times 10^8[\text{m}\cdot\text{s}^{-1}]}{2 \times 10^9[\text{s}^{-1}]} = 0.15 \qquad [\text{m}].$$

Substituting these values into Equation (14.123) yields a maximum effective aperture of

$$A_{em} = \frac{(.15)^2}{4\pi} \times 3.16 = 5.66 \times 10^{-3} \qquad [\text{m}^2].$$

The power density of the incident wave is

$$\mathcal{S}_{\text{ave}} = \frac{1}{2}\frac{(500 \times 10^{-3}[\text{V}])^2}{377[\Omega]} = 3.32 \times 10^{-4} \qquad [\text{W/m}^2].$$

Substituting these values of A_{em} and $\mathcal{S}_{\text{ave}}$ into Equation (14.116), we obtain

$$P_{\text{rec}} = 3.32 \times 10^{-4} \quad [\text{W/m}^2] \times 5.66 \times 10^{-3} \quad [\text{m}^2] = 1.88 \qquad [\mu\text{W}].$$

Some readers may be wondering why the wavelength-squared term appears in the expressions that relate the effective aperture A_{em} to the directivity D_o of an antenna (Equations (14.122) and (14.123)). One explanation is that such a term is needed in order for these expressions to be dimensionally correct. Another, more insightful, explanation is that short-wavelength (i.e., high-frequency) antennas are smaller than low-frequency antennas with the same directivity. Since the physical sizes of high-frequency antennas are smaller, their effective apertures are also smaller.

As a final comment, we note that Equations (14.122) and (14.123) were derived for lossess antennas. Under this circumstance, all the received power is delivered to the load. It is relatively simple to show that these same expressions hold for lossy antennas, except that the directive gain $D_g(\theta, \phi)$ is replaced by the power gain $G_g(\theta, \phi)$, and the directivity D_o is replaced by the gain G_o.

14-7-3 ANTENNA LINKS AND THE FRISS TRANSMISSION EQUATION

Now that we have discussed both the transmitting and receiving characteristics of antennas, we can determine the complete transfer function between transmitting and receiving antennas. Figure 14-46 shows a typical antenna link, consisting of transmitting and receiving antennas, spaced a distance R apart. We will assume that the antennas are polarization matched and oriented such that each antenna lies in the direction of maximum radiation of the other.

If the transmitting antenna has a gain G_t, the time-averaged power density at the receiving antenna is

$$\mathcal{S}_{\text{ave}} = \frac{P_t}{4\pi R^2} G_t \qquad [\text{W/m}^2],$$

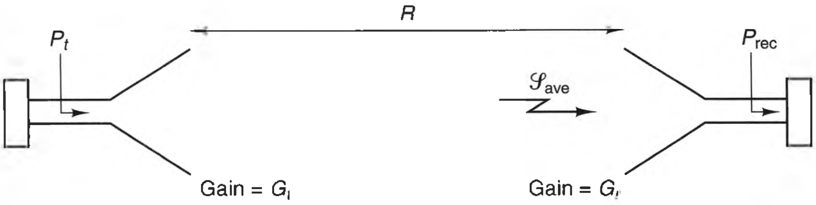


Figure 14-46 Geometry for deriving the Friss transmission equation.

where P_t is the input power to the transmitting antenna. From Equation (14.116), the power P_{rec} delivered to a conjugate-matched load attached to the terminals of the receiving antenna is

$$P_{\text{out}} = \mathcal{S}_{\text{ave}} A_{er} = \frac{P_t}{4\pi R^2} G_t A_{er},$$

where A_{er} is the maximum effective aperture of the receiving antenna. Thus, the transfer function that relates the transmitted and received powers is given by the expression

$$\frac{P_{\text{rec}}}{P_t} = \frac{G_t A_{er}}{4\pi R^2}.$$

Using Equation (14.122), we can replace the effective aperture A_{er} in this expression with $(\lambda^2/4\pi)\, D_r$, where D_r is the receiving antenna directivity, yielding

$$\frac{P_{\text{rec}}}{P_t} = \left(\frac{\lambda}{4\pi R}\right)^2 G_t G_r. \tag{14.124}$$

Equation (14.124) is called the ***Friss transmission equation***, which states that

The amount of power transferred between two antennas is proportional to the product of the antenna gains.

According to the Friss equation, the deficiencies of a low-gain transmitting antenna can be compensated for by using a high-gain receiving antenna, and vice versa. This is an important consideration in many practical applications, since it is often necessary for one antenna of a transmit–receive link to have a low gain due to size or weight constraints, such as when antennas are placed on spacecraft or satellites.

Example 14-10

Design a transmit–receive link that delivers 1 [μW] of power to a load when 10 [W] is supplied to the transmitting antenna. If space requirements demand that one antenna has a gain G_a of only

5 [dB], find the necessary gain G_b of the other antenna if the distance between the antennas is 100 [km] and the operating frequency is 6 [GHz].

Solution:

In linear units, we have

$$G_a = 10^{\frac{5}{10}} = 3.16,$$

and the operating wavelength is

$$\lambda = \frac{3 \times 10^8}{6 \times 10^9} = 0.05 \qquad [\text{m}].$$

Substituting these values into Equation (14.124) and solving for G_b, we obtain the required gain of the second antenna:

$$G_b = \frac{1 \times 10^{-6}\,[\text{W}]}{10\,[\text{W}]} \times \left(\frac{4\pi \times 100 \times 10^3\,[\text{m}]}{.05\,[\text{m}]}\right)^2 = 6.32 \times 10^7 = 78\,[\text{dB}].$$

14-7-4 RADAR CROSS SECTION AND THE RADAR RANGE EQUATION

Antennas are key components in radars, which are used to determine the position and velocity of objects such as aircraft and ships. Figure 14-47 shows a simple radar configuration. Here, a transmitting antenna radiates a field that illuminates a target. The target then reradiates (or scatters) a portion of this energy back to a receiving antenna. To simplify matters, we will assume that both the transmitting and receiving antennas are the same distance R from the target. Radars whose transmit and receive antennas are colocated are called ***monostatic radars***, whereas those with transmit and receive antennas in different locations are called ***bistatic radars***. Most radars are monostatic, since it is usually more convenient to use the same antenna for both transmitting and receiving.

To determine the power received by the receiving antenna of a radar, we must first model the scattering characteristics of the target. This is done using a parameter called the ***radar cross section*** σ (also called the ***echo area***), which is defined as follows:

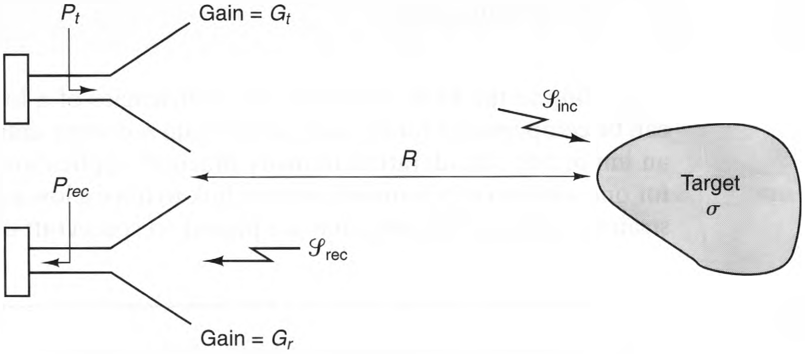


Figure 14-47 Geometry for deriving the radar range equation.

The radar cross section σ of a scatterer is the cross-sectional area of an idealized isotropic scatterer that reradiates the same power density to the receiver as does the actual target.

Using this definition, we find that the power P_c captured by the equivalent scatterer is

$$P_c = \mathcal{S}_{\text{inc}}\sigma \qquad [\text{W}],$$

where $\mathcal{S}_{\text{inc}}$ is the power density incident upon the scatterer. Also, since the equivalent scatterer reradiates this power isotropically, the scattered power density $\mathcal{S}_s$ at the receiver is

$$\mathcal{S}_{\text{rcc}} = \frac{P_c}{4\pi R^2} = \frac{\mathcal{S}_{\text{inc}}\sigma}{4\pi R^2} \qquad [\text{W/m}^2]. \tag{14.125}$$

Solving this expression for σ, we find that the radar cross section is given by

$$\sigma = \lim_{R\to\infty}\left[\frac{4\pi R^2\mathcal{S}_{\text{rec}}}{\mathcal{S}_{\text{inc}}}\right] \qquad [\text{m}^2]. \tag{14.126}$$

If the gain of the transmitting antenna is G_t, the incident power density at the scatterer is

$$\mathcal{S}_{\text{inc}} = \frac{P_t}{4\pi R^2}G_t,$$

where P_t is the power input to the transmitter. From Equation (14.125), the scattered power density at the receiving antenna is

$$\mathcal{S}_{\text{rec}} = \frac{\sigma P_t G_t}{(4\pi R^2)^2}.$$

If the effective aperture of the receiving antenna is A_{er}, the power P_r delivered to the load is

$$P_{\text{rec}} = A_{er}\mathcal{S}_{\text{rec}} = A_{er}\frac{\sigma P_t G_t}{(4\pi R^2)^2}.$$

Finally, using $A_{er} = \lambda^2/(4\pi)\,G_r$, where G_r is the gain of the receiving antenna, we obtain

$$\frac{P_{\text{rec}}}{P_t} = \sigma\frac{G_t G_r}{4\pi}\left[\frac{\lambda}{4\pi R^2}\right]^2. \tag{14.127}$$

Equation (14.127) is called the ***radar range equation***. This equation is used routinely in radar system calculations to estimate target sizes from a knowledge of the received power and the antenna gains. It can also be used to determine the transmitted power and antenna gains necessary to "see" a specified target at given range (distance).

Example 14-11

A radar system is to be built that is capable of detecting targets with radar cross sections of 10 [m^2] at a range of 100 [km] at a frequency of 2 [GHz]. If the transmitting and receiving antennas both have gains of 40 [dB], and the minimum detectable signal at the receiver is 0.5 [nW], find the minimum transmitted power.

Solution:

In linear units, the antenna gains are

$$G_{\text{ot}} = G_{\text{or}} = 10^{\frac{40}{10}} = 10^4.$$

Also, the free-space wavelength λ is

$$\lambda = \frac{3 \times 10^8}{2 \times 10^9} = 0.15 \qquad [\text{m}].$$

Substituting these values into Equation (14.127), we find that the minimum transmitted power is

$$P_t \geqslant \frac{0.5 \times 10^{-9}}{10} \times \left[\frac{4\pi \times (100 \times 10^3)^2}{.15}\right]^2 \times \frac{4\pi}{10^4 \times 10^4} = 4.41 \qquad [\text{MW}].$$

14-8 Summation

In this chapter, we have discussed various aspects of radiation, including its causes, ways that it can be controlled, and a number of structures that can initiate it. Although this discussion was certainly not exhaustive, the reader should have a good background for tackling radiating systems and problems.

Perhaps the key concept that should be remembered when dealing with radiation is that radiated fields are totally dependent on the fields and currents on and near the radiating structure. If these quantities can be controlled, the radiated fields can also be controlled. This is true for antennas, in which we want radiation, and also for systems in which radiation is undesirable. In both cases, the way to control the far-zone behavior of a structure is to pay close attention to the fields and currents close to the structure, as well as the shape of the structure itself.

Problems

14-1 For the vector potential $\mathbf{A} = A_0\,(e^{-jkz}/\rho)\,\hat{\mathbf{a}}_\rho$, where $k = \omega\sqrt{\mu\epsilon}$,

(a) Show that $\mathbf{A}$ satisfies Equation (14.15) in the dielectric region of a coaxial cable.

(b) Find the scalar potential V that corresponds to $\mathbf{A}$ using the Lorentz gauge.

(c) Find the E- and H-fields that are associated with these potentials, and show that they are the TEM-fields of a coaxial cable.

14-2 In Example 14-1 it was shown that the potential pair $\mathbf{A}_1 = A_o e^{-jkz}\,\hat{\mathbf{a}}_x$ and $V_1 = 0$ corresponds to a linearly polarized planc wave. But this potential pair is not unique. Another vector potential that produces the same E- and H-fields is

$$\mathbf{A}_2 = A_o(\hat{\mathbf{a}}_x + \hat{\mathbf{a}}_z)e^{-jkz}.$$

(a) Using the Lorentz gauge, find the scalar potential V_2 that corresponds to $\mathbf{A}_2$.
(b) Show that $\mathbf{A}_2$ and V_2 correspond to the same E- and H-fields as do $\mathbf{A}_1$ and V_1.

14-3 A Hertzian dipole is located at the origin. If a measurement shows that $|\mathbf{E}| = 10$ [mV/m] at a range of $r = 50$ [m] along angular coordinates (θ, ϕ), find $|\mathbf{E}|$ at a range $r = 8$ [m] along the same angular coordinates. Assume that both ranges are in the far zone of the dipole.

14-4 A 1.0 [cm] length wire segment lies along the z-axis and carries a uniform, 100 [MHz] current. If the wire is centered about the origin, calculate the ratio E_θ/H_ϕ in the xy-plane at a) $r = 0.1$ [m], b) $r = 0.5$ [m], and c) $r = 5.0$ [m].

14-5 Find the distance (in terms of wavelength) where the r^{-1}, r^{-2}, and r^{-3} factors in the expression for the E_θ of a Hertzian dipole are equal.

14-6 If an antenna has a far-zone E-field given by

$$\mathbf{E} = \frac{\sin 2\theta}{r} e^{-jkr}\,\hat{\mathbf{a}}_\theta,$$

calculate the radiated power.

14-7 Find the directivity of an antenna whose normalized radiation intensity is

$$U = \begin{cases} \sin\theta\sin\phi \quad \text{[W/sr]} & 0 < \theta < \pi \quad \text{and} \quad 0 < \phi < \pi \\ 0 & \text{otherwise} \end{cases}.$$

14-8 A sinusoidal current of with a peak amplitude of 0.5 [A] is applied to the input terminals of a lossless antenna. Calculate the input resistance if the resulting radiation pattern is given by

$$U = \begin{cases} 4\sin^2\theta\sin^2\phi \quad \text{[W/sr]} & 0 < \theta < 180^\circ \quad \text{and} \quad 0 < \phi < 180^\circ \\ 0 & \text{otherwise} \end{cases}.$$

14-9 A lossless antenna has a directivity of 12 [dB]. If its input resistance is 100 [Ω], calculate the rms current at the input terminals necessary to produce a radiation intensity of 100 [W/sr] in the direction of maximum radiation.

14-10 An 800 [MHz] current with a peak amplitude of 100 [mA] is applied to the terminals of an $\ell = 0.3$ [m], center-fed dipole. What is the maximum power density in [W/m^2] radiated by this antenna at a range of 60 [km].

14-11 A common "trick" for raising the input impedance of a half-wave dipole is to feed it off center. Using the assumption that the shape of the current distribution on a half-wave dipole antenna is insensitive to where the feed point is, find the input resistance of a half-wave dipole when the feed point is offset by a distance $h = \lambda/8$ from the center.

14-12 The length of typical automobile "whip" antenna is 85 [cm]. Find the input resistance to such an antenna at 1100 [KHz] (the middle of the AM broadcast band), assuming that the current induced along its length is uniform and the automobile body acts as an infinite ground plane.

14-13 Consider a circular loop antenna that consists of 100 turns of wire around a cylindrical core with radius of 1 [cm]. The windings are AWG 20 copper wire, which has a radius of 0.406 [mm] and a conductivity of $\sigma = 5.8 \times 10^7$ [S/m]. Assuming that the current flows uniformly within the skin depth of the wire, calculate the radiation efficiency of this antenna at $f = 500$ [kHz] if the permeability of the core is a) $\mu_r = 1.0$, b) $\mu_r = 200$.

14-14 Find the number of turns N required to fabricate a loop-stick antenna for use in the AM broadcast band (525–1610 [kHz]) that has a minimum in-band radiation resistance of 20 [Ω]. Assume that the windings are formed around a 0.25 [cm] diameter core that has a relative permeability of 500.

14-15 Determine the width (in wavelengths) of a long slot antenna with a uniform aperture field that radiates a main-beam width (null to null) of a) 30°, b) 15°, and c) 5°.

14-16 Derive the pattern function (Equation (14.91)) of the dipole-reflector antenna shown in Figure 14-29.

14-17 Determine the minimum number of elements necessary to obtain a null-to-null beam width of 20° for a broadside array of isotropic elements if the element spacing is $\lambda/4$.

14-18 A simple technique that is often used to produce a scanning radar antenna is to form arrays out of waveguide slots. Each slot radiates a field that depends on the amplitude and phase of the waveguide mode beneath it. Since waveguide modes are dispersive, the phase shift between adjacent elements is a function of frequency. This causes the main beam to scan with frequency without the need for mechanical gimbals or electric phase shifters for each element. For the five element slot array shown in Figure P14-18, calculate the range of angles $\Delta\theta$ scanned as the transmitting frequency is varied from $1.1 f_c$ to $1.95 f_c$, where f_c is the cutoff frequency of the dominant waveguide mode.

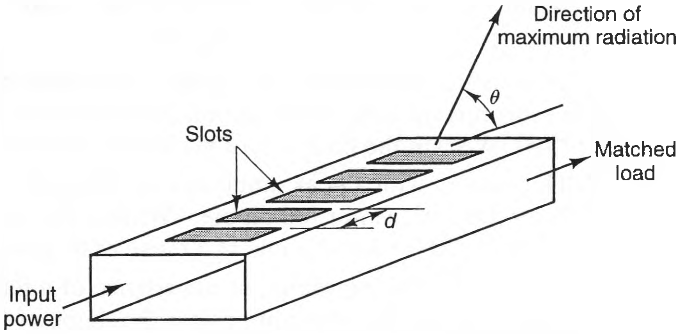


Figure P14-18

14-19 A 10-element broadside array has an element spacing of $\lambda/3$. Calculate the angles (with respect to the array axis) where pattern nulls appear.

14-20 Repeat Problem 14-19 for the case where the array is fed as an ordinary end-fire array.

14-21 Plot and compare the pattern functions for a four-element broadside array with element spacing 0.4λ when the elements are a) isotropic radiators, and b) $\lambda/2$ dipoles parallel to the array axis.

14-22 Plot and compare the pattern functions for a five-element, end-fire array with element spacing 0.2λ when the elements are a) isotropic radiators, and b) parallel $\lambda/2$ dipoles, mounted perpendicular to the array axis.

14-23 Find the relative feed-current magnitudes required to create a seven-element broadside, binomial array.

14-24 Using the identity

$$\left|\frac{\sin(N\psi/2)}{\sin(\psi/2)}\right|^2 = \frac{1}{N} + \frac{2}{N^2}\sum_{m=1}^{N-1}(N-m)\cos m\psi,$$

one can derive the following closed-form expression for the directivity of a linear array of N identical elements (see Stutzman and Thiele *Antenna Theory and Design*, New York):

$$D_o = \frac{1}{\dfrac{1}{N} + \dfrac{2}{N^2}\displaystyle\sum_{m=1}^{N-1}\frac{(N-m)}{mkd}\sin mkd\cos m\delta}.$$

In this equation, d is the element spacing, k is the free-space wave number, and δ is the incremental phase shift between elements. Compare the values obtained from the preceding formula with the approximate formula given by Equation (14.101) for broadside arrays with $d/\lambda = 0.4$ that have a) $N = 2$ and b) $N = 4$ elements.

14-25 Show that a six-element, linear broadside array can be arranged as either a two- or three-element array of compound elements. Specify what these complex elements are in terms of the original isotropic elements, and show that the array factors of these equivalent arrays are identical to the factors of the original array.

14-26 Calculate the open-circuit voltage V_{oc} of an antenna with gain 22 [dB] that is illuminated by a 3 [GHz] polarization-matched plane wave with a power density of 2.5×10^{-12} [W/m^2] along the direction of the antenna's maximum sensitivity. Assume that the antenna has an input impedance of 50 [Ω].

14-27 Calculate the directive gain D_g and effective aperture A_e of a $\lambda/2$ dipole at $f =$ 500 [MHz] and an observation angle of $\theta = 45°$, where θ is measured with respect to the dipole axis.

14-28 A well-known formula from antenna theory states that when a wire antenna is illuminated by an incident field E^i, the open circuit voltage V_{oc} is given by:

$$V_{oc} = \frac{1}{I_o}\int I(\ell)E^i(\ell)\,d\ell.$$

In this formula, $E^i(\ell)$ is the tangential component of the incident field along the wire, and $I(\ell)$ is the current induced on the wire when it is driven as a transmitter by a current source I_o at its terminals. The formula is proven in a number of texts (for instance, John Kraus, *Antennas*) and is a result of the reciprocity theorem of electromagnetics. For a half-wave dipole,

(a) use the formula to find V_{oc} when the incident field is a plane wave with peak amplitude E^i that is polarized parallel to the wire and propagates perpendicular to the wire,

(b) use the result in part a) to find the effective aperture of the antenna (*Hint:* Use Equation (14.117).)

(c) show that the effective aperture value found in b) corresponds to the directivity of a half-wavelength dipole, $D_o = 1.64 = 2.15$ [dB].

14-29 If an antenna with a 30 [dB] directivity is illuminated along its direction of maximum sensitivity with a polarization-matched, 2 [GHz] plane wave with a power density of 100 [mW/m^2], calculate the power that this antenna will deliver to a matched load.

14-30 A certain antenna has a directivity of 24 [dB] at 12 [GHz]. Find the power this antenna delivers to a matched load when a 1 [μW/m^2], linearly-polarized plane wave of the same frequency is incident upon the antenna along its direction of maximum sensitivity when

(a) the antenna is polarization matched with the incident plane wave,

(b) the antenna is circularly polarized.

14-31 Suppose that a communication channel must deliver a minimum of 3×10^{-14} [W] to the input terminals of its receiver in order to maintain the channel. If the transmitting and receiving antennas have gains of 15 [dB] and 10 [dB], respectively, and are 62 [km] apart, find the minimum power P_t that must be delivered to the transmitting antenna if the antennas are polarization matched, the channel frequency is 2 [GHz], and the receiver input impedance is conjugate matched to the receiving antenna.

14-32 A J-band radar uses the same antenna for transmitting and receiving, with a gain of 37 [dB] at a frequency of 18.6 [GHz]. If the peak transmitted power is 1.2 [MW] and the minimum detectable signal at the receiver is 14×10^{-15} [W], find the maximum range at which a target with radar cross section $\sigma = 5$ [m^2] can be detected.

14-33 Find the radar cross section σ of a target that returns a peak power of 1×10^{-10} [W] to a radar receiver when the peak transmitted power is 100 [kW], the frequency is 15 [GHz], and the range is 1.5 [km]. Assume that the transmitter and receiver use the same 25 [dB] antenna, which is conjugate matched to the receiver's input impedance.

14-34 Antenna side lobes can pose serious target-identification problems in radar systems. Consider a radar antenna with a 50 [dB] directive gain along its direction of maximum radiation. It also has a side lobe 30° off that direction with a directive gain of 20 [dB]. If a target with radar cross section σ_1 located along the angle of maximum directivity at a range of 10 [km] produces a received power P, find the radar cross section σ_2 of a target along the side-lobe angle at a range of 1 [km] that produces the same received power.

14-35 In the derivation of Equation (14.17), it was stated that:

$$\nabla^2\left[\frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}\right] = -k^2\frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|} - 4\pi\delta(x-x')\,\delta(y-y')\,\delta(z-z'),$$

where $\delta(x)$ is the Dirac delta function.

(a) Show that the first term on the right-hand side of this expression follows directly from the Lapacian operator (Equation (2.123)) by expanding **r** and **r**′ in Cartesian coordinates.

(b) Prove that the second term is needed at **r** = **r**′ by considering the integral $\int_{\Delta V}\nabla^2(e^{-jk|\mathbf{r}-\mathbf{r}'|}/|\mathbf{r}-\mathbf{r}'|)\,dv$, where ΔV is the small spherical volume of radius δ surrounding the point **r**′ shown in Figure P14-35.

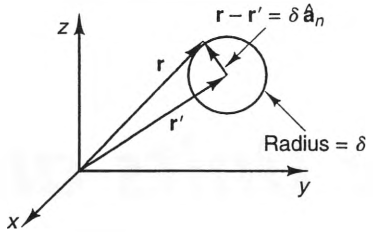


Figure P14-35

Use the divergence theorem to evaluate this integral using a surface integral around the spherical surface, noting that $\mathbf{r}-\mathbf{r}' = \delta\,\hat{\mathbf{a}}_n$ points outward from the surface at each point

Appendix A: Units and Symbols

TABLE A-1 MKSA (RATIONALIZED) UNITS

Quantity	Typical Symbols	Unit	Abbreviation
Length	ℓ, r, R	Meter	m
Mass	m	Kilogram	kg
Time	t	Second	s
Current	I, i	Ampere	A

TABLE A-2 DERIVED UNITS USED IN ELECTROMAGNETICS

Quantity	Typ. Symbol	Primary Unit	Common Alt. Unit	MKSA Equivalent Unit
Admittance	Y	siemens (S)	ampere/volt	$\mathrm{m}^2 \cdot \mathrm{kg}^{-1}\mathrm{s}^3 \cdot \mathrm{A}^2$
Attenuation	α	neper/meter(Np/m)	—	m^{-1}
Conductance	G	siemens (S)	ampere/volt	$\mathrm{m}^{-2} \cdot \mathrm{kg}^{-1} \cdot \mathrm{s}^3 \cdot \mathrm{A}^2$
Capacitance	C	farad (F)	coulomb/volt	$\mathrm{m}^{-2} \cdot \mathrm{kg}^{-1} \cdot \mathrm{s}^4 \cdot \mathrm{A}^2$
Charge	Q, q	coulomb (C)	—	$\mathrm{s} \cdot \mathrm{A}$
Charge density (volume)	ρ_v	coulomb/meter^3 (C/m^3)	—	$\mathrm{m}^{-3} \cdot \mathrm{s} \cdot \mathrm{A}$
Charge density (surface)	ρ_s	coulomb/meter^2 (C/m^2)	—	$\mathrm{m}^{-2} \cdot \mathrm{s} \cdot \mathrm{A}$
Charge density (line)	ρ_ℓ	coulomb/meter (C/m)	—	$\mathrm{m}^{-1} \cdot \mathrm{s} \cdot \mathrm{A}$
Conductivity	σ	siemens/meter (S/m)	—	$\mathrm{m} \cdot \mathrm{kg}^{-1} \cdot \mathrm{s}^3 \cdot \mathrm{A}^2$
Energy (work)	W	joule (J)	newton-meter	$\mathrm{m}^2 \cdot \mathrm{kg} \cdot \mathrm{s}^{-2}$
Current	I	ampere (A)	—	A
Current density (volume)	$\mathbf{J}$	ampere/meter^2 (A/m^2)	—	$\mathrm{m}^{-2} \cdot \mathrm{A}$
Current density (surface)	$\mathbf{J}_s$	amper/meter (A/m)	—	$\mathrm{m}^{-1} \cdot \mathrm{A}$
Electric dipole moment	$\mathbf{p}, p$	coulomb-meter (C · m)	—	$\mathrm{m} \cdot \mathrm{s} \cdot \mathrm{A}$
Electric flux density	$\mathbf{D}$	coulomb/meter^2 (C/m^2)	—	$\mathrm{m}^{-2} \cdot \mathrm{s} \cdot \mathrm{A}$
Electric field intensity	$\mathbf{E}$	volt/meter (V/m)	newton/coulomb	$\mathrm{m} \cdot \mathrm{kg} \cdot \mathrm{s}^{-3} \cdot \mathrm{A}^{-1}$
Electric potential	V	volt (V)	—	$\mathrm{m}^2 \cdot \mathrm{kg} \cdot \mathrm{s}^{-3} \cdot \mathrm{A}^{-1}$
Energy (work)	W	joule (J)	—	$\mathrm{m}^2 \cdot \mathrm{kg} \cdot \mathrm{s}^{-2}$
Energy density	W	joule/meter^3(J/m^3)	—	$\mathrm{m}^{-1} \cdot \mathrm{kg} \cdot \mathrm{s}^{-2}$
Electromotive force	V	volt (V)	—	$\mathrm{m}^2 \cdot \mathrm{kg} \cdot \mathrm{s}^{-3} \cdot \mathrm{A}^{-1}$
Force	F	newton (N)	—	$\mathrm{m} \cdot \mathrm{kg} \cdot \mathrm{s}^{-2}$
Frequency	f	hertz (Hz)	—	s^{-1}
Impedance	Z, η	ohm (Ω)	—	$\mathrm{m}^2 \cdot \mathrm{kg} \cdot \mathrm{s}^{-3} \cdot \mathrm{A}^{-2}$
Inductance	L	henry (H)	—	$\mathrm{m}^2 \cdot \mathrm{kg} \cdot \mathrm{s}^{-2} \cdot \mathrm{A}^{-2}$
Magnetic dipole moment	$\mathbf{m}, m$	ampere-meter^2 (A · m^2)	—	$\mathrm{m}^2 \cdot \mathrm{A}$
Magnetic field Intensity	$\mathbf{H}$	ampere/meter (A/m)	gauss	$\mathrm{m}^{-1} \cdot \mathrm{A}$
Magnetic flux	Φ	weber (Wb)	tesla-meter^2	$\mathrm{m}^2 \cdot \mathrm{kg} \cdot \mathrm{s}^{-2} \cdot \mathrm{A}^{-1}$
Magnetic flux density	$\mathbf{B}$	tesla (T)	weber/meter^2	$\mathrm{kg} \cdot \mathrm{s}^{-2} \cdot \mathrm{A}^{-1}$
Magnetic vector potential	$\mathbf{A}$	tesla-meter (T · m)	weber/meter	$\mathrm{m} \cdot \mathrm{kg} \cdot \mathrm{s}^{-2} \cdot \mathrm{A}^{-1}$
Magnetization	$\mathbf{M}$	ampere/meter (A/m)	—	$\mathrm{m}^{-1} \cdot \mathrm{A}$
Magnetomotive force	V_m	ampere (A)	—	A
Permeability	μ, μ_o	henry/meter (H/m)	—	$\mathrm{m} \cdot \mathrm{kg} \cdot \mathrm{s}^{-2} \cdot \mathrm{A}^{-2}$
Permittivity	ϵ, ϵ_o	farad/meter (F/m)	—	$\mathrm{m}^{-3} \cdot \mathrm{kg}^{-1} \cdot \mathrm{s}^4 \cdot \mathrm{A}^2$
Phase	ϕ	radian	—	(dimensionless)
Phase constant	β	radian/meter	meter^{-1}	m^{-1}
Power	P	watt (W)	joule/second	$\mathrm{m}^2 \cdot \mathrm{kg} \cdot \mathrm{s}^{-3}$
Propagation constant	γ	meter^{-1} (m^{-1})	—	m^{-1}
Radiation intensity	U	watt/steradian	—	$\mathrm{m}^2 \cdot \mathrm{kg} \cdot \mathrm{s}^{-3}$
Reactance	X	ohm (Ω)	—	$\mathrm{m}^2 \cdot \mathrm{kg} \cdot \mathrm{s}^{-3} \cdot \mathrm{A}^{-2}$
Reluctance	$\mathfrak{R}$	henry^{-1} (H^{-1})	—	$\mathrm{m}^{-2} \cdot \mathrm{kg}^{-1} \cdot \mathrm{s}^2 \cdot \mathrm{A}^2$
Resistance	R	ohm (Ω)	—	$\mathrm{m}^2 \cdot \mathrm{kg} \cdot \mathrm{s}^{-3} \cdot \mathrm{A}^{-2}$
Solid angle	Ω	steradian	—	(dimensionless)
Susceptance	B	siemens (S)	ohm^{-1}	$\mathrm{m}^{-2} \cdot \mathrm{kg}^{-1} \cdot \mathrm{s}^3 \cdot \mathrm{A}^2$
Torque	T	newton-meter (N · m)	—	$\mathrm{m}^2 \cdot \mathrm{kg} \cdot \mathrm{s}^{-2}$
Velocity	u	meter/second (m/s)	—	$\mathrm{m} \cdot \mathrm{s}^{-1}$
Voltage	V, v	volt (V)	joule/coulomb	$\mathrm{m}^2 \cdot \mathrm{kg} \cdot \mathrm{s}^{-3} \cdot \mathrm{A}^{-1}$
Wavelength	λ	meter (m)	—	m
Wave number	k	meter^{-1} (m^{-1})	—	m^{-1}

TABLE A-3 LIST OF PREFIXES (MULTIPLIERS) USED WITH UNITS

Prefix	Symbol	Magnitude	Prefix	Symbol	Magnitude
Tera	T	10^{12}	Centi[†]	c	10^{-2}
Giga	G	10^{9}	Mili	m	10^{-3}
Mega	M	10^{6}	Micro	μ	10^{-6}
Kilo	k	10^{3}	Nano	n	10^{-9}
Hecto[†]	h	10^{2}	Pico	p	10^{-12}
Deca[†]	da	10^{1}	Femto	f	10^{-15}
Deci[†]	d	10^{-1}	Atto	a	10^{-18}

[†] These prefixes are usually used only for measurements of length.

Appendix B: Coordinate System Relationships and Vector Identities

TABLE B-1 RELATIONSHIPS BETWEEN COORDINATE VARIABLES IN CARTESIAN, CYLINDRICAL, AND SPHERICAL COORDINATE SYSTEMS

	=	Cartesian	Cylindrical	Spherical
Cartesian	x	x	$\rho \cos \phi$	$r \sin \theta \cos \phi$
	y	y	$\rho \sin \phi$	$r \sin \theta \sin \phi$
	z	z	z	$r \cos \theta$
Cylindrical	ρ	$\sqrt{x^2 + y^2}$	ρ	$r \sin \theta$
	ϕ	$\tan^{-1}\left\{\frac{y}{x}\right\}$	ϕ	ϕ
	z	z	z	$r \cos \theta$
Spherical	r	$\sqrt{x^2 + y^2 + z^2}$	$\frac{\rho}{\sin \theta}$	r
	θ	$\cos^{-1}\left\{\frac{z}{\sqrt{x^2 + y^2 + z^2}}\right\}$	$\tan^{-1}\left\{\frac{\rho}{z}\right\}$	θ
	ϕ	$\tan^{-1}\left\{\frac{y}{x}\right\}$	ϕ	ϕ

TABLE B-2 DOT PRODUCTS OF THE BASE UNIT VECTORS OF THE CARTESIAN, CYLINDRICAL, AND SPHERICAL COORDINATE SYSTEMS

	Cartesian			Cylindrical			Spherical		
•	$\hat{\mathbf{a}}_x$	$\hat{\mathbf{a}}_y$	$\hat{\mathbf{a}}_z$	$\hat{\mathbf{a}}_\rho$	$\hat{\mathbf{a}}_\phi$	$\hat{\mathbf{a}}_z$	$\hat{\mathbf{a}}_r$	$\hat{\mathbf{a}}_\theta$	$\hat{\mathbf{a}}_\phi$
$\hat{\mathbf{a}}_x$	1	0	0	$\cos\phi$	$-\sin\phi$	0	$\sin\theta\cos\phi$	$\cos\theta\cos\phi$	$-\sin\phi$
$\hat{\mathbf{a}}_y$	0	1	0	$\sin\phi$	$\cos\phi$	0	$\sin\theta\sin\phi$	$\cos\theta\sin\phi$	$\cos\phi$
$\hat{\mathbf{a}}_z$	0	0	1	0	0	1	$\cos\theta$	$-\sin\theta$	0
$\hat{\mathbf{a}}_\rho$	$\cos\phi$	$\sin\phi$	0	1	0	0	$\sin\theta$	$\cos\theta$	0
$\hat{\mathbf{a}}_\phi$	$-\sin\phi$	$\cos\phi$	0	0	1	0	0	0	1
$\hat{\mathbf{a}}_z$	0	0	1	0	0	1	$\cos\theta$	$-\sin\theta$	0
$\hat{\mathbf{a}}_r$	$\sin\theta\cos\phi$	$\sin\theta\sin\phi$	$\cos\theta$	$\sin\theta$	0	$\cos\theta$	1	0	0
$\hat{\mathbf{a}}_\theta$	$\cos\theta\cos\phi$	$\cos\theta\sin\phi$	$-\sin\theta$	$\cos\theta$	0	$-\sin\theta$	0	1	0
$\hat{\mathbf{a}}_\phi$	$-\sin\phi$	$\cos\phi$	0	0	1	0	0	0	1

TABLE B-3: RELATIONSHIPS BETWEEN VECTOR COMPONENTS IN THE CARTESIAN, CYLINDRICAL, AND SPHERICAL COORDINATE SYSTEMS.

=	Cartesian	Cylindrical	Spherical
A_x	A_x	$A_\rho\cos\phi - A_\phi\sin\phi$	$A_r\sin\theta\cos\phi + A_\theta\cos\theta\cos\phi$ $-A_\phi\sin\phi$
A_y	A_y	$A_\rho\sin\phi + A_\phi\cos\phi$	$A_r\sin\theta\sin\phi + A_\theta\cos\theta\sin\phi$ $+A_\phi\cos\phi$
A_z	A_z	A_z	$A_r\cos\theta - A_\theta\sin\theta$
A_ρ	$A_x\cos\phi + A_y\sin\phi$	A_ρ	$A_r\sin\theta + A_\theta\cos\theta$
A_ϕ	$-A_x\sin\phi + A_y\cos\phi$	A_ϕ	A_ϕ
A_z	A_z	A_z	$A_r\cos\theta - A_\theta\sin\theta$
A_r	$A_x\sin\theta\cos\phi + A_y\sin\theta\sin\phi$ $+ A_z\cos\theta$	$A_\rho\sin\theta + A_z\cos\theta$	A_r
A_θ	$A_x\cos\theta\cos\phi + A_y\cos\theta\sin\phi$ $- A_z\sin\theta$	$A_\rho\cos\theta - A_z\sin\theta$	A_θ
A_ϕ	$-A_x\sin\phi + A_y\cos\phi$	A_ϕ	A_ϕ

Vector Identities:

$$(\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C} = (\mathbf{B} \times \mathbf{C}) \cdot \mathbf{A} = (\mathbf{C} \times \mathbf{A}) \cdot \mathbf{B} \tag{B.1}$$

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \cdot \mathbf{C})\mathbf{B} - (\mathbf{A} \cdot \mathbf{B})\mathbf{C} \tag{B.2}$$

$$\nabla \cdot (V\mathbf{A}) = \mathbf{A} \cdot \nabla \mathbf{V} + V \nabla \cdot \mathbf{A} \tag{B.3}$$

$$\nabla \times (V\mathbf{A}) = \nabla V \times \mathbf{A} + V \nabla \times \mathbf{A} \tag{B.4}$$

$$\nabla \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot \nabla \times \mathbf{A} - \mathbf{A} \cdot \nabla \times \mathbf{B} \tag{B.5}$$

$$\nabla(\mathbf{A} \cdot \mathbf{B}) = (\mathbf{A} \cdot \nabla)\mathbf{B} + (\mathbf{B} \cdot \nabla)\mathbf{A} + \mathbf{A} \times (\nabla \times \mathbf{B}) + \mathbf{B} \times (\nabla \times \mathbf{A}) \tag{B.6}$$

$$\nabla \cdot \nabla \equiv \nabla^2 \tag{B.7}$$

$$\nabla \cdot (\nabla \times \mathbf{A}) = 0 \tag{B.8}$$

$$\nabla \times (\nabla V) = 0 \tag{B.9}$$

$$\nabla \times \nabla \times \mathbf{A} = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} \tag{B.10}$$

$$\int_V (\nabla \cdot \mathbf{A})\,dv = \oint_S \mathbf{A} \cdot \mathbf{ds} \tag{B.11}$$

$$\int_S (\nabla \times \mathbf{A}) \cdot \mathbf{ds} = \oint_C \mathbf{A} \cdot \mathbf{d\ell} \tag{B.12}$$

$$\int_V \nabla \times \mathbf{F}\,dv = -\oint_S \mathbf{F} \times \mathbf{ds} \tag{B.13}$$

Appendix C: Fundamental Constants and Material Parameters

TABLE C-1 FUNDAMENTAL PHYSICAL CONSTANTS

Physical Quantity	Symbol	Value
Electron charge	e	-1.602×10^{-19} [C]
Electron rest mass	m_e	9.107×10^{-31} [kg]
Electron radius	R_e	2.81×10^{-15} [m]
Proton rest mass	m_p	1.673×10^{-27} [kg]
Velocity of light (in a vacuum)	c	$\sim 3 \times 10^8$ [m/s]
Permittivity of free space	ϵ_o	$8.854 \times 10^{-12} \sim \frac{1}{36\pi} \times 10^{-9}$ [F/m]
Permeability of free space	μ_o	$4\pi \times 10^{-7}$ [H/m]
Intrinsic Impedance of free space	η_o	$\sim 120\pi$ or 377 [Ω]

TABLE C-2 CONDUCTIVITIES OF SOME COMMON MATERIALS AT 20° C

Material	σ, [S/m]
Aluminum	3.817×10^{7}
Carbon (diamond)	2.0×10^{-13}
Carbon (graphite)	7.143×10^{4}
Copper (commercial annealed)	5.8×10^{7}
Germanium	2.22×10^{6}
Gold	4.1×10^{7}
Iron	1.03×10^{7}
Lead (solid)	4.56×10^{6}
Mercury (liquid)	1.04×10^{6}
Magnesium	2.242×10^{7}
Nickel	1.45×10^{7}
Silicon	1.176×10^{3}
Steel (0.4–0.5 C, balance Fe)	$4.5 - 7.7 \times 10^{6}$
Tin	8.77×10^{6}
Titanium	2.09×10^{6}
Tungsten	1.825×10^{7}
Zinc	1.66×10^{7}
Water (distilled)	2×10^{-4}
Water (fresh)	1.0×10^{-3}
Seawater	4.0
Soil (dry)	10^{-5}
Rubber	2.7×10^{-11}
Polyethylene	1.5×10^{-12}

TABLE C-3 DIELECTRIC CONSTANTS, LOSS TANGENTS, AND DIELECTRIC STRENGTHS OF SOME COMMON DIELECTRICS

Material	Dielectric Constant ϵ_r	Loss Tangent $\tan \phi = \epsilon''/\epsilon_r$			Dielectric Strength in [MV/m]
		10^3 Hz	10^6 Hz	10^8 Hz	
Air (atmospheric pressure)	1.0006	≈0	≈0	≈0	3
Ice (fresh or salt)	3.3–4.2*	—	0.12	0.035	—
Glass	4–7	0.005	0.004	0.003	30
Mica	7.45	0.0019	0.0013	—	200
Paper	2–4	0.008	0.04	0.07	12
Polyethylene	2.26	<0.0002	<0.0002	0.0002	47
Polystyrene	2.56	<0.00005	<0.00005	0.00007	20–30
Polytetrafluoroethylene (teflon)	2.1	<0.0005	<0.0003	<0.0002	20
Porcelain	5.5	0.014	0.0075	.0078	4
Seawater	72	—	0.9	—	—
Silicone–rubber	3.35	0.0067	0.003	0.0032	10–40
Snow	1.2-3.3*	0.49	0.0215	—	—
Soil, sandy dry	2.91	0.08	0.017	—	—
Soil, loamy dry	2.83	0.05	0.018	—	—
Water (distilled)	80	$<10^{-6}$	0.04	0.005	—
Wood (Douglas fir)	2.1	0.008	0.026	0.033	10

* Decreases with increasing frequency.

The relative permittivity ϵ_r of most materials is relatively constant from dc through the rf and microwave frequency ranges. Ice and water are notable exceptions. For them ϵ_r decreases with increasing frequency in the rf region.

TABLE C-4 RELATIVE PERMEABILITIES OF SOME COMMON MAGNETIC MATERIALS

Material	Relative Permeability μ_r
Ferromagnetics	
Iron	4,000
Permalloy	70,000
Supermalloy	1,000,000
Permendur	5,000
Cobalt	600
Ferrimagnetics	
Manganese–zinc ferrite	750
Nickel–zinc ferrite	650
Diamagnetics	
Bismuth	0.999833
Mercury	0.999968
Copper	0.9999906
Water	0.9999912
Paramagnetics	
Air	1.00000037
Tungsten	1.00008
Manganese	1.001

Appendix D: Transmission-Line Parameters

The most fundamental parameters of a transmission line are its inductance L, capacitance C, resistance R, and conductance G. These parameters are distributed on a per-unit-length basis along the length of the transmission line and are functions of the cross-sectional dimensions of the transmission line, the materials used, and the operating frequency (or bandwidth). From these fundamental parameters, the following operating parameters can be determined:

$$\text{Characteristic impedance: } Z_0 = R_0 + jX_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \quad [\Omega] \tag{D.1}$$

$$\text{Phase constant: } \beta = \mathrm{Im}[\sqrt{(R + j\omega L)(G + j\omega C)}] \quad [\mathrm{m}^{-1}] \tag{D.2}$$

$$\text{Attenuation constant: } \alpha = \mathrm{Re}[\sqrt{(R + j\omega L)(G + j\omega C)}] \quad [\mathrm{Np/m}] \tag{D.3}$$

$$\text{Wavelength: } \lambda = \frac{2\pi}{\beta} \quad [\mathrm{m}] \tag{D.4}$$

$$\text{Phase velocity: } u_p = \frac{\omega}{\beta} \quad [\mathrm{m/s}] \tag{D.5}$$

$$\text{Group velocity: } u_g = \frac{\partial\omega}{\partial\beta} \quad [\mathrm{m/s}] \tag{D.6}$$

All of the preceding operating parameters are interrelated, so it is common to characterize a transmission line using just a few parameters. The parameters most often chosen are the characteristic impedance Z_o, the phase velocity u_p, and the attenuation constant α. Formulas for these parameters have been derived for a wide range of different types of transmission lines, and the reader can find exhaustive presentations of these formulas in the references cited at the end of this appendix.

Most transmission lines fall into one of two classes. The first and simplest are transverse electromagnetic (TEM) lines. These are transmission lines whose dominant mode is a TEM wave. In the strictest sense, only transmission lines with a homogeneous dielectric and no losses can support a TEM mode. However, even when loss is present, the dominant modes on transmission lines with uniform dielectrics are so close to being TEM that they are generally classed as TEM lines. Examples of TEM lines are coaxial cables, two-wire lines (with uniform dielectrics), and strip lines. On TEM lines, the phase velocity u_p is governed only by the dielectric constant of the dielectric. If the dielectric is nonmagnetic,

$$u_p = \frac{c}{\sqrt{\epsilon_r}} \qquad \text{(TEM transmission lines)}, \tag{D.7}$$

where c is the speed of light in a vacuum (3×10^8 [m/s]) and $\epsilon_r = \mathrm{Re}(\epsilon/\epsilon_o)$ is the dielectric constant.

The second class of transmission lines consists of those with nonuniform dielectrics. These are called quasi-TEM lines, since their dominant modes are nearly TEM, but always have at least one longitudinal component when the lines are operated above dc. These lines can be characterized by the same types of operating parameters as for TEM lines. But unlike TEM lines, where the phase velocity is controlled only by the dielectric constant of the dielectric, the phase velocity of quasi-TEM lines is controlled not only by the dielectric constants of the materials used, but also by how they are configured. An effective dielectric constant ϵ_{eff} is typically used to specify the velocity of propagation:

$$u_p = \frac{c}{\sqrt{\epsilon_{\text{eff}}}} \qquad \text{(Quasi-TEM transmission lines)}. \tag{D.8}$$

Examples of quasi-TEM transmission lines are microstrip transmission lines and slot line transmission lines.

The attenuation constant α of both classes of transmission lines can usually be expressed as the sum of a dielectric loss constant α_d and a conductor loss constant α_c:

$$\alpha = \alpha_d + \alpha_c, \tag{D.9}$$

For all low-loss TEM lines, α_d is given by[1]

$$\alpha_d = \frac{\omega\epsilon''}{2\epsilon'}\sqrt{\mu\epsilon'} \qquad \text{(Low loss TEM transmission lines)}, \tag{D.10}$$

[1] This expression was derived in Chapter 12 (see Equation (12.79)) for the case of a plane wave, but it is applicable to any TEM wave in a homogeneous dielectric.

where ϵ' and $-\epsilon''$ are the real and imaginary parts of the complex permittivity, respectively. (See Sections 10-4-3 and 12-5 for a discussion of the complex permittivity.) Formulas for α_d on quasi-TEM lines must be derived for the specific geometry of the line.

Formulas for α_c always involve the cross sectional dimensions of the line. Also, these formulas always involve the conductor skin depth δ, which is given by

$$\delta = \frac{1}{\sqrt{\pi f \mu \sigma}}, \tag{D.11}$$

where f is the frequency of operation, and μ and σ are the permeability and conductivity of the conductors. (Equation (D.11) is derived in Section 12-5.)

The remainder of this appendix presents formulas for the operating parameters of a number of common types of transmission lines.

Coaxial Cables. Coaxial cables are the most popular type of TEM transmission line. They are particularly attractive for their shielding properties, which makes them resistant to outside sources of interference. Figure D-1 shows the geomtery of a coaxial cable.

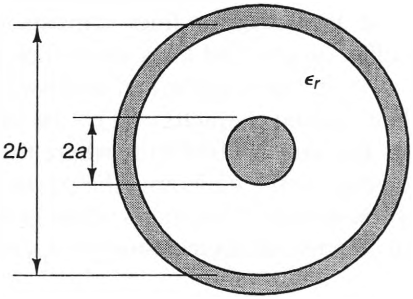


Figure D-1 A coaxial cable.

The characteristic impedance and conductor attenuation constant are respectively given by

$$Z_o = \frac{1}{2\pi}\sqrt{\frac{\mu_o}{\epsilon'}}\ln\left(\frac{b}{a}\right) \qquad [\Omega] \tag{D.12}$$

and

$$\alpha = \frac{\omega\epsilon''}{2\epsilon'}\sqrt{\mu\epsilon'} + \frac{2.11\times 10^{-4}}{\sigma\delta}\left(\frac{1}{a}+\frac{1}{b}\right) \qquad [\text{Np/m}], \tag{D.13}$$

where δ is the skin depth of the conductors. Equation (D.12) is derived using the per-unit-length capacitance and inductance formulas derived in Chapters 6 and 9, respectively. Equation (D.13) can be derived using Equation (11.115) and the surface resistance of the conductors.

Two-Wire Lines. In two-wire transmission lines, both conductors have the same relationship to ground: so they are balanced transmission lines. This makes them attractive for feeding balanced loads, such as dipole antennas or the inputs of balanced amplifiers. Although they are not shielded, it is often possible to limit their inductive coupling to outside sources of interference by twisting them.

Figure D-2 shows a two-wire transmission line.

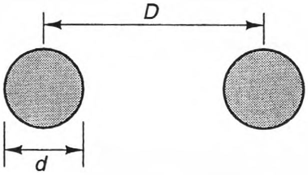


Figure D-2 A two-wire transmission line.

Expressions for the characteristic impedance and conductor attenuation constant are respectively given by (Wadell p. 66)

$$Z_o = \frac{1}{\pi}\sqrt{\frac{\mu_o}{\epsilon'}}\cosh^{-1}\left(\frac{D}{d}\right) \qquad [\Omega] \tag{D.14}$$

and

$$\alpha = \frac{\omega\epsilon''}{2\epsilon'}\sqrt{\mu\epsilon'} + \frac{7.48\times 10^{-4}}{d\sigma\delta}\times\frac{D/d}{\sqrt{(D/d)^2-1}}\times\frac{1}{\cosh^{-1}\left(\frac{D}{d}\right)} \qquad [\text{Np/m}], \tag{D.15}$$

where δ is the skin depth of the conductors.

Strip Lines. Strip lines are common in microwave circuits. Like coaxial cables, they are shielded, but the losses on strip lines are usually lower, since the dielectric is usually air. Figure D-3 shows the cross section of a typical strip line. Approximate expressions for the characteristic impedance Z_o (Wadel, p. 126–128),

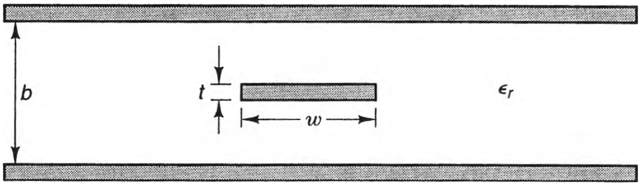


Figure D-3 A strip line transmission line.

$$Z_o = \frac{377}{2\pi\sqrt{\epsilon_r}}\ln\left\{1 + \frac{4b}{\pi w'}\left[\frac{8b}{\pi w'} + \sqrt{\left(\frac{8b}{\pi w'}\right)^2 + 6.27}\right]\right\} \qquad [\Omega] \tag{D.16}$$

where

$$w' = w + \frac{t}{3.2}\ln\left(\frac{5b}{t}\right). \tag{D.17}$$

This approximation is valid for

$t/b < 0.1$

$w/t > 2.5$

$w/b > 0.1$.

Microstrip Lines. Microstrip transmission lines are the most popular type of transmission used on printed circuit boards. These transmission lines are very easy to fabricate using automated processes and offer relatively attractive electrical properties. Figure D-4 shows a typical microstrip transmission line.

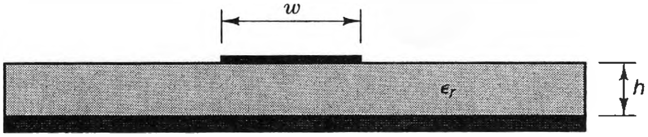


Figure D-4 A microstrip transmission line.

Approximate expressions for the effective dielectric constant and the characteristic impedance are as follows (Gardiol pp. 92–93):

For $\frac{w}{h} \leqslant 1$,

$$\epsilon_{\text{eff}} \approx \frac{1}{2}(\epsilon_r + 1) + \frac{1}{2}(\epsilon_r - 1)\left[\left(1 + 12\frac{h}{w}\right)^{-1/2} + 0.04\left(1 - \frac{w}{h}\right)^2\right] \quad \text{(D.18)}$$

$$Z_o \approx \frac{60}{\sqrt{\epsilon_{\text{eff}}}} \ln\left(8\frac{h}{w} + \frac{w}{4h}\right). \quad \text{(D.19)}$$

For $\frac{w}{h} > 1$,

$$\epsilon_{\text{eff}} \approx \frac{1}{2}(\epsilon_r + 1) + \frac{1}{2}(\epsilon_r - 1)\left(1 + 12\frac{h}{w}\right)^{-1/2} \quad \text{(D.20)}$$

$$Z_o \approx \frac{120\pi}{\sqrt{\epsilon_{\text{eff}}}}\left[\frac{w}{h} + 1.393 + 0.667\ln\left(\frac{w}{h} + 1.444\right)\right]^{-1}. \quad \text{(D.21)}$$

Formulas have also been derived that predict the w/h ratio necessary to achieve a particular value of Z_o (Wheeler: see Gardiol, p. 93).

For $\frac{w}{h} \leqslant 2$,

$$\frac{w}{h} \approx 4\left[\frac{1}{2}\exp(A) - \exp(-A)\right]^{-1}, \quad \text{(D.22)}$$

where

$$A = \frac{\pi Z_o}{377}\sqrt{2(\epsilon_r + 1)} + \frac{\epsilon_r - 1}{\epsilon_r + 1}(0.23 + 0.11/\epsilon_r) \quad \text{(D.23)}$$

Similarly, for $\dfrac{w}{h} \geqslant 2$,

$$\frac{w}{h} \approx \frac{\epsilon_r - 1}{\pi\epsilon_r}\left(\ln\,(B-1) + 0.39 - \frac{0.61}{\epsilon_r}\right) + \frac{2}{\pi}\left(B - 1 - \ln(2B-1)\right), \tag{D.24}$$

where

$$B = \frac{\pi}{2\sqrt{\epsilon_r}}\frac{377\ [\Omega]}{Z_o}. \tag{D.25}$$

Slot Lines. Figure D-5 shows a typical slot line, which consists of two metal traces on the same side of a dielectric slab. Slot lines are attractive on printed circuit boards when it is not convenient to have a conducting sheet on the bottom of the dielectric (such as for microstrip lines).

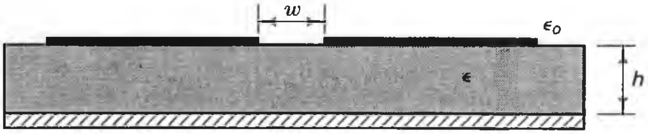


Figure D-5 A slot line transmission line.

Approximate expressions for ϵ_{eff} and Z_o are as follows:

For $0.02 < w/h < 0.2$,

$$(\epsilon_{\text{eff}})^{-1} \approx [0.923 - 0.448\log\epsilon_r + 0.2w/h \\ - (0.29w/h + 0.047)\log(h/\lambda_o \times 10^2)]^{-2} \tag{D.26}$$

$$Z_o \approx 72.62 - 35.19\log\epsilon_r + \frac{50(w/h - 0.02)\,(w/h - 0.1)}{w/h} \\ + \log(w/h \times 10^2)[44.28 - 19.58\log\epsilon_r] \\ -[0.32\log\epsilon_r - 0.11 + w/h(1.07\log\epsilon_r + 1.44)] \\ \times [11.4 - 6.07\log\epsilon_r - h/\lambda_o \times 10^2]^2. \tag{D.27}$$

For $0.2 < w/h < 1.0$,

$$(\epsilon_{\text{eff}})^{-1} \approx 0.987 - 0.483\log\epsilon_r + \frac{w}{h}(0.111 - 0.0022\epsilon_r) \\ -\left(0.121 + \frac{0.094w}{h} - 0.0032\epsilon_r\right)\log(h/\lambda_o \times 10^2) \tag{D.28}$$

$$
\begin{aligned}
Z_0 \approx\ & 113.19 - 53.55 \log \epsilon_r + 1.25w/h(114.59 - 51.88 \log \epsilon_r) \\
& + 20.0(w/h - 0.2)(1.0 - w/h) \\
& - [0.15 + 0.23 \log \epsilon_r + w/h(-0.79 + 2.07 \log \epsilon_r)] \\
& \times [10.25 - 5\epsilon_r + w/h(2.1 - 1.42 \log \epsilon_r) - h/\lambda \times 10^2]^2,
\end{aligned}
\tag{D.29}
$$

where "log" signifies the base-10 logarithm. These expressions are valid for $9.7 < \epsilon_r < 20.0$ and $0.01 < h/\lambda < 0.25/\sqrt{\epsilon_r - 1.0}$. (See R. Garg and K. C. Gupta, pp. 156–161.)

References

[1] B. C. Wadell. *Transmission Line Design Handbook*. Norwood, MA: Artech House, Inc., 1991.

[2] K. C. Gupta, R. Garg, & I. J. Bahl. *Microstrip Lines and Slotlines*. Dedham, MA: Artech House, Inc., 1979.

[3] *Reference Data for Radio Engineers*. Howard W. Sams Co., 1975. New York.

[4] R. Garg, & K. C. Gupta. "Expressions for Wavelength and Impedance of a Slotline," *IEEE Trans. Microwave Theory and Tech.*, vol. MTT-24, August, 1976.

[5] F. E. Gardiol. *Introduction to Microwaves*. Dedham, MA: Artech Hours, Inc., 1984.

[6] H. A. Wheeler. "Transmission Line Properties of Parallel Strips Separated by a Dielectric Sheet," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-13, No. 3, March 1965, pp. 172–185.

Appendix E: Answers to Selected Problems

CHAPTER 2

2-2 a) $\sqrt{14}$, b) $\sqrt{17}$, c) $-3\hat{\mathbf{a}}_x$, d) $-6\hat{\mathbf{a}}_y - 9\hat{\mathbf{a}}_z$, e) $\theta_{\min} = 44.52°$.

2-5 a) $\mathbf{A} = (2\cos\phi - 3\sin\phi)\hat{\mathbf{a}}_\rho + (-2\sin\phi - 3\cos\phi)\hat{\mathbf{a}}_\phi + 2\hat{\mathbf{a}}_z$, b) at P_1, $\mathbf{A} = -1.6\hat{\mathbf{a}}_\rho - 3.2\hat{\mathbf{a}}_\phi + 2\hat{\mathbf{a}}_z$, at P_2, $\mathbf{A} = 0.23\hat{\mathbf{a}}_\rho - 3.6\hat{\mathbf{a}}_\phi + 2\hat{\mathbf{a}}_z$.

2-7 $\mathbf{C} = 6/\sqrt{2}\,\hat{\mathbf{a}}_2 + \hat{\mathbf{a}}_3$.

2-10 volume $= \dfrac{1}{3}abh$.

2-13 a) 4.5, b) 25/6.

2-15 $\dfrac{8}{\sqrt{3}}$.

2-18 a) 7.48, b) −2.41

2-20 a) $\nabla f = 2y\hat{\mathbf{a}}_x + 2x\hat{\mathbf{a}}_y$, b) $\nabla f = 2\rho\sin 2\phi\hat{\mathbf{a}}_\rho + 2\rho\cos 2\phi\hat{\mathbf{a}}_\phi$, c) proof

2-23 a) −4, b) −4

2-25 a) $\nabla f = \sin\theta\cos\phi\hat{\mathbf{a}}_r + \cos\theta\cos\phi\hat{\mathbf{a}}_\theta - \sin\phi\hat{\mathbf{a}}_\phi$, b) $\nabla \times \nabla f = 0$, c) $\nabla \cdot \nabla f = 0$.

2-26 a) 6π, b) 6π.

CHAPTER 3

3-1 $Q = 8.73\ [\mu\text{C}]$.

3-4 $I = 1/6\ [\text{A}]$.

3-6 $\rho_v = -2t + Const\ [\text{mC/m}^3]$

3-8 $\mathbf{dF}_1 = 6.93 \times 10^{-9}\ \hat{\mathbf{a}}_x\ [\text{N}]$

3-10 $\mathbf{E} = 693.3\hat{\mathbf{a}}_x + 50\hat{\mathbf{a}}_y - 500\hat{\mathbf{a}}_z\ [\text{V/m}]$

3-12 $u = 0.918 \times 10^8\ [\text{m/s}]$.

CHAPTER 4

4-2 $\mathbf{E} = 0.575\hat{\mathbf{a}}_x - 0.575\hat{\mathbf{a}}_y + 1.47\hat{\mathbf{a}}_z\ [\text{kV/m}]$.

4-4 $\mathbf{E} = \dfrac{-a^2\rho_{\ell o}}{4\epsilon_o[z^2 + a^2]^{3/2}}\,\hat{\mathbf{a}}_x$.

4-6 $|z| \leqslant \dfrac{a}{19.975}$

4-8 $\mathbf{E} = \dfrac{1 - e^{-\rho}}{\rho\epsilon_o}\,\hat{\mathbf{a}}_\rho$

4-13 a) $\mathbf{E} = \dfrac{\rho_{vo}[1 - (r + 1)e^{-r}]}{\epsilon_o r^2}\,\hat{\mathbf{a}}_r$, b) $\lim\limits_{r\to\infty} \mathbf{E} = \dfrac{\rho_{vo}}{\epsilon_o r^2}\,\hat{\mathbf{a}}_r$, which is the same as a point charge at the origin.

4-16 $V_{ab} = 4.4\ [\text{V}]$

4-18 $V_{ab} = \dfrac{d}{2\epsilon_o}(\rho_{sa} - \rho_{sb})$

CHAPTER 5

5-2 $N_p = 10^7\ [\text{cm}^{-3}]$.

5-5 $P_{\text{dis}} = \dfrac{\sigma E_o^2}{2\alpha}\ [\text{W/m}^2]$.

5-8 a) $3.4 \times 10^{-6}\%$, b) $3.4 \times 10^{-3}\%$.

5-12 $R = \dfrac{b - a}{2\pi\ell k}$.

5-15 $\epsilon_r = 1.56$.

5-17 a) $\rho_s = \dfrac{-2Qd}{4\pi[x^2 + y^2 + d^2]^{3/2}}$, b) $Q_{\text{surface}} = -Q$.

5-19 $\mathbf{E} = \left(\dfrac{\gamma}{\epsilon_o} - \dfrac{A}{\rho}\right)\hat{\mathbf{a}}_\rho$, where $A = \dfrac{\dfrac{\gamma}{\epsilon_o}(b - a) + V_o}{\ln\dfrac{b}{a}}$.

CHAPTER 6

6-2 $C = 12.1$ [pF/m].
6-4 $C = 473$ [pF/m].
6-10 $\sigma = 1.77 \times 10^{-6}$ [S/m].
6-13 a) $V_2 = 0$, b) $V_2 = 99$ [mV].

CHAPTER 7

7-4 $\mathbf{B} = \mu_o J_\phi \hat{\mathbf{a}}_z$ for $\rho < a$, $\mathbf{B} = \dfrac{\mu_o a J_z}{\rho} \hat{\mathbf{a}}_\phi$ for $\rho > a$.

7-6 $\mathbf{B} = \dfrac{4\mu_o I}{\pi w \sqrt{2}} \hat{\mathbf{a}}_z$.

7-9 $\mathbf{B} = \dfrac{2 \times 10^{-9}}{\rho} \hat{\mathbf{a}}_\phi$ [T], where ρ is specified in meters.

7-12 a) $400 \hat{\mathbf{a}}_\phi$ [μT], b) $-23.5 \hat{\mathbf{a}}_\phi$ [μT].

7-14 $\mathbf{B} = \dfrac{\mu_o d J_o}{2} \hat{\mathbf{a}}_y$ inside hole.

7-15 $\mathbf{A} = \dfrac{\mu_o I\ell}{4\pi r} \hat{\mathbf{a}}_z$, $\mathbf{B} = \dfrac{\mu_o I\ell}{4\pi r^2} \sin\theta \hat{\mathbf{a}}_\phi$.

7-18 $V_m = \dfrac{1}{\pi\mu_o} \cos\dfrac{\pi x}{a} \cos\dfrac{\pi y}{b}$.

CHAPTER 8

8-1 $\rho_v = 10^5$ [C/m^3].
8-4 $m = 5.2 \times 10^{-22}$ [A • m^2].
8-7 $|B_2| = 0.15$ [T], $\theta_2 = 86.2°$.
8-9 $\mathbf{B} = 2 \times 10^4 \mu_o I \hat{\mathbf{a}}_z$ for $\rho < a$, $\mathbf{B} = 2 \times 10^4 \mu_r \mu_o I \hat{\mathbf{a}}_z$ for $a < \rho < b$,

$\mathbf{B} = 0$ for $\rho > 0$, $\Phi = 2\pi I \times 10^4 \mu_o \left[a^2 + \mu_r(b^2 - a^2)\right]$.

8-15 $\mathbf{M} = 1.91 \times 10^7 \hat{\mathbf{a}}_z$ [A/m], and $\mathbf{J}_{sm} = 1.91 \times 10^7 \hat{\mathbf{a}}_\phi$ [A/m].
8-16 $\Phi_{\text{left}} = 10.5$ [μWb] and $\Phi_{\text{right}} = 65.8$ [μWb], both upward.
8-18 $B = 1.9$ [T].

CHAPTER 9

9-2 $V_m = -8.33 \cos \omega t$ [V].

9-4 a) $P_{\text{ave}} = \dfrac{\pi a^4 L \sigma (\omega B_o)^2}{16}$ [W], $P_{\text{lam}} = \dfrac{\pi a^4 (0.9)^2 L \sigma (\omega B_o)^2}{16 N^2}$ [W],

c) $\textit{Ratio} = \dfrac{(0.9)^2}{N}$.

9-8 $i(t) = 0.24 \sin \omega t$ [mA].

9-11 $L_{12} = 0.158$ [mH].

9-14 $F_m = 15.5$ [mN] (attractive).

9-16 $F_m = 27.4$ [N] when $x = 0.7$ [cm], and $F_m = 13.4$ [N] when $x = 1.0$ [cm]

CHAPTER 10

10-1 $\epsilon = \dfrac{2}{\mu_o(\alpha x)^2}$

10-3 a) $\mathbf{H} = \dfrac{\beta E_o}{\omega \mu r}\sin(\omega t - \beta r)\hat{\mathbf{a}}_\phi$, b) $\beta = \omega\sqrt{\mu\epsilon}$, c) an $\hat{\mathbf{a}}_r$ must be present when βr is small.

10-4 a) $J_c/J_d = 1.5 \times 10^7$, b) $J_c/J_d = 0.562$.

10-7 $\mathbf{H}_o = \mathbf{H}_o \cos(k_x x)e^{-j\beta z}\hat{\mathbf{a}}_z$.

10-10 $\mathbf{J}_d = -j\beta H_o \cos(k_x x)e^{-j\beta z}\hat{\mathbf{a}}_y$

10-13 $\mathbf{E}_2 = 1.52\hat{\mathbf{a}}_x + 0.04\hat{\mathbf{a}}_y - 3\hat{\mathbf{a}}_z$ [V/m]

$\mathbf{H}_2 = -1.58\hat{\mathbf{a}}_x + 0.83\hat{\mathbf{a}}_y - 4\hat{\mathbf{a}}_z$ [A/m]

CHAPTER 11

11-2 a) $C = 91.4$ [pF/m], b) $L = 275$ [nH/m], c) $R_o = 54.9$ [Ω], d) $u = 1.99 \times 10^8$ [m/s].

11-5 $V_{\text{in}} = 33.3$ [V], $I_{\text{in}} = 0.66$ [A].

11-7 $\dfrac{w}{h} = 2.26$.

11-12 $v_L = 0$ for $t < 4$ [ns], $v_L = 3.5$ [V] for $4 < t < 12$ [ns], $v_L = 5$ [V] for $t > 12$ [ns]. The steady state value is $v_L = 5$ [V].

11-14 $v_{\text{ref}}(t) = [2.5 + 2.5e^{-(t-T)/\tau}]U(t - 2T)$ [mV], where $\tau = 6$ [ns] and $T = 10.3$ [ns].

11-16 $u_p = \dfrac{c}{a + b\omega}$, $u_g = \dfrac{c}{a + 2b\omega}$

11-18 a) $Z_L = 174.76 \angle 23.8°$ [Ω], b) $|v_L| = 14.25$ [V].

11-21 $P_{75} = 13.3$ [W], $P_{25} = 9.97$ [W], $P_{100} = 39.9$ [W].

11-25 $Z_L = 42.83 - j27.99$ [Ω].

11-28 $Z_L = 91.12 - j15.0$ [Ω].

11-30 $Z_{\text{in}} = 80.88$ [Ω].

CHAPTER 12

12-1 $f = 100$ [MHz], $\mathbf{H} = \dfrac{-3}{\eta_o}e^{-j(2.094y + \pi/4)}\hat{\mathbf{a}}_z$ [A/m],

$\mathbf{E}(t) = 3\cos(\omega t - 2.094y - \pi/4)\hat{\mathbf{a}}_x$, $\mathbf{H}(t) = \dfrac{-3}{\eta_o}\cos(\omega t - 2.094y - \pi/4)\hat{\mathbf{a}}_z$.

12-2 $f = 150$ [MHz], $\lambda = 2$ [m].

12-4 $\mathbf{E} = -1.69(2\hat{\mathbf{a}}_x + \hat{\mathbf{a}}_z)\cos\left(\omega t - \frac{k_o}{\sqrt{5}}x - \frac{2k_o}{\sqrt{5}}z\right)$ [kV/m],

$\mathbf{H} = 10\hat{\mathbf{a}}_y\cos\left(\omega t - \frac{k_o}{\sqrt{5}}x - \frac{2k_o}{\sqrt{5}}z\right)$ [A/m].

12-8 AR = 9.64, $\tau = 17.58°$.

12-12 $\epsilon = \epsilon_o[2.5 - j0.025]$.

12.15 $\epsilon = \epsilon_o[4.99 - j2.45 \times 10^{-7}]$.

12-18 a) $P_{\text{diss}} = \frac{J_o^2 d}{\sigma}$ [W/m^2], b) $P_{\text{diss}} = \frac{J_o^2 d}{\sigma}$ [W/m^2].

12-19 $t = 57.1$ [s].

12-22 $P_{\text{diss}} = 2.1$ [mW/m^2].

12-25 $\ell = 0.049$ [mm].

12-27 AR = 2.2.

12-30 $h = 0.3$ [m].

CHAPTER 13

13-4 $\text{TE}_{10}, \text{TE}_{20}, \text{TE}_{11}, \text{TM}_{11}, \text{TE}_{21}, \text{TM}_{21}, \text{TE}_{01}$

13-6 $\alpha = 928.6$ [dB/m].

13-9 $d = 370.8$ [m].

13-12 a) $\alpha = 6 \times 10^{-2}$ [dB/m], b) $\alpha = 4.5 \times 10^{-2}$ [dB/m].

13-14 $\ell = 1.75$ [cm], $d = 0.659$ [cm].

13-17 $\Delta\tau = 0.114$ [ns/m].

13-20 $d_{\text{max}} = 3.3218$ [μm].

13-22 $d = 4.81$ [mm].

13-24 a) 31.45%, b) 1.89%, c) 0.019%.

13-26 $P >$ 2.66 [mW].

13-30 a) $d = 2.36$ [cm], b) $d = 1.925$ [mm].

CHAPTER 14

14-3 $E = 62.5$ [mV/m].

14-6 $P_{\text{rad}} = 8.89$ [mW].

14-8 $R_{\text{in}} = 67.02$ [Ω].

14-11 $R_{\text{in}} = 146$ [Ω].

14-13 a) $\eta_r = 1.04\%$, b) $\eta_r = 99.76\%$.

14-17 $N_{\text{min}} = 24$.

14-18 $30.85° < \theta < 65.38°$.

14-20 Nulls occur at $\theta = \pm 45.57°, \pm 66.42°, \pm 84.26°, \pm 101.5°, \pm 120°, \pm 143.1°$.

14-26 $V_{oc} = 11.2$ [μV].
14-29 $P_r = 179$ [mW].
14-32 $R_{max} = 193.6$ [km].
14-34 $\sigma_2 = 100\sigma_1$.

Selected Bibliography

In addition to the references cited throughout the text, the following works may be helpful in understanding the material.

INTRODUCTORY

Aahn, Markus. *Electromagnetic Field Theory, a Problem Solving Approach*, New York: John Wiley & Sons, 1979.

Coren, Richard L. *Basic Engineering Electromagnetics, an Applied Approach*. New Jersey: Prentice-Hall, Inc., Englewood Cliffs, 1989.

Frankl, Daniel R. *Electromagnetic Theory*. Prentice-Hall, Inc. New Jersey: Englewood Cliffs, 1986.

Hayt, William H. *Engineering Electromagnetics*, 3d ed. New York: McGraw-Hill, Inc, 1974.

Johnk, Carl T. A. *Engineering Electromagnetic Fields and Waves*, 2d ed. New York: John Wiley & Sons, 1988.

Kraus, John D. *Electromagnetics*, 4th ed. New York: McGraw-Hill, Inc. 1992.

Marshall, Stanley V., & Skitek, Gabrial G. *Electromagnetic Concepts & Applications*, 2d ed. Prentice-Hall, Inc., New Jersey: Englewood Cliffs, 1987.

Neff, Herbert P., Jr. *Basic Electromagnetic Fields*, 2d ed. New York: Harper & Row, Publishers, 1987.

Neff, Herbert P., Jr. *Introductory Electromagnetics*. New York: John Wiley & Sons, 1991.

Paul, Clayton R., & Nasar, Syed A. *Introduction to Electromagnetic Fields*. New York: McGraw-Hill, 1982.

Rao, N. Narayana. *Elements of Engineering Electromagnetics.* New Jersey: Prentice-Hall, Inc., Englewood Cliffs, 1977.

Sadiku, Matthew N.O. *Elements of Electromagnetics.* Fort Worth: Saunders College Publishing, 1989.

Schwarz, Steven E. *Electromagnetics for Engineers.* Philadelphia: Saunders College Publishing, 1990.

Setian, Leo. *Engineering Field Theory with Applications.* Cambridge: Cambridge University Press, 1992.

Shen, Liang C., & Kong, Jin A. *Applied Electromagnetism*, 3d ed. Boston: PWS Publishing Company, an International Thomson Publishing Company, 1987.

INTERMEDIATE

Bohn, Eric V. *Introduction To Electromagnetic Fields and Waves.* Reading, MA: Addison-Wesley Publishing Company, 1968.

Della Torre, Edward, & Longo, Charles V. *The Electromagnetic Field.* Boston: Allyn and Bacon, Inc., 1969.

Ferraro, Vincenzo C. A. *Electromagnetic Theory.* London: Athlone Press, 1954.

Haus, Hermann, A., & Melcher, James R. *Electromagnetic Fields and Energy.* Englewood Cliffs, NJ: Prentice-Hall, 1989.

Jordan, Edward C., & Balmain, Keith G. *Electromagnetic Waves and Radiating Systems*, 2d ed. Englewood Cliffs, New Jersey: Prentice-Hall, 1968.

Lorrain, Paul, & Corson, Dale R. *Electromagnetism, Principles and Applications.* New York: W. H. Freeman and Company, 1990.

Plonsey, Robert, & Collin, Robert E. *Principles and Applications of Electromagnetic Fields.* New York: McGraw-Hill, 1961.

Ramo, Simon, & Whinnery, John R. Van Duzer, Theodore, *Fields and Waves in Communication Electronics.* New York: John Wiley & Sons, Inc., 1965.

Seely, Samuel, & Poularikas, Alexander D. *Electromagnetics, Classical and Modern Theory and Applications.* New York: Marcel Dekker, Inc., 1979.

Wangsness, Roald K. *Electromagnetic Fields*, 2d ed., New York: John Wiley & Sons, 1986.

ADVANCED

Balanis, Constantine A. *Advanced Engineering Electromagnetics.* New York: John Wiley & Sons, 1989.

Harrington, Roger F. *Time-Harmonic Electromagnetic Fields.* New York: Mc-Graw Hill, 1961.

Ishimaru, Akira. *Electromagnetic Wave Propagation, Radiation, and Scattering.* Englewood Cliffs, NJ: Prentice-Hall, 1991.

Jackson, John D. *Classical Electrodynamics.* New York: John Wiley & Sons, Inc., 1975.

Jones, D. S. *The Theory of Electromagnetism.* New York: Pergamon Press, 1964.

Kong, Jin A. *Electromagnetic Wave Theory.* New York: John Wiley & Sons, 1986.

Mason, Max, & Weaver, Warren. *The Electromagnetic Field.* Chicago: The University of Chicago Press, 1929.

Maxwell, James Clerk. *A Treatise on Electricity and Magnetism*, vols. 1 and 2, NY: Dover Publications, 1954. (Originally published in 1873).

Schwartz, Melvin. *Principles of Electrodynamics.* New York: McGraw-Hill, 1972.

Smythe, W. R. *Static and Dynamic Electricity*, 3d ed. New York: McGraw-Hill, 1968.

Stratton, Julius A. *Electromagnetic Theory.* New York: McGraw-Hill, 1941.

Van Bladel, J. *Electromagnetic Theory.* New York: McGraw-Hill, 1964.

SELECTED TOPICS

Amerasekera, Ajith, & Duvvury, Charvaka. *ESD In Silicon Integrated Circuits.* New York: John Wiley & Sons, 1995.

Baden Fuller, A. J. *Microwaves.* 2d ed. Oxford: Pergamon Press, 1979.

Bahl, I. J., & Bhartia, P. *Microstrip Antennas.* Dedham, MA: Artech House, 1980.

Balanis, Constantine A. *Antenna Theory.* New York: Harper & Row, 1982.

Boll, Richard. *Soft Magnetic Materials.* Munich and Berlin: Siemens-Aktiengesellschaft, 1979.

Booton, Richard C., Jr. *Computational Methods for Electromagnetics and Microwaves.* New York: John Wiley & Sons, 1992.

Burke, Harry. *Handbook of Magnetic Phenomena.* New York: Van Nostrand Reinhold Co., 1986.

Chang, Jen-Shih, Kelly, Arnold, & Crowley, Joseph. *Handbook of Electrostatic Processes.* New York: Marcel Dekker, Inc., 1995.

Collin, Robert E. *Field Theory of Guided Waves.* 2d ed. New York: IEEE Press, 1991.

Collin, Robert E. *Foundations of Microwave Engineering*, 2d ed. New York: McGraw Hill, 1992.

Diament, Paul. *Wave Transmission and fiber Optics.* New York: Macmillan Publishing Company, 1990.

Green, Paul E., Jr. *Fiber Optic Networks.* Prentice Hall, NJ: Englewood Cliffs, 1993.

Gupta, K. C., Garg, Ramesh I., & Bahl, I. J. *Microstrip Line and Slotlines*, Dedham, MA: Artech House, 1979.

Heck, Carl. *Magnetic Materials and their Applications.* Crane, New York: Rusak & Company, 1974.

Jones, William B., Jr. *Introduction to Optical Fiber Communication Systems.* New York: Holt, Rinehart and Winston, Inc., 1988.

Keiser, Gerd. *Optical Fiber Communications*, 2d ed. New York: McGraw-Hill, 1991.

Liao, Samuel Y. *Microwave Devices and Circuits*, 3d ed. Englewood Cliffs, NJ: Prentice-Hall, 1990.

Marcuvitz, Nathan. *Waveguide Handbook.* London: Peter Peregrinus Ltd. on behalf of The Institution of Electrical Engineers, 1986.

Mills, Jeffrey P. *Electro-Magnetic Interference Reduction in Electronic Systems.* Englewood Cliffs, NJ: Prentice-Hall, 1993.

Moore, A. D. *Electrostatics and Its Applications.* New York: John Wiley & Sons, 1973.

Palais, Joseph C. *Fiber Optic Communications*, 2d ed. Englewood Cliffs, NJ: Prentice Hall, 1988.

Paul, Clayton R. *Introduction to Electromagnetic Compatibility.* New York: John Wiley & Sons, 1992.

Pozar, David M. *Antenna Design Using Personal Computers*, Dedham, MA: Artech House, 1985.

Pozar, David M. *Microwave Engineering.* Reading, MA: Addison-Wesley, 1990.

Sokolnikoff, Ivan Stephen, & Sokolnikoff, Elizabeth S. *Higher Mathematics for Engineers and Physicists*, 2d ed. New York: McGraw-Hill, 1941.

Sokoinikoff, Ivan S., & Redheffer, Raymond M. *Mathematics of Physics and Modern Engineering*, 2d ed. New York: McGraw-Hill, 1966.

Stutzman, Warren L., & Thiele, Gary A. *Antenna Theory and Design.* New York: John Wiley & Sons, 1981.

Wadell, Brian C. *Transmission Line Design Handbook.* MA: Artech House, Dedham, 1991.

Index

Cylindrical Coordinates (ρ, ϕ, z)

$$\nabla f = \frac{\partial f}{\partial \rho}\hat{\mathbf{a}}_\rho + \frac{1}{\rho}\frac{\partial f}{\partial \phi}\hat{\mathbf{a}}_\phi + \frac{\partial f}{\partial z}\hat{\mathbf{a}}_z$$

$$\nabla \cdot \mathbf{A} = \frac{1}{\rho}\left[\frac{\partial}{\partial \rho}(\rho A_\rho)\right] + \frac{1}{\rho}\frac{\partial A_\phi}{\partial \phi} + \frac{\partial A_z}{\partial z}$$

$$\nabla \times \mathbf{A} = \left[\frac{1}{\rho}\frac{\partial A_z}{\partial \phi} - \frac{\partial A_\phi}{\partial z}\right]\hat{\mathbf{a}}_\rho + \left[\frac{\partial A_\rho}{\partial z} - \frac{\partial A_z}{\partial \rho}\right]\hat{\mathbf{a}}_\phi + \frac{1}{\rho}\left[\frac{\partial}{\partial \rho}(\rho A_\phi) - \frac{\partial A_\rho}{\partial \phi}\right]\hat{\mathbf{a}}_z$$

$$\nabla^2 f = \frac{1}{\rho}\frac{\partial}{\partial \rho}\left(\rho\frac{\partial f}{\partial \rho}\right) + \frac{1}{\rho^2}\frac{\partial^2 f}{\partial \phi^2} + \frac{\partial^2 f}{\partial z^2}$$

Spherical Coordinates (r, θ, ϕ)

$$\nabla f = \frac{\partial f}{\partial r}\hat{\mathbf{a}}_r + \frac{1}{r}\frac{\partial f}{\partial \theta}\hat{\mathbf{a}}_\theta + \frac{1}{r\sin\theta}\frac{\partial f}{\partial \phi}\hat{\mathbf{a}}_\phi$$

$$\nabla \cdot \mathbf{A} = \frac{1}{r^2}\left[\frac{\partial}{\partial r}(r^2 A_r)\right] + \frac{1}{r\sin\theta}\left[\frac{\partial}{\partial \theta}(A_\theta \sin\theta)\right] + \frac{1}{r\sin\theta}\frac{\partial A_\phi}{\partial \phi}$$

$$\nabla \times \mathbf{A} = \frac{1}{r\sin\theta}\left[\frac{\partial}{\partial \theta}(A_\phi \sin\theta) - \frac{\partial A_\theta}{\partial \phi}\right]\hat{\mathbf{a}}_r$$

$$+ \frac{1}{r}\left[\frac{1}{\sin\theta}\frac{\partial A_r}{\partial \phi} - \frac{\partial}{\partial r}(rA_\phi)\right]\hat{\mathbf{a}}_\theta + \frac{1}{r}\left[\frac{\partial}{\partial r}(rA_\theta) - \frac{\partial A_r}{\partial \theta}\right]\hat{\mathbf{a}}_\phi$$

$$\nabla^2 f = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial f}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta\frac{\partial f}{\partial \theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 f}{\partial \phi^2}$$

Vector Identities

$$(\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C} = (\mathbf{B} \times \mathbf{C}) \cdot \mathbf{A} = (\mathbf{C} \times \mathbf{A}) \cdot \mathbf{B}$$

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \cdot \mathbf{C})\mathbf{B} - (\mathbf{A} \cdot \mathbf{B})\mathbf{C}$$

$$\nabla \cdot (V\mathbf{A}) = \mathbf{A} \cdot \nabla \mathbf{V} + V\nabla \cdot \mathbf{A}$$

$$\nabla \times (V\mathbf{A}) = \nabla V \times \mathbf{A} + V\nabla \times \mathbf{A}$$

$$\nabla \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot \nabla \times \mathbf{A} - \mathbf{A} \cdot \nabla \times \mathbf{B}$$

$$\nabla(\mathbf{A} \cdot \mathbf{B}) = (\mathbf{A} \cdot \nabla)\mathbf{B} + (\mathbf{B} \cdot \nabla)\mathbf{A} + \mathbf{A} \times (\nabla \times \mathbf{B}) + \mathbf{B} \times (\nabla \times \mathbf{A})$$

$$\nabla \cdot \nabla \equiv \nabla^2$$

$$\nabla \cdot (\nabla \times \mathbf{A}) = 0$$

$$\nabla \times (\nabla V) = 0$$

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

$$\int_V (\nabla \cdot \mathbf{A}) dv = \oint_S \mathbf{A} \cdot \mathbf{ds}$$

$$\int_S (\nabla \times \mathbf{A}) \cdot ds = \oint_C \mathbf{A} \cdot \mathbf{d\ell}$$

$$\int_V \nabla \times \mathbf{F} dv = -\oint_S \mathbf{F} \times \mathbf{ds}$$

Gradient, Divergence, Curl, and Laplacian Operations

Cartesian Coordinates (x, y, z)

$$\nabla f = \frac{\partial f}{\partial x}\hat{\mathbf{a}}_x + \frac{\partial f}{\partial y}\hat{\mathbf{a}}_y + \frac{\partial f}{\partial z}\hat{\mathbf{a}}_z$$

$$\nabla \cdot \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}$$

$$\nabla \times \mathbf{A} = \left[\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right]\hat{\mathbf{a}}_x + \left[\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right]\hat{\mathbf{a}}_y + \left[\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right]\hat{\mathbf{a}}_z$$

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$$

PRENTICE HALL
Upper Saddle River, NJ 07458

http://www.prenhall.com

ISBN 0-02-328521-4

90000

9 780023 285219