FOUNDATIONS FOR RADIO FREQUENCY ENGINEERING

Wen Geyi

World Scientific

FOUNDATIONS FOR RADIO FREQUENCY ENGINEERING

This page intentionally left blank

FOUNDATIONS FOR RADIO FREQUENCY ENGINEERING

Wen Geyi

Nanjing University of Information Science & Technology, China

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data

Wen, Geyi.

Foundations for radio frequency engineering / by Wen Geyi (Nanjing University of Information Science & Technology, China).

pages cm

Includes bibliographical references and index.

ISBN 978-9814578707 (alk. paper)

1. Radio wave propagation--Mathematics. 2. Electromagnetic waves--Mathematical models. 3. Microwaves. I. Title. II. Title: Radio frequency engineering.

QC665.T7W46 2015

621.384--dc23

2014045142

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

In-house Editors: Sutha Surenddar/Steven Patt

Typeset by Stallion Press

Email: enquiries@stallionpress.com

Printed in Singapore

To my parents
To Jun and Lan

This page intentionally left blank

Preface

In an age of knowledge explosion, students have to meet the challenges of maintaining perspective amid a deluge of information and change in their developments of expertise. The traditional university courses and their contents must therefore be designed, planned or merged accordingly so that the students can master the core materials that are needed in their future careers, while having enough time to study the new courses to be frequently added to the curriculum.

With the rapid development of wireless communication technologies, the demand on wireless spectrum has been growing dramatically. This results in extensive and intensive research in radio frequency (RF) theory and techniques, and substantial advancements in the area of radio engineering, both in theory and practice, have emerged in recent years. RF engineering deals with various devices that are designed to operate in the frequency range from 3 kHz to 300 GHz, and therefore covers all areas where electromagnetic fields must be transmitted or received as a carrier wave. For this reason, a good RF engineer must have in-depth knowledge in mathematics and physics, as well as specialized training in the areas of applied electromagnetics such as guided structures and microwave circuits, antenna and wave propagation, and electromagnetic compatibility (EMC) designs of electronic circuits.

RF engineering is closely linked to three IEEE (Institute of Electrical and Electronics Engineers) professional societies: Microwave Theory and Techniques (MTT), Antennas and Propagation (AP) and EMC. Traditionally, different courses have been created to meet the needs for different societies. For example, the students oriented to the MTT society must take the courses such as *Microwave Engineering*, or *Field Theory of*

Guided Waves. For the students specialized in AP and EMC societies, the courses *Antenna Theory and Design* and *Electromagnetic Compatibility* are compulsory. Nevertheless, these three professional societies are intimately related as they have numerous things in common in terms of theories and techniques. Many professionals are often active in the three societies at the same time. In fact, a typical RF department in a wireless company has engineers working in different areas belonging to these societies, and most of the time, they have to work together to solve an engineering problem as a team.

The above trends indicate that it is necessary and also possible to create a new course or a book that provides the fundamentals of microwaves, antennas and propagation, and EMC in a common framework for the students, engineers and applied physicists dedicated to the three IEEE societies. The topics in RF engineering are enormous. The contents of the book have been selected on the basis of their fundamentality and importance to suit various needs arising in RF engineering. All areas of RF engineering are established on the solutions of Maxwell equations, which can be solved either analytically or numerically. Before the invention of computers, analytical methods were the dominant tools for the analysis of electromagnetic phenomena, often involving the applications of sophisticated mathematics and closed-form solutions. Nowadays, computer technology plays a tremendous role in our daily life as well as in scientific research activities. By taking advantage of the capabilities of modern computer technologies and the state-of-the-art computer-aided design (CAD) tools, the numerical methods are capable of solving many complicated problems encountered in practice and they occupy a significant piece of current academic research. In electromagnetic society, numerical methods have been treated in many references. For this reason, this book essentially examines analytical techniques while typical numerical methods and their applications will also be discussed.

One of the important research areas of RF engineering is the microwave field theory which may be applied to the analysis of guided waves, resonances, radiations and scattering. In many situations, a microwave field problem can be reduced to a network or circuit problem, allowing us to apply the circuit and network methods to solve the original field problem. The network formulation has eliminated unnecessary details in the field theory while reserving useful global information, such as the terminal voltages and currents. As a consequence, many RF engineers now largely rely on CAD tools and circuit analysis with little or no field analysis.

This procedure, however, is not always successful. In fact, the initial RF circuits resulting from CAD tools usually bear little resemblance to the final design, and revisions are needed to achieve the required performances. One should always remember that the field theory is the foundation of the circuit analysis, and its importance cannot be overemphasized in order for best innovation practices. For this reason, both microwave field theory and circuit design theory are discussed in this book, which features a wide coverage of the fundamental topics such as electromagnetic boundary value problems, waveguide theory, microwave resonators, microwave circuits, antennas and wave propagation, EMC techniques, and information theory and typical application systems.

The book consists of 8 chapters. Chapter 1 reviews the fundamental electromagnetic theory. The basic properties and important theorems derived from Maxwell equations are summarized. When applied properly, these properties and theorems may bring deep physical insight into the practical problems and simplify them dramatically. Various solution methods for the boundary value problems related to Maxwell equations are discussed, which includes the method of separation of variables, the method of Green's functions, and the method of variations. Some important topics such as numerical techniques and potential theory are also addressed.

Chapter 2 deals with the waveguide theory. Waveguides are the cornerstone of microwave engineering and their counterparts are connecting wires in low frequency circuits. The waveguide theory can be formulated as an eigenvalue problem with the cut-off wavenumbers being the eigenvalues and eigenfunctions being the corresponding vector modal functions. A variational principle exists for the cut-off wavenumbers and can be expressed as a Rayleigh quotient. The vector modal functions are the extremal functions that make the Rayleigh quotient stationary and constitute a complete set. The typical waveguide eigenvalue problems are solved by the method of separation of variables as well as by various numerical methods. The waveguide discontinuities or waveguide junctions are analyzed by the field expansions in terms of the vector modal functions as well as numerical methods. Also presented in this chapter are inhomogeneous waveguides such as dielectric waveguides and microstrip lines, transient phenomena in waveguides, and periodic structures.

A resonator is a device that oscillates at some frequencies (called resonant frequencies) with greater amplitude than at others, and it is used to either generate waves of specific frequencies or to select specific frequencies from a signal. Its counterpart is the LC resonant circuit at

low frequency. A microwave resonator is an important building block for microwave circuits. A cavity resonator also constitutes an eigenvalue problem and there exists a variational principle for the resonant frequencies. The theory of cavity resonators parallels that of waveguides. Chapter 3 investigates the theory of microwave resonators. It includes the solutions of vector modal functions for typical cavity resonators, coupling between the waveguide and cavity resonator, dielectric resonators, microstrip patch resonators and open resonators.

A microwave circuit is composed of distributed elements with dimensions comparable to the wavelength. The amplitudes and phases of the voltage and current may vary significantly over the length of the circuit. The essential tools for the analysis and design of microwave circuits include the theory of transmission lines, network analysis and synthesis, impedance matching, and filter theory. Many design problems in microwave integrated circuit systems can be reduced to a circuit problem without too much involvement of the field theory. According to their functions, the RF circuit components may be classified as frequency-related, impedance-related and power-related. The fundamental aspects of the network theory and the design principles for various passive and active RF components, such as phase shifters, attenuators, power combiners and dividers, directional couplers, filters, amplifiers, oscillators, and mixers, are summarized and elucidated in Chapter 4.

An antenna is a device which converts a guided wave in a feeding line into a radio wave in free space, and vice versa. Antennas are essential to all wireless systems and play a role in linking the components in the systems through radio waves. In transmission mode, a radio transmitter supplies RF power to the antenna's terminals through a transmission line, and the antenna radiates the energy into space as an electromagnetic wave. In reception mode, the antenna intercepts the power from the incoming electromagnetic wave, inducing a weak voltage signal at its terminal. The induced voltage signal is then amplified for further processing. Antennas can be designed to transmit or receive radio waves in a particular direction (directional antennas) or in all directions equally (omnidirectional antennas). Chapter 5 is devoted to the antenna theory as well as the radiation mechanisms of typical antenna systems, including wire antennas, slot and aperture antennas, broadband antennas, and array antennas. It features a wide coverage of advanced topics, such as the spherical vector wavefunctions, Foster reactance theorem for ideal antennas, equivalent circuits for transmitting and receiving antennas, the physical limitations

of antennas, and the methods for the evaluation of antenna quality factor.

A radio propagation model is a mathematical formulation for the characterization of radio wave propagation as a function of frequency, distance and other conditions. Many factors may change propagation properties of radio waves. The atmosphere, the ground, mountains, buildings, and weather conditions all have major influences on wave propagations and cause variation in receiving signal strength. The propagation models can be roughly divided into statistical and deterministic models. The statistical models are derived from extensive field measurements and statistical analysis, and are valid for similar environments where the measurements were carried out. Sometimes site-specific deterministic propagation models are preferred for more accurate predictions of radio wave propagations. Chapter 6 is concerned with the propagation of radio waves in atmosphere and the ray tracing techniques, the statistical models for mobile channels, and the propagation models for deterministic multiple-input multiple-output (MIMO) systems.

EMC studies the unintentional generation, transmission and reception of electromagnetic energy, and deals with the electromagnetic interferences (EMI) or disturbance that the unintentional electromagnetic energy (as an external source) may induce. Its aim is to ensure that the electronic equipment will not interfere with each other's normal operation. Compliance with national or international standards is usually required by laws passed by individual nations before the electronic devices are brought to the market. Chapter 7 investigates the EMC in modern electronic circuits and systems. The relationship between the fields and circuits is discussed. The basic rules for emission reduction are expounded. The transmission line models for the study of susceptibility are introduced. The effective techniques to cope with the EMC issues are investigated.

The pioneering work on information theory by Shannon (1948) and Wiener (1949) has laid the foundation of modern communication theory and information systems. The fundamental theorem of information theory states that it is possible to transmit information through a noise channel at any rate less than channel capacity with an arbitrarily small probability of error. A signal chosen from a specified class is to be transmitted through a communication channel as an input and is received at the output of the channel. During the transmission, the signal may be altered by noise and distortion. For each permissible input, the output is determined statistically by a probability distribution assigned by the channel. At the output, a

statistical decision must be made to identify the transmitted signal as reliably as possible. Chapter 8 gives a brief introduction to the information theory and typical information systems. It covers the probability theory and random process, information theory, digital communication systems, and radar systems.

The book can be used as a graduate-level text or as a reference for researchers and engineers in applied electromagnetics. The prerequisite for the book is advanced calculus. The SI unit system is used throughout the book. A $e^{j\omega t}$ time variation is assumed for time-harmonic fields. A special symbol '□' is used to indicate the end of an example or a remark. The references are not meant to be complete but the author has tried to indicate the origins of the important results included in the book.

In preparing this book, I have benefited from suggestions from many colleagues and friends, and would like to take this occasion to extend my sincere thanks to all of them. Particular thanks go to Prof. Thomas T. Y. Wong of Illinois Institute of Technology and Prof. Jun Xiang Ge of Nanjing University of Information Science and Technology. Last but not least, I would like to express my deepest gratitude to my family members for their constant encouragement and support.

Wen Geyi
Nanjing, China
May 2014

Contents

Chapter 1

Solutions of Electromagnetic Field Problems

> One scientific epoch ended and another began with James Clerk Maxwell.
>
> —Albert Einstein

The electromagnetic theory is the foundation of radio frequency (RF) engineering. In 1873, J.C. Maxwell (1831–1879) summarized the theory on electricity and magnetism discovered by many great physicists including H.C. Oersted (1777–1851), A.M. Ampère (1775–1836), and M. Faraday (1791–1861), and formulated a set of equations since known as Maxwell equations, representing one of the great achievements in physics. The Maxwell equations describe the behavior of electric and magnetic fields, as well as their interactions with matter, and they are the starting point for the investigation of all macroscopic electromagnetic phenomena.

Radio frequency (RF) refers to the frequency range from 3 KHz to 300 GHz. RF engineering deals with various wireless systems, and is an important subject in electrical engineering. RF technologies are widely used in fixed and mobile communication, broadcasting, radar and navigation systems, satellite communication, computer networks and innumerable other applications. Different frequencies of radio waves have different propagation characteristics in the Earth's atmosphere. Table 1.1 shows various frequency bands and their major applications.

Microwave frequency often refers to the frequency range from 1 GHz to 300 GHz. Table 1.2 gives the old and new names for typical microwave frequency bands. At the low end of the microwave spectrum, the traditional lumped circuit theory starts to become ineffective, and the field theory thus enters the picture. Microwave field theory is one of the important research areas of RF engineering, which may be applied to solve various boundary

Table 1.1 RF spectrum

Frequency/ wavelength	Designation	Applications
3 Hz–30 Hz/ 10^5 km–10^4 km	ELF (Extremely low frequency)	Submarines
30 Hz–300 Hz/ 10^4 km–10^3 km	SLF (Super low frequency)	Power grids, submarines
300 Hz–3 kHz/ 10^3 km–10^2 km	ULF (Ultra low frequency)	Earthquake studies
3 kHz–30 kHz/ 100 km–10 km	VLF (Very low frequency)	Submarines near the surface
30 kHz–300 kHz/ 10 km–1 km	LF (Low frequency)	Submarines, aircraft beacons, AM broadcast, navigation
300 kHz–3 MHz/ 1 km–100 m	MF (Medium frequency)	AM broadcast, navigation
3 MHz–30 MHz/ 100 m–10 m	HF (High frequency)	Shortwave broadcast, over the horizon radar
30 MHz–300 MHz/ 10 m–1 m	VHF (Very high frequency)	FM, TV
300 MHz–3 GHz/ 1 m–10 cm	UHF (Ultra high frequency)	TV, LAN, cellular, GPS
3 GHz–30 GHz/ 10 cm–1 cm	SHF (Super high frequency)	Radar, GSO satellites, data communications
30 GHz–300 GHz/ 1 cm–1 mm	EHF (Extremely high frequency)	Radar, automotive, data communications
300 GHz–3 THz/ 1 mm–0.1 mm	THF (Tremendously high frequency)	Sensing and imaging, security screening, high-altitude communications

Table 1.2 Microwave frequency bands

Frequency (GHz)	Old names	New names
1–2	L	D
2–4	S	E, F
4–8	C	G, H
8–12	X	I, J
12–18	Ku	J
18–26	K	J
26–40	Ka	K

value problems such as the analysis of guided waves, resonances, radiations and scattering.

In mathematics, a boundary value problem consists of a differential equation together with a set of additional constraints on the boundary

Table 1.3 Some trinities for differential equations

Trinity	Description
Three types of differential equations:	Elliptical, hyperbolic and parabolic.
Three types of problems:	Boundary value problems, initial value problems, and eigenvalue problems.
Three types of boundary conditions:	Dirichlet boundary condition, named after the German mathematician Johann Peter Gustav Lejeune Dirichlet (1805–1859); Neumann boundary condition, named after the German mathematician Carl Gottfried Neumann (1832–1925); and Robin boundary condition, named after the French mathematician Victor Gustave Robin (1855–1897).
Three important mathematical tools:	Divergence theorem, inequalities and convergence theorems.
Three analytical solution methods:	Separation of variables, Green's function method, and variational method.
Three numerical solution methods:	Finite element method, finite difference method, moment method.

of the domain of the equation (called the boundary conditions). Various methods for the solution of differential equations have been proposed. Linear differential equations are generally solved by means of variational method, the method of separation of variables, and the method of Green's function, named after the British mathematician George Green (1793–1841). Some usual trinities for differential equations are summarized in Table 1.3 (Gustafson, 1987).

1.1 Maxwell Equations

Maxwell equations are a set of partial differential equations that form the foundation of electrical and optical engineering. Maxwell equations describe how electric and magnetic fields are generated by charges and currents and altered by each other. Maxwell equations have been proved to be very successful in explaining and predicting a variety of macroscopic phenomena. However, in some special situations such as extremely strong fields and extremely short distances, they may fail and can be noticeably inaccurate. Moreover, Maxwell equations must be replaced by quantum electrodynamics in order for dealing with microscopic phenomena.

1.1.1 Maxwell Equations and Boundary Conditions

Maxwell equations in the time domain can be expressed as follows

$$\begin{aligned} \nabla \times \mathbf{H}(\mathbf{r},t) &= \frac{\partial \mathbf{D}(\mathbf{r},t)}{\partial t} + \mathbf{J}(\mathbf{r},t), \\ \nabla \times \mathbf{E}(\mathbf{r},t) &= -\frac{\partial \mathbf{B}(\mathbf{r},t)}{\partial t}, \\ \nabla \cdot \mathbf{D}(\mathbf{r},t) &= \rho(\mathbf{r},t), \\ \nabla \cdot \mathbf{B}(\mathbf{r},t) &= 0. \end{aligned} \tag{1.1}$$

In (1.1), $\mathbf{r}$ is the observation point of the fields in meter and t is the time in second; $\mathbf{H}$ is the magnetic field intensity measured in amperes per meter (A/m); $\mathbf{B}$ is the magnetic induction intensity measured in tesla (N/A·m); $\mathbf{E}$ is electric field intensity measured in volts per meter (V/m); $\mathbf{D}$ is the electric induction intensity measured in coulombs per square meter ($\mathrm{C/m^2}$); $\mathbf{J}$ is electric current density measured in amperes per square meter ($\mathrm{A/m^2}$); ρ is the electric charge density measured in coulombs per cubic meter ($\mathrm{C/m^3}$). The first equation is Ampère's law, and it describes how the electric field changes according to the current density and magnetic field. The second equation is Faraday's law, and it characterizes how the magnetic field varies according to the electric field. The minus sign is required by Lenz's law, i.e., when an electromotive force is generated by a change of magnetic flux, the polarity of the induced electromotive force is such that it produces a current whose magnetic field opposes the change, which produces it. The third equation is Coulomb's law, and it says that the electric field depends on the charge distribution and obeys the inverse square law. The last equation shows that there are no free magnetic monopoles and that the magnetic field also obeys the inverse square law. It should be understood that none of the experiments had anything to do with waves at the time when Maxwell derived his equations. Maxwell equations imply more than the experimental facts. The continuity equation can be derived from (1.1) as

$$\nabla \cdot \mathbf{J}(\mathbf{r},t) = -\frac{\partial \rho(\mathbf{r},t)}{\partial t}. \tag{1.2}$$

The charge density ρ and the current density $\mathbf{J}$ in Maxwell equations are free charge density and currents and they exclude charges and currents forming part of the structure of atoms and molecules. The bound charges and currents are regarded as material, which are not included in ρ and $\mathbf{J}$. The current density normally consists of two parts: $\mathbf{J} = \mathbf{J}_{\mathrm{con}} + \mathbf{J}_{\mathrm{imp}}$. Here $\mathbf{J}_{\mathrm{imp}}$ is referred to as external or impressed current source, which

is independent of the fields and delivers energy to electric charges in a system. The impressed current source can be of electric and magnetic type as well as of non-electric or non-magnetic origin. $\mathbf{J}_{\text{con}} = \sigma\mathbf{E}$, where σ is the conductivity of the medium in mhos per meter, denotes the conduction current induced by the impressed source $\mathbf{J}_{\text{imp}}$. Sometimes it is convenient to introduce an external or impressed electric field $\mathbf{E}_{\text{imp}}$ defined by $\mathbf{J}_{\text{imp}} = \sigma\mathbf{E}_{\text{imp}}$. In more general situation, one may write $\mathbf{J} = \mathbf{J}_{\text{ind}}(\mathbf{E}, \mathbf{B}) + \mathbf{J}_{\text{imp}}$, where $\mathbf{J}_{\text{ind}}(\mathbf{E}, \mathbf{B})$ is the induced current by the impressed current $\mathbf{J}_{\text{imp}}$.

Sometimes it is convenient to introduce magnetic current $\mathbf{J}_m$ and magnetic charges ρ_m, which are related by

$$\nabla \cdot \mathbf{J}_m(\mathbf{r}, t) = -\frac{\partial \rho_m(\mathbf{r}, t)}{\partial t} \tag{1.3}$$

and Maxwell equations must be modified as

$$\begin{aligned} \nabla \times \mathbf{H}(\mathbf{r}, t) &= \frac{\partial \mathbf{D}(\mathbf{r}, t)}{\partial t} + \mathbf{J}(\mathbf{r}, t), \\ \nabla \times \mathbf{E}(\mathbf{r}, t) &= -\frac{\partial \mathbf{B}(\mathbf{r}, t)}{\partial t} - \mathbf{J}_m(\mathbf{r}, t), \\ \nabla \cdot \mathbf{D}(\mathbf{r}, t) &= \rho(\mathbf{r}, t), \\ \nabla \cdot \mathbf{B}(\mathbf{r}, t) &= \rho_m(\mathbf{r}, t). \end{aligned} \tag{1.4}$$

The inclusions of $\mathbf{J}_m$ and ρ_m make Maxwell equations more symmetric although there has been no evidence that the magnetic current and charge are physically present. The validity of introducing such concepts in Maxwell equations is justified by the equivalence principle, i.e., they are introduced as a mathematical equivalent to electromagnetic fields.

If all the sources are of magnetic type, Equations (1.4) reduce to

$$\begin{aligned} \nabla \times \mathbf{H}(\mathbf{r}, t) &= \frac{\partial \mathbf{D}(\mathbf{r}, t)}{\partial t}, \\ \nabla \times \mathbf{E}(\mathbf{r}, t) &= -\frac{\partial \mathbf{B}(\mathbf{r}, t)}{\partial t} - \mathbf{J}_m(\mathbf{r}, t), \\ \nabla \cdot \mathbf{D}(\mathbf{r}, t) &= 0, \\ \nabla \cdot \mathbf{B}(\mathbf{r}, t) &= \rho_m(\mathbf{r}, t). \end{aligned} \tag{1.5}$$

Mathematically (1.1) and (1.5) are similar. One can obtain one of them by simply interchanging symbols between the left and right columns in Table 1.4, where μ and ε denote the permeability and permittivity of the medium respectively. This property is called **duality**. The importance of

Table 1.4 Duality

Electric source	Magnetic source
$\mathbf{E}$	$\mathbf{H}$
$\mathbf{H}$	$-\mathbf{E}$
$\mathbf{J}$	$\mathbf{J}_m$
ρ	ρ_m
μ	ε
ε	μ

the duality is that one can obtain the solution of magnetic type from the solution of electric type by interchanging symbols and vice versa.

For the time-harmonic (sinusoidal) fields, Equations (1.1) and (1.2) can be expressed as

$$\begin{aligned} \nabla \times \mathbf{H}(\mathbf{r}) &= j\omega \mathbf{D}(\mathbf{r}) + \mathbf{J}(\mathbf{r}), \\ \nabla \times \mathbf{E}(\mathbf{r}) &= -j\omega \mathbf{B}(\mathbf{r}), \\ \nabla \cdot \mathbf{D}(\mathbf{r}) &= \rho(\mathbf{r}), \\ \nabla \cdot \mathbf{B}(\mathbf{r}) &= 0, \\ \nabla \cdot \mathbf{J}(\mathbf{r}) &= -j\omega \rho(\mathbf{r}), \end{aligned} \tag{1.6}$$

where the field quantities denote the complex amplitudes (phasors) defined by

$$\mathbf{E}(\mathbf{r}, t) = \mathrm{Re}[\mathbf{E}(\mathbf{r}) e^{j\omega t}], \text{ etc.}$$

We use the same notations for both time-domain and frequency-domain quantities.

The force acting on a point charge q, moving with a velocity $\mathbf{v}$ with respect to an observer, by the electromagnetic field is given by

$$\mathbf{F}(\mathbf{r}, t) = q[\mathbf{E}(\mathbf{r}, t) + \mathbf{v}(\mathbf{r}, t) \times \mathbf{B}(\mathbf{r}, t)] \tag{1.7}$$

where $\mathbf{E}$ and $\mathbf{B}$ are the total fields, including the field generated by the moving charge q. Equation (1.7) is referred to as **Lorentz force equation**, named after Dutch physicist Hendrik Antoon Lorentz (1853–1928). It is known that there are two different formalisms in classical physics. One is mechanics that deals with particles, and the other is electromagnetic field theory that deals with radiated waves. The particles and waves are coupled through Lorentz force equation, which usually appears as an assumption separate from Maxwell equations. The Lorentz force is the only means

to detect electromagnetic fields. For a continuous charge distribution, the Lorentz force equation becomes

$$\mathbf{f}(\mathbf{r},t) = \rho\mathbf{E}(\mathbf{r},t) + \mathbf{J}(\mathbf{r},t) \times \mathbf{B}(\mathbf{r},t) \tag{1.8}$$

where $\mathbf{f}$ is the force density acting on the charge distribution ρ, i.e., the force acting on the charge distribution per unit volume. Maxwell equations, Lorentz force equation and continuity equation constitute the fundamental equations in electrodynamics.

The boundary conditions on the surface between two different media can be easily obtained as follows

$$\begin{aligned} \mathbf{u}_n \times (\mathbf{H}_1 - \mathbf{H}_2) &= \mathbf{J}_s, \\ \mathbf{u}_n \times (\mathbf{E}_1 - \mathbf{E}_2) &= 0, \\ \mathbf{u}_n \cdot (\mathbf{D}_1 - \mathbf{D}_2) &= \rho_s, \\ \mathbf{u}_n \cdot (\mathbf{B}_1 - \mathbf{B}_2) &= 0, \end{aligned} \tag{1.9}$$

where $\mathbf{u}_n$ is the unit normal of the boundary directed from medium 2 to medium 1; $\mathbf{J}_s$ and ρ_s are the surface current density and surface charge density respectively.

Remark 1.1: To derive the boundary conditions (1.9), we may draw a small cylinder of height Δh and base area ΔS so that the boundary S between medium 1 and medium 2 intersects the middle section of the cylinder as illustrated in Figure 1.1. If the base area is sufficiently small the fields may be assumed to be a constant value over each end of the cylinder. Taking the integration of the first equation of (1.4) over the surface of the cylinder, we obtain

$$\mathbf{u}_n \times \mathbf{H}_1 \Delta S - \mathbf{u}_n \times \mathbf{H}_2 \Delta S + \mathbf{K}\Delta h - \frac{\partial \mathbf{D}}{\partial t}\Delta S \Delta h = \mathbf{J}\Delta S \Delta h, \tag{1.10}$$

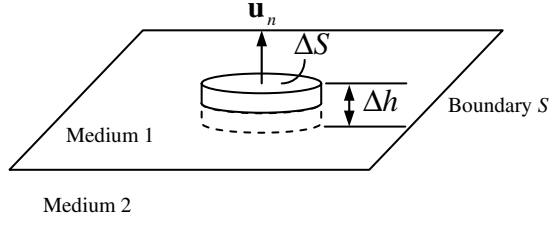


Figure 1.1 Derivation of boundary conditions.

where $\mathbf{K}\Delta h$ denotes the integral of $\mathbf{u}_n \times \mathbf{H}$ over the side walls of the cylinder. In the limit as $\Delta h \to 0$, the ends of the cylinder lie just on either side of the boundary S and the integral over the side walls becomes vanishingly small. Thus

$$\lim_{\Delta h \to 0} \left(\mathbf{K}\Delta h - \frac{\partial \mathbf{D}}{\partial t} \Delta S \Delta h \right) = 0, \quad \lim_{\Delta h \to 0} \mathbf{J}\Delta h = \mathbf{J}_s.$$

Here $\mathbf{J}_s$ stands for the surface current density. Equation (1.10) can be written as

$$\mathbf{u}_n \times (\mathbf{H}_1 - \mathbf{H}_2) = \mathbf{J}_s. \tag{1.11}$$

The rest of the equations in (1.9) can be derived in a similar way. □

1.1.2 Constitutive Relations

Maxwell equations are a set of 7 equations involving 16 unknowns (i.e., five vector functions $\mathbf{E}, \mathbf{H}, \mathbf{B}, \mathbf{D}, \mathbf{J}$ and one scalar function ρ and the last equation of (1.1) is not independent). To determine the fields, nine more equations are needed, and they are given by the **generalized constitutive relations**:

$$\mathbf{D} = f_1(\mathbf{E}, \mathbf{H}), \quad \mathbf{B} = f_2(\mathbf{E}, \mathbf{H})$$

together with the **generalized Ohm's law**:

$$\mathbf{J} = f_3(\mathbf{E}, \mathbf{H})$$

if the medium is conducting. The constitutive relations establish the connections between field quantities and reflect the properties of the medium, and they are totally independent of the Maxwell equations. An **anisotropic medium** is defined by

$$D_i(\mathbf{r}, t) = \sum_{j=x,y,z} [a_i^j(\mathbf{r}) E_j(\mathbf{r}, t) + (G_i^j * E_j)(\mathbf{r}, t)],$$

$$B_i(\mathbf{r}, t) = \sum_{j=x,y,z} [d_i^j(\mathbf{r}) H_j(\mathbf{r}, t) + (F_i^j * H_j)(\mathbf{r}, t)],$$

where $i = x, y, z$; $*$ denotes the convolution with respect to time; a_i^j, d_i^j are independent of time; and G_i^j, F_i^j are functions of $(\mathbf{r}, t)$. A **biisotropic medium** is defined by

$$\mathbf{D}(\mathbf{r}, t) = a(\mathbf{r})\mathbf{E}(\mathbf{r}, t) + b(\mathbf{r})\mathbf{H}(\mathbf{r}, t) + (G * \mathbf{E})(\mathbf{r}, t) + (K * \mathbf{H})(\mathbf{r}, t)$$

$$\mathbf{B}(\mathbf{r}, t) = c(\mathbf{r})\mathbf{E}(\mathbf{r}, t) + d(\mathbf{r})\mathbf{H}(\mathbf{r}, t) + (L * \mathbf{E})(\mathbf{r}, t) + (F * \mathbf{H})(\mathbf{r}, t)$$

where a, b, c, d are independent of time and G, K, L, F are functions of $(\mathbf{r}, t)$. An **isotropic medium** is defined by

$$\mathbf{D}(\mathbf{r}, t) = a(\mathbf{r})\mathbf{E}(\mathbf{r}, t) + (G * \mathbf{E})(\mathbf{r}, t),$$
$$\mathbf{B}(\mathbf{r}, t) = d(\mathbf{r})\mathbf{H}(\mathbf{r}, t) + (F * \mathbf{H})(\mathbf{r}, t).$$

For monochromatic fields, the constitutive relations for an anisotropic medium are usually expressed by

$$\mathbf{D} = \overleftrightarrow{\boldsymbol{\varepsilon}} \cdot \mathbf{E}, \quad \mathbf{B} = \overleftrightarrow{\boldsymbol{\mu}} \cdot \mathbf{H},$$

where $\overleftrightarrow{\boldsymbol{\mu}}$ and $\overleftrightarrow{\boldsymbol{\varepsilon}}$ are dyads. For an introduction of dyadic analysis, please refer to Bladel (2007).

The constitutive relations are often written as

$$\begin{aligned} \mathbf{D}(\mathbf{r}, t) &= \varepsilon_0 \mathbf{E}(\mathbf{r}, t) + \mathbf{P}(\mathbf{r}, t) + \cdots, \\ \mathbf{B}(\mathbf{r}, t) &= \mu_0[\mathbf{H}(\mathbf{r}, t) + \mathbf{M}(\mathbf{r}, t) + \cdots], \end{aligned} \tag{1.12}$$

where μ_0 and ε_0 are permeability and permittivity in vacuum respectively; $\mathbf{M}$ is the magnetization vector and $\mathbf{P}$ is the polarization vector. Equations (1.12) may contain higher order terms, which have been omitted since in most cases only the magnetization and polarization vectors are significant. The vector $\mathbf{M}$ and $\mathbf{P}$ reflect the effects of the Lorentz force on elemental particles in the medium and therefore they depend on both $\mathbf{E}$ and $\mathbf{B}$ in general. Since the elemental particles in the medium have finite masses and are mutually interacting, $\mathbf{M}$ and $\mathbf{P}$ are also functions of time derivatives of $\mathbf{E}$ and $\mathbf{B}$ as well as their magnitudes. The same applies for the current density $\mathbf{J}_{\text{ind}}$. In most cases, $\mathbf{M}$ is only dependent on the magnetic field $\mathbf{B}$ and its time derivatives while $\mathbf{P}$ and $\mathbf{J}$ are only dependent on the electric field $\mathbf{E}$ and its time derivatives. If these dependences are linear, the medium is said to be **linear**. These linear dependences are usually expressed as

$$\begin{aligned} \mathbf{D} &= \tilde{\varepsilon}\mathbf{E} + \tilde{\varepsilon}_1 \frac{\partial \mathbf{E}}{\partial t} + \tilde{\varepsilon}_2 \frac{\partial^2 \mathbf{E}}{\partial t^2} + \cdots, \\ \mathbf{B} &= \tilde{\mu}\mathbf{H} + \tilde{\mu}_1 \frac{\partial \mathbf{H}}{\partial t} + \tilde{\mu}_2 \frac{\partial^2 \mathbf{H}}{\partial t^2} + \cdots, \\ \mathbf{J}_{\text{ind}} &= \tilde{\sigma}\mathbf{E} + \tilde{\sigma}_1 \frac{\partial \mathbf{E}}{\partial t} + \tilde{\sigma}_2 \frac{\partial^2 \mathbf{E}}{\partial t^2} + \cdots, \end{aligned} \tag{1.13}$$

where all the scalar coefficients are constants. For the monochromatic fields, the first two expressions of (1.13) reduce to

$$\mathbf{D} = \varepsilon\mathbf{E}, \quad \mathbf{B} = \mu\mathbf{H}$$

where

$$\begin{aligned} \varepsilon &= \varepsilon' - j\varepsilon'', & \mu &= \mu' - j\mu'', \\ \varepsilon' &= \tilde{\varepsilon} - \omega^2\tilde{\varepsilon}_2 + \cdots, & \mu' &= \tilde{\mu} - \omega^2\tilde{\mu}_2 + \cdots, \\ \varepsilon'' &= -\omega\tilde{\varepsilon}_1 + \omega^3\tilde{\varepsilon}_3 - \cdots, & \mu'' &= -\omega\tilde{\mu}_1 + \omega^3\tilde{\mu}_3 - \cdots. \end{aligned} \tag{1.14}$$

The parameters ε' and ε'' are real and are called **capacitivity** and **dielectric loss factor** respectively. The parameters μ' and μ'' are real and are called **inductivity** and **magnetic loss factor** respectively.

1.1.3 Wave Equations

The electromagnetic wave equations are second order partial differential equations that describe the propagation of electromagnetic waves through a medium. If the medium is homogeneous and isotropic and non-dispersive, we have $\mathbf{B} = \mu\mathbf{H}$ and $\mathbf{D} = \varepsilon\mathbf{E}$, where μ and ε are constants. On elimination of $\mathbf{E}$ or $\mathbf{H}$ in the generalized Maxwell equations, we obtain

$$\begin{aligned} \nabla\times\nabla\times\mathbf{E} + \mu\varepsilon\frac{\partial^2\mathbf{E}}{\partial t^2} &= -\nabla\times\mathbf{J}_m - \mu\frac{\partial\mathbf{J}}{\partial t}, \\ \nabla\times\nabla\times\mathbf{H} + \mu\varepsilon\frac{\partial^2\mathbf{H}}{\partial t^2} &= \nabla\times\mathbf{J} - \varepsilon\frac{\partial\mathbf{J}_m}{\partial t}. \end{aligned} \tag{1.15}$$

These are known as the **wave equations**. For the time-harmonic fields, Equations (1.15) reduce to

$$\begin{aligned} \nabla\times\nabla\times\mathbf{E} - k^2\mathbf{E} &= -\nabla\times\mathbf{J}_m - j\omega\mu\mathbf{J}, \\ \nabla\times\nabla\times\mathbf{H} - k^2\mathbf{H} &= \nabla\times\mathbf{J} - j\omega\varepsilon\mathbf{J}_m, \end{aligned} \tag{1.16}$$

where $k = \omega\sqrt{\mu\varepsilon}$ is the wavenumber. It can be seen that the source terms on the right-hand side of (1.15) and (1.16) are very complicated. To simplify the analysis, the electromagnetic potential functions may be introduced (see Section 1.5). The wave equations may be used to solve the following three different field problems:

(1) Electromagnetic fields in source-free region: Wave propagations in space and waveguides, wave oscillation in cavity resonators, etc.
(2) Electromagnetic fields generated by known source distributions: Antenna radiations, excitations in waveguides and cavity resonators, etc.
(3) Interaction of fields and sources: Wave propagation in plasma, coupling between electron beams and propagation mechanism, etc.

If the medium is inhomogeneous and anisotropic so that $\mathbf{D} = \overleftrightarrow{\boldsymbol{\varepsilon}} \cdot \mathbf{E}$ and $\mathbf{B} = \overleftrightarrow{\boldsymbol{\mu}} \cdot \mathbf{H}$, the wave equations for the time-harmonic fields are

$$\begin{aligned} \nabla \times \overleftrightarrow{\boldsymbol{\mu}}^{-1} \cdot \nabla \times \mathbf{E}(\mathbf{r}) - \omega^2 \overleftrightarrow{\boldsymbol{\varepsilon}} \cdot \mathbf{E}(\mathbf{r}) &= -j\omega \mathbf{J}(\mathbf{r}) - \nabla \times \overleftrightarrow{\boldsymbol{\mu}}^{-1} \cdot \mathbf{J}_m, \\ \nabla \times \overleftrightarrow{\boldsymbol{\varepsilon}}^{-1} \cdot \nabla \times \mathbf{H}(\mathbf{r}) - \omega^2 \overleftrightarrow{\boldsymbol{\mu}} \cdot \mathbf{H}(\mathbf{r}) &= -j\omega \mathbf{J}_m(\mathbf{r}) + \nabla \times \overleftrightarrow{\boldsymbol{\varepsilon}}^{-1} \cdot \mathbf{J}. \end{aligned} \tag{1.17}$$

1.1.4 Dispersion

If the speed of the wave propagation and the wave attenuation in a medium depend on the frequency, the medium is said to be dispersive. Dispersion arises from the fact that the polarization and magnetization and the current density cannot follow the rapid changes of the electromagnetic fields, which implies that the electromagnetic energy can be absorbed by the medium. Thus, the dissipation or absorption always occurs whenever the medium shows the dispersive effects. In reality, all media show some dispersive effects. The medium can be divided into normal dispersive and anomalous dispersive. A **normal dispersive medium** refers to the situation where the refractive index increases as the frequency increases. Most naturally occurring transparent media exhibit normal dispersion in the visible range of electromagnetic spectrum. In an **anomalous dispersive medium**, the refractive index decreases as frequency increases. The dispersive effects are usually recognized by the existence of elementary solutions (plane wave solution) of Maxwell equations in source-free region

$$A(\mathbf{k}) e^{j(\omega t - \mathbf{k} \cdot \mathbf{r})}, \tag{1.18}$$

where $A(\mathbf{k})$ is the amplitude, $\mathbf{k}$ is wave vector and ω is the frequency. When the elementary solutions are introduced into Maxwell equations, a relationship between $\mathbf{k}$ and ω may be found as follows

$$f(\omega, \mathbf{k}) = 0. \tag{1.19}$$

This is called **dispersion relation**. For a single linear differential equation with constant coefficients, there is a one–one correspondence between the equation and the dispersion relation. We only need to consider the following correspondences:

$$\frac{\partial}{\partial t} \leftrightarrow j\omega, \quad \nabla \leftrightarrow -j\mathbf{k},$$

which yield a polynomial dispersion relation. To find the dispersion relation of the medium, the plane wave solutions may be assumed for Maxwell

equations as follows

$$\mathbf{E}(\mathbf{r},t) = \mathrm{Re}[\mathbf{E}(\mathbf{r})e^{j\omega t - j\mathbf{k}\cdot\mathbf{r}}], \text{ etc.} \tag{1.20}$$

Similar expressions hold for other quantities. In the following, it is assumed that the wave vector $\mathbf{k}$ is allowed to be a complex vector and there is no impressed source inside the medium. Introducing (1.20) into Maxwell equations (1.6) with $\mathbf{J}_{\text{imp}} = 0$ and using the calculation $\nabla e^{-j\mathbf{k}\cdot\mathbf{r}} = -j\mathbf{k}e^{-j\mathbf{k}\cdot\mathbf{r}}$, we obtain

$$\begin{aligned} -j\mathbf{k}\times\mathbf{H}(\mathbf{r}) + \nabla\times\mathbf{H}(\mathbf{r}) &= j\omega\mathbf{D}(\mathbf{r}) + \mathbf{J}_{\text{con}}(\mathbf{r}), \\ -j\mathbf{k}\times\mathbf{E}(\mathbf{r}) + \nabla\times\mathbf{E}(\mathbf{r}) &= -j\omega\mathbf{B}(\mathbf{r}). \end{aligned}$$

In most situations, the complex amplitudes of the fields are slowly varying functions of space coordinates. The above equations may thus reduce to

$$\begin{aligned} \mathbf{k}\times\mathbf{H}(\mathbf{r}) &= -\omega\mathbf{D}(\mathbf{r}) + j\mathbf{J}_{\text{con}}(\mathbf{r}), \\ \mathbf{k}\times\mathbf{E}(\mathbf{r}) &= \omega\mathbf{B}(\mathbf{r}). \end{aligned} \tag{1.21}$$

If the medium is isotropic, dispersive and lossy, we may write

$$\mathbf{J}_{\text{con}} = \sigma\mathbf{E}, \quad \mathbf{D} = (\varepsilon' - j\varepsilon'')\mathbf{E}, \quad \mathbf{B} = (\mu' - j\mu'')\mathbf{H}.$$

Substituting these into (1.21) yields

$$\mathbf{k}\cdot\mathbf{k} = \omega^2(\mu' - j\mu'')[\varepsilon' - j(\varepsilon'' + \sigma/\omega)].$$

Assuming $\mathbf{k} = \mathbf{u}_k(\beta - j\alpha)$($\mathbf{u}_k$ is a unit vector), then we have

$$\beta - j\alpha = \omega\sqrt{(\mu' - j\mu'')[\varepsilon' - j(\varepsilon'' + \sigma/\omega)]}$$

from which we may find that

$$\beta = \frac{\omega}{\sqrt{2}}\sqrt{(A^2 + B^2)^{1/2} + A}, \quad \alpha = \frac{\omega}{\sqrt{2}}\sqrt{(A^2 + B^2)^{1/2} - A}$$

where $A = \mu'\varepsilon' - \mu''(\varepsilon'' + \sigma/\omega)$, $B = \mu''\varepsilon' + \mu'(\varepsilon'' + \sigma/\omega)$.

1.1.5 Electromagnetic Field Theorems

A number of theorems can be derived from Maxwell equations, and they usually bring deep physical insight into the electromagnetic field

problems. When applied properly, these theorems can simplify the problems dramatically.

1.1.5.1 *Superposition Theorem*

Superposition theorem applies to all linear systems. Suppose that the impressed current source $\mathbf{J}_{\text{imp}}$ can be expressed as a linear combination of independent impressed current sources $\mathbf{J}^k_{\text{imp}}(k = 1, 2, \ldots, n)$

$$\mathbf{J}_{\text{imp}} = \sum_{k=1}^{n} a_k \mathbf{J}^k_{\text{imp}},$$

where $a_k(k = 1, 2, \ldots, n)$ are arbitrary constants. If $\mathbf{E}^k$ and $\mathbf{H}^k$ are fields produced by the source $\mathbf{J}^k_{\text{imp}}$, the **superposition theorem** for electromagnetic fields asserts that the fields $\mathbf{E} = \sum_{k=1}^{n} a_k \mathbf{E}^k$ and $\mathbf{H} = \sum_{k=1}^{n} a_k \mathbf{H}^k$ are a solution of Maxwell equations produced by the source $\mathbf{J}_{\text{imp}}$.

1.1.5.2 *Conservation of Electromagnetic Energy*

The law of **conservation of electromagnetic energy** is known as the **Poynting theorem**, named after the English physicist John Henry Poynting (1852–1914). It can be found from (1.1) that

$$-\mathbf{J}_{\text{imp}} \cdot \mathbf{E} - \mathbf{J}_{\text{ind}} \cdot \mathbf{E} = \nabla \cdot \mathbf{S} + \mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} + \mathbf{H} \cdot \frac{\partial \mathbf{B}}{\partial t}. \tag{1.22}$$

In a region V bounded by S, the integral form of (1.22) is

$$-\int_V \mathbf{J}_{\text{imp}} \cdot \mathbf{E}\, dV = \int_V \mathbf{J}_{\text{ind}} \cdot \mathbf{E}\, dV + \int_S \mathbf{S} \cdot \mathbf{u}_n\, dS + \int_V \left(\mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} + \mathbf{H} \cdot \frac{\partial \mathbf{B}}{\partial t} \right) dV, \tag{1.23}$$

where $\mathbf{u}_n$ is the unit outward normal of S, and $\mathbf{S} = \mathbf{E} \times \mathbf{H}$ is the **Poynting vector** representing the electromagnetic power-flow density measured in watts per square meter (W/m^2). It is assumed that this explanation holds for all media. Thus, the left-hand side of the above equation stands for the power supplied by the impressed current source. The first term on the right-hand side is the work done per second by the electric field to maintain the current in the conducting part of the system. The second term denotes the electromagnetic power flowing out of S. The last term can be interpreted

as the work done per second by the impressed source to establish the fields. The energy density w required to establish the electromagnetic fields may be defined as follows

$$dw = \left(\mathbf{E}\cdot\frac{\partial \mathbf{D}}{\partial t} + \mathbf{H}\cdot\frac{\partial \mathbf{B}}{\partial t}\right)dt. \tag{1.24}$$

Assuming all the sources and fields are zero at $t = -\infty$, we have

$$w = w_e + w_m, \tag{1.25}$$

where w_e and w_m are the **electric field energy density** and **magnetic field energy density** respectively

$$w_e = \frac{1}{2}\mathbf{E}\cdot\mathbf{D} + \int_{-\infty}^{t}\frac{1}{2}\left(\mathbf{E}\cdot\frac{\partial \mathbf{D}}{\partial t} - \mathbf{D}\cdot\frac{\partial \mathbf{E}}{\partial t}\right)dt,$$

$$w_m = \frac{1}{2}\mathbf{H}\cdot\mathbf{B} + \int_{-\infty}^{t}\frac{1}{2}\left(\mathbf{H}\cdot\frac{\partial \mathbf{B}}{\partial t} - \mathbf{B}\cdot\frac{\partial \mathbf{H}}{\partial t}\right)dt.$$

Equation (1.23) can be written as

$$-\int_V \mathbf{J}_{\text{imp}}\cdot\mathbf{E}\,dV = \int_V \mathbf{J}_{\text{ind}}\cdot\mathbf{E}\,dV + \int_S \mathbf{S}\cdot\mathbf{u}_n\,dS + \frac{\partial}{\partial t}\int_V (w_e + w_m)dV. \tag{1.26}$$

In general, the energy density w does not represent the stored energy density in the fields: the energy temporarily located in the fields and completely recoverable when the fields are reduced to zero. The energy density w given by (1.25) can be considered as the stored energy density only if the medium is lossless (i.e., $\nabla\cdot\mathbf{S} = 0$). If the medium is isotropic and time-invariant, we have

$$w_e = \frac{1}{2}\mathbf{E}\cdot\mathbf{D}, \quad w_m = \frac{1}{2}\mathbf{H}\cdot\mathbf{B}.$$

If the fields are time-harmonic, the Poynting theorem takes the following form

$$-\frac{1}{2}\int_V \mathbf{E}\cdot\bar{\mathbf{J}}_{\text{imp}}\,dV = \frac{1}{2}\int_V \mathbf{E}\cdot\bar{\mathbf{J}}_{\text{ind}}\,dV + \int_S \frac{1}{2}(\mathbf{E}\times\bar{\mathbf{H}})\cdot\mathbf{u}_n\,dS + j2\omega\int_V\left(\frac{1}{4}\mathbf{B}\cdot\bar{\mathbf{H}} - \frac{1}{4}\mathbf{E}\cdot\bar{\mathbf{D}}\right)dV, \tag{1.27}$$

where the bar denotes complex conjugate. The time averages of Poynting vector, energy densities over one period of the sinusoidal wave $e^{j\omega t}$, denoted

T, are

$$\frac{1}{T}\int_0^T \mathbf{E}\times\mathbf{H}dt = \frac{1}{2}\mathrm{Re}(\mathbf{E}\times\bar{\mathbf{H}}),$$

$$\frac{1}{T}\int_0^T \frac{1}{2}\mathbf{E}\cdot\mathbf{D}dt = \frac{1}{4}\mathrm{Re}(\mathbf{E}\cdot\bar{\mathbf{D}}),$$

$$\frac{1}{T}\int_0^T \frac{1}{2}\mathbf{H}\cdot\mathbf{B}dt = \frac{1}{4}\mathrm{Re}(\mathbf{H}\cdot\bar{\mathbf{B}}).$$

It should be noted that the Poynting theorem (1.23) in time domain and the Poynting theorem (1.27) in frequency domain are independent. This property can be used to find the stored energies around a small antenna (see Chapter 5).

1.1.5.3 *Uniqueness Theorems*

It is important to know the conditions under which the solution of Maxwell equations is unique. Let us consider a multiple-connected region V bounded by $S = \sum_{i=0}^{N} S_i$, as shown in Figure 1.2. Assume that the medium inside V is linear, isotropic and time invariant, and it may contain some impressed source $\mathbf{J}_{\mathrm{imp}}$. So we have $\mathbf{D} = \varepsilon\mathbf{E}, \mathbf{B} = \mu\mathbf{H}$, and $\mathbf{J}_{\mathrm{ind}} = \sigma\mathbf{E}$. The uniqueness theorem for time-domain fields can be expressed as follows:

Uniqueness theorem for time-domain fields: Suppose that the electromagnetic sources are turned on at $t = 0$. The electromagnetic fields in a region are uniquely determined by the sources within the region, the initial electric field and the initial magnetic field at $t = 0$ inside the region,

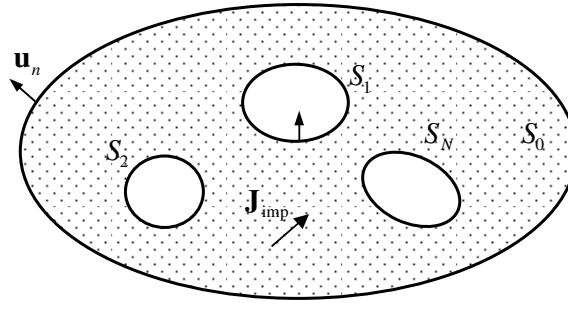


Figure 1.2 Multiple-connected region.

together with the tangential electric field (or the tangential magnetic field) on the boundary for $t > 0$, or together with the tangential electric field on part of the boundary and the tangential magnetic field on the rest of the boundary for $t > 0$.

The uniqueness theorem for time-harmonic fields may be stated as follows:

Uniqueness theorem for time-harmonic fields: For a region that contains the dissipation loss or radiation loss, the electromagnetic fields are uniquely determined by the sources within the region, together with the tangential electric field (or the tangential magnetic field) on the boundary, or together with the tangential electric field on part of the boundary and the tangential magnetic field on the rest of the boundary.

The uniqueness for time-harmonic fields is guaranteed if the system has radiation loss, regardless of the medium is lossy or not. This property has been widely validated by the study of antenna radiation problems, in which the surrounding medium is often assumed to be lossless. Note that the uniqueness for time-harmonic fields fails for a system that contains no dissipation loss and radiation loss. The uniqueness in a lossless medium is usually obtained by considering the fields in a lossless medium to be the limit of the corresponding fields in a lossy medium as the loss goes to zero, which is based on an assumption that the limit of a unique solution is also unique. However, this limiting process may lead to physically unacceptable solutions. Also notice that there is no need to introduce losses for a unique solution in the time-domain analysis (Geyi, 2010).

1.1.5.4 *Equivalence Theorems*

It is known that there is no answer to the question of whether field or source is primary. The equivalence principles just indicate that the distinction between the field and source is kind of blurred. Let V be an arbitrary region bounded by S, as shown in Figure 1.3. Two sources that produce the same fields inside a region are said to be equivalent within that region. Similarly, two electromagnetic fields $\{\mathbf{E}_1, \mathbf{D}_1, \mathbf{H}_1, \mathbf{B}_1\}$ and $\{\mathbf{E}_2, \mathbf{D}_2, \mathbf{H}_2, \mathbf{B}_2\}$ are said to be equivalent inside a region if they both satisfy the Maxwell equations and are equal in that region.

The main application of the equivalence theorem is to find equivalent sources to replace the influences of substance (the medium is homogenized), so that the formulae for retarding potentials can be used. The equivalent sources may be located inside S (equivalent volume sources) or on S

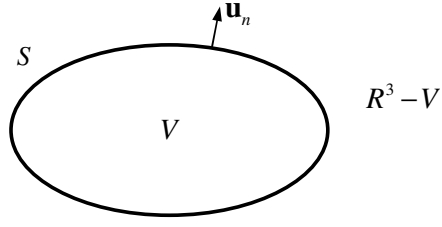


Figure 1.3 Equivalence theorem.

(equivalent surface sources). The most general form of the equivalent principles is stated as follows.

General equivalence principle: Let us consider two electromagnetic field problems in two different media:

$$\text{Problem 1:} \begin{cases} \nabla \times \mathbf{H}_1(\mathbf{r},t) = \partial \mathbf{D}_1(\mathbf{r},t)/\partial t + \mathbf{J}_1(\mathbf{r},t), \\ \nabla \times \mathbf{E}_1(\mathbf{r},t) = -\partial \mathbf{B}_1(\mathbf{r},t)/\partial t - \mathbf{J}_{m1}(\mathbf{r},t), \\ \nabla \cdot \mathbf{D}_1(\mathbf{r},t) = \rho_1(\mathbf{r},t),\ \nabla \cdot \mathbf{B}_1(\mathbf{r},t) = \rho_{m1}(\mathbf{r},t), \\ \mathbf{D}_1(\mathbf{r},t) = \varepsilon_1(\mathbf{r})\mathbf{E}_1(\mathbf{r},t), \mathbf{B}_1(\mathbf{r},t) = \mu_1(\mathbf{r})\mathbf{H}_1(\mathbf{r},t) \end{cases}$$

$$\text{Problem 2:} \begin{cases} \nabla \times \mathbf{H}_2(\mathbf{r},t) = \partial \mathbf{D}_2(\mathbf{r},t)/\partial t + \mathbf{J}_2(\mathbf{r},t), \\ \nabla \times \mathbf{E}_2(\mathbf{r},t) = -\partial \mathbf{B}_2(\mathbf{r},t)/\partial t - \mathbf{J}_{m2}(\mathbf{r},t), \\ \nabla \cdot \mathbf{D}_2(\mathbf{r},t) = \rho_2(\mathbf{r},t), \nabla \cdot \mathbf{B}_2(\mathbf{r},t) = \rho_{m2}(\mathbf{r},t), \\ \mathbf{D}_2(\mathbf{r},t) = \varepsilon_2(\mathbf{r})\mathbf{E}_2(\mathbf{r},t), \mathbf{B}_2(\mathbf{r},t) = \mu_2(\mathbf{r})\mathbf{H}_2(\mathbf{r},t). \end{cases}$$

If a new set of electromagnetic fields $\{\mathbf{E}, \mathbf{D}, \mathbf{H}, \mathbf{B}\}$ satisfying

$$\begin{cases} \nabla \times \mathbf{H}(\mathbf{r},t) = \partial \mathbf{D}(\mathbf{r},t)/\partial t + \mathbf{J}(\mathbf{r},t), \\ \nabla \times \mathbf{E}(\mathbf{r},t) = -\partial \mathbf{B}(\mathbf{r},t)/\partial t - \mathbf{J}_m(\mathbf{r},t), \\ \nabla \cdot \mathbf{D}(\mathbf{r},t) = \rho(\mathbf{r},t), \nabla \cdot \mathbf{B}(\mathbf{r},t) = \rho_m(\mathbf{r},t), \\ \mathbf{D}(\mathbf{r},t) = \varepsilon(\mathbf{r})\mathbf{E}(\mathbf{r},t), \mathbf{B}(\mathbf{r},t) = \mu(\mathbf{r})\mathbf{H}(\mathbf{r},t), \end{cases} \tag{1.28}$$

is constructed in such a way that the sources of the fields $\{\mathbf{E}, \mathbf{D}, \mathbf{H}, \mathbf{B}\}$ and the parameters of the medium satisfy

$$\begin{cases} \mathbf{J} = \mathbf{J}_1, \mathbf{J}_m = \mathbf{J}_{m1} \\ \rho = \rho_1, \rho_m = \rho_{m1}, \ \ \mathbf{r} \in V; \\ \mu = \mu_1, \varepsilon = \varepsilon_2 \end{cases} \qquad \begin{cases} \mathbf{J} = \mathbf{J}_2, \mathbf{J}_m = \mathbf{J}_{m2} \\ \rho = \rho_2, \rho_m = \rho_{m2}, \ \ \mathbf{r} \in R^3 - V \\ \mu = \mu_2, \varepsilon = \varepsilon_2 \end{cases}$$

and

$$\begin{cases} \mathbf{J} = \mathbf{u}_n \times (\mathbf{H}_{2+} - \mathbf{H}_{1-}) \\ \mathbf{J}_m = -\mathbf{u}_n \times (\mathbf{E}_{2+} - \mathbf{E}_{1-}) \\ \rho = \mathbf{u}_n \cdot (\mathbf{D}_{2+} - \mathbf{D}_{1-}) \\ \rho_m = \mathbf{u}_n \cdot (\mathbf{B}_{2+} - \mathbf{B}_{1-}) \end{cases}, \quad \mathbf{r} \in S$$

where $\mathbf{u}_n$ is the unit outward normal to S, and the subscripts $+$ and $-$ signify the values obtained as S is approached from outside S and inside S respectively, then we have

$$\{\mathbf{E}, \mathbf{D}, \mathbf{H}, \mathbf{B}\} = \{\mathbf{E}_1, \mathbf{D}_1, \mathbf{H}_1, \mathbf{B}_1\}, \quad \mathbf{r} \in V$$
$$\{\mathbf{E}, \mathbf{D}, \mathbf{H}, \mathbf{B}\} = \{\mathbf{E}_2, \mathbf{D}_2, \mathbf{H}_2, \mathbf{B}_2\}, \quad \mathbf{r} \in R^3 - V$$

By the equivalence principle, the magnetic current $\mathbf{J}_m$ and magnetic charge ρ_m, introduced in the generalized Maxwell equations, are justified in the sense of equivalence. If $\mathbf{E}_1 = \mathbf{D}_1 = \mathbf{H}_1 = \mathbf{B}_1 = \mathbf{J}_1 = \mathbf{J}_{m1} = 0$ in the general equivalence theorem, we may choose $\mu = \mu_2$, $\varepsilon = \varepsilon_2$ in (1.28) inside S. If all the sources for Problem 2 are contained inside S, the following sources

$$\begin{cases} \mathbf{J}_s = \mathbf{u}_n \times \mathbf{H}_{2+}, \mathbf{J}_{ms} = -\mathbf{u}_n \times \mathbf{E}_{2+} \\ \rho_s = \mathbf{u}_n \cdot \mathbf{D}_{2+}, \rho_{ms} = \mathbf{u}_n \cdot \mathbf{B}_{2+} \end{cases}, \quad \mathbf{r} \in S$$

produce the electromagnetic fields $\{\mathbf{E}, \mathbf{D}, \mathbf{H}, \mathbf{B}\}$ in (1.28). In other words, the above sources generate the fields $\{\mathbf{E}_2, \mathbf{D}_2, \mathbf{H}_2, \mathbf{B}_2\}$ in $R^3 - V$ and a zero field in V. Thus we have:

Schelkunoff–Love equivalence (named after the American mathematician Sergei Alexander Schelkunoff, 1897–1992; and the English mathematician Augustus Edward Hough Love, 1863–1940): Let $\{\mathbf{E}, \mathbf{D}, \mathbf{H}, \mathbf{B}\}$ be the electromagnetic fields with source confined in S. The following surface sources

$$\begin{cases} \mathbf{J}_s = \mathbf{u}_n \times \mathbf{H}, \ \mathbf{J}_{ms} = -\mathbf{u}_n \times \mathbf{E} \\ \rho_s = \mathbf{u}_n \cdot \mathbf{D}, \ \rho_{ms} = \mathbf{u}_n \cdot \mathbf{B} \end{cases}, \quad \mathbf{r} \in S \tag{1.29}$$

produce the same fields $\{\mathbf{E}, \mathbf{D}, \mathbf{H}, \mathbf{B}\}$ outside S and a zero field inside S.

Since the sources in (1.29) produce a zero field inside S, the interior of S may be filled with a perfect electric conductor. By use of the Lorentz reciprocity theorem [see (1.32)], it can be shown that the surface electric current pressed tightly on the perfect conductor does not produce fields. As a result, only the surface magnetic current is needed in (1.29). Similarly,

the interior of S may be filled with a perfect magnetic conductor, and in this case the surface magnetic current does not produce fields and only the surface electric current is needed in (1.29). In both cases, one cannot directly apply the vector potential formula even if the medium outside S is homogeneous.

1.1.5.5 *Reciprocity*

A linear system is said to be **reciprocal** if the response of the system with a particular load and a source is the same as the response when the source and the load are interchanged. Consider two sets of time-harmonic sources, $\mathbf{J}_1, \mathbf{J}_{m1}$ and $\mathbf{J}_2, \mathbf{J}_{m2}$, of the same frequency in the same linear medium. The fields produced by the two sources are respectively denoted by $\mathbf{E}_1, \mathbf{H}_1$ and $\mathbf{E}_2, \mathbf{H}_2$, and they satisfy the Maxwell equations

$$\begin{cases} \nabla \times \mathbf{H}_i(\mathbf{r}) = j\omega\varepsilon \mathbf{E}_i(\mathbf{r}) + \mathbf{J}_i(\mathbf{r}) \\ \nabla \times \mathbf{E}_i(\mathbf{r}) = -j\omega\mu \mathbf{H}_i(\mathbf{r}) - \mathbf{J}_{mi}(\mathbf{r}) \end{cases}, \quad (i = 1, 2).$$

The reciprocity can be stated as

$$\int_V (\mathbf{E}_2 \cdot \mathbf{J}_1 - \mathbf{H}_2 \cdot \mathbf{J}_{m1}) dV = \int_V (\mathbf{E}_1 \cdot \mathbf{J}_2 - \mathbf{H}_1 \cdot \mathbf{J}_{m2}) dV + \int_S (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS, \quad (1.30)$$

where V is a finite region bounded by S. If both sources are outside S, the surface integral in (1.30) is zero. If both sources are inside S, it can be shown that the surface integral is also zero by using the radiation condition. Therefore, we obtain the Lorentz form of reciprocity

$$\int_S (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS = 0 \quad (1.31)$$

and the Rayleigh–Carson form of reciprocity

$$\int_V (\mathbf{E}_2 \cdot \mathbf{J}_1 - \mathbf{H}_2 \cdot \mathbf{J}_{m1}) dV = \int_V (-\mathbf{H}_1 \cdot \mathbf{J}_{m2} + \mathbf{E}_1 \cdot \mathbf{J}_2) dV. \quad (1.32)$$

If the surface S only contains the sources $\mathbf{J}_1(\mathbf{r})$ and $\mathbf{J}_{m1}(\mathbf{r})$, (1.30) becomes

$$\int_V (\mathbf{E}_2 \cdot \mathbf{J}_1 - \mathbf{H}_2 \cdot \mathbf{J}_{m1}) dV = \int_S (\mathbf{E}_2 \cdot \mathbf{u}_n \times \mathbf{H}_1 - \mathbf{H}_2 \cdot \mathbf{E}_1 \times \mathbf{u}_n) dS.$$

This is the familiar form of Huygens' principle. The electromagnetic reciprocity theorem can also be generalized to an anisotropic medium (Kong, 1990; Tai, 1961; Harrington, 1958). Let us consider a special case where the region V does not contain any sources. We denote the fields inside the region V by $(\mathbf{E}_1, \mathbf{H}_1)$ or $(\mathbf{E}_2, \mathbf{H}_2)$ when it is endowed with medium parameters $(\overleftrightarrow{\boldsymbol{\mu}}, \overleftrightarrow{\boldsymbol{\varepsilon}})$ or with transposed medium parameters $(\overleftrightarrow{\boldsymbol{\mu}}^t, \overleftrightarrow{\boldsymbol{\varepsilon}}^t)$. It follows from the Maxwell equations in source-free region that

$$\begin{aligned}
\nabla \cdot (\mathbf{E}_1 \times \mathbf{H}_2) &= \mathbf{H}_2 \cdot \nabla \times \mathbf{E}_1 - \mathbf{E}_1 \cdot \nabla \times \mathbf{H}_2 \\
&= -j\omega \mathbf{H}_2 \cdot \overleftrightarrow{\boldsymbol{\mu}} \cdot \mathbf{H}_1 - j\omega \mathbf{E}_1 \cdot \overleftrightarrow{\boldsymbol{\varepsilon}}^t \cdot \mathbf{E}_2, \\
\nabla \cdot (\mathbf{E}_2 \times \mathbf{H}_1) &= \mathbf{H}_1 \cdot \nabla \times \mathbf{E}_2 - \mathbf{E}_2 \cdot \nabla \times \mathbf{H}_1 \\
&= -j\omega \mathbf{H}_1 \cdot \overleftrightarrow{\boldsymbol{\mu}}^t \cdot \mathbf{H}_2 - j\omega \mathbf{E}_2 \cdot \overleftrightarrow{\boldsymbol{\varepsilon}} \cdot \mathbf{E}_1.
\end{aligned}$$

This gives

$$\nabla \cdot (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) = 0.$$

After integration over V, we obtain (1.31).

1.2 Method of Separation of Variables

Let us consider a differential equation

$$\hat{L}u = f, \tag{1.33}$$

where $\hat{L}$ is a differential operator, f is a known source function and u is the unknown function. One method of solving (1.33) is to find the spectral representation of $\hat{L}$ by studying the solution of the following eigenvalue equation

$$\hat{L}u = \lambda u,$$

where λ is called **eigenvalue** and u is the corresponding **eigenfunction**. The **method of eigenfunction expansion** is also called the **method of separation of variables** if $\hat{L}$ is a partial differential operator. The basic idea of separation of variables is to seek a solution in the form of a product of functions, each of which depends on one variable only, so that the solution of original partial differential equations may reduce to the solution of ordinary differential equations. We will use the Helmholtz equation to illustrate the procedure in this section. The **Helmholtz equation**, named after the

German physicist Hermann Ludwig Ferdinand von Helmholtz (1821–1894), is the time-independent form of wave equation, and is defined by

$$(\nabla^2 + k^2)u = 0, \tag{1.34}$$

where k is a constant. When k is zero, the Helmholtz equation reduces to the Laplace equation, named after the French mathematician Pierre-Simon marquis de Laplace (1749–1827). The Helmholtz equation is separable in 11 orthogonal coordinate systems (Eisenhart, 1934).

1.2.1 Eigenvalue Problem of Sturm–Liouville Type

First let us consider the most common eigenvalue problem for the ordinary differential equation known as the **Sturm–Liouville equation** [named after the French mathematicians Jacques Charles François Sturm (1803–1855) and Joseph Liouville (1809–1882)]

$$\left[-\frac{d}{dx}p(x)\frac{d}{dx} + q(x)\right] v_n(x) = \lambda_n w(x) v_n(x), \quad a < x < b \tag{1.35}$$

subject to the homogeneous boundary conditions of **impedance type**:

$$p(x)\frac{dv_n(x)}{dx} + \alpha(x)v_n(x) = 0, \quad x = a, b. \tag{1.36}$$

In the above, λ_n is the eigenvalue and v_n is the corresponding eigenfunction. The functions p, q and the weight function w are assumed to be real functions of x in $[a, b]$ and furthermore $w > 0$. Multiplying (1.35) by v_n, integrating over x between a and b and using integration by parts, we obtain

$$\lambda_n = \frac{\int_a^b p\left(\frac{dv_n}{dx}\right)^2 dx + \int_a^b q v_n^2\, dx + \alpha(b)v_n^2(b) - \alpha(a)v_n^2(a)}{\int_a^b w v_n^2\, dx}. \tag{1.37}$$

This indicates that λ_n is real. We now multiply (1.35) by the eigenfunction v_m and integrate over the x domain to obtain

$$\int_a^b v_m \frac{d}{dx}\left(p\frac{dv_n}{dx}\right)dx - \int_a^b q v_m v_n\, dx + \lambda_n \int_a^b w v_m v_n\, dx = 0. \tag{1.38}$$

Interchanging m and n gives another equation

$$\int_a^b v_n \frac{d}{dx}\left(p\frac{dv_m}{dx}\right)dx - \int_a^b q v_n v_m\,dx + \lambda_m \int_a^b w v_n v_m\,dx = 0. \tag{1.39}$$

Subtracting (1.38) from (1.39) yields

$$(\lambda_m - \lambda_n)\int_a^b w v_n v_m\,dx = \left[p\left(v_n\frac{dv_m}{dx} - v_m\frac{dv_n}{dx}\right)\right]_a^b.$$

In view of the boundary conditions (1.36), we obtain the following orthogonal relationship

$$\int_a^b w v_n v_m\,dx = 0, \quad m \neq n. \tag{1.40}$$

The eigenfunctions may be normalized as follows

$$\int_a^b w(x) v_n^2(x)dx = 1. \tag{1.41}$$

The set of eigenfunctions $\{v_n\}$ is said to be **orthonormal** if both (1.40) and (1.41) are satisfied. If we assume $v_n(a) = v_n(b) = 0$, then all the eigenvalues are positive from (1.37). Suppose that the eigenfunctions are complete and therefore every square integrable function $f(x)$ in $[a, b]$ can be represented by

$$f(x) = \sum_n f_n v_n, \tag{1.42}$$

where the sum is over all eigenfunctions, and

$$f_n = \int_a^b w(x) f(x) v_n(x) dx.$$

The completeness and orthonormality of the set $\{v_n\}$ can be expressed concisely in a symbolic manner by choosing $f(x) = \delta(x - x')$ in (1.42)

$$\frac{\delta(x - x')}{w(x')} = \sum_n v_n(x) v_n(x'), \quad a < x,\ \ x' < b. \tag{1.43}$$

1.2.2 Rectangular Coordinate System

In rectangular coordinate system (x, y, z), Helmholtz equation (1.34) becomes

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} + k^2 u = 0. \tag{1.44}$$

We seek a solution in the form of product of three functions of one coordinate each

$$u = X(x)Y(y)Z(z). \tag{1.45}$$

Substituting (1.45) into (1.44) gives

$$\frac{1}{X}\frac{d^2X}{dx^2} + \frac{1}{Y}\frac{d^2Y}{dy^2} + \frac{1}{Z}\frac{d^2Z}{dz^2} + k^2 = 0. \tag{1.46}$$

Since k is a constant and each term depends on one variable only and can change independently, the left-hand side of (1.46) can sum to zero for all coordinate values only if each term is a constant. Thus we have

$$\begin{aligned} &\frac{d^2X}{dx^2} + k_x^2 X = 0, \\ &\frac{d^2Y}{dy^2} + k_y^2 Y = 0, \\ &\frac{d^2Z}{dz^2} + k_z^2 Z = 0, \end{aligned} \tag{1.47}$$

where k_x, k_y and k_z are separation constants and satisfy

$$k_x^2 + k_y^2 + k_z^2 = k^2. \tag{1.48}$$

The solutions of (1.47) are harmonic functions, denoted by $X(k_x x)$, $Y(k_y y)$ and $Z(k_z z)$, and they are any linear combination of the following independent **harmonic functions**:

$$e^{ik_\alpha \alpha}, \quad e^{-ik_\alpha \alpha}, \quad \cos k_\alpha \alpha, \quad \sin k_\alpha \alpha \, (\alpha = x, y, z). \tag{1.49}$$

Consequently, the solution (1.45) may be expressed as

$$u = X(k_x x)Y(k_y y)Z(k_z z). \tag{1.50}$$

The separation constants k_x, k_y and k_z are also called eigenvalues, and they are determined by the boundary conditions. The corresponding solutions (1.50) are called eigenfunctions or elementary wavefunctions. The general

solution of (1.44) can be expressed as a linear combination of the elementary wavefunctions.

1.2.3 Cylindrical Coordinate System

In cylindrical coordinate system (ρ, φ, z), (1.34) can be written as

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial u}{\partial\rho}\right)+\frac{1}{\rho^2}\frac{\partial^2 u}{\partial\varphi^2}+\frac{\partial^2 u}{\partial z^2}+k^2u=0. \tag{1.51}$$

By the method of separation of variables, the solutions may be assumed to be

$$u=R(\rho)\Phi(\varphi)Z(z). \tag{1.52}$$

Introducing (1.52) into (1.51) yields

$$\begin{aligned}&\frac{d^2R}{d\rho^2}+\frac{1}{\rho}\frac{dR}{d\rho}+\left(\mu^2-\frac{p^2}{\rho^2}\right)R=0,\\&\frac{d^2\Phi}{d\varphi^2}+p^2\Phi=0,\\&\frac{d^2Z}{dz^2}+\beta^2Z=0,\end{aligned} \tag{1.53}$$

where μ, p and β are separation constants and satisfy

$$\beta^2+\mu^2=k^2. \tag{1.54}$$

The first equation of (1.53) is **Bessel equation**, named after the German mathematician Friedrich Wilhelm Bessel (1784–1846), whose solutions are **Bessel functions**:

$$J_p(\mu\rho),\quad N_p(\mu\rho),\quad H_p^{(1)}(\mu\rho),\quad H_p^{(2)}(\mu\rho),$$

where $J_p(\mu\rho)$ and $N_p(\mu\rho)$ are the Bessel functions of the first and second kind, $H_p^{(1)}(\mu\rho)$ and $H_p^{(2)}(\mu\rho)$ are the Bessel functions of the third and fourth kind, also called **Hankel functions** of first and second kind respectively, named after German mathematician Hermann Hankel (1839–1873). The Bessel function of the first kind is defined by

$$J_p(\mu z)=\sum_{m=0}^{\infty}\frac{(-1)^m}{\Gamma(m+1)\Gamma(p+m+1)}\left(\frac{\mu z}{2}\right)^{p+2m}, \tag{1.55}$$

where $\Gamma(\alpha)$ is the gamma function defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \quad \alpha > 0.$$

If p is not an integer, a second independent solution is $J_{-p}(\mu z)$. If $p = n$ is an integer, $J_{-n}(\mu z)$ is related to $J_n(\mu z)$ by

$$J_{-n}(z) = (-1)^n J_n(z).$$

The Bessel function of the second kind (also known as **Neumann function**) defined by

$$N_p(\mu z) = \frac{\cos p\pi J_p(\mu z) - J_{-p}(\mu z)}{\sin p\pi}, \tag{1.56}$$

and the Bessel functions of the third (Hankel function of the first kind) and fourth kind (Hankel function of the second kind) are defined by

$$\begin{aligned} H_p^{(1)}(\mu z) &= J_p(\mu z) + jN_p(\mu z), \\ H_p^{(2)}(\mu z) &= J_p(\mu z) - jN_p(\mu z). \end{aligned} \tag{1.57}$$

The solutions of second and third equation of (1.53) are harmonic functions. Note that only $J_p(\mu\rho)$ is finite at $\rho = 0$. The separation constants μ and p are determined by the boundary conditions. For example, if the field u is finite and satisfies homogeneous Dirichlet boundary condition $u = 0$ at $\rho = a$, the separation constant μ is determined by $J_p(\mu\rho) = 0$. If the cylindrical region contains all φ from 0 to 2π, the separation constant p is usually determined by the requirement that the field is single-valued, i.e., $\Phi(0) = \Phi(2\pi)$. In this case, p must be integers. If the cylindrical region only contains a circular sector, p will be fractional numbers.

Let $R_p(\mu z) = AJ_p(\mu z) + BN_p(\mu z)$, where A and B are constant. We have the recurrence relations

$$\frac{2p}{\mu z} R_p(\mu z) = R_{p-1}(\mu z) + R_{p+1}(\mu z),$$

$$\frac{1}{\mu}\frac{d}{dz} R_p(\mu z) = \frac{1}{2}\left[R_{p-1}(\mu z) - R_{p+1}(\mu z)\right],$$

$$z\frac{d}{dz} R_p(\mu z) = pR_p(\mu z) - \mu z R_{p+1}(\mu z),$$

$$\frac{d}{dz}[z^p R_p(\mu z)] = \mu z^p R_{p-1}(\mu z),$$
$$\frac{d}{dz}[z^{-p} R_p(\mu z)] = -\mu z^{-p} R_{p+1}(\mu z).$$

1.2.4 Spherical Coordinate System

In spherical coordinate system (r, θ, φ), (1.34) can be expressed as

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial u}{\partial\theta}\right) + \frac{1}{r^2\sin\theta}\frac{\partial^2 u}{\partial\varphi^2} + k^2 u = 0. \tag{1.58}$$

By means of the separation of variables, we may let

$$u = R(r)\Theta(\theta)\Phi(\varphi). \tag{1.59}$$

Substitution of (1.59) into (1.58) leads to

$$\begin{aligned} &\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) + k^2 r^2 = \beta^2, \\ &\frac{1}{\Theta\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) - \frac{m^2}{\sin^2\theta} = -\beta^2, \\ &\frac{d^2\Phi}{d\varphi^2} + m^2\Phi = 0. \end{aligned} \tag{1.60}$$

Let $x = \cos\theta$ and $P(x) = \Theta(\theta)$, the second equation of (1.60) becomes

$$(1-x^2)\frac{d^2P}{dx^2} - 2x\frac{dP}{dx} + \left(\beta^2 - \frac{m^2}{1-x^2}\right)P = 0. \tag{1.61}$$

This is called **Legendre equation**, named after the French mathematician Adrien-Marie Legendre (1752–1833). The points $x = \pm 1$ are singular. Equation (1.61) has two linearly independent solutions and can be expressed as a power series at $x = 0$. In general, the series solution diverges at $x = \pm 1$. But if we let $\beta^2 = n(n+1), n = 0, 1, 2, \ldots$, the series will be finite at $x = \pm 1$ and have finite terms. Thus the separation constant β is determined

naturally and (1.60) can be written as

$$\begin{aligned}
&\frac{d}{dr}\left(r^2\frac{dR}{dr}\right)+\left[k^2r^2-n(n+1)\right]R=0,\\
&(1-x^2)\frac{d^2P}{dx^2}-2x\frac{dP}{dx}+\left[n(n+1)-\frac{m^2}{1-x^2}\right]P=0,\\
&\frac{d^2\Phi}{d\varphi^2}+m^2\Phi=0.
\end{aligned}\tag{1.62}$$

The solutions of the first equation of (1.62) are **spherical Bessel functions**

$$\begin{aligned}
j_n(kr)&=\sqrt{\frac{\pi}{2kr}}J_{n+1/2}(kr),\quad n_n(kr)=\sqrt{\frac{\pi}{2kr}}N_{n+1/2}(kr),\\
h_n^{(1)}(kr)&=\sqrt{\frac{\pi}{2kr}}H_{n+1/2}^{(1)}(kr),\quad h_n^{(2)}(kr)=\sqrt{\frac{\pi}{2kr}}H_{n+1/2}^{(2)}(kr).
\end{aligned}\tag{1.63}$$

Let $z_n(kr)=Aj_n(kr)+Bn_n(kr)$, where A and B are constants. We have the recurrence relations:

$$\begin{aligned}
\frac{2n+1}{kr}z_n(kr)&=z_{n-1}(kr)+z_{n+1}(kr),\\
\frac{2n+1}{k}\frac{d}{dr}z_n(kr)&=nz_{n-1}(kr)-(n+1)z_{n+1}(kr),\\
\frac{d}{dr}[r^{n+1}z_n(kr)]&=kr^{n+1}z_{n-1}(kr),\\
\frac{d}{dr}[r^{-n}z_n(kr)]&=-kr^{-n}z_{n+1}(kr).
\end{aligned}$$

The solutions of the second equation of (1.62) are **associated Legendre functions** of first and second kind defined by

$$P_n^m(x)=\frac{(1-x^2)^{m/2}}{2^n n!}\frac{d^{m+n}}{dx^{m+n}}(x^2-1)^n,\tag{1.64}$$

and

$$Q_n^m(x)=(1-x^2)^{\frac{m}{2}}\frac{d^m}{dx^m}Q_n(x),\quad m\le n,\tag{1.65}$$

respectively, with

$$Q_n(x) = \frac{1}{2}P_n^0(x)\ln\frac{1+x}{1-x} - \sum_{r=1}^{n}\frac{1}{r}P_{r-1}^0(x)P_{n-r}^0(x)$$

being the **Legendre function of the second kind**.

The following integrations are useful

$$\int_{-1}^{1}\frac{P_n^m(x)P_n^k(x)}{1-x^2}dx = \frac{1}{m}\frac{(n+m)!}{(n-m)!}\delta_{mk},$$

$$\int_{-1}^{1}P_k^m(x)P_n^m(x)dx = \frac{2}{2k+1}\frac{(k+m)!}{(k-m)!}\delta_{kn},$$

$$\int_{0}^{\pi}\left[\frac{dP_n^m(\cos\theta)}{d\theta}\frac{dP_k^m(\cos\theta)}{d\theta} + \frac{m^2}{\sin^2\theta}P_n^m(\cos\theta)P_k^m(\cos\theta)\right]\sin\theta\,d\theta$$

$$= \frac{2}{2n+1}\frac{(n+m)!}{(n-m)!}n(n+1)\delta_{nk}.$$

The solutions of the third equation of (1.62) are harmonic functions. Note that the separation coefficients are not related in spherical coordinate system.

1.3 Method of Green's Functions

Physically, the Green's function represents the field produced by a point source, and provides a general method to solve differential equations. Through the use of the Green's function, the solution of a differential equation can be represented by an integral defined over the source region or on a closed surface enclosing the source. Mathematically, the solution of a partial differential equation in the source region V

$$\hat{L}u(\mathbf{r}) = f(\mathbf{r}), \quad \mathbf{r}\in V \tag{1.66}$$

can be expressed as

$$u(\mathbf{r}) = \hat{L}^{-1}f(\mathbf{r}),$$

where $\hat{L}^{-1}$ stands for the inverse of $\hat{L}$ and is often represented by an integral operator whose kernel is the Green's function. Let us assume that there exists a function G such that

$$\hat{L}^{-1}f(\mathbf{r}) = -\int_V G(\mathbf{r},\mathbf{r}')f(\mathbf{r}')dV(\mathbf{r}'), \tag{1.67}$$

Applying $\hat{L}$ to both sides of the above equation yields

$$\hat{L}\hat{L}^{-1}f(\mathbf{r}) = f(\mathbf{r}) = -\int_V \hat{L}G(\mathbf{r},\mathbf{r}')f(\mathbf{r}')dV(\mathbf{r}').$$

This equation implies that the function G satisfies

$$\hat{L}G(\mathbf{r},\mathbf{r}')f(\mathbf{r}) = -\delta(\mathbf{r}-\mathbf{r}') \tag{1.68}$$

where δ denotes the delta function. The function G is called the **fundamental solution** or **Green's function** of the Equation (1.66).

1.3.1 Green's Functions for Helmholtz Equation

Let $\boldsymbol{\rho} = (x, y)$, $\mathbf{r} = (x, y, z)$ and v be a constant. The fundamental solutions of wave equations are summarized in Table 1.5, where $H(x)$ is the unit step function. It can be seen that the Green's functions are symmetric $G(\mathbf{r},\mathbf{r}') = G(\mathbf{r}',\mathbf{r})$.

Example 1.1: The Green's function for one-dimensional Helmholtz equation satisfies

$$\begin{aligned} &\frac{d^2G(z,z')}{dz^2} + k^2G(z,z') = -\delta(z-z'), \\ &\lim_{z\to\pm\infty}\left(\frac{dG}{dz} \pm jkG\right) = 0. \end{aligned} \tag{1.69}$$

The second equation denotes the radiation condition at infinity. Let

$$G(z,z') = \begin{cases} G_1(z,z'), & z < z' \\ G_2(z,z'), & z > z' \end{cases}.$$

Then we may write

$$\begin{aligned} G_1(z,z') &= a_1e^{-jk(z-z')} + b_1e^{jk(z-z')}, \\ G_2(z,z') &= a_2e^{-jk(z-z')} + b_2e^{jk(z-z')}, \end{aligned}$$

where a_1, b_1, a_2, b_2 are constants to be determined. Taking the radiation condition into account, we have $a_1 = b_2 = 0$. Thus

$$\begin{aligned} G_1(z,z') &= b_1e^{jk(z-z')}, \\ G_2(z,z') &= a_2e^{-jk(z-z')}. \end{aligned} \tag{1.70}$$

Table 1.5 Green's functions

Equations	Green's functions
2D Laplace equation: $\nabla^2 G(\boldsymbol{\rho},\boldsymbol{\rho}') = -\delta(\boldsymbol{\rho}-\boldsymbol{\rho}')$	$G(\boldsymbol{\rho},\boldsymbol{\rho}') = -\dfrac{1}{2\pi}\ln\lvert\boldsymbol{\rho}-\boldsymbol{\rho}'\rvert$
3D Laplace equation: $\nabla^2 G(\mathbf{r},\mathbf{r}') = -\delta(\mathbf{r}-\mathbf{r}')$	$G(\mathbf{r},\mathbf{r}') = \dfrac{1}{4\pi\lvert\mathbf{r}-\mathbf{r}'\rvert}$
2D Helmholtz equation: $(\nabla^2+k^2)G(\boldsymbol{\rho},\boldsymbol{\rho}') = -\delta(\boldsymbol{\rho}-\boldsymbol{\rho}')$	$G(\boldsymbol{\rho},\boldsymbol{\rho}') = \dfrac{1}{4j}H_0^{(2)}(k\lvert\boldsymbol{\rho}-\boldsymbol{\rho}'\rvert)$
3D Helmholtz equation: $(\nabla^2+k^2)G(\mathbf{r},\mathbf{r}') = -\delta(\mathbf{r}-\mathbf{r}')$	$G(\mathbf{r},\mathbf{r}') = \dfrac{e^{-jk\lvert\mathbf{r}-\mathbf{r}'\rvert}}{4\pi\lvert\mathbf{r}-\mathbf{r}'\rvert}$
1D wave equation: $\begin{cases}\left(\dfrac{\partial^2}{\partial z^2}-\dfrac{1}{v^2}\dfrac{\partial^2}{\partial t^2}\right)G(z,z';t,t') \\ \quad = -\delta(z-z')\delta(t-t') \\ G(z,z';t,t') = 0, t<t'\end{cases}$	$G(z,z';t,t') = \dfrac{v}{2}H(t-t'-\lvert z-z'\rvert/v)$
2D wave equation: $\begin{cases}\left(\nabla^2-\dfrac{1}{v^2}\dfrac{\partial^2}{\partial t^2}\right)G(\boldsymbol{\rho},\boldsymbol{\rho}';t,t') \\ \quad = -\delta(\boldsymbol{\rho}-\boldsymbol{\rho}')\delta(t-t') \\ G(\boldsymbol{\rho},\boldsymbol{\rho}';t,t') = 0, t<t'\end{cases}$	$G(\boldsymbol{\rho},\boldsymbol{\rho}';t,t') = \dfrac{H(t-t'-\lvert\boldsymbol{\rho}-\boldsymbol{\rho}'\rvert/v)}{2\pi\sqrt{(t-t')^2-\lvert\boldsymbol{\rho}-\boldsymbol{\rho}'\rvert/v}}$
3D wave equation: $\begin{cases}\left(\nabla^2-\dfrac{1}{v^2}\dfrac{\partial^2}{\partial t^2}\right)G(\mathbf{r},\mathbf{r}';t,t') \\ \quad = -\delta(\mathbf{r}-\mathbf{r}')\delta(t-t') \\ G(\mathbf{r},\mathbf{r}';t,t') = 0, t<t'\end{cases}$	$G(\mathbf{r},\mathbf{r}';t,t') = \dfrac{\delta(t-t'-\lvert\mathbf{r}-\mathbf{r}'\rvert/v)}{4\pi\lvert\mathbf{r}-\mathbf{r}'\rvert}$

From the first equation of (1.69), we have

$$G_1(z,z') = G_2(z,z'),$$
$$\frac{dG_2(z,z')}{dz} - \frac{dG_1(z,z')}{dz} = -1.$$

at $z = z'$. Introducing (1.70) into the above equations gives

$$a_2 = b_1 = \frac{1}{j2k}.$$

Finally, we have

$$G(z,z') = \begin{cases} \dfrac{1}{j2k}e^{jk(z-z')}, & z < z' \\ \dfrac{1}{j2k}e^{-jk(z-z')}, & z > z' \end{cases} = \frac{1}{j2k}e^{-jk|z-z'|}.$$

This is the Green's function for one-dimensional Helmholtz equation.

□

1.3.2 Partial Differential Equations and Integral Equations

The partial differential equations can be converted into integral equations by means of Green's functions. Let V be a bounded region in space and S its boundary. Let $\hat{L}$ be a linear differential operator. The operator $\hat{L}^*$ is called the **formal adjoint** of $\hat{L}$ if there exists a vector function $\mathbf{U}(\mathbf{r})$ such that the relation

$$v(\mathbf{r})\hat{L}u(\mathbf{r}) - u(\mathbf{r})\hat{L}^*v(\mathbf{r}) = \nabla \cdot \mathbf{U}(\mathbf{r})$$

holds for arbitrary functions u and v($\mathbf{U}$ may vary with u and v). The operator $\hat{L}$ becomes **self-adjoint** if the right-hand side of the above equation vanishes. Consider a differential equation

$$\hat{L}u(\mathbf{r}) = f(\mathbf{r}), \quad \mathbf{r} \in \Omega,$$

where u is the unknown function and f is a known source function. We may use the direct method to establish the integral equation. From integration by parts, we obtain

$$\int_V \Big[v(\mathbf{r})\hat{L}u(\mathbf{r}) - u(\mathbf{r})\hat{L}^*v(\mathbf{r})\Big]dV = \int_S b[u(\mathbf{r}), v(\mathbf{r})]dS, \tag{1.71}$$

where u and v are two arbitrary smooth functions and $b(\cdot,\cdot)$ is a bilinear form. If $G(\mathbf{r},\mathbf{r}')$ is the Green's function of $\hat{L}^*$

$$\hat{L}^*G(\mathbf{r},\mathbf{r}') = -\delta(\mathbf{r}-\mathbf{r}'),$$

we may let $v(\mathbf{r}) = G(\mathbf{r},\mathbf{r}')$ in (1.71), yielding

$$\int_\Omega u(\mathbf{r})\delta(\mathbf{r}-\mathbf{r}')dV(\mathbf{r}) - \int_S b[G(\mathbf{r},\mathbf{r}'), u(\mathbf{r})]dS(\mathbf{r})$$

$$= -\int_V G(\mathbf{r},\mathbf{r}')f(\mathbf{r})dV(\mathbf{r}).$$

If S is smooth we may let $\mathbf{r}' \to S$ to obtain

$$\frac{1}{2}u(\mathbf{r}') - \int_S b[G(\mathbf{r},\mathbf{r}'),u(\mathbf{r})]dS(\mathbf{r}) = -\int_V G(\mathbf{r},\mathbf{r}')f(\mathbf{r})dV(\mathbf{r}).$$

By use of the symmetric property of the Green's function $G(\mathbf{r},\mathbf{r}') = G(\mathbf{r}',\mathbf{r})$, the above equation can be written as

$$\frac{1}{2}u(\mathbf{r}) - \int_S b[G(\mathbf{r},\mathbf{r}'),u(\mathbf{r}')]dS(\mathbf{r}') = -\int_V G(\mathbf{r},\mathbf{r}')f(\mathbf{r}')dV(\mathbf{r}').$$

This is the integral equation defined on the boundary S.

Example 1.2: For the differential operator defined by

$$\hat{L}u = a_{11}(x,y)\frac{\partial^2 u}{\partial x^2} + a_{22}(x,y)\frac{\partial^2 u}{\partial y^2} + 2a_{12}(x,y)\frac{\partial^2 u}{\partial x \partial y}$$
$$+ b_1(x,y)\frac{\partial u}{\partial x} + b_2(x,y)\frac{\partial u}{\partial y} + c(x,y)u,$$

the formal adjoint is given by

$$\hat{L}^* v = \frac{\partial^2}{\partial x^2}(a_{11}v) + \frac{\partial^2}{\partial y^2}(a_{22}v) + 2\frac{\partial^2}{\partial x \partial y}(a_{12}v) - \frac{\partial}{\partial x}(b_1 v) - \frac{\partial}{\partial y}(b_2 v) + cv,$$

with

$$\mathbf{U}(x,y) = \mathbf{u}_x \left[a_{11}\left(v\frac{\partial u}{\partial x} - u\frac{\partial v}{\partial x}\right) + a_{12}\left(v\frac{\partial u}{\partial y} - u\frac{\partial v}{\partial y}\right) \right.$$
$$\left. + \left(b_1 - \frac{\partial a_{11}}{\partial x} - \frac{\partial a_{12}}{\partial y}\right) uv \right] + \mathbf{u}_y \left[a_{12}\left(v\frac{\partial u}{\partial x} - u\frac{\partial v}{\partial x}\right) \right.$$
$$\left. + a_{22}\left(v\frac{\partial u}{\partial y} - u\frac{\partial v}{\partial y}\right) + \left(b_2 - \frac{\partial a_{12}}{\partial x} - \frac{\partial a_{22}}{\partial y}\right) uv \right].$$

Apparently, we have $\hat{L} = \nabla^2 = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) = \hat{L}^*$. $\square$

1.3.3 Dyadic Green's Functions

A dyadic is a second order tensor formed by putting two vectors side by side. Its manipulation rules are analogous to that for matrix algebra. Dyadic notation was first established by Josiah Willard Gibbs (1839–1903) in 1884. The application of dyadic Green's function in solving electromagnetic boundary value problem can be traced back to Schwinger's work in early 1940s. In 1950, Levine and Schwinger applied the dyadic Green's

function to investigate the diffraction problem by an aperture in an infinite plane conducting screen (Levine and Schwinger, 1950). In 1953, Morse and Feshbach discussed various applications of dyadic Green's functions (Morse and Feshbach, 1953). A more systematic study of dyadic Green's functions and their applications in electromagnetic engineering can be found in Tai (1994). Consider an electric current element in the direction of $\alpha(\alpha = x, y, z)$ located at $\mathbf{r}'$

$$\mathbf{J}^{(\alpha)}(\mathbf{r}) = -\frac{1}{j\omega\mu}\delta(\mathbf{r} - \mathbf{r}')\mathbf{u}_\alpha,$$

which produces electromagnetic fields $\mathbf{E}^{(\alpha)}(\mathbf{r})$ and $\mathbf{H}^{(\alpha)}(\mathbf{r})$ at $\mathbf{r}$. Let

$$\begin{aligned} \mathbf{G}_e^{(\alpha)}(\mathbf{r}, \mathbf{r}') &= \mathbf{E}^{(\alpha)}(\mathbf{r}), \\ \mathbf{G}_m^{(\alpha)}(\mathbf{r}, \mathbf{r}') &= -j\omega\mu\mathbf{H}^{(\alpha)}(\mathbf{r}). \end{aligned} \tag{1.72}$$

$\mathbf{G}_e^{(\alpha)}(\mathbf{r}, \mathbf{r}')$ and $\mathbf{G}_m^{(\alpha)}(\mathbf{r}, \mathbf{r}')$ are respectively referred to as **electric and magnetic Green's function** along direction α in free space. It follows from Maxwell equations that

$$\begin{aligned} \nabla \times \mathbf{G}_e^{(\alpha)}(\mathbf{r}, \mathbf{r}') &= \mathbf{G}_m^{(\alpha)}(\mathbf{r}, \mathbf{r}'), \\ \nabla \times \mathbf{G}_m^{(\alpha)}(\mathbf{r}, \mathbf{r}') &= \mathbf{u}_\alpha\delta(\mathbf{r} - \mathbf{r}') + k^2\mathbf{G}_e^{(\alpha)}(\mathbf{r}, \mathbf{r}'). \end{aligned}$$

The dyads defined by

$$\overleftrightarrow{\mathbf{G}}_e(\mathbf{r}, \mathbf{r}') = \sum_{\alpha=x,y,z} \mathbf{G}_e^{(\alpha)}(\mathbf{r}, \mathbf{r}')\mathbf{u}_\alpha, \quad \overleftrightarrow{\mathbf{G}}_m(\mathbf{r}, \mathbf{r}') = \sum_{\alpha=x,y,z} \mathbf{G}_m^{(\alpha)}(\mathbf{r}, \mathbf{r}')\mathbf{u}_\alpha$$

are respectively called **electric and magnetic dyadic Green's functions** in free space. Apparently, we have

$$\begin{aligned} \nabla \times \overleftrightarrow{\mathbf{G}}_e(\mathbf{r}, \mathbf{r}') &= \overleftrightarrow{\mathbf{G}}_m(\mathbf{r}, \mathbf{r}'), \\ \nabla \times \overleftrightarrow{\mathbf{G}}_m(\mathbf{r}, \mathbf{r}') &= \overleftrightarrow{\mathbf{I}}\,\delta(\mathbf{r} - \mathbf{r}') + k^2\overleftrightarrow{\mathbf{G}}_e(\mathbf{r}, \mathbf{r}'), \end{aligned} \tag{1.73}$$

where $\overleftrightarrow{\mathbf{I}}$ is the identity dyad. From (1.73), we obtain

$$\begin{aligned} \nabla \times \nabla \times \overleftrightarrow{\mathbf{G}}_e(\mathbf{r}, \mathbf{r}') - k^2\overleftrightarrow{\mathbf{G}}_e(\mathbf{r}, \mathbf{r}') &= \overleftrightarrow{\mathbf{I}}\,\delta(\mathbf{r} - \mathbf{r}'), \\ \nabla \times \nabla \times \overleftrightarrow{\mathbf{G}}_m(\mathbf{r}, \mathbf{r}') - k^2\overleftrightarrow{\mathbf{G}}_m(\mathbf{r}, \mathbf{r}') &= \nabla \times [\overleftrightarrow{\mathbf{I}}\,\delta(\mathbf{r} - \mathbf{r}')]. \end{aligned} \tag{1.74}$$

The free space electric dyadic Green's function $\overleftrightarrow{\mathbf{G}}_e(\mathbf{r},\mathbf{r}')$ may be represented by free space Green's function $G(\mathbf{r},\mathbf{r}')$ as follows

$$\overleftrightarrow{\mathbf{G}}_e(\mathbf{r},\mathbf{r}') = \left(\overleftrightarrow{\mathbf{I}} + \frac{1}{k^2}\nabla\nabla\right) G(\mathbf{r},\mathbf{r}'). \tag{1.75}$$

In fact, the first equation of (1.74) may be written as

$$-\nabla^2\overleftrightarrow{\mathbf{G}}_e(\mathbf{r},\mathbf{r}') - k^2\overleftrightarrow{\mathbf{G}}_e(\mathbf{r},\mathbf{r}') + \nabla\nabla\cdot\overleftrightarrow{\mathbf{G}}_e(\mathbf{r},\mathbf{r}') = \overleftrightarrow{\mathbf{I}}\,\delta(\mathbf{r}-\mathbf{r}'). \tag{1.76}$$

Taking the divergence of the first equation of (1.74) yields

$$\nabla\cdot\overleftrightarrow{\mathbf{G}}_e(\mathbf{r},\mathbf{r}') = -\frac{1}{k^2}\nabla\cdot[\overleftrightarrow{\mathbf{I}}\,\delta(\mathbf{r}-\mathbf{r}')] = -\frac{1}{k^2}\nabla\delta(\mathbf{r}-\mathbf{r}')$$

Substituting the above into (1.76), we obtain

$$\nabla^2\overleftrightarrow{\mathbf{G}}_e(\mathbf{r},\mathbf{r}') + k^2\overleftrightarrow{\mathbf{G}}_e(\mathbf{r},\mathbf{r}') = -\left(\overleftrightarrow{\mathbf{I}} + \frac{1}{k^2}\nabla\nabla\right)\delta(\mathbf{r}-\mathbf{r}').$$

Obviously (1.75) satisfies the above equation.

The free space magnetic dyadic Green's function may be expressed as

$$\overleftrightarrow{\mathbf{G}}_m(\mathbf{r},\mathbf{r}') = \nabla\times\overleftrightarrow{\mathbf{G}}_e(\mathbf{r},\mathbf{r}') = \nabla\times[G(\mathbf{r},\mathbf{r}')\overleftrightarrow{\mathbf{I}}] = \nabla\times[\overleftrightarrow{\mathbf{G}}_0(\mathbf{r},\mathbf{r}')].$$

where $\overleftrightarrow{\mathbf{G}}_0(\mathbf{r},\mathbf{r}') = G(\mathbf{r},\mathbf{r}')\overleftrightarrow{\mathbf{I}}$ satisfies Helmholtz equation

$$(\nabla^2 + k^2)\overleftrightarrow{\mathbf{G}}_0(\mathbf{r},\mathbf{r}') = -\overleftrightarrow{\mathbf{I}}\,\delta(\mathbf{r}-\mathbf{r}'). \tag{1.77}$$

Making use of the Green's identity, we obtain

$$\begin{aligned}\mathbf{E}(\mathbf{r}) = &-j\omega\mu\int_V \overleftrightarrow{\mathbf{G}}_e(\mathbf{r},\mathbf{r}')\cdot\mathbf{J}(\mathbf{r}')dV(\mathbf{r}')\\ &+\int_S j\omega\mu\overleftrightarrow{\mathbf{G}}_e(\mathbf{r},\mathbf{r}')\cdot[\mathbf{u}_n\times\mathbf{H}(\mathbf{r}')]dS(\mathbf{r}')\\ &-\int_S \nabla\times\overleftrightarrow{\mathbf{G}}_e(\mathbf{r},\mathbf{r}')\cdot[\mathbf{u}_n\times\mathbf{E}(\mathbf{r}')]dS(\mathbf{r}'),\end{aligned}$$

$$\mathbf{H}(\mathbf{r}) = \int_V \overleftrightarrow{\mathbf{G}}_e(\mathbf{r},\mathbf{r}') \cdot \nabla' \times \mathbf{J}(\mathbf{r}') dV(\mathbf{r}')$$
$$- \int_S j\omega\varepsilon \overleftrightarrow{\mathbf{G}}_e(\mathbf{r},\mathbf{r}') \cdot [\mathbf{u}_n \times \mathbf{E}(\mathbf{r}')] dS(\mathbf{r}')$$
$$+ \int_S \nabla \times \overleftrightarrow{\mathbf{G}}_e(\mathbf{r},\mathbf{r}') \cdot [\mathbf{u}_n \times \mathbf{H}(\mathbf{r}')] dS(\mathbf{r}').$$

1.3.4 Green's Functions and Spectral Representation

Consider the following eigenvalue problem

$$\hat{L}v = \lambda v.$$

Let $v_1, v_2, \ldots, v_n$ be eigenfunctions corresponding to different eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ associated with the operator $\hat{L}$. We assume that the eigenfunctions are orthonormal

$$(v_m, v_n) = \delta_{mn} = \begin{cases} 1 & m = n \\ 0 & m \neq n \end{cases}.$$

The operator equation

$$(\hat{L} - \lambda)u = f, \tag{1.78}$$

where λ is a complex parameter, can be solved by substituting the expansions

$$u = \sum_n \alpha_n v_n, \quad f = \sum_n \beta_n v_n$$

into (1.78). We may find that

$$\alpha_n = -\frac{\beta_n}{\lambda - \lambda_n}$$

and

$$u = -\sum_n \frac{\beta_n}{\lambda - \lambda_n} v_n. \tag{1.79}$$

Let C_R be a circle of radius R at the origin in the complex λ-plane. Then we have

$$\int_{C_R} u(\lambda) d\lambda = -\sum_n \beta_n v_n \int_{C_R} \frac{d\lambda}{\lambda - \lambda_n},$$

where the sum is over those eigenvalues λ_n contained within the circle. The singularities of the integrand are simple poles with residue of unity at all $\lambda = \lambda_n$ within the contour. Taking the limit as $R \to \infty$ and using the Residue Theorem we obtain

$$\lim_{R\to\infty} \int_{C_R} u(\lambda)d\lambda = -2\pi j \sum_n \beta_n v_n,$$

where the sum is now over all of the eigenfunctions. Therefore

$$f = -\frac{1}{2\pi j} \int_C u(\lambda)d\lambda \tag{1.80}$$

where C is the contour at infinity obtained in the limit operation. As a special case, we consider the Green's function problem

$$(\hat{L} - \lambda)G(\mathbf{r}, \mathbf{r}'; \lambda) = -\delta(\mathbf{r} - \mathbf{r}').$$

From (1.80) we have

$$\delta(\mathbf{r} - \mathbf{r}') = \frac{1}{2\pi j} \int_C G(\mathbf{r}, \mathbf{r}'; \lambda)d\lambda \tag{1.81}$$

which is called the **spectral representation of the delta function** for the operator $\hat{L}$. It can be shown that the Green's function $G(\mathbf{r}, \mathbf{r}'; \lambda)$ is an analytic function of λ except for poles and branch-point singularities. Therefore, the right-hand side of (1.81) reduces to a sum of residuals at the poles (eigenvalues) plus integrals along the branch cut (continuous spectrum) (Friedman, 1956).

Example 1.3: Let $\hat{L} = -\frac{d^2}{dx^2}$. The domain of $\hat{L}$ consist of twice-differentiable functions satisfying the boundary conditions

$$v(0) = v(1) = 0. \tag{1.82}$$

Then we have

$$v_n(x) = \sin(\sqrt{\lambda_n}x), \quad \lambda_n = n^2\pi^2 (n = 1, 2, \ldots).$$

If the boundary conditions (1.82) are replaced by

$$v(0) = 0, \quad v'(1) = \frac{1}{2}v(1), \tag{1.83}$$

then we have

$$v_n(x) = \sin(\sqrt{\lambda_n}x), \quad n = 1, 2, \ldots,$$

where λ_n satisfies

$$\tan\sqrt{\lambda_n} = 2\sqrt{\lambda_n}. \quad \square$$

1.4 Variational Method and Numerical Techniques

Variational method or **calculus of variations** deals with maximizing or minimizing functionals, which are often expressed as definite integrals involving functions and their derivatives. The extremal functions that make the functional stationary (i.e., attain a maximum or minimum value) can be obtained by assuming that the rate of change of the functional is zero. The variational method has found wide applications in mathematical physics. In physics, the **principle of least action** (or more accurately, the **principle of stationary action**) is a variational principle. When the action of a mechanical system is required to be stationary, the equations of motion for the system can be obtained. The principle of least action leads to the development of the Lagrangian and Hamiltonian formulations of classical mechanics. Although these formulations seem difficult to grasp at first, they have some merits that Newton's formulation does not have. For example, they can be easily transferred to the frameworks of relativistic and quantum-mechanical physics. The principle of least action is considered as the core strategy of modern physics. In terms of the principle of least action, the differential equations of a given physical system (i.e., the equations of motion) can be derived by minimizing the action of the system. The original problem, governed by the differential equations, is thus replaced by an equivalent variational problem. Such a procedure is also called the **energy method**. It is commonly believed that the theoretical formulation of a physical law is not complete until the law can be reformulated as a variational problem.

1.4.1 Functional Derivative

Let F be a functional, i.e., a map from some function space into the real axis. Let v be an arbitrary function in the space. The **gradient** or **functional derivative** of F at u, denoted by $\nabla F(u) = \delta F(u)/\delta u$ and used to describe

the rate of change of the functional, is defined by

$$\left(\frac{\delta F(u)}{\delta u}, v\right) = \frac{d}{dt}F(u+tv)\bigg|_{t=0}, \tag{1.84}$$

where $(\cdot,\cdot)$ is an **inner product** defined by the following rules:

(1) Positive definiteness: $(u,u) \geq 0$ and $(u,u) = 0$ if and only if $u = 0$
(2) Hermitian property: $(u,v) = \overline{(v,u)}$
(3) Homogeneity: $(\alpha u, v) = \alpha(u,v)$
(4) Additivity: $(u+v,w) = (u,w) + (v,w)$

where u, v, w are functions, and α is a number. A linear space equipped with an inner product is called an **inner product space**.

Extremum theorem: A necessary condition for a functional F to have an extremum at u is

$$\frac{\delta F(u)}{\delta u} = 0. \tag{1.85}$$

This equation is referred to as the **Lagrangian equation**.

Example 1.4: The **action** of the system is an integral over time of a function called **Lagrangian function** L

$$S(\mathbf{q}) = \int_{t_1}^{t_2} L[t, \mathbf{q}(t), \dot{\mathbf{q}}(t)]dt,$$

where $\mathbf{q}(t) = [q_1(t), q_2(t), q_3(t)]$ are known as the **generalized coordinates**. Suppose $\mathbf{q}(t)$ is the path that renders S to be a minimum. The functional derivative then vanishes

$$\frac{\delta S}{\delta q_i} = 0, \quad i = 1, 2, 3.$$

Let the path $\mathbf{q}(t)$ be changed to $\mathbf{q}(t) + \varepsilon\Delta\mathbf{q}(t)$, where $\Delta\mathbf{q} = (\Delta q_1, \Delta q_2, \Delta q_3)$ is small everywhere in the time interval $[t_1, t_2]$ and the endpoints of the path are assumed to be fixed: $\Delta\mathbf{q}(t_1) = \Delta\mathbf{q}(t_2) = 0$. Assume that the inner product between two real scalar functions u and v is defined by

$$(u,v) = \int_{t_1}^{t_2} uv\, dt.$$

The functional derivative with respect to q_1 can be found by (1.84) as

$$\begin{aligned}\left(\frac{\delta S}{\delta q_1}, \Delta q_1\right) &= \int_{t_1}^{t_2} \frac{\delta S}{\delta q_1} \Delta q_1 \, dt \\ &= \frac{d}{d\varepsilon} \int_{t_1}^{t_2} L(t, q_1 + \varepsilon \Delta q_1, q_2, q_3, \dot{q}_1 + \varepsilon \Delta \dot{q}_1, \dot{q}_2, \dot{q}_3) dt \Big|_{\varepsilon=0} \\ &= \int_{t_1}^{t_2} \left(\Delta q_1 \frac{\partial L}{\partial q_1} - \Delta q_1 \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_1} \right) dt + \Delta q_1 \frac{\partial L}{\partial q_1} \Big|_{t_1}^{t_2} \\ &= \int_{t_1}^{t_2} \left(\frac{\partial L}{\partial q_1} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_1} \right) \Delta q_1 \, dt. \end{aligned} \tag{1.86}$$

Similar expressions can be obtained for the functional derivative with respect to q_2 and q_3. Hence if S has a local extremum at $\mathbf{q}(t)$, we have

$$\frac{\delta S}{\delta q_i} = \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} - \frac{\partial L}{\partial q_i} = 0, \quad i = 1, 2, 3. \tag{1.87}$$

These are the well-known Lagrangian equations in classical mechanics.

□

Example 1.5: An operator $\hat{A}$ on an inner product space is defined as a map from the inner product space to itself. The operator $\hat{A}$ is called **self-adjoint** if it satisfies

$$(\hat{A}u, v) = (u, \hat{A}v).$$

We may introduce the following functional (called **Rayleigh quotient**)

$$\lambda(u) = \frac{(\hat{A}u, u)}{(u, u)}. \tag{1.88}$$

Since $\hat{A}$ is self adjoint, λ can be shown to be real. Thus, for an arbitrary v, we have

$$\begin{aligned}\left(\frac{\delta \lambda}{\delta u}, v\right) &= \frac{d}{dt} \lambda(u + \varepsilon v) \Big|_{\varepsilon=0} = \frac{d}{dt} \frac{(\hat{A}(u + \varepsilon v), u + \varepsilon v)}{(u + \varepsilon v, u + \varepsilon v)} \Big|_{\varepsilon=0} \\ &= \frac{1}{(u, u)} 2\mathrm{Re}(\hat{A}u - \lambda u, v). \end{aligned}$$

If $\delta\lambda/\delta u = 0$, we have

$$\hat{A}u = \lambda u. \tag{1.89}$$

This is the Lagrangian equation for the functional (1.88). □

1.4.2 Variational Expressions and Rayleigh–Ritz Method

For an operator $\hat{A}$ defined on an inner product space, the adjoint operator $\hat{A}^*$ is defined by

$$(\hat{A}u, v) = (u, \hat{A}^* v), \tag{1.90}$$

where u, v are elements in the inner product space. Consider the following operator equations

$$\hat{A}u = f, \tag{1.91}$$

$$\hat{A}^* v = g. \tag{1.92}$$

Then

$$(u, g) = (u, \hat{A}^* v) = (\hat{A}u, v) = (f, v), \tag{1.93}$$

from which we obtain

$$(u, g) = \frac{(u, g)(f, v)}{(\hat{A}u, v)}. \tag{1.94}$$

The expression

$$I_1(u, v) = \frac{(u, g)(f, v)}{(\hat{A}u, v)} \tag{1.95}$$

for (u, g) can be demonstrated to have variational properties. In fact (1.91) can be obtained by letting $\frac{\delta I_1(u,v)}{\delta v} = 0$ and (1.92) can be obtained by letting $\frac{\delta I_1(u,v)}{\delta u} = 0$. Another variational expression is

$$I_2(u, v) = (\hat{A}u, v) - (u, g) - (f, v). \tag{1.96}$$

Similarly (1.91) can be obtained by letting $\frac{\delta I_2(u,v)}{\delta v} = 0$ and (1.92) can be obtained by letting $\frac{\delta I_2(u,v)}{\delta u} = 0$. In the special case where $\hat{A} = \hat{A}^*$ ($\hat{A}$ is a

self-adjoint operator) and $f = g$, (1.95) and (1.96) reduce to

$$I_1(u) = \frac{(u, f)(f, u)}{(\hat{A}u, u)}, \tag{1.97}$$

$$I_2(u) = (\hat{A}u, u) - (u, f) - (f, u). \tag{1.98}$$

Example 1.6: Assume that the inner product between two real scalar functions u and v is defined by

$$(u, v) = \int_\Omega uv \, d\Omega, \tag{1.99}$$

where Ω is a finite region bounded by Γ in (x, y)-plane. Let $\hat{A} = -(\nabla^2 + k^2)$ and f be a real function. Find the variational expression for Helmholtz equation

$$\begin{cases} \hat{A}u = -(\nabla^2 + k^2)u = f, \\ u|_\Gamma = 0 \quad \text{or} \quad \left.\dfrac{\partial u}{\partial n}\right|_\Gamma = 0. \end{cases} \tag{1.100}$$

Solution: According to (1.98), the variational expression can be expressed as

$$I(u) = \int_\Omega -u\nabla^2 u \, d\Omega - \int_\Omega k^2 u^2 \, d\Omega - 2\int_\Omega uf \, d\Omega.$$

Use of integration by parts yields

$$I(u) = \int_\Omega \nabla u \cdot \nabla u \, d\Omega - \int_\Gamma u\frac{\partial u}{\partial n} d\Gamma - \int_\Omega k^2 u^2 \, d\Omega - 2\int_\Omega uf \, d\Omega.$$

Making use of the boundary condition we obtain

$$I(u) = \int_\Omega (\nabla u \cdot \nabla u - k^2 u^2) d\Omega - 2\int_\Omega uf \, d\Omega. \tag{1.101}$$

This is the variational expression for Helmholtz equation (1.100). Note that the smoothness requirement on the unknown function u has been relaxed. We now demonstrate that (1.101) is indeed a variational expression. Assume

that v satisfies the homogeneous boundary condition. We have

$$\begin{aligned}\frac{d}{dt}I(u+tv) &= \frac{d}{dt}\int_\Omega \nabla(u+tv)\cdot\nabla(u+tv)d\Omega \\ &\quad - \frac{d}{dt}\int_\Omega k^2(u+tv)^2 d\Omega - 2\frac{d}{dt}\int_\Omega (u+tv)f\, d\Omega \\ &= \int_\Omega (2\nabla v\cdot\nabla u + 2t\nabla v\cdot\nabla v - 2k^2uv - 2tv^2)d\Omega - 2\int_\Omega vf\, d\Omega.\end{aligned}$$

Thus

$$\begin{aligned}\left(\frac{\delta I(u)}{\delta u}, v\right) &= \frac{d}{dt}I(u+tv)\Big|_{t=0} \\ &= \int_\Omega (2\nabla v\cdot\nabla u - 2k^2uv)d\Omega - 2\int_\Omega vf\, d\Omega \\ &= \int_\Gamma 2v\frac{\partial u}{\partial n}d\Omega - \int_\Omega 2v\nabla^2 u\, d\Omega - \int_\Omega 2k^2uv\, d\Omega - 2\int_\Omega vf\, d\Omega \\ &= -\int_\Omega 2v\nabla^2 u\, d\Omega - \int_\Omega 2k^2uv\, d\Omega - 2\int_\Omega vf\, d\Omega. \qquad (1.102)\end{aligned}$$

The first equation of (1.100) follows from the Lagrangian equation $\frac{\delta I(u)}{\delta u} = 0$. □

Example 1.7: Assume that the inner product is defined by (1.99). Consider the following integral equation

$$\hat{A}u(\mathbf{r}) = u(\mathbf{r}) + \int_\Gamma G(\mathbf{r},\mathbf{r}')u(\mathbf{r}')d\Gamma(\mathbf{r}') = f(\mathbf{r}). \qquad (1.103)$$

It follows from (1.98) that

$$\begin{aligned}I(u) &= \int_\Gamma\int_\Gamma u(\mathbf{r})G(\mathbf{r},\mathbf{r}')u(\mathbf{r}')d\Gamma(\mathbf{r})d\Gamma(\mathbf{r}') \\ &\quad + \int_\Gamma u^2(\mathbf{r})d\Gamma(\mathbf{r}) - 2\int_\Gamma u(\mathbf{r})f(\mathbf{r})d\Gamma(\mathbf{r}).\end{aligned}$$

This is the variational expression for the integral equation (1.103). □

Once the variational expression is given, the **Rayleigh–Ritz method** can be used to obtain the approximate solution of the operator equation (1.91) We use the variational expression (1.98) to illustrate the procedure. Let

$$u = \sum_{j=1}^{N} a_j u_j, \tag{1.104}$$

where $\{u_j | j = 1, 2, \ldots, N\}$ is a set of known basis functions and $a_j (j = 1, 2, \ldots, N)$ are unknowns to be determined. Inserting (1.104) into (1.98) yields

$$I = \sum_{j=1}^{N} \sum_{i=1}^{N} a_j \bar{a}_i (\hat{A} u_j, u_i) - \sum_{j=1}^{N} a_j (u_j, f) - \sum_{j=1}^{N} \bar{a}_j (f, u_j). \tag{1.105}$$

This can be written in matrix form as

$$I = [\bar{u}]^T [A][u] - 2\mathrm{Re}[\bar{u}]^T [f], \tag{1.106}$$

where $[u] = [a_1, a_2, \ldots, a_N]^T$, $[f] = [(f, u_1), (f, u_2), \ldots, (f, u_N)]^T$ and

$$[A] = \begin{bmatrix} (\hat{A}u_1, u_1) & (\hat{A}u_2, u_1) & \cdots & (\hat{A}u_N, u_1) \\ (\hat{A}u_1, u_2) & (\hat{A}u_2, u_2) & \cdots & (\hat{A}u_N, u_2) \\ \vdots & \vdots & \ddots & \vdots \\ (\hat{A}u_1, u_N) & (\hat{A}u_2, u_N) & \cdots & (\hat{A}u_N, u_N) \end{bmatrix}. \tag{1.107}$$

Since $\hat{A}$ is self adjoint, the matrix $[A]$ is a Hermitian matrix

$$[A] = [\bar{A}]^T.$$

By letting $\frac{\delta I}{\delta [u]} = 0$ from (1.106), we obtain

$$[A][u] = [f]. \tag{1.108}$$

So the operator equation (1.91) has been transformed into a matrix equation after Rayleigh–Ritz procedure.

1.4.3 Numerical Techniques: A General Approach

Consider the solution of operator equation (1.91). Whenever an exact solution is not available, we have to seek numerical solution. Let $\{u_j \,|\, j = 1, 2, \ldots, N\}$ be a set of orthonormal basis functions in the domain of operator $\hat{A}$, called **basis** or **trial functions**, which satisfy

$$(u_i, u_j) = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}.$$

As an approximation, the unknown function u may be expanded as follows

$$u = \sum_{j=1}^{N} a_j u_j. \tag{1.109}$$

Let $\{v_j \,|\, j = 1, 2, \ldots, N\}$ be a another set of orthonormal basis functions in the range of $\hat{A}$, called **weighting functions** or **testing functions**. The residual left by the above approximation is given by

$$R = \hat{A}u - f = \sum_{j=1}^{N} a_j \hat{A} u_j - f,$$

and may be expanded in terms of the weighting functions

$$R = \sum_{i=1}^{N} b_i v_i,$$

where $b_i = (R, v_i)$, $i = 1, 2, \ldots, N$. The absolute value of the residual should be as small as possible. In other words, the squared residual

$$|R|^2 = \sum_{i=1}^{N} |b_i|^2$$

must reach a minimum. A derivative with respect to $|b_i|$ gives $b_i = 0$ for $i = 1, 2, \ldots, N$. Thus we have

$$\sum_{j=1}^{N} a_j (\hat{A} u_j, v_i) = (f, v_i), \quad i = 1, 2, \ldots, N.$$

This can be written in matrix form as

$$[A][u] = [f], \tag{1.110}$$

where $[u] = [a_1, a_2, \ldots, a_N]^T$, $[f] = [(f, v_1), (f, v_2), \ldots, (f, v_N)]^T$ and

$$[A] = \begin{bmatrix} (\hat{A}u_1, v_1) & (\hat{A}u_2, v_1) & \cdots & (\hat{A}u_N, v_1) \\ (\hat{A}u_1, v_2) & (\hat{A}u_2, v_2) & \cdots & (\hat{A}u_N, v_2) \\ \vdots & \vdots & \ddots & \vdots \\ (\hat{A}u_1, v_N) & (\hat{A}u_2, v_N) & \cdots & (\hat{A}u_N, v_N) \end{bmatrix}. \tag{1.111}$$

Apparently, (1.110) reduces to (1.108) if $u_i = v_i$, $i = 1, 2, \ldots, N$. The above procedure is called **projection method** or the **method of weighted residuals**. When the problem is posed in the above general manner, (1.110) may not have any solutions. If (1.110) has a unique solution for each N, the approximate solution (1.109) may not converge to the exact solution of equation (1.91) as $N \to \infty$. If we choose $u^i = \hat{B}v^i$, where $\hat{B}$ is an operator, the projection method reduces to **moment method**. Especially if we choose $\hat{B} = \hat{A}$, the projection method is equivalent to the **method of least squares**. If we choose $\hat{B} = \hat{I}$ (a unit operator), the projection method reduces to **Galerkin's method**. For most applications, the solution u of the operator equation (1.91) is defined in a region $\Omega \in R^m (m = 1, 2, 3)$. If we choose $v = \delta(\mathbf{r} - \mathbf{r}^i)$, where $\{\mathbf{r}^i \,|\, i = 1, 2, \ldots, N\}$ is a selected set of points in the defining region of the unknown function u, the projection method reduces to **collocation method**. The selected points $\mathbf{r}^i$ are called **collocation points**.

The practical implementation of numerical methods depends on how to construct the numerical basis or trial functions. We may choose a set of points $\{\mathbf{r}^i \,|\, i = 1, 2, \ldots, N\}$ with $\mathbf{r}^i = (x_1^i, \ldots, x_m^i)$, which are called **global nodes**. The node numbering system is called the **global numbering system**. Consider a set of functions $\{u_i(\mathbf{r}) \,|\, i = 1, 2, \ldots, N\}$, which satisfies

(1) For each i, there exists a positive number ε_i such that

$$u_i(\mathbf{r}) = \begin{cases} \neq 0, & |\mathbf{r} - \mathbf{r}^i| \leq \varepsilon_i \\ = 0, & |\mathbf{r} - \mathbf{r}^i| > \varepsilon_i \end{cases}$$

(2) $u_i(\mathbf{r})(i = 1, 2, \ldots, N)$ are continuous and $u_i(\mathbf{r}^j) = \delta_{ij}$.

It is easy to show that the set $\{u_i(\mathbf{r}) \,|\, i = 1, 2, \ldots, N\}$ is linearly independent. The approximate solution u can then be expressed by

$$u(\mathbf{r}) = \sum_{i=1}^{N} a_i u_i(\mathbf{r}).$$

The set $\{u_i(\mathbf{r}) \mid i = 1, 2, \ldots, N\}$ forms a global basis. Basis functions defined over the whole region Ω are difficult to obtain if the region is of a complex shape. To construct these global basis functions, we may divide the region Ω into n subregions (called **elements**) Ω_e $(e = 1, 2, \ldots, n)$ such that the intersection of any two elements is either empty or consists of a common boundary curve or points (see Figure 1.4). For each element, we choose N_e nodes $\mathbf{r}^\alpha (\alpha = 1, 2, \ldots, N_e)$ (the node numbering system α is called **local numbering system**) and introduce the **Lagrange shape functions** $l_e^\alpha (\alpha = 1, 2, \ldots, N_e)$, which are smooth and satisfy

$$\begin{aligned} l_e^\alpha(\mathbf{r}) &= 0, \quad \mathbf{r} \notin \Omega_e, \ \ \alpha = 1, 2, \ldots, N_e \\ l_e^\alpha(\mathbf{r}^\beta) &= \delta_{\alpha\beta}, \quad \alpha, \beta = 1, 2, \ldots, N_e \end{aligned} .$$

The nodes that are not on the boundaries of elements are called **internal nodes**. Otherwise, they are called **boundary nodes**. If m elements meet at $\mathbf{r}$ we say that $\mathbf{r}$ has m-multiplicity, denoted by $m(\mathbf{r})$. Let $\mathbf{r}^i$ be a node of m-multiplicity, i.e., there exist m elements $\Omega_{(e_j)} (j = 1, 2, \ldots, m)$ that meet at $\mathbf{r}^i$. Then the global basis functions can be constructed as follows

$$u_i(\mathbf{r}) = \sum_{j=1}^{m(\mathbf{r}^i)} \frac{1}{m(\mathbf{r})} l_{e_j}^{\alpha_j}(\mathbf{r}), \quad i = 1, 2, \ldots, N,$$

where α_j is the local numbering for the node $\mathbf{r}^i$.

The construction of Lagrange shape function is an interpolation process. A **line element** is shown in Figure 1.5(a). The global coordinates

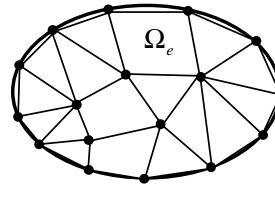


Figure 1.4 Discretization of the solution region.

Figure 1.5 (a) Linear element. (b) Standard element.

of the two end points p_1 and p_2 are denoted by x_1 and x_2 respectively. It is common practice to regard each line element as having been transformed into a standard line element in a local coordinate system λ (Figure 1.5(b)). The transformation may be assumed to be linear

$$x = a + b\lambda, \tag{1.112}$$

where the constants a and b are determined by requiring that x takes the correct values x_1 and x_2 at p_1 and p_2 respectively. Therefore we have

$$x_1 = a, \quad x_2 = a + b, \tag{1.113}$$

which gives

$$a = x_1, \quad b = x_2 - x_1. \tag{1.114}$$

Thus (1.112) can be written as

$$x = (1 - \lambda)x_1 + \lambda x_2, \tag{1.115}$$

and the Lagrange shape functions for the line element are then given by

$$\begin{cases} l_e^1(\lambda) = 1 - \lambda \\ l_e^2(\lambda) = \lambda \end{cases}, \quad 0 \le \lambda \le 1. \tag{1.116}$$

The line element with the above shape functions is called a **linear element**. The local coordinate λ is also called **natural coordinate system**, which can be expressed as

$$\lambda = \frac{|p_1 p|}{|p_1 p_2|},$$

where p is an arbitrary point in the element with coordinate x. The inverse transform of (1.115) is given by

$$\lambda = \frac{x - x_1}{x_2 - x_1}. \tag{1.117}$$

In terms of this inverse transform, an arbitrary function u defined over the linear element can be expressed as

$$u = (1 - \lambda)u(1) + \lambda u(2) = \alpha_1(x)u(1) + \alpha_2(x)u(2), \tag{1.118}$$

where $u(j)$ denotes the values of u at node $j(j = 1, 2)$ and

$$\alpha_1(x) = \frac{x_2 - x}{x_2 - x_1}, \quad \alpha_2(x) = \frac{x - x_1}{x_2 - x_1}. \tag{1.119}$$

To better represent the unknowns, we may use higher order interpolation. For a quadratic interpolation, one more node p_3 must be introduced in the middle of the standard line element (Figure 1.6). The Lagrange shape functions for the quadratic element can be easily found as follows

$$\begin{cases} l_e^1 = (\lambda - 1)(2\lambda - 1) \\ l_e^2 = \lambda(2\lambda - 1) \\ l_e^3 = 4\lambda(1 - \lambda) \end{cases}, \quad 0 \le \lambda \le 1.$$

Consider a planar triangular element $\Delta p_1 p_2 p_3$ in (x, y)-plane shown in Figure 1.7(a), which can be transformed into a standard triangle in the local coordinate system (λ_1, λ_2) as shown in Figure 1.7(b). The transformation can be assumed to be of the form

$$w = a + b\lambda_1 + c\lambda_2, \quad (w = x, y), \tag{1.120}$$

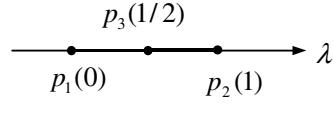


Figure 1.6 Quadratic element.

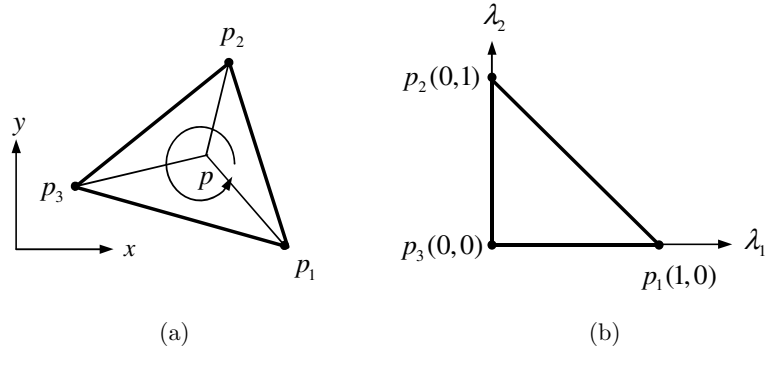


Figure 1.7 (a) Linear triangular element. (b) Standard triangle.

where the constants a, b and c are determined by requiring that $w(w = x, y)$ takes the correct values w_1, w_2 and w_3 at p_1, p_2 and p_3 respectively. Therefore, we have

$$w_1 = a + b, \quad w_2 = a + c, \quad w_3 = a, \tag{1.121}$$

which gives

$$a = w_3, \quad b = w_1 - w_3, \quad c = w_2 - w_3. \tag{1.122}$$

It follows from (1.120) and (1.122) that

$$\begin{aligned} x &= x_1\lambda_1 + x_2\lambda_2 + x_3(1 - \lambda_1 - \lambda_2), \\ y &= y_1\lambda_1 + y_2\lambda_2 + y_3(1 - \lambda_1 - \lambda_2). \end{aligned} \tag{1.123}$$

The Lagrange shape functions are thus found to be

$$l_e^i(\lambda_1, \lambda_2) = \lambda_i, \quad 0 \leq \lambda_i \leq 1, \quad i = 1, 2, 3.$$

where $\lambda_3 = 1 - \lambda_1 - \lambda_2$. Let the area of the planar triangular element $\Delta p_1 p_2 p_3$ in (x, y)-plane be denoted by Δ. The global coordinates for vertex p_i are denoted by $(x_i, y_i)(i = 1, 2, 3)$. The triangle is then divided into three small triangles using an arbitrary point p inside the triangle as a common vortex. The local coordinate system can be expressed as

$$\begin{aligned} \lambda_1 &= \Delta_1/\Delta, \\ \lambda_2 &= \Delta_2/\Delta, \\ \lambda_3 &= \Delta_3/\Delta, \end{aligned} \tag{1.124}$$

where Δ_1, Δ_2 and Δ_3 are areas of the subtriangle $\Delta p_2 p_3 p$, $\Delta p_3 p_1 p$ and $\Delta p_1 p_2 p$ respectively. Note that $\sum_{i=1}^{3} \lambda_i = 1, 0 \leq \lambda_i \leq 1$. The coordinate system (λ_1, λ_2) is called natural or area coordinate system. The inverse transform of (1.123) is given by

$$\begin{aligned} \lambda_1 &= \frac{x_2y_3 - x_3y_2 + x(y_2 - y_3) - y(x_2 - x_3)}{(x_1 - x_3)(y_2 - y_3) - (x_2 - x_3)(y_1 - y_3)}, \\ \lambda_2 &= \frac{x_3y_1 - x_1y_3 + x(y_3 - y_1) - y(x_3 - x_1)}{(x_1 - x_3)(y_2 - y_3) - (x_2 - x_3)(y_1 - y_3)}. \end{aligned} \tag{1.125}$$

In terms of the above inverse transformation, an arbitrary function u may be expressed in terms of the shape functions over the triangular element as

$$\begin{aligned} u &= \lambda_1 u(1) + \lambda_2 u(2) + (1 - \lambda_1 - \lambda_2) u(3) \\ &= \sum_{j=1}^{3} \alpha_j(x, y) u(j), \end{aligned} \tag{1.126}$$

where $u(j)$ denotes the values of u at node j $(j = 1, 2, 3)$ and

$$\begin{aligned} \alpha_1(x, y) &= \frac{x_2 y_3 - x_3 y_2 + x(y_2 - y_3) - y(x_2 - x_3)}{(x_1 - x_3)(y_2 - y_3) - (x_2 - x_3)(y_1 - y_3)} \\ \alpha_2(x, y) &= \frac{x_3 y_1 - x_1 y_3 + x(y_3 - y_1) - y(x_3 - x_1)}{(x_1 - x_3)(y_2 - y_3) - (x_2 - x_3)(y_1 - y_3)} \\ \alpha_3(x, y) &= \frac{x_1 y_2 - x_2 y_1 + x(y_2 - y_1) - y(x_2 - x_1)}{(x_1 - x_3)(y_2 - y_3) - (x_2 - x_3)(y_1 - y_3)}. \end{aligned} \tag{1.127}$$

To achieve higher accuracy, higher-order interpolation can be used. In this case, more nodes other than the vertices must be inserted to the triangle. For example, if we use quadratic interpolation, the mid-points 4, 5, and 6 of the sides of the standard right triangle may be introduced (Figure 1.8). The Lagrange shape functions are then given by

$$\begin{cases} l_e^1 = \lambda_1(2\lambda_1 - 1), & l_e^4 = 4\lambda_1\lambda_2 \\ l_e^2 = \lambda_2(2\lambda_2 - 1), & l_e^5 = 4\lambda_2\lambda_3 \\ l_e^3 = \lambda_3(2\lambda_3 - 1), & l_e^6 = 4\lambda_1\lambda_3 \end{cases}, \quad 0 \le \lambda_i \le 1.$$

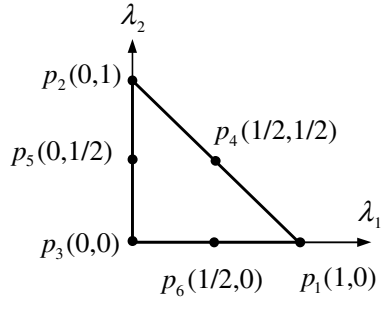


Figure 1.8 Quadratic triangular element.

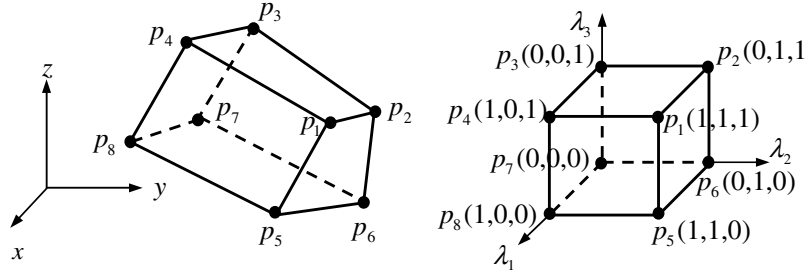


Figure 1.9 (a) Hexahedral element. (b) Standard cube.

In three dimensions, the element can be chosen as tetrahedron, triangular prism or hexahedron. The construction of shape functions can be carried out in a similar way. We use a linear hexahedral element as an example to illustrate the process (Figure 1.9). The transformation from a hexahedron to the standard cube may be assumed to be

$$\begin{aligned} w = {} & a + b\lambda_1 + c\lambda_2 + d\lambda_3 + e\lambda_1\lambda_2 + f\lambda_1\lambda_3 \\ & + g\lambda_2\lambda_3 + h\lambda_1\lambda_2\lambda_3, \quad (w = x, y, z) \end{aligned} \tag{1.128}$$

where the constants a, b, c, d, e, f, g, h are determined by requiring that w $(w = x, y, z)$ takes the correct values w_i at $p_i (i = 1, 2, \ldots, 8)$ respectively. Thus we have

$$\begin{aligned} w_1 &= a + b + c + d + e + f + g + h, & w_5 &= a + b + c + e, \\ w_2 &= a + c + d + g, & w_6 &= a + c, \\ w_3 &= a + d, & w_7 &= a, \\ w_4 &= a + b + d + f, & w_8 &= a + b, \end{aligned}$$

which yields

$$\begin{aligned} a &= w_7, & e &= w_5 - w_6 + w_7 - w_8, \\ b &= w_8 - w_7, & f &= -w_3 + w_4 - w_8 + w_7, \\ c &= w_6 - w_7, & g &= w_2 - w_3 - w_6 + w_7, \\ d &= w_3 - w_7, & h &= w_1 - w_2 + w_3 - w_4 - w_5 + w_6 - w_7 + w_8. \end{aligned}$$

Introducing the above solutions into (1.128), we obtain

$$\begin{aligned} w = {} & w_1\lambda_1\lambda_2\lambda_3 + w_2(1-\lambda_1)\lambda_2\lambda_3 + w_3(1-\lambda_1)(1-\lambda_2)\lambda_3 \\ & + w_4\lambda_1(1-\lambda_2)\lambda_3 + w_5\lambda_1\lambda_2(1-\lambda_3) + w_6(1-\lambda_1)\lambda_2(1-\lambda_3) \\ & + w_7(1-\lambda_1)(1-\lambda_2)(1-\lambda_3) + w_8\lambda_1(1-\lambda_2)(1-\lambda_3). \end{aligned}$$

The Lagrange shape functions are then given by

$$\begin{cases} l_e^1 = \lambda_1\lambda_2\lambda_3, & l_e^5 = \lambda_1\lambda_2(1-\lambda_3), \\ l_e^2 = (1-\lambda_1)\lambda_2\lambda_3, & l_e^6 = (1-\lambda_1)\lambda_2(1-\lambda_3), \\ l_e^3 = (1-\lambda_1)(1-\lambda_2)\lambda_3, & l_e^7 = (1-\lambda_1)(1-\lambda_2)(1-\lambda_3), \\ l_e^4 = \lambda_1(1-\lambda_2)\lambda_3, & l_e^8 = \lambda_1(1-\lambda_2)(1-\lambda_3) \end{cases} \quad 0 \le \lambda_i \le 1.$$

Higher-order elements involving nodes on the faces can be derived in a similar manner.

The **finite element method** (FEM) is commonly introduced as a special case of Galerkin's method for solving partial differential equations. The basic step of FEM is to divide the domain of the problem into a collection of subdomains (called elements) with each subdomain represented by a set of element equations to the original problem (called partial stiffness matrix equation in mechanics), followed by recombining all sets of element equations into a global system of equations for the final calculation (called global stiffness matrix equation in mechanics). The well-known **finite difference method** (FDM) may be regarded as a special case of projection method. Both the FEM and the FDM are called **domain method** since the governing equation has to be solved over the entire defining region of the problem. On the other hand, the integral equations are defined on the boundary of the defining region and the numerical methods used to solve them are called **boundary method**, such as **boundary element method** (BEM), which combines the element concept and collocation method to solve the boundary integral equations. The domain method can be easily applied to non-linear, inhomogeneous and time varying problems. The numerical accuracy of domain methods is generally lower than the boundary method because the discretization error is limited only on the boundary for the latter. The numerical methods commonly used in RF engineering will be discussed with applications in the forthcoming chapters.

1.5 Potential Theory

Potential theory is the mathematical treatment of the potential-energy functions used in physics to study gravitation and electromagnetism, and has developed into a major field of mathematical research (Kellogg, 1953; MacMillan, 1958). In 19th century, it was believed that all forces in nature could be derived from a potential which satisfies Laplace equation. These days, the term 'potential' is used in a broad sense, and the potential is not necessarily a solution of Laplace equation. As long as the solution of a partial differential equation can be expressed as the first derivative of a new function, this new function can be considered as a potential. Usually the solution and its potential function satisfy the same type of equation, while the equation for the latter has a simpler source terms.

1.5.1 Vector Potential, Scalar Potential, and Gauge Conditions

From the equations $\nabla \cdot \mathbf{B} = 0$ and $\nabla \times \mathbf{E} = -\partial \mathbf{B}/\partial t$, a vector potential $\mathbf{A}$ and a scalar potential ϕ can be introduced such that

$$\mathbf{E} = -\nabla\phi - \frac{\partial \mathbf{A}}{\partial t}, \quad \mathbf{B} = \nabla \times \mathbf{A}. \tag{1.129}$$

If the medium is isotropic and homogeneous, we may substitute (1.129) into $\nabla \times \mathbf{H} = \mathbf{J} + \partial \mathbf{D}/\partial t$, and insert the first of (1.129) into $\nabla \cdot \mathbf{D} = \rho$ to obtain

$$\begin{aligned} \left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right)\mathbf{A} &= -\mu\mathbf{J} + \nabla\left(\nabla \cdot \mathbf{A} + \frac{1}{v^2}\frac{\partial \phi}{\partial t}\right), \\ \left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right)\phi &= -\frac{\rho}{\varepsilon} - \frac{\partial}{\partial t}\left(\nabla \cdot \mathbf{A} + \frac{1}{v^2}\frac{\partial \phi}{\partial t}\right), \end{aligned} \tag{1.130}$$

where $v = 1/\sqrt{\mu\varepsilon}$. The term $\nabla \cdot \mathbf{A} + \partial\phi/v^2\partial t$ on the right-hand sides can be set to zero by means of the gauge transform. In fact, we may define a new vector potential $\mathbf{A}'$ and a new scalar potential ϕ' through

$$\mathbf{A}' = \mathbf{A} + \nabla\psi, \tag{1.131}$$

$$\phi' = \phi - \frac{\partial \psi}{\partial t}, \tag{1.132}$$

where ψ is called the **gauge function**. The transformation from $(\mathbf{A}, \phi)$ to $(\mathbf{A}', \phi')$ defined by (1.131) and (1.132) is called **gauge transformation**. The electromagnetic fields remain unchanged under the gauge transformation. The new vector potential $\mathbf{A}'$ and scalar potential ϕ' satisfy

$$\nabla \cdot \mathbf{A}' + \frac{1}{v^2}\frac{\partial \phi'}{\partial t} = \nabla \cdot \mathbf{A} + \frac{1}{v^2}\frac{\partial \phi}{\partial t} + \nabla^2 \psi - \frac{1}{v^2}\frac{\partial^2 \psi}{\partial t^2}.$$

If the term $\nabla \cdot \mathbf{A} + \partial\phi/v^2\partial t$ is not zero, the left-hand side can be sent to zero by forcing the gauge function ψ to satisfy

$$\nabla^2 \psi - \frac{1}{v^2}\frac{\partial^2 \psi}{\partial t^2} = -\left(\nabla \cdot \mathbf{A} + \frac{1}{v^2}\frac{\partial \phi}{\partial t}\right).$$

Thus the equation

$$\nabla \cdot \mathbf{A} + \frac{1}{v^2}\frac{\partial \phi}{\partial t} = 0 \tag{1.133}$$

may be assumed and is called **Lorenz gauge condition**, named after the Danish physicist Ludvig Valentin Lorenz (1829–1891). If $\mathbf{A}$ and ϕ satisfy the Lorenz gauge condition, Equations (1.130) reduce to

$$\left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right)\mathbf{A} = -\mu\mathbf{J}, \quad \left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right)\phi = -\frac{\rho}{\varepsilon}, \tag{1.134}$$

and they become uncoupled. The retarded solutions of (1.134) are given by

$$\mathbf{A}(\mathbf{r}, t) = \int_{V_0} \frac{\mu \mathbf{J}(\mathbf{r}', t - R/v)}{4\pi R} dV(\mathbf{r}'), \quad \phi(\mathbf{r}, t) = \int_{V_0} \frac{\rho(\mathbf{r}', t - R/v)}{4\pi\varepsilon R} dV(\mathbf{r}'),$$

where V_0 denotes the source region. Note that the Lorenz gauge condition implies the continuity equation of the current.

Remark 1.2 (Magnetic sources): When the sources are purely magnetic, Maxwell equations (1.4) become

$$\begin{aligned}
\nabla \times \mathbf{H}(\mathbf{r}, t) &= \frac{\partial \mathbf{D}(\mathbf{r}, t)}{\partial t}, \\
\nabla \times \mathbf{E}(\mathbf{r}, t) &= -\frac{\partial \mathbf{B}(\mathbf{r}, t)}{\partial t} - \mathbf{J}_m(\mathbf{r}, t), \\
\nabla \cdot \mathbf{D}(\mathbf{r}, t) &= 0, \\
\nabla \cdot \mathbf{B}(\mathbf{r}, t) &= \rho_m(\mathbf{r}, t).
\end{aligned}$$

In a similar way, we may introduce a vector potential $\mathbf{A}_m$ and a scalar potential ϕ_m so that

$$\mathbf{H} = -\nabla\phi_m - \frac{\partial \mathbf{A}_m}{\partial t}, \quad \mathbf{D} = \nabla \times \mathbf{A}_m.$$

Assuming that the potentials satisfy the Lorenz gauge condition

$$\nabla \cdot \mathbf{A}_m + \frac{1}{v^2}\frac{\partial \phi_m}{\partial t} = 0,$$

then

$$\left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right)\mathbf{A}_m = -\varepsilon \mathbf{J}_m, \quad \left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right)\phi_m = -\frac{\rho_m}{\mu}.$$

The retarded solutions of the above equations are

$$\mathbf{A}_m(\mathbf{r}, t) = \int_{V_0} \frac{\varepsilon \mathbf{J}_m(\mathbf{r}', t - R/v)}{4\pi R} dV(\mathbf{r}'),$$

$$\phi_m(\mathbf{r}, t) = \int_{V_0} \frac{\rho_m(\mathbf{r}', t - R/v)}{4\pi\mu R} dV(\mathbf{r}'),$$

where V_0 denotes the source region. □

1.5.2 Hertz Vectors and Debye Potentials

In addition to vector potential $\mathbf{A}$ and scalar potential ϕ, other potential functions can be introduced to simplify the problems. The current source $\mathbf{J}$ can be divided into the sum of two components

$$\mathbf{J}(\mathbf{r}, t) = \int_{V_0} \mathbf{J}(\mathbf{r}', t)\delta(\mathbf{r} - \mathbf{r}')dV(\mathbf{r}') = -\nabla^2 \int_{V_0} \frac{\mathbf{J}(\mathbf{r}', t)}{4\pi R} dV(\mathbf{r}') = \mathbf{J}^{\parallel} + \mathbf{J}^{\perp} \tag{1.135}$$

where we have used $\nabla^2(1/4\pi R) = -\delta(\mathbf{r} - \mathbf{r}')$, and

$$\mathbf{J}^{\parallel}(\mathbf{r}, t) = -\nabla\nabla \cdot \int_{V_0} \frac{\mathbf{J}(\mathbf{r}', t)}{4\pi R} dV(\mathbf{r}'), \mathbf{J}^{\perp}(\mathbf{r}, t) = \nabla \times \nabla \times \int_{V_0} \frac{\mathbf{J}(\mathbf{r}', t)}{4\pi R} dV(\mathbf{r}') \tag{1.136}$$

are referred to as the **irrotational component** and **solenoidal component** of $\mathbf{J}$ respectively. If the current source $\mathbf{J}$ is irrotational, it only has a

longitudinal component and can be written as

$$\begin{aligned}\mathbf{J}^{\parallel}(\mathbf{r},t) &= -\nabla \int_{V_0} \nabla' \cdot \left[\frac{\mathbf{J}(\mathbf{r}',t)}{4\pi R}\right] dV(\mathbf{r}') - \nabla \int_{V_0} \frac{\nabla' \cdot \mathbf{J}(\mathbf{r}',t)}{4\pi R} dV(\mathbf{r}') \\ &= \nabla \int_{\partial V_0} \frac{\mathbf{J}(\mathbf{r}',t)}{4\pi R} \cdot \mathbf{u}_n(\mathbf{r}') dV(\mathbf{r}') - \nabla \int_{V_0} \frac{\nabla' \cdot \mathbf{J}(\mathbf{r}',t)}{4\pi R} dV(\mathbf{r}') \\ &= -\nabla \int_{V_0} \frac{\nabla' \cdot \mathbf{J}(\mathbf{r}',t)}{4\pi R} dV(\mathbf{r}') = \nabla \frac{\partial}{\partial t} \int_{V_0} \frac{\rho(\mathbf{r}',t)}{4\pi R} dV(\mathbf{r}') = \frac{\partial \mathbf{P}}{\partial t}.\end{aligned} \tag{1.137}$$

Here

$$\mathbf{P}(\mathbf{r},t) = \nabla \int_{V_0} \frac{\rho(\mathbf{r}',t)}{4\pi R} dV(\mathbf{r}')$$

is the equivalent polarization vector. From the continuity equation, the corresponding polarization charge density is given by $\rho = -\nabla \cdot \mathbf{P}$. Substituting (1.137) into the first equation of (1.134), we have

$$\left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right)\mathbf{A} = -\mu \frac{\partial \mathbf{P}}{\partial t}.$$

To get rid of the differential operation on the source term, we may introduce a new potential function $\mathbf{\Pi}_e$ such that

$$\mathbf{A} = \frac{1}{v^2}\frac{\partial \mathbf{\Pi}_e}{\partial t}. \tag{1.138}$$

The new potential function $\mathbf{\Pi}_e$ is called **electric Hertz vector** and satisfies

$$\left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right)\mathbf{\Pi}_e = -\frac{\mathbf{P}}{\varepsilon}. \tag{1.139}$$

From (1.133) and (1.138), we obtain

$$\phi = -\nabla \cdot \mathbf{\Pi}_e. \tag{1.140}$$

In terms of the electric Hertz vector, the electromagnetic fields may be represented by

$$\mathbf{B} = \frac{1}{v^2}\nabla \times \frac{\partial \mathbf{\Pi}_e}{\partial t}, \quad \mathbf{E} = \nabla(\nabla \cdot \mathbf{\Pi}_e) - \frac{1}{v^2}\frac{\partial^2 \mathbf{\Pi}_e}{\partial t^2}. \tag{1.141}$$

If the current source $\mathbf{J}$ is solenoidal, it only has a transverse component and may be written as

$$\mathbf{J} = \nabla \times \mathbf{M} \tag{1.142}$$

where $\mathbf{M}$ is the equivalent magnetization vector

$$\mathbf{M}(\mathbf{r}, t) = \nabla \times \int_{V_0} \frac{\mathbf{J}(\mathbf{r}', t)}{4\pi R} dV(\mathbf{r}')$$

by (1.136). Introducing (1.142) into the first equation of (1.134) gives

$$\left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right)\mathbf{A} = -\mu\nabla \times \mathbf{M}.$$

To get rid of the differential operation on the source term, we can introduce a new potential function $\mathbf{\Pi}_m$, called **magnetic Hertz vector** such that $\mathbf{A} = -\mu\nabla \times \mathbf{\Pi}_m$. The magnetic Hertz vector satisfies

$$\left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right)\mathbf{\Pi}_m = -\mathbf{M}. \tag{1.143}$$

Since $\nabla \cdot \mathbf{A} = 0$ implies $\phi = 0$, the electromagnetic fields can be expressed as

$$\mathbf{B} = \mu\nabla \times \nabla \times \mathbf{\Pi}_m, \quad \mathbf{E} = -\mu\nabla \times \frac{\partial \mathbf{\Pi}_m}{\partial t}. \tag{1.144}$$

In general, the current source $\mathbf{J}$ is of the form $\mathbf{J} = \partial\mathbf{P}/\partial t + \nabla \times \mathbf{M}$ from (1.136). For a linear medium, the superposition theorem applies and the electromagnetic fields for a general current source can be expressed as the sum of (1.141) and (1.144):

$$\mathbf{E} = \nabla(\nabla \cdot \mathbf{\Pi}_e) - \frac{1}{v^2}\frac{\partial^2 \mathbf{\Pi}_e}{\partial t^2} - \mu\nabla \times \frac{\partial \mathbf{\Pi}_m}{\partial t},$$
$$\mathbf{H} = \varepsilon\nabla \times \frac{\partial \mathbf{\Pi}_e}{\partial t} + \nabla \times \nabla \times \mathbf{\Pi}_m.$$

In a source-free region, these equations may be written as

$$\mathbf{E} = \nabla \times \nabla \times \mathbf{\Pi}_e - \mu\nabla \times \frac{\partial \mathbf{\Pi}_m}{\partial t}, \quad \mathbf{H} = \varepsilon\nabla \times \frac{\partial \mathbf{\Pi}_e}{\partial t} + \nabla \times \nabla \times \mathbf{\Pi}_m,$$

by use of (1.139) and (1.143). Note that

$$\begin{aligned}\mathbf{\Pi}_e(\mathbf{r},t) &= \int_{V_0} \frac{\mathbf{P}(\mathbf{r}', t - R/v)}{4\pi\varepsilon R} dV(\mathbf{r}'),\\ \mathbf{\Pi}_m(\mathbf{r},t) &= \int_{V_0} \frac{\mathbf{M}(\mathbf{r}', t - R/v)}{4\pi R} dV(\mathbf{r}').\end{aligned} \tag{1.145}$$

It can be shown that the electromagnetic fields in source-free region can be represented by two scalar potential functions (Jones, 1964). Hence, we may use the spherical coordinate system (r, θ, φ) and choose $\mathbf{\Pi}_e = \mathbf{r}u_e$ and $\mathbf{\Pi}_m = \mathbf{r}u_m$ to represent the electromagnetic fields. Here u_e and u_m satisfy the homogeneous wave equations:

$$\left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right) u_e = 0, \quad \left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right) u_m = 0,$$

and they are called **Debye potentials**, named after the Dutch physicist Peter Joseph William Debye (1884–1966). Let $\mathbf{u}_r$, $\mathbf{u}_\theta$ and $\mathbf{u}_\varphi$ denote the unit vectors in the direction of increasing r, θ and φ respectively. A simple calculation gives

$$\begin{aligned}\nabla \times \nabla \times (\mathbf{r}u_e) &= \left(-\frac{1}{r}\nabla^2_{\theta\varphi} u_e\right)\mathbf{u}_r + \nabla_{\theta\varphi}\left[\frac{1}{r}\frac{\partial(ru_e)}{\partial r}\right],\\ \nabla \times (\mathbf{r}u_m) &= \nabla_{\theta\varphi} u_m \times \mathbf{u}_r,\end{aligned}$$

where

$$\begin{aligned}\nabla_{\theta\varphi} &= \mathbf{u}_\theta \frac{\partial}{\partial\theta} + \mathbf{u}_\varphi \frac{1}{\sin\theta}\frac{\partial}{\partial\varphi},\\ \nabla^2_{\theta\varphi} &= \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}.\end{aligned}$$

Thus the electromagnetic fields in source-free region can be expressed as

$$\begin{aligned}\mathbf{E} &= -\left(\frac{1}{r}\nabla^2_{\theta\varphi} u_e\right)\mathbf{u}_r + \nabla_{\theta\varphi}\left[\frac{1}{r}\frac{\partial(ru_e)}{\partial r}\right] + \mu\mathbf{u}_r \times \frac{\partial}{\partial t}\nabla_{\theta\varphi} u_m,\\ \mathbf{H} &= -\left(\frac{1}{r}\nabla^2_{\theta\varphi} u_m\right)\mathbf{u}_r + \nabla_{\theta\varphi}\left[\frac{1}{r}\frac{\partial(ru_m)}{\partial r}\right] - \varepsilon\mathbf{u}_r \times \frac{\partial}{\partial t}\nabla_{\theta\varphi} u_e.\end{aligned} \tag{1.146}$$

1.5.3 Jump Relations

In electromagnetic theory, we are often encountered with the following potential integrals

$$\mathbf{A}(\mathbf{r}) = \int_S \mathbf{a}(\mathbf{r}')G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}'), \quad \varphi(\mathbf{r}) = \int_S f(\mathbf{r}')G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}'),$$

where $G(\mathbf{r},\mathbf{r}') = e^{-jkR}/4\pi R$. When the field point $\mathbf{r}$ is on the surface S, these potential integrals are defined as improper but convergent integrals as follows

$$\mathbf{A}(\mathbf{r}) = \lim_{\delta\to 0} \int_{S-S_\delta} \mathbf{a}(\mathbf{r}')G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}'),$$

$$\varphi(\mathbf{r}) = \lim_{\delta\to 0} \int_{S-S_\delta} f(\mathbf{r}')G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}'),$$

where $\mathbf{r} \in S$, S_δ is a small area of arbitrary shape containing $\mathbf{r}$ and δ is the maximum chord of S_δ. For $\mathbf{r} \in S$, the following **jump relations** can be established

$$\begin{aligned}
\nabla\cdot\mathbf{A}_\pm(\mathbf{r}) &= \int_S \nabla G(\mathbf{r},\mathbf{r}')\cdot\mathbf{a}(\mathbf{r}')dS(\mathbf{r}') \mp \frac{1}{2}\mathbf{u}_n(\mathbf{r})\cdot\mathbf{a}(\mathbf{r}),\\
\nabla\times\mathbf{A}_\pm(\mathbf{r}) &= \int_S \nabla G(\mathbf{r},\mathbf{r}')\times\mathbf{a}(\mathbf{r}')dS(\mathbf{r}') \mp \frac{1}{2}\mathbf{u}_n(\mathbf{r})\times\mathbf{a}(\mathbf{r}),\\
\nabla\varphi_\pm(\mathbf{r}) &= \int_S f(\mathbf{r}')\nabla G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}') \mp \frac{1}{2}\mathbf{u}_n(\mathbf{r})f(\mathbf{r}),
\end{aligned} \tag{1.147}$$

where $\mathbf{u}_n(\mathbf{r})$ is the unit outward normal of S at $\mathbf{r}$ and

$$\nabla\cdot\mathbf{A}_\pm(\mathbf{r}) \equiv \lim_{h\to+0} \nabla\cdot\mathbf{A}[\mathbf{r}\pm h\mathbf{u}_n(\mathbf{r})],$$

$$\nabla\times\mathbf{A}_\pm(\mathbf{r}) \equiv \lim_{h\to+0} \nabla\times\mathbf{A}[\mathbf{r}\pm h\mathbf{u}_n(\mathbf{r})],$$

$$\nabla\varphi_\pm(\mathbf{r}) \equiv \lim_{h\to+0} \nabla\varphi[\mathbf{r}\pm h\mathbf{u}_n(\mathbf{r})].$$

The subscripts $+$ and $-$ respectively indicate the limit values as $\mathbf{r}$ approaches S from the exterior and interior of S. All the integrals in (1.147)

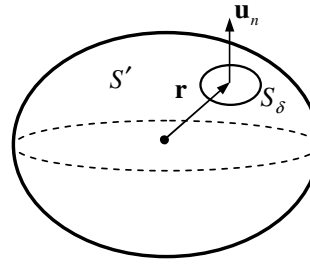


Figure 1.10 Cauchy principal value.

stand for the Cauchy principal values. Moreover, we have

$$\lim_{h\to+0} \mathbf{u}_n(\mathbf{r}) \times \{\nabla\times\nabla\times\mathbf{A}[\mathbf{r}+h\mathbf{u}_n(\mathbf{r})] - \nabla\times\nabla\times\mathbf{A}[\mathbf{r}-h\mathbf{u}_n(\mathbf{r})]\} = 0, \qquad \mathbf{r}\in S.$$

We only show the derivation of the last relation in (1.147). Let the closed surface S be split into two parts, S' and S_δ, of which S_δ is a small region surrounding $\mathbf{r}$, and S' the remainder of S. If S is smooth around $\mathbf{r}$, S_δ may be considered as a circular disk of radius δ centered at $\mathbf{r}$, as shown in Figure 1.10. Thus

$$\begin{aligned}\nabla\varphi_\pm(\mathbf{r}) &= \lim_{h\to0}\nabla\int_S f(\mathbf{r}')G(\mathbf{r}\pm h\mathbf{u}_n(\mathbf{r}),\mathbf{r}')dS(\mathbf{r}')\\ &= \lim_{h\to0}\nabla\int_{S'} f(\mathbf{r}')G(\mathbf{r}\pm h\mathbf{u}_n(\mathbf{r}),\mathbf{r}')dS(\mathbf{r}')\\ &\quad+\lim_{h\to0}\nabla\int_{S_\delta} f(\mathbf{r}')G(\mathbf{r}\pm h\mathbf{u}_n(\mathbf{r}),\mathbf{r}')dS(\mathbf{r}').\end{aligned}$$

The first integral on the right-hand side approaches to the principal value as $\delta\to0$. The integral over S_δ can be calculated through the approximation

$$\begin{aligned}&\lim_{h\to0}\nabla\int_{S_\delta} f(\mathbf{r}')G(\mathbf{r}\pm h\mathbf{u}_n(\mathbf{r}),\mathbf{r}')dS(\mathbf{r}')\\ &\quad= f(\mathbf{r})\lim_{h\to0}\frac{1}{4\pi}\left(\nabla_t \pm \mathbf{u}_n\frac{\partial}{\partial h}\right)\int_{S_\delta}\frac{dS(\mathbf{r}')}{|\mathbf{r}\pm h\mathbf{u}_n(\mathbf{r})-\mathbf{r}'|}\end{aligned}$$

$$\approx f(\mathbf{r}) \lim_{h\to 0} \frac{1}{4\pi} \left(\nabla_t \pm \mathbf{u}_n \frac{\partial}{\partial h}\right) \int_0^{2\pi} \int_0^{\delta} \frac{\rho d\rho d\phi}{\sqrt{\rho^2 + h^2}}$$

$$= \frac{1}{2} f(\mathbf{r}) \lim_{h\to 0} \left(\nabla_t \pm \mathbf{u}_n \frac{\partial}{\partial h}\right) \left(\delta - h + \frac{h^2}{2\delta}\right) = \mp \frac{1}{2} \mathbf{u}_n f(\mathbf{r}),$$

which gives the last expression of (1.147).

The function

$$\varphi(\mathbf{r}) = \int_S f(\mathbf{r}') G(\mathbf{r}, \mathbf{r}') dS(\mathbf{r}'), \quad \mathbf{r} \in R^3 - S$$

is called a **single-layer potential** with density f and the function

$$\psi(\mathbf{r}) = \int_S f(\mathbf{r}') \frac{\partial G(\mathbf{r}, \mathbf{r}')}{\partial n(\mathbf{r}')} dS(\mathbf{r}'), \quad \mathbf{r} \in R^3 - S$$

is called a **double-layer potential** with density f.

1.5.4 Multipole Expansion

Let us consider the following integral

$$\mathbf{F} = \int_V \psi(\mathbf{r}) \mathbf{J}(\mathbf{r}) dV(\mathbf{r}), \tag{1.148}$$

where $\mathbf{J}$ is a current source distribution, ψ is a slowly varying function over the region V bounded by a closed surface S. The function ψ can be expanded as a Taylor series about an origin inside V

$$\psi(\mathbf{r}) \approx \psi(0) + (\mathbf{r} \cdot \nabla)\psi(0). \tag{1.149}$$

Introducing (1.149) into (1.148), we obtain

$$\mathbf{F} = \psi(0) \int_V \mathbf{J}(\mathbf{r}) dV(\mathbf{r}) + \int_V (\mathbf{r} \cdot \nabla)\psi(0) \mathbf{J}(\mathbf{r}) dV(\mathbf{r}). \tag{1.150}$$

The first term in (1.150) can be written as

$$\psi(0) \int_V \mathbf{J}(\mathbf{r}) dV(\mathbf{r}) = \psi(0) j\omega \mathbf{p}. \tag{1.151}$$

Here $\mathbf{p}$ is the **electric dipole moment** of the source $\mathbf{J}$ defined by

$$\mathbf{p} = \frac{1}{j\omega}\int_V \mathbf{J}(\mathbf{r})dV(\mathbf{r}) = \int_V \mathbf{r}\rho(\mathbf{r})dV(\mathbf{r}) + \int_S \mathbf{r}\rho_s(\mathbf{r})dS(\mathbf{r}), \quad (1.152)$$

where ρ and ρ_s are the volume charge and surface charge density respectively

$$\rho = -\frac{1}{j\omega}\nabla\cdot\mathbf{J}, \quad \rho_s = \frac{1}{j\omega}\mathbf{u}_n\cdot\mathbf{J}.$$

The second term in (1.150) can be written as

$$\begin{aligned}\int_V (\mathbf{r}\cdot\nabla)\psi(0)\mathbf{J}(\mathbf{r})dV(\mathbf{r}) &= \nabla\psi(0)\cdot\int_V \mathbf{r}\mathbf{J}(\mathbf{r})dV(\mathbf{r})\\ &= \frac{1}{2}\nabla\psi(0)\cdot\int_V [\mathbf{r}\mathbf{J}(\mathbf{r}) + \mathbf{J}(\mathbf{r})\mathbf{r}]dV(\mathbf{r})\\ &\quad + \frac{1}{2}\nabla\psi(0)\cdot\int_V [\mathbf{r}\mathbf{J}(\mathbf{r}) - \mathbf{J}(\mathbf{r})\mathbf{r}]dV(\mathbf{r}).\end{aligned} \quad (1.153)$$

Introducing the **magnetic dipole moment** $\mathbf{m}$ and the **dyadic electric quadrupole** $\overleftrightarrow{\mathbf{Q}}^e$

$$\begin{aligned}\mathbf{m} &= \frac{1}{2}\int_V \mathbf{r}\times\mathbf{J}(\mathbf{r})dV(\mathbf{r}),\\ \overleftrightarrow{\mathbf{Q}}^e &= \int_V \mathbf{r}\mathbf{r}\rho(\mathbf{r})dV(\mathbf{r}) + \int_S \mathbf{r}\mathbf{r}\rho_s(\mathbf{r})dS(\mathbf{r})\\ &= \frac{1}{j\omega}\int_V [\mathbf{r}\mathbf{J}(\mathbf{r}) + \mathbf{J}(\mathbf{r})\mathbf{r}]dV(\mathbf{r}),\end{aligned}$$

(1.153) may be rewritten as

$$\int_V (\mathbf{r}\cdot\nabla)\psi(0)\mathbf{J}(\mathbf{r})dV(\mathbf{r}) = -\nabla\psi(0)\times\mathbf{m} + \frac{1}{2}j\omega\nabla\psi(0)\cdot\overleftrightarrow{\mathbf{Q}}^e. \quad (1.154)$$

Substituting (1.151) and (1.154) into (1.150), we have

$$\mathbf{F} = j\omega\psi(0)\mathbf{p} - \nabla\psi(0)\times\mathbf{m} + \frac{1}{2}j\omega\nabla\psi(0)\cdot\overleftrightarrow{\mathbf{Q}}^e. \quad (1.155)$$

This relationship is useful in studying low-frequency radiation problems.

For a magnetic current $\mathbf{J}_m$, we may introduce the **electric dipole moment p**, the **magnetic dipole moment m** and the **dyadic magnetic quadrupole** $\overleftrightarrow{\mathbf{Q}}^m$, defined as follows

$$\begin{aligned}
\mathbf{p} &= \frac{1}{2}\int_V \mathbf{J}_m(\mathbf{r}) \times \mathbf{r} dV(\mathbf{r}), \\
\mathbf{m} &= \frac{1}{j\omega}\int_V \mathbf{J}_m(\mathbf{r}) dV(\mathbf{r}) = \int_V \mathbf{r}\rho_m(\mathbf{r}) dV(\mathbf{r}) + \int_S \mathbf{r}\rho_{ms}(\mathbf{r}) dS(\mathbf{r}), \\
\overleftrightarrow{\mathbf{Q}}^m &= \int_V \mathbf{r}\mathbf{r}\rho_m(\mathbf{r}) dV(\mathbf{r}) + \int_S \mathbf{r}\mathbf{r}\rho_{ms}(\mathbf{r}) dS(\mathbf{r}) \\
&= \frac{1}{j\omega}\int_V [\mathbf{r}\mathbf{J}_m(\mathbf{r}) + \mathbf{J}_m(\mathbf{r})\mathbf{r}] dV(\mathbf{r}),
\end{aligned} \tag{1.156}$$

where

$$\rho_m = -\frac{1}{j\omega}\nabla \cdot \mathbf{J}_m, \quad \rho_{ms} = \frac{1}{j\omega}\mathbf{u}_n \cdot \mathbf{J}_m.$$

> We do not really deal with mathematical physics, but with physical mathematics; not with the mathematical formulation of physical facts, but with the physical motivation of mathematical methods.
>
> —Arnold Sommerfeld (German physicist, 1868–1951)

This page intentionally left blank

Chapter 2

Waveguides

> We have seen that electromagnetic fields may be held together and guided by the surface of a non-conducting rod and that they protect themselves against outward radiation by a skin effect. The protection will be complete if we embed the non-conductor in a metallic tube, whereupon the condition of a sufficiently high dielectric constant may be omitted and the dielectric within the tube may also be air. We thus arrive at the configuration of the wave guides, which have become important in high frequency practice.
>
> —Arnold Sommerfeld (German physicist, 1868–1951)

A waveguide is a structure that guides high frequency electromagnetic signals. In order to decrease the distortion while the signals propagate in the waveguide, the frequency of the transmitted signals and the waveguide dimensions must be properly chosen. Mathematically, the theory of waveguide reduces to eigenvalue problems of partial differential equations. The eigenvalues are the cut-off wavenumbers (i.e., the cut-off frequency) of the corresponding eigenvectors or guided modes. When a signal propagates in a waveguide, the frequency components below the cut-off frequency of the dominant mode are quickly attenuated and the higher order modes are excited by the frequency components above the cut-off frequency of the higher order modes, which may cause signal distortion and interference. For this reason, the normal operation of a waveguide is limited to the frequency range between the cut-off frequency of the dominant mode and that of the next higher mode.

The waveguide theory is the cornerstone of electromagnetic engineering. The early history of the waveguides has been summarized by Packard and Oliner (Packard, 1984; Oliner, 1984). The essential basis of modern waveguide theory was developed by Oliver Heaviside in the late 19th century,

who considered various possibilities for waves along wire lines and found that a single conductor line was not feasible, and a guided wave needs two wires. Heaviside also introduced the term 'impedance', which is defined as the ratio of voltage to current in a circuit. The concept of the impedance was then extended to fields and waves by Schelkunoff in 1938 in a systematic way. The impedance is regarded as the characteristic of the field as well as the medium, and has a direction. In 1897, Rayleigh showed that waves could propagate within a hollow conducting cylinder and found that such waves existed only in a set of well-defined normal modes, and to support the modes in the hollow cylinder, the operating frequency must exceed the cut-off frequencies of the corresponding modes. The theory of dielectric waveguide was first studied by Sommerfeld in 1899 and then extended by the Greek physicist Demetrius Hondros (1882–1962) in 1909. The guided wave in a single dielectric rod is based on the fact that the discontinuity surface between two different media is likely to bind the wave to that surface, thus guiding the wave. The possible use of hollow waveguides was investigated during 1930s by the American radio engineers George Clark Southworth (1890–1972) and Wilmer Lanier Barrow (1903–1975). Most important results on waveguide theory obtained in the first half of last century have been included in the '*Waveguide Handbook*' (Marcuvitz, 1951). Since then, many new guiding structures have been proposed, such as microstrip line (Grieg and Engelmann, 1952), finline (Meier, 1972), etc., and they are widely used in microwave integrated circuits. Figure 2.1 shows some waveguide structures commonly used in microwave engineering.

In practice, the waveguides are used as signal paths among different components. When the propagating mode hits an obstacle or a discontinuity, such as a conducting post across the guide, a number of higher order modes will be excited so that the fields, which are expressed as a linear superimposition of higher order modes, can satisfy the boundary conditions. These higher order modes are not propagating and die out rapidly away from the discontinuity, and they exist as stored energies. Various discontinuities in a waveguide may be regarded as a multi-port network characterized by network parameters or represented by a lumped-element circuit.

2.1 Modal Theory for Metal Waveguides

The modal theory of waveguide is the foundation of microwave engineering, and is best understood by variational analysis.

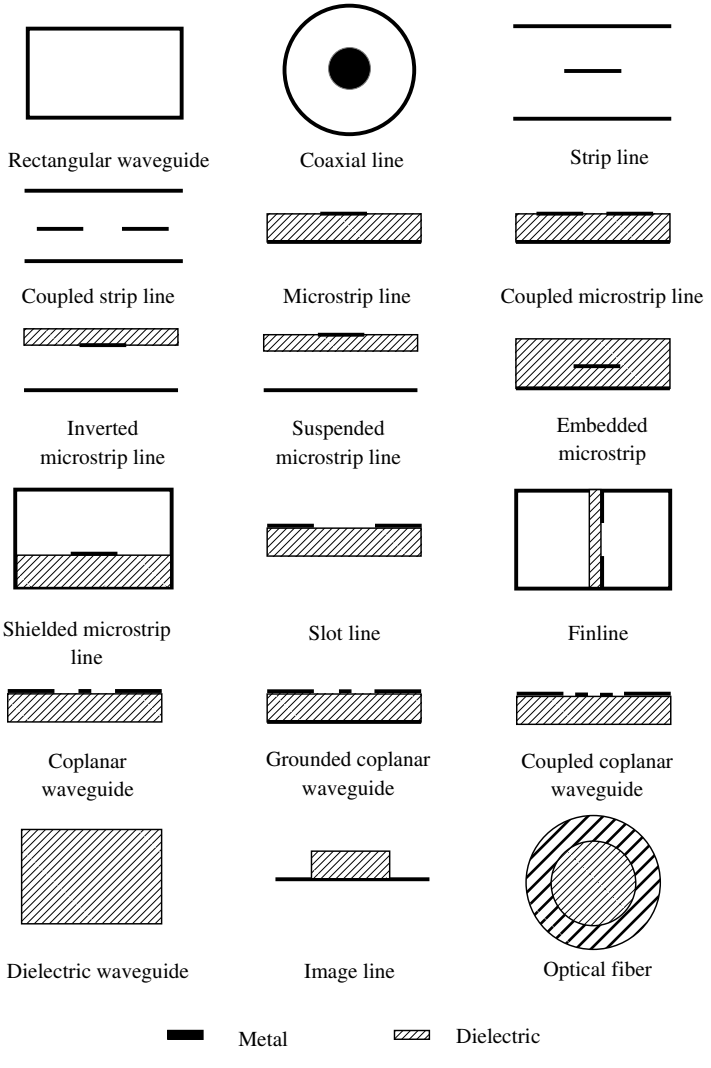


Figure 2.1 Waveguides.

2.1.1 Eigenvalue Equation

An arbitrary metal waveguide is shown in Figure 2.2. The waveguide is assumed to be uniform along z direction and is filled with homogeneous medium with medium parameters μ, ε and σ. The cross section of the waveguide is denoted by Ω, its boundary by Γ. For a waveguide with

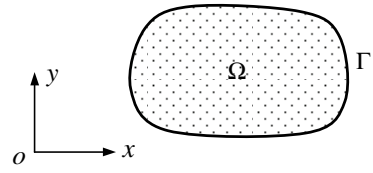


Figure 2.2 An arbitrary waveguide.

perfectly conducting walls, the electric field in the waveguide satisfies the wave equation

$$\begin{aligned} &\nabla^2\mathbf{E}(\mathbf{r},t) - \frac{1}{v^2}\frac{\partial^2\mathbf{E}(\mathbf{r},t)}{\partial t^2} - \sigma\frac{\eta}{v}\frac{\partial\mathbf{E}(\mathbf{r},t)}{\partial t} = 0, \quad \mathbf{r}\in\Omega, \\ &\nabla\cdot\mathbf{E}(\mathbf{r},t) = 0, \quad \mathbf{r}\in\Omega, \\ &\mathbf{u}_n\times\mathbf{E}(\mathbf{r},t) = 0, \quad \mathbf{r}\in\Gamma, \end{aligned} \tag{2.1}$$

where $v = 1/\sqrt{\mu\varepsilon}$, $\eta = \sqrt{\mu/\varepsilon}$, and $\mathbf{u}_n$ is the unit outward normal to the boundary Γ. The electric field can be decomposed into the sum of a transverse component and a longitudinal component. Both components may be assumed to be the separable functions of the transverse coordinates $\boldsymbol{\rho} = (x, y)$ and longitudinal coordinate z with time

$$\mathbf{E}(\mathbf{r},t) = [\mathbf{e}(\boldsymbol{\rho}) + \mathbf{u}_z e_z(\boldsymbol{\rho})]u(z,t) \tag{2.2}$$

where $\mathbf{u}_z$ is the unit vector along $+z$ direction. Substituting (2.2) into (2.1) and considering the boundary conditions, we may obtain

$$\begin{aligned} &\nabla\times\nabla\times\mathbf{e} - \nabla(\nabla\cdot\mathbf{e}) = k_c^2\mathbf{e}, \quad \boldsymbol{\rho}\in\Omega, \\ &\mathbf{u}_n\times\mathbf{e} = 0, \quad \nabla\cdot\mathbf{e} = 0, \quad \boldsymbol{\rho}\in\Gamma, \end{aligned} \tag{2.3}$$

where k_c^2 is the separation constant. The function $u(z,t)$ satisfies the modified Klein–Gordon equation,

$$\frac{\partial^2 u(z,t)}{\partial z^2} - \frac{1}{v^2}\frac{\partial^2 u(z,t)}{\partial t^2} - \sigma\frac{\eta}{v}\frac{\partial u(z,t)}{\partial t} - k_c^2 u(z,t) = 0,$$

named after the Swedish physicist Oskar Benjamin Klein (1894–1977) and German physicist Walter Gordon (1893–1939), who proposed the equation in 1927. For $\sigma = 0$, the above equation reduces to Klein–Gordon equation.

Remark 2.1: For time-harmonic fields, we may let

$$\begin{aligned}\mathbf{E}(\mathbf{r}) &= [\mathbf{e}(\boldsymbol{\rho}) + \mathbf{u}_z e_z(\boldsymbol{\rho})]e^{-j\beta z},\\ \mathbf{H}(\mathbf{r}) &= [\mathbf{h}(\boldsymbol{\rho}) + \mathbf{u}_z h_z(\boldsymbol{\rho})]e^{-j\beta z}.\end{aligned} \tag{2.4}$$

Substituting these into Maxwell equation in a source-free region

$$\begin{aligned}\nabla \times \mathbf{H}(\mathbf{r}) &= j\omega\varepsilon_e \mathbf{E}(\mathbf{r}),\\ \nabla \times \mathbf{E}(\mathbf{r}) &= -j\omega\mu \mathbf{H}(\mathbf{r}),\\ \nabla \cdot \mathbf{E}(\mathbf{r}) &= 0,\\ \nabla \cdot \mathbf{H}(\mathbf{r}) &= 0,\end{aligned}$$

where $\varepsilon_e = \varepsilon(1 - j\frac{\sigma}{\omega\varepsilon})$, we obtain

$$\begin{aligned}&\nabla \times \mathbf{h} = j\omega\varepsilon_e \mathbf{u}_z e_z,\\ &j\beta \mathbf{u}_z \times \mathbf{h} + \mathbf{u}_z \times \nabla h_z = -j\omega\varepsilon_e \mathbf{e},\\ &\nabla \times \mathbf{e} = -j\omega\mu \mathbf{u}_z h_z,\\ &j\beta \mathbf{u}_z \times \mathbf{e} + \mathbf{u}_z \times \nabla e_z = j\omega\mu \mathbf{h},\\ &\nabla \cdot \mathbf{e} = j\beta e_z,\\ &\nabla \cdot \mathbf{h} = j\beta h_z.\end{aligned} \tag{2.5}$$

It follows from (2.5) that

$$\begin{aligned}&\nabla \times \nabla \times \mathbf{e} - \nabla(\nabla \cdot \mathbf{e}) = k_c^2 \mathbf{e}, \quad \boldsymbol{\rho} \in \Omega,\\ &\mathbf{u}_n \times \mathbf{e} = 0, \nabla \cdot \mathbf{e} = 0, \quad \boldsymbol{\rho} \in \Gamma,\end{aligned}$$

where $k_c^2 = \omega^2\mu\varepsilon_e - \beta^2$. □

2.1.2 Properties of Modal Functions

Multiplying the first equation of (2.3) by $\mathbf{e}$ and taking the integration over the cross section Ω, we obtain

$$k_c^2(\mathbf{e}) = \frac{\int\limits_{\Omega} (|\nabla \times \mathbf{e}|^2 + |\nabla \cdot \mathbf{e}|^2) d\Omega}{\int\limits_{\Omega} |\mathbf{e}|^2 d\Omega}. \tag{2.6}$$

It is easy to verify that the vector function $\mathbf{e}$ that renders (2.6) a minimum and the corresponding constant k_c^2 satisfy (2.3). Let $\mathbf{e}_1$ be the first eigenfunction that minimizes (2.6) and k_{c1}^2 be the corresponding eigenvalue;

$\mathbf{e}_2$ be the eigenfunction that minimizes (2.6) under the supplementary condition that $\mathbf{e}_2$ is perpendicular to $\mathbf{e}_1$, and k_{c2}^2 be the corresponding eigenvalue; and $\mathbf{e}_n$ be the nth eigenfunction that minimizes (2.6) under the supplementary conditions $(\mathbf{e}_n, \mathbf{e}_1) = (\mathbf{e}_n, \mathbf{e}_2) = \cdots (\mathbf{e}_n, \mathbf{e}_{n-1}) = 0$, and k_{cn}^2 be the corresponding eigenvalue, and so on. This procedure generates a set of orthogonal eigenfunctions $\{\mathbf{e}_1, \mathbf{e}_2, \ldots\}$, and the corresponding eigenvalues satisfy $0 \leq k_{c1}^2 \leq k_{c2}^2 \leq \cdots$. The eigenfunction $\mathbf{e}_n$ is called nth **vector modal function**, and the corresponding eigenvalue $k_{cn}(n = 1, 2, \ldots)$ is called **cut-off wavenumber** of the nth vector modal functions. It can be shown that $\lim_{n\to\infty} k_{cn}^2 = \infty$, and the set of vector modal functions is complete (Kurokawa, 1969; Geyi, 2010). From now on, we assume that all vector modal functions are orthonormal, i.e.,

$$\int_{\Omega} \mathbf{e}_m \cdot \mathbf{e}_n \, d\Omega = \delta_{mn}. \tag{2.7}$$

An arbitrary vector function $\mathbf{f}$ can then be expanded as

$$\mathbf{f} = \sum_{n=1}^{\infty} a_n \mathbf{e}_n$$

with $a_n = \int_\Omega \mathbf{f} \cdot \mathbf{e}_n d\Omega$. The vector modal function $\mathbf{e}_n$ belongs to one of the following three categories

$$\begin{aligned} &1.\ \nabla \times \mathbf{e}_n = 0, \quad \nabla \cdot \mathbf{e}_n = 0, \\ &2.\ \nabla \times \mathbf{e}_n \neq 0, \quad \nabla \cdot \mathbf{e}_n = 0, \\ &3.\ \nabla \times \mathbf{e}_n = 0, \quad \nabla \cdot \mathbf{e}_n \neq 0. \end{aligned}$$

The vector modal functions belonging to the first category are called **Transverse electromagnetic** (TEM) **modes**, which satisfy $\nabla \times \mathbf{e}_n = 0$ and $\nabla \cdot \mathbf{e}_n = 0$. For TEM modes, a scalar potential function $\varphi(\boldsymbol{\rho})$ may be introduced such that $\mathbf{e}_n = -\nabla\phi$ and

$$\begin{aligned} &\nabla \cdot \nabla\phi = 0, \quad \boldsymbol{\rho} \in \Omega, \\ &\mathbf{u}_n \times \nabla\phi = 0, \quad \boldsymbol{\rho} \in \Gamma. \end{aligned} \tag{2.8}$$

The second equation implies that the potential function ϕ is constant along the boundary Γ. If Ω is simply connected, the above equations imply $\mathbf{e}_n = 0$ and a hollow waveguide does not support a TEM mode. If Ω is a multiply-connected region (such as a coaxial cable), ϕ may take different values on different conductors. In this case, the waveguide can support a TEM mode. If $\mathbf{e}_n$ is a TEM mode, we have $k_{cn}^2 = 0$.

The vector modal functions belonging to second category are called **transverse electric** (TE) modes, which satisfy $\nabla \times \mathbf{e}_n \neq 0$ and $\nabla \cdot \mathbf{e}_n = 0$. A new scalar function h_{zn} may be introduced

$$\nabla \times \mathbf{e}_n = -\mathbf{u}_z k_{cn} h_{zn}, \tag{2.9}$$

which is proportional to the longitudinal magnetic field. It follows from (2.3) and (2.9) that

$$\begin{aligned} \nabla^2 h_{zn} + k_{cn}^2 h_{zn} = 0, \quad \boldsymbol{\rho} \in \Omega, \\ \mathbf{u}_n \cdot \nabla h_{zn} = 0, \quad \boldsymbol{\rho} \in \Gamma. \end{aligned} \tag{2.10}$$

By definition, we have

$$\int_\Omega h_{zm} h_{zn}\, d\Omega = \frac{1}{k_{cm} k_{cn}} \int_\Omega \nabla \times \mathbf{e}_m \cdot \nabla \times \mathbf{e}_n\, d\Omega = \frac{k_{cm}}{k_{cn}} \int_\Omega \mathbf{e}_m \cdot \mathbf{e}_n\, d\Omega.$$

Hence the set $\{\mathbf{e}_n\}$ is orthonormal if and only if the set $\{h_{zn}\}$ is orthonormal.

The vector modal functions belonging to the third category are called **transverse magnetic** (TM) modes, which satisfies $\nabla \times \mathbf{e}_n = 0$, $\nabla \cdot \mathbf{e}_n \neq 0$. A new scalar function e_{zn} may be introduced such that

$$\nabla \cdot \mathbf{e}_n = k_{cn} e_{zn}. \tag{2.11}$$

The new function e_{zn} is proportional to the longitudinal electric field. It follows from (2.3) and (2.11) that

$$\begin{aligned} &\nabla^2 e_{zn} + k_{cn}^2 e_{zn} = 0, \quad \boldsymbol{\rho} \in \Omega, \\ &e_{zn} = 0, \quad \boldsymbol{\rho} \in \Gamma. \end{aligned} \tag{2.12}$$

Similarly, we have

$$\int_\Omega e_{zm} e_{zn}\, d\Omega = \frac{1}{k_{cm} k_{cn}} \int_\Omega \nabla \cdot \mathbf{e}_m \cdot \nabla \cdot \mathbf{e}_n\, d\Omega = \frac{k_{cm}}{k_{cn}} \int_\Omega \mathbf{e}_m \cdot \mathbf{e}_n\, d\Omega.$$

Thus the set $\{\mathbf{e}_n\}$ is orthonormal if and only if the set $\{e_{zn}\}$ is orthonormal. From the orthonormal set $\{\mathbf{e}_n\}$, we may obtain the following three orthonormal set:

$$\begin{aligned} &\{\mathbf{u}_z \times \mathbf{e}_n \,|\, \mathbf{u}_n \cdot \mathbf{u}_z \times \mathbf{e}_n = 0, \nabla \cdot \mathbf{e}_n = 0, \boldsymbol{\rho} \in \Gamma\}, \\ &\{e_{zn} \,|\, e_{zn} = 0, \boldsymbol{\rho} \in \Gamma\}, \\ &\{h_{zn}, \tilde{c} \,|\, \mathbf{u}_n \cdot \nabla h_{zn} = 0, \boldsymbol{\rho} \in \Gamma\}, \end{aligned}$$

where $\tilde{c}$ is a constant. According to the boundary conditions, $\{\mathbf{e}_n\}$ is most suitable for the expansion of transverse electric field; $\{\mathbf{u}_z \times \mathbf{e}_n\}$ is best suited to the expansion of the transverse magnetic field; $\{e_{zn}\}$ is most appropriate for the expansion of longitudinal electric field; $\{h_{zn}\}$ is most proper for the expansion of longitudinal magnetic field. Introducing the **modal voltage** and **modal current**

$$V_n = \int_\Omega \mathbf{E} \cdot \mathbf{e}_n \, d\Omega, \quad I_n = \int_\Omega \mathbf{H} \cdot \mathbf{u}_z \times \mathbf{e}_n \, d\Omega, \tag{2.13}$$

the electromagnetic fields in the waveguide can be expanded as follows

$$\mathbf{E} = \sum_{n=1}^{\infty} \mathbf{e}_n V_n + \mathbf{u}_z \sum_{n=1}^{\infty} e_{zn} \int_\Omega \mathbf{u}_z \cdot \mathbf{E} e_{zn} \, d\Omega,$$

$$\mathbf{H} = \sum_{n=1}^{\infty} \mathbf{u}_z \times \mathbf{e}_n I_n + \mathbf{u}_z \frac{1}{\Omega} \int_\Omega \mathbf{u}_z \cdot \mathbf{H} \, d\Omega + \sum_{n=1}^{\infty} h_{zn} \int_\Omega \mathbf{H} \cdot h_{zn} \, d\Omega,$$

$$\nabla \times \mathbf{E} = \sum_{n=1}^{\infty} \mathbf{u}_z \times \mathbf{e}_n \int_\Omega \nabla \times \mathbf{E} \cdot \mathbf{u}_z \times \mathbf{e}_n \, d\Omega$$

$$+ \mathbf{u}_z \frac{1}{\Omega} \int_\Omega \mathbf{u}_z \cdot \nabla \times \mathbf{E} \, d\Omega + \sum_{n=1}^{\infty} \mathbf{u}_z h_{zn} \int_\Omega \nabla \times \mathbf{E} \cdot \mathbf{u}_z h_{zn} \, d\Omega,$$

$$\nabla \times \mathbf{H} = \sum_{n=1}^{\infty} \mathbf{e}_n \int_\Omega \nabla \times \mathbf{H} \cdot \mathbf{e}_n \, d\Omega + \mathbf{u}_z \sum_{n=1}^{\infty} e_{zn} \int_\Omega \mathbf{u}_z \cdot \nabla \times \mathbf{H} e_{zn} \, d\Omega.$$

Substituting the above expansions into Maxwell equations and comparing the similar terms we obtain (Geyi, 2010)

$$-\frac{dI_n}{dz} + \int_\Omega \mathbf{H} \cdot \nabla \times \mathbf{e}_n \, d\Omega = j\omega\varepsilon_e V_n,$$

$$k_{cn}^2 I_n = j\omega\varepsilon_e \int_\Omega \mathbf{u}_z \cdot \mathbf{E} \nabla \cdot \mathbf{e}_n \, d\Omega,$$

$$\frac{dV_n}{dz} + \int_\Omega \mathbf{u}_z \cdot \mathbf{E} \nabla \cdot \mathbf{e}_n \, d\Omega = -j\omega\mu I_n,$$

$$k_{cn}^2 V_n = -j\omega\mu \int_\Omega \mathbf{H} \cdot \nabla \times \mathbf{e}_n \, d\Omega,$$

$$\int_\Omega \mathbf{H} \cdot \frac{\mathbf{u}_z}{\Omega^{1/2}} \, d\Omega = 0,$$

where $\varepsilon_e = \varepsilon(1 - j\frac{\sigma}{\omega\varepsilon})$ and the waveguide is assumed to be perfectly conducting. The modal voltage and modal current satisfy the transmission line equation

$$\frac{dV_n}{dz} = -j\beta_n Z_{wn} I_n(z), \quad \frac{dI_n}{dz} = -j\beta_n Y_{wn} V_n(z), \tag{2.14}$$

where $Y_{wn} = 1/Z_{wn}$ and

$$\beta_n = \begin{cases} k, & \text{TEM mode} \\ \sqrt{k^2 - k_{cn}^2}, & \text{TE or TM mode} \end{cases}, \quad Z_{wn} = \begin{cases} \eta, & \text{TEM mode} \\ \eta k/\beta_n, & \text{TE mode} \\ \eta\beta_n/k, & \text{TM mode} \end{cases}. \tag{2.15}$$

Here $k = \omega\sqrt{\mu\varepsilon_e}$, $\eta = \sqrt{\mu/\varepsilon_e}$, and Z_{wn} is called the **wave impedance** of the nth mode. If $\beta_n \neq 0$, the solutions of (2.14) can be expressed as

$$\begin{aligned} V_n(z) &= V_n^+(z) + V_n^-(z) = A_n e^{-j\beta_n z} + B_n e^{j\beta_n z}, \\ I_n(z) &= I_n^+(z) - I_n^-(z) = (A_n e^{-j\beta_n z} - B_n e^{j\beta_n z}) Z_{wn}^{-1}, \end{aligned} \tag{2.16}$$

where the superscript $+$ and $-$ represent the wave propagating in $+z$ and $-z$ direction respectively

$$\begin{aligned} V_n^+(z) &= A_n e^{-j\beta_n z}, \quad V_n^-(z) = B_n e^{j\beta_n z}, \\ I_n^+(z) &= A_n Z_{wn}^{-1} e^{-j\beta_n z}, \quad I_n^-(z) = B_n Z_{wn}^{-1} e^{j\beta_n z}. \end{aligned}$$

The **characteristic impedance** for the nth mode is defined by

$$Z_{0n} = \frac{V_n^+}{I_n^+} = \frac{V_n^-}{I_n^-} = Z_{wn}.$$

The **guide wavelength** for the nth mode is defined by

$$\lambda_n = \frac{2\pi}{\beta_n}.$$

Other expansion coefficients may be represented by

$$\int_\Omega \mathbf{u}_z \cdot \mathbf{H} h_{zn} d\Omega = \frac{k_{cn}}{j\beta_n} \frac{V_n(z)}{Z_{wn}},$$

$$\int_\Omega \mathbf{u}_z \cdot \mathbf{E} e_{zn} d\Omega = \frac{k_{cn}}{j\beta_n} I_n(z) Z_{wn}.$$

Finally, the total fields in the waveguide may be written as

$$\begin{aligned}\mathbf{E} &= \sum_{n=1}^{\infty}\left(V_n\mathbf{e}_n + I_n Z_{wn}\frac{k_{cn}}{j\beta_n}\mathbf{u}_z\frac{\nabla\cdot\mathbf{e}_n}{k_{cn}}\right),\\ \mathbf{H} &= \sum_{n=1}^{\infty}\left(I_n\mathbf{u}_z\times\mathbf{e}_n - \frac{V_n}{Z_{wn}}\frac{k_{cn}}{j\beta_n}\frac{\nabla\times\mathbf{e}_n}{k_{cn}}\right).\end{aligned} \tag{2.17}$$

The above expansions are fundamental in the study of waveguide discontinuities. The power along the waveguide is given by

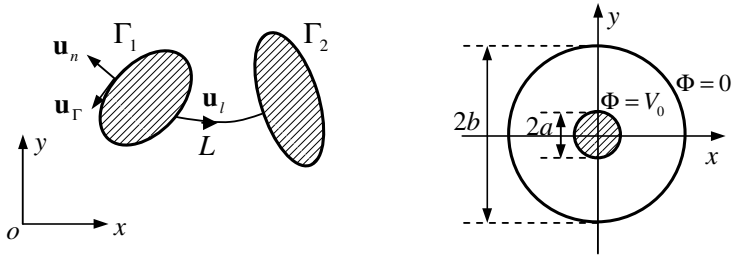
$$P = \frac{1}{2}\mathrm{Re}\int_{\Omega}\mathbf{E}\times\bar{\mathbf{H}}\cdot\mathbf{u}_z\,d\Omega = \sum_{n=1}^{\infty}\frac{1}{2}\mathrm{Re}(V_n\bar{I}_n). \tag{2.18}$$

Remark 2.2 (Equivalent voltage and current)**:** At microwave frequencies, there are no voltmeter and ammeters for the measurement of voltages and currents. In addition, the definition of voltage or current is not unique in most situations. For a TEM transmission line, it is possible to define a voltage and a current, which are uniquely related to the transverse electric and transverse magnetic fields respectively. Consider a TEM transmission line consisting of two conductors [Figure 2.3(a)], and assume that the transverse electric field for the TEM mode propagating in the $+z$ direction is

$$\mathbf{E}_t = -\nabla\Phi e^{-j\beta z}. \tag{2.19}$$

The voltage propagating in the $+z$ direction may be defined by

$$V^+ = \int_L \mathbf{E}_t\cdot\mathbf{u}_l\,dl, \tag{2.20}$$



(a) (b)

Figure 2.3 TEM waveguides. (a) An arbitrary TEM line. (b) A coaxial transmission line.

where L is an arbitrary path connecting the two conductors. Substituting (2.19) into (2.20) gives

$$V^+ = V_0 e^{-j\beta z}, \tag{2.21}$$

where $V_0 = \Phi_1 - \Phi_2$ is the potential difference between the two conductors. Note that the potential function Φ is proportional to ϕ defined in (2.8). The current propagating in the $+z$ direction can be defined by

$$I^+ = \int_\Gamma \mathbf{J}_s \cdot \mathbf{u}_z \, d\Gamma = \int_\Gamma J_s d\Gamma = \int_\Gamma \mathbf{H}_t \cdot \mathbf{u}_\Gamma \, d\Gamma, \tag{2.22}$$

where Γ represents the boundary of one of the conductors, $\mathbf{H}_t$ is the transverse magnetic field for the TEM mode propagating in the $+z$ direction, and $\mathbf{J}_s = \mathbf{u}_n \times \mathbf{H}_t = \mathbf{u}_z J_s$ is the surface current flowing in the $+z$ direction.

For the coaxial transmission line shown in Figure 2.3(b), the potential function Φ satisfies the boundary conditions $\Phi = V_0$ at $\rho = a$ and $\Phi = 0$ at $\rho = b$ and is given by

$$\Phi = V_0 \frac{\ln(\rho/b)}{\ln(a/b)}.$$

Thus the electric and magnetic fields for the TEM mode propagating in the $+z$ direction are

$$\begin{aligned} \mathbf{E}_t &= -\nabla\Phi e^{-j\beta z} = V_0 e^{-j\beta z}\mathbf{e}_0, \\ \mathbf{H}_t &= \frac{1}{\eta}\mathbf{u}_z \times \mathbf{E}_t = \frac{1}{\eta} V_0 e^{-j\beta z}\mathbf{u}_z \times \mathbf{e}_0, \end{aligned} \tag{2.23}$$

where $\mathbf{e}_0$ is the vector modal function (not normalized)

$$\mathbf{e}_0 = \frac{1}{\rho \ln(b/a)}\mathbf{u}_\rho. \tag{2.24}$$

The current propagating in the $+z$ direction is then given by

$$I^+ = \int_\Gamma \mathbf{H}_t \cdot \mathbf{u}_\Gamma d\Gamma = I_0 e^{-j\beta z}, \tag{2.25}$$

where

$$I_0 = \frac{2\pi}{\eta \ln(b/a)} V_0.$$

The characteristic impedance may then be defined by

$$Z_0 = \frac{V^+}{I^+} = \frac{\eta \ln(b/a)}{2\pi}. \tag{2.26}$$

The power along the line is

$$P = \frac{1}{2}\mathrm{Re}\int_{\Omega} \mathbf{E}_t \times \bar{\mathbf{H}}_t \cdot \mathbf{u}_z \, d\Omega = \frac{1}{2}\mathrm{Re}\int_a^b \int_0^{2\pi} \mathbf{E}_t \times \bar{\mathbf{H}}_t \cdot \mathbf{u}_z \rho \, d\rho \, d\varphi = \frac{\pi V_0^2}{\ln(b/a)}. \tag{2.27}$$

The power is also given by the following expression

$$P = \frac{1}{2}\mathrm{Re}V^+\bar{I}^+ = \frac{1}{2}|I^+|^2 Z_0 = \frac{1}{2}\frac{|V^+|^2}{Z_0} = \frac{\pi V_0^2}{\ln(b/a)}. \quad \square$$

2.1.3 Mode Excitation

Consider a uniform waveguide excited by the electric current source $\mathbf{J}$ and the magnetic current $\mathbf{J}_m$ confined in the region $z_1 < z < z_2$, as shown in Figure 2.4. According to (2.17), the fields for $z \geq z_2$ and $z \leq z_1$ in the waveguide may be respectively expanded in terms of the vector modal functions as follows

$$\begin{aligned} \mathbf{E}^+ &= \sum_{n=1}^{\infty} \left(V_n^+ \mathbf{e}_n + I_n^+ Z_{wn} \frac{k_{cn}}{j\beta_n} \mathbf{u}_z \frac{\nabla \cdot \mathbf{e}_n}{k_{cn}} \right) = \sum_{n=1}^{\infty} A_n \mathbf{E}_n^+, \\ \mathbf{H}^+ &= \sum_{n=1}^{\infty} \left(I_n^+ \mathbf{u}_z \times \mathbf{e}_n - \frac{V_n^+}{Z_{wn}} \frac{k_{cn}}{j\beta_n} \frac{\nabla \times \mathbf{e}_n}{k_{cn}} \right) = \sum_{n=1}^{\infty} A_n \mathbf{H}_n^+, \end{aligned} \tag{2.28}$$

$$\begin{aligned} \mathbf{E}^- &= \sum_{n=1}^{\infty} \left(V_n^- \mathbf{e}_n - \frac{k_{cn}}{j\beta_n} I_n^- Z_{wn} \mathbf{u}_z \frac{\nabla \cdot \mathbf{e}_n}{k_{cn}} \right) = \sum_{n=1}^{\infty} B_n \mathbf{E}_n^-, \\ \mathbf{H}^- &= \sum_{n=1}^{\infty} \left(-I_n^- \mathbf{u}_z \times \mathbf{e}_n - \frac{k_{cn}}{j\beta_n} \frac{V_n^-}{Z_{wn}} \frac{\nabla \times \mathbf{e}_n}{k_{cn}} \right) = \sum_{n=1}^{\infty} B_n \mathbf{H}_n^-, \end{aligned} \tag{2.29}$$

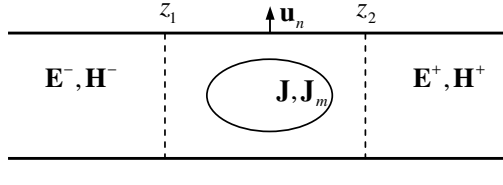


Figure 2.4 Mode excitation.

where

$$\begin{aligned} \mathbf{E}_n^+ &= (\mathbf{e}_n + \mathbf{u}_z e_{zn}) e^{-j\beta_n z}, \quad & \mathbf{H}_n^+ &= (\mathbf{h}_n + \mathbf{u}_z h_{zn}) e^{-j\beta_n z}, \\ \mathbf{E}_n^- &= (\mathbf{e}_n - \mathbf{u}_z e_{zn}) e^{j\beta_n z}, & \mathbf{H}_n^- &= (-\mathbf{h}_n + \mathbf{u}_z h_{zn}) e^{j\beta_n z}, \end{aligned} \tag{2.30}$$

with

$$\mathbf{h}_n = \frac{\mathbf{u}_z \times \mathbf{e}_n}{Z_{wn}}, \quad e_{zn} = \frac{\nabla \cdot \mathbf{e}_n}{j\beta_n}, \quad h_{zn}\mathbf{u}_z = -\frac{\nabla \times \mathbf{e}_n}{j\beta_n Z_{wn}}. \tag{2.31}$$

Note that

$$\int_\Omega (\mathbf{e}_n \times \mathbf{h}_n) \cdot \mathbf{u}_z d\Omega = \frac{1}{Z_{wn}}, \tag{2.32}$$

where Ω is the cross section of the waveguide. The expansion coefficients in (2.28) and (2.29) can be determined by using the Lorentz reciprocity theorem. Since the fields $\mathbf{E}_n^\pm, \mathbf{H}_n^\pm$ satisfy the source-free Maxwell equations, we have

$$\begin{aligned} \nabla \cdot (\mathbf{E}_n^\pm \times \mathbf{H} - \mathbf{E} \times \mathbf{H}_n^\pm) &= \mathbf{H} \cdot \nabla \times \mathbf{E}_n^\pm - \mathbf{E}_n^\pm \cdot \nabla \times \mathbf{H} - \mathbf{H}_n^\pm \cdot \nabla \\ &\quad \times \mathbf{E} + \mathbf{E} \cdot \nabla \times \mathbf{H}_n^\pm \\ &= \mathbf{H} \cdot (-j\omega\mu_0 \mathbf{H}_n^\pm) - \mathbf{E}_n^\pm \cdot (\mathbf{J} + j\omega\varepsilon_0 \mathbf{E}) \\ &\quad - \mathbf{H}_n^\pm \cdot (-j\omega\mu_0 \mathbf{H} - \mathbf{J}_m) + \mathbf{E} \cdot (j\omega\varepsilon_0 \mathbf{E}_n^\pm) \\ &= -\mathbf{J} \cdot \mathbf{E}_n^\pm + \mathbf{J}_m \cdot \mathbf{H}_n^\pm. \end{aligned}$$

Integrating over the volume V bounded by the perfectly conducting walls and the two cross-sectional planes $z = z_1$ and $z = z_2$, we have

$$\begin{aligned} &- \int_{z=z_1} (\mathbf{E}_n^\pm \times \mathbf{H} - \mathbf{E} \times \mathbf{H}_n^\pm) \cdot \mathbf{u}_z \, dS + \int_{z=z_2} (\mathbf{E}_n^\pm \times \mathbf{H} - \mathbf{E} \times \mathbf{H}_n^\pm) \cdot \mathbf{u}_z \, dS \\ &\quad = \int_V (-\mathbf{J} \cdot \mathbf{E}_n^\pm + \mathbf{J}_m \cdot \mathbf{H}_n^\pm) dV. \end{aligned}$$

It is readily found that

$$\begin{aligned} &\int_{z=z_1} (\mathbf{E}_n^+ \times \mathbf{H} - \mathbf{E} \times \mathbf{H}_n^+) \cdot \mathbf{u}_z \, dS \\ &\quad = \int_{z=z_1} B_n (\mathbf{E}_n^+ \times \mathbf{H}_n^- - \mathbf{E}_n^- \times \mathbf{H}_n^+) \cdot \mathbf{u}_z \, dS = -\frac{2B_n}{Z_{wn}}, \end{aligned}$$

$$\int_{z=z_2} (\mathbf{E}_n^+ \times \mathbf{H} - \mathbf{E} \times \mathbf{H}_n^+) \cdot \mathbf{u}_z \, dS$$
$$= \int_{z=z_2} A_n (\mathbf{E}_n^+ \times \mathbf{H}_n^+ - \mathbf{E}_n^+ \times \mathbf{H}_n^+) \cdot \mathbf{u}_z \, dS = 0,$$
$$\int_{z=z_1} (\mathbf{E}_n^- \times \mathbf{H} - \mathbf{E} \times \mathbf{H}_n^-) \cdot \mathbf{u}_z \, dS$$
$$= \int_{z=z_1} B_n (\mathbf{E}_n^- \times \mathbf{H}_n^- - \mathbf{E}_n^- \times \mathbf{H}_n^-) \cdot \mathbf{u}_z \, dS = 0,$$
$$\int_{z=z_2} (\mathbf{E}_n^- \times \mathbf{H} - \mathbf{E} \times \mathbf{H}_n^-) \cdot \mathbf{u}_z \, dS$$
$$= \int_{z=z_2} A_n (\mathbf{E}_n^- \times \mathbf{H}_n^+ - \mathbf{E}_n^+ \times \mathbf{H}_n^-) \cdot \mathbf{u}_z \, dS = \frac{2A_n}{Z_{wn}}.$$

Thus we have

$$A_n = \frac{Z_{wn}}{2} \int_V (-\mathbf{J} \cdot \mathbf{E}_n^- + \mathbf{J}_m \cdot \mathbf{H}_n^-) dV,$$
$$B_n = \frac{Z_{wn}}{2} \int_V (-\mathbf{J} \cdot \mathbf{E}_n^+ + \mathbf{J}_m \cdot \mathbf{H}_n^+) dV. \tag{2.33}$$

2.2 Vector Modal Functions

The fields in the waveguide can be expanded as a linear combination of various modes. The transverse components of the modes satisfy the vector equation (2.3). For a waveguide filled with homogeneous medium, the transverse components can be derived from the longitudinal components that satisfy the Helmholtz equation and the related boundary conditions. For typical waveguides, the method of separation of variables may be used to solve the Helmholtz equation.

2.2.1 Rectangular Waveguide

A homogeneous rectangular waveguide shown in Figure 2.5 only supports TE or TM modes, which can be determined from (2.10) and (2.12). Let both $e_{zn}(x, y)$ and $h_{zn}(x, y)$ be a product of functions of separate x and y

$$e_{zn}(x, y) = e_{p_e}(x) e_{q_e}(y), \quad h_{zn}(x, y) = h_{p_h}(x) h_{q_h}(y). \tag{2.34}$$

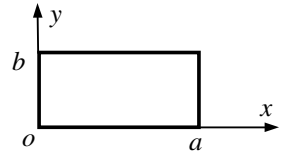


Figure 2.5 Rectangular waveguide.

Substituting these back into (2.10) and (2.12) gives

$$\left(\frac{d^2}{dx^2}+p_e^2\right)e_{p_e}(x)=0,\quad e_{p_e}(0)=e_{p_e}(a)=0,$$

$$\left(\frac{d^2}{dy^2}+q_e^2\right)e_{q_e}(y)=0,\quad e_{q_e}(0)=e_{q_e}(b)=0,$$

$$\left(\frac{d^2}{dx^2}+p_h^2\right)h_{p_h}(x)=0,\quad \left.\frac{\partial h_{p_h}}{\partial x}\right|_{x=0}=\left.\frac{\partial h_{p_h}}{\partial x}\right|_{x=a}=0,$$

$$\left(\frac{d^2}{dy^2}+q_h^2\right)h_{q_h}(y)=0,\quad \left.\frac{\partial h_{q_h}}{\partial y}\right|_{y=0}=\left.\frac{\partial h_{q_h}}{\partial y}\right|_{y=b}=0,$$

where p and q are separation constants satisfying

$$k_{cn,e}^2=p_e^2+q_e^2,\quad k_{cn,h}^2=p_h^2+q_h^2.$$

The normalized eigenfunctions are given by

$$e_{p_e}(x)=\sqrt{\frac{2}{a}}\sin\frac{p\pi}{a}x,\quad p_e=\frac{p\pi}{a},\quad p=1,2,\ldots,$$

$$e_{q_e}(y)=\sqrt{\frac{2}{b}}\sin\frac{q\pi}{b}y,\quad q_e=\frac{q\pi}{b},\quad q=1,2,\ldots,$$

$$h_{p_h}(x)=\sqrt{\frac{\varepsilon_p}{a}}\cos\frac{p\pi}{a}x,\quad p_h=\frac{p\pi}{a},\quad p=0,1,2,\ldots,$$

$$h_{q_h}(y)=\sqrt{\frac{\varepsilon_q}{b}}\cos\frac{q\pi}{b}y,\quad q_h=\frac{q\pi}{b},\quad q=0,1,2,\ldots,$$

where $\varepsilon_m=\begin{cases}1, m=0\\ 2, m\geq 1\end{cases}$. These eigenfunctions satisfy

$$\delta(x-x')=\sum_{p=1}^{\infty}e_{p_e}(x)e_{p_e}(x'),$$

$$\delta(y-y') = \sum_{q=1}^{\infty} e_{q_e}(y)e_{q_e}(y'),$$

$$\delta(x-x') = \sum_{p=0}^{\infty} h_{p_h}(x)h_{p_h}(x'),$$

$$\delta(y-y') = \sum_{q=0}^{\infty} h_{q_h}(y)h_{q_h}(y'), \tag{2.35}$$

and

$$\begin{aligned}\delta(x-x')\delta(y-y') &= \sum_{p=1}^{\infty}\sum_{q=1}^{\infty} e_{p_e}(x)e_{q_e}(y)e_{p_e}(x')e_{q_e}(y') \\ &= \sum_{n=1}^{\infty} e_{zn}(x,y)e_{zn}(x',y'),\end{aligned} \tag{2.36}$$

$$\begin{aligned}\delta(x-x')\delta(y-y') &= \sum_{p=0}^{\infty}\sum_{q=0}^{\infty} h_{p_h}(x)h_{q_h}(y)h_{p_h}(x')h_{q_h}(y') \\ &= \sum_{n=0}^{\infty} h_{zn}(x,y)h_{zn}(x',y'),\end{aligned} \tag{2.37}$$

where

$$e_{zn}(x,y) = e_{p_e}(x)e_{q_e}(y) = \sqrt{\frac{4}{ab}}\sin\frac{p\pi}{a}x\sin\frac{q\pi}{b}y,$$

$$h_{zn}(x,y) = h_{p_h}(x)h_{q_h}(y) = \sqrt{\frac{\varepsilon_p\varepsilon_q}{ab}}\cos\frac{p\pi}{a}x\cos\frac{q\pi}{b}y,$$

are the modal solutions. Note that the subscript n represents the multiple index (p,q). The cut-off wavenumbers for TM and TE modes are respectively given by

$$k_{cn,e}^2 = \left(\frac{p\pi}{a}\right)^2 + \left(\frac{q\pi}{b}\right)^2; \quad p,q = 1,2,\ldots,$$

$$k_{cn,h}^2 = \left(\frac{p\pi}{a}\right)^2 + \left(\frac{q\pi}{b}\right)^2; \quad p,q = 0,1,2,\ldots.$$

The vector modal functions can be obtained from (2.3). For TM modes, the vector modal functions are given by

$$\mathbf{e}_n = \mathbf{u}_x \frac{1}{k_{cn}} \frac{p\pi}{a} \sqrt{\frac{4}{ab}} \cos\frac{p\pi}{a}x \sin\frac{q\pi}{b}y + \mathbf{u}_y \frac{1}{k_{cn}} \frac{q\pi}{b} \sqrt{\frac{4}{ab}} \sin\frac{p\pi}{a}x \cos\frac{q\pi}{b}y, \tag{2.38}$$

and the vector modal functions for TE modes are

$$\mathbf{e}_n = \mathbf{u}_x \frac{1}{k_{cn}} \frac{q\pi}{b} \sqrt{\frac{\varepsilon_p \varepsilon_q}{ab}} \cos\frac{p\pi}{a}x \sin\frac{q\pi}{b}y - \mathbf{u}_y \frac{1}{k_{cn}} \frac{p\pi}{a} \sqrt{\frac{\varepsilon_p \varepsilon_q}{ab}} \sin\frac{p\pi}{a}x \cos\frac{q\pi}{b}y \tag{2.39}$$

where $\mathbf{u}_x$ and $\mathbf{u}_y$ are unit vectors along x and y direction respectively. The dominant mode in the rectangular waveguide is TE_{10} mode and is the most commonly used one. The field components of this mode are (the wave is assumed to be propagating in $+z$ direction)

$$\begin{aligned} H_z &= A\cos\frac{\pi x}{a} e^{-j\beta_{10}z}, \\ H_x &= \frac{j\beta_{10}}{k_{c10}} A\sin\frac{\pi x}{a} e^{-j\beta_{10}z}, \\ E_y &= -jZ_{w10}\frac{\beta_{10}}{k_{c10}} A\sin\frac{\pi x}{a} e^{-j\beta_{10}z}, \end{aligned}$$

where

$$k_{c10} = \frac{\pi}{a}, \quad \beta_{10} = \sqrt{k^2 - k_{c10}^2}, \quad Z_{w10} = \frac{\eta k}{\beta_{10}}.$$

The guide wavelength for TE_{10} mode is

$$\lambda_{g10} = \frac{2\pi}{\beta_{10}} = \frac{\lambda}{\sqrt{1-(\lambda/2a)^2}}.$$

To maintain a single dominant mode operation, the dimensions of the rectangular waveguide must satisfy

$$b < \frac{\lambda}{2} < a < \lambda.$$

Usually, we choose $a = 2b$, $a = 0.7\lambda$.

2.2.2 Circular Waveguide

A uniform waveguide of circular cross section of radius a is shown in Figure 2.6. The waveguide is best described by the cylindrical coordinate system (ρ, φ, z), in which we have

$$\nabla_t^2 = \frac{1}{\rho}\frac{\partial}{\partial\rho}\rho\frac{\partial}{\partial\rho} + \frac{1}{\rho^2}\frac{\partial^2}{\partial\varphi^2}. \tag{2.40}$$

Let both $e_{zn}(\rho, \varphi)$ and $h_{zn}(\rho, \varphi)$ be a product of functions of separate ρ and φ

$$e_{zn}(\rho, \varphi) = e_{p_e}(\rho)e_{q_e}(\varphi), \quad h_{zn}(\rho, \varphi) = h_{p_h}(\rho)h_{q_h}(\varphi). \tag{2.41}$$

Substituting these into (2.10) and (2.12), we obtain

$$\left(\frac{d^2}{d\rho^2} + \frac{1}{\rho}\frac{d}{d\rho} + k_{cn,e}^2 - \frac{q^2}{\rho^2}\right) e_{p_e}(\rho) = 0, \quad e_{p_e}(a) = 0,$$

$$\left(\frac{d^2}{d\varphi^2} + q^2\right) e_{q_e}(\varphi) = 0, \quad e_{q_e}(0) = e_{q_e}(2\pi), \quad \frac{de_{q_e}(0)}{d\varphi} = \frac{de_{q_e}(2\pi)}{d\varphi},$$

$$\left(\frac{d^2}{d\rho^2} + \frac{1}{\rho}\frac{d}{d\rho} + k_{cn,h}^2 - \frac{q^2}{\rho^2}\right) h_{p_h}(\rho) = 0, \quad \frac{dh_{p_h}(a)}{d\rho} = 0,$$

$$\left(\frac{d^2}{d\varphi^2} + q^2\right) h_{q_h}(\varphi) = 0, \quad h_{q_h}(0) = h_{q_h}(2\pi), \quad \frac{dh_{q_h}(0)}{d\varphi} = \frac{dh_{q_h}(2\pi)}{d\varphi}.$$

The normalized eigenfunctions for TM modes are

$$e_{p_e}\left(\chi_{qp}\frac{\rho}{a}\right) = \frac{\sqrt{2}J_q\left(\chi_{qp}\frac{\rho}{a}\right)}{\chi_{qp}J_{q+1}\left(\chi_{qp}\right)}, \quad q = 0, 1, 2, \ldots,$$

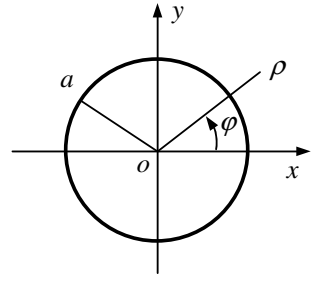


Figure 2.6 Circular waveguide.

$$e_{q_e}(\varphi) = \sqrt{\frac{\varepsilon_q}{2\pi}} \begin{pmatrix} \cos q\varphi \\ \sin q\varphi \end{pmatrix}, \quad q = 0, 1, 2, \ldots,$$

where χ_{qp} is the pth non-vanishing root of the equation

$$J_q(\chi_{qp}) = 0.$$

The cut-off wavenumbers for TM modes are given by

$$k_{cn,e} = \frac{\chi_{qp}}{a}.$$

The normalized vector modal functions for TM modes may be determined by (2.3) as follows

$$\mathbf{e}_n = -\mathbf{u}_\rho \sqrt{\frac{\varepsilon_q}{\pi}} \frac{J_q'\left(\chi_{qp}\frac{\rho}{a}\right)}{aJ_{q+1}(\chi_{qp})} \begin{pmatrix} \cos q\varphi \\ \sin q\varphi \end{pmatrix} \pm \mathbf{u}_\varphi \sqrt{\frac{\varepsilon_q}{\pi}} \frac{q}{\chi_{qp}} \frac{J_q\left(\chi_{qp}\frac{\rho}{a}\right)}{\rho J_{q+1}(\chi_{qp})} \begin{pmatrix} \sin q\varphi \\ \cos q\varphi \end{pmatrix}. \tag{2.42}$$

where $\mathbf{u}_\rho$ and $\mathbf{u}_\varphi$ are unit vectors along ρ and φ direction respectively. The normalized eigenfunctions for TE modes are

$$h_{p_h}\left(\chi_{qp}'\frac{\rho}{a}\right) = \frac{\sqrt{2}J_q\left(\chi_{qp}'\frac{\rho}{a}\right)}{\sqrt{\chi_{qp}'^2 - q^2}J_q(\chi_{qp}')}, \quad q = 0, 1, 2, \ldots,$$

$$h_{q_h}(\varphi) = \sqrt{\frac{\varepsilon_q}{2\pi}} \begin{pmatrix} \cos q\varphi \\ \sin q\varphi \end{pmatrix}, \quad q = 0, 1, 2, \ldots,$$

where χ_{qp}' is the pth non-vanishing root of the equation

$$J_q'(\chi_{qp}') = 0.$$

The cut-off wavenumbers for the TE modes are given by

$$k_{cn,h} = \frac{\chi_{qp}'}{a}.$$

The normalized vector modal functions for TE modes are

$$\mathbf{e}_n = \pm\mathbf{u}_\rho \sqrt{\frac{\varepsilon_q}{\pi}} \frac{q}{\sqrt{\chi_{qp}'^2 - q^2}} \frac{J_q\left(\chi_{qp}'\frac{\rho}{a}\right)}{\rho J_q\left(\chi_{qp}'\right)} \begin{pmatrix} \sin q\varphi \\ \cos q\varphi \end{pmatrix}$$

$$+ \mathbf{u}_\varphi \sqrt{\frac{\varepsilon_q}{\pi}} \frac{\chi_{qp}'}{\sqrt{\chi_{qp}'^2 - q^2}} \frac{J_q'\left(\chi_{qp}'\frac{\rho}{a}\right)}{aJ_q(\chi_{qp}')} \begin{pmatrix} \cos q\varphi \\ \sin q\varphi \end{pmatrix}. \tag{2.43}$$

2.2.3 Coaxial Waveguide

A coaxial waveguide is shown in Figure 2.7. The dominant mode for coaxial waveguide is the TEM mode. The potential function ϕ for the TEM mode can be determined from the Laplace equation (2.8), and it may be written as

$$\phi = \frac{\ln \rho}{\sqrt{2\pi \ln c_1}},$$

where $c_1 = b/a$. The normalized vector modal function for the TEM mode may be obtained from

$$\mathbf{e}_n = \nabla\phi = \mathbf{u}_\rho \frac{l}{\sqrt{2\pi \ln c_1}} \frac{1}{\rho}. \tag{2.44}$$

For the higher order modes, we need to solve the Helmholtz equation. Let both $e_{zn}(\rho, \varphi)$ and $h_{zn}(\rho, \varphi)$ be a product of functions of separate ρ and φ

$$e_{zn}(\rho, \varphi) = e_{p_e}(\rho) e_{q_e}(\varphi), \quad h_{zn}(\rho, \varphi) = h_{p_h}(\rho) h_{q_h}(\varphi). \tag{2.45}$$

Substituting these into (2.10) and (2.12), we obtain

$$\left(\frac{d^2}{d\rho^2} + \frac{1}{\rho}\frac{d}{d\rho} + k_{cn,e}^2 - \frac{q^2}{\rho^2}\right) e_{p_e}(\rho) = 0, \quad e_{p_e}(a) = e_{p_e}(b) = 0,$$

$$\left(\frac{d^2}{d\varphi^2} + q^2\right) e_{q_e}(\varphi) = 0, \quad e_{q_e}(0) = e_{q_e}(2\pi), \quad \frac{de_{q_e}(0)}{d\varphi} = \frac{de_{q_e}(2\pi)}{d\varphi},$$

$$\left(\frac{d^2}{d\rho^2} + \frac{1}{\rho}\frac{d}{d\rho} + k_{cn,h}^2 - \frac{q^2}{\rho^2}\right) h_{p_h}(\rho) = 0, \quad \frac{dh_{p_h}(a)}{d\rho} = \frac{dh_{p_h}(b)}{d\rho} = 0,$$

$$\left(\frac{d^2}{d\varphi^2} + q^2\right) h_{q_h}(\varphi) = 0, \quad h_{q_h}(0) = h_{q_h}(2\pi), \quad \frac{dh_{q_h}(0)}{d\varphi} = \frac{dh_{q_h}(2\pi)}{d\varphi}.$$

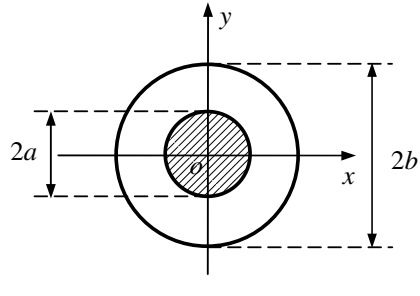


Figure 2.7 Coaxial waveguide.

The normalized eigenfunctions for TM modes are easily found to be

$$e_{p_e}\left(\chi_{qp}\frac{\rho}{a}\right) = \frac{\pi}{\sqrt{2}}\frac{J_q\left(\chi_{qp}\frac{\rho}{a}\right)N_q(\chi_{qp}) - N_q\left(\chi_{qp}\frac{\rho}{a}\right)J_q(\chi_{qp})}{\sqrt{J_q^2(\chi_{qp})/J_q^2(c\chi_{qp}) - 1}}, \quad q = 0, 1, 2, \ldots,$$

$$e_{q_e}(\varphi) = \sqrt{\frac{\varepsilon_q}{2\pi}}\begin{pmatrix}\cos q\varphi\\ \sin q\varphi\end{pmatrix}, \quad q = 0, 1, 2, \ldots,$$

where χ_{qp} is the pth non-vanishing root of the equation

$$J_q(c_1\chi_{qp})N_q(\chi_{qp}) - N_q(c_1\chi_{qp})J_q(\chi_{qp}) = 0.$$

The cut-off wavenumbers for TM modes are given by

$$k_{cn,e} = \frac{\chi_{qp}}{a} = \frac{(c_1 - 1)\chi_{qp}}{b - a} \approx \frac{\pi p}{b - a}, \quad p = 1, 2, \ldots.$$

The normalized vector modal functions for TM modes may be determined from (2.3) as follows

$$\begin{aligned}\mathbf{e}_n = &-\mathbf{u}_\rho\frac{\chi_{qp}}{a}e'_{p_e}\left(\chi_{qp}\frac{\rho}{a}\right)\sqrt{\frac{\varepsilon_q}{2\pi}}\begin{pmatrix}\cos q\varphi\\ \sin q\varphi\end{pmatrix}\\ &\pm\mathbf{u}_\varphi\frac{q}{\rho}e_{p_e}\left(\chi_{qp}\frac{\rho}{a}\right)\sqrt{\frac{\varepsilon_q}{2\pi}}\begin{pmatrix}\sin q\varphi\\ \cos q\varphi\end{pmatrix},\end{aligned} \tag{2.46}$$

where e'_{p_e} denotes the derivative with respect to its argument.

The normalized eigenfunctions for TE modes are

$$h_{p_h}\left(\chi'_{qp}\frac{\rho}{a}\right) = \frac{\pi}{\sqrt{2}}\frac{J_q\left(\chi'_{qp}\frac{\rho}{a}\right)N'_q(\chi_{qp}) - N_q\left(\chi'_{qp}\frac{\rho}{a}\right)J'_q(\chi'_{qp})}{\sqrt{\frac{J_q^2(\chi'_{qp})}{J_q^2(c_1\chi'_{qp})}\left[1 - \left(\frac{q}{c_1\chi'_{qp}}\right)^2\right] - \left[1 - \left(\frac{q}{\chi'_{qp}}\right)^2\right]}},$$

$$q = 0, 1, 2, \ldots,$$

$$h_{q_h}(\varphi) = \sqrt{\frac{\varepsilon_q}{2\pi}}\begin{pmatrix}\cos q\varphi\\ \sin q\varphi\end{pmatrix}, \quad q = 0, 1, 2, \ldots,$$

where χ'_{qp} is the pth non-vanishing root of the equation

$$J'_q(c_1\chi'_{qp})N'_q(\chi'_{qp}) - N'_q(c_1\chi'_{qp})J'_q(\chi'_{qp}) = 0.$$

The cut-off wavenumbers for TE modes are given by

$$k_{cn,h} = \frac{\chi'_{q1}}{a} = \frac{(c_1+1)\chi'_{q1}}{b+a} \approx \frac{2q}{b+a}, \quad q = 1, 2, \ldots,$$

$$k_{cn,h} = \frac{\chi'_{qp}}{a} = \frac{(c_1-1)\chi'_{qp}}{b-a} \approx \frac{(p-1)\pi}{b-a}, \quad p = 2, 3, \ldots.$$

The dominant TE mode is $\mathrm{TE}_{11}(q=1, p=1)$. The normalized vector modal functions for TE modes may be written as

$$\mathbf{e}_n = \pm \mathbf{u}_\rho \frac{q}{\rho} h_{p_e}\left(\chi'_{qp}\frac{\rho}{a}\right)\sqrt{\frac{\varepsilon_q}{2\pi}}\begin{pmatrix}\sin q\varphi \\ \cos q\varphi\end{pmatrix} + \mathbf{u}_\varphi \frac{\chi'_{qp}}{a} h'_{p_e}\left(\chi'_{qp}\frac{\rho}{a}\right)\sqrt{\frac{\varepsilon_q}{2\pi}}\begin{pmatrix}\cos q\varphi \\ \sin q\varphi\end{pmatrix}. \tag{2.47}$$

2.2.4 Numerical Analysis for Metal Waveguides

For an arbitrary metal waveguide, we have to resort to numerical methods in order to find the modal solutions. It suffices to investigate the numerical solution of (2.10) for TE modes and (2.12) for TM modes.

2.2.4.1 *Boundary Element Method*

The **boundary element method** (BEM) is based on the discretization of an integral equation that is derived from original partial differential equation and defined on the boundary of the defining domain. The advantage of the BEM lies in the fact that only the boundary of the defining domain requires sub-division to produce a boundary mesh. Thus the dimension of the problem is reduced by one, which makes BEM much easier to handle and more computationally efficient than the domain methods.

Let ϕ be a scalar field representing either the longitudinal magnetic field h_z for TE mode or longitudinal electric field e_z for TM mode. Then we may write

$$(\nabla_t^2 + k_c^2)\phi(\boldsymbol{\rho}) = 0, \tag{2.48}$$

where k_c is the cut-off wavenumber. For TE modes, ϕ satisfies the Neumann boundary condition

$$\left.\frac{\partial \phi}{\partial n}\right|_\Gamma = 0. \tag{2.49}$$

For TM modes, ϕ satisfies the Dirichlet boundary condition

$$\phi|_{\Gamma} = 0. \tag{2.50}$$

The boundary integral equations for the waveguide problems may be easily derived by using the Green's identity

$$\int_{\Omega} (u\nabla^2 v - v\nabla^2 u)d\Omega = \int_{\Gamma} \left(u\frac{\partial v}{\partial n} - v\frac{\partial v}{\partial n}\right)d\Gamma \tag{2.51}$$

and the Green's function G defined by

$$(\nabla_t^2 + k_c^2)G(\boldsymbol{\rho}, \boldsymbol{\rho}') = -\delta(\boldsymbol{\rho} - \boldsymbol{\rho}'). \tag{2.52}$$

For TE modes, we may let $u = \phi$ and $v = G(\boldsymbol{\rho}, \boldsymbol{\rho}')$ in (2.51) and make use of (2.48) and (2.49) to obtain

$$C(\boldsymbol{\rho})\phi(\boldsymbol{\rho}) + \int_{\Gamma} \phi(\boldsymbol{\rho}')\frac{\partial G(\boldsymbol{\rho}, \boldsymbol{\rho}')}{\partial n(\boldsymbol{\rho}')}d\Gamma(\boldsymbol{\rho}') = 0, \tag{2.53}$$

where $C(\boldsymbol{\rho}) = \theta/2\pi$, and θ is the angle formed by the two half tangents at the boundary point $\boldsymbol{\rho}$, as illustrated in Figure 2.8. Equation (2.53) is the boundary integral equation for TE modes. For TM modes, we may let $u = \phi$ and $v = G(\boldsymbol{\rho}, \boldsymbol{\rho}')$ in (2.51) and make use of (2.48) and (2.50) to obtain

$$\int_{\Gamma} G(\boldsymbol{\rho}, \boldsymbol{\rho}')q(\boldsymbol{\rho}')d\Gamma(\boldsymbol{\rho}') = 0, \tag{2.54}$$

where $q = \frac{\partial \phi}{\partial n}$. Equation (2.54) is the boundary integral equation for TM modes.

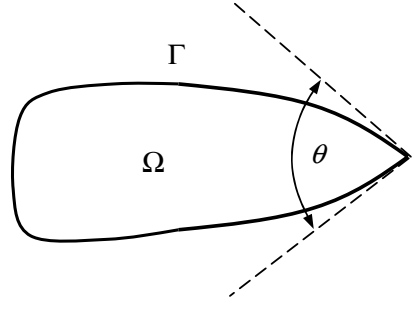


Figure 2.8 An arbitrary boundary point.

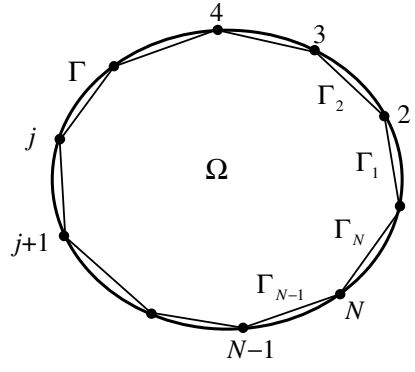


Figure 2.9 Boundary approximated by line elements.

To numerically solve the boundary integral equations (2.53) and (2.54), we first choose N nodes $\boldsymbol{\rho}^j = (x^j, y^j)(j = 1, 2, \ldots, N)$ on the boundary Γ. We then connect these nodes successively by linear segments (called **boundary elements**) as illustrated by Figure 2.9. Thus (2.53) and (2.54) can be approximated by

$$C(\boldsymbol{\rho})\phi(\boldsymbol{\rho}) + \sum_{j=1}^{N} \int_{\Gamma_j} \phi(\boldsymbol{\rho}') \frac{\partial G(\boldsymbol{\rho}, \boldsymbol{\rho}')}{\partial n(\boldsymbol{\rho}')} d\Gamma(\boldsymbol{\rho}') = 0, \tag{2.55}$$

$$\sum_{j=1}^{N} \int_{\Gamma_j} G(\boldsymbol{\rho}, \boldsymbol{\rho}') q(\boldsymbol{\rho}') d\Gamma(\boldsymbol{\rho}') = 0. \tag{2.56}$$

(1) Constant Element Equations

For the method of constant elements, the unknowns ϕ and q on element Γ_j are treated as a constant, respectively denoted by $\phi_c(j)$ and $q_c(j)$, $j = 1, 2, \ldots, N$. Equations (2.55) and (2.56) are then required to be exactly satisfied at the collocation points, selected as the middle points of the elements

$$\boldsymbol{\rho}_c^i = \frac{\boldsymbol{\rho}^i + \boldsymbol{\rho}^{i+1}}{2}, \quad i = 1, 2, \ldots, N, \tag{2.57}$$

where $\boldsymbol{\rho}^{N+1} = \boldsymbol{\rho}^1$. This process leads to the **constant element equations** for TE and TM modes

$$[M^{TE}(k_c)][\phi_c] = 0, \tag{2.58}$$

$$[M^{TM}(k_c)][q_c] = 0, \tag{2.59}$$

where

$$M_{ij}^{TE}(k_c) = \frac{1}{2}\delta_{ij} + \int_{\Gamma_j} \frac{\partial G(\boldsymbol{\rho}_c^i, \boldsymbol{\rho}')}{\partial n(\boldsymbol{\rho}')} d\Gamma(\boldsymbol{\rho}'), \quad i,j = 1,2,\ldots,N,$$
$$[\phi_c] = [\phi_c(1), \phi_c(2), \ldots, \phi_c(N)]^T. \tag{2.60}$$

$$M_{ij}^{TM}(k_c) = \int_{\Gamma_j} G(\boldsymbol{\rho}_c^i, \boldsymbol{\rho}') d\Gamma(\boldsymbol{\rho}'), \quad i,j = 1,2,\ldots,N,$$
$$[q_c] = [q_c(1), q_c(2), \ldots, q_c(N)]^T. \tag{2.61}$$

The necessary and sufficient conditions for the existence of a non-trivial solution of (2.58) and (2.59) are respectively given by

$$\det[M^{TE}(k_c)] = 0, \tag{2.62}$$

$$\det[M^{TM}(k_c)] = 0. \tag{2.63}$$

These are the generalized eigenvalue equations, from which the cut-off wavenumber k_c for TE and TM modes can be determined. The corresponding eigenvectors can be found from (2.58) and (2.59). The Green's function in (2.60) and (2.61) may be chosen as

$$G(\boldsymbol{\rho}, \boldsymbol{\rho}') = -\frac{1}{4} N_0(k_c|\boldsymbol{\rho} - \boldsymbol{\rho}'|) + g(\boldsymbol{\rho}, \boldsymbol{\rho}'), \tag{2.64}$$

where N_0 is the Neumann function, and $g(\boldsymbol{\rho}, \boldsymbol{\rho}')$ is an arbitrary solution of homogeneous Helmholtz equation

$$(\nabla_t^2 + k_c^2) g(\boldsymbol{\rho}, \boldsymbol{\rho}') = 0. \tag{2.65}$$

(2) Linear Element Equations

For the method of linear element, the unknowns ϕ and q on element Γ_j can be approximated by

$$\phi = \phi(j)(1-\lambda) + \phi(j+1)\lambda, \tag{2.66}$$

$$q = q(j)(1-\lambda) + q(j+1)\lambda, \tag{2.67}$$

where $\phi(j)$ and $q(j)$ denote the values of ϕ and q at node j ($j = 1,2,\ldots,N$). Substituting the linear interpolations (2.66) and (2.67) into (2.55) and (2.56) respectively and using the nodes as collocation points, we obtain the **linear element equations** for TE and TM modes

$$[M^{TE}(k_c)][\phi] = 0, \tag{2.68}$$

$$[M^{TM}(k_c)][q] = 0, \tag{2.69}$$

where

$$\begin{aligned} M_{ij}^{TE}(k_c) &= C(\boldsymbol{\rho}^i)\delta_{ij} + G_n(i,j) - \tilde{G}_n(i,j) + \tilde{G}_n(i,j-1), \\ [\phi] &= [\phi(1), \phi(2), \ldots, \phi(N)]^T, \\ G_n(i,j) &= \int_{\Gamma_j} \frac{\partial G(\boldsymbol{\rho}^i, \boldsymbol{\rho}')}{\partial n(\boldsymbol{\rho}')} d\Gamma(\boldsymbol{\rho}'), \\ \tilde{G}_n(i,j) &= \int_{\Gamma_j} \lambda \frac{\partial G(\boldsymbol{\rho}^i, \boldsymbol{\rho}')}{\partial n(\boldsymbol{\rho}')} d\Gamma(\boldsymbol{\rho}'), \quad i,j = 1,2,\ldots,N. \end{aligned} \tag{2.70}$$

$$\begin{aligned} M_{ij}^{TM}(k_c) &= G(i,j) - \tilde{G}(i,j) + \tilde{G}(i,j-1), \\ [q] &= [q(1), q(2), \ldots, q(N)]^T, \\ G(i,j) &= \int_{\Gamma_j} G(\boldsymbol{\rho}^i, \boldsymbol{\rho}') d\Gamma(\boldsymbol{\rho}'), \\ \tilde{G}(i,j) &= \int_{\Gamma_j} \lambda G(\boldsymbol{\rho}^i, \boldsymbol{\rho}') d\Gamma(\boldsymbol{\rho}'), \quad i,j = 1,2,\ldots,N. \end{aligned} \tag{2.71}$$

Remark 2.3 (Spurious solutions)**:** When a differential equation is transformed into an integral equation, the requirement of smoothness of the unknown functions is relaxed and this raises the question of whether the integral equation is equivalent to the original differential equation. In fact, the spurious solutions may occur in the integral equation formulation. It can be shown that the spurious solutions may be avoided if the Green's function (2.64) satisfies the radiation condition. If the Green's function does not satisfy the radiation condition, the spurious solutions will occur and they are eigenvalues of exterior Dirichlet problem. Based on this property, a criterion for discriminating the spurious solutions may be developed for a waveguide with edges (Geyi, 1990a; 2010). □

2.2.4.2 *Finite Difference Method*

The **finite difference method** (FDM) is a domain method for solving differential equations by using finite difference to approximate derivatives. For generality, the waveguide cross section Ω is discretized into a number of polygonal elements with N nodes $\{n_1, n_2, \ldots, n_N\}$, as illustrated in Figure 2.10.

Let $n_i (i = 1,2,\ldots,N)$ be an arbitrary interior node, which is assumed to have e_i neighboring nodes $m_j (j = 1,2,\ldots,e_i)$. Let q_j be the middle point

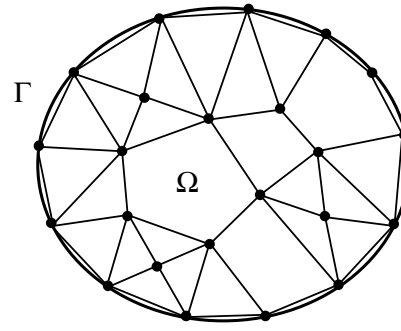


Figure 2.10 Polygonal discretization.

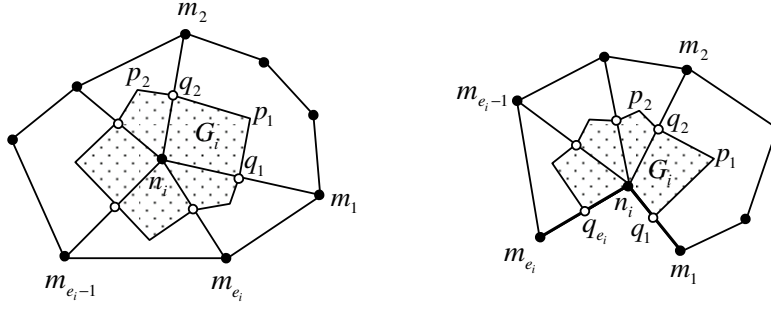


Figure 2.11 (a) Dual element for interior node. (b) Dual element for boundary node.

of line segment $\overline{n_i m_j}$. In each polygon using n_i as a vertex, we choose a point $p_j (j = 1, 2, \ldots, e_i)$ and then connect points $q_1, p_1, q_2, p_2, \ldots$ successively, obtaining a polygonal region G_i [Figure 2.11(a)]. This is referred to as the **dual element** of n_i. Taking the integral of (2.48) over the dual element G_i, we have

$$\int_{G_i} (\nabla_t^2 + k_c^2)\phi \, d\Omega = 0. \tag{2.72}$$

Utilizing the Green's identity (2.51), (2.72) can be transformed into the following form

$$\int_{\partial G_i} \frac{\partial \phi}{\partial n} d\Gamma + \int_{G_i} k_c^2 \phi \, d\Omega = 0. \tag{2.73}$$

where ∂G_i denotes the boundary of the dual element G_i. The line integral along ∂G_i can be expressed as the sum of the integrals along the broken

segments $p_1q_2p_2, \ldots, p_{e_i}q_1p_1$; then we have

$$\sum_{j=1}^{e_i} \int_{\overline{p_{j-1}q_jp_j}} \frac{\partial \phi}{\partial n} d\Gamma + \int_{G_i} k_c^2 \phi \, d\Omega = 0. \tag{2.74}$$

Making use of the following approximation

$$\int_{\overline{p_{j-1}q_jp_j}} \frac{\partial \phi}{\partial n} d\Gamma = -[\phi(n_i) - \phi(m_j)] \frac{\overline{p_{j-1}q_j} + \overline{q_jp_j}}{\overline{m_jn_i}},$$

$$\int_{G_i} k_c^2 \phi \, d\Omega = k_c^2 \phi(n_i) S_i,$$

where S_i represents the area of G_i and $\phi(n_i)$ denotes the value of ϕ at node n_i, (2.74) can be written as

$$\sum_{j=1}^{e_i} [\phi(n_i) - \phi(m_j)] \frac{\overline{p_{j-1}q_j} + \overline{q_jp_j}}{\overline{m_jn_i}} - k_c^2 \phi(n_i) S_i = 0. \tag{2.75}$$

Here we have used the convention $p_0 = p_{e_i}$. Equation (2.75) is the node equation for an interior node.

We now construct the node equation for a boundary node. For the Dirichlet boundary condition, the node equation for a boundary node n_i is trivial and $\phi(n_i) = 0$. For the Neumann boundary condition, we may introduce the dual element G_i for a boundary node n_i as illustrated in Figure 2.11(b), where the node n_i itself is also a vortex of the polygonal region G_i. Taking the integration of (2.48) over the region G_i and using the Green's identity, we obtain

$$\int_{\overline{q_1p_1}} \frac{\partial \phi}{\partial n} d\Gamma + \int_{\overline{p_1q_2p_2}} \frac{\partial \phi}{\partial n} d\Gamma + \cdots + \int_{\overline{p_{e_i-1}q_{e_i}}} \frac{\partial \phi}{\partial n} d\Gamma$$
$$+ \int_{G_i} k_c^2 \phi \, d\Omega + \int_{\overline{n_iq_1}} \frac{\partial \phi}{\partial n} d\Gamma + \int_{\overline{n_iq_{e_i}}} \frac{\partial \phi}{\partial n} d\Gamma = 0.$$

On account of the Neumann boundary condition, the last two terms of the above expression vanish. Hence

$$\sum_{j=2}^{e_i-1} [\phi(n_i) - \phi(m_j)] \frac{\overline{p_{j-1}q_j} + \overline{q_jp_j}}{\overline{m_jn_i}} + [\phi(n_i) - \phi(m_1)] \frac{\overline{p_1q_1}}{\overline{m_1n_i}}$$
$$+ [\phi(n_i) - \phi(m_{e_i})] \frac{\overline{p_{e_i-1}q_{e_i}}}{\overline{m_{e_i}n_i}} - k_c^2 \phi(n_i) S_i = 0. \tag{2.76}$$

This is the node equation for the boundary node n_i. Combining the node equations for the interior nodes and boundary nodes, we obtain the standard algebraic eigenvalue equation

$$[A][\phi] - k_c^2[B][\phi] = 0. \tag{2.77}$$

Here $[A]$ is a band matrix and $[B]$ is a diagonal matrix, and $[\phi] = [\phi(n_1), \phi(n_2), \ldots, \phi(n_N)]^T$. Equation (2.77) can be used to determine the cut-off wavenumbers.

Remark 2.4: Based on the polygon discretization, a network model for the waveguide problem may be established to transform the original boundary value problem into a circuit problem (Geyi, 1990b). □

2.2.4.3 *Finite Element Method*

The **finite element method** (FEM) uses the variational method to minimize a functional derived from a partial differential equation, producing a numerical solution for the partial differential equation. It has been shown that the solution of (2.48) is equivalent to minimizing the following functional (see Section 1.4)

$$I(\phi) = \int_{\Omega} (\nabla_t \phi \cdot \nabla_t \phi - k_c^2 \phi^2) d\Omega. \tag{2.78}$$

The cross section Ω of the waveguide may be discretized into N_e triangular elements Ω_e $(e = 1, 2, \ldots, N_e)$ with N_n nodes, as illustrated in Figure 2.12. Thus

$$I(\phi) = \sum_{e=1}^{N_e} I(\phi_e), \tag{2.79}$$

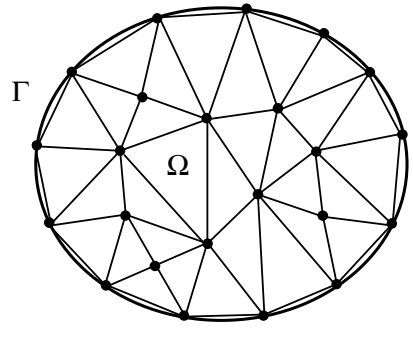


Figure 2.12 Triangular element discretization.

where ϕ_e denotes the unknown scalar field ϕ over a triangular element Ω_e, and can be expressed in terms of the linear shape functions as

$$\phi_e = \sum_{m=1}^{3} \alpha_m(x,y)\phi_e(m),$$

with

$$\alpha_1(x,y) = \frac{x_2y_3 - x_3y_2 + x(y_2-y_3) - y(x_2-x_3)}{(x_1-x_3)(y_2-y_3) - (x_2-x_3)(y_1-y_3)}$$

$$\alpha_2(x,y) = \frac{x_3y_1 - x_1y_3 + x(y_3-y_1) - y(x_3-x_1)}{(x_1-x_3)(y_2-y_3) - (x_2-x_3)(y_1-y_3)}$$

$$\alpha_3(x,y) = \frac{x_1y_2 - x_2y_1 + x(y_2-y_1) - y(x_2-x_1)}{(x_1-x_3)(y_2-y_3) - (x_2-x_3)(y_1-y_3)}.$$

Note that

$$\begin{aligned} I(\phi_e) &= \int_{\Omega_e} (\nabla_t\phi_e \cdot \nabla_t\phi_e - k_c^2\phi_e^2)d\Omega \\ &= \sum_{m=1}^{3}\sum_{n=1}^{3} \phi_e(m)\phi_e(n) \int_{\Omega_e} \nabla_t\alpha_m(x,y) \cdot \nabla_t\alpha_n(x,y)d\Omega \\ &\quad - k_c^2 \sum_{m=1}^{3}\sum_{n=1}^{3} \phi_e(m)\phi_e(n) \int_{\Omega_e} \alpha_m(x,y) \cdot \alpha_n(x,y)d\Omega. \end{aligned}$$

This can be written in matrix form as

$$I(\phi_e) = [\phi_e]^T[C^e][\phi_e] - k_c^2[\phi_e]^T[D^e][\phi_e], \tag{2.80}$$

where

$$\begin{aligned} [\phi_e] &= [\phi_e(1), \phi_e(2), \phi_e(3)]^T, \\ C_{mn}^e &= \int_{\Omega_e} \nabla_t\alpha_m(x,y) \cdot \nabla_t\alpha_n(x,y)d\Omega, \\ D_{mn}^e &= \int_{\Omega_e} \alpha_m(x,y) \cdot \alpha_n(x,y)d\Omega \end{aligned} \tag{2.81}$$

are local matrices. Substituting (2.80) into (2.79), we obtain

$$I(\phi) = [\phi]^T[C][\phi] - k_c^2[\phi]^T[D][\phi], \tag{2.82}$$

where $[\phi] = [\phi(1), \phi(2), \ldots, \phi(N_n)]^T$, $[C]$ and $[D]$ are global matrices formed by assembling the local matrices $[C^e]$ and $[D^e]$.

The nodes can be divided into free nodes where the values of the scalar field ϕ are yet to be determined and fixed nodes where the values of the scalar field ϕ are known. If we number the free nodes first and the fixed nodes last, (2.82) can be written as

$$I(\phi) = \begin{bmatrix}\phi_f & \phi_g\end{bmatrix} \begin{bmatrix}C_{ff} & C_{fg} \\ C_{gf} & C_{gg}\end{bmatrix} \begin{bmatrix}\phi_f \\ \phi_g\end{bmatrix} - k_c^2 \begin{bmatrix}\phi_f & \phi_g\end{bmatrix} \begin{bmatrix}D_{ff} & D_{fg} \\ D_{gf} & D_{gg}\end{bmatrix} \begin{bmatrix}\phi_f \\ \phi_g\end{bmatrix} \tag{2.83}$$

where subscripts f and g denote the free nodes and fixed nodes respectively. Letting $\frac{\delta I}{\delta \phi_f} = 0$ yields

$$\begin{bmatrix}C_{ff} & C_{fg}\end{bmatrix} \begin{bmatrix}\phi_f \\ \phi_g\end{bmatrix} - k_c^2 \begin{bmatrix}D_{ff} & D_{fg}\end{bmatrix} \begin{bmatrix}\phi_f \\ \phi_g\end{bmatrix} = 0. \tag{2.84}$$

For TM modes, we have $\phi_g = 0$. For TE modes, all nodes are free. So we have

$$[C_{ff}][\phi_f] - k_c^2 [D_{ff}][\phi_f] = 0. \tag{2.85}$$

This equation can be used to determine the cut-off wavenumbers and the corresponding modal fields.

2.3 Inhomogeneous Metal Waveguides

Inhomogeneously filled waveguides, such as a rectangular waveguide partially filled with dielectric slabs, are used in a number of waveguide components. The determination of the propagation constants of the modes in the waveguides is the major focus of our interest.

2.3.1 General Field Relationships

Consider a metal waveguide, which is uniform along z-axis. The cross section of the waveguide is denoted by Ω and its boundary is assumed to be a perfect conductor and is denoted by $\Gamma = \Gamma_1 + \Gamma_2$, as shown in Figure 2.13. The waveguide is filled with inhomogeneous medium in which μ and ε are functions of transverse positions but are constant along z-axis. Assume that the fields in the waveguide have a z-dependence

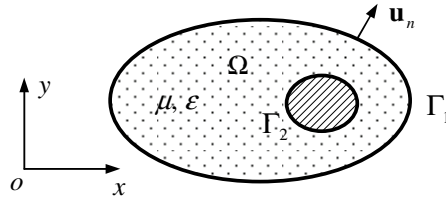


Figure 2.13 Inhomogeneous waveguide.

of the form $e^{-j\beta z}$

$$\mathbf{E}(\mathbf{r}) = \mathbf{e}(\boldsymbol{\rho})e^{-j\beta z}, \quad \mathbf{H}(\mathbf{r}) = \mathbf{h}(\boldsymbol{\rho})e^{-j\beta z}, \tag{2.86}$$

where $\boldsymbol{\rho} = (x, y) \in \Omega$ denotes the transverse position. Introducing these into Maxwell equations, we obtain

$$\begin{aligned} \nabla_\beta \times \mathbf{h} = j\omega\varepsilon\mathbf{e}, \quad \nabla_\beta \times \mathbf{e} = -j\omega\mu\mathbf{h}, \\ \nabla_\beta \cdot \varepsilon\mathbf{e} = 0, \quad \nabla_\beta \cdot \mu\mathbf{h} = 0. \end{aligned} \tag{2.87}$$

Here $\nabla_\beta = \nabla_t - j\beta\mathbf{u}_z$ denotes an operator obtained from ∇ by replacing the derivative with respect to z with multiplication by $-j\beta$, and ∇_t is transverse gradient operator. For an arbitrary vector function $\mathbf{f}(\boldsymbol{\rho})$ and a scalar function $u(\boldsymbol{\rho})$, we have

$$\begin{aligned} \nabla_\beta \cdot (\nabla_\beta \times \mathbf{f}) &= 0, \\ \nabla_\beta \cdot (u\mathbf{f}) &= u\nabla_\beta \cdot \mathbf{f} + \mathbf{f} \cdot \nabla_t u, \\ \nabla_\beta \cdot (\nabla_\beta u) &= \nabla_t^2 u - \beta^2 u, \\ \nabla_\beta \times (\nabla_\beta u) &= 0, \\ \nabla_\beta \times \nabla_\beta \times \mathbf{f} &= -\nabla_t^2 \mathbf{f} + \beta^2 \mathbf{f} + \nabla_\beta(\nabla_\beta \cdot \mathbf{f}). \end{aligned} \tag{2.88}$$

It follows from (2.87) that

$$\begin{aligned} \nabla_\beta \times \varepsilon_r^{-1} \nabla_\beta \times \mathbf{h} = k_0^2 \mu_r \mathbf{h}, \\ \nabla_\beta \times \mu_r^{-1} \nabla_\beta \times \mathbf{e} = k_0^2 \varepsilon_r \mathbf{e}, \\ \nabla_\beta \cdot \varepsilon_r \mathbf{e} = 0, \quad \nabla_\beta \cdot \mu_r \mathbf{h} = 0, \end{aligned} \tag{2.89}$$

where $\mu_r = \mu/\mu_0$, $\varepsilon_r = \varepsilon/\varepsilon_0$ and $k_0 = \omega\sqrt{\mu_0\varepsilon_0}$. A solution of (2.89) is called a **guided mode** of the waveguide if the field is non-trivial and has finite energy:

$$(\beta, k_0) \in R^2, \ \ (\mathbf{e}, \mathbf{h}) \neq 0, \quad \text{and} \quad \int\limits_{R^2} |\mathbf{e}|^2 d\Omega < \infty, \ \int\limits_{R^2} |\mathbf{h}|^2 d\Omega < \infty.$$

2.3.2 Symmetric Formulation

It follows from (2.89) that the fields satisfy

$$\begin{aligned} &\nabla_\beta \times \varepsilon_r^{-1} \nabla_\beta \times \mathbf{h} = k_0^2 \mu_r \mathbf{h}, \quad \boldsymbol{\rho} \in \Omega, \\ &\mathbf{u}_n \cdot \mu_r \mathbf{h} = 0, \quad \mathbf{u}_n \times \varepsilon_r^{-1} \nabla_\beta \times \mathbf{h} = 0, \quad \boldsymbol{\rho} \in \Gamma, \end{aligned} \tag{2.90}$$

and

$$\begin{aligned} &\nabla_\beta \times \mu_r^{-1} \nabla_\beta \times \mathbf{e} = k_0^2 \varepsilon_r \mathbf{e}, \quad \boldsymbol{\rho} \in \Omega, \\ &\mathbf{u}_n \times \mathbf{e} = 0, \quad \mathbf{u}_n \cdot \mu_r^{-1} \nabla_\beta \times \mathbf{e} = 0, \quad \boldsymbol{\rho} \in \Gamma, \end{aligned} \tag{2.91}$$

where $\mathbf{u}_n$ is the unit outward normal on Γ. In (2.90) and (2.91), the propagation constant is considered as a parameter while the wavenumber k_0 is taken as the eigenvalue that is a function of β. For a given β, both (2.90) and (2.91) define a symmetric eigenvalue problem respectively. From (2.90) we obtain

$$\begin{aligned} k_0^2 \int_\Omega \mu_r |\mathbf{h}|^2 d\Omega &= \int_\Omega \frac{1}{n^2} (\nabla_\beta \times \mathbf{h}) \cdot (\overline{\nabla_\beta \times \mathbf{h}}) d\Omega \\ &\geq \frac{1}{n_+^2} \int_\Omega (\nabla_\beta \times \mathbf{h}) \cdot (\overline{\nabla_\beta \times \mathbf{h}}) d\Omega \\ &= \frac{1}{n_+^2} \int_\Omega (\nabla_\beta \times \nabla_\beta \times \mathbf{h}) \cdot \bar{\mathbf{h}} d\Omega, \end{aligned}$$

where $n = \sqrt{\varepsilon_r}$ and $n_+ = \max_{\boldsymbol{\rho} \in \Omega} n(\boldsymbol{\rho})$. Making use of the last equation of (2.88) and integration by parts, we have

$$\begin{aligned} k_0^2 \int_\Omega \mu_r |\mathbf{h}|^2 \, d\Omega &\geq \frac{1}{n_+^2} \int_\Omega (-\nabla_t^2 \mathbf{h} + \beta^2 \mathbf{h} + \nabla_\beta \nabla_\beta \cdot \mathbf{h}) \cdot \bar{\mathbf{h}} \, d\Omega \\ &= \frac{1}{n_+^2} \int_\Omega (|\nabla_t \times \mathbf{h}|^2 + |\nabla_t \cdot \mathbf{h}|^2 - |\nabla_\beta \cdot \mathbf{h}|^2) d\Omega + \frac{\beta^2}{n_+^2} \int_\Omega |\mathbf{h}|^2 \, d\Omega. \end{aligned}$$

If μ_r is a constant, we have $\nabla_\beta \cdot \mathbf{h} = 0$ and the above is equivalent to

$$\int_\Omega (|\nabla_t \times \mathbf{h}|^2 + |\nabla_t \cdot \mathbf{h}|^2) d\Omega + (\beta^2 - k_0^2 \mu_r n_+^2) \int_\Omega |\mathbf{h}|^2 \, d\Omega \leq 0. \tag{2.92}$$

As a result, if $|\beta| \geq k_0 \sqrt{\mu_r} n_+$, then $\mathbf{h} = 0$ and (2.90) has a trivial solution. In other words, no guided modes exist in this case. Hence the solution (β, k_0) of (2.90) or (2.91) must satisfy

$$k_0 > \frac{|\beta|}{\sqrt{\mu_r} n_+}. \tag{2.93}$$

This is the **guidance condition** for an inhomogeneously filled waveguide.

2.3.3 Asymmetric Formulation

In engineering, the propagation constant β is usually considered as the eigenvalue while the frequency or the wavenumber k_0 is taken as a parameter. This arrangement often yields a non-symmetric eigenvalue problem, and is more difficult to study. The guided modes in the waveguide may be decomposed into a transverse and a longitudinal component

$$\mathbf{E}(\mathbf{r}) = [\mathbf{e}(\boldsymbol{\rho}) + \mathbf{u}_z e_z(\boldsymbol{\rho})]e^{-j\beta z}, \quad \mathbf{H}(\mathbf{r}) = [\mathbf{h}(\boldsymbol{\rho}) + \mathbf{u}_z h_z(\boldsymbol{\rho})]e^{-j\beta z}. \tag{2.94}$$

Introducing these into Maxwell equations, we obtain

$$\begin{aligned} &\nabla \times \mathbf{h} = j\omega\varepsilon \mathbf{u}_z e_z, \quad \nabla \times \mathbf{e} = -j\omega\mu \mathbf{u}_z h_z, \\ &j\beta \mathbf{u}_z \times \mathbf{h} + \mathbf{u}_z \times \nabla h_z = -j\omega\varepsilon \mathbf{e}, \\ &j\beta \mathbf{u}_z \times \mathbf{e} + \mathbf{u}_z \times \nabla e_z = j\omega\mu \mathbf{h}, \\ &\nabla \cdot \varepsilon \mathbf{e} = j\beta\varepsilon e_z, \quad \nabla \cdot \mu \mathbf{h} = j\beta\mu h_z. \end{aligned} \tag{2.95}$$

By eliminating $\mathbf{h}$, e_z and h_z, we have the following eigenvalue problem

$$\begin{aligned} \mu\nabla \times \mu^{-1}\nabla \times \mathbf{e} - \nabla\varepsilon^{-1}\nabla \cdot \varepsilon\mathbf{e} - (\omega^2\mu\varepsilon - \beta^2)\mathbf{e} = 0, \quad \boldsymbol{\rho} \in \Omega, \\ \mathbf{u}_n \times \mathbf{e} = 0, \quad \nabla \cdot \varepsilon\mathbf{e} = 0, \quad \boldsymbol{\rho} \in \Gamma. \end{aligned} \tag{2.96}$$

In (2.96), β^2 is taken as the eigenvalue and ω^2 as the parameter. The differential operator in (2.96) is not symmetric. Let $\mathbf{e}_m$ and $\mathbf{e}_n$ be two different eigenfunctions corresponding to the eigenvalues β_m^2 and β_n^2 respectively. Then

$$\mu\nabla \times \mu^{-1}\nabla \times \mathbf{e}_m - \nabla\varepsilon^{-1}\nabla \cdot \varepsilon\mathbf{e}_m - (\omega^2\mu\varepsilon - \beta_m^2)\mathbf{e}_m = 0.$$

Taking the scalar product of the above equation with $\nabla \times \mu^{-1}\nabla \times \mathbf{e}_n - \omega^2\varepsilon\mathbf{e}_n$ and integrating the resultant equation over Ω yield

$$\begin{aligned} &\int_\Omega \mu(\nabla \times \mu^{-1}\nabla \times \mathbf{e}_m - \omega^2\varepsilon\mathbf{e}_m) \cdot (\nabla \times \mu^{-1}\nabla \times \mathbf{e}_n - \omega^2\varepsilon\mathbf{e}_n)d\Omega \\ &\quad - \int_\Omega \omega^2\varepsilon^{-1}(\nabla \cdot \varepsilon\mathbf{e}_m)(\nabla \cdot \varepsilon\mathbf{e}_n)d\Omega \\ &\quad + \beta_m^2 \int_\Omega (\mu^{-1}\nabla \times \mathbf{e}_m \cdot \nabla \times \mathbf{e}_n - \omega^2\varepsilon\mathbf{e}_m \cdot \mathbf{e}_n)d\Omega = 0. \end{aligned} \tag{2.97}$$

Interchanging m and n and subtracting the result from (2.97) gives

$$(\beta_m^2 - \beta_n^2) \int_\Omega (\mu^{-1}\nabla \times \mathbf{e}_m \cdot \nabla \times \mathbf{e}_n - \omega^2\varepsilon\mathbf{e}_m \cdot \mathbf{e}_n)d\Omega = 0. \tag{2.98}$$

This implies the following orthogonality relation

$$\int_{\Omega} (\mu^{-1}\nabla \times \mathbf{e}_m \cdot \nabla \times \mathbf{e}_n - \omega^2 \varepsilon \mathbf{e}_m \cdot \mathbf{e}_n) d\Omega = 0, \tag{2.99}$$

if $\beta_m^2 \neq \beta_n^2$. From (2.95), the transverse magnetic field can be expressed in terms of the transverse electric field

$$\mathbf{u}_z \times \mathbf{h} = \frac{1}{\omega\beta}(\nabla \times \mu^{-1}\nabla \times \mathbf{e} - \omega^2 \varepsilon \mathbf{e}),$$

and the orthogonality relation (2.99) can be written as

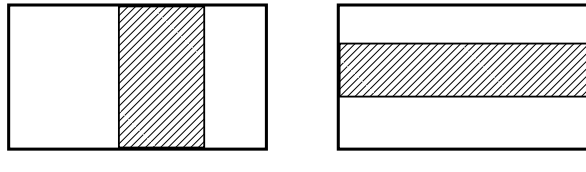
$$\int_{\Omega} (\mathbf{e}_m \times \mathbf{h}_n) \cdot \mathbf{u}_z d\Omega = 0, \quad m \neq n. \tag{2.100}$$

This is the most general form of the orthogonality relation in a waveguide. The modes in a waveguide filled with homogeneous medium can be classified into TEM, TE and TM modes. In an inhomogeneous waveguide, such classification is impossible since the modes contain both e_z and h_z components.

2.3.4 Dielectric-Slab-Loaded Rectangular Waveguides

Dielectric-slab-loaded rectangular waveguides (Figure 2.14) have significant advantage in bandwidth and power handling capacity over unloaded rectangular waveguide. By proper choice of dimensions and dielectrics, the bandwidth and power handling capacity can be significantly increased (Vartanian *et al.*, 1958). The modes in the dielectric loaded waveguide are not either TE or TM modes, but combinations of both, called **longitudinal section electric** (LSE) modes or **longitudinal section magnetic** (LSM) modes.

Consider an asymmetrically loaded waveguide as illustrated in Figure 2.15. The dielectric slab of thickness t is along the sidewall of the waveguide. The refractive index of the material filled in the waveguide is



(a) (b)

Figure 2.14 Dielectric-slab-loaded rectangular waveguides.

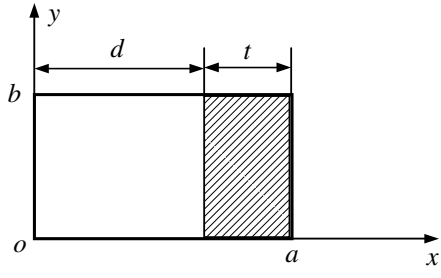


Figure 2.15 Asymmetrically loaded waveguide.

assumed to be $n(\mathbf{r})$. The LSE modes can be derived from a magnetic Hertz vector

$$\mathbf{\Pi}_m = \mathbf{u}_x \Pi_m(x, y) e^{-\gamma z}.$$

The electromagnetic fields are given by

$$\begin{aligned} \mathbf{E} &= -j\omega\mu_0 \nabla \times \mathbf{\Pi}_m, \\ \mathbf{H} &= \nabla \times \nabla \times \mathbf{\Pi}_m = n^2 k_0^2 \mathbf{\Pi}_m + \nabla\nabla \cdot \mathbf{\Pi}_m, \end{aligned} \tag{2.101}$$

where $k_0 = \omega\sqrt{\mu_0 \varepsilon_0}$. The potential function $\Pi_m(x, y)$ satisfies

$$\nabla_t^2 \Pi_m + (n^2 k_0^2 + \gamma^2)\Pi_m = 0, \tag{2.102}$$

where

$$n(x) = \begin{cases} 1, & 0 < x < d \\ \sqrt{\varepsilon_r}, & t < x < a \end{cases}.$$

To satisfy the boundary conditions on the conducting walls, we have

$$\Pi_m = \begin{cases} A \sin k_{x1} x \cos \dfrac{n\pi}{b} y, & 0 \le x \le d \\ A \sin k_{x2}(a - x) \cos \dfrac{n\pi}{b} y, & d \le x \le a \end{cases},$$

with

$$\gamma^2 = k_{x1}^2 + \left(\frac{m\pi}{b}\right)^2 - k_0^2 = k_{x2}^2 + \left(\frac{m\pi}{b}\right)^2 - n^2 k_0^2. \tag{2.103}$$

It follows from (2.101) that

$$E_z = j\omega\mu_0 e^{-\gamma z} \frac{\partial \Pi_m}{\partial y}, \quad H_y = e^{-\gamma z} \frac{\partial^2 \Pi_m}{\partial y \partial x}.$$

These field components must be continuous at the air–dielectric interface. So we have

$$\begin{aligned} A \sin k_{x1} d &= B \sin k_{x2} t, \\ A k_{x1} \cos k_{x1} d &= -B k_{x2} \cos k_{x2} t, \end{aligned} \tag{2.104}$$

which yields

$$k_{x2} \tan k_{x1} d = -k_{x1} \tan k_{x2} t. \tag{2.105}$$

The above equation together with (2.103) can be used to determine the wavenumbers k_{x1} and k_{x2}.

2.4 Waveguide Discontinuities

In practice, a uniform waveguide is often discontinued by components and junctions. The waveguide discontinuities are used as various passive components and their introduction will distort the fields in the original uniform waveguide. One of the important tasks of microwave field theory is to establish the circuit parameters or network parameters for various waveguide discontinuities. In most applications, the waveguide supports a single dominant propagating mode. When a discontinuity exists, such as discontinuity in cross-sectional shape or an obstacle in the waveguide, an infinite number of non-propagating modes will be excited in the vicinity of the discontinuity by the incident dominant propagating mode.

2.4.1 Network Representation of Waveguide Discontinuities

A typical n-port waveguide discontinuity is shown in Figure 2.16(a), which consists of n uniform waveguides and a discontinuity (a junction). The reference planes $T_1, T_2, \ldots,$ and T_n are assumed to be far away from the discontinuity so that only the dominant modes exist at the reference planes.

The modal voltage V and the modal current I at a reference plane are proportional to the transverse electric field and transverse magnetic field in the waveguide respectively. The uniqueness theorem indicates that the modal voltages at the reference planes $V_1, V_2, \ldots, V_n$ can be determined by the modal currents $I_1, I_2, \ldots, I_n$ at the reference planes. If the medium is linear, the modal voltages and currents are linearly related. So we may write

$$[V] = [Z][I], \tag{2.106}$$

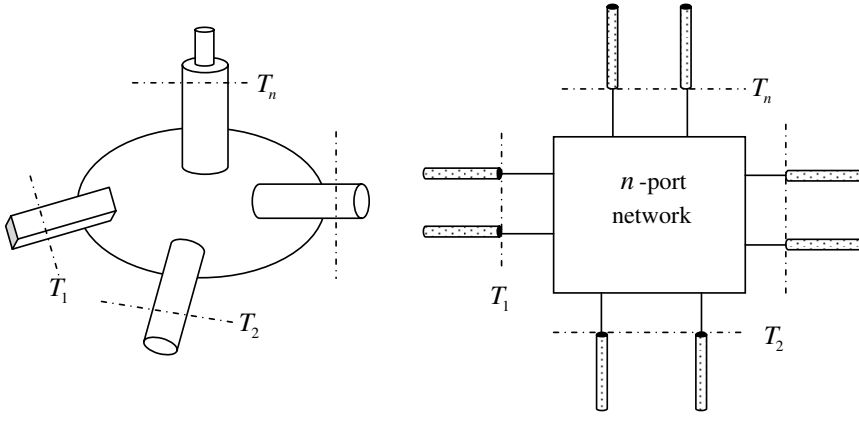


Figure 2.16 (a) Waveguide junction. (b) Equivalent circuit.

where

$$[V] = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{bmatrix}, \quad [I] = \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_n \end{bmatrix}, \quad [Z] = \begin{bmatrix} Z_{11} & Z_{12} & \cdots & Z_{1n} \\ Z_{21} & Z_{22} & \cdots & Z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n1} & Z_{n2} & \cdots & Z_{nn} \end{bmatrix},$$

and $Z_{ij}(i, j = 1, 2, \ldots, n)$ are called **impedance parameters**. It follows from (2.106) that

$$Z_{ii} = \left.\frac{V_i}{I_i}\right|_{I_l=0, l\neq i}, \quad \mathrm{Z}_{ij} = \left.\frac{V_i}{I_j}\right|_{I_l=0, l\neq j}.$$

Hence the impedance parameters are also called **open circuit parameters**. If the net power delivered into the network, denoted as P, is zero:

$$P = \frac{1}{4}[I]^T[Z^T + \bar{Z}][\bar{I}] = 0,$$

the network is lossless and satisfies the lossless condition

$$[Z^T + \bar{Z}] = 0. \tag{2.107}$$

To determine the network parameters, the field distribution in the waveguide junction must be known. There are a number of analytical methods, which can be applied to solve the waveguide junction problems (e.g., Collin, 1991; Schwinger and Saxon, 1968; Lewin, 1951). The variational method is

the most commonly used analytic technique that can handle a large variety of discontinuity problems.

2.4.2 Diaphragms in Waveguide-Variational Method

Figure 2.17 shows a general diaphragm in a uniform waveguide. The shadowed region denotes the conducting diaphragm, which is perpendicular to the waveguide axis and is located at $z = 0$. This structure covers the common inductive and capacitive windows, and their combination, the resonant window. Suppose the waveguide only supports the dominant mode and the waveguide extends to infinity in $\pm z$ directions. The dominant mode of unit amplitude is incident upon the diaphragm from $z = -\infty$, which excites a number of higher order modes in the neighborhood of the diaphragm. In the region $z < 0$, the transverse electromagnetic fields can be expanded as follows

$$\mathbf{E}_t^- = (e^{-j\beta_1 z} + \Gamma e^{j\beta_1 z})\mathbf{e}_1 + \sum_{n=2}^{\infty} V_n e^{j\beta_n z}\mathbf{e}_n,$$

$$\mathbf{H}_t^- = (e^{-j\beta_1 z} - \Gamma e^{j\beta_1 z})Z_{w1}^{-1}\mathbf{u}_z \times \mathbf{e}_1 - \sum_{n=2}^{\infty} V_n Z_{wn}^{-1} e^{j\beta_n z}\mathbf{u}_z \times \mathbf{e}_n,$$

where V_n are the modal voltages; Γ is the refection coefficient of the dominant mode at $z = 0$; β_n and Z_{wn} are given by (2.15). Similarly, the

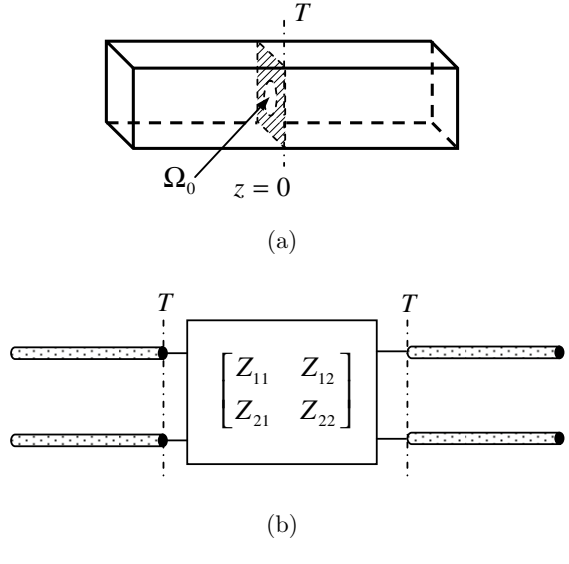


Figure 2.17 (a) A diaphragm. (b) Equivalent circuit.

fields in the region $z > 0$ can be expanded as

$$\mathbf{E}_t^+ = V_1' e^{-j\beta_1' z}\mathbf{e}_1 + \sum_{n=2}^{\infty} \mathbf{e}_n V_n' e^{-j\beta_n' z},$$

$$\mathbf{H}_t^+ = V_1' e^{-j\beta_1' z} Z_{w1}^{-1}\mathbf{u}_z \times \mathbf{e}_1 + \sum_{n=2}^{\infty} \mathbf{u}_z \times \mathbf{e}_n V_n' Z_{wn}^{-1} e^{-j\beta_n' z}.$$

The continuity of the tangent electric field at $z = 0$ gives

$$1 + \Gamma = V_1' = \int_{\Omega_0} \mathbf{E}_t(0) \cdot \mathbf{e}_1 \, d\Omega, \qquad V_n = V_n' = \int_{\Omega_0} \mathbf{E}_t(0) \cdot \mathbf{e}_n \, d\Omega, \quad (n \geq 2), \tag{2.108}$$

where Ω_0 denotes the aperture at $z = 0$. Considering the symmetry property of the structure and that the tangential electric field must be continuous at $z = 0$, we have $Z_{11} = Z_{22}$ and the equivalent circuit shown in Figure 2.17(b) can be simplified to a T-type circuit shown in Figure 2.18(a). The first expression of (2.108) indicates that the two terminal voltages of

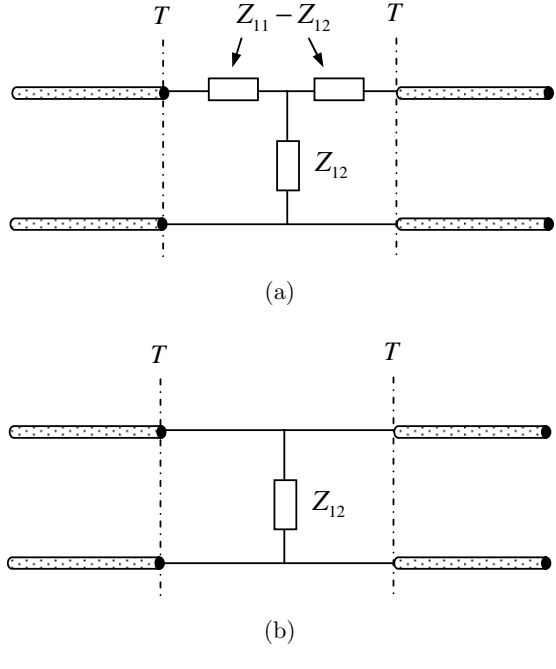


Figure 2.18 Equivalent circuit for the diaphragm.

the equivalent circuit are equal, which implies $Z_{11} = Z_{12}$, and the final equivalent circuit is shown in Figure 2.18(b). Note that the tangential magnetic field must also be continuous at the aperture

$$(1-\Gamma)Z_{w1}^{-1}\mathbf{u}_z \times \mathbf{e}_1 - \sum_{n=2}^{\infty} \mathbf{u}_z \times \mathbf{e}_n V_n Z_{wn}^{-1}$$

$$= V_1' Z_{w1}^{-1}\mathbf{u}_z \times \mathbf{e}_1 + \sum_{n=2}^{\infty} \mathbf{u}_z \times \mathbf{e}_n V_n' Z_{wn}^{-1}.$$

Substituting (2.108) into the above equation, we obtain the following integral equation

$$\mathbf{e}_1 = \mathbf{e}_1 \int_{\Omega_0} \mathbf{E}_t(0)\cdot\mathbf{e}_1\, d\Omega + \sum_{n=2}^{\infty} \mathbf{e}_n Z_{w1} Z_{wn}^{-1} \int_{\Omega_0} \mathbf{E}_t(0)\cdot\mathbf{e}_n\, d\Omega, \quad \text{in } \Omega_0. \tag{2.109}$$

This equation can be used to determine the aperture field $\mathbf{E}_t(0)$. The input admittance is

$$Y = \frac{1}{Z_{w1}} + \frac{1}{Z_{12}} = \frac{1}{Z_{w1}}\frac{1-\Gamma}{1+\Gamma}.$$

From the first equation of (2.108), we obtain

$$\frac{1}{Z_{12}} = \frac{1}{Z_{w1}}\frac{-2\Gamma}{1+\Gamma} = \frac{2}{Z_{w1}}\frac{1-\int_{\Omega_0}\mathbf{E}_t(0)\cdot\mathbf{e}_1\, d\Omega}{\int_{\Omega_0}\mathbf{E}_t(0)\cdot\mathbf{e}_1\, d\Omega}. \tag{2.110}$$

Multiplying both sides of (2.109) by $\bar{\mathbf{E}}_t(0)$ and taking integration over Ω_0 yield

$$1 - \int_{\Omega_0}\mathbf{E}_t(0)\cdot\mathbf{e}_1\, d\Omega = \frac{\sum_{n=2}^{\infty} Z_{w1}Z_{wn}^{-1}\left|\int_{\Omega_0}\mathbf{E}_t(0)\cdot\mathbf{e}_n\, d\Omega\right|^2}{\int_{\Omega_0}\bar{\mathbf{E}}_t(0)\cdot\mathbf{e}_1\, d\Omega}.$$

Introducing this into (2.110) yields

$$\frac{1}{Z_{12}} = \frac{1}{Z_{w1}}\frac{-2\Gamma}{1+\Gamma} = \frac{1}{Z_{w1}}\frac{2\sum_{n=2}^{\infty} Z_{w1}Z_{wn}^{-1}\left|\int_{\Omega_0}\mathbf{E}_t(0)\cdot\mathbf{e}_n\, d\Omega\right|^2}{\left|\int_{\Omega_0}\mathbf{E}_t(0)\cdot\mathbf{e}_1\, d\Omega\right|^2}. \tag{2.111}$$

This is a variational expression on the aperture field $\mathbf{E}_t(0)$ (Kurokawa, 1969), i.e., (2.111) is stationary with respect to the aperture field $\mathbf{E}_t(0)$. Figure 2.19 shows some typical capacitive and inductive diaphragms in

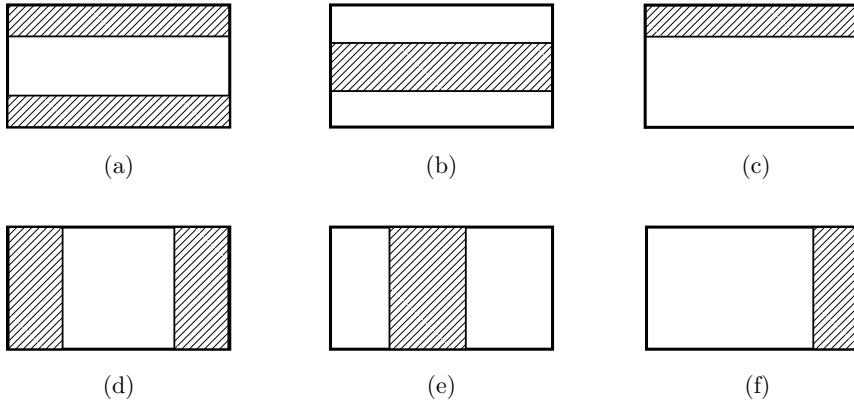


Figure 2.19 (a–c) Capacitive diaphragms. (d–f) Inductive diaphragms.

a rectangular waveguide. The solution to these diaphragms can all be obtained from (2.111).

Remark 2.5: The variatonal method is capable of analyzing a large variety of waveguide discontinuity problem, and it takes various forms. The method was introduced by Schwinger during the period from 1940 to 1945 and many useful results have been summarized in the *Discontinuities in Waveguides* (Schwinger and Saxon, 1968). □

2.4.3 Conducting Posts in Waveguide — Method of Green's Function

The post in a waveguide is often used as a matching element, a filter or a phase shifter. Figure 2.20 shows a circular conducting post across the narrow side of a rectangular waveguide. The dominant TE_{10} mode is incident upon the post and induces current on the post, which generates a number of higher order modes around the post. Since the electric field of the TE_{10} mode has a y component only and is independent of the y coordinate, and the whole structure is uniform in the y direction, the higher order modes excited must be independent of y, and thus are TE_{m0} modes. For TE_{m0} modes the magnetic energy is higher than electric energy, the post is thus equivalent to an inductor. This boundary value problem may be approached by the method of Green's function. Consider the line current source located at $x = x_0, z = z_0$

$$\mathbf{J} = I\delta(x - x_0)\delta(z - z_0)\mathbf{u}_y.$$

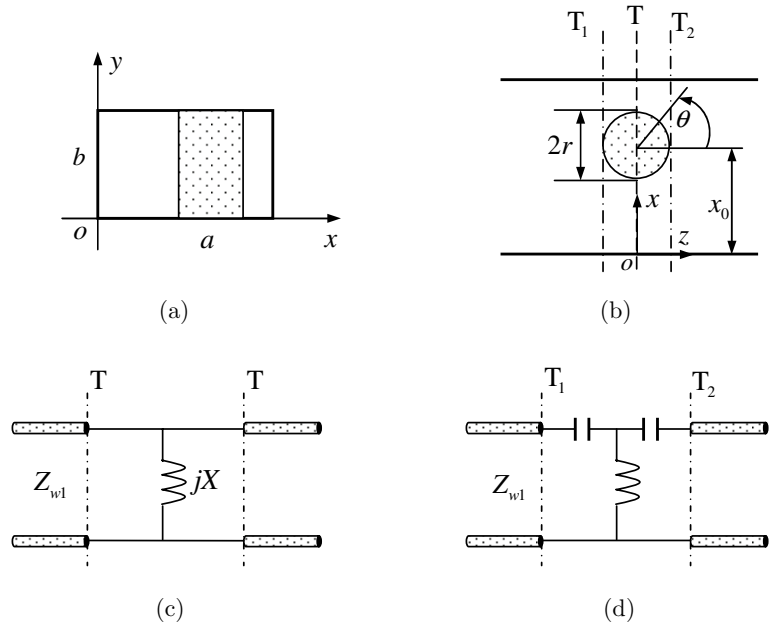


Figure 2.20 Inductive post in rectangular waveguide and its equivalent circuit.

Since the fields are independent of y, the electric field E_y generated by the line source satisfies

$$\begin{cases} \dfrac{\partial^2 E_y}{\partial x^2} + \dfrac{\partial^2 E_y}{\partial z^2} + k^2 E_y = j\omega\mu_0 I\delta(x - x_0)\delta(z - z_0) \\ E_y|_{x=0} = E_y|_{x=a} = 0 \end{cases}.$$

The solution of the equation (the Green's functions) may be written as

$$E_y = -\frac{\omega\mu_0 I}{a}\sum_{n=1}^{\infty}\frac{1}{\beta_n}\sin\frac{n\pi}{a}x\sin\frac{n\pi}{a}x_0 e^{-j\beta_n|z-z_0|}, \tag{2.112}$$

where $\beta_n^2 = k^2 - (n\pi/a)^2$. An integral equation may be obtained by using the Green's function (2.112). For a very thin post, the surface current on the post may be regarded as centered at the axis of the post $x = x_0, z = 0$. Let the incident field $E_{y,in}$ be a TE_{10} mode of unit amplitude $E_{y,in} = \sin\frac{\pi}{a}xe^{-j\beta_1 z}$. The current induced on the post produces a scattered field determined by (2.112). On the surface of the post, the tangential electric field must vanish. Introducing a polar coordinate system (r, θ) as shown in

Figure 2.20(b), we may write

$$\sin\frac{\pi}{a}(x_0 + r\sin\theta)e^{-j\beta_1 r\cos\theta} - \frac{\omega\mu_0 I}{a} \times \sum_{n=1}^{\infty}\frac{1}{\beta_n}\sin\frac{n\pi}{a}(x_0 + r\sin\theta)\sin\frac{n\pi}{a}x_0 e^{-j\beta_n|r\cos\theta|} = 0. \quad (2.113)$$

For a very thin post, the fields may be considered as constant along θ direction. For convenience, we may let $\theta = \pi/2 (z = 0)$ in (2.113) and make use of the following relation (Jones, 1964)

$$\sum_{n=1}^{\infty}\frac{1}{n}\sin\frac{n\pi}{a}(x_0 + r)\sin\frac{n\pi}{a}x_0 \approx \frac{1}{2}\ln\left(\frac{2a}{\pi r}\sin\frac{\pi}{a}x_0\right), \quad r \to 0$$

to find that

$$\sin\frac{\pi}{a}x_0 - \frac{\omega\mu_0 I}{a}\frac{1}{\beta_1}\sin^2\frac{\pi}{a}x_0\left(1 + 2\frac{jX}{Z_{w1}}\right) = 0 \quad (2.114)$$

where

$$\frac{X}{Z_{w1}} = \frac{a\beta_1}{4\pi\sin^2\frac{\pi}{a}x_0}\left[\ln\left(\frac{2a}{\pi r}\sin\frac{\pi}{a}x_0\right) - 2\sin^2\frac{\pi}{a}x_0 + 2\sum_{n=2}^{\infty}\left(\frac{-j\pi}{a\beta_n} - \frac{1}{n}\right)\sin^2\frac{n\pi}{a}x_0\right]. \quad (2.115)$$

Here Z_{w1} is the wave impedance of TE_{10} mode. Equation (2.114) can be used to determine I. The voltage reflection coefficient for the dominant mode at $z = 0$ is given by

$$\Gamma = -\frac{\omega\mu_0 I}{a}\frac{1}{\beta_1}\sin\frac{n\pi}{a}x_0 = \frac{-1}{1 + 2jX/Z_{w1}}.$$

The input admittance at $z = 0$ is

$$Y = \frac{1}{Z_{w1}}\frac{1-\Gamma}{1+\Gamma} = \frac{1}{jX} + \frac{1}{Z_{w1}}.$$

The equivalent circuit for the thin conducting post is shown in Figure 2.20(c). Since the logarithmic term in (2.115) dominates, we have $X > 0$, and the thin post is an inductor. For a general thick conducting post, the equivalent circuit is shown in Figure 2.20(d), where two capacitors must be introduced to take the thickness into account.

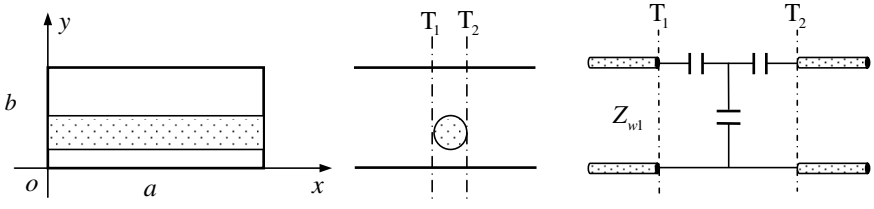


Figure 2.21 Capacitive post and its equivalent circuit.

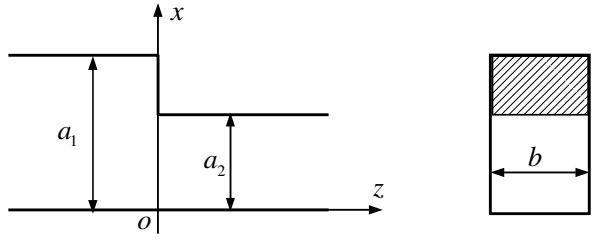


Figure 2.22 Waveguide step.

A circular conducting post across the broadside of a rectangular waveguide is shown in Figures 2.21(a) and 2.21(b), which is equivalent to a capacitive circuit as shown in Figure 2.21(c).

2.4.4 Waveguide Steps — Mode Matching Method

A waveguide step in broadside direction (H-plane step) is shown in Figure 2.22 and the discontinuity occurs at $z = 0$. A dominant TE_{10} mode of unit amplitude is assumed and is incident upon the step from the region $z < 0$ (first waveguide). The fields in this region may be expanded as a series using first N vector modal functions

$$\mathbf{E}_t = (e^{-j\beta_1 z} + \Gamma e^{j\beta_1 z})\mathbf{e}_1 + \sum_{n=2}^{N} V_n e^{j\beta_n z}\mathbf{e}_n,$$

$$\mathbf{H}_t = (e^{-j\beta_1 z} - \Gamma e^{j\beta_1 z})Z_{w1}^{-1}\mathbf{u}_z \times \mathbf{e}_1 - \sum_{n=2}^{N} V_n Z_{wn}^{-1} e^{j\beta_n z}\mathbf{u}_z \times \mathbf{e}_n,$$

where Z_{wn} is the wave impedance, and

$$\mathbf{e}_n = -\mathbf{u}_y \sqrt{\frac{2}{a_1 b}} \sin \frac{n\pi}{a_1} x, \quad \beta_n^2 = k^2 - (n\pi/a_1)^2.$$

Similarly the fields in the region $z > 0$ (second waveguide) may be written as

$$\mathbf{E}'_t = V'_1 e^{-j\beta'_1 z} \mathbf{e}'_1 + \sum_{n=2}^{N} \mathbf{e}'_n V'_n e^{-j\beta'_n z},$$

$$\mathbf{H}'_t = V'_1 e^{-j\beta'_1 z} Z'^{-1}_{w1} \mathbf{u}_z \times \mathbf{e}'_1 + \sum_{n=2}^{N} \mathbf{u}_z \times \mathbf{e}'_n V'_n Z'^{-1}_{wn} e^{-j\beta'_n z},$$

where only the dominant mode is assumed to be propagating, and

$$\mathbf{e}'_n = -\mathbf{u}_y \sqrt{\frac{2}{a_2 b}} \sin \frac{n\pi}{a_2} x, \quad \beta'^2_n = k^2 - (n\pi/a_2)^2.$$

The tangential fields must be continuous at $z = 0$, which yields

$$(1+\Gamma)\mathbf{e}_1 + \sum_{n=2}^{N} V_n \mathbf{e}_n = V'_1 \mathbf{e}'_1 + \sum_{n=2}^{N} V'_n \mathbf{e}'_n, \tag{2.116}$$

$$(1-\Gamma) Z_{w1}^{-1} \mathbf{u}_z \times \mathbf{e}_1 - \sum_{n=2}^{N} V_n Z_{wn}^{-1} \mathbf{u}_z \times \mathbf{e}_n = V'_1 Z'^{-1}_{w1} \mathbf{u}_z \times \mathbf{e}'_1 + \sum_{n=2}^{N} V'_n Z'^{-1}_{wn} \mathbf{u}_z \times \mathbf{e}'_n. \tag{2.117}$$

Multiplying both sides of (2.116) by $\mathbf{e}_m$ and taking the integration over the cross section Ω of the first waveguide yield

$$(1+\Gamma)\delta_{m1} + V_m(1-\delta_{m1}) = \sum_{n=1}^{N} V'_n B_{mn}, \quad m = 1, 2, \ldots, N, \tag{2.118}$$

where

$$B_{mn} = \int_{\Omega} \mathbf{e}_m \cdot \mathbf{e}'_n \, d\Omega.$$

Multiplying both sides of (2.117) by $\mathbf{u}_z \times \mathbf{e}'_m$ and taking the integration over the cross section Ω' of the second waveguide lead to

$$(1-\Gamma) Z_{w1}^{-1} D_{m1} - \sum_{n=2}^{N} V_n Z_{wn}^{-1} D_{mn} = V'_m Z'^{-1}_{wm}, \quad m = 1, 2, \ldots, N, \tag{2.119}$$

where

$$D_{mn} = \int_{\Omega'} \mathbf{e}'_m \cdot \mathbf{e}_n \, d\Omega.$$

The expansion coefficients Γ, V_n, V'_n can be determined from (2.118) and (2.119).

2.4.5 Coupling by Small Apertures

Consider a system of two waveguides coupled by a small aperture S_a bounded by Γ, as shown in Figure 2.23. The impressed electric current $\mathbf{J}_{\text{imp}}$ and magnetic current $\mathbf{J}_{m,\text{imp}}$ are assumed to be located in waveguide 1 only and there are no impressed sources in waveguide 2. By Schelkunoff–Love equivalence principle, the original problem can be separated into two equivalent problems as shown in Figures 2.23(b) and 2.23(c). In waveguide 1, the fields are produced by the impressed sources $\mathbf{J}_{\text{imp}}$, $\mathbf{J}_{m,\text{imp}}$

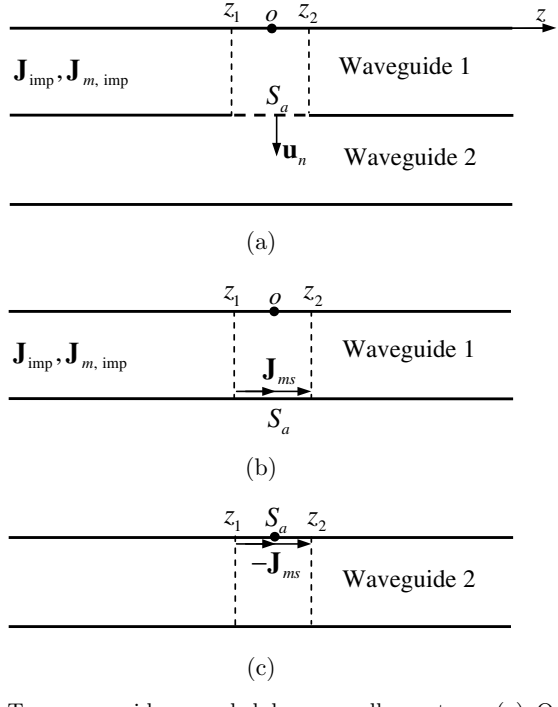


Figure 2.23 Two waveguides coupled by a small aperture. (a) Original problem. (b) Equivalent problem for waveguide 1. (c) Equivalent problem for waveguide 2.

and the equivalent magnetic current $\mathbf{J}_{ms} = \mathbf{u}_n \times \mathbf{E}$ over the aperture region S_a with the aperture covered by an electric conductor. In waveguide 2, the fields are produced by the equivalent magnetic current $-\mathbf{J}_{ms}$ with the aperture covered by an electric conductor. When the aperture is absent (i.e., closed by a perfect conductor), the incident fields generated by the impressed sources in waveguide 1 are denoted by $\mathbf{E}_{in}^{(1)}$, $\mathbf{H}_{in}^{(1)}$ (the superscripts 1 and 2 will be used to designate waveguide 1 and 2 respectively). Then

$$\mathbf{u}_n \times \mathbf{E}_{in}^{(1)} = 0, \quad \mathbf{u}_n \cdot \mathbf{H}_{in}^{(1)} = 0.$$

The total fields in waveguide 1 can be expressed as

$$\mathbf{E}^{(1)} = \mathbf{E}_{in}^{(1)} + \mathbf{E}_s^{(1)} = 0, \quad \mathbf{H}^{(1)} = \mathbf{H}_{in}^{(1)} + \mathbf{H}_s^{(1)},$$

where $\mathbf{E}_s^{(1)}$ and $\mathbf{H}_s^{(1)}$ are the scattered fields produced by the magnetic current $\mathbf{J}_{ms}$ and satisfy

$$\begin{aligned} \nabla \times \mathbf{H}_s^{(1)} &= j\omega\varepsilon_0 \mathbf{E}_s^{(1)}, \\ \nabla \times \mathbf{E}_s^{(1)} &= -j\omega\mu_0 \mathbf{H}_s^{(1)} - \mathbf{J}_{ms}. \end{aligned}$$

Assume that the magnetic current element $\mathbf{J}_{ms}$ is located between z_1 and z_2. According to (2.17), the scattered fields for $z \geq z_2$ and $z \leq z_1$ in waveguide 1 may be respectively expanded in terms of the vector modal functions as follows

$$\begin{aligned} \mathbf{E}_s^{(1)} &= \sum_{n=1}^{\infty} \left(V_n^{(1)+} \mathbf{e}_n^{(1)} + \frac{k_{cn}}{j\beta_n} I_n^{(1)+} Z_{wn} \mathbf{u}_z \frac{\nabla \cdot \mathbf{e}_n^{(1)}}{k_{cn}} \right) = \sum_{n=1}^{\infty} A_n^{(1)} \mathbf{E}_n^{(1)+}, \\ \mathbf{H}_s^{(1)} &= \sum_{n=1}^{\infty} \left(I_n^{(1)+} \mathbf{u}_z \times \mathbf{e}_n^{(1)} - \frac{k_{cn}}{j\beta_n} \frac{V_n^{(1)+}}{Z_{wn}} \frac{\nabla \times \mathbf{e}_n^{(1)}}{k_{cn}} \right) = \sum_{n=1}^{\infty} A_n^{(1)} \mathbf{H}_n^{(1)+}, \end{aligned} \tag{2.120}$$

$$\begin{aligned} \mathbf{E}_s^{(1)} &= \sum_{n=1}^{\infty} \left(V_n^{(1)-} \mathbf{e}_n^{(1)} - \frac{k_{cn}}{j\beta_n} I_n^{(1)-} Z_{wn} \mathbf{u}_z \frac{\nabla \cdot \mathbf{e}_n^{(1)}}{k_{cn}} \right) = \sum_{n=1}^{\infty} B_n^{(1)} \mathbf{E}_n^{(1)-}, \\ \mathbf{H}_s^{(1)} &= \sum_{n=1}^{\infty} \left(-I_n^{(1)-} \mathbf{u}_z \times \mathbf{e}_n^{(1)} - \frac{k_{cn}}{j\beta_n} \frac{V_n^{(1)-}}{Z_{wn}} \frac{\nabla \times \mathbf{e}_n^{(1)}}{k_{cn}} \right) = \sum_{n=1}^{\infty} B_n^{(1)} \mathbf{H}_n^{(1)-}, \end{aligned} \tag{2.121}$$

where

$$\mathbf{E}_n^{(1)+} = (\mathbf{e}_n^{(1)} + \mathbf{u}_z e_{zn}^{(1)})e^{-j\beta_n^{(1)}z}, \quad \mathbf{H}_n^{(1)+} = (\mathbf{h}_n^{(1)} + \mathbf{u}_z h_{zn}^{(1)})e^{-j\beta_n^{(1)}z},$$
$$\mathbf{E}_n^{(1)-} = (\mathbf{e}_n^{(1)} - \mathbf{u}_z e_{zn}^{(1)})e^{j\beta_n^{(1)}z}, \quad \mathbf{H}_n^{(1)-} = (-\mathbf{h}_n^{(1)} + \mathbf{u}_z h_{zn}^{(1)})e^{j\beta_n^{(1)}z}, \tag{2.122}$$

with

$$\mathbf{h}_n^{(1)} = \frac{\mathbf{u}_z \times \mathbf{e}_n^{(1)}}{Z_{wn}^{(1)}}, \quad e_{zn}^{(1)} = \frac{\nabla \cdot \mathbf{e}_n^{(1)}}{j\beta_n^{(1)}}, \quad h_{zn}^{(1)}\mathbf{u}_z = -\frac{\nabla \times \mathbf{e}_n^{(1)}}{j\beta_n^1 Z_{wn}^{(1)}}. \tag{2.123}$$

Note that

$$\int_{\Omega_1} (\mathbf{e}_n^{(1)} \times \mathbf{h}_n^{(1)}) \cdot \mathbf{u}_z \, d\Omega = \frac{1}{Z_{wn}^{(1)}}, \tag{2.124}$$

where Ω_1 is the cross section of waveguide 1. The expansion coefficients in (2.120) and (2.121) can be determined by (2.33)

$$A_n^{(1)} = \frac{Z_{wn}^{(1)}}{2} \int_{S_a} \mathbf{J}_{ms} \cdot \mathbf{H}_n^{(1)-} dS, \quad B_n^{(1)} = \frac{Z_{wn}^{(1)}}{2} \int_{S_a} \mathbf{J}_{ms} \cdot \mathbf{H}_n^{(1)+} dS. \tag{2.125}$$

We now introduce a local coordinate system (ξ, ζ) with the origin at the center of the aperture as illustrated in Figure 2.24. For a small aperture, the field $\mathbf{H}_n^{(1)\pm}$ may be expanded into a Taylor series about the origin

$$\begin{aligned}\mathbf{H}_n^{(1)\pm}(\xi,\zeta) &= \mathbf{H}_n^{(1)\pm}(0,0) + \xi\frac{\partial \mathbf{H}_n^{(1)\pm}(0,0)}{\partial \xi} + \zeta\frac{\partial \mathbf{H}_n^{(1)\pm}(0,0)}{\partial \zeta} \\ &= \mathbf{H}_n^{(1)\pm}(0,0) + \mathbf{r}_a \cdot \nabla \mathbf{H}_n^{(1)\pm}(0,0),\end{aligned}$$

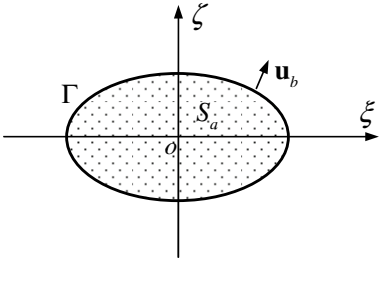


Figure 2.24 Aperture coordinates.

where $\mathbf{r}_a = \xi\mathbf{u}_\xi + \zeta\mathbf{u}_\zeta$. Equation (2.125) can be written as

$$A_n^{(1)} = \frac{Z_{wn}^{(1)}}{2}\mathbf{H}_n^{(1)-}(0,0)\cdot\int\limits_{S_a}\mathbf{J}_{ms}\,dS + \frac{Z_{wn}^{(1)}}{2}\int\limits_{S_a}\mathbf{r}_a\cdot\nabla\mathbf{H}_n^{(1)-}(0,0)\cdot\mathbf{J}_{ms}\,dS,$$

$$B_n^{(1)} = \frac{Z_{wn}^{(1)}}{2}\mathbf{H}_n^{(1)+}(0,0)\cdot\int\limits_{S_a}\mathbf{J}_{ms}\,dS + \frac{Z_{wn}^{(1)}}{2}\int\limits_{S_a}\mathbf{r}_a\cdot\nabla\mathbf{H}_n^{(1)+}(0,0)\cdot\mathbf{J}_{ms}\,dS. \tag{2.126}$$

Since the magnetic current is confined in the aperture, for an arbitrary function ϕ, we have

$$\begin{aligned}\int\limits_{S_a}\nabla\cdot(\phi\mathbf{J}_{ms})dS &= \int\limits_{S_a}(\phi\nabla\cdot\mathbf{J}_{ms} + \mathbf{J}_{ms}\cdot\nabla\phi)dS\\ &= \int\limits_{\partial S_a}(\phi\mathbf{J}_{ms})\cdot\mathbf{u}_b d\Gamma = 0,\end{aligned} \tag{2.127}$$

with $\mathbf{u}_b$ being the unit outward normal to the aperture boundary Γ. Making use of (2.127), the first integral on the right-hand side of (2.126) can be written as

$$\int\limits_{S_a}\mathbf{J}_{ms}\,dS = -\int\limits_{S_a}\mathbf{r}_a\nabla\cdot\mathbf{J}_{ms}\,dS = j\omega\int\limits_{S_a}\mathbf{r}_a\rho_{ms}\,dS = j\omega\mu_0\mathbf{m},$$

where

$$\mathbf{m} = \frac{1}{\mu_0}\int\limits_{S_a}\mathbf{r}_a\rho_{ms}\,dS = \frac{1}{j\omega\mu_0}\int\limits_{S_a}\mathbf{J}_{ms}\,dS = \frac{1}{j\omega\mu_0}\int\limits_{S_a}(\mathbf{u}_n\times\mathbf{E})dS \tag{2.128}$$

is the magnetic dipole moment. Note that

$$\begin{aligned}&\mathbf{r}_a\cdot\nabla\mathbf{H}_n^{(1)\pm}(0,0)\cdot\mathbf{J}_{ms}\\ &\quad= (\xi\mathbf{u}_\xi+\zeta\mathbf{u}_\zeta)\cdot\left(\mathbf{u}_\xi\mathbf{u}_\xi\frac{\partial H_{n\xi}^{(1)\pm}(0,0)}{\partial\xi} + \mathbf{u}_\xi\mathbf{u}_\zeta\frac{\partial H_{n\zeta}^{(1)\pm}(0,0)}{\partial\xi}\right.\\ &\qquad\left.+\,\mathbf{u}_\zeta\mathbf{u}_\xi\frac{\partial H_{n\xi}^{(1)\pm}(0,0)}{\partial\zeta} + \mathbf{u}_\zeta\mathbf{u}_\zeta\frac{\partial H_{n\zeta}^{(1)\pm}(0,0)}{\partial\zeta}\right)\cdot(J_{ms\xi}\mathbf{u}_\xi + J_{ms\zeta}\mathbf{u}_\zeta).\end{aligned}$$

This can be written as

$$\begin{aligned}\mathbf{r}_a\cdot\nabla\mathbf{H}_n^{(1)\pm}(0,0)\cdot\mathbf{J}_{ms} &= \xi J_{ms\xi}\frac{\partial H_{n\xi}^{(1)\pm}(0,0)}{\partial\xi} + \xi J_{ms\zeta}\frac{\partial H_{n\zeta}^{(1)\pm}(0,0)}{\partial\xi}\\ &\quad+\zeta J_{ms\xi}\frac{\partial H_{n\xi}^{(1)\pm}(0,0)}{\partial\zeta} + \zeta J_{ms\zeta}\frac{\partial H_{n\zeta}^{(1)\pm}(0,0)}{\partial\zeta}.\end{aligned}$$

Subtracting and adding similar terms, we get

$$\begin{aligned}
&\mathbf{r}_a \cdot \nabla \mathbf{H}_n^{(1)\pm}(0,0) \cdot \mathbf{J}_{ms} \\
&\quad = \left(\frac{\xi}{2} J_{ms\zeta} - \frac{\zeta}{2} J_{ms\xi}\right) \left(\frac{\partial H_{n\zeta}^{(1)\pm}(0,0)}{\partial \xi} - \frac{\partial H_{n\xi}^{(1)\pm}(0,0)}{\partial \zeta}\right) \\
&\qquad + \frac{\xi}{2} J_{ms\zeta} \frac{\partial H_{n\xi}^{(1)\pm}(0,0)}{\partial \zeta} + \frac{\zeta}{2} J_{ms\xi} \frac{\partial H_{n\zeta}^{(1)\pm}(0,0)}{\partial \xi} + \xi J_{ms\xi} \frac{\partial H_{n\xi}^{(1)\pm}(0,0)}{\partial \xi} \\
&\qquad + \zeta J_{ms\zeta} \frac{\partial H_{n\zeta}^{(1)\pm}(0,0)}{\partial \zeta} + \frac{\xi}{2} J_{ms\zeta} \frac{\partial H_{n\zeta}^{(1)\pm}(0,0)}{\partial \xi} + \frac{\zeta}{2} J_{ms\xi} \frac{\partial H_{n\xi}^{(1)\pm}(0,0)}{\partial \zeta}.
\end{aligned} \tag{2.129}$$

The first term on the right-hand side can be written as $j\omega\varepsilon_0 \mathbf{E}_n^{(1)\pm} \cdot (\mathbf{r}_a \times \mathbf{J}_{ms})/2$. This gives

$$j\omega\varepsilon_0 \mathbf{E}_n^{(1)\pm}(0,0) \cdot \int\limits_{S_a} \frac{\mathbf{r}_a \times \mathbf{J}_{ms}}{2} dS = -j\omega \mathbf{E}_n^{(1)\pm}(0,0) \cdot \mathbf{p}, \tag{2.130}$$

where

$$\mathbf{p} = \varepsilon_0 \frac{1}{2} \int\limits_{S_a} (\mathbf{J}_{ms} \times \mathbf{r}_a) dS \tag{2.131}$$

is the equivalent electric dipole moment of the magnetic current. Setting $\phi = \xi^2/2$, $\zeta^2/2$ and $\xi\zeta$ in (2.127), we obtain respectively

$$\begin{aligned}
\int\limits_{S_a} \xi J_{ms\xi}\, dS &= \frac{j\omega}{2} \int\limits_{S_a} \xi^2 \rho_{ms}\, dS, \\
\int\limits_{S_a} \zeta J_{ms\zeta}\, dS &= \frac{j\omega}{2} \int\limits_{S_a} \zeta^2 \rho_{ms}\, dS, \\
\int\limits_{S_a} (\xi J_{ms\zeta} + \zeta J_{ms\xi}) dS &= j\omega \int\limits_{S_a} \xi\zeta \rho_{ms}\, dS.
\end{aligned}$$

Introducing the dyadic magnetic quadrupole $\overleftrightarrow{\mathbf{Q}}^m$ defined by

$$\begin{aligned}
Q_{\xi\xi}^m &= \frac{1}{\mu_0} \int\limits_{S_a} \xi^2 \rho_{ms}\, dS, \quad Q_{\zeta\zeta}^m = \frac{1}{\mu_0} \int\limits_{S_a} \zeta^2 \rho_{ms}\, dS, \\
Q_{\xi\zeta}^m &= Q_{\zeta\xi}^m = \frac{1}{\mu_0} \int\limits_{S_a} \xi\zeta \rho_{ms}\, dS
\end{aligned}$$

we obtain

$$\int_{S_a} \mathbf{r}_a \cdot \nabla \mathbf{H}_n^{(1)\pm}(0,0) \cdot \mathbf{J}_{ms} dS = j\omega \mathbf{E}_n^{(1)\pm}(0,0) \cdot \mathbf{p} + \frac{j\omega\mu_0}{2} \nabla \mathbf{H}_n^{(1)\pm}(0,0) : \overleftrightarrow{\mathbf{Q}}^m, \tag{2.132}$$

where the double dot denotes the double product of two dyads. The expansion coefficients in (2.126) are then given by

$$\begin{aligned} A_n^{(1)} &= \frac{Z_{wn}^{(1)}}{2} \Big[- \mathbf{E}_n^{(1)-}(0,0) \cdot j\omega \mathbf{p} \\ &\qquad + \mathbf{H}_n^{(1)-}(0,0) \cdot j\omega\mu_0 \mathbf{m} + \frac{\mu_0}{2} \nabla \mathbf{H}_n^{(1)-}(0,0) : \overleftrightarrow{\mathbf{Q}}^m \Big], \\ B_n^{(1)} &= \frac{Z_{wn}^{(1)}}{2} \Big[- \mathbf{E}_n^{(1)+}(0,0) \cdot j\omega \mathbf{p} \\ &\qquad + \mathbf{H}_n^{(1)+}(0,0) \cdot j\omega\mu_0 \mathbf{m} + \frac{\mu_0}{2} \nabla \mathbf{H}_n^{(1)+}(0,0) : \overleftrightarrow{\mathbf{Q}}^m \Big]. \end{aligned} \tag{2.133}$$

The fields in waveguide 2 are generated by the equivalent magnetic current $-\mathbf{J}_{ms}$. The fields in the regions $z \geq z_2$ and $z \leq z_1$ in waveguide 2 may be respectively expanded in terms of the vector modal functions as follows

$$\mathbf{E}_s^{(2)} = \sum_{n=1}^{\infty} A_n^{(2)} \mathbf{E}_n^{(2)+}, \quad \mathbf{H}_s^{(2)} = \sum_{n=1}^{\infty} A_n^{(2)} \mathbf{H}_n^{(2)+}, \tag{2.134}$$

$$\mathbf{E}_s^{(2)} = \sum_{n=1}^{\infty} B_n^{(2)} \mathbf{E}_n^{(2)-}, \quad \mathbf{H}_s^{(2)} = \sum_{n=1}^{\infty} B_n^{(2)} \mathbf{H}_n^{(2)-}, \tag{2.135}$$

where

$$\begin{aligned} \mathbf{E}_n^{(2)+} &= (\mathbf{e}_n^{(2)} + \mathbf{u}_z e_{zn}^{(2)}) e^{-j\beta_n^{(2)} z}, \quad \mathbf{H}_n^{(2)+} = (\mathbf{h}_n^{(2)} + \mathbf{u}_z h_{zn}^{(2)}) e^{-j\beta_n^{(2)} z}, \\ \mathbf{E}_n^{(2)-} &= (\mathbf{e}_n^{(2)} - \mathbf{u}_z e_{zn}^{(2)}) e^{j\beta_n^{(2)} z}, \quad \mathbf{H}_n^{(2)-} = (-\mathbf{h}_n^{(2)} + \mathbf{u}_z h_{zn}^{(2)}) e^{j\beta_n^{(2)} z}, \end{aligned} \tag{2.136}$$

with

$$\mathbf{h}_n^{(2)} = \frac{\mathbf{u}_z \times \mathbf{e}_n^{(2)}}{Z_{wn}^{(1)}}, \quad e_{zn}^{(2)} = \frac{\nabla \cdot \mathbf{e}_n^{(2)}}{j\beta_n^{(1)}}, \quad h_{zn}^{(2)} = -\frac{\nabla \times \mathbf{e}_n^{(2)}}{j\beta_n^{(1)} Z_{wn}^{(2)}}. \tag{2.137}$$

Similarly, we have

$$\begin{aligned} A_n^{(2)} &= -\frac{Z_{wn}^{(2)}}{2} \Big[- \mathbf{E}_n^{(2)-}(0,0) \cdot j\omega \mathbf{p} \\ &\qquad + \mathbf{H}_n^{(2)-}(0,0) \cdot j\omega\mu_0 \mathbf{m} + \frac{\mu_0}{2} \nabla \mathbf{H}_n^{(2)-}(0,0) : \overleftrightarrow{\mathbf{Q}}^m \Big], \end{aligned}$$

$$B_n^{(2)} = -\frac{Z_{wn}^{(2)}}{2}\Big[- \mathbf{E}_n^{(2)+}(0,0)\cdot j\omega\mathbf{p} + \mathbf{H}_n^{(2)+}(0,0)\cdot j\omega\mu_0\mathbf{m} + \frac{\mu_0}{2}\nabla\mathbf{H}_n^{(2)+}(0,0) : \overleftrightarrow{\mathbf{Q}}^m\Big], \quad (2.138)$$

where $\mathbf{p}$, $\mathbf{m}$ and $\overleftrightarrow{\mathbf{Q}}^m$ are same as defined before. In most applications, the quadrupole terms in (2.138) can be ignored. The magnetic current $\mathbf{J}_{ms}$ and the charge ρ_{ms} may be determined by using numerical methods. Cohn proposed an electrolytic-tank method for determining the aperture parameters of arbitrary shape (Cohn, 1951). For very small apertures, the static field solution for the dipole moments can be readily found (Bethe, 1944; Stratton, 1941). For a small circular aperture of radius a, the dipole moments can be obtained by solving integral equations as follows (see Section 7.3)

$$\mathbf{m} = \frac{8}{3}a_0^3\mathbf{H}_{in}^{(1)}(0), \quad \frac{\mathbf{p}}{\varepsilon_0} = -\frac{4a^3}{3}\mathbf{E}_{in}^{(1)}(0). \quad (2.139)$$

2.4.6 Numerical Analysis — Finite Difference Method

The exact solution of Maxwell equations for a waveguide discontinuity problem is generally very difficult. In this section, we discuss the numerical solution of an arbitrarily shaped two-dimensional waveguide junction by means of FDM. The waveguide junction consists of n rectangular waveguide ports and is shown in Figure 2.25, where the reference planes Γ_p $(p = 1, 2, \ldots, n)$ and the metallic wall Γ_0 completely enclose the waveguide discontinuity region Ω; d_p is the width a_p or height b_p of the waveguide p for the H-plane (the plane containing magnetic field) or E-plane (the plane containing electric field) junction. The waveguide p is assumed to be filled with a dielectric of relative permittivity ε_{rp}. If the excitation by the dominant TE_{10} mode is assumed, the waveguide discontinuity can then be described by the following equations.

$$(\nabla_t^2 + k^2)\phi = 0, \quad (2.140)$$

where $\nabla_t^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$, $k^2 = \tau k_0^2$, $k_0^2 = \omega^2\mu_0\varepsilon_0$, and

$$\tau = \begin{cases} \varepsilon_r, & \text{for H-plane junction} \\ \varepsilon_r - (\pi/k_0 a)^2, & \text{for E-plane junction} \end{cases}, \quad (2.141)$$

$$\phi = \begin{cases} E_z, & \text{for H-plane junction} \\ H_z, & \text{for E-plane junction} \end{cases}, \quad (2.142)$$

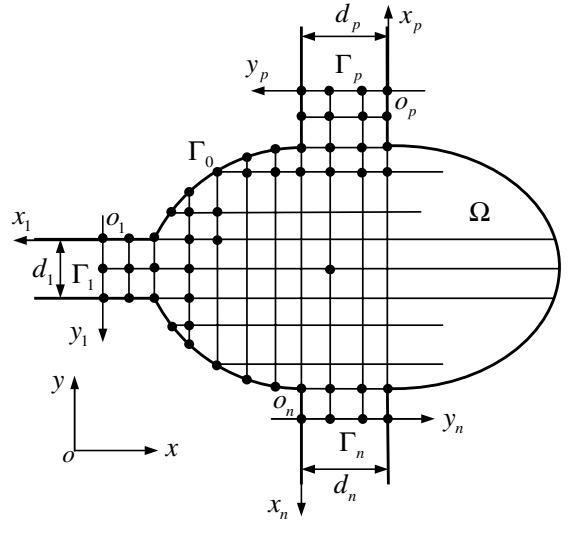


Figure 2.25 An arbitrarily shaped waveguide junction.

where E_z and H_z are the z-components of electric and magnetic field respectively.

Similar to the finite difference solution of the waveguide eigenvalue problem, the discontinuity region Ω may be divided into a number of polygons (Geyi, 1991). For the interior nodes n_i as shown in Figure 2.11(a), the node equation can be written as

$$\sum_{j=1}^{e_i} [\phi(n_i) - \phi(m_j)] \frac{\overline{p_{j-1}q_j} + \overline{q_j p_j}}{\overline{m_j n_i}} - k^2 \phi(n_i) S_i = 0. \tag{2.143}$$

For the boundary node n_i shown in Figure 2.11(b), the following relation can be derived in a similar way

$$\sum_{j=2}^{e_i-1} [\phi(n_i) - \phi(m_j)] \frac{\overline{p_{j-1}q_j} + \overline{q_j p_j}}{\overline{m_j n_i}} + [\phi(n_i) - \phi(m_1)] \frac{\overline{p_1 q_1}}{\overline{m_1 n_i}}$$
$$+ [\phi(n_i) - \phi(m_{e_i})] \frac{\overline{p_{e_i-1}q_{e_i}}}{\overline{m_{e_i} n_i}} - k_c^2 \phi(n_i) S_i - S(i) = 0, \tag{2.144}$$

where

$$S(i) = \int_{\overline{n_i q_1}} \frac{\partial \phi}{\partial n} \, d\Gamma + \int_{\overline{n_i q_{e_i}}} \frac{\partial \phi}{\partial n} \, d\Gamma. \tag{2.145}$$

Note that $S(i)$ vanishes if the boundary node is on the metallic wall Γ_0. In order to evaluate $S(i)$ for the boundary nodes on the reference planes $\Gamma_p (p = 1, 2, \ldots, n)$, we may assume that the dominant TE_{10} mode of amplitude $B_1^{(q)}$ is incident from waveguide $q (q = 1, 2, \ldots, n)$. The scalar field ϕ on $\Gamma_p (p = 1, 2, \ldots, n)$ may be expressed as

$$\phi(x_p, y_p) = B_1^{(q)} \delta_{pq} e^{j\beta_1^{(q)} x_p} f_1^{(q)}(y_q) + \sum_{m=1}^{\infty} a_m^{(p)} e^{-j\beta_m^{(p)} x_p} f_m^{(p)}(y_p), \quad (2.146)$$

where

$$f_m^{(p)}(y_p) = \sqrt{\frac{\sigma_m - 1}{b_p}} \cos \frac{(m-1)\pi}{b_p} y_p,$$

$$\beta_m^{(p)} = \sqrt{\varepsilon_{rp} k_0^2 - \left(\frac{\pi}{a}\right)^2 - \left[\frac{(m-1)\pi}{b_p}\right]^2},$$

$$\sigma_m = \begin{cases} 1, & m = 0 \\ 2, & m \neq 0 \end{cases}$$

for E-plane junction and

$$f_m^{(p)}(y_p) = \sqrt{\frac{2}{a_p}} \sin \frac{m\pi}{a_p} y_p,$$

$$\beta_m^{(p)} = \sqrt{\varepsilon_{rp} k_0^2 - \left(\frac{m\pi}{a_p}\right)^2}$$

for H-plane junction. Introducing (2.146) into (2.145), we obtain

$$S(i) = -j2B_1^{(q)} \delta_{pq} \beta_1^{(q)} I_1^{(q)}(i) + \sum_{m=1}^{\infty} j\beta_m^{(p)} a_m^{(p)} I_m^{(p)}(i), \quad (2.147)$$

where

$$a_m^{(p)} = \int_0^{d_p} \phi(x_p, y_p)|_{x_p=0} dy_p,$$

$$I_m^{(p)}(i) = \int_{\overline{q_1 q_{e_i}}} f_m^{(p)}(y_p) dy_p.$$

Let the number of nodes on Γ_p be denoted by $n^{(p)}$. The scalar field ϕ on Γ_p can be approximated by a constant over the integration interval $\overline{q_1 q_{e_i}}$,

denoted by $\phi(n_i)$. Then

$$a_m^{(p)} = \sum_{l=1}^{n^{(p)}} \phi(n_l) I_m^{(p)}(l).$$

Substituting this into (2.147) yields

$$S(i) = -j2B_1^{(q)}\delta_{pq}\beta_1^{(q)} I_1^{(q)}(i) + \sum_{l=1}^{n^{(p)}} \phi(n_l) \sum_{m=1}^{\infty} j\beta_m^{(p)} I_m^{(p)}(l) I_m^{(p)}(i). \quad (2.148)$$

Combining (2.144) and (2.148), we obtain the node equation for the boundary nodes.

Once the node values of the scalar field ϕ are known, the scattering parameter for the TE_{10} mode can be calculated as follows

$$\begin{aligned} S_{pp} &= \frac{1}{B_1^{(q)}} \int_0^{d_p} \phi(x_p, y_p)|_{x_p=0} f_1^{(p)}(y_p) dy_p - 1, \\ S_{pq} &= \sqrt{\frac{\beta_1^{(p)}\tau_q}{\beta_1^{(q)}\tau_p}} \frac{1}{B_1^{(q)}} \int_0^{d_p} \phi(x_p, y_p)|_{x_p=0} f_1^{(p)}(y_p) dy_p, \end{aligned} \quad (2.149)$$

where τ_p and τ_q stand for the parameter τ in waveguide p and q respectively, and they should be replaced by 1 for H-plane junctions.

2.5 Transient Fields in Waveguides

According to the linear system theory and Fourier analysis, the response of the system to an arbitrary pulse can be obtained by superimposing its responses to all the real frequencies. In other words, the solution to the time-domain problem can be expressed in terms of the time-harmonic solution through the use of the Fourier transform. This process can be assisted by the fast Fourier transform and has been used extensively in studying the transient responses of electromagnetic systems. The procedure, however, is not always most effective and is not a trivial exercise since the time-harmonic problem must be solved for a large range of frequencies, and only an approximate time-harmonic solution valid over a finite frequency band can be obtained.

Moreover, the time-harmonic solution may not be able to give the correct physical picture in some situations. The time-harmonic field theory is founded on the assumption that a monotonic electromagnetic source turns on at $t = -\infty$ and the initial conditions of the fields produced by the source are ignored. This assumption does not cause any problem if the system has dissipation or radiation loss. When the system is lossless, the assumption may lead to physically unacceptable solutions. For example, the time-harmonic theory predicts that the field response of a lossless metal cavity is sinusoidal if the excitation source is sinusoidal. The time-domain theory, however, shows that a sinusoidal response can be built up only if the cavity is excited by a sinusoidal source whose frequency coincides with one of the resonant frequencies. In addition, the field responses in a lossless cavity predicted by the time-harmonic theory are singular everywhere inside the cavity if the frequency of the sinusoidal excitation source coincides with one of the resonant frequencies of the cavity, while the time-domain theory always gives finite field responses. Therefore, we are forced to seek a solution in the time domain in some situations.

2.5.1 Field Expansions

Assume that the medium in the waveguide is homogeneous and isotropic with medium parameters μ, ε and σ. The cross section of the waveguide is denoted by Ω and its boundary by Γ, which is assumed to be a perfect conductor. The fields in the waveguide may be expanded in terms of vector modal functions as follows

$$\begin{aligned}
\mathbf{E}(\mathbf{r},t) &= \sum_{n=1}^{\infty} v_n(z,t)\mathbf{e}_n + \mathbf{u}_z \sum_{n=1}^{\infty} e'_{zn}(z,t)\frac{\nabla\cdot\mathbf{e}_n}{k_{cn}},\\
\mathbf{H}(\mathbf{r},t) &= \sum_{n=1}^{\infty} i_n(z,t)\mathbf{u}_z\times\mathbf{e}_n + \mathbf{u}_z\frac{1}{\sqrt{\Omega}}\int_{\Omega}\frac{\mathbf{u}_z\cdot\mathbf{H}(\mathbf{r},t)}{\sqrt{\Omega}}\,d\Omega\\
&\quad + \sum_{n=1}^{\infty} h'_{zn}(z,t)\frac{\nabla\times\mathbf{e}_n}{k_{cn}},
\end{aligned} \tag{2.150}$$

$$\begin{aligned}
\nabla\times\mathbf{E} &= \sum_{n=1}^{\infty}\left(\frac{\partial v_n}{\partial z} + k_{cn}e'_{zn}\right)\mathbf{u}_z\times\mathbf{e}_n + \sum_{n=1}^{\infty} k_{cn}v_n\frac{\nabla\times\mathbf{e}_n}{k_{cn}},\\
\nabla\times\mathbf{H} &= \sum_{n=1}^{\infty}\left(-\frac{\partial i_n}{\partial z} + k_{cn}h'_{zn}\right)\mathbf{e}_n + \mathbf{u}_z\sum_{n=1}^{\infty} k_{cn}i_n\frac{\nabla\cdot\mathbf{e}_n}{k_{cn}},
\end{aligned} \tag{2.151}$$

where v_n and i_n are **time-domain modal voltage** and **time-domain modal current** defined by

$$\begin{aligned} v_n(z,t) &= \int_\Omega \mathbf{E}(\mathbf{r},t)\cdot\mathbf{e}_n\, d\Omega, \\ i_n(z,t) &= \int_\Omega \mathbf{H}(\mathbf{r},t)\cdot\mathbf{u}_z\times\mathbf{e}_n\, d\Omega, \end{aligned} \tag{2.152}$$

and e'_{zn} and h'_{zn} are given by

$$\begin{aligned} e'_{zn}(z,t) &= \int_\Omega \mathbf{u}_z\cdot\mathbf{E}(\mathbf{r},t)\frac{\nabla\cdot\mathbf{e}_n}{k_{cn}}\, d\Omega, \\ h'_{zn}(z,t) &= \int_\Omega \mathbf{H}(\mathbf{r},t)\cdot\frac{\nabla\times\mathbf{e}_n}{k_{cn}}\, d\Omega. \end{aligned}$$

Inserting (2.150) and (2.151) into generalized Maxwell equations

$$\begin{aligned} \nabla\times\mathbf{E}(\mathbf{r},t) &= -\mu\frac{\partial\mathbf{H}(\mathbf{r},t)}{\partial t} - \mathbf{J}_m(\mathbf{r},t), \\ \nabla\times\mathbf{H}(\mathbf{r},t) &= \varepsilon\frac{\partial\mathbf{E}(\mathbf{r},t)}{\partial t} + \mathbf{J}(\mathbf{r},t) + \sigma\mathbf{E}(\mathbf{r},t), \end{aligned}$$

and comparing the transverse and longitudinal components, we obtain

$$\begin{aligned} &\frac{\partial^2 v_n^{TEM}}{\partial z^2} - \frac{1}{v^2}\frac{\partial^2 v_n^{TEM}}{\partial t^2} - \sigma\frac{\eta}{v}\frac{\partial v_n^{TEM}}{\partial t} \\ &\quad = \mu\frac{\partial}{\partial t}\int_\Omega \mathbf{J}\cdot\mathbf{e}_n\, d\Omega - \frac{\partial}{\partial z}\int_\Omega \mathbf{J}_m\cdot\mathbf{u}_z\times\mathbf{e}_n\, d\Omega. \end{aligned} \tag{2.153}$$

$$\frac{\partial v_n^{TEM}}{\partial z} = -\mu\frac{\partial i_n^{TEM}}{\partial t} - \int_\Omega \mathbf{J}_m\cdot\mathbf{u}_z\times\mathbf{e}_n\, d\Omega, \tag{2.154}$$

$$\begin{aligned} &\frac{\partial^2 v_n^{TE}}{\partial z^2} - \frac{1}{v^2}\frac{\partial^2 v_n^{TE}}{\partial t^2} - \sigma\frac{\eta}{v}\frac{\partial v_n^{TE}}{\partial t} - k_{cn}^2 v_n^{TE} \\ &\quad = \mu\frac{\partial}{\partial t}\int_\Omega \mathbf{J}\cdot\mathbf{e}_n\, d\Omega - \frac{\partial}{\partial z}\int_\Omega \mathbf{J}_m\cdot\mathbf{u}_z\times\mathbf{e}_n\, d\Omega \\ &\qquad + k_{cn}\int_\Omega \mathbf{u}_z\cdot\mathbf{J}_m\frac{\mathbf{u}_z\cdot\nabla\times\mathbf{e}_n}{k_{cn}}\, d\Omega. \end{aligned} \tag{2.155}$$

$$\frac{\partial v_n^{TE}}{\partial z} = -\mu \frac{\partial i_n^{TE}}{\partial t} - \int\limits_{\Omega} \mathbf{J}_m \cdot \mathbf{u}_z \times \mathbf{e}_n d\Omega, \tag{2.156}$$

$$\begin{aligned} &\frac{\partial^2 i_n^{TM}}{\partial z^2} - \frac{1}{v^2}\frac{\partial^2 i_n^{TM}}{\partial t^2} - \sigma\frac{\eta}{v}\frac{\partial i_n^{TM}}{\partial t} - k_{cn}^2 i_n^{TM} \\ &\quad = -\frac{\partial}{\partial z}\int\limits_{\Omega} \mathbf{J}\cdot\mathbf{e}_n\, d\Omega - k_{cn}\int\limits_{\Omega} \mathbf{u}_z \cdot \mathbf{J}\frac{\nabla\cdot\mathbf{e}_n}{k_{cn}}\, d\Omega \\ &\qquad + \varepsilon\frac{\partial}{\partial t}\int\limits_{\Omega} \mathbf{J}_m \cdot \mathbf{u}_z \times \mathbf{e}_n\, d\Omega + \sigma\int\limits_{\Omega} \mathbf{J}_m \cdot \mathbf{u}_z \times \mathbf{e}_n\, d\Omega. \end{aligned} \tag{2.157}$$

$$\frac{\partial i_n^{TM}}{\partial z} = -\varepsilon\frac{\partial v_n^{TM}}{\partial t} - \sigma v_n^{TM} - \int\limits_{\Omega} \mathbf{J}\cdot\mathbf{e}_n\, d\Omega. \tag{2.158}$$

The excitation problem in the waveguide is now reduced to the solution of a series of inhomogeneous modified Klein–Gordon equations. The modal currents i_n can be determined by the time integration of v_n. Other expansion coefficients can be determined similarly.

2.5.2 Solution of Modified Klein–Gordon Equation

To find the complete solution of the transient fields in the waveguide, we need to solve the modified Klein–Gordon equation. This can be done by the use of retarded Green's function (Geyi, 2006a). The retarded Green's function of the modified Klein–Gordon equation is defined by

$$\begin{aligned} \left(\frac{\partial^2}{\partial z^2} - \frac{1}{v^2}\frac{\partial^2}{\partial t^2} - \sigma\frac{\eta}{v}\frac{\partial}{\partial t} - k_{cn}^2\right) G_n(z,t;z',t') &= -\delta(z-z')\delta(t-t'), \\ G_n(z,t;z',t')|_{t<t'} &= 0. \end{aligned} \tag{2.159}$$

The second equation represents the causality condition. The solution of the above equation is given by (Geyi, 2010)

$$\begin{aligned} G_n(z,t;z',t') = \frac{v}{2}e^{-\gamma(t-t')}&H[(t-t') - |z-z'|/v] \\ &\cdot J_0\left[(k_{cn}^2v^2-\gamma^2)^{1/2}\sqrt{(t-t')^2 - |z-z'|^2/v^2}\right], \end{aligned} \tag{2.160}$$

where $J_0(x)$ is the Bessel function of first kind and $H(x)$ is the unit step function. The retarded Green's function can now be used to solve the

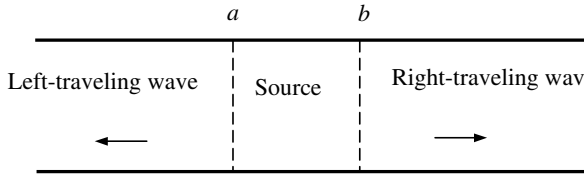


Figure 2.26 Left-traveling wave and right-traveling wave in a waveguide.

modified Klein–Gordon equation with the known source function $f(z,t)$:

$$\left(\frac{\partial^2}{\partial z^2} - \frac{1}{v^2}\frac{\partial^2}{\partial t^2} - \sigma\frac{\eta}{v}\frac{\partial}{\partial t} - k_{cn}^2\right) u_n(z,t) = f(z,t).$$

If the source function $f(z,t)$ is limited in a finite interval (a,b), as shown in Figure 2.26, the solution of the above equation may be expressed, in terms of Green's function, as

$$\begin{aligned} u_n(z,t) = & \int_{-\infty}^{\infty} G_n(z,t;z',t')\frac{\partial u_n(z',t')}{\partial z'}dt'\Bigg|_{z=a}^{b} \\ & - \int_{-\infty}^{\infty} u_n(z',t')\frac{\partial G_n(z,t;z',t')}{\partial z'}dt'\Bigg|_{z=a}^{b} \\ & - \int_{a}^{b}\int_{-\infty}^{\infty} f(z',t')G_n(z,t;z',t')dt'dz', \quad z \in (a,b), \end{aligned} \tag{2.161}$$

where the symmetry of Green's function about z and z' has been used. If we let $a \to -\infty$ and $b \to \infty$, the above expression becomes

$$u_n(z,t) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(z',t')G_n(z,t;z',t')dt'dz', \quad z \in (-\infty,\infty). \tag{2.162}$$

It can be shown that, in the source-free region, we have

$$u_n(z,t) + vJ_0(k_{cn}vt)H(t) * \frac{\partial u_n(z,t)}{\partial z} = 0, \quad z \geq b > 0, \tag{2.163}$$

$$u_n(z,t) - vJ_0(k_{cn}vt)H(t) * \frac{\partial u_n(z,t)}{\partial z} = 0, \quad z \leq a < 0, \tag{2.164}$$

where $*$ denotes the convolution. Both (2.163) and (2.164) are integro-differential equations and are respectively called **right-traveling condition** and **left-traveling condition** of the wave. If the source is turned on at $t = 0$, all the fields must be zero when $t < 0$, Equations (2.163) and (2.164) can be solved by the single-sided Laplace transform and their solutions are

$$u_n(z,t) = u_n\left(b, t - \frac{z-b}{v}\right) - ck_{cn}(z-b) \times \int_0^{t-\frac{z-b}{v}} \frac{J_1\left[k_{cn}v\sqrt{(t-\tau)^2 - (z-b)^2/v^2}\right]}{\sqrt{(t-\tau)^2 - (z-b)^2/v^2}} u_n(b,\tau)d\tau, \quad z \geq b > 0,$$

$$u_n(z,t) = u_n\left(a, t + \frac{z-a}{v}\right) + ck_{cn}(z-a) \times \int_0^{t+\frac{z-a}{v}} \frac{J_1\left[k_{cn}v\sqrt{(t-\tau)^2 - (z-a)^2/v^2}\right]}{\sqrt{(t-\tau)^2 - (z-a)^2/v^2}} u_n(a,\tau)d\tau, \quad z \leq a < 0.$$

Once the input signal is known the output signal after traveling a certain distance in the waveguide can be determined by the above convolution integral.

2.6 Dielectric Waveguides

The theory of dielectric waveguide or optical fiber is the foundation of microwave, millimeter-wave and optical integrated circuits. An optical fiber consists of a core of dielectric material surrounded by a cladding of another dielectric material which has lower refractive index than that of the core. The electromagnetic fields are confined in the core region due to the total internal reflection and the fiber acts as a waveguide. The optical fibers have been widely used in fiber-optic communication, and they carry much more information and travel longer distances than conventional metal wires. Moreover, they are immune to electromagnetic interferences.

2.6.1 Guidance Condition

Let Ω_c denote the cross section of the core region of an arbitrary fiber, and the exterior region (the cladding) be denoted by Ω_∞, as shown in

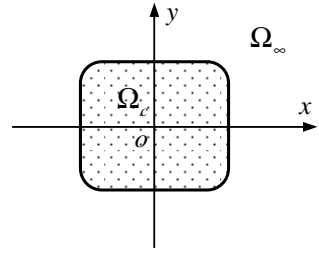


Figure 2.27 An optical fiber.

Figure 2.27. The medium parameters of the fiber are given by μ_0, $\varepsilon_r\varepsilon_0$. The refractive index of the fiber is denoted by $n(\boldsymbol{\rho}) = \sqrt{\varepsilon_r}$, which is a positive function of the transverse coordinates $\boldsymbol{\rho} = (x, y)$ only. We assume that the fiber cladding is homogeneous and extends infinitely in the transverse (x, y)-plane. This assumption is reasonable since the radius of the core is very small compared to the radius of the cladding in practice. The refractive index of the cladding is thus a constant, denoted by $n(\boldsymbol{\rho}) = n_\infty$. If the index n is piecewise constant, the fiber is called a **step-index fiber**. If the index n is a continuous function, the fiber is called a **graded-index fiber**. Assume that the fields have a z dependence of the form $e^{-j\beta z}$

$$\mathbf{E}(\mathbf{r}) = \mathbf{e}(\boldsymbol{\rho})e^{-j\beta z}, \quad \mathbf{H}(\mathbf{r}) = \mathbf{h}(\boldsymbol{\rho})e^{-j\beta z}. \tag{2.165}$$

Introducing these into Maxwell equations, we obtain

$$\begin{aligned} &\nabla_\beta \times \mathbf{h} = j\omega\varepsilon_r\varepsilon_0\mathbf{e}, \quad \nabla_\beta \times \mathbf{e} = -j\omega\mu_0\mathbf{h}, \\ &\nabla_\beta \cdot \varepsilon_r\varepsilon_0\mathbf{e} = 0, \quad \nabla_\beta \cdot \mu_0\mathbf{h} = 0. \end{aligned} \tag{2.166}$$

Here $\nabla_\beta = \nabla_t - j\beta\mathbf{u}_z$. It follows from (2.166) that

$$\begin{aligned} \nabla_\beta \times n^{-2}\nabla_\beta \times \mathbf{h} = k_0^2\mathbf{h}, \quad \nabla_\beta \times \nabla_\beta \times \mathbf{e} = k_0^2 n^2\mathbf{e}, \\ \nabla_\beta \cdot n^2\mathbf{e} = 0, \quad \nabla_\beta \cdot \mathbf{h} = 0, \end{aligned} \tag{2.167}$$

where $k_0 = \omega\sqrt{\mu_0\varepsilon_0}$. A solution of (2.167) is called a **guided mode** of the waveguide if the field is non-trivial and has finite energy. Multiplying the first equation by $\bar{\mathbf{h}}$ and taking integration over the whole (x, y)-plane,

we obtain

$$\begin{aligned} k_0^2 \int\limits_{R^2} |\mathbf{h}|^2 d\Omega &= \int\limits_{R^2} \frac{1}{n^2} (\nabla_\beta \times \mathbf{h}) \cdot (\overline{\nabla_\beta \times \mathbf{h}}) d\Omega \\ &\geq \frac{1}{n_+^2} \int\limits_{R^2} (\nabla_\beta \times \mathbf{h}) \cdot (\overline{\nabla_\beta \times \mathbf{h}}) d\Omega \\ &= \frac{1}{n_+^2} \int\limits_{R^2} (\nabla_\beta \times \nabla_\beta \times \mathbf{h}) \cdot \bar{\mathbf{h}} d\Omega, \end{aligned}$$

where $n_+ = \max\limits_{\boldsymbol{\rho} \in R^2} n(\boldsymbol{\rho})$. From the last equation of (2.88), we obtain

$$\begin{aligned} k_0^2 \int\limits_{R^2} |\mathbf{h}|^2 \, d\Omega &\geq \frac{1}{n_+^2} \int\limits_{R^2} (-\nabla_t^2 \mathbf{h} + \beta^2 \mathbf{h}) \cdot \bar{\mathbf{h}} \, d\Omega \\ &= \frac{1}{n_+^2} \int\limits_{R^2} (|\nabla_t \times \mathbf{h}|^2 + |\nabla_t \cdot \mathbf{h}|^2) d\Omega + \frac{\beta^2}{n_+^2} \int\limits_{R^2} |\mathbf{h}|^2 \, d\Omega. \end{aligned}$$

This is equivalent to

$$\int\limits_{R^2} (|\nabla_t \times \mathbf{h}|^2 + |\nabla_t \cdot \mathbf{h}|^2) d\Omega + (\beta^2 - k_0^2 n_+^2) \int\limits_{R^2} |\mathbf{h}|^2 \, d\Omega \leq 0. \tag{2.168}$$

If $|\beta| \geq k_0 n_+$, the above inequality implies $\mathbf{h} = \mathbf{e} = 0$, i.e., the fiber does not support guided modes in this case. As a result, a necessary condition for the existence of a guided mode is

$$k_0 > \frac{|\beta|}{n_+}. \tag{2.169}$$

In the region Ω_∞, the first equation of (2.167) becomes

$$\nabla_t^2 \mathbf{h} + (k_0^2 n_\infty^2 - \beta^2) \mathbf{h} = 0, \quad \boldsymbol{\rho} \in \Omega_\infty.$$

If $k_0 > |\beta|/n_\infty$ and $\mathbf{h}$ is a guided mode, the above relation implies that $\mathbf{h}$ must be zero from the uniqueness theorem for Helmholtz equation. Therefore, another necessary condition for the existence of a guided mode is

$$k_0 \leq \frac{|\beta|}{n_\infty}. \tag{2.170}$$

Combining (2.169) and (2.170) yields

$$\frac{|\beta|}{n_+} < k_0 \leq \frac{|\beta|}{n_\infty}. \tag{2.171}$$

This implies

$$n_+ > n_\infty, \tag{2.172}$$

which is called the **guidance condition** for the optical fiber.

2.6.2 Circular Optical Fiber

In a cylindrical system, the total fields can be decomposed into a sum of the transverse component and longitudinal component

$$\mathbf{E} = \mathbf{E}_t + \mathbf{u}_z E_z, \quad \mathbf{H} = \mathbf{H}_t + \mathbf{u}_z H_z.$$

If the fields have a z dependence of the form $e^{-j\beta z}$, the transverse fields may be expressed in terms of the longitudinal fields as

$$\begin{aligned} \mathbf{E}_t &= \frac{1}{k_c^2}\left[-j\omega\mu\nabla_t \times (\mathbf{u}_z H_z) - j\beta\nabla_t E_z\right], \\ \mathbf{H}_t &= \frac{1}{k_c^2}\left[j\omega\varepsilon\nabla_t \times (\mathbf{u}_z E_z) - j\beta\nabla_t H_z\right], \end{aligned} \tag{2.173}$$

where $k_c^2 = \omega^2\mu\varepsilon - \beta^2$. The longitudinal components satisfy the two-dimensional Helmholtz equation

$$(\nabla_t^2 + k_c^2)E_z = 0, \quad (\nabla_t^2 + k_c^2)H_z = 0. \tag{2.174}$$

Consider the circular optical fiber shown in Figure 2.28. The core is the circular region of radius a with medium parameters μ_0, $\varepsilon_{r1}\varepsilon_0$. The external

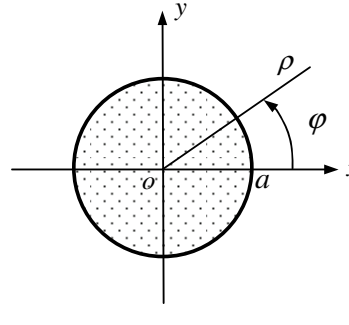


Figure 2.28 Circular optical fiber.

region is the cladding with medium parameters μ_0, $\varepsilon_{r2}\varepsilon_0$. The refractive indices $n_i = \sqrt{\varepsilon_{ri}}$ $(i = 1, 2)$ are assumed to be constants. The solutions of (2.174) in the core region must be finite, and may be written as

$$E_{z1} = A_1 \frac{J_m(k_{c1}\rho)}{J_m(k_{c1}a)} e^{jm\varphi} e^{-j\beta z} = A_1 \frac{J_m(u\rho')}{J_m(u)} e^{jm\varphi} e^{-j\beta z},$$

$$H_{z1} = B_1 \frac{J_m(k_{c1}\rho)}{J_m(k_{c1}a)} e^{jm\varphi} e^{-j\beta z} = B_1 \frac{J_m(u\rho')}{J_m(u)} e^{jm\varphi} e^{-j\beta z},$$

where $k_{c1}^2 = k_0^2 n_1^2 - \beta^2$; $k_0^2 = \omega^2 \mu_0 \varepsilon_0$; $u = k_{c1}a$; $\rho' = \rho/a$; A_1 and B_1 are constants to be determined by boundary conditions. The solutions of (2.174) in the cladding region must decrease as ρ increases to guarantee that the fields are square-integrable (i.e., the energy is finite), and they are given by

$$E_{z2} = A_2 \frac{K_m(\gamma\rho)}{K_m(\gamma a)} e^{jm\varphi} e^{-j\beta z} = A_2 \frac{K_m(v\rho')}{K_m(v)} e^{jm\varphi} e^{-j\beta z},$$

$$H_{z2} = B_2 \frac{K_m(\gamma\rho)}{K_m(\gamma a)} e^{jm\varphi} e^{-j\beta z} = B_2 \frac{K_m(v\rho')}{K_m(v)} e^{jm\varphi} e^{-j\beta z},$$

where K_m are modified Bessel functions of the second kind, and $\gamma^2 = \beta^2 - k_0^2 n_2^2$ and $v = \gamma a$. The transverse field components can then be determined from (2.173). In the core region, the transverse field components are

$$E_{\rho 1} = -j \left(\frac{a}{u}\right)^2 \left[\frac{u\beta J_m'(u\rho')}{aJ_m(u)} A_1 + \frac{j\omega\mu_0 m J_m(u\rho')}{\rho J_n(u)} B_1\right] e^{jm\varphi} e^{-j\beta z},$$

$$E_{\varphi 1} = -j \left(\frac{a}{u}\right)^2 \left[\frac{j\beta m J_m(u\rho')}{\rho J_m(u)} A_1 - \frac{\omega\mu_0 u J_m'(u\rho')}{aJ_m(u)} B_1\right] e^{jm\varphi} e^{-j\beta z},$$

$$H_{\rho 1} = -j \left(\frac{a}{u}\right)^2 \left[\frac{u\beta J_m'(u\rho')}{aJ_m(u)} B_1 - \frac{j\omega\varepsilon_0 n_1^2 m J_m(u\rho')}{\rho J_m(u)} A_1\right] e^{jm\varphi} e^{-j\beta z},$$

$$H_{\varphi 1} = -j \left(\frac{a}{u}\right)^2 \left[\frac{j\beta m J_m(u\rho')}{\rho J_m(u)} B_1 + \frac{\omega\varepsilon_0 n_1^2 u J_m'(u\rho')}{aJ_m(u)} A_1\right] e^{jm\varphi} e^{-j\beta z}.$$

In the cladding region, the transverse field components are

$$E_{\rho 2} = j \left(\frac{a}{v}\right)^2 \left[\frac{v\beta K_m'(v\rho')}{aK_m(v)} A_2 + \frac{j\omega\mu_0 m K_n(v\rho')}{\rho K_m(v)} B_2\right] e^{jm\varphi} e^{-j\beta z},$$

$$E_{\varphi 2} = j \left(\frac{a}{v}\right)^2 \left[\frac{j\beta m K_m(v\rho')}{\rho K_m(v)} A_2 - \frac{\omega\mu_0 v K_m'(v\rho')}{aK_m(v)} B_2\right] e^{jm\varphi} e^{-j\beta z},$$

$$H_{\rho 2} = j\left(\frac{a}{v}\right)^2\left[\frac{v\beta K_m'(v\rho')}{aK_m(v)}B_2 - \frac{j\omega\varepsilon_0 n_2^2 m K_m(v\rho')}{\rho K_m(v)}A_2\right]e^{jm\varphi}e^{-j\beta z},$$

$$H_{\varphi 2} = j\left(\frac{a}{v}\right)^2\left[\frac{j\beta m K_m(v\rho')}{\rho K_m(v)}B_2 + \frac{\omega\varepsilon_0 n_2^2 v K_m'(v\rho')}{aK_m(v)}A_2\right]e^{jm\varphi}e^{-j\beta z}.$$

The boundary conditions at $\rho = a$ require that the tangential fields must be continuous, and this leads to

$$\begin{aligned}
&A_1 = A_2, \quad B_1 = B_2,\\
&-\left(\frac{a}{u}\right)^2\left[\frac{j\beta m}{\rho}A_1 - \frac{\omega\mu_0 u J_m'(u)}{aJ_m(u)}B_1\right]\\
&\qquad = \left(\frac{a}{v}\right)^2\left[\frac{j\beta m}{\rho}A_2 - \frac{\omega\mu_0 v K_m'(v)}{aK_m(v)}B_2\right],\\
&-\left(\frac{a}{u}\right)^2\left[\frac{j\beta m}{\rho}B_1 + \frac{\omega\varepsilon_0 n_1^2 u J_n'(u)}{aJ_m(u)}A_1\right]\\
&\qquad = \left(\frac{a}{v}\right)^2\left[\frac{j\beta m}{\rho}B_2 + \frac{\omega\varepsilon_0 n_2^2 v K_m'(v)}{aK_m(v)}A_2\right].
\end{aligned}$$

These equations can be reduced to

$$A_1\left(\frac{1}{u^2}+\frac{1}{v^2}\right)\frac{j\beta m}{a} - B_1\frac{\omega\mu_0}{a}\left[\frac{1}{u}\frac{J_m'(u)}{J_m(u)} + \frac{1}{v}\frac{K_m'(v)}{K_m(v)}\right] = 0,$$

$$A_1\frac{\omega\varepsilon_0}{a}\left[\frac{n_1^2}{u}\frac{J_m'(u)}{J_m(u)} + \frac{n_2^2}{v}\frac{K_m'(v)}{K_m(v)}\right] + B_1\frac{j\beta m}{a}\left(\frac{1}{u^2}+\frac{1}{v^2}\right) = 0.$$

A non-trivial solution of the above set of equations requires that the determinant of the coefficient matrix vanishes, yielding

$$\left[\frac{1}{u}\frac{J_m'(u)}{J_m(u)} + \frac{1}{v}\frac{K_m'(v)}{K_m(v)}\right]\left[\frac{n_1^2}{u}\frac{J_m'(u)}{J_m(u)} + \frac{n_2^2}{v}\frac{K_m'(v)}{K_m(v)}\right] = \frac{m^2\beta^2}{k_0^2}\left(\frac{1}{u^2}+\frac{1}{v^2}\right),$$

which can be used to determine the propagation constant β. For the guided modes, both k_{c1} and γ must be positive. This requires

$$k_0 n_2 \le \beta \le k_0 n_1. \tag{2.175}$$

When $\beta = k_0 n_2$, we have $\gamma = 0$, which is called the **cut-off condition**. Note that the propagation constant β is not equal to zero when the optical

fiber is at cut-off. This is different from a hollow metal waveguide. If $\gamma < 0$, the fields will radiate in ρ direction, and at same time, they still propagate along z direction. Such field distributions are called **radiation modes**.

2.6.3 Dielectric Slab Waveguide

The dielectric slab waveguide is an infinite planar slab, as shown in Figure 2.29. The whole space is divided into three regions. Region I is the dielectric slab of thickness h with relative dielectric constant ε_r. Regions II and III are free space. Assume that the wave is propagating in the z direction and the fields are independent of x. According to (2.173), the electromagnetic fields can be expressed as

$$\begin{aligned} E_x &= \frac{1}{k_c^2}\left(-j\beta\frac{\partial E_z}{\partial x} - j\omega\mu\frac{\partial H_z}{\partial y}\right), \\ E_y &= \frac{1}{k_c^2}\left(-j\beta\frac{\partial E_z}{\partial y} + j\omega\mu\frac{\partial H_z}{\partial x}\right), \\ H_x &= \frac{1}{k_c^2}\left(j\omega\varepsilon\frac{\partial E_z}{\partial y} - j\beta\frac{\partial H_z}{\partial x}\right), \\ H_y &= \frac{1}{k_c^2}\left(-j\omega\varepsilon\frac{\partial E_z}{\partial x} - j\beta\frac{\partial H_z}{\partial y}\right). \end{aligned} \tag{2.176}$$

Since the fields are independent of x, the above expressions reduce to

$$\text{TE:}\begin{cases} E_x = -j\omega\mu\dfrac{1}{k_c^2}\dfrac{\partial H_z}{\partial y} \\ H_y = -j\beta\dfrac{1}{k_c^2}\dfrac{\partial H_z}{\partial y} \end{cases}, \quad \text{TM:}\begin{cases} E_y = -j\beta\dfrac{1}{k_c^2}\dfrac{\partial E_z}{\partial y} \\ H_x = j\omega\varepsilon\dfrac{1}{k_c^2}\dfrac{\partial E_z}{\partial y} \end{cases}. \tag{2.177}$$

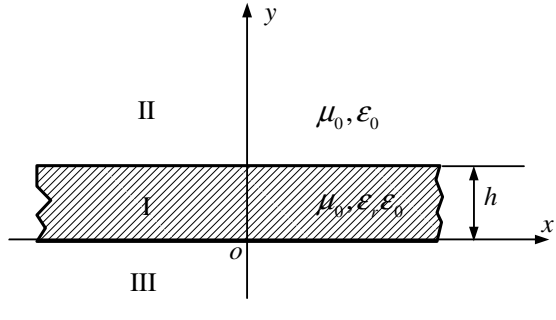


Figure 2.29 Planar slab.

For TE modes, we may let

$$H_z = \begin{cases} Ae^{-\gamma(y-h)}, & y > h \\ Be^{-jk_{c1}y} + Ce^{jk_{c1}y}, & 0 < y < h \\ De^{\gamma y}, & y < 0 \end{cases}$$

where $k_{c1} = \sqrt{\varepsilon_r k_0^2 - \beta^2}$, $\gamma = \sqrt{\beta^2 - k_0^2}$. At the interfaces $y = 0$ and $y = h$, E_x and H_z are continuous. Thus we have

$$A = Be^{-jk_{c1}h} + Ce^{jk_{c1}h}, \quad D = B + C,$$
$$k_{c1}D = j\gamma B - j\gamma C, \quad k_{c1}A = -j\gamma Be^{-jk_{c1}h} + j\gamma Ce^{jk_{c1}h}.$$

It follows that

$$\tan(k_{c1}h - \alpha) = \frac{\gamma}{k_{c1}}, \tag{2.178}$$

where $\alpha = \arctan(\gamma/k_{c1})$. The propagation constant β may be determined from (2.178).

For TM modes, we may write

$$E_z = \begin{cases} Ae^{-\gamma(y-h)}, & y > h \\ Be^{-jk_{c1}y} + Ce^{jk_{c1}y}, & 0 < y < h \\ De^{\gamma y}, & y < 0 \end{cases}.$$

At the interfaces $y = 0$ and $y = h$, H_x and E_z are continuous. Thus

$$A = Be^{-jk_{c1}h} + Ce^{jk_{c1}h}$$
$$D = B + C$$
$$k_{c1}D = j\gamma\varepsilon_r B - j\gamma\varepsilon_r C$$
$$k_{c1}A = -j\gamma\varepsilon_r Be^{-jk_{c1}h} + j\gamma\varepsilon_r Ce^{jk_{c1}h}.$$

From these equations, we obtain

$$\tan(k_{c1}h - \alpha) = \frac{\gamma\varepsilon_r}{k_{c1}}, \tag{2.179}$$

where $\alpha = \arctan(\gamma\varepsilon_r/k_{c1})$. The above equation can be used to find the propagation constant β.

2.7 Microstrip Lines

Microstrip line was first proposed by Grieg and Engelmann (1952), which consists of a conducting strip separated from a ground plane by a dielectric

layer known as the substrate (e.g., FR-4). The microstrip is the building block of many microwave components, such as filters, couplers, power dividers and antennas. These components can be integrated together as a pattern of metallization on the substrate. Compared to the traditional waveguide, the microstrip line is less expensive, lighter and more compact. The disadvantages of microstrip line are lower power handling capacity, higher losses, and are susceptible to interferences due to its openness.

The microstrip line is often used as a TEM transmission line, which is a valid approximation when the frequency is relatively low (e.g., below a few gigahertz). In fact, the microstrip can only support hybrid modes, a combination of TE and TM modes. Figure 2.30(a) shows a standard microstrip line, where region I is the dielectric substrate and region II is the outside. The width of the strip is W and its thickness is assumed to be zero.

The fields in the microstrip can be decomposed into the sum of a transverse component and a longitudinal component

$$\mathbf{E} = \mathbf{E}_t + \mathbf{u}_z E_z, \quad \mathbf{H} = \mathbf{H}_t + \mathbf{u}_z H_z.$$

If the fields are assumed to change according to $e^{-j\beta z}$ in the longitudinal direction, we may write $\nabla = \nabla_t - j\beta \mathbf{u}_z$ and (2.173) applies. Thus

$$E_{xi} = \frac{1}{k_{ci}^2}\left(-j\beta\frac{\partial E_{zi}}{\partial x} - j\omega\mu_i\frac{\partial H_{zi}}{\partial y}\right),$$

$$E_{yi} = \frac{1}{k_{ci}^2}\left(-j\beta\frac{\partial E_{zi}}{\partial y} + j\omega\mu_i\frac{\partial H_{zi}}{\partial x}\right),$$

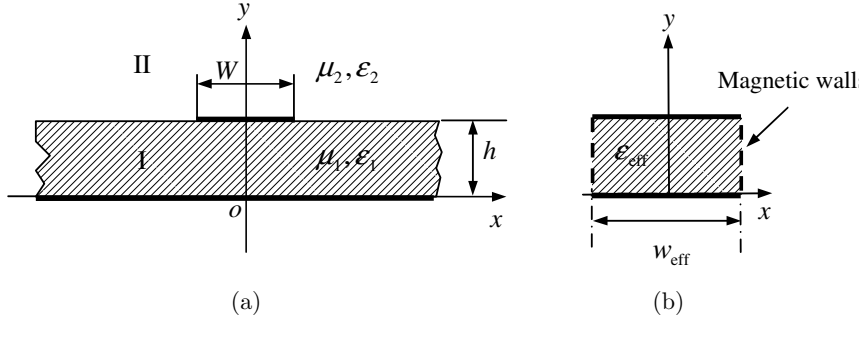


Figure 2.30 (a) Standard microstrip. (b) Waveguide model.

$$
\begin{aligned}
H_{xi} &= \frac{1}{k_{ci}^2}\left(j\omega\varepsilon_i\frac{\partial E_{zi}}{\partial y} - j\beta\frac{\partial H_{zi}}{\partial x}\right),\\
H_{yi} &= \frac{1}{k_{ci}^2}\left(-j\omega\varepsilon_i\frac{\partial E_{zi}}{\partial x} - j\beta\frac{\partial H_{zi}}{\partial y}\right),
\end{aligned}
\tag{2.180}
$$

where $i = \mathrm{I}, \mathrm{II}$, $k_{ci}^2 = k_i^2 - \beta^2$, $k_i^2 = k^2\varepsilon_{ri}$, $k^2 = \omega^2\mu_0\varepsilon_0$. The longitudinal components satisfy the Helmholtz equations

$$
(\nabla_t^2 + k_{ci}^2)E_{zi} = 0, \quad (\nabla_t^2 + k_{ci}^2)H_{zi} = 0. \tag{2.181}
$$

2.7.1 Spectral-Domain Analysis

The microstrip may be investigated by spectral domain method (Itoh and Mittra, 1973). Introducing the Fourier transforms along x-axis

$$
\begin{aligned}
\tilde{E}_{zi}(\varpi, y, z) &= \int_{-\infty}^{+\infty} E_{zi}(x, y, z)e^{-j\varpi x}dx,\\
\tilde{H}_{zi}(\varpi, y, z) &= \int_{-\infty}^{+\infty} H_{zi}(x, y, z)e^{-j\varpi x}dx,
\end{aligned}
$$

into (2.181), we have

$$
\begin{aligned}
\frac{d^2\tilde{E}_{zi}(\varpi, y, z)}{dy^2} - \alpha_i^2\tilde{E}_{zi}(\varpi, y, z) = 0,\\
\frac{d^2\tilde{H}_{zi}(\varpi, y, z)}{dy^2} - \alpha_i^2\tilde{H}_{zi}(\varpi, y, z) = 0,
\end{aligned}
\tag{2.182}
$$

where $\alpha_i^2 = \varpi^2 - k_{ci}^2$. The Fourier transforms of (2.180) are

$$
\begin{aligned}
\tilde{E}_{xi} &= \frac{1}{k_{ci}^2}\left(\varpi\beta\tilde{E}_{zi} - j\omega\mu_i\frac{\partial\tilde{H}_{zi}}{\partial y}\right),\\
\tilde{E}_{yi} &= \frac{1}{k_{ci}^2}\left(-j\beta\frac{\partial\tilde{E}_{zi}}{\partial y} - \varpi\omega\mu_i\tilde{H}_{zi}\right),\\
\tilde{H}_{xi} &= \frac{1}{k_{ci}^2}\left(j\omega\varepsilon_i\frac{\partial\tilde{E}_{zi}}{\partial y} + \varpi\beta\tilde{H}_{zi}\right),\\
\tilde{H}_{yi} &= \frac{1}{k_{ci}^2}\left(\varpi\omega\varepsilon_i\tilde{E}_{zi} - j\beta\frac{\partial\tilde{H}_{zi}}{\partial y}\right).
\end{aligned}
\tag{2.183}
$$

The solutions of (2.182) may be written as

$$\begin{aligned}
\tilde{E}_{z1}(\varpi, y, z) &= A_e \mathrm{sh}\alpha_1 y e^{-j\beta z}, \\
\tilde{E}_{z2}(\varpi, y, z) &= B_e e^{-\alpha_2(y-h)} e^{-j\beta z}, \\
\tilde{H}_{z1}(\varpi, y, z) &= A_h \mathrm{ch}\alpha_1 y e^{-j\beta z}, \\
\tilde{H}_{z2}(\varpi, y, z) &= B_h e^{-\alpha_2(y-h)} e^{-j\beta z}.
\end{aligned} \tag{2.184}$$

From (2.183) and (2.184), we may obtain the field expressions in each region. In region I, we have

$$\begin{aligned}
\tilde{E}_{x1}(\varpi, y, z) &= \frac{1}{k_{c1}^2} \left(\varpi\beta A_e \mathrm{sh}\alpha_1 y - j\omega\mu_1\alpha_1 A_h \mathrm{sh}\alpha_1 y\right) e^{-j\beta z}, \\
\tilde{E}_{y1}(\varpi, y, z) &= \frac{1}{k_{c1}^2} \left(-j\beta A_e \alpha_1 \mathrm{ch}\alpha_1 y - \varpi\omega\mu_1 A_h \mathrm{ch}\alpha_1 y\right) e^{-j\beta z}, \\
\tilde{H}_{x1}(\varpi, y, z) &= \frac{1}{k_{c1}^2} \left(j\omega\varepsilon_1\alpha_1 A_e \mathrm{ch}\alpha_1 y + \varpi\beta A_h \mathrm{ch}\alpha_1 y\right) e^{-j\beta z}, \\
\tilde{H}_{y1}(\varpi, y, z) &= \frac{1}{k_{c1}^2} \left(\varpi\omega\varepsilon_1 A_e \mathrm{sh}\alpha_1 y - j\beta A_h \alpha_1 \mathrm{sh}\alpha_1 y\right) e^{-j\beta z}.
\end{aligned}$$

In region II, we have

$$\begin{aligned}
\tilde{E}_{x2}(\varpi, y, z) &= \frac{1}{k_{c2}^2} \left(\varpi\beta B_e e^{-\alpha_2(y-h)} + j\omega\mu_2\alpha_2 B_h e^{-\alpha_2(y-h)}\right) e^{-j\beta z}, \\
\tilde{E}_{y2}(\varpi, y, z) &= \frac{1}{k_{c2}^2} \left(j\beta\alpha_2 B_e e^{-\alpha_2(y-h)} - \varpi\omega\mu_2 B_h e^{-\alpha_2(y-h)}\right) e^{-j\beta z}, \\
\tilde{H}_{x2}(\varpi, y, z) &= \frac{1}{k_{c2}^2} \left(-j\omega\varepsilon_2\alpha_2 B_e e^{-\alpha_2(y-h)} + \varpi\beta B_h e^{-\alpha_2(y-h)}\right) e^{-j\beta z}, \\
\tilde{H}_{y2}(\varpi, y, z) &= \frac{1}{k_{c2}^2} \left(\varpi\omega\varepsilon_2 B_e e^{-\alpha_2(y-h)} + j\beta\alpha_2 B_h e^{-\alpha_2(y-h)}\right) e^{-j\beta z}.
\end{aligned}$$

At the interface $y = h$, the fields must satisfy the boundary conditions

$$\begin{aligned}
\tilde{E}_{x1}(\varpi, h, z) &= \tilde{E}_{x2}(\varpi, h, z) \equiv \tilde{E}_x(\varpi, h) e^{-j\beta z}, \\
\tilde{E}_{z1}(\varpi, h, z) &= \tilde{E}_{z2}(\varpi, h, z) \equiv \tilde{E}_z(\varpi, h) e^{-j\beta z}, \\
\tilde{H}_{x1}(\varpi, h, z) - \tilde{H}_{x2}(\varpi, h, z) &= \tilde{J}_z(\varpi) e^{-j\beta z}, \\
\tilde{H}_{z1}(\varpi, h, z) - \tilde{H}_{z2}(\varpi, h, z) &= -\tilde{J}_x(\varpi) e^{-j\beta z},
\end{aligned} \tag{2.185}$$

where $\tilde{J}_x(\varpi)e^{-j\beta z}$, $\tilde{J}_z(\varpi)e^{-j\beta z}$ are the current distribution on the conducting strip. From (2.185) we obtain

$$\begin{aligned}\tilde{G}_{11}\tilde{J}_x(\varpi)+\tilde{G}_{12}\tilde{J}_z(\varpi)&=\tilde{E}_x(\varpi,h)\\ \tilde{G}_{21}\tilde{J}_x(\varpi)+\tilde{G}_{22}\tilde{J}_z(\varpi)&=\tilde{E}_z(\varpi,h)\end{aligned}\tag{2.186}$$

where $\tilde{G}_{ij}(i,j=1,2)$ are the functions of ϖ and the propagation constant β. The terms $\tilde{E}_x(\varpi,h)$ and $\tilde{E}_z(\varpi,h)$ on the right-hand sides of (2.186) can be eliminated by using Galerkin's method. The current on the strip may be expanded as follows

$$\tilde{J}_x(\varpi)=\sum_{m=1}^{M}c_m\tilde{J}_{xm}(\varpi),\quad \tilde{J}_z(\varpi)=\sum_{m=1}^{M}d_m\tilde{J}_{zm}(\varpi),$$

where M is a positive integer; $\tilde{J}_{xm}(\varpi)$ and $\tilde{J}_{zm}(\varpi)$ are the basis functions. Substituting these into (2.186), and taking the inner product of $\tilde{J}_{xn}(\varpi)$, $\tilde{J}_{xn}(\varpi)(n=1,2,\ldots,M)$ with (2.186), we obtain

$$\begin{cases}\displaystyle\sum_{m=1}^{M}K_{nm}^{(x,x)}c_m+\sum_{m=1}^{M}K_{nm}^{(x,z)}d_m=0\\ \displaystyle\sum_{m=1}^{M}K_{nm}^{(z,x)}c_m+\sum_{m=1}^{M}K_{nm}^{(z,z)}d_m=0\end{cases},\quad n=1,2,\ldots,M,\tag{2.187}$$

where

$$\begin{aligned}K_{nm}^{(x,x)}&=\int_{-\infty}^{\infty}\tilde{J}_{xn}(\varpi)\tilde{G}_{11}\tilde{J}_{xm}(\varpi)d\varpi,\\ K_{nm}^{(x,z)}&=\int_{-\infty}^{\infty}\tilde{J}_{xn}(\varpi)\tilde{G}_{12}\tilde{J}_{zm}(\varpi)d\varpi,\\ K_{nm}^{(z,x)}&=\int_{-\infty}^{\infty}\tilde{J}_{zn}(\varpi)\tilde{G}_{21}\tilde{J}_{xm}(\varpi)d\varpi,\\ K_{nm}^{(z,z)}&=\int_{-\infty}^{\infty}\tilde{J}_{zn}(\varpi)\tilde{G}_{22}\tilde{J}_{zm}(\varpi)d\varpi.\end{aligned}$$

In deriving (2.187), the Parseval identity

$$\int_{-\infty}^{\infty} \tilde{J}_{xn}(\varpi)\tilde{E}_x(\varpi,h)d\varpi = 2\pi \int_{-\infty}^{\infty} J_{xn}(x)E_x(x,h)dx,$$

has been used. Since the current distribution vanishes except on the strip while the tangential electric fields vanish on the strip, the right-hand side of the above identity is zero. The propagation constant β may be determined from (2.187) by requiring that the determinant of the coefficient matrix is zero.

2.7.2 Closed Form Formulae for Microstrip Lines

In the following, we assume $\mu_1 = \mu_2 = \mu_0$ and $\varepsilon_1 = \varepsilon_r\varepsilon_0, \varepsilon_2 = \varepsilon_0$ (see Figure 2.30). The **effective relative dielectric constant** $\varepsilon_{r,\text{eff}}$ may be introduced for the microstrip line such that its phase velocity and propagation constant are the same as the microstrip line filled with an equivalent homogeneous medium with relative dielectric constant $\varepsilon_{r,\text{eff}}$ (i.e., $\mu_1 = \mu_2 = \mu_0$, $\varepsilon_1 = \varepsilon_2 = \varepsilon_{r,\text{eff}}\varepsilon_0$), and can be expressed as

$$v_p = \frac{c}{\sqrt{\varepsilon_{r,\text{eff}}}}, \quad \beta = k_0\sqrt{\varepsilon_{r,\text{eff}}}, \tag{2.188}$$

where $c = 1/\sqrt{\mu_0\varepsilon_0}$, $k_0 = \omega\sqrt{\mu_0\varepsilon_0}$.

The closed form formulae for microstrip line parameters can be obtained by approximate analytic solutions such as conformal mapping (Wheeler, 1965), along with empirical adjustment of various numerical constants in the analytic solutions.

2.7.2.1 *Analysis Formulae*

Given the dimensions of the microstrip line, the characteristic impedance and effective relative dielectric constant are given by (Owens, 1976)

$$Z_0 = \begin{cases} \dfrac{119.9}{\sqrt{2(\varepsilon_r+1)}}\left\{H_1 - \dfrac{\varepsilon_r - 1}{2(\varepsilon_r+1)}\left(\ln\dfrac{\pi}{2} + \dfrac{1}{\varepsilon_r}\ln\dfrac{4}{\pi}\right)\right\}, & W < 2h, \\ \dfrac{119.9\pi}{\sqrt{\varepsilon_r}}\left\{\dfrac{W}{h} + \dfrac{2\ln 4}{\pi} + \dfrac{\varepsilon_r - 1}{\varepsilon_r^2}\cdot\dfrac{2}{\pi}\ln\left(\dfrac{e\pi^2}{16}\right) + \dfrac{\varepsilon_r+1}{\pi\varepsilon_r}H_2\right\}^{-1}, & W \geq 2h, \end{cases} \tag{2.189}$$

$$\varepsilon_{r,\text{eff}} = \begin{cases} \dfrac{\varepsilon_r+1}{2}\left[1-\dfrac{\varepsilon_r-1}{2H'(\varepsilon_r+1)}\left(\ln\dfrac{\pi}{2}+\dfrac{1}{\varepsilon_r}\cdot\ln\dfrac{4}{\pi}\right)\right]^{-2}, & W<h \\ \dfrac{\varepsilon_r+1}{2}+\dfrac{\varepsilon_r-1}{2}\left(1+\dfrac{10h}{W}\right)^{-0.555}, & W>h, \end{cases} \tag{2.190}$$

where

$$H_1 = \ln\left[\frac{4h}{W}+\sqrt{\left(\frac{4h}{W}\right)^2+2}\right], \quad H_2 = \ln\frac{e\pi}{2}+\ln\left(\frac{W}{2h}+0.94\right).$$

2.7.2.2 *Synthesis Formulae*

Given the characteristic impedance Z_0 and relative dielectric constant ε_r, the ratio W/h can be found as (Owens, 1976)

$$\frac{W}{h} = \begin{cases} \dfrac{8}{e^A-2e^{-A}}, & W/h<2, \\ \dfrac{2}{\pi}\left\{B-1-\ln(2B-1)+\dfrac{\varepsilon_r-1}{2\varepsilon_r}\left[\ln(B-1)+0.293-\dfrac{0.517}{\varepsilon_r}\right]\right\}, & W/h>2, \end{cases}$$

where

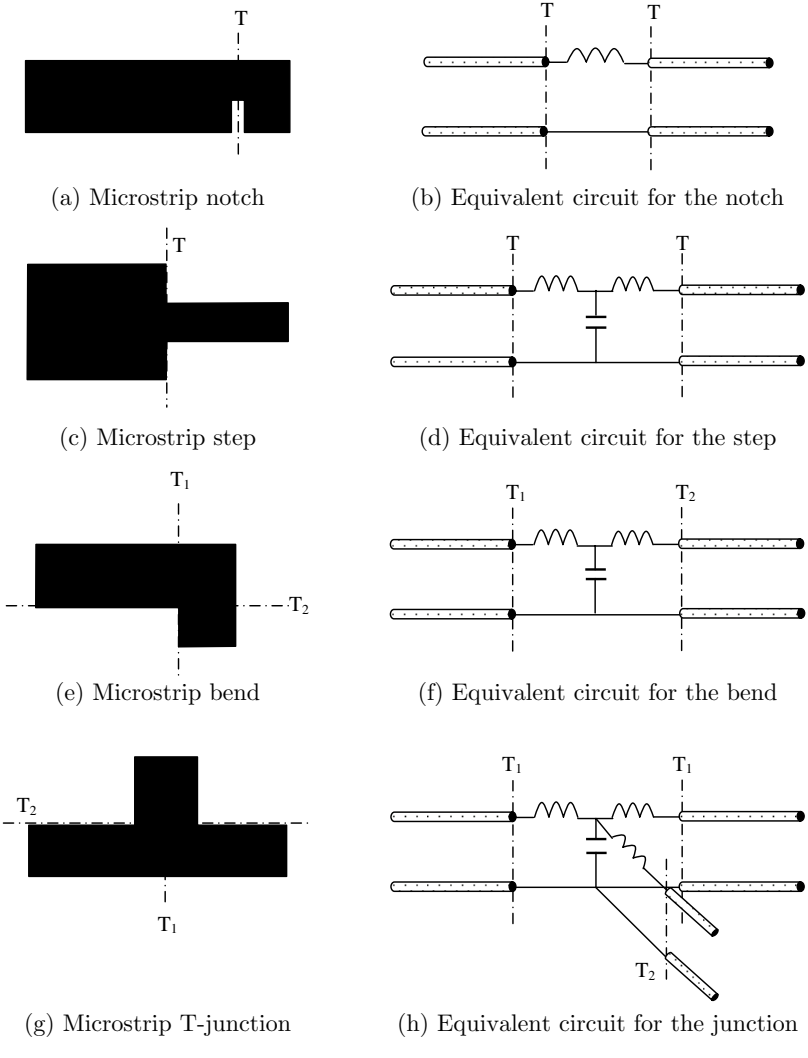
$$A = \frac{Z_0\sqrt{2(\varepsilon_r+1)}}{119.9}+\frac{\varepsilon_r-1}{2(\varepsilon_r+1)}\left(\ln\frac{\pi}{2}+\frac{1}{\varepsilon_r}\cdot\ln\frac{4}{\pi}\right), \quad B = \frac{377\pi}{2Z_0\sqrt{\varepsilon_r}}.$$

2.7.3 Microstrip Discontinuities

Microstrip discontinuities refer to the discontinuity structures in the strip conductor, such as open ends, gaps, notches, steps, bends, and T-junctions, crossings, and so on, and they are often used in microwave integrated circuits. Several typical discontinuities are shown in Figure 2.31. The analysis of various microstrip discontinuities can be carried out by analytical techniques such as static approximations and waveguide models, or numerical methods.

2.7.3.1 *Waveguide Models*

The waveguide models are applicable to a number of microstrip discontinuities (Menzel and Wolff, 1977). The waveguide model for a microstrip



(a) Microstrip notch (b) Equivalent circuit for the notch

(c) Microstrip step (d) Equivalent circuit for the step

(e) Microstrip bend (f) Equivalent circuit for the bend

(g) Microstrip T-junction (h) Equivalent circuit for the junction

Figure 2.31 Typical microstrip discontinuities.

line consists of a parallel plate waveguide of width w_{eff} and height h with magnetic side walls, as shown in Figure 2.30(b). It is filled with a dielectric medium of the relative dielectric constant $\varepsilon_{r,\text{eff}}$ which may be determined by (2.190) or by (Hammerstad and Jensen, 1980)

$$\varepsilon_{r,\text{eff}} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2} \frac{1}{\sqrt{1 + 12h/w}}.$$

As an approximation, the effective width w_{eff} may be assumed to be frequency independent and was given by Wheeler (1965).

If the height of the waveguide model is small, the fields may be assumed to be independent of the y coordinate. Considering the boundary conditions that the tangential fields must satisfy, the waveguide only supports TEM mode and TE_{n0} modes, which have non-vanishing components E_y, H_x and H_z. The longitudinal magnetic field that satisfies the boundary conditions may be written as

$$h_{zn} = \sqrt{\frac{\varepsilon_n}{w_{\text{eff}} h}} \left(\frac{n\pi}{w_{\text{eff}}}\right)^{-1} \sin\left(\frac{n\pi}{w_{\text{eff}}} x\right), \tag{2.191}$$

where $\varepsilon_n = \begin{cases} 1, & m = 0 \\ 2, & m \geq 1 \end{cases}$. It follows from (2.17) that the transverse electromagnetic fields can be expressed as

$$\begin{aligned} \mathbf{E}_t &= \sum_{n=1}^{\infty} \mathbf{e}_n (A_n e^{-j\beta_n z} + B_n e^{j\beta_n z}) \\ \mathbf{H}_t &= \sum_{n=1}^{\infty} \frac{1}{Z_{wn}} \mathbf{u}_z \times \mathbf{e}_n (A_n e^{-j\beta_n z} - B_n e^{j\beta_n z}) \end{aligned} \tag{2.192}$$

where $\mathbf{e}_n = (\mathbf{u}_z \times \nabla h_{zn})/k_{cn}$. Similar to the study of diaphragm problem, the field expressions (2.192) can be used to analyze the microstrip discontinuities shown in Figure 2.31 by matching the transverse fields at the interfaces between different regions.

2.7.3.2 *Method of Green's Function*

A microstrip circuit may be considered as a planar circuit in the sense that one of its dimensions is much smaller than a wavelength. Figure 2.32 shows an arbitrary discontinuity (a metallic patch) with connecting ports. Similar to the waveguide models, all sides of the discontinuity region will be enclosed by magnetic walls except at the reference plane of the ports. If the fields are assumed to be independent of z, we have $E_x = E_y = 0$. Thus the electric field satisfies

$$\left(\nabla_t^2 + k^2\right) E_z(\boldsymbol{\rho}) = 0, \quad k^2 = \omega^2 \mu \varepsilon, \tag{2.193}$$

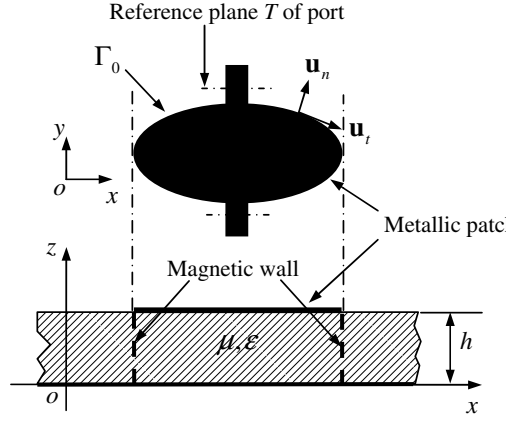


Figure 2.32 Microstrip discontinuity.

where $\nabla_t^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$, $\boldsymbol{\rho} = (x, y)$. The magnetic field can be expressed as

$$\mathbf{H} = \frac{1}{j\omega\mu}\left(-\frac{\partial E_z}{\partial y}\mathbf{u}_x + \frac{\partial E_z}{\partial x}\mathbf{u}_y\right),$$

and the current induced on the metallic patch is

$$\begin{aligned}\mathbf{J} = -\mathbf{u}_z \times \mathbf{H} &= \frac{1}{j\omega\mu}\left(\frac{\partial E_z}{\partial x}\mathbf{u}_x + \frac{\partial E_z}{\partial y}\mathbf{u}_y\right) \\ &= \frac{1}{j\omega\mu}\nabla E_z = \frac{1}{j\omega\mu}\left(\frac{\partial E_z}{\partial t}\mathbf{u}_t + \frac{\partial E_z}{\partial n}\mathbf{u}_n\right).\end{aligned}$$

where $\mathbf{u}_t$ and $\mathbf{u}_n$ are the unit tangent vector and unit outward normal of the boundary of the metallic patch. Let Γ denote the boundary of the metallic patch, and Γ_0 the portion of Γ less the intersections of the reference planes with the ports. Then the normal component of the current on Γ_0 must be zero

$$\left.\frac{\partial E_z}{\partial n}\right|_{\Gamma_0} = 0. \tag{2.194}$$

One may introduce a Green's function G

$$\begin{cases} (\nabla_t^2 + k^2)G(\boldsymbol{\rho}, \boldsymbol{\rho}') = -\delta(\boldsymbol{\rho} - \boldsymbol{\rho}') \\ \left.\dfrac{\partial G}{\partial n}\right|_{\Gamma} = 0 \end{cases}. \tag{2.195}$$

From the Green's identity, the following integral equation can be obtained

$$E_z(\boldsymbol{\rho}) = \int_{\Gamma} G(\boldsymbol{\rho}, \boldsymbol{\rho}') \frac{\partial E_z(\boldsymbol{\rho}')}{\partial n(\boldsymbol{\rho}')} d\Gamma(\boldsymbol{\rho}').$$

Taking the boundary condition (2.194) into account, we may write the above equation as

$$E_z(\boldsymbol{\rho}) = j\omega\mu \sum_{i=1}^{m} \int_{T_i} G(\boldsymbol{\rho}, \boldsymbol{\rho}') J_i(\boldsymbol{\rho}') d\Gamma(\boldsymbol{\rho}'),$$

where $J_i(\boldsymbol{\rho}) = \frac{1}{j\omega\mu} \frac{\partial E_z(\boldsymbol{\rho})}{\partial n(\boldsymbol{\rho})}$ denotes the current density flowing away from the discontinuity at port i ($i = 1, 2, \ldots, m$) and m is the number of ports. If the widths of the ports, denoted by W_i ($i = 1, 2, \ldots, m$), are small, the current density $J_i(\boldsymbol{\rho})$ may be considered to be constant, denoted by $\tilde{J}_i$. The distributed voltage at port j is defined by $V_i(\boldsymbol{\rho}) = -hE_z(\boldsymbol{\rho})$, and can thus be written as

$$V_j(\boldsymbol{\rho}) = -j\omega\mu h \sum_{i=1}^{m} \tilde{J}_i \int_{T_i} G(\boldsymbol{\rho}, \boldsymbol{\rho}') d\Gamma(\boldsymbol{\rho}'). \tag{2.196}$$

The current flowing into port i is given by

$$\tilde{I}_i = \int_{T_i} J_i(\boldsymbol{\rho}') d\Gamma(\boldsymbol{\rho}') = -\tilde{J}_i W_i. \tag{2.197}$$

The voltage $\tilde{V}_j$ at port j may be defined as the average of the distributed voltage

$$\tilde{V}_j = \frac{1}{W_j} \int_{T_j} V_j(\boldsymbol{\rho}') d\Gamma(\boldsymbol{\rho}'). \tag{2.198}$$

It follows from (2.196), (2.197) and (2.198) that

$$\tilde{V}_j = \sum_{i=1}^{m} \tilde{I}_i \frac{j\omega\mu h}{W_i W_j} \int_{T_j} \int_{T_i} G(\boldsymbol{\rho}, \boldsymbol{\rho}') d\Gamma(\boldsymbol{\rho}) d\Gamma(\boldsymbol{\rho}'). \tag{2.199}$$

The impedance matrix elements for the discontinuity are thus given by

$$Z_{ij} = \frac{j\omega\mu h}{W_i W_j} \int_{T_j} \int_{T_i} G(\boldsymbol{\rho}, \boldsymbol{\rho}') d\Gamma(\boldsymbol{\rho}) d\Gamma(\boldsymbol{\rho}'). \tag{2.200}$$

Remark 2.6: Microwave circuits can be integrated in either hybrid or monolithic form. In a hybrid integrated circuit, the circuit interconnections are formed by microstrip lines deposited on an insulating substrate or printed circuit board (PCB) while the active devices and passive components in the circuit are attached to the substrate. In a microwave monolithic integrated circuit (MMIC), active devices, passive components and microstrip lines are fabricated in a single block of semiconductor material. □

2.8 Waveguide with Lossy Walls

Our previous discussions are based on the assumption that the waveguide wall is a perfect conductor. In practice, the waveguide wall has finite conductivity σ and thus has heat loss, which causes the attenuation of electromagnetic wave as it propagates in the waveguide. If the waveguide wall is smooth, the surface impedance Z_s may be introduced such that the tangential electric field and the tangential magnetic field on the wall are related by

$$\mathbf{u}_n \times \mathbf{E} = Z_s \mathbf{u}_n \times (\mathbf{u}_n \times \mathbf{H}). \tag{2.201}$$

In most cases, the surface impedance is given by

$$Z_s = \frac{1+j}{\sigma \delta_s}, \tag{2.202}$$

where $\delta_s = (2/\omega\mu\sigma)^{1/2}$ is the skin depth. At the surface, the tangential electric field must have a tangential component equal to $Z_s \mathbf{J}_s$, where $\mathbf{J}_s = \mathbf{u}_n \times \mathbf{H}$ is the surface current density. This implies that an axial component of electric field must be present, which gives rise to a component of Poynting vector directed into the conductor and caused power loss in the conductor. Substituting the field decompositions

$$\begin{aligned} \mathbf{E}(\mathbf{r}) &= [\mathbf{e}(\boldsymbol{\rho}) + \mathbf{u}_z e_z(\boldsymbol{\rho})] e^{-\gamma z}, \\ \mathbf{H}(\mathbf{r}) &= [\mathbf{h}(\boldsymbol{\rho}) + \mathbf{u}_z h_z(\boldsymbol{\rho})] e^{-\gamma z}, \end{aligned} \tag{2.203}$$

into Maxwell equations

$$\begin{aligned} \nabla \times \mathbf{E}(\mathbf{r}) &= -j\omega\mu \mathbf{H}(\mathbf{r}), \\ \nabla \times \mathbf{H}(\mathbf{r}) &= j\omega\varepsilon \mathbf{E}(\mathbf{r}), \end{aligned}$$

we may obtain

$$\begin{aligned}
&\nabla \times \mathbf{e} = -j\omega\mu \mathbf{u}_z h_z, \quad \nabla \times \mathbf{h} = j\omega\varepsilon \mathbf{u}_z e_z, \\
&\gamma \mathbf{u}_z \times \mathbf{e} + \mathbf{u}_z \times \nabla e_z = j\omega\mu \mathbf{h}, \\
&\gamma \mathbf{u}_z \times \mathbf{h} + \mathbf{u}_z \times \nabla h_z = -j\omega\varepsilon \mathbf{e}, \\
&\nabla \cdot \mathbf{e} = \gamma e_z, \quad \nabla \cdot \mathbf{h} = \gamma h_z.
\end{aligned} \tag{2.204}$$

It follows from the above equations that the transverse electric field satisfies

$$\nabla \times \nabla \times \mathbf{e} - \nabla(\nabla \cdot \mathbf{e}) - (k^2 + \gamma^2)\mathbf{e} = 0, \quad \rho \in \Omega. \tag{2.205}$$

From (2.201), we obtain

$$\mathbf{u}_t \cdot \mathbf{e} = Z_s h_z, \quad e_z = -Z_s \mathbf{u}_t \cdot \mathbf{h}, \quad \rho \in \Gamma, \tag{2.206}$$

where $\mathbf{u}_t$ denotes the unit tangent vector along the boundary Γ of the cross section of the waveguide (see Figure 2.33). Making use of (2.204), (2.206) can be written as

$$\begin{aligned}
\mathbf{u}_t \cdot \mathbf{e} &= -\frac{Z_s}{j\omega\mu} \mathbf{u}_z \cdot \nabla \times \mathbf{e}, \quad \rho \in \Gamma \\
\nabla \cdot \mathbf{e} &= -\frac{Z_s}{j\omega\mu} \mathbf{u}_t \cdot (\gamma^2 \mathbf{u}_z \times \mathbf{e} + \gamma \mathbf{u}_z \times \nabla e_z) \\
&= \frac{Z_s}{j\omega\mu} \mathbf{u}_n \cdot (\gamma^2 \mathbf{e} + \nabla\nabla \cdot \mathbf{e}) \\
&= \frac{Z_s}{j\omega\mu} \mathbf{u}_n \cdot (\nabla \times \nabla \times \mathbf{e} - k^2 \mathbf{e}), \quad \rho \in \Gamma.
\end{aligned} \tag{2.207}$$

Equation (2.205) and the boundary conditions (2.207) constitute an eigenvalue problem. The solution of this eigenvalue problem can be

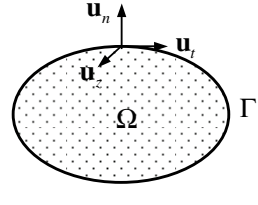


Figure 2.33 Cross section of waveguide.

expressed as a linear combination of the vector modal functions derived from (2.3) as follows

$$\mathbf{e} = \sum_{i=1}^{\infty} a_i \mathbf{e}_i, \tag{2.208}$$

where $\mathbf{e}_i$ satisfy

$$\nabla \times \nabla \times \mathbf{e}_i - \nabla(\nabla \cdot \mathbf{e}_i) - (k^2 + \gamma_i^2)\mathbf{e}_i = 0, \quad \rho \in \Omega \tag{2.209}$$

where $\gamma_i^2 = k_{ci}^2 - k^2$. Multiplying the above equation by $\mathbf{e}$ and (2.205) by $\mathbf{e}_n$, subtracting and integrating the resultant equations over Ω and converting the surface integral into line integral, we obtain

$$\begin{aligned}(\gamma^2 - \gamma_n^2)a_n = -\frac{Z_s}{j\omega\mu} \sum_{i=1}^{\infty} a_i \int_{\Gamma} [&(\nabla \times \mathbf{e}_n) \cdot (\nabla \times \mathbf{e}_i) \\ &-(\mathbf{u}_n \cdot \mathbf{e}_n)\mathbf{u}_n \cdot (\nabla \times \nabla \times \mathbf{e}_i - k^2 \mathbf{e}_i)] d\Gamma, \quad n = 1, 2, 3, \ldots .\end{aligned} \tag{2.210}$$

This is an algebraic equation of infinite dimension. To ensure a non-zero solution the determinant of the coefficient matrix of (2.210) must be zero, which gives an infinite number of solutions for γ^2. To every γ^2, there corresponds a set of coefficients a_i, which can be used to determine the modal solution $\mathbf{e}$ for the lossy waveguide through (2.208). The above procedure is called the **method of perturbation**.

When the loss is present, we may write $\gamma = \alpha + j\beta$, where α is the attenuation constant and β is the phase constant. Since the loss is present, the power P in the line must decrease according to a factor $e^{-2\alpha z}$. Thus the power loss per unit length is given by

$$P_{\text{loss}} = -\frac{\partial P}{\partial z} = 2\alpha P. \tag{2.211}$$

The power loss per unit length due to the surface impedance is

$$P_{\text{loss}} = \frac{1}{2}\text{Re}\, Z_s \int_{\Gamma} \mathbf{J}_s \cdot \bar{\mathbf{J}}_s \, d\Gamma = \frac{1}{2}\text{Re}\, Z_s \int_{\Gamma} (\mathbf{u}_n \times \mathbf{H}) \cdot (\mathbf{u}_n \times \bar{\mathbf{H}}) d\Gamma. \tag{2.212}$$

As a first-order perturbation, we may use the field for the loss-free case, still denoted as $\mathbf{H}$, to evaluate the above integral. Thus (2.212) can be written as

$$P_{\text{loss}} = \frac{1}{2}\text{Re}\, Z_s \int_{\Gamma} \mathbf{H} \cdot \bar{\mathbf{H}} \, d\Gamma. \tag{2.213}$$

Similarly, the power propagated along the line can be evaluated using loss-free fields as follows

$$P = \frac{1}{2}\mathrm{Re}\int_{\Omega} \mathbf{E}\times\bar{\mathbf{H}}\cdot\mathbf{u}_z\, d\Omega. \tag{2.214}$$

Equations (2.211), (2.213) and (2.214) can be used to determine the attenuation constant.

Example 2.1 (Lossy coaxial line)**:** Let the conductors of the coaxial line in Figure 2.3(b) have finite conductivity σ. For the loss-free case, the fields are given by (2.23). The power propagated in the line is

$$P = \frac{1}{2}\mathrm{Re}\int_0^{2\pi}\int_a^b \mathbf{E}\times\bar{\mathbf{H}}\cdot\mathbf{u}_z\rho\, d\rho\, d\varphi = \frac{\pi V_0^2}{\eta\ln(b/a)}. \tag{2.215}$$

The power loss from the finite conductivity is

$$\begin{aligned} P_{\text{loss}} &= \frac{1}{2}\mathrm{Re}\, Z_s \int_{\Gamma_1+\Gamma_2} \mathbf{H}\cdot\bar{\mathbf{H}}\, d\Gamma \\ &= \frac{1}{2}\mathrm{Re}\, Z_s \frac{V_0^2}{\eta^2[\ln(b/a)]^2}\int_0^{2\pi}\left(\frac{1}{a}+\frac{1}{b}\right)d\varphi \\ &= \mathrm{Re}\, Z_s \frac{\pi V_0^2}{\eta^2[\ln(b/a)]^2}\frac{a+b}{ab}. \end{aligned} \tag{2.216}$$

The attenuation constant is given by

$$\alpha = \frac{P_{\text{loss}}}{2P} = \frac{\mathrm{Re}\, Z_s}{2\eta\ln(b/a)}\frac{a+b}{ab}. \quad \square \tag{2.217}$$

2.9 Periodic Structures

Periodic structures refer to the structures either with periodically electrical properties or periodic boundary conditions, and they have wide applications in electronic devices, microwave circuits and antennas. A periodic structure may be considered as a cascade of identical discontinuity (called **elementary structure**) as illustrated in Figure 2.34.

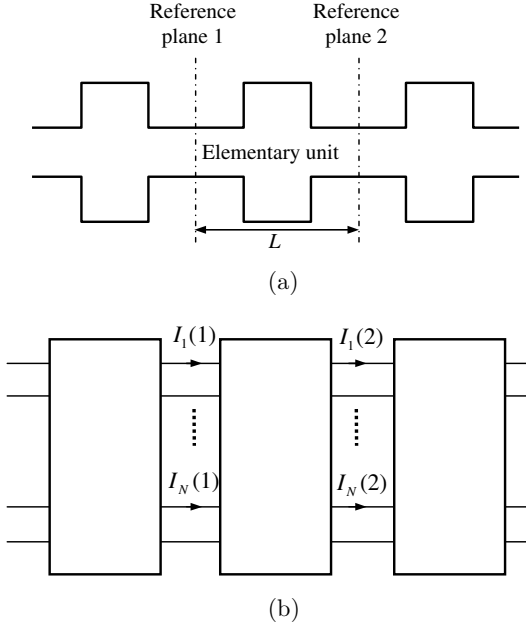


Figure 2.34 Periodic structure and its equivalent circuit.

2.9.1 Properties of Periodic Structures

The field in a periodic structure in a waveguide can be assumed to be

$$\mathbf{E}(x, y, z) = \tilde{\mathbf{E}}(x, y, z)e^{-\gamma z}, \tag{2.218}$$

where γ is the propagation constant, $\tilde{\mathbf{E}}(x, y, z)$ is a periodic function along z direction with period L. Then we have

$$\mathbf{E}(x, y, z + L) = \mathbf{E}(x, y, z)e^{-\gamma L}. \tag{2.219}$$

This is called **Floquet theorem**, after French mathematician Gaston Floquet (1847–1920). The periodic function $\tilde{\mathbf{E}}(x, y, z)$ may be expanded as a Fourier series

$$\mathbf{E}(x, y, z) = \sum_{n=-\infty}^{\infty} \mathbf{E}_n(x, y)e^{-\gamma z}e^{-j\frac{2\pi n}{L}z}, \tag{2.220}$$

where

$$\mathbf{E}_n(x, y) = \frac{1}{L}\int_{z}^{z+L} \tilde{\mathbf{E}}(x, y, z)e^{j\frac{2\pi n}{L}z}dz.$$

Each term in (2.220) is called a **spatial harmonic**. For a lossless system, we may let $\gamma = j\beta_0$. The propagation constant for nth spatial harmonic is

$$\beta_n = \beta_0 + \frac{2\pi n}{L}.$$

The phase velocity of the nth spatial harmonic is given by

$$v_{pn} = \frac{\omega}{\beta_n} = \frac{\omega}{\beta_0 + \frac{2\pi n}{L}}.$$

Since n can take negative integers, the phase velocity of the spatial harmonics can be negative. The group velocity of the nth spatial harmonic is

$$v_{gn} = \frac{d\omega}{d\beta_n} = \left(\frac{d\beta_n}{d\omega}\right)^{-1} = \left(\frac{d\beta_0}{d\omega}\right)^{-1} = v_{g0}.$$

Hence the spatial harmonics have the same group velocity.

2.9.2 Equivalent Circuit for Periodic Structures

Consider an elementary structure between two reference planes as shown in Figure 2.34(a). When two elementary structures are in close proximity, the higher order modes at the reference planes cannot be ignored. The incident fields into each elementary structure are superposition of the dominant mode and the higher order modes. Assume that the periodic structure may be considered as a uniform waveguide in the vicinity of the reference planes. The transverse fields at the reference planes i ($i = 1, 2$) can be expressed in terms of the vector modal functions $\mathbf{e}_n$ of the uniform waveguide as follows

$$\mathbf{E}_t(i) = \sum_{n=1}^{N} V_n(i)\mathbf{e}_n, \quad \mathbf{H}_t(i) = \sum_{n=1}^{N} I_n(i)\mathbf{u}_z \times \mathbf{e}_n, \quad i = 1, 2, \qquad (2.221)$$

where $V_n(i)$ and $I_n(i)$ are the modal voltage and modal current at the reference plane i. The N modes in the expansions (2.221) may be propagating or evanescent. Define the voltage vector $[V]$ and current vector $[I]$

$$[V(i)] = \begin{bmatrix} V_1(i) \\ V_2(i) \\ \vdots \\ V_N(i) \end{bmatrix}, \quad [I(i)] = \begin{bmatrix} I_1(i) \\ I_2(i) \\ \vdots \\ I_N(i) \end{bmatrix}$$

and the elementary structure is then equivalent to a $2N$ port network, which has N inputs and N outputs. If the medium is linear, the transverse electric field at the reference plane is determined by the transverse magnetic field and vice versa (uniqueness theorem). For convenience, we introduce the transfer matrix $[T]$

$$[T] = \begin{bmatrix} [T_{11}] & [T_{12}] \\ [T_{21}] & [T_{22}] \end{bmatrix}$$

such that

$$\begin{bmatrix} [V(2)] \\ [I(2)] \end{bmatrix} = \begin{bmatrix} [T_{11}] & [T_{12}] \\ [T_{21}] & [T_{22}] \end{bmatrix} \begin{bmatrix} [V(1)] \\ [I(1)] \end{bmatrix}. \tag{2.222}$$

If the network is reciprocal the transfer matrix is **symplectic**

$$[T]^T[J][T] = [J], \tag{2.223}$$

where $[J]$ is a block matrix

$$[J] = \begin{bmatrix} 0 & [1] \\ -[1] & 0 \end{bmatrix},$$

and $[1]$ is $N \times N$ identity matrix. If the network is lossless (i.e., the net power into the network is zero), the transfer matrix satisfies

$$[T]^+[K][T] = [K], \tag{2.224}$$

where $[K]$ is defined by

$$[K] = \begin{bmatrix} 0 & [1] \\ [1] & 0 \end{bmatrix}.$$

By Floquet theorem, we may write

$$\begin{bmatrix} [V(2)] \\ [I(2)] \end{bmatrix} = \begin{bmatrix} [1]e^{-\gamma L} & 0 \\ 0 & [1]e^{-\gamma L} \end{bmatrix} \begin{bmatrix} [V(1)] \\ [I(1)] \end{bmatrix}. \tag{2.225}$$

Subtracting (2.222) from (2.225) yields the following eigenvalue problem

$$\left\{ \begin{bmatrix} [T_{11}] & [T_{12}] \\ [T_{21}] & [T_{22}] \end{bmatrix} - e^{-\gamma L} \begin{bmatrix} [1] & 0 \\ 0 & [1] \end{bmatrix} \right\} \begin{bmatrix} [V(1)] \\ [I(1)] \end{bmatrix} = 0. \tag{2.226}$$

Let

$$\lambda = e^{-\gamma L}, \quad [X] = \begin{bmatrix} [V(1)] \\ [I(1)] \end{bmatrix}.$$

We have

$$([T] - \lambda[1])[X] = 0. \tag{2.227}$$

The eigenvalue λ is determined by the following algebraic equation

$$\det([T] - \lambda[1]) = 0. \tag{2.228}$$

By (2.223), it is easy to show that the following properties hold for the eigenvalue problem (2.227) (Kurokawa, 1969):

(1) If all the eigenvalues of (2.228) are distinct, an arbitrary vector of $2N$ dimension can be represented as a linear combination of the corresponding eigenvectors.
(2) If λ is an eigenvalue, λ^{-1} is also an eigenvalue.
(3) Let λ_i and λ_j be eigenvalues, $[X_i]$ and $[X_j]$ their corresponding eigenvectors. If $\lambda_i\lambda_j \neq 1$, then

$$[X_i]^T[J][X_j] = 0. \tag{2.229}$$

If $\lambda_i\lambda_j = 1$, then

$$[X_i]^T[J][X_j] \neq 0. \tag{2.230}$$

If (2.230) holds, we may let

$$[\widetilde{X}_i] = \begin{cases} \tilde{c}[X_i]^T, & \lambda_i^2 = 1 \\ \tilde{c}[X_j]^T, & \lambda_i^2 \neq 1 \end{cases},$$

where $[X_j]$ is the eigenvector corresponding to $\lambda_j = \lambda_i^{-1}$. By properly selecting the constant $\tilde{c}$, we may have

$$[\widetilde{X}_i][J][X_j] = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}. \tag{2.231}$$

An arbitrary vector of $2N$ dimension may be represented by

$$[X] = \sum_{i=1}^{2N} [X_i]([\widetilde{X}_i][J][X]). \tag{2.232}$$

Applying $[T]$ to the above equation gives

$$[T][X] = \sum_{i=1}^{2N} \lambda_i [X_i]([\widetilde{X}_i][J][X]). \tag{2.233}$$

Thus the spectral representation of $[T]$ is

$$[T](\cdot) = \sum_{i=1}^{2N} \lambda_i [X_i][\widetilde{X}_i][J](\cdot). \tag{2.234}$$

In general, we have

$$[T]^n(\cdot) = \sum_{i=1}^{2N} \lambda_i^n [X_i][\widetilde{X}_i][J](\cdot). \tag{2.235}$$

The transverse electromagnetic fields at the reference plane can be constructed from the eigenvector $[X_i]$ of $[T]$. Equation (2.235) implies that the electromagnetic fields in the nth period is the nth power of eigenvalue times the electromagnetic fields in the first period. The electromagnetic field pattern with this property is called a **mode** of the periodic structure.

2.9.3 ω–β Diagram

A waveguide is loaded at regular intervals by diaphragms, as shown in Figure 2.35(a). The elementary structure is shown in Figure 2.35(b). Assume that the impedance Z is purely reactive, denoted by $Z = jb_0Y_0$, where Y_0 is the wave admittance. The voltage and current at the reference plane T_1 can be written as

$$V_1 = A_1 + B_1, \quad I_1 = Y_0(A_1 - B_1). \tag{2.236}$$

On the left-hand side of the reference plane T, the voltage and current are given by

$$V = A_1 e^{-j\theta} + B_1 e^{j\theta}, \quad I = Y_0(A_1 e^{-j\theta} - B_1 e^{j\theta}), \tag{2.237}$$

where $\theta = \beta_0 L/2$, and $\beta_0 = \sqrt{k^2 - k_c^2}$ is the propagation constant. On the right-hand side of the reference plane, the voltage and current are respectively V and $I - ZV$

$$V = A + B, \quad I - ZV = Y_0(A - B). \tag{2.238}$$

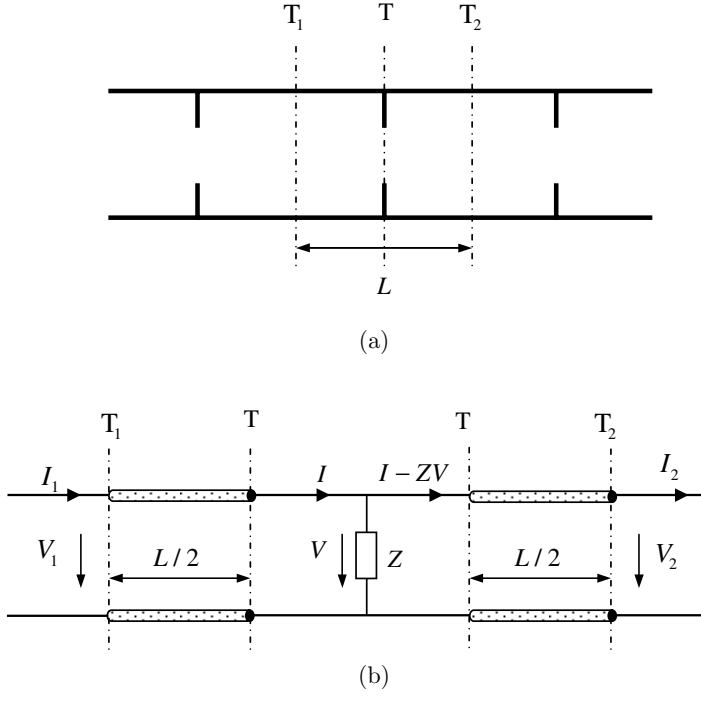


Figure 2.35 Equivalent circuit for the elementary structure.

The voltage and current at the reference plane T_2 are

$$V_2 = Ae^{-j\theta} + Be^{j\theta}, \quad I_2 = Y_0(Ae^{-j\theta} - Be^{j\theta}). \tag{2.239}$$

It follows from (2.236) to (2.239) that

$$\begin{bmatrix} V_2 \\ I_2 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = [T] \begin{bmatrix} V_1 \\ I_1 \end{bmatrix}, \tag{2.240}$$

where

$$T_{11} = T_{22} = \cos\beta_0 L - \frac{1}{2} b_0 \sin\beta_0 L,$$

$$T_{12} = j\frac{1}{Y_0}\left(\frac{1}{2} b_0 - \sin\beta_0 L - \frac{1}{2} b_0 \cos\beta_0 L\right),$$

$$T_{21} = -jY_0\left(\frac{1}{2} b_0 + \sin\beta_0 L + \frac{1}{2} b_0 \cos\beta_0 L\right).$$

By Floquet theorem, (2.240) can be written as

$$\begin{bmatrix} V_2 \\ I_2 \end{bmatrix} = \begin{bmatrix} e^{-\gamma L} & 0 \\ 0 & e^{-\gamma L} \end{bmatrix} \begin{bmatrix} V_1 \\ I_1 \end{bmatrix}. \tag{2.241}$$

Subtracting (2.240) from (2.241) yields

$$\left\{ \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} - e^{-\gamma L} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\} \begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = 0. \tag{2.242}$$

The propagation constant γ may be determined from

$$\det \left\{ \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} - e^{-\gamma L} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\} = 0.$$

This equation has two roots for the propagation constant

$$e^{-\gamma_{1,2} L} = \cos \beta_0 L - \frac{1}{2} b_0 \sin \beta_0 L \pm \left[\left(\frac{1}{2} b_0 \right)^2 - \left(\sin \beta_0 L + \frac{1}{2} b_0 \cos \beta_0 L \right)^2 \right]^{1/2}. \tag{2.243}$$

If the condition

$$\left(\frac{1}{2} b_0 \right)^2 < \left(\sin \beta_0 L + \frac{1}{2} b_0 \cos \beta_0 L \right)^2$$

is met, the eigenvalue $e^{-\gamma_{1,2} L}$ is a complex number of unit amplitude and the periodic structure is in passband. When condition

$$\left(\frac{1}{2} b_0 \right)^2 > \left(\sin \beta_0 L + \frac{1}{2} b_0 \cos \beta_0 L \right)^2$$

is met, the eigenvalue is a real number and the periodic structure is in stopband. If the periodic structure is in passband, we may let

$$\gamma_1 = -j\beta L, \quad \gamma_2 = j\beta L.$$

Inserting these into (2.243), we obtain

$$\cos \beta L = \cos \beta_0 L - \frac{1}{2} b_0 \sin \beta_0 L.$$

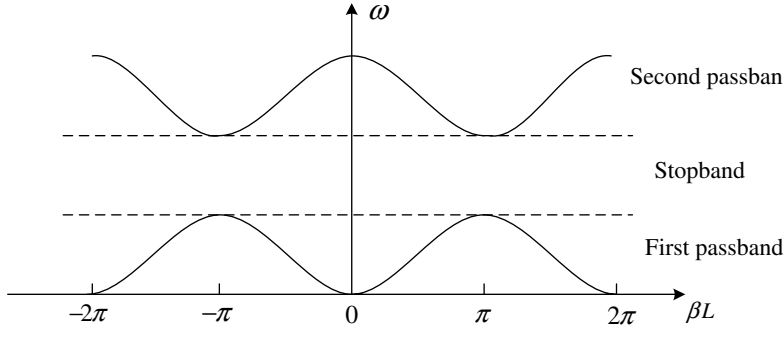


Figure 2.36 ω–β diagram.

This gives the relationship between frequency and βL. The resultant plot is called ω–β **diagram**. Such a plot is shown in Figure 2.36, where the passbands and stopbands occur alternately.

In the new era, thought itself will be transmitted by radio.

—Guglielmo Marconi

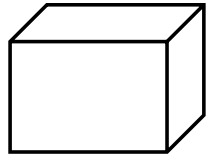
Chapter 3

Microwave Resonators

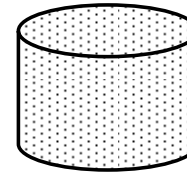
> The further a mathematical theory is developed, the more harmoniously and uniformly does its construction proceed, and unsuspected relations are disclosed between hitherto separated branches of the science.
>
> —David Hilbert (German mathematician, 1862–1943)

A **resonator** is a device or system that naturally oscillates at some frequencies, called its **resonant frequencies**, with greater amplitude than at others. The oscillations in a resonator can be either electromagnetic or mechanical. Resonators are used to either generate waves of specific frequencies or to select specific frequencies from a signal. An LC circuit in electrical engineering is a typical resonator, where the energy oscillates back and forth between the capacitor and the inductor and the oscillation (resonant) frequency is determined by the capacitance and inductance values. The number of resonant frequencies of a system corresponds to its degree of freedom.

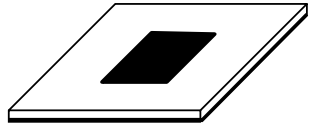
Microwave resonators, such as metal cavity resonators, dielectric resonators, microstrip resonators, and open resonators as shown in Figure 3.1, are widely used in microwave circuit designs. A **metal cavity resonator** consists of a closed metallic structure that confines electromagnetic energy in a specified region. The electromagnetic fields in the cavity are excited by an external power source, which is coupled to the cavity by a small aperture, a small probe or a loop. A **dielectric resonator**, first proposed by Richtmyer (1939), is a piece of dielectric material that exhibits resonance just as a metallic resonator. Compared to metal cavity resonators, dielectric resonators have many advantages such as lower weight, lower cost, smaller size, and ease of manufacturing, etc. Although the electromagnetic fields are not zero outside the dielectric walls of the resonator, they decay very rapidly while away from the resonator



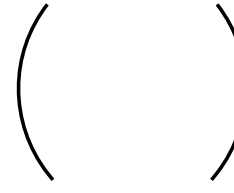
(a) Metal cavity resonator



(b) Dielectric resonator



(c) Microstrip resonator



(d) Open resonator

Figure 3.1 Microwave resonators.

walls and most of the energy is stored in the dielectric region if the dielectric constant is sufficiently high. Dielectric resonators can exhibit extremely high Q factor that is comparable to a metal cavity resonator.

A metal cavity resonator is often used in the frequency range where its dimensions are of the order of wavelength. As the operating frequency increases, the size of the cavity resonator becomes either very small (if it is kept to the order of wavelength), or very large compared to the wavelength. In the latter case, a large number of resonant modes will be excited and their resonant frequencies will be squeezed together or overlap (in other words, the density of the resonant modes increases). This phenomenon is called **mode competition**. To reduce the number of the modes, one can remove some of the walls of the cavity resonator to form an open resonator. If the remaining walls (called **mirrors**) are properly designed, only a small portion of the modes will exist (weakly damped) while other modes will become strongly damped and thus the modal density is reduced. An **open resonator** (or **optical resonator)** is an arrangement of mirrors that forms a standing wave cavity resonator for high frequency (light) waves.

3.1 Theory of Metal Cavity Resonators

For a metal cavity resonator, the electric and magnetic energies are stored in the cavity and the only losses are due to finite conductivity of cavity

walls and dielectric losses of material filling the cavity. A metal cavity resonator has an infinite number of resonant frequencies that correspond to electromagnetic field modes satisfying necessary boundary conditions on the walls of the cavity. If the set of modes is ordered with increasing resonant frequencies, there is always a lowest resonant frequency but, in general, no highest resonant frequency. As the resonant frequencies increase, the density of the modes increases accordingly (becoming infinitely dense at infinite frequencies). For this reason, only the first few modes are useful in practice for a closed cavity resonator.

3.1.1 Field Expansions for Cavity Resonators

Assume that the cavity is filled with homogeneous medium with medium parameters μ and ε. The enclosed region by a perfectly conducting wall is denoted by V, its boundary by S, as shown in Figure 3.2. The fields inside the cavity satisfy the Maxwell equations

$$\begin{aligned} \nabla \times \mathbf{H}(\mathbf{r}) &= j\omega\varepsilon \mathbf{E}(\mathbf{r}), \\ \nabla \times \mathbf{E}(\mathbf{r}) &= -j\omega\mu \mathbf{H}(\mathbf{r}), \\ \nabla \cdot \mathbf{E}(\mathbf{r}) = 0, &\quad \nabla \cdot \mathbf{H}(\mathbf{r}) = 0, \end{aligned} \tag{3.1}$$

and the boundary conditions $\mathbf{u}_n \times \mathbf{E} = 0$ and $\mathbf{u}_n \cdot \mathbf{H} = 0$ on S, where $\mathbf{u}_n$ is the unit outward normal of S. It follows from (3.1) that

$$\begin{aligned} &\nabla \times \nabla \times \mathbf{E}(\mathbf{r}) - k^2 \mathbf{E}(\mathbf{r}) = 0, \quad \mathbf{r} \in V, \\ &\mathbf{u}_n \times \mathbf{E}(\mathbf{r}) = 0, \quad \mathbf{r} \in S, \end{aligned} \tag{3.2}$$

$$\begin{aligned} &\nabla \times \nabla \times \mathbf{H}(\mathbf{r}) - k^2 \mathbf{H}(\mathbf{r}) = 0, \quad \mathbf{r} \in V, \\ &\mathbf{u}_n \cdot \mathbf{H}(\mathbf{r}) = 0, \quad \mathbf{u}_n \times \nabla \times \mathbf{H}(\mathbf{r}) = 0, \quad \mathbf{r} \in S. \end{aligned} \tag{3.3}$$

As a result, we may introduce the following eigenvalue problems

$$\begin{aligned} &\nabla \times \nabla \times \mathbf{e} - k_e^2 \mathbf{e} = 0, \quad \nabla \cdot \mathbf{e} = 0, \quad \mathbf{r} \in V, \\ &\mathbf{u}_n \times \mathbf{e} = 0, \quad \mathbf{r} \in S, \end{aligned} \tag{3.4}$$

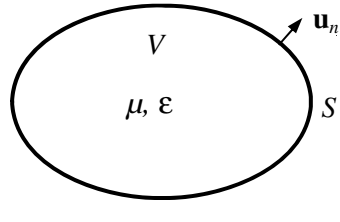


Figure 3.2 An arbitrary metal cavity resonator.

$$\begin{aligned} &\nabla \times \nabla \times \mathbf{h} - k_h^2 \mathbf{h} = 0, \quad \nabla \cdot \mathbf{h} = 0, \quad \mathbf{r} \in V, \\ &\mathbf{u}_n \cdot \mathbf{h} = 0, \quad \mathbf{u}_n \times \nabla \times \mathbf{h} = 0, \quad \mathbf{r} \in S, \end{aligned} \tag{3.5}$$

where k_e^2 and k_h^2 are the eigenvalues to be determined. The eigenfunctions of (3.4) and (3.5), however, do not form a complete set. Considering the relations $\nabla \cdot \mathbf{e} = 0$ and $\nabla \cdot \mathbf{h} = 0$, (3.4) and (3.5) may be regularized as follows

$$\begin{aligned} &\nabla \times \nabla \times \mathbf{e} - \nabla \nabla \cdot \mathbf{e} - k_e^2 \mathbf{e} = 0, \quad \mathbf{r} \in V, \\ &\mathbf{u}_n \times \mathbf{e} = 0, \quad \nabla \cdot \mathbf{e} = 0, \quad \mathbf{r} \in S, \end{aligned} \tag{3.6}$$

$$\begin{aligned} &\nabla \times \nabla \times \mathbf{h} - \nabla \nabla \cdot \mathbf{h} - k_h^2 \mathbf{h} = 0, \quad \mathbf{r} \in V, \\ &\mathbf{u}_n \cdot \mathbf{h} = 0, \quad \mathbf{u}_n \times \nabla \times \mathbf{h} = 0, \quad \mathbf{r} \in S. \end{aligned} \tag{3.7}$$

These are the eignevalue equations for the metal cavity resonator, which are similar to those for a metal waveguide and can be studied in an exact manner. Equation (3.6) has an infinite number of eigenvalues $0 \leq k_{e1}^2 \leq k_{e2}^2 \leq \cdots \leq k_{en}^2 \leq \cdots$, and $k_{en}^2 \to \infty$ $(n \to \infty)$. The corresponding eigenfunctions $\{\mathbf{e}_n\}$ are called **electric vector modal functions**, and they form a complete set. The electric vector modal functions fall into the following three categories:

$$\begin{aligned} &1.\ \nabla \times \mathbf{e}_n = 0, \quad \nabla \cdot \mathbf{e}_n = 0, \\ &2.\ \nabla \times \mathbf{e}_n \neq 0, \quad \nabla \cdot \mathbf{e}_n = 0, \\ &3.\ \nabla \times \mathbf{e}_n = 0, \quad \nabla \cdot \mathbf{e}_n \neq 0. \end{aligned}$$

Similar to the waveguide theory, the electric vector modal functions belonging to the first category only exist in a multiply-connected region. For the modes in the first category, we may introduce a scalar potential function φ_n and let $\mathbf{e}_n = \nabla \varphi_n$. Thus

$$\begin{aligned} &\nabla^2 \varphi_n = 0, \quad r \in V, \\ &\varphi_n = \text{const}, \quad r \in S. \end{aligned} \tag{3.8}$$

For the modal functions in the third category, we may introduce a scalar potential function ϕ_n and let $\mathbf{e}_n = \nabla \phi_n$. It is easy to find that

$$\begin{aligned} &\nabla^2 \phi_n + k_{en}^2 \phi_n = 0, \quad r \in V, \\ &\phi_n = 0, \quad r \in S \end{aligned} \tag{3.9}$$

and

$$\int_V |\mathbf{e}_n|^2 \, dV = k_{en}^2 \int_V \phi_n^2 \, dV. \tag{3.10}$$

Similarly, it can be shown that the magnetic vector modal functions of (3.7) form a complete set and fall into the following three categories

$$1.\ \nabla \times \mathbf{h}_n = 0, \quad \nabla \cdot \mathbf{h}_n = 0,$$

$$2.\ \nabla \times \mathbf{h}_n \neq 0, \quad \nabla \cdot \mathbf{h}_n = 0,$$

$$3.\ \nabla \times \mathbf{h}_n = 0, \quad \nabla \cdot \mathbf{h}_n \neq 0.$$

For the vector modal functions belonging to the first and third category, we may introduce a scalar potential function ψ_n and let $\mathbf{h}_n = \nabla \psi_n$ to find that

$$\begin{aligned} &\nabla^2 \psi_n + k_{hn}^2 \psi_n = 0, \quad r \in V \\ &\frac{\partial \psi_n}{\partial n} = 0, \quad r \in S \end{aligned} \tag{3.11}$$

and

$$\int_V |\mathbf{h}_n|^2 \, dV = k_{hn}^2 \int_V \psi_n^2 \, dV. \tag{3.12}$$

The vector modal functions belonging to the second category in the two sets of vector modal functions $\{\mathbf{e}_n\}$ and $\{\mathbf{h}_n\}$ are related to each other. In fact, if $\mathbf{e}_n$ is in the second category, we can define a function $\mathbf{h}_n$ through

$$\nabla \times \mathbf{e}_n = k_{en} \mathbf{h}_n, \tag{3.13}$$

and $\mathbf{h}_n$ belongs to Category 2. Furthermore, we have

$$\nabla \times \nabla \times \mathbf{h}_n - k_{en}^2 \mathbf{h}_n = k_{en}^{-1} \nabla \times \left(\nabla \times \nabla \times \mathbf{e}_n - k_{en}^2 \mathbf{e}_n \right) = 0, \quad \mathbf{r} \in V,$$

$$\mathbf{u}_n \times \nabla \times \mathbf{h}_n = k_{en}^{-1} \mathbf{u}_n \times \nabla \times \nabla \times \mathbf{e}_n = k_{en}^{-1} \mathbf{u}_n \times k_{en}^2 \mathbf{e}_n = 0, \quad \mathbf{r} \in S.$$

Consider the integration of $\mathbf{u}_n \cdot \mathbf{h}_n$ over an arbitrary part of S, denoted ΔS

$$\int_{\Delta S} \mathbf{u}_n \cdot \mathbf{h}_n \, dS = k_{en}^{-1} \int_{\Delta S} \mathbf{u}_n \cdot \nabla \times \mathbf{e}_n \, dS = k_{en}^{-1} \int_{\Delta \Gamma} \mathbf{e}_n \cdot \mathbf{u}_\Gamma \, d\Gamma, \tag{3.14}$$

where $\Delta\Gamma$ is the closed contour around ΔS and $\mathbf{u}_\Gamma$ is the unit tangent vector along the contour. The right-hand side of (3.14) vanishes. So we have $\mathbf{u}_n \cdot \mathbf{h}_n = 0$ for ΔS is arbitrary. Therefore, $\mathbf{h}_n$ satisfies (3.7), and the corresponding eigenvalue is k_{en}^2. If $\mathbf{h}_m$ is another vector modal function corresponding to $\mathbf{e}_m$ belonging to the second category, then

$$\begin{aligned}\int_V \mathbf{h}_m \cdot \mathbf{h}_n \, dV &= (k_{em}k_{en})^{-1} \int_V \nabla \times \mathbf{e}_m \cdot \nabla \times \mathbf{e}_n \, dV \\ &= (k_{em}k_{en})^{-1} \int_S \mathbf{u}_n \times \mathbf{e}_m \cdot \nabla \times \mathbf{e}_n \, dS \\ &\quad + (k_{en}/k_{em}) \int_V \mathbf{e}_m \cdot \mathbf{e}_n \, dV = \delta_{mn}.\end{aligned}$$

Consequently, the vector modal functions $\mathbf{h}_n$ in the second category can be derived from the vector modal functions $\mathbf{e}_n$ in Category 2 and they are orthonormal. Conversely, if $\mathbf{h}_n$ is in the second category, one can define $\mathbf{e}_n$ through

$$\nabla \times \mathbf{h}_n = k_{hn}\mathbf{e}_n. \tag{3.15}$$

A similar discussion shows that $\mathbf{e}_n$ is an eigenfunction of (3.4) with k_{hn} being the eigenvalue. So the completeness of the two sets is still guaranteed if the vector modal functions belonging to the second category in $\{\mathbf{e}_n\}$ and $\{\mathbf{h}_n\}$ are related through either (3.13) or (3.15). Hereafter, (3.13) and (3.15) will be assumed and $k_{e,n} = k_{h,n}$ will be denoted by k_n. Note that the complete set $\{\mathbf{e}_n\}$ is most appropriate for the expansion of electric field, and the complete set $\{\mathbf{h}_n\}$ is most appropriate for the expansion of the magnetic field.

The Maxwell equations inside the cavity filled with homogeneous medium with medium parameters σ, μ, and ε can be written as

$$\begin{aligned}\nabla \times \mathbf{H}(\mathbf{r}, t) &= (\sigma + j\omega\varepsilon)\mathbf{E}(\mathbf{r}, t), \\ \nabla \times \mathbf{E}(\mathbf{r}, t) &= -j\omega\mu\mathbf{H}(\mathbf{r}, t).\end{aligned} \tag{3.16}$$

The fields inside the cavity resonator may be expanded as follows

$$\begin{aligned}\mathbf{E}(\mathbf{r}) &= \sum_n V_n\mathbf{e}_n(\mathbf{r}) + \sum_\nu V_\nu\mathbf{e}_\nu(\mathbf{r}), \\ \mathbf{H}(\mathbf{r}) &= \sum_n I_n\mathbf{h}_n(\mathbf{r}) + \sum_\tau I_\tau\mathbf{h}_\tau(\mathbf{r}),\end{aligned} \tag{3.17}$$

$$\begin{aligned}
\nabla \times \mathbf{E}(\mathbf{r}) &= \sum_n \mathbf{h}_n(\mathbf{r}) \int_V \nabla \times \mathbf{E}(\mathbf{r}) \cdot \mathbf{h}_n(\mathbf{r}) dV \\
&\quad + \sum_\tau \mathbf{h}_\tau(\mathbf{r}) \int_V \nabla \times \mathbf{E}(\mathbf{r}) \cdot \mathbf{h}_\tau(\mathbf{r}) dV, \\
\nabla \times \mathbf{H}(\mathbf{r}) &= \sum_n \mathbf{e}_n(\mathbf{r}) \int_V \nabla \times \mathbf{H}(\mathbf{r}) \cdot \mathbf{e}_n(\mathbf{r}) dV \\
&\quad + \sum_\nu \mathbf{e}_\nu(\mathbf{r}) \int_V \nabla \times \mathbf{H}(\mathbf{r}) \cdot \mathbf{e}_\nu(\mathbf{r}) dV,
\end{aligned} \tag{3.18}$$

where the subscript n denotes the modes belonging to second category, and the greek subscript ν and τ for the modes belonging to first or third category, and

$$V_{n(\nu)} = \int_V \mathbf{E}(\mathbf{r}) \cdot \mathbf{e}_{n(\nu)}(\mathbf{r}) dV, \quad I_{n(\tau)} = \int_V \mathbf{H}(\mathbf{r}) \cdot \mathbf{h}_{n(\tau)}(\mathbf{r}) dV. \tag{3.19}$$

Making use of the following calculations

$$\begin{aligned}
\int_V \nabla \times \mathbf{E} \cdot \mathbf{h}_n \, dV &= \int_V \mathbf{E} \cdot \nabla \times \mathbf{h}_n \, dV + \int_S (\mathbf{E} \times \mathbf{h}_n) \cdot \mathbf{u}_n \, dS \\
&= k_n V_n + \int_S (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_n \, dS, \\
\int_V \nabla \times \mathbf{E} \cdot \mathbf{h}_\tau \, dV &= \int_V \mathbf{E} \cdot \nabla \times \mathbf{h}_\tau \, dV + \int_S (\mathbf{E} \times \mathbf{h}_\tau) \cdot \mathbf{u}_n \, dS \\
&= \int_S (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_\tau \, dS, \\
\int_V \nabla \times \mathbf{H} \cdot \mathbf{e}_n \, dS &= \int_V \mathbf{H} \cdot \nabla \times \mathbf{e}_n \, dV + \int_S (\mathbf{H} \times \mathbf{e}_n) \cdot \mathbf{u}_n \, dS = k_n I_n, \\
\int_V \nabla \times \mathbf{H} \cdot \mathbf{e}_\nu \, dS &= \int_V \mathbf{H} \cdot \nabla \times \mathbf{e}_\nu \, dV + \int_S (\mathbf{H} \times \mathbf{e}_\nu) \cdot \mathbf{u}_n \, dS = 0.
\end{aligned}$$

Equation (3.18) may be written as

$$\nabla \times \mathbf{E} = \sum_n \mathbf{h}_n \left[k_n V_n + \int_S (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_n \, dS \right] + \sum_\tau \mathbf{h}_\tau(\mathbf{r}) \int_S (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_\tau \, dS$$

$$\nabla \times \mathbf{H} = \sum_n k_n I_n \mathbf{e}_n.$$

It follows from the above equations, (3.16) and (3.17) that

$$\sum_n k_n I_n \mathbf{e}_n = (\sigma + j\omega\varepsilon) \left[\sum_n V_n \mathbf{e}_n(\mathbf{r}) + \sum_\nu V_\nu \mathbf{e}_\nu(\mathbf{r}) \right],$$

$$\sum_n \mathbf{h}_n \left[k_n V_n + \int_S (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_n \, dS \right] + \sum_\tau \mathbf{h}_\tau(\mathbf{r}) \int_S (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_\tau \, dS$$

$$= -j\omega\mu \left[\sum_n I_n \mathbf{h}_n(\mathbf{r}) + \sum_\tau I_\tau \mathbf{h}_\tau(\mathbf{r}) \right].$$

Comparing the expansion coefficients gives

$$k_n I_n = (\sigma + j\omega\varepsilon) V_n,$$

$$V_\nu = 0,$$

$$k_n V_n + \int_S (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_n \, dS = -j\omega\mu I_n,$$

$$\int_S (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_\tau \, dS = -j\omega\mu I_\tau.$$

From these equations, we obtain

$$\begin{aligned} V_n &= \frac{k_n}{\sigma + j\omega\varepsilon} I_n, \\ V_\nu &= 0, \\ I_n &= -\left[j\omega\mu + \frac{k_n^2}{(\sigma + j\omega\varepsilon)} \right]^{-1} \int_S (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_n \, dS, \\ I_\tau &= \frac{1}{-j\omega\mu} \int_S (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_\tau \, dS. \end{aligned} \tag{3.20}$$

Thus, once the tangential electric fields on the boundary S are known, the fields inside the cavity resonator are then fully determined.

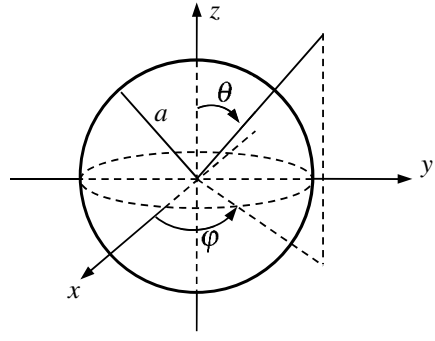


Figure 3.3 Spherical cavity resonator.

Example 3.1 (Vector modal functions of spherical cavity)**:** For a spherical cavity resonator of radius a shown in Figure 3.3, the vector modal functions in the third category can be obtained by solving (3.9). In spherical coordinate system (r, θ, φ), the solutions of (3.9) are given by

$$\phi_{nmq}(\mathbf{r}) = j_n(k_{nq}r)P_n^m(\cos\theta)\begin{Bmatrix}\cos m\varphi\\ \sin m\varphi\end{Bmatrix}, \tag{3.21}$$

where k_{nq} satisfies the equation

$$j_n(k_{nq}a) = 0, \tag{3.22}$$

and the eigenvalues k_{en} in (3.9) are given by

$$k_{en} = k_{nq}. \tag{3.23}$$

The normalization integral is

$$\int_V \phi_{nmq}^2\, dV = \frac{a^3 2\pi(m+n)!}{\varepsilon_m(2n+1)(n-m)!}[j_{n+1}(k_{nq}a)]^2. \tag{3.24}$$

The corresponding vector modal functions $\mathbf{e}_n$ in the third category, denoted by $\mathbf{L}_{nmq}$, are

$$\begin{aligned}\mathbf{L}_{nmq} &= \nabla\phi_{nmq}\\ &= \frac{dj_n(k_{nq}r)}{dr}P_n^m(\cos\theta)\begin{Bmatrix}\cos m\varphi\\ \sin m\varphi\end{Bmatrix}\mathbf{u}_r\\ &\quad + \frac{1}{r}j_n(k_{nq}r)\frac{dP_n^m(\cos\theta)}{d\theta}\begin{Bmatrix}\cos m\varphi\\ \sin m\varphi\end{Bmatrix}\mathbf{u}_\theta\\ &\quad + \frac{m}{r\sin\theta}j_n(k_{nq}r)P_n^m(\cos\theta)\begin{Bmatrix}-\sin m\varphi\\ \cos m\varphi\end{Bmatrix}\mathbf{u}_\varphi\end{aligned} \tag{3.25}$$

where $\mathbf{u}_r$, $\mathbf{u}_\theta$ and $\mathbf{u}_\varphi$ are unit vectors along r, θ and φ direction respectively. The vector modal functions in the second category satisfy

$$\begin{aligned} &\nabla \times \nabla \times \mathbf{e}_n - \nabla\nabla \cdot \mathbf{e}_n - k_e^2 \mathbf{e}_n = 0, \quad \mathbf{r} \in V, \\ &\nabla \cdot \mathbf{e}_n = 0, \quad \mathbf{r} \in V, \\ &\mathbf{u}_n \times \mathbf{e}_n = 0, \quad \mathbf{r} \in S, \end{aligned} \tag{3.26}$$

and may be constructed by (We use $\mathbf{M}$ and $\mathbf{N}$ to denote the two different classes of vector modal functions in the second category.)

$$\mathbf{M} = \nabla \times (\mathbf{r}\psi), \quad \mathbf{N} = \frac{1}{k}\nabla \times \nabla \times (\mathbf{r}\psi),$$

where ψ satisfies

$$(\nabla^2 + k^2)\psi = 0. \tag{3.27}$$

The solutions of (3.27) inside the spherical cavity are given by

$$\psi_{nm}(\mathbf{r}) = j_n(kr) P_n^m(\cos\theta) \begin{Bmatrix} \cos m\varphi \\ \sin m\varphi \end{Bmatrix}.$$

Thus, we may write

$$\begin{aligned} \mathbf{M}_{nmq} &= \nabla \times \left[\mathbf{r} j_n(k_{nq} r) P_n^m(\cos\theta) \begin{Bmatrix} \cos m\varphi \\ \sin m\varphi \end{Bmatrix} \right] \\ &= \frac{m}{\sin\theta} j_n(k_{nq} r) P_n^m(\cos\theta) \begin{Bmatrix} -\sin m\varphi \\ \cos m\varphi \end{Bmatrix} \mathbf{u}_\theta \\ &\quad - j_n(k_{nq} r) \frac{dP_n^m(\cos\theta)}{d\theta} \begin{Bmatrix} \cos m\varphi \\ \sin m\varphi \end{Bmatrix} \mathbf{u}_\varphi, \end{aligned} \tag{3.28}$$

where k_{nq} satisfy (3.22). Also we have

$$\begin{aligned} \mathbf{N}_{nmq} &= \frac{1}{k_{nq}} \nabla \times \nabla \times \left[\mathbf{r} j_n(k_{nq} r) P_n^m(\cos\theta) \begin{Bmatrix} \cos m\varphi \\ \sin m\varphi \end{Bmatrix} \right] \\ &= \frac{n(n+1)}{k_{nq} r} j_n(k_{nq} r) P_n^m(\cos\theta) \begin{Bmatrix} -\sin m\varphi \\ \cos m\varphi \end{Bmatrix} \mathbf{u}_r \\ &\quad + \frac{1}{k_{nq} r} \frac{d[r j_n(k_{nq} r)]}{dr} \frac{dP_n^m(\cos\theta)}{d\theta} \begin{Bmatrix} \cos m\varphi \\ \sin m\varphi \end{Bmatrix} \mathbf{u}_\theta \\ &\quad + \frac{m}{k_{nq} r \sin\theta} \frac{d[r j_n(k_{nq} r)]}{dr} P_n^m(\cos\theta) \begin{Bmatrix} -\sin m\varphi \\ \cos m\varphi \end{Bmatrix} \mathbf{u}_\varphi, \end{aligned} \tag{3.29}$$

where k_{nq} satisfy

$$\left.\frac{d[rj_n(k_{nq}r)]}{dr}\right|_{r=a} = 0.$$

The vector modal functions $\mathbf{M}_{nmq}$ and $\mathbf{N}_{nmq}$ are called transverse electric modes and transverse magnetic modes respectively. □

3.1.2 Vector Modal Functions for Waveguide Cavity Resonators

The evaluation of the vector modal functions in an arbitrary metal cavity is not an easy task. When the metal cavity consists of a section of a uniform metal waveguide shorted at both ends, the method of separation of variables may be applied to determine the vector modal functions of simple cavity geometries such as rectangular waveguide cavity resonator, circular waveguide cavity resonator and coaxial waveguide cavity resonator.

3.1.2.1 *Field Expansions in Waveguide Cavity Resonator*

Consider a waveguide cavity with a perfect electric wall of length L, as shown in Figure 3.4. The electromagnetic fields inside the waveguide cavity can be expanded in terms of the transverse vector modal functions $\mathbf{e}_n$ in the waveguide

$$\begin{aligned}\mathbf{E}(\mathbf{r}) &= \sum_{n=1}^{\infty} v_n(z)\mathbf{e}_n(\boldsymbol{\rho}) + \mathbf{u}_z \sum_{n=1}^{\infty} \frac{\nabla \cdot \mathbf{e}_n(\boldsymbol{\rho})}{k_{cn}} e'_{zn}(z),\\ \mathbf{H}(\mathbf{r}) &= \sum_{n=1}^{\infty} i_n(z)\mathbf{u}_z \times \mathbf{e}_n(\boldsymbol{\rho}) + \mathbf{u}_z \frac{1}{\sqrt{\Omega}} \int_{\Omega} \frac{\mathbf{u}_z \cdot \mathbf{H}}{\sqrt{\Omega}} d\Omega + \sum_{n=1}^{\infty} \frac{\nabla \times \mathbf{e}_n(\boldsymbol{\rho})}{k_{cn}} h'_{zn}(z),\end{aligned} \tag{3.30}$$

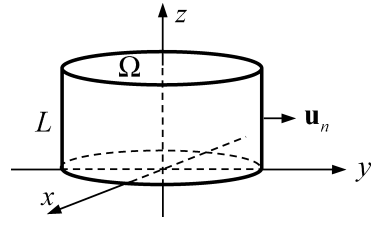


Figure 3.4 A metal cavity formed by a section of waveguide.

where $\boldsymbol{\rho} = (x, y)$ is the position vector in the waveguide cross-section Ω, and

$$v_n(z) = \int_\Omega \mathbf{E} \cdot \mathbf{e}_n \, d\Omega, \quad i_n(z) = \int_\Omega \mathbf{H} \cdot \mathbf{u}_z \times \mathbf{e}_n \, d\Omega,$$

$$h'_{zn}(z) = \int_\Omega \mathbf{H} \cdot \left(\frac{\nabla \times \mathbf{e}_{tn}}{k_{cn}} \right) d\Omega, \quad e'_{zn}(z) = \int_\Omega \mathbf{u}_z \cdot \mathbf{E} \left(\frac{\nabla \cdot \mathbf{e}_{tn}}{k_{cn}} \right) d\Omega.$$

For the TEM modes, we have

$$\frac{dv_n^{\mathrm{TEM}}}{dz} = -j\omega\mu i_n^{\mathrm{TEM}}, \quad \frac{di_n^{\mathrm{TEM}}}{dz} = -j\omega\varepsilon v_n^{\mathrm{TEM}}. \tag{3.31}$$

The modal voltages for the TEM modes satisfy

$$\frac{d^2 v_n^{\mathrm{TEM}}}{dz^2} + k^2 v_n^{\mathrm{TEM}} = 0. \tag{3.32}$$

For the TE modes, we have

$$\begin{aligned} &\frac{dv_n^{\mathrm{TE}}}{dz} = -j\omega\mu i_n^{\mathrm{TE}}, \\ &\frac{di_n^{\mathrm{TE}}}{dz} - k_{cn} h'_{zn} = -j\omega\varepsilon v_n^{\mathrm{TE}}, \\ &j\omega\mu h'_{zn} = -k_{cn} v_n^{\mathrm{TE}}. \end{aligned} \tag{3.33}$$

The modal voltages v_n^{TE} satisfy

$$\frac{d^2 v_n^{\mathrm{TE}}}{dz^2} + (k^2 - k_{cn}^2) v_n^{\mathrm{TE}} = 0. \tag{3.34}$$

For the TM modes, we have

$$\begin{aligned} &\frac{dv_n^{\mathrm{TM}}}{dz} + k_{cn} e'_{zn} = -j\omega\mu i_n^{\mathrm{TM}}, \\ &\frac{di_n^{\mathrm{TM}}}{dz} = -j\omega\varepsilon v_n^{\mathrm{TM}}, \\ &j\omega\varepsilon e'_{zn} = k_{cn} i_n^{\mathrm{TM}}. \end{aligned} \tag{3.35}$$

The modal voltages v_n^{TM} satisfy

$$\frac{\partial^2 v_n^{\mathrm{TM}}}{\partial z^2} + (k^2 - k_{cn}^2) v_n^{\mathrm{TM}} = 0. \tag{3.36}$$

Since the tangential electric field on the electric conductor must be zero, the voltages satisfy the homogeneous Dirichlet boundary conditions

$$v_n(z)|_{z=0} = v_n(z)|_{z=L} = 0. \tag{3.37}$$

Thus, we may write

$$v_n^{\mathrm{TEM}}(z) = a_n^{\mathrm{TEM}}\sqrt{\frac{2}{L}}\sin\frac{l\pi}{L}z, \quad k = \frac{l\pi}{L},$$
$$i_n^{\mathrm{TEM}}(z) = \frac{1}{-jk\eta}\frac{dv_n^{\mathrm{TEM}}}{dz} = \frac{a_n^{\mathrm{TEM}}}{-j\eta}\sqrt{\frac{2}{L}}\cos\frac{l\pi}{L}z,$$

$$v_n^{\mathrm{TE}}(z) = a_n^{\mathrm{TE}}\sqrt{\frac{2}{L}}\sin\frac{l\pi}{L}z, \quad k^2 = \left(\frac{l\pi}{L}\right)^2 + k_{cn}^2,$$
$$i_n^{\mathrm{TE}}(z) = \frac{1}{-jk\eta}\frac{dv_n^{\mathrm{TE}}}{dz} = \frac{a_n^{\mathrm{TE}}}{-j\eta\sqrt{(l\pi/L)^2 + k_{cn}^2}}\sqrt{\frac{2}{L}}\frac{l\pi}{L}\cos\frac{l\pi}{L}z,$$
$$v_n^{\mathrm{TM}}(z) = a_n^{\mathrm{TM}}\sqrt{\frac{2}{L}}\sin\frac{l\pi}{L}z, \quad k^2 = \left(\frac{l\pi}{L}\right)^2 + k_{cn}^2,$$
$$i_n^{\mathrm{TM}}(z) = \frac{jk}{(k^2 - k_{cn}^2)}\frac{dv_n^{\mathrm{TM}}}{dz} = \frac{j\sqrt{(l\pi/L)^2 + k_{cn}^2}}{\eta(l\pi/L)}a_n^{\mathrm{TM}}\sqrt{\frac{2}{L}}\cos\frac{l\pi}{L}z,$$
$$e'_{zn}(z) = \frac{k_{cn}\eta}{jk}i_n^{\mathrm{TM}} = \frac{a_n^{\mathrm{TM}}k_{cn}}{k^2 - k_{cn}^2}\sqrt{\frac{2}{L}}\frac{l\pi}{L}\cos\frac{l\pi}{L}z = a_n^{\mathrm{TM}}\frac{k_{cn}L}{n\pi}\sqrt{\frac{2}{L}}\cos\frac{l\pi}{L}z.$$

As a result, the first equation of (3.30) can be written as

$$\mathbf{E}(\mathbf{r}) = \sum_{n=1}^{\infty} a_n^{\mathrm{TEM}}\mathbf{e}_n^{\mathrm{TEM}}(\boldsymbol{\rho}, z) + \sum_{n=1}^{\infty} a_n^{\mathrm{TE}}\mathbf{e}_n^{\mathrm{TE}}(\boldsymbol{\rho}, z) + \sum_{n=1}^{\infty} a_n^{\mathrm{TM}}\mathbf{e}_n^{\mathrm{TM}}(\boldsymbol{\rho}, z), \tag{3.38}$$

where

$$\mathbf{e}_n^{\mathrm{TEM}}(\boldsymbol{\rho}, z) = \mathbf{e}_n(\boldsymbol{\rho})\sqrt{\frac{2}{L}}\sin\frac{l\pi}{L}z,$$
$$\mathbf{e}_n^{\mathrm{TE}}(\boldsymbol{\rho}, z) = \mathbf{e}_n(\boldsymbol{\rho})\sqrt{\frac{2}{L}}\sin\frac{l\pi}{L}z, \tag{3.39}$$
$$\mathbf{e}_n^{\mathrm{TM}}(\boldsymbol{\rho}, z) = \mathbf{e}_n(\boldsymbol{\rho})\sqrt{\frac{2}{L}}\sin\frac{l\pi}{L}z + \mathbf{u}_z\nabla\cdot\mathbf{e}_n(\boldsymbol{\rho})\frac{L}{l\pi}\sqrt{\frac{2}{L}}\cos\frac{l\pi}{L}z.$$

These are the vector modal functions for waveguide cavity resonators. Note that both the vector modal functions $\mathbf{e}_n^{\mathrm{TE}}$ and $\mathbf{e}_n^{\mathrm{TM}}$ belong to the second category.

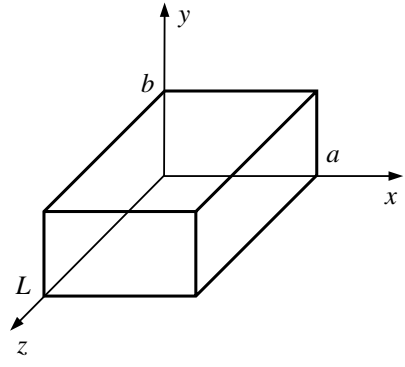


Figure 3.5 Rectangular waveguide cavity.

3.1.2.2 *Rectangular Waveguide Cavity Resonator*

For a rectangular waveguide cavity resonator shown in Figure 3.5, the vector modal functions for the rectangular cavity resonator can be obtained from the vector modal functions for the rectangular waveguide. The calculations are straightforward and we merely list the results:

$$\begin{aligned}
\mathbf{e}_n^{\mathrm{TE}}(\boldsymbol{\rho}, z) &= \mathbf{u}_x \frac{1}{k_{cn}} \frac{q\pi}{b} \sqrt{\frac{2\varepsilon_p \varepsilon_q}{\mathrm{Lab}}} \cos\frac{p\pi}{a}x \sin\frac{q\pi}{b}y \sin\frac{l\pi}{L}z \\
&\quad - \mathbf{u}_y \frac{1}{k_{cn}} \frac{p\pi}{a} \sqrt{\frac{2\varepsilon_p \varepsilon_q}{\mathrm{Lab}}} \sin\frac{p\pi}{a}x \cos\frac{q\pi}{b}y \sin\frac{l\pi}{L}z, \\
\mathbf{e}_n^{\mathrm{TM}}(\boldsymbol{\rho}, z) &= \mathbf{u}_x \frac{1}{k_{cn}} \frac{p\pi}{a} \sqrt{\frac{8}{\mathrm{Lab}}} \cos\frac{p\pi}{a}x \sin\frac{q\pi}{b}y \sin\frac{l\pi}{L}z \\
&\quad + \mathbf{u}_y \frac{1}{k_{cn}} \frac{q\pi}{b} \sqrt{\frac{8}{\mathrm{Lab}}} \sin\frac{p\pi}{a}x \cos\frac{q\pi}{b}y \sin\frac{l\pi}{L}z \\
&\quad - \mathbf{u}_z k_{cn} \frac{L}{n\pi} \sqrt{\frac{8}{\mathrm{Lab}}} \sin\frac{p\pi}{a}x \sin\frac{q\pi}{b}y \cos\frac{l\pi}{L}z.
\end{aligned}$$

3.1.2.3 *Circular Waveguide Cavity Resonator*

The vector modal functions for circular waveguide cavity resonator (Figure 3.6) are given by

$$\begin{aligned}
\mathbf{e}_n^{\mathrm{TE}}(\boldsymbol{\rho}, z) &= \pm\mathbf{u}_\rho \sqrt{\frac{2\varepsilon_m}{L\pi}} \frac{m}{\sqrt{\chi_{mn}'^2 - m^2}} \frac{1}{\rho} \frac{J_m\left(\chi_{mn}' \frac{\rho}{a}\right)}{J_m\left(\chi_{mn}'\right)} \begin{pmatrix} \sin m\varphi \\ \cos m\varphi \end{pmatrix} \sin\frac{l\pi}{L}z \\
&\quad + \mathbf{u}_\varphi \sqrt{\frac{2\varepsilon_m}{L\pi}} \frac{\chi_{mn}'}{\sqrt{\chi_{mn}'^2 - m^2}} \frac{1}{a} \frac{J_m'\left(\chi_{mn}' \frac{\rho}{a}\right)}{J_m\left(\chi_{mn}'\right)} \begin{pmatrix} \cos m\varphi \\ \sin m\varphi \end{pmatrix} \sin\frac{l\pi}{L}z,
\end{aligned}$$

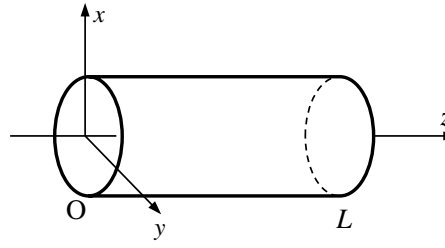


Figure 3.6 Circular waveguide cavity resonator.

$$\begin{aligned}\mathbf{e}_n^{\mathrm{TM}}(\boldsymbol{\rho}, z) = &-\mathbf{u}_\rho \sqrt{\frac{2\varepsilon_m}{L\pi}} \frac{1}{a} \frac{J_m'\left(\chi_{mn}\frac{\rho}{a}\right)}{J_{m+1}(\chi_{mn})} \begin{pmatrix}\cos m\varphi \\ \sin m\varphi\end{pmatrix} \sin\frac{l\pi}{L}z \\ &\pm \mathbf{u}_\varphi \sqrt{\frac{2\varepsilon_m}{L\pi}} \frac{m}{\rho\chi_{mn}} \frac{J_m\left(\chi_{mn}\frac{\rho}{a}\right)}{J_{m+1}(\chi_{mn})} \begin{pmatrix}\sin m\varphi \\ \cos m\varphi\end{pmatrix} \sin\frac{l\pi}{L}z \\ &+ \mathbf{u}_z \frac{L}{n\pi} \sqrt{\frac{2\varepsilon_m}{L\pi}} \frac{\chi_{mn}}{a^2} \frac{J_m\left(\chi_{mn}\frac{\rho}{a}\right)}{J_{m+1}(\chi_{mn})} \begin{pmatrix}\cos m\varphi \\ \sin m\varphi\end{pmatrix} \cos\frac{l\pi}{L}z,\end{aligned}$$

where we have used the following calculation for TM modes

$$\begin{aligned}\nabla \cdot \mathbf{e}_n(\boldsymbol{\rho}) &= \left\{ -\left[\frac{1}{z_1} J_m'(z_1) + J_m''(z_1)\right] \frac{\chi_{mn}}{a^2 J_{m+1}(\chi_{mn})} \right. \\ &\qquad \left. + \frac{1}{\chi_{mn}} \frac{m^2 J_m(z_1)}{\rho^2 J_{m+1}(\chi_{mn})} \right\} \sqrt{\frac{\varepsilon_m}{\pi}} \begin{pmatrix}\cos m\varphi \\ \sin m\varphi\end{pmatrix} \\ &= \left\{ \left[1 - \frac{a^2 m^2}{(\rho\chi_{mn})^2}\right] \frac{\chi_{mn} J_m(z_1)}{a^2 J_{m+1}(\chi_{mn})} \right. \\ &\qquad \left. + \frac{1}{\chi_{mn}} \frac{m^2 J_m(z_1)}{\rho^2 J_{m+1}(\chi_{mn})} \right\} \sqrt{\frac{\varepsilon_m}{\pi}} \begin{pmatrix}\cos m\varphi \\ \sin m\varphi\end{pmatrix} \\ &= \frac{\chi_{mn} J_m(z_1)}{a^2 J_{m+1}(\chi_{mn})} \sqrt{\frac{\varepsilon_m}{\pi}} \begin{pmatrix}\cos m\varphi \\ \sin m\varphi\end{pmatrix}, \quad z_1 = \chi_{mn}\frac{\rho}{a}.\end{aligned}$$

3.1.2.4 *Coaxial Waveguide Cavity Resonator*

The vector modal functions for coaxial waveguide cavity resonator (Figure 3.7) are listed below:

$$\mathbf{e}_n^{\mathrm{TEM}}(\boldsymbol{\rho}, z) = \mathbf{u}_\rho \frac{l}{\sqrt{2\pi \ln c_1}} \frac{1}{\rho} \sqrt{\frac{2}{L}} \sin\frac{l\pi}{L}z,$$

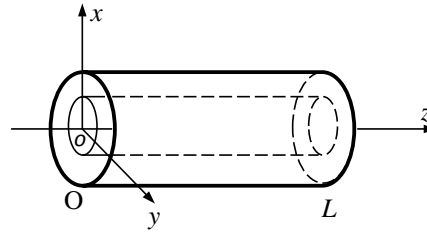


Figure 3.7 Coaxial waveguide cavity resonator.

$$\begin{aligned}
\mathbf{e}_n^{\mathrm{TE}}(\boldsymbol{\rho}, z) &= \pm \mathbf{u}_\rho \frac{m}{\rho} h\left(\chi'_{mn}\frac{\rho}{a}\right)\sqrt{\frac{\varepsilon_m}{L\pi}}\begin{pmatrix}\sin m\varphi\\ \cos m\varphi\end{pmatrix}\sin\frac{l\pi}{L}z \\
&\quad + \mathbf{u}_\varphi \frac{\chi'_{mn}}{a} h'\left(\chi'_{mn}\frac{\rho}{a}\right)\sqrt{\frac{\varepsilon_m}{L\pi}}\begin{pmatrix}\cos m\varphi\\ \sin m\varphi\end{pmatrix}\sin\frac{l\pi}{L}z, \\
\mathbf{e}_n^{\mathrm{TM}}(\boldsymbol{\rho}, z) &= -\mathbf{u}_\rho \frac{\chi_{mn}}{a} e'\left(\chi_{mn}\frac{\rho}{a}\right)\sqrt{\frac{\varepsilon_m}{L\pi}}\begin{pmatrix}\cos m\varphi\\ \sin m\varphi\end{pmatrix}\sin\frac{l\pi}{L}z \\
&\quad \pm \mathbf{u}_\varphi \frac{m}{\rho} e\left(\chi_{mn}\frac{\rho}{a}\right)\sqrt{\frac{\varepsilon_m}{L\pi}}\begin{pmatrix}\sin m\varphi\\ \cos m\varphi\end{pmatrix}\sin\frac{l\pi}{L}z \\
&\quad + \mathbf{u}_z \frac{L}{n\pi}\left(\frac{\chi_{mn}}{a}\right)^2 e\left(\chi_{mn}\frac{\rho}{a}\right)\sqrt{\frac{\varepsilon_m}{L\pi}}\begin{pmatrix}\cos m\varphi\\ \sin m\varphi\end{pmatrix}\cos\frac{l\pi}{L}z,
\end{aligned}$$

where we have used the following calculation for TM modes

$$\begin{aligned}
\nabla\cdot\mathbf{e}_n &= \left\{-\left(\frac{\chi_{mn}}{a}\right)^2\left[\frac{1}{z_1}e'(z_1)+e''(z_1)\right]+\frac{m^2}{\rho^2}e(z_1)\right\}\sqrt{\frac{\varepsilon_m}{2\pi}}\begin{pmatrix}\cos m\varphi\\ \sin m\varphi\end{pmatrix} \\
&= \left\{\left(\frac{\chi_{mn}}{a}\right)^2\left[1-\frac{a^2m^2}{(\rho\chi_{mn})^2}\right]+\frac{m^2}{\rho^2}\right\}e(z_1)\sqrt{\frac{\varepsilon_m}{2\pi}}\begin{pmatrix}\cos m\varphi\\ \sin m\varphi\end{pmatrix} \\
&= \left(\frac{\chi_{mn}}{a}\right)^2 e(z_1)\sqrt{\frac{\varepsilon_m}{2\pi}}\begin{pmatrix}\cos m\varphi\\ \sin m\varphi\end{pmatrix},
\end{aligned}$$

and

$$e\left(\chi_{mn}\frac{\rho}{a}\right) = \frac{\pi}{\sqrt{2}}\frac{J_m\left(\chi_{mn}\frac{\rho}{a}\right)N_m(\chi_{mn}) - N_m\left(\chi_{mn}\frac{\rho}{a}\right)J_m(\chi_{mn})}{\sqrt{J_m^2(\chi_{mn})/J_m^2(c\chi_{mn})-1}},$$

$$m = 0, 1, 2, \ldots,$$

$$h\left(\chi'_{mn}\frac{\rho}{a}\right)=\frac{\pi}{\sqrt{2}}\frac{J_m\left(\chi'_{mn}\frac{\rho}{a}\right)N'_m(\chi_{mn})-N_m\left(\chi'_{mn}\frac{\rho}{a}\right)J'_m(\chi'_{mn})}{\sqrt{\frac{J_m^2(\chi'_{mn})}{J_m^2(c_1\chi'_{mn})}\left[1-\left(\frac{m}{c_1\chi'_{mn}}\right)^2\right]-\left[1-\left(\frac{m}{\chi'_{mn}}\right)^2\right]}},$$

$$m=0,1,2,\ldots,$$

$$z_1=\chi_{mn}\frac{\rho}{a}.$$

3.1.3 Integral Equation for Cavity Resonators

A resonant cavity filled with homogeneous medium and completely enclosed by a perfect conductor can be studied through the integral equation formulation. To derive an integral equation for the resonant cavity problem, we may use the integral representation of the magnetic field inside the metal cavity region V

$$\mathbf{H}(\mathbf{r})=-\int_S[\mathbf{u}_n(\mathbf{r}')\times\mathbf{H}(\mathbf{r}')]\times\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}'),\quad \mathbf{r}\in V, \tag{3.40}$$

where $G(\mathbf{r},\mathbf{r}')=e^{-jk|\mathbf{r}-\mathbf{r}'|}/4\pi\,|\mathbf{r}-\mathbf{r}'|$ and $k=\omega\sqrt{\mu\varepsilon}$. In deriving (3.40), the boundary condition $\mathbf{u}_n(\mathbf{r})\times\mathbf{E}(\mathbf{r})=0$ has been used. Letting $\mathbf{r}$ approach S from inside V and using the jump relation yield

$$\frac{1}{2}\mathbf{H}_-(\mathbf{r})=-\int_S[\mathbf{u}_n(\mathbf{r}')\times\mathbf{H}_-(\mathbf{r}')]\times\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}'),$$

where $\mathbf{H}_-(\mathbf{r})$ denotes the limit value of $\mathbf{H}(\mathbf{r})$ when $\mathbf{r}$ approaches the boundary S from inside V. Introducing the surface current density $\mathbf{J}_s(\mathbf{r})=-\mathbf{u}_n(\mathbf{r})\times\mathbf{H}_-(\mathbf{r})$, the above equation can be written as

$$\frac{1}{2}\mathbf{J}_s(\mathbf{r})+\mathbf{u}_n(\mathbf{r})\times\int_S\mathbf{J}_s(\mathbf{r}')\times\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')=0. \tag{3.41}$$

The condition that (3.41) has a nontrivial solution can be used to determine the resonant frequencies. Numerical discretization of (3.41) is straightforward (Geyi and Hongshi, 1988a). Evidently any resonant frequencies of (3.1) satisfy the integral equation (3.41). It can be shown that the converse is also true. In fact, if $\mathbf{J}_s(\mathbf{r})$ is a nontrivial solution corresponding to a frequency ω obtained from (3.41), one may construct the fields

$$\begin{aligned}\mathbf{E}(\mathbf{r})&=\int_S\left[j\omega\mu\mathbf{J}_s(\mathbf{r}')G(\mathbf{r},\mathbf{r}')-\frac{\rho_s(\mathbf{r}')}{\varepsilon}\nabla' G(\mathbf{r},\mathbf{r}')\right]dS(\mathbf{r}'),\\ \mathbf{H}(\mathbf{r})&=-\int_S\mathbf{J}_s(\mathbf{r}')\times\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}'),\end{aligned} \tag{3.42}$$

where $\mathbf{r} \in R^3$, $\rho_s = \nabla_s \cdot \mathbf{J}_s / j\omega$ and ∇_s denotes the surface divergence. From the jump relations, we obtain

$$\mathbf{E}_+(\mathbf{r}) = -\frac{\rho_s(\mathbf{r})}{2\varepsilon}\mathbf{u}_n(\mathbf{r}) + \int_S \left[j\omega\mu \mathbf{J}_s(\mathbf{r}')G(\mathbf{r},\mathbf{r}') - \frac{\rho_s(\mathbf{r}')}{\varepsilon}\nabla' G(\mathbf{r},\mathbf{r}') \right] dS(\mathbf{r}'),$$

$$\mathbf{E}_-(\mathbf{r}) = \frac{\rho_s(\mathbf{r})}{2\varepsilon}\mathbf{u}_n(\mathbf{r}) + \int_S \left[j\omega\mu \mathbf{J}_s(\mathbf{r}')G(\mathbf{r},\mathbf{r}') - \frac{\rho_s(\mathbf{r}')}{\varepsilon}\nabla' G(\mathbf{r},\mathbf{r}') \right] dS(\mathbf{r}'),$$

$$\mathbf{H}_+(\mathbf{r}) = -\frac{1}{2}\mathbf{J}_s(\mathbf{r}) \times \mathbf{u}_n(\mathbf{r}) - \int_S \mathbf{J}_s(\mathbf{r}') \times \nabla' G(\mathbf{r},\mathbf{r}') dS(\mathbf{r}'),$$

$$\mathbf{H}_-(\mathbf{r}) = \frac{1}{2}\mathbf{J}_s(\mathbf{r}) \times \mathbf{u}_n(\mathbf{r}) - \int_S \mathbf{J}_s(\mathbf{r}') \times \nabla' G(\mathbf{r},\mathbf{r}') dS(\mathbf{r}'),$$

where $+$ and $-$ denote the limit value from inside V and outside V respectively. These equations imply

$$\mathbf{E}_+(\mathbf{r}) - \mathbf{E}_-(\mathbf{r}) = -\frac{\rho_s(\mathbf{r})}{\varepsilon}\mathbf{u}_n(\mathbf{r}), \tag{3.43}$$

$$\mathbf{u}_n(\mathbf{r}) \times \mathbf{H}_+(\mathbf{r}) = 0, \quad \mathbf{r} \in S. \tag{3.44}$$

It is easy to show that the fields defined by (3.42) satisfy the Maxwell equations in whole space and the radiation condition at infinity. From (3.44) and the uniqueness theorem of Maxwell equations, the electromagnetic fields defined by (3.42) are zero outside Ω. Therefore, we have $\mathbf{E}_+(\mathbf{r}) = 0$, $\mathbf{r} \in S$ and from (3.43) we obtain $\mathbf{u}_n(\mathbf{r}) \times \mathbf{E}_-(\mathbf{r}) = 0$, $\mathbf{r} \in S$, which shows that ω and the fields defined by (3.42) satisfy (3.1). Hence ω is a resonant frequency of the cavity resonator.

3.2 Coupling between Waveguide and Cavity Resonator

The cavity resonators can be coupled to an external circuit (such as a waveguide) through apertures, probes, loops and gaps, etc., depending on the nature of the cavity resonator and the external circuit under consideration. These couplings may change the field distribution inside the cavity and can be investigated by the modal theory developed before. Figure 3.8 shows some typical coupling mechanisms and their equivalent circuits.

3.2.1 One-Port Microwave Network as a RLC Circuit

Consider a one-port network fed by a waveguide as shown in Figure 3.8(a). Let Ω be the cross-section of the waveguide at a reference plane T (input

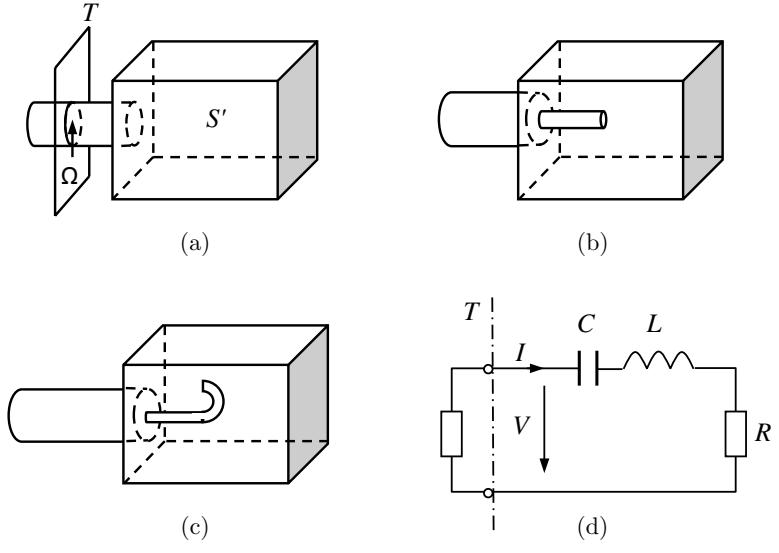


Figure 3.8 Typical coupling mechanisms and their equivalent circuits.

terminal). The cross-section Ω and the perfectly conducting wall S' form a closed region V bounded by S, in which the medium parameters are σ, μ, ε. Applying the Poynting theorem to the fields in V, we obtain

$$\int\limits_{S=\Omega+S'} (\mathbf{E} \times \bar{\mathbf{H}}) \cdot \mathbf{u}_n \, dS = -j4\omega(\widetilde{W}_m - \widetilde{W}_e) + 2P, \tag{3.45}$$

where $\mathbf{u}_n$ is the unit outward normal to S and

$$\widetilde{W}_m = \frac{1}{4}\int\limits_V \mu \, |\mathbf{H}|^2 \, dV, \quad \widetilde{W}_e = \frac{1}{4}\int\limits_V \varepsilon \, |\mathbf{E}|^2 \, dV, \quad P = \frac{1}{2}\int\limits_V \sigma \, |\mathbf{E}|^2 \, dV.$$

Since S' is assumed to be a perfect conductor, we have

$$\int\limits_{S=\Omega+S'} (\mathbf{E} \times \bar{\mathbf{H}}) \cdot \mathbf{u}_n \, dS = -V\bar{I} \tag{3.46}$$

for waveguide in a single-mode state. It follows from (3.45) and (3.46) that the input impedance of the one-port network is given by

$$Z = \frac{V}{I} = R + j\left(\omega L - \frac{1}{\omega C}\right),$$

where

$$R = \frac{2P}{|I|^2}, \quad L = \frac{4\widetilde{W}_m}{|I|^2}, \quad C = \frac{|I|^2}{4\omega^2 \widetilde{W}_e}.$$

Therefore, a one-port network is equivalent to a RLC circuit. In general cases, the component values of the equivalent RLC circuit depend on the frequency.

3.2.2 Properties of RLC Resonant Circuit

Let us consider a RLC circuit with constant components shown in Figure 3.9. When the RLC circuit is at resonance the reactive part of the input impedance vanishes. In this case, the input current reaches the maximum for a given impressed voltage. The input impedance of the RLC circuit can be written as

$$Z = R + j\left(\omega L - \frac{1}{\omega C}\right).$$

The reactive part vanishes at the resonant frequency ω_r, which is given by

$$\omega_r = \frac{1}{\sqrt{LC}}.$$

Then the input impedance can be expressed as

$$Z = Z_0\left[\Delta + j\left(\frac{\omega}{\omega_r} - \frac{\omega_r}{\omega}\right)\right], \quad |Z| = Z_0\left[\Delta^2 + \left(\frac{\omega}{\omega_r} - \frac{\omega_r}{\omega}\right)^2\right]^{1/2},$$

where

$$Z_0 = \omega_r L = \frac{1}{\omega_r C} = \sqrt{\frac{L}{C}}, \quad \Delta = \frac{R}{\sqrt{L/C}} = \frac{R}{Z_0}.$$

The magnitude of the current in response to a unit voltage is

$$|I| = \frac{1}{|Z|}, \quad |I|_{\max} = \frac{1}{Z_0\Delta}.$$

If the input voltage is fixed, the current reaches maximum at the resonant frequency. Note that we have assumed that all the component values are

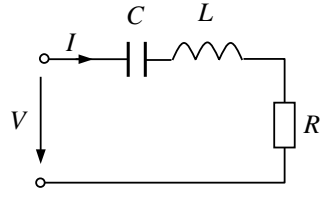


Figure 3.9 RLC circuit.

independent of frequency. The response of the current (the resonant curve) drops to $1/\sqrt{2}$ fraction of its maximum value when

$$\left|\frac{\omega}{\omega_r} - \frac{\omega_r}{\omega}\right| = \Delta,$$

which has two solutions

$$\frac{\omega_L}{\omega_r} = -\frac{1}{2}\Delta + \left(1 + \frac{\Delta^2}{4}\right)^{1/2}, \quad \frac{\omega_H}{\omega_r} = \frac{1}{2}\Delta + \left(1 + \frac{\Delta^2}{4}\right)^{1/2}.$$

The bandwidth of the resonant curve is defined by

$$B = \frac{\omega_H - \omega_L}{\omega_r} = \Delta.$$

The stored energies in the inductor and capacitor can be written as

$$\widetilde{W}_m = \frac{1}{4}|I|^2 L, \quad \widetilde{W}_e = \frac{1}{4}\frac{|I|^2}{\omega^2 C}$$

respectively. The quality factor at resonance is given by

$$\begin{aligned} Q &= \omega_r \frac{\widetilde{W}_m + \widetilde{W}_e}{P} = \frac{2\omega_r \widetilde{W}_m}{P} = \frac{2\omega_r}{(1/2)|I|^2_{\max} R}\frac{1}{4}L|I|^2_{\max} \\ &= \frac{\omega_r L}{R} = \frac{Z_0}{R} = \frac{1}{\Delta} = \frac{1}{B}. \end{aligned}$$

It will be shown in Chapter 5 that the above relationship approximately holds for a high Q antenna. The input complex power can be expressed as

$$\begin{aligned} P_{\text{in}} &= \frac{1}{2}V\bar{I} = \frac{1}{2}|I|^2 Z = \frac{1}{2}|I|^2\left[R - j\left(\omega L - \frac{1}{\omega C}\right)\right] \\ &= P_R + j2\omega(\widetilde{W}_m - \widetilde{W}_e). \end{aligned}$$

3.2.3 Aperture Coupling to Cavity Resonator

The electromagnetic energy may be coupled to a cavity resonator by a waveguide through an aperture. Let Ω be the cross section of the waveguide at $z = 0$ where the reference plane T (input terminal) intersects with the waveguide, as shown in Figure 3.10. The cross section Ω and metallic wall S' form the cavity resonator region.

Suppose the waveguide only supports the dominant mode and the waveguide extends to infinity in $-z$ direction. A wave of unit amplitude is incident from $z = -\infty$, which excites a number of higher order modes in the neighborhood of the reference plane. The transverse fields in the

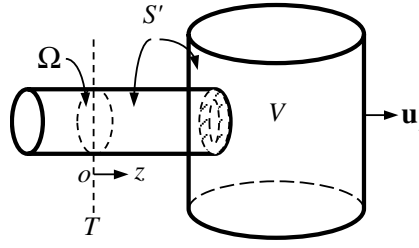


Figure 3.10 Coupling between waveguide and cavity resonator.

waveguide region $z < 0$ can be expanded as follows

$$\begin{aligned} -\mathbf{u}_z \times \mathbf{E} &= -(e^{-j\beta_1 z} + \Gamma e^{j\beta_1 z})\mathbf{u}_z \times \mathbf{e}_{1,g} - \sum_{m=2}^{\infty} V_{m,g} e^{j\beta_m z} \mathbf{u}_z \times \mathbf{e}_{m,g}, \\ -\mathbf{u}_z \times \mathbf{H} &= (e^{-j\beta_1 z} - \Gamma e^{j\beta_1 z}) Z_{w1}^{-1} \mathbf{e}_{1,g} - \sum_{m=2}^{\infty} V_{m,g} Z_{wm}^{-1} e^{j\beta_m z} \mathbf{e}_{m,g}, \end{aligned} \tag{3.47}$$

where the subscript g signifies the waveguide modes; $V_{m,g}$ are the modal voltages; Γ is the refection coefficient of the dominant mode at $z = 0$. The fields in the cavity region $z > 0$ can be expanded as

$$\begin{aligned} \mathbf{E} &= \sum_n \frac{k_n}{\sigma + j\omega\varepsilon} I_{n,r} \mathbf{e}_{n,r}, \\ \mathbf{H} &= \sum_n I_{n,r} \mathbf{h}_{n,r} + \sum_\tau I_{\tau,r} \mathbf{h}_{\tau,r}, \end{aligned} \tag{3.48}$$

where the subscript r denotes the resonator modes, and we have used (3.20). The expansion coefficients $I_{n,r}$ are determined by (3.20). Note that

$$\int_S (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_{n,r}\, dS = \int_\Omega (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_{n,r}\, dS + \int_{S'} (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_{n,r}\, dS,$$

$$\int_S (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_{\tau,r}\, dS = \int_\Omega (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_{\tau,r}\, dS + \int_{S'} (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_{\tau,r}\, dS.$$

Assuming that the cavity wall is a perfect conductor and making use of (3.47), we have

$$\begin{aligned} \int_S (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_{n,r}\, dS &= \int_\Omega (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_{n,r}\, dS \\ &= -(1 + \Gamma) \int_\Omega \mathbf{u}_z \times \mathbf{e}_{1,g} \cdot \mathbf{h}_{n,r}\, dS \\ &\quad - \sum_{m=2}^{\infty} V_{m,g} \int_\Omega \mathbf{u}_z \times \mathbf{e}_{m,g} \cdot \mathbf{h}_{n,r}\, dS, \end{aligned}$$

$$\int_S (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_{\tau,r}\, dS = \int_\Omega (\mathbf{u}_n \times \mathbf{E}) \cdot \mathbf{h}_{\tau,r}\, dS$$
$$= -(1+\Gamma) \int_\Omega \mathbf{u}_z \times \mathbf{e}_{1,g} \cdot \mathbf{h}_{\tau,r}\, dS$$
$$- \sum_{m=2}^{\infty} V_{m,g} \int_\Omega \mathbf{u}_z \times \mathbf{e}_{m,g} \cdot \mathbf{h}_{\tau,r}\, dS.$$

Substituting these into (3.20), we obtain

$$I_{n,r} = \frac{\left(j + \frac{\sigma}{\omega\varepsilon}\right) \sum_{m=1}^{\infty} V_{m,g} i_{nm}}{j\omega_n\mu \left[\frac{1}{Q_n} + j\left(\frac{\omega}{\omega_n} - \frac{\omega_n}{\omega}\right)\right]}, \quad I_{\tau,r} = \frac{1}{j\omega\mu} \sum_{m=1}^{\infty} V_{m,g} i_{\tau m}, \tag{3.49}$$

where $Q_n = \omega_n\varepsilon/\sigma$, $V_{1,g} = 1 + \Gamma$ and

$$i_{nm} = \int_\Omega \mathbf{u}_z \times \mathbf{e}_{m,g} \cdot \mathbf{h}_{n,r}\, dS, \quad i_{\tau m} = \int_\Omega \mathbf{u}_z \times \mathbf{e}_{m,g} \cdot \mathbf{h}_{\tau,r}\, dS.$$

Note that the tangential magnetic field must be continuous at $z = 0$. Thus, it follows from (3.47) and (3.48) that

$$(1-\Gamma) Z_{w1}^{-1} \mathbf{e}_{1,g} - \sum_{m=2}^{\infty} V_{m,g} Z_{wm}^{-1} \mathbf{e}_{m,g}$$
$$= -\sum_n I_{n,r} \mathbf{u}_z \times \mathbf{h}_{n,r} - \sum_\tau I_{\tau,r} \mathbf{u}_z \times \mathbf{h}_{\tau,r}.$$

Multiplying both sides by $\mathbf{e}_{1,g}$ respectively, and taking the integration over the reference plane, we get

$$(1-\Gamma) Z_{w1}^{-1} = \sum_n I_{n,r} i_{n1} + \sum_\tau I_{\tau,r} i_{\tau 1}.$$

The input admittance is given by

$$Y = \frac{1}{Z_{w1}} \frac{1-\Gamma}{1+\Gamma} = \sum_n \frac{\left(j + \frac{\sigma}{\omega\varepsilon}\right)(i_{n1})^2}{j\omega_n\mu \left[\frac{1}{Q_n} + j\left(\frac{\omega}{\omega_n} - \frac{\omega_n}{\omega}\right)\right]} + \sum_\tau \frac{(i_{\tau 1})^2}{j\omega\mu}$$
$$+ \sum_n \sum_{m=2}^{\infty} \frac{V_{m,g}}{V_{1,g}} \frac{\left(j + \frac{\sigma}{\omega\varepsilon}\right) i_{nm} i_{n1}}{j\omega_n\mu \left[\frac{1}{Q_n} + j\left(\frac{\omega}{\omega_n} - \frac{\omega_n}{\omega}\right)\right]} + \frac{1}{j\omega\mu} \sum_\tau \sum_{m=2}^{\infty} \frac{V_{m,g}}{V_{1,g}} i_{\tau m} i_{\tau 1}. \tag{3.50}$$

If only the dominant mode exists at the input terminal, this reduces to

$$Y = \frac{1}{Z_{w1}} \frac{1-\Gamma}{1+\Gamma} = \sum_n \frac{\left(j + \frac{\sigma}{\omega\varepsilon}\right)(i_{n1})^2}{j\omega_n\mu \left[\frac{1}{Q_n} + j\left(\frac{\omega}{\omega_n} - \frac{\omega_n}{\omega}\right)\right]} + \sum_\tau \frac{(i_{\tau 1})^2}{j\omega\mu}. \quad (3.51)$$

3.2.4 Probe Coupling to Cavity Resonator

Consider a metal cavity with a perfectly conducting wall, and assume that the medium in the cavity is homogeneous and isotropic with medium parameters σ, μ and ε. The volume occupied by the cavity is denoted by V and its boundary by S. If the cavity contains an impressed electric current source $\mathbf{J}$ and a magnetic current source $\mathbf{J}_m$ representing a probe, the fields excited by these sources satisfy the Maxwell equations in the cavity:

$$\begin{aligned} \nabla \times \mathbf{H}(\mathbf{r}, t) &= \varepsilon \frac{\partial \mathbf{E}(\mathbf{r}, t)}{\partial t} + \sigma \mathbf{E}(\mathbf{r}, t) + \mathbf{J}(\mathbf{r}, t), \\ \nabla \times \mathbf{E}(\mathbf{r}, t) &= -\mu \frac{\partial \mathbf{H}(\mathbf{r}, t)}{\partial t} - \mathbf{J}_m(\mathbf{r}, t), \end{aligned} \quad (3.52)$$

with the boundary conditions $\mathbf{u}_n \times \mathbf{E} = 0$ and $\mathbf{u}_n \cdot \mathbf{H} = 0$ on the boundary S. Here $\mathbf{u}_n$ is the unit outward normal to the boundary. The fields inside the cavity can be expanded in terms of its vector modal functions as follows

$$\begin{aligned} \mathbf{E}(\mathbf{r}, t) &= \sum_n V_n(t)\mathbf{e}_n(\mathbf{r}) + \sum_\nu V_\nu(t)\mathbf{e}_\nu(\mathbf{r}), \\ \mathbf{H}(\mathbf{r}, t) &= \sum_n I_n(t)\mathbf{h}_n(\mathbf{r}) + \sum_\tau I_\tau(t)\mathbf{h}_\tau(\mathbf{r}), \end{aligned} \quad (3.53)$$

$$\begin{aligned} \nabla \times \mathbf{E}(\mathbf{r}, t) &= \sum_n \mathbf{h}_n(\mathbf{r}) \int_V \nabla \times \mathbf{E}(\mathbf{r}, t) \cdot \mathbf{h}_n(\mathbf{r}) dV \\ &\quad + \sum_\tau \mathbf{h}_\tau(\mathbf{r}) \int_V \nabla \times \mathbf{E}(\mathbf{r}, t) \cdot \mathbf{h}_\tau(\mathbf{r}) dV, \\ \nabla \times \mathbf{H}(\mathbf{r}, t) &= \sum_n \mathbf{e}_n(\mathbf{r}) \int_V \nabla \times \mathbf{H}(\mathbf{r}, t) \cdot \mathbf{e}_n(\mathbf{r}) dV \\ &\quad + \sum_\nu \mathbf{e}_\nu(\mathbf{r}) \int_V \nabla \times \mathbf{H}(\mathbf{r}, t) \cdot \mathbf{e}_\nu(\mathbf{r}) dV, \end{aligned} \quad (3.54)$$

where the subscript n denotes the vector modal functions belonging to Category 2, and the greek subscript ν and τ for the vector modal functions

belonging to Category 1 or 3, and

$$V_{n(\nu)}(t) = \int_V \mathbf{E}(\mathbf{r},t)\cdot\mathbf{e}_{n(\nu)}(\mathbf{r})dV, \quad I_{n(\tau)}(t) = \int_V \mathbf{H}(\mathbf{r},t)\cdot\mathbf{h}_{n(\tau)}(\mathbf{r})dV. \tag{3.55}$$

Considering the following calculations

$$\begin{aligned}
\int_V \nabla\times\mathbf{E}\cdot\mathbf{h}_n\,dV &= \int_V \mathbf{E}\cdot\nabla\times\mathbf{h}_n\,dV + \int_S (\mathbf{E}\times\mathbf{h}_n)\cdot\mathbf{u}_n\,dS = k_n V_n,\\
\int_V \nabla\times\mathbf{E}\cdot\mathbf{h}_\tau\,dV &= \int_V \mathbf{E}\cdot\nabla\times\mathbf{h}_\tau\,dV + \int_S (\mathbf{E}\times\mathbf{h}_\tau)\cdot\mathbf{u}_n\,dS = 0,\\
\int_V \nabla\times\mathbf{H}\cdot\mathbf{e}_n\,dS &= \int_V \mathbf{H}\cdot\nabla\times\mathbf{e}_n\,dV + \int_S (\mathbf{H}\times\mathbf{e}_n)\cdot\mathbf{u}_n\,dS = k_n I_n,\\
\int_V \nabla\times\mathbf{H}\cdot\mathbf{e}_\nu\,dS &= \int_V \mathbf{H}\cdot\nabla\times\mathbf{e}_\nu\,dV + \int_S (\mathbf{H}\times\mathbf{e}_\nu)\cdot\mathbf{u}_n\,dS = 0,
\end{aligned}$$

(3.54) can be written as

$$\nabla\times\mathbf{E} = \sum_n k_n V_n \mathbf{h}_n, \quad \nabla\times\mathbf{H} = \sum_n k_n I_n \mathbf{e}_n.$$

Substituting the above expansions into (3.52) and equating the expansion coefficients of the vector modal functions, we obtain

$$\begin{aligned}
&\frac{\partial V_n}{\partial t} + \frac{\sigma}{\varepsilon}V_n - \frac{k_n}{\varepsilon}I_n = -\frac{1}{\varepsilon}\int_V \mathbf{J}\cdot\mathbf{e}_n\,dV,\\
&\frac{\partial V_\nu}{\partial t} + \frac{\sigma}{\varepsilon}V_\nu = -\frac{1}{\varepsilon}\int_V \mathbf{J}\cdot\mathbf{e}_\nu\,dV,\\
&\frac{\partial I_n}{\partial t} + \frac{k_n}{\mu}V_n = -\frac{1}{\mu}\int_V \mathbf{J}_m\cdot\mathbf{h}_n\,dV,\\
&\frac{\partial I_\tau}{\partial t} = -\frac{1}{\mu}\int_V \mathbf{J}_m\cdot\mathbf{h}_\tau\,dV.
\end{aligned} \tag{3.56}$$

From these equations, we may find that

$$\begin{aligned}
&\frac{\partial^2 I_n}{\partial t^2} + 2\gamma\frac{\partial I_n}{\partial t} + \omega_n^2 I_n = \omega_n S_n^I,\\
&\frac{\partial^2 V_n}{\partial t^2} + 2\gamma\frac{\partial V_n}{\partial t} + \omega_n^2 V_n = \omega_n S_n^V,
\end{aligned} \tag{3.57}$$

where $\omega_n = k_n v$, $v = 1/\sqrt{\mu\varepsilon}$, $\gamma = \sigma/2\varepsilon$ and

$$S_n^I = v \int_V \mathbf{J} \cdot \mathbf{e}_n \, dV - \frac{1}{k_n \eta} \frac{\partial}{\partial t} \int_V \mathbf{J}_m \cdot \mathbf{h}_n \, dV - \frac{\sigma v}{k_n} \int_V \mathbf{J}_m \cdot \mathbf{h}_n \, dV,$$

$$S_n^V = -\frac{\eta}{k_n} \frac{\partial}{\partial t} \int_V \mathbf{J} \cdot \mathbf{e}_n \, dV - v \int_V \mathbf{J}_m \cdot \mathbf{h}_n \, dV$$

where $\eta = \sqrt{\mu/\varepsilon}$. The expansion coefficients I_n and V_n may be determined by use of the retarded Green's function defined by

$$\begin{aligned} &\frac{\partial^2 G_n(t,t')}{\partial t^2} + 2\gamma \frac{\partial G_n(t,t')}{\partial t} + \omega_n^2 G_n(t,t') = -\delta(t-t'), \\ &G_n(t,t')|_{t<t'} = 0. \end{aligned} \tag{3.58}$$

The solution of (3.58) is readily found to be

$$G_n(t,t') = -\frac{e^{-\gamma(t-t')}}{\sqrt{\omega_n^2 - \gamma^2}} \sin\sqrt{\omega_n^2 - \gamma^2}(t-t') H(t-t'). \tag{3.59}$$

Therefore, the general solution of I_n may be written as

$$\begin{aligned} I_n(t) = &- \int_{-\infty}^{\infty} G_n(t,t') \omega_n S_n^I(t') dt' \\ &+ e^{-\gamma t}\Big(c_1 \cos\sqrt{\omega_n^2 - \gamma^2}\, t + c_2 \sin\sqrt{\omega_n^2 - \gamma^2}\, t\Big), \end{aligned} \tag{3.60}$$

where c_1 and c_2 are two arbitrary constants. If the source is turned on at $t = 0$, both $V_n(0^-)$ and $I_n(0^-)$ may be assumed to be zero due to causality. Considering the third equation of (3.56), the second term of (3.60) vanishes. Thus

$$\begin{aligned} I_n(t) = &\frac{\omega_n}{\sqrt{\omega_n^2 - \gamma^2}} \int_{0^-}^{t} e^{-\gamma(t-t')} \sin\sqrt{\omega_n^2 - \gamma^2}(t-t') \\ &\times \left[v \int_V \mathbf{J} \cdot \mathbf{e}_n \, dV - \frac{1}{k_n \eta} \frac{\partial}{\partial t'} \int_V \mathbf{J}_m \cdot \mathbf{h}_n \, dV - \frac{\sigma v}{k_n} \int_V \mathbf{J}_m \cdot \mathbf{h}_n \, dV \right] dt'. \end{aligned} \tag{3.61}$$

Similarly, we have

$$V_n(t) = \frac{\omega_n}{\sqrt{\omega_n^2 - \gamma^2}} \int_{0^-}^{t} e^{-\gamma(t-t')} \sin\sqrt{\omega_n^2 - \gamma^2}(t-t') \times \left[-\frac{\eta}{k_n} \frac{\partial}{\partial t'} \int_V \mathbf{J} \cdot \mathbf{e}_n \, dV - v \int_V \mathbf{J}_m \cdot \mathbf{h}_n \, dV \right] dt'. \tag{3.62}$$

and

$$\begin{aligned} V_\nu(t) &= -\frac{1}{\varepsilon} e^{-2\gamma t} \int_{0^-}^{t} e^{2\gamma t'} dt' \int_V \mathbf{J} \cdot \mathbf{e}_\nu \, dV, \\ I_\tau(t) &= -\frac{1}{\mu} \int_{0^-}^{t} dt' \int_V \mathbf{J}_m \cdot \mathbf{h}_\tau \, dV. \end{aligned} \tag{3.63}$$

Substituting (3.61), (3.62) and (3.63) into (3.53), we may find out the field distributions inside the metal cavity.

Assume that the current source is sinusoidal and is turned on at $t = 0$

$$\mathbf{J}(\mathbf{r}, t) = \mathbf{J}'(\mathbf{r}) H(t) \sin \omega t = \mathrm{Re}\big[-j\mathbf{J}'(\mathbf{r}) H(t) e^{j\omega t}\big] \tag{3.64}$$

and $\mathbf{J}_m(\mathbf{r}, t) = 0$. It follows from (3.61)–(3.63) that

$$I_n(t) = \omega_n v \int_V \mathbf{J}' \cdot \mathbf{e}_n \, dV \left[\frac{-(\omega_n^2 - \omega^2)\sin\omega t + 2\omega\gamma\cos\omega t}{(\omega_n^2 - \omega^2)^2 + 4\omega^2\gamma^2} + \frac{1}{\beta_n} \frac{-(\omega_n^2 - \omega^2)\omega \sin\beta_n t + 2\omega\gamma(\beta_n \cos\omega t + \gamma \sin\beta_n t)}{(\omega_n^2 - \omega^2)^2 + 4\omega^2\gamma^2} e^{-\gamma t} \right], \tag{3.65}$$

$$V_n(t) = -\frac{\eta\omega\omega_n}{k_n} \int_V \mathbf{J}' \cdot \mathbf{e}_n \, dV \left[\frac{(\omega_n^2 - \omega^2)\cos\omega t + 2\omega\gamma\sin\omega t}{(\omega_n^2 - \omega^2)^2 + 4\omega^2\gamma^2} + \frac{1}{\beta_n} \frac{-(\omega_n^2 - \omega^2)(\gamma \sin\beta_n t + \beta_n \cos\beta_n t) - 2\omega^2\gamma \sin\beta_n t}{(\omega_n^2 - \omega^2)^2 + 4\omega^2\gamma^2} e^{-\gamma t} \right], \tag{3.66}$$

$$V_\nu(t) = -\frac{1}{\varepsilon}\int_V \mathbf{J}' \cdot \mathbf{e}_\nu \, dV \left[\frac{2\gamma \sin\omega t - \omega\cos\omega t}{\omega^2 + 4\gamma^2} + \frac{\omega e^{-2\gamma t}}{\omega^2 + 4\gamma^2}\right],$$

$$I_\tau(t) = 0.$$

The time-domain electromagnetic fields are given by

$$\begin{aligned}\mathbf{E}(\mathbf{r},t) = &-\sum_n \frac{\eta\omega\omega_n}{k_n}\mathbf{e}_n(\mathbf{r})\int_V \mathbf{J}' \cdot \mathbf{e}_n \, dV \left[\frac{(\omega_n^2 - \omega^2)\cos\omega t + 2\omega\gamma\sin\omega t}{(\omega_n^2 - \omega^2)^2 + 4\omega^2\gamma^2}\right.\\ &\left. - \frac{1}{\beta_n}\frac{(\omega_n^2 - \omega^2)(\gamma\sin\beta_n t + \beta_n\cos\beta_n t) + 2\omega^2\gamma\sin\beta_n t}{(\omega_n^2 - \omega^2)^2 + 4\omega^2\gamma^2}e^{-\gamma t}\right]\\ &+ \sum_\nu \frac{1}{\varepsilon}\mathbf{e}_\nu(\mathbf{r})\int_V \mathbf{J}' \cdot \mathbf{e}_\nu \, dV \left[\frac{\omega\cos\omega t - 2\gamma\sin\omega t}{\omega^2 + 4\gamma^2} + \frac{\omega e^{-2\gamma t}}{\omega^2 + 4\gamma^2}\right],\end{aligned} \tag{3.67}$$

$$\begin{aligned}\mathbf{H}(\mathbf{r},t) = &\sum_n \omega_n v\mathbf{h}_n(\mathbf{r})\int_V \mathbf{J}' \cdot \mathbf{e}_n \, dV \left[\frac{(\omega_n^2 - \omega^2)\sin\omega t - 2\omega\gamma\cos\omega t}{(\omega_n^2 - \omega^2)^2 + 4\omega^2\gamma^2}\right.\\ &\left. + \frac{1}{\beta_n}\frac{-(\omega_n^2 - \omega^2)\omega\sin\beta_n t + 2\omega\gamma(\beta_n\cos\omega t + \gamma\sin\beta_n t)}{(\omega_n^2 - \omega^2)^2 + 4\omega^2\gamma^2}e^{-\gamma t}\right].\end{aligned} \tag{3.68}$$

Hence the response in a metal cavity resonator can be separated into the sum of a steady-state response and a transient response if the medium is lossy. The transient response tends to zero with increasing time. For more details, please refer to Geyi (2008a).

Example 3.2 (loop coupling)**:** Consider a small closed loop l of cross section Ω_0 and denote the area bounded by the loop by S_l as shown in Figure 3.11.

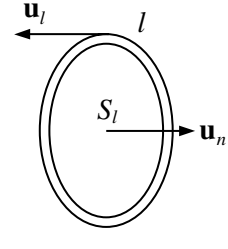


Figure 3.11 Loop coupling.

We have

$$\int_V \mathbf{J}' \cdot \mathbf{e}_n \, dV = \int_V \mathbf{J}' \cdot \mathbf{e}_n \, dS\mathbf{u}_l \cdot dl\mathbf{u}_l = \int_{\Omega_0} \mathbf{J}' \cdot dS\mathbf{u}_l \int_l \mathbf{e}_n \cdot dl\mathbf{u}_l$$

$$= jI \int_l \mathbf{e}_n \cdot dl\mathbf{u}_l = jI \int_{S_l} \nabla \times \mathbf{e}_n \cdot dS\mathbf{u}_n = jIk_n \int_{S_l} \mathbf{h}_n \cdot dS\mathbf{u}_n,$$

$$\int_V \mathbf{J}' \cdot \mathbf{e}_\nu \, dV = 0,$$

where I is the current in the loop. Thus, in terms of phasor notations, the steady state responses for the fields are

$$\mathbf{E}(\mathbf{r}) = \sum_n \frac{j\eta\omega\omega_n I}{\omega^2 - \omega_n^2 - j2\omega\gamma} \mathbf{e}_n(\mathbf{r}) \int_{S_l} \mathbf{h}_n \cdot dS\mathbf{u}_n, \tag{3.69}$$

$$\mathbf{H}(\mathbf{r}) = \sum_n \frac{\omega_n\omega_n I}{\omega^2 - \omega_n^2 - j2\omega\gamma} \mathbf{h}_n(\mathbf{r}) \int_{S_l} \mathbf{h}_n \cdot dS\mathbf{u}_n. \tag{3.70}$$

The induced voltage in the loop may be calculated by

$$V = \int_l \mathbf{E} \cdot \mathbf{u}_l \, dl = \int_{S_l} \nabla \times \mathbf{E} \cdot \mathbf{u}_n \, dS = -j\omega\mu \int_{S_l} \mathbf{H} \cdot \mathbf{u}_n \, dS$$

$$= \sum_n \frac{-jk\eta\omega_n^2 I}{\omega^2 - \omega_n^2 - j2\omega\gamma} \left(\int_{S_l} \mathbf{h}_n \cdot dS\mathbf{u}_n \right)^2,$$

which gives

$$Z = \frac{V}{I} = \sum_n \frac{\eta k_n}{\frac{1}{Q_n} + j\left(\frac{\omega}{\omega_n} - \frac{\omega_n}{\omega}\right)} \left(\int_{S_l} \mathbf{h}_n \cdot dS\mathbf{u}_n \right)^2. \tag{3.71}$$

This is the input impedance of the resonator. □

3.3 Dielectric Resonator

Microwave integrated circuits often use dielectric resonators and other open waveguide structures. A dielectric resonator is a structure of high dielectric constant, which exhibits resonance at some frequencies like a cavity resonator. The electromagnetic energy is confined to the resonator region due to the high dielectric constant. This can be understood as follows.

Let a plane wave be normally incident from upper half space of high ε_r to the free space. The reflection coefficient at the interface is

$$R = \frac{\sqrt{\varepsilon_r} - 1}{\sqrt{\varepsilon_r} + 1}.$$

As $\varepsilon_r \to \infty$, we have $R \to 1$. In this case, total reflection occurs and all energy is confined in the high dielectric constant region. The dielectric resonators are widely used in the design of filters, oscillators and antennas. The Q factor is an important figure of merit for the dielectric resonator, which can be expressed as

$$Q = \frac{\omega \widetilde{W}}{P} = \frac{1}{\tan\delta}$$

where $\widetilde{W}$ is the stored energy in the dielectric resonator, P is power dissipation, ω is resonant radian frequency, and $\tan\delta = \sigma/\omega\varepsilon_r\varepsilon_0$ is the loss tangent.

3.3.1 Representation of the Fields in a Cylindrical System

In a source-free region, the electromagnetic fields may be represented by two scalar functions. We may let $v_3 = z, h_3 = 1$ in an arbitrary curvilinear coordinate system (v_1, v_2, v_3) with metric coefficients (h_1, h_2, h_3) to obtain a cylindrical system (v_1, v_2, z) with $\frac{\partial h_{1,2}}{\partial z} = 0$. Assume that the fields in the cylindrical system have a z-dependence of the form $e^{-\gamma z}$. For TE wave, all the field components may be expressed in terms of longitudinal magnetic field H_z as

$$\begin{aligned} E_1 &= -\frac{j\omega\mu}{h_2(k^2+\gamma^2)}\frac{\partial H_z}{\partial v_2}, \quad E_2 = \frac{j\omega\mu}{h_1(k^2+\gamma^2)}\frac{\partial H_z}{\partial v_1}, \quad E_z = 0, \\ H_1 &= \frac{1}{h_1(k^2+\gamma^2)}\frac{\partial^2 H_z}{\partial v_1 \partial z}, \quad H_2 = \frac{1}{h_2(k^2+\gamma^2)}\frac{\partial^2 H_z}{\partial v_2 \partial z}. \end{aligned} \tag{3.72}$$

For TM wave, all the field components may be expressed in terms of longitudinal electric field E_z as

$$\begin{aligned} E_1 &= \frac{1}{h_1(k^2+\gamma^2)}\frac{\partial^2 E_z}{\partial v_1 \partial z}, \quad E_2 = \frac{1}{h_2(k^2+\gamma^2)}\frac{\partial^2 E_z}{\partial v_2 \partial z}, \\ H_1 &= \frac{j\omega\varepsilon}{h_2(k^2+\gamma^2)}\frac{\partial E_z}{\partial v_2}, \quad H_2 = -\frac{j\omega\varepsilon}{h_1(k^2+\gamma^2)}\frac{\partial E_z}{\partial v_1}, \quad H_z = 0. \end{aligned} \tag{3.73}$$

Note that both the field components E_z and H_z satisfy the Helmholtz equation

$$(\nabla^2 + k^2)\begin{pmatrix} E_z \\ H_z \end{pmatrix} = 0.$$

3.3.2 Circular Cylindrical Dielectric Resonator — Mixed Magnetic Wall Model

A circular cylindrical dielectric resonator of radius a and height L is shown in Figure 3.12. The origin of the circular cylindrical coordinate system is chosen at the center of the resonator. The dielectric resonator may be analyzed by using the **mixed magnetic wall model** introduced by Cohn (1968), in which the cylindrical surface $\{\rho = a, -\infty < z < \infty\}$ is assumed to be a perfect magnetic wall. The two air-filled hollow waveguides $|z| > L/2$ operate below the cut-off frequencies so that the fields decay exponentially in the z-direction away from each end of the dielectric resonator.

3.3.2.1 *TE Modes*

The TE modes in the circular cylindrical system enclosed by the magnetic wall may be determined from the longitudinal magnetic field component H_z.

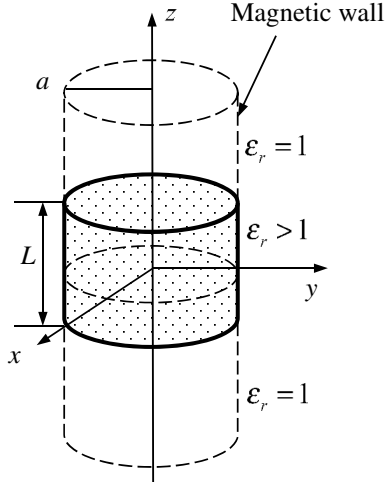


Figure 3.12 Circular cylindrical dielectric resonator.

We may write

$$H_{z1} = A_m J_m(k_c\rho)\begin{pmatrix}\cos m\varphi\\ \sin m\varphi\end{pmatrix}\cos(\beta z+\psi), \quad |z| < \frac{L}{2},$$

$$H_{z2} = B_m J_m(k_c\rho)\begin{pmatrix}\cos m\varphi\\ \sin m\varphi\end{pmatrix}e^{-\alpha(|z|-L/2)}, \quad |z| > \frac{L}{2},$$

where

$$\beta^2 = \varepsilon_r k_0^2 - k_c^2, \quad \alpha^2 = k_c^2 - k_0^2, \quad k_0^2 = \omega^2\mu_0\varepsilon_0,$$

and k_c is the separation constant. The magnetic field component H_z must vanish on the magnetic wall $\rho = a$, which yields

$$J_m(k_c a) = 0.$$

Thus the cut-off wavenumbers are given by

$$k_{cmn} = \frac{\chi_{mn}}{a},$$

where χ_{mn} are the nth zeros of the Bessel functions of the first kind of order m. The tangential fields E_ρ, H_φ must be continuous at the interface $|z| = L/2$. Equivalently

$$H_{z1} = H_{z2}, \quad \frac{\partial H_{z1}}{\partial z} = \frac{\partial H_{z2}}{\partial z}, \quad z = \pm\frac{L}{2}.$$

These conditions can be met by requiring

$$A_m\cos\left(\frac{\beta L}{2}+\psi\right) = B_m, \quad A_m\beta\sin\left(\frac{\beta L}{2}+\psi\right) = \alpha B_m,$$

which yield

$$\beta\tan\left(\frac{\beta L}{2}+\psi\right) = \alpha. \tag{3.74}$$

Considering the symmetry of the structure, H_z must be either symmetrical or asymmetrical about the plane $z = 0$. Thus we have

$$\psi = -\frac{p\pi}{2}, \quad p = 0, 1, 2, \ldots.$$

and (3.74) becomes

$$\beta L = p\pi + 2\arctan\frac{\alpha}{\beta} = (p+\delta)\pi, \tag{3.75}$$

where $\delta = \frac{2}{\pi}\arctan\frac{\alpha}{\beta}$ with $0 < \delta < 1$. The TE modes in the circular cylindrical dielectric resonator are called $\mathrm{TE}_{mn(p+\delta)}$ modes, and the lowest mode is $\mathrm{TE}_{01\delta}$. Equation (3.75) can be written as

$$\sqrt{\varepsilon_r k_0^2 - k_c^2}L = p\pi + 2\arctan\frac{\sqrt{k_c^2 - k_0^2}}{\sqrt{\varepsilon_r k_0^2 - k_c^2}}. \tag{3.76}$$

This can be used to determine the resonant frequencies of the $\mathrm{TE}_{mn(p+\delta)}$ modes.

3.3.2.2 *TM Modes*

For the TM modes, we may write

$$E_{z1} = A_m J_m(k_c\rho)\begin{pmatrix}\cos m\varphi\\ \sin m\varphi\end{pmatrix}\cos(\beta z + \psi), \quad |z| < \frac{L}{2},$$

$$E_{z2} = B_m J_m(k_c\rho)\begin{pmatrix}\cos m\varphi\\ \sin m\varphi\end{pmatrix}e^{-\alpha(|z|-L/2)}, \quad |z| > \frac{L}{2}.$$

The condition that the derivative $\partial E_z/\partial\rho$ must vanish on the magnetic wall leads to

$$J'_m(k_c a) = 0.$$

Hence, the cut-off wavenumbers are

$$k_{cmn} = \frac{\chi'_{mn}}{a},$$

where χ'_{mn} are the nth zeros of the derivative of the Bessel functions of the first kind of mth order, i.e., $J'_m(\chi'_{mn}) = 0$. The tangential fields E_ρ, H_φ must be continuous at the interface $|z| = L/2$. Equivalently

$$\varepsilon_r E_{z1} = E_{z2}, \quad \frac{\partial E_{z1}}{\partial z} = \frac{\partial E_{z2}}{\partial z}, \quad z = \pm\frac{L}{2}.$$

From the above equations, we obtain

$$A_m\varepsilon_r\cos\left(\frac{\beta L}{2} + \psi\right) = B_m,$$

$$A_m\beta\sin\left(\frac{\beta L}{2} + \psi\right) = \alpha B_m.$$

Thus

$$\beta\tan\left(\frac{\beta L}{2} + \psi\right) = \varepsilon_r\alpha. \tag{3.77}$$

Similarly, we may let

$$\psi = -\frac{p\pi}{2}, \quad p = 0, 1, 2, \ldots.$$

It follows from (3.77) that

$$\beta L = p\pi + 2\arctan\frac{\varepsilon_r \alpha}{\beta} = (p + \delta)\pi, \tag{3.78}$$

where $\delta = \frac{2}{\pi}\arctan\frac{\varepsilon_r\alpha}{\beta}$ with $0 < \delta < 1$. The TM modes in the circular cylindrical dielectric resonator are called $\mathrm{TM}_{mn(p+\delta)}$ modes and the lowest mode is $\mathrm{TM}_{11\delta}$.

3.3.3 Integral Equation for Dielectric Resonators

It will be assumed that the dielectric resonator with medium parameters μ and ε is finite and homogeneous, which occupies the region V bounded by S, as shown in Figure 3.13. The total fields outside and inside the resonator can be represented respectively by

$$\begin{aligned}
\mathbf{E}(\mathbf{r}) = &-\int_S j\omega\mu_0 G_0(\mathbf{r}, \mathbf{r}')\mathbf{J}_s(\mathbf{r}')dS(\mathbf{r}') - \int_S \mathbf{J}_{ms}(\mathbf{r}') \times \nabla' G_0(\mathbf{r}, \mathbf{r}')dS(\mathbf{r}') \\
&- \frac{1}{j\omega\varepsilon_0}\int_S \nabla'_s \cdot \mathbf{J}_s(\mathbf{r}')\nabla' G_0(\mathbf{r}, \mathbf{r}')dS(\mathbf{r}'), \\
\mathbf{H}(\mathbf{r}) = &-\int_S j\omega\varepsilon_0 G_0(\mathbf{r}, \mathbf{r}')\mathbf{J}_{ms}(\mathbf{r}')dS(\mathbf{r}') + \int_S \mathbf{J}_s(r') \times \nabla' G_0(\mathbf{r}, \mathbf{r}')dS(\mathbf{r}') \\
&- \frac{1}{j\omega\mu_0}\int_S \nabla'_s \cdot \mathbf{J}_{ms}(\mathbf{r}')\nabla' G_0(\mathbf{r}, \mathbf{r}')dS(\mathbf{r}'),
\end{aligned}$$

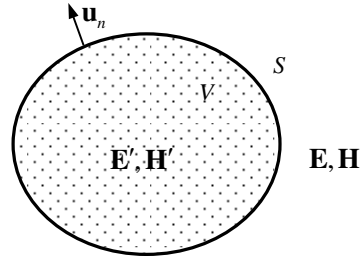


Figure 3.13 An arbitrary dielectric resonator.

and

$$\mathbf{E}'(\mathbf{r}) = \int_S j\omega\mu \mathbf{J}_s(\mathbf{r}')G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}') + \int_S \mathbf{J}_{ms}(\mathbf{r}') \times \nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$+ \frac{1}{j\omega\varepsilon} \int_S \nabla'_s \cdot \mathbf{J}_s(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}'),$$

$$\mathbf{H}'(\mathbf{r}) = \int_S j\omega\varepsilon \mathbf{J}_{ms}(\mathbf{r}')G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}') - \int_S \mathbf{J}_s(\mathbf{r}') \times \nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$+ \frac{1}{j\omega\mu} \int_S \nabla'_s \cdot \mathbf{J}_{ms}(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}').$$

Here

$$G_0(\mathbf{r},\mathbf{r}') = \frac{e^{-jk_0|\mathbf{r}-\mathbf{r}'|}}{4\pi\,|\mathbf{r}-\mathbf{r}'|}, \quad G(\mathbf{r},\mathbf{r}') = \frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{4\pi\,|\mathbf{r}-\mathbf{r}'|},$$
$$\mathbf{J}_s = \mathbf{u}_n \times \mathbf{H} = \mathbf{u}_n \times \mathbf{H}', \quad \mathbf{J}_{ms} = -\mathbf{u}_n \times \mathbf{E} = -\mathbf{u}_n \times \mathbf{E}',$$
$$k_0 = \omega\sqrt{\mu_0\varepsilon_0}, \quad k = \omega\sqrt{\mu\varepsilon},$$

where use is made of the fact that the tangential components of the total fields must be continuous across S

$$(\mathbf{u}_n \times \mathbf{E})_+ = (\mathbf{u}_n \times \mathbf{E}')_-, \quad (\mathbf{u}_n \times \mathbf{H})_+ = (\mathbf{u}_n \times \mathbf{H}')_-. \tag{3.79}$$

If the observation point $\mathbf{r}$ approaches to a point of S and the jump relations are used, we obtain

$$\mathbf{E}(\mathbf{r}) = \frac{1}{2}\mathbf{u}_n(\mathbf{r}) \times \mathbf{J}_{ms}(\mathbf{r}) - \frac{1}{j2\omega\varepsilon_0}\mathbf{u}_n(\mathbf{r})\nabla_s \cdot \mathbf{J}_s(\mathbf{r})$$
$$- \int_S j\omega\mu_0 G_0(\mathbf{r},\mathbf{r}')\mathbf{J}_s(\mathbf{r}')dS(\mathbf{r}') - \int_S \mathbf{J}_{ms}(\mathbf{r}') \times \nabla' G_0(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$- \frac{1}{j\omega\varepsilon_0} \int_S \nabla'_s \cdot \mathbf{J}_s(\mathbf{r}')\nabla' G_0(\mathbf{r},\mathbf{r}')dS(\mathbf{r}'),$$

$$\mathbf{H}(\mathbf{r}) = -\frac{1}{2}\mathbf{u}_n(\mathbf{r}) \times \mathbf{J}_s(\mathbf{r}) - \frac{1}{j2\omega\mu_0}\mathbf{u}_n(\mathbf{r})\nabla_s \cdot \mathbf{J}_{ms}(\mathbf{r})$$
$$- \int_S j\omega\varepsilon_0 G_0(\mathbf{r},\mathbf{r}')\mathbf{J}_{ms}(\mathbf{r}')dS(\mathbf{r}') + \int_S \mathbf{J}_s(\mathbf{r}') \times \nabla' G_0(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$- \frac{1}{j\omega\mu_0} \int_S \nabla'_s \cdot \mathbf{J}_{ms}(\mathbf{r}')\nabla' G_0(\mathbf{r},\mathbf{r}')dS(\mathbf{r}'),$$

$$\mathbf{E}'(\mathbf{r}) = \frac{1}{2}\mathbf{u}_n(\mathbf{r}) \times \mathbf{J}_{ms}(\mathbf{r}) - \frac{1}{j2\omega\varepsilon}\mathbf{u}_n(\mathbf{r})\nabla_s \cdot \mathbf{J}_s(\mathbf{r})$$
$$+ \int_S j\omega\mu\mathbf{J}_s(\mathbf{r}')G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}') + \int_S \mathbf{J}_{ms}(\mathbf{r}') \times \nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$+ \frac{1}{j\omega\varepsilon}\int_S \nabla'_s \cdot \mathbf{J}_s(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}'),$$

$$\mathbf{H}'(\mathbf{r}) = -\frac{1}{2}\mathbf{u}_n(\mathbf{r}) \times \mathbf{J}_s(\mathbf{r}) - \frac{1}{j2\omega\mu}\mathbf{u}_n(\mathbf{r})\nabla_s \cdot \mathbf{J}_{ms}(\mathbf{r})$$
$$+ \int_S j\omega\varepsilon\mathbf{J}_{ms}(\mathbf{r}')G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}') - \int_S \mathbf{J}_s(\mathbf{r}') \times \nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$+ \frac{1}{j\omega\mu}\int_S \nabla'_s \cdot \mathbf{J}_{ms}(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}').$$

Multiplying these equations vectorially by $\mathbf{u}_n$ yields

$$-\frac{1}{2}\varepsilon_0\mathbf{J}_{ms}(\mathbf{r}) = -\mathbf{u}_n(\mathbf{r}) \times \int_S j\omega\mu_0\varepsilon_0 G_0(\mathbf{r},\mathbf{r}')\mathbf{J}_s(\mathbf{r}')dS(\mathbf{r}')$$
$$- \mathbf{u}_n(\mathbf{r}) \times \int_S \mathbf{J}_{ms}(\mathbf{r}') \times \varepsilon_0\nabla' G_0(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$- \frac{1}{j\omega}\mathbf{u}_n(\mathbf{r}) \times \int_S \nabla'_s \cdot \mathbf{J}_s(\mathbf{r}')\nabla' G_0(\mathbf{r},\mathbf{r}')dS(\mathbf{r}'), \qquad (3.80)$$

$$\frac{1}{2}\mu_0\mathbf{J}_s(\mathbf{r})(\mathbf{r}) = -\mathbf{u}_n(\mathbf{r}) \times \int_S j\omega\mu_0\varepsilon_0 G_0(\mathbf{r},\mathbf{r}')\mathbf{J}_{ms}(\mathbf{r}')dS(\mathbf{r}')$$
$$+ \mathbf{u}_n(\mathbf{r}) \times \int_S \mathbf{J}_s(\mathbf{r}') \times \mu_0\nabla' G_0(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$- \frac{1}{j\omega}\mathbf{u}_n(\mathbf{r}) \times \int_S \nabla'_s \cdot \mathbf{J}_{ms}(\mathbf{r}')\nabla' G_0(\mathbf{r},\mathbf{r}')dS(\mathbf{r}'), \qquad (3.81)$$

$$-\frac{1}{2}\varepsilon\mathbf{J}_{ms}(\mathbf{r}) = \mathbf{u}_n(\mathbf{r}) \times \int_S j\omega\mu\varepsilon\mathbf{J}_s(\mathbf{r}')G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$+ \mathbf{u}_n(\mathbf{r}) \times \int_S \mathbf{J}_{ms}(\mathbf{r}') \times \varepsilon\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$+ \frac{1}{j\omega}\mathbf{u}_n(\mathbf{r}) \times \int_S \nabla'_s \cdot \mathbf{J}_s(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}'), \tag{3.82}$$

$$\frac{1}{2}\mu\mathbf{J}_s(\mathbf{r}) = \mathbf{u}_n(\mathbf{r}) \times \int_S j\omega\mu\varepsilon\mathbf{J}_{ms}(\mathbf{r}')G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$- \mathbf{u}_n(\mathbf{r}) \times \int_S \mathbf{J}_s(\mathbf{r}') \times \mu\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$+ \frac{1}{j\omega}\mathbf{u}_n(\mathbf{r}) \times \int_S \nabla'_s \cdot \mathbf{J}_{ms}(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}'). \tag{3.83}$$

Adding (3.80) and (3.82) gives

$$-\frac{1}{2}(\varepsilon_0+\varepsilon)\mathbf{J}_{ms}(\mathbf{r}) + j\mathbf{u}_n(\mathbf{r}) \times \int_S \left[k_0^2 G_0(\mathbf{r},\mathbf{r}') - k^2 G(\mathbf{r},\mathbf{r}')\right]\frac{1}{\omega}\mathbf{J}_s(\mathbf{r}')dS(\mathbf{r}')$$
$$+ \mathbf{u}_n(\mathbf{r}) \times \int_S \mathbf{J}_{ms}(\mathbf{r}') \times [\varepsilon_0\nabla' G_0(\mathbf{r},\mathbf{r}') - \varepsilon\nabla' G(\mathbf{r},\mathbf{r}')]dS(\mathbf{r}')$$
$$+ \frac{1}{j\omega}\mathbf{u}_n(\mathbf{r}) \times \int_S \nabla'_s \cdot \mathbf{J}_s(\mathbf{r}')\left[\nabla' G_0(\mathbf{r},\mathbf{r}') - \nabla' G(\mathbf{r},\mathbf{r}')\right]dS(\mathbf{r}') = 0. \tag{3.84}$$

Adding (3.81) and (3.83) gives

$$\frac{1}{2}(\mu_0+\mu)\mathbf{J}_s(\mathbf{r}) + j\mathbf{u}_n(\mathbf{r}) \times \int_S \left[k_0^2 G_0(\mathbf{r},\mathbf{r}') - k^2 G(\mathbf{r},\mathbf{r}')\right]\frac{1}{\omega}\mathbf{J}_{ms}(\mathbf{r}')dS(\mathbf{r}')$$
$$+ \mathbf{u}_n(\mathbf{r}) \times \int_S \mathbf{J}_s(\mathbf{r}') \times [\mu\nabla' G(\mathbf{r},\mathbf{r}') - \mu_0\nabla' G_0(\mathbf{r},\mathbf{r}')]dS(\mathbf{r}')$$
$$+ \frac{1}{j\omega}\mathbf{u}_n(\mathbf{r}) \times \int_S \nabla'_s \cdot \mathbf{J}_{ms}(\mathbf{r}')\left[\nabla' G_0(\mathbf{r},\mathbf{r}') - \nabla' G(\mathbf{r},\mathbf{r}')\right]dS(\mathbf{r}') = 0. \tag{3.85}$$

Making use of the relation $\int_S \nabla_s \cdot \mathbf{F}(\mathbf{r}) dS(\mathbf{r}) = 0$ for an arbitrary vector field $\mathbf{F}(\mathbf{r})$, the last integral in (3.84) and (3.85) may be written as

$$\begin{aligned}
&\int_S \nabla'_s \cdot \mathbf{J}_s(\mathbf{r}') \left[\nabla' G_0(\mathbf{r},\mathbf{r}') - \nabla' G(\mathbf{r},\mathbf{r}')\right] dS(\mathbf{r}') \\
&\qquad = -\nabla \int_S \nabla'_s \cdot \mathbf{J}_s(\mathbf{r}') \left[G_0(\mathbf{r},\mathbf{r}') - G(\mathbf{r},\mathbf{r}')\right] dS(\mathbf{r}') \\
&\qquad = -\int_S \left[\mathbf{J}_s(\mathbf{r}') \cdot \nabla'\right] \nabla' \left[G_0(\mathbf{r},\mathbf{r}') - G(\mathbf{r},\mathbf{r}')\right] dS(\mathbf{r}').
\end{aligned}$$

Therefore, (3.84) and (3.85) become

$$\begin{aligned}
&-\frac{1}{2}(\varepsilon_0 + \varepsilon)\mathbf{J}_{ms}(\mathbf{r}) + j\mathbf{u}_n(\mathbf{r}) \times \int_S \left[k_0^2 G_0(\mathbf{r},\mathbf{r}') - k^2 G(\mathbf{r},\mathbf{r}')\right] \frac{1}{\omega}\mathbf{J}_s(\mathbf{r}') dS(\mathbf{r}') \\
&\quad + \mathbf{u}_n(\mathbf{r}) \times \int_S \mathbf{J}_{ms}(\mathbf{r}') \times \left[\varepsilon_0 \nabla' G_0(\mathbf{r},\mathbf{r}') - \varepsilon \nabla' G(\mathbf{r},\mathbf{r}')\right] dS(\mathbf{r}') \\
&\quad - \frac{1}{j\omega}\mathbf{u}_n(\mathbf{r}) \times \int_S \left[\mathbf{J}_s(\mathbf{r}') \cdot \nabla'\right] \left[\nabla' G_0(\mathbf{r},\mathbf{r}') - \nabla' G(\mathbf{r},\mathbf{r}')\right] dS(\mathbf{r}') = 0,
\end{aligned} \tag{3.86}$$

$$\begin{aligned}
&\frac{1}{2}(\mu_0 + \mu)\mathbf{J}_s(\mathbf{r}) + j\mathbf{u}_n(\mathbf{r}) \times \int_S \left[k_0^2 G_0(\mathbf{r},\mathbf{r}') - k^2 G(\mathbf{r},\mathbf{r}')\right] \frac{1}{\omega}\mathbf{J}_{ms}(\mathbf{r}') dS(\mathbf{r}') \\
&\quad + \mathbf{u}_n(\mathbf{r}) \times \int_S \mathbf{J}_s(\mathbf{r}') \times \left[\mu \nabla' G(\mathbf{r},\mathbf{r}') - \mu_0 \nabla' G_0(\mathbf{r},\mathbf{r}')\right] dS(\mathbf{r}') \\
&\quad - \frac{1}{j\omega}\mathbf{u}_n(\mathbf{r}) \times \int_S \left[\mathbf{J}_{ms}(\mathbf{r}') \cdot \nabla'\right] \left[\nabla' G_0(\mathbf{r},\mathbf{r}') - \nabla' G(\mathbf{r},\mathbf{r}')\right] dS(\mathbf{r}') = 0.
\end{aligned} \tag{3.87}$$

Equations (3.86) and (3.87) are the integral equations for an arbitrary dielectric resonator, and may be rewritten as

$$\begin{aligned}
&-\frac{1}{2}\left(\frac{k_0}{\eta_0} + \frac{k}{\eta}\right)\mathbf{J}_{ms}(\mathbf{r}) + j\mathbf{u}_n(\mathbf{r}) \times \int_S \left[k_0^2 G_0(\mathbf{r},\mathbf{r}') - k^2 G(\mathbf{r},\mathbf{r}')\right] \mathbf{J}_s(\mathbf{r}') dS(\mathbf{r}') \\
&\quad + \mathbf{u}_n(\mathbf{r}) \times \int_S \mathbf{J}_{ms}(\mathbf{r}') \times \left[\frac{k_0}{\eta_0} \nabla' G_0(\mathbf{r},\mathbf{r}') - \frac{k}{\eta} \nabla' G(\mathbf{r},\mathbf{r}')\right] dS(\mathbf{r}') \\
&\quad - \frac{1}{j}\mathbf{u}_n(\mathbf{r}) \times \int_S \left[\mathbf{J}_s(\mathbf{r}') \cdot \nabla'\right] \left[\nabla' G_0(\mathbf{r},\mathbf{r}') - \nabla' G(\mathbf{r},\mathbf{r}')\right] dS(\mathbf{r}') = 0,
\end{aligned} \tag{3.88}$$

$$\frac{1}{2}\left(k_0\eta_0 + k\eta\right)\mathbf{J}_s(\mathbf{r}) + j\mathbf{u}_n(\mathbf{r}) \times \int_S \left[k_0^2 G_0(\mathbf{r},\mathbf{r}') - k^2 G(\mathbf{r},\mathbf{r}')\right]\mathbf{J}_{ms}(\mathbf{r}')dS(\mathbf{r}')$$
$$+ \mathbf{u}_n(\mathbf{r}) \times \int_S \mathbf{J}_s(\mathbf{r}') \times [k\eta\nabla' G(\mathbf{r},\mathbf{r}') - k_0\eta_0\nabla' G_0(\mathbf{r},\mathbf{r}')]dS(\mathbf{r}')$$
$$- \frac{1}{j}\mathbf{u}_n(\mathbf{r}) \times \int_S \left[\mathbf{J}_{ms}(\mathbf{r}') \cdot \nabla'\right]\left[\nabla' G_0(\mathbf{r},\mathbf{r}') - \nabla' G(\mathbf{r},\mathbf{r}')\right]dS(\mathbf{r}') = 0, \tag{3.89}$$

where $\eta_0 = \sqrt{\mu_0/\varepsilon_0},\ \eta = \sqrt{\mu/\varepsilon}$. The integral equations (3.88) and (3.89) can be used to determine the resonant frequencies of an arbitrary dielectric resonator (Geyi and Hongshi, 1988b).

3.4 Microstrip Resonators

A microstrip resonator consists of a metallic patch on a grounded dielectric substrate, as shown in Figure 3.14. It has found wide applications in filters, oscillators, mixers, and circulators, and can also be used as an antenna. The microstrip resonator can be analyzed by using the magnetic wall model. We use the rectangular microstrip resonator as an example to illustrate the procedure.

The magnetic wall model is based on the assumption that all the side faces of the rectangular microstrip resonator are magnetic walls while the top and the bottom faces are electric walls, as shown in Figure 3.15. The width and length of the rectangular patch are denoted by W and L, respectively. The height h of the substrate of relative dielectric constant ε_r is assumed to be much smaller than a wavelength. In this case, the field inside the resonator is basically a TM wave with respect to y direction, and we have $E_y \neq 0, H_y = 0$ and the fields are independent of y. As a result, we have

$$H_x = -j\frac{\omega\varepsilon}{k^2}\frac{\partial E_y}{\partial z}, \quad H_z = j\frac{\omega\varepsilon}{k^2}\frac{\partial E_y}{\partial x}$$

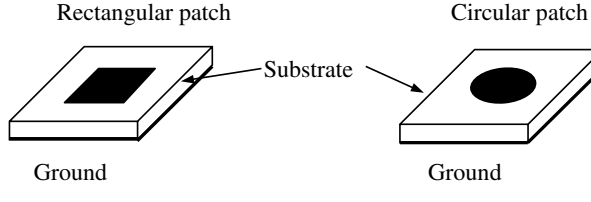


Figure 3.14 Rectangular microstrip resonator and circular microstrip disk resonator.

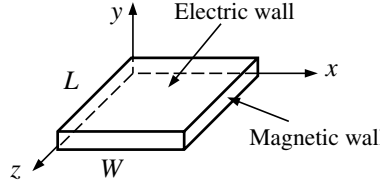


Figure 3.15 Magnetic wall model for the rectangular microstrip resonator.

with $E_x = E_z = H_y = 0$. The normal derivative of E_y on the magnetic walls must be zero and this leads to

$$E_y = A_{mn} \cos\Big(\frac{m\pi}{W}x\Big) \cos\Big(\frac{n\pi}{L}z\Big), \quad m, n = 0, 1, 2, \ldots,$$

where m and n are not zero simultaneously. The resonant wavenumbers are

$$k_{m0n} = \sqrt{\Big(\frac{m\pi}{W}\Big)^2 + \Big(\frac{n\pi}{L}\Big)^2},$$

and the resonant frequencies are given by

$$f_0 = \frac{c}{2\sqrt{\varepsilon_r}} \sqrt{\Big(\frac{m\pi}{W}\Big)^2 + \Big(\frac{n\pi}{L}\Big)^2},$$

where $c = 1/\sqrt{\mu_0 \varepsilon_0}$.

3.5 Open Resonators

Dicke (1958), Prokhorov (1958), and Schawlow and Townes (1958) independently proposed to use the Fabry–Perot interferometer as a laser resonator (Fabry and Perot, 1899). An open resonator in its simplest form consists of two mirrors facing each other, between which the field is reflected back and forth to form a standing wave. The open resonators are widely used in microwave and optical frequency range as an oscillatory system. The theory of the open resonator has been investigated by many researchers (Fox and Li, 1961; Kogelnik and Li, 1966).

3.5.1 Paraxial Approximations

The coherent radiation generated by lasers or masers operating in the optical or infrared wavelength regions usually appears as a beam whose transverse extent is large compared to the wavelength. The properties of such beam in the resonant structure have been studied extensively

(Boyd and Gordon, 1961; Goubau and Schwering, 1961). A laser beam is very similar to a plane wave but its transverse amplitude distribution is not uniform and is concentrated near the axis of propagation with phase front being slightly curved. A component ϕ of the electromagnetic fields satisfies the Helmholtz equation

$$(\nabla^2 + k^2)\phi(x, y, z) = 0, \tag{3.90}$$

where k is the wavenumber. The scalar wave function ϕ may be represented by two dimensional Fourier transform as follows

$$\phi(x, y, z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tilde{\phi}(k_x, k_y, z)e^{-j(k_x x + k_y y)}dk_x\, dk_y, \tag{3.91}$$

where

$$\tilde{\phi}(k_x, k_y, z) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \phi(x, y, z)e^{j(k_x x + k_y y)}dx\, dy \tag{3.92}$$

is the spatial spectrum of the scalar wave function ϕ in the (x, y)-plane. On substituting (3.91) into (3.90), we may find that the spatial spectrum $\tilde{\phi}$ satisfies

$$\frac{d^2\tilde{\phi}}{dz^2} + (k^2 - k_x^2 - k_y^2)\tilde{\phi} = 0. \tag{3.93}$$

The solution of the above equation can be written as

$$\tilde{\phi}(k_x, k_y, z) = \tilde{\phi}^a(k_x, k_y)e^{-j\sqrt{k^2 - k_x^2 - k_y^2}z}, \tag{3.94}$$

where $\tilde{\phi}^a(k_x, k_y)$ is the amplitude independent of z. If the scalar wave function ϕ is a spatial wave packet propagating in the z-direction of a rectangular system, the spatial spectrum $\tilde{\phi}(k_x, k_y, z)$ is significant only for $|k_x| \ll k,\ |k_y| \ll k$. Thus, we have

$$j\sqrt{k^2 - k_x^2 - k_y^2}z \approx jkz - \frac{j}{2k}\left(k_x^2 + k_y^2\right)z. \tag{3.95}$$

Inserting this into (3.94) yields

$$\tilde{\phi}(k_x, k_y, z) = \tilde{\phi}^a(k_x, k_y)e^{\frac{j}{2k}\left(k_x^2 + k_y^2\right)z}e^{-jkz}.$$

Introducing the above expression into (3.91), we obtain

$$\phi(x,y,z) = \phi^a(x,y,z)e^{-jkz}, \tag{3.96}$$

where

$$\phi^a(x,y,z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tilde{\phi}^a(k_x,k_y)e^{\frac{j}{2k}\left(k_x^2+k_y^2\right)z}e^{-j(k_x x+k_y y)}dk_x\,dk_y \tag{3.97}$$

is slowly varying amplitude with respect to the coordinate z. Note that

$$\phi(x,y,0) = \phi^a(x,y,0) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tilde{\phi}^a(k_x,k_y)e^{-j(k_x x+k_y y)}dk_x\,dk_y \tag{3.98}$$

and this implies

$$\tilde{\phi}^a(k_x,k_y) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \phi^a(x,y,0)e^{j(k_x x+k_y y)}dx\,dy. \tag{3.99}$$

Substituting (3.99) into (3.97), we obtain

$$\phi^a(x,y,z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x-\xi,y-\eta)\phi^a(\xi,\eta,0)d\xi\,d\eta, \tag{3.100}$$

where

$$\begin{aligned} &G(x-\xi,y-\eta)\\ &\quad= \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left\{\frac{j\left(k_x^2+k_y^2\right)z}{2k} - j\left[k_x(x-\xi)+k_y(y-\eta)\right]\right\}dk_x\,dk_y\\ &\quad= \frac{jk}{2\pi z}\exp\left\{-\frac{jk}{2z}\left[(x-\xi)^2+(y-\eta)^2\right]\right\}. \end{aligned} \tag{3.101}$$

On substituting (3.96) into (3.90), we may find the slowly varying amplitude $\phi^a(x,y,z)$ satisfies the **paraxial wave equation**

$$\frac{\partial^2\phi^a}{\partial x^2} + \frac{\partial^2\phi^a}{\partial y^2} - 2jk\frac{\partial\phi^a}{\partial z} = 0, \tag{3.102}$$

which is a parabolic equation. In circular cylindrical system, the paraxial wave equation may be written as

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial\phi^a}{\partial\rho}\right)+\frac{1}{\rho^2}\frac{\partial^2\phi^a}{\partial\varphi^2}-2jk\frac{\partial\phi^a}{\partial z}=0. \tag{3.103}$$

As an example, we assume that there is a point source at $z=0$

$$\phi^a(\xi,\eta,0)=C\delta\left(\frac{\xi}{a}\right)\delta\left(\frac{\eta}{a}\right),$$

where C is a constant. From (3.100), we obtain

$$\begin{aligned}\phi^a(x,y,z)&=\frac{jka^2}{2\pi z}\exp\left(-jk\tfrac{x^2+y^2}{2z}\right),\\ \phi(x,y,z)&=C\frac{jka^2}{2\pi z}\exp\left[-jk\left(z+\frac{x^2+y^2}{2z}\right)\right].\end{aligned} \tag{3.104}$$

As a result of paraxial approximation, the spherical wave front generated by a point source has been replaced by a parabolic wave front. This approximation is appropriate when the wave varies only gradually along the propagating axis.

Another example is to assume a plane wave with a transverse Gaussian distribution at $z=0$

$$\phi^a(\xi,\eta,0)=C\exp\left(-\frac{\rho^2}{a^2}\right),$$

where $\rho^2=x^2+y^2$ and a is the Gaussian beam width. Substituting this into (3.100), we have

$$\begin{aligned}\phi^a(x,y,z)&=C\frac{jk}{2\pi z}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{-\frac{jk}{2z}\left[(x-\xi)^2+(y-\eta)^2\right]}e^{-\frac{\xi^2+\eta^2}{a^2}}d\xi\,d\eta\\ &=\frac{C}{1-jD}\exp\left[-\frac{\rho^2}{a^2(1-jD)}\right]\\ &=\frac{C}{\sqrt{1+D^2}}\exp\left[-\frac{\rho^2}{a^2(1+D^2)}-j\left(\frac{\rho^2}{a^2}\frac{D}{1+D^2}-\arctan D\right)\right],\end{aligned} \tag{3.105}$$

where $D=2z/(ka)^2$. It can be seen that the beam width $a^2(1+D^2)$ increases with the propagating distance z. The wave front distortion is also observed as the distance z increases.

3.5.2 Modes in Open Resonators

In rectangular coordinate system, the modal solutions of (3.102) are given by

$$\phi^a_{mn}(x,y,z) = \frac{w_0}{w(z)} H_m\left(\frac{\sqrt{2}x}{w(z)}\right) H_n\left(\frac{\sqrt{2}y}{w(z)}\right) \exp\left[-\frac{x^2+y^2}{w^2(z)}\right] \times \exp\left[j(m+n+1)\arctan\left(\frac{z}{z_0}\right) - jk\frac{x^2+y^2}{2R(z)}\right], \quad (3.106)$$

where H_m and H_n are Hermite polynomials of mth and nth order respectively with m and n being non-negative integers. Other parameters are defined by

$$z_0 = \frac{kw_0^2}{2}, \quad w(z) = w_0\left(1+\frac{z^2}{z_0^2}\right)^{1/2}, \quad R(z) = z + \frac{z_0^2}{z}.$$

In circular cylindrical system, the modal solution of (3.103) are given by

$$\phi^a_{mn}(x,y,z) = \frac{w_0}{w(z)}\left(\frac{\sqrt{2}\rho}{w(z)}\right)^m L^n_m\left(\frac{\sqrt{2}\rho}{w(z)}\right) \exp\left[-\frac{\rho^2}{w^2(z)}\right] \times \exp\left[j(2n+m+1)\arctan\left(\frac{z}{z_0}\right) - jk\frac{\rho^2}{2R(z)} - m\varphi\right], \quad (3.107)$$

where L^n_m are the generalized Laguerre polynomials, m is an integer and n is a non-negative integer. Lasers are often made to operate in the lowest mode which corresponds to $m = n = 0$ and is called TEM_{00} mode. For the lowest TEM_{00} mode, both (3.106) and (3.107) reduce to

$$\phi^a_{00}(x,y,z) = \frac{w_0}{w(z)} \exp\left[-\frac{x^2+y^2}{w^2(z)}\right] \exp\left[j\arctan\left(\frac{z}{z_0}\right) - jk\frac{x^2+y^2}{2R(z)}\right]. \quad (3.108)$$

Let us consider an open resonator consisting of two identical parallel square metal plate (plane mirror) of width $2a$ separated by a distance $2l$, as shown in Figure 3.16. We use ϕ to denote either field component E_y or E_x. Then ϕ satisfies (3.90) and the following boundary conditions

$$\phi = 0, \quad |x| \le a, \quad |y| \le b, \quad z = \pm l. \quad (3.109)$$

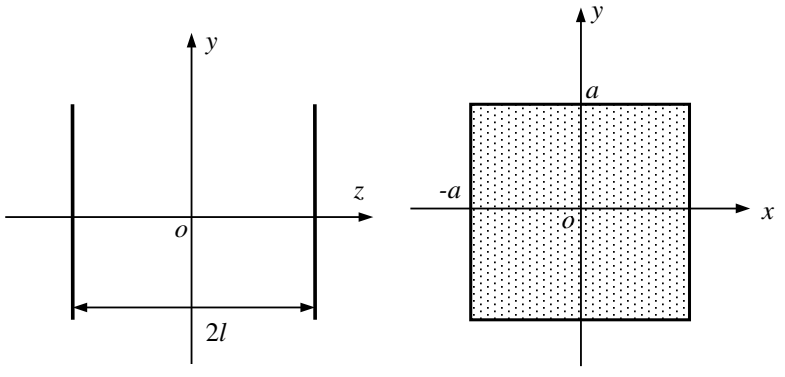


Figure 3.16 Parallel-plate resonator.

If the mirrors are sufficiently large compared to wavelength, the total reflection happens for the modes of any order. The modal field inside the open resonator can be written as sum of the modes of left-traveling and right-traveling

$$\phi(x,y,z) = A_{mn}\phi^a_{mn}(x,y,z)e^{-jkz} + B_{mn}\phi^a_{mn}(x,y,-z)e^{jkz}.$$

The application of the boundary conditions (3.109) yields

$$\phi(x,y,-l) = A_{mn}\phi^a_{mn}(x,y,-l)e^{jkl} + B_{mn}\phi^a_{mn}(x,y,l)e^{-jkl} = 0,$$
$$\phi(x,y,l) = A_{mn}\phi^a_{mn}(x,y,l)e^{-jkl} + B_{mn}\phi^a_{mn}(x,y,-l)e^{jkl} = 0,$$

from which we obtain

$$2kl - 2(m+n+1)\arctan\left(\frac{l}{z_0}\right) = \pi q, \quad q = 0,1,2,\ldots, \tag{3.110}$$

where q is the number of nodes of the axial standing wave pattern. The above equation can be used to determine the resonant frequencies of various modes

$$k_{mnq} = \frac{\pi q}{2l} + \frac{m+n+1}{l}\arctan\left(\frac{l}{z_0}\right). \tag{3.111}$$

In the preceding discussions, the diffraction effects due to the finite size of the mirrors have been ignored.

> Mathematical analysis is as extensive as nature itself; it defines all perceptible relations, measures times, spaces, forces, temperatures; this

difficult science is formed slowly, but it preserves every principle which it has once acquired; it grows and strengthens itself incessantly in the midst of the many variations and errors of the human mind.

—Jean Baptiste Joseph Fourier
(French mathematician, 1768–1830)

Chapter 4

Microwave Circuits

> The ability to simplify means eliminating the unnecessary so that the necessary can speak.
>
> —Hans Hofmann (American artist, 1880–1966)

At frequencies where the wavelength is considerably larger than the greatest dimensions of an electronic system being examined, the system can be built with the usual lumped circuit elements such as resistors, inductors and capacitors, and can be analyzed by the standard loop and node voltages without considering the time delay at different points in the system. In the microwave frequency range where the wavelength is comparable to the largest dimensions of the system, the time delay, the wave propagation effects, the radiation by the current distribution in the circuits, and the effects of the distributed capacitance and inductance in the connecting leads and terminals can no longer be ignored. As a result, the conventional lumped circuit elements no longer behave in the desired manner as usual and must be replaced by transmission line and waveguide components. These distributed circuit elements may take a variety of forms and their equivalent circuit parameters or network matrix elements may be obtained from the field analysis.

A microwave circuit may consist of a number of distributed circuit elements, including passive or active devices, and is usually studied by scattering parameters. A typical microwave circuit may contain devices to achieve various functions such as frequency conversions, impedance matching, and power manipulations. In general, these devices can be classified according to their functions as frequency-related, impedance-related and power-related as illustrated in Table 4.1. The frequency is a fundamental parameter in RF engineering and it determines the implementing structures of RF circuits and the materials to be used. The frequency-related devices

Table 4.1 Classification of microwave devices

Microwave devices	Examples
Frequency-related devices	Oscillators, resonators, frequency synthesizers, frequency dividers and multipliers, mixers, filters, etc.
Impedance-related devices	Impedance transformers, impedance matching circuits, phase shifters, antennas, etc.
Power-related devices	Power dividers, directional couplers, attenuators, amplifiers, switches, etc.

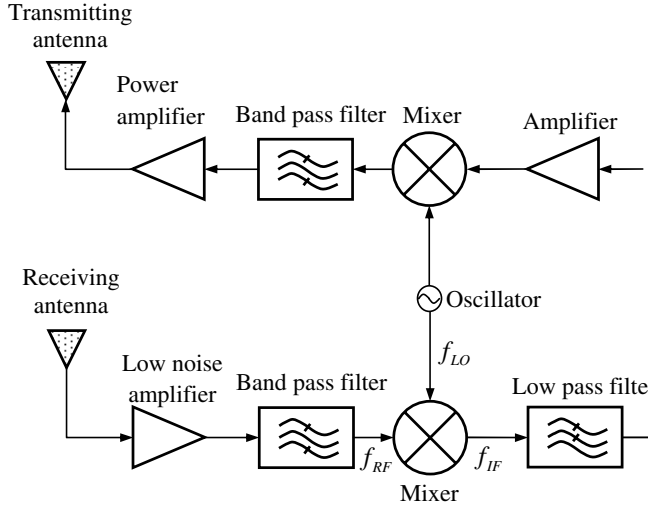


Figure 4.1 A typical RF heterodyne transceiver.

include signal generators, frequency converters and frequency selective circuits. The impedance is another important parameter that is used to characterize the effects of circuits on the transmission of microwave signals. In RF engineering, a central task is to design various matching circuits to achieve maximum power transfer between different devices. The power is used to measure the strength of signals and the final design target of various microwave circuits or system is to realize optimal power transfer from one part to another in the system.

The block diagram of a typical RF heterodyne transmitter/receiver system is shown in Figure 4.1. The performance of this system is determined by the power delivered to the transmitting antenna as well as the sensitivity of the receiver, which is defined as the minimum signal level that the system

can detect with acceptable signal-to-noise ratio. Most of the microwave circuit devices contained in Figure 4.1 will be discussed in this chapter.

4.1 Circuit Theory of Transmission Lines

The transmission lines or waveguides are used to transmit microwave signals. They are also used extensively in microwave circuit designs, such as directional couplers, filters, and power dividers. In most practical situations, the waveguides are in a state of single-mode operation, and can be characterized by the conventional circuit theory of transmission lines.

4.1.1 Transmission Line Equations

In the time domain, the voltage and current along a short section of transmission line, as shown in Figure 4.2, satisfy the **transmission line equations**:

$$\frac{\partial v(z,t)}{\partial z} = -Ri(z,t) - L\frac{\partial i(z,t)}{\partial t}, \quad \frac{\partial i(z,t)}{\partial z} = -Gv(z,t) - C\frac{\partial v(z,t)}{\partial t},$$

where R, L, G and C are the resistance, inductance, conductance and capacitance per unit length of the transmission line respectively. For

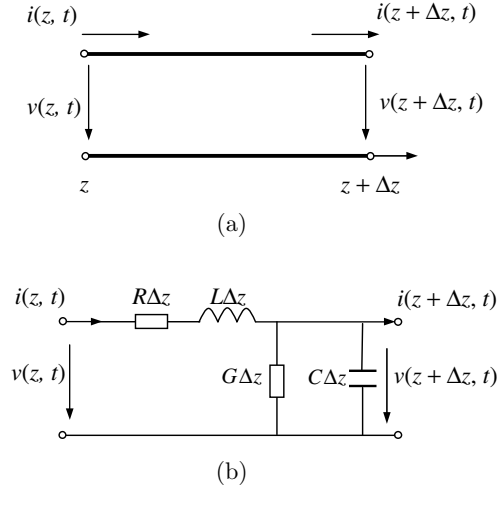


Figure 4.2 (a) A short section of transmission line. (b) Equivalent circuit.

time-harmonic fields, these equations reduce to

$$\frac{dV}{dz} = -IZ_{\text{unit}}, \quad \frac{dI}{dz} = -VY_{\text{unit}}, \tag{4.1}$$

where V and I are phasors, and $Z_{\text{unit}} = R+j\omega L$ and $Y_{\text{unit}} = G+j\omega C$ are the series impedance and shunt admittance per unit length of the transmission line. From the above equations, we obtain

$$\frac{d^2V}{dz^2} - \gamma^2 V = 0, \quad \frac{d^2I}{dz^2} - \gamma^2 I = 0. \tag{4.2}$$

The quantity $\gamma = \sqrt{Z_{\text{unit}}Y_{\text{unit}}} = \alpha + j\beta$ is called the **propagation constant**. The solutions for the voltage and current can be obtained from (4.1) and (4.2) as

$$\begin{aligned} V &= V^+ + V^- = Ae^{-\gamma z} + Be^{\gamma z}, \\ I &= I^+ - I^- = \frac{1}{Z_0}(Ae^{-\gamma z} - Be^{\gamma z}), \end{aligned} \tag{4.3}$$

where $V^+ = Ae^{-\gamma z}$, $V^- = Be^{\gamma z}$, $I^+ = Ae^{-\gamma z}/Z_0$, and $I^- = Be^{\gamma z}/Z_0$ are the incident voltage wave, the reflected voltage wave, the incident current wave, and the reflected current wave respectively; and

$$Z_0 = \sqrt{\frac{Z_{\text{unit}}}{Y_{\text{unit}}}} = \frac{V^+}{I^+} = \frac{V^-}{I^-}$$

is called the **characteristic impedance**. The minus sign in front of I^- implies that the reference direction of I^- is opposite that of I^+. When the time factor is restored, we have

$$V^+ = Ae^{-\alpha z}e^{j(\omega t - \beta z)},$$

which stands for a wave moving along the positive z-direction with an exponential damping factor determined by the attenuation constant α. The phase velocity is the speed of points of constant phase and is given by $v_p = \omega/\beta$. As a result, $\beta = 2\pi/\lambda$, where λ is the wavelength. The **reflection coefficient** Γ at position z is defined by

$$\Gamma = \frac{V^-}{V^+} = \frac{B}{A}e^{2\gamma z} = \Gamma_L e^{2\gamma z},$$

where $\Gamma_L = B/A$ is the reflection coefficient at $z = 0$ (Figure 4.3), called the **load reflection coefficient**. The input impedance at z can be obtained

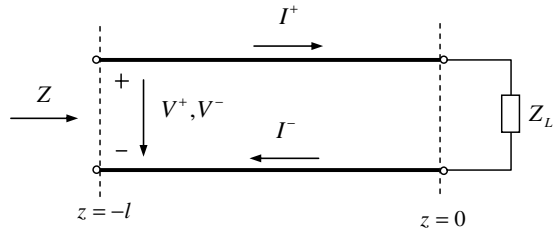


Figure 4.3 Transmission line terminated in a load.

from (4.3)

$$Z = \frac{V}{I} = \frac{V^+ + V^-}{I^+ - I^-} = \frac{V^+}{I^+}\frac{1 + V^-/V^+}{1 - I^-/I^+} = Z_0\frac{1+\Gamma}{1-\Gamma} = Z_0\frac{1+\Gamma_L e^{2\gamma z}}{1-\Gamma_L e^{2\gamma z}}. \tag{4.4}$$

The reflection coefficient is

$$\Gamma = \frac{Z - Z_0}{Z + Z_0}. \tag{4.5}$$

It follows from (4.4) and (4.5) that

$$Z = Z_0\frac{Z_L + Z_0\tanh(-\gamma z)}{Z_0 + Z_L\tanh(-\gamma z)}. \tag{4.6}$$

For a lossless transmission line, the input impedance at $z = -l$ becomes

$$Z = Z_0\frac{Z_L + jZ_0\tan(\beta l)}{Z_0 + jZ_L\tan(\beta l)}. \tag{4.7}$$

Example 4.1: For the matched case: $Z_L = Z_0$, we have $Z = Z_0$. For the open circuit: $Z_L = \infty$, we have $Z = Z_0/j\tan\beta l$. For the short circuit: $Z_L = 0$, we have $Z = jZ_0\tan\beta l$. When $l = \lambda/4$, we have $Z = Z_0^2/Z_L$, which is called the **quarter wavelength transform**. □

Remark 4.1 (TEM transmission line parameters)**:** The circuit parameters R, G, L, C for a TEM transmission line shown in Figure 4.4 can be determined from the field analysis. For the TEM mode in the line, the electric field $\mathbf{E}$ propagating in $+z$-direction can be expressed as the gradient of a scalar function ϕ, i.e., $\mathbf{E} = -\nabla\phi$. If the medium surrounding the conductors has finite conductivity σ_m and a complex permittivity $\varepsilon = \varepsilon' - j\varepsilon''$, the charge density on the conductors can be found from the normal component of the electric field at the surface: $\rho_s = \varepsilon'\mathbf{u}_n \cdot \mathbf{E} = -\varepsilon'\partial\phi/\partial n$. The total

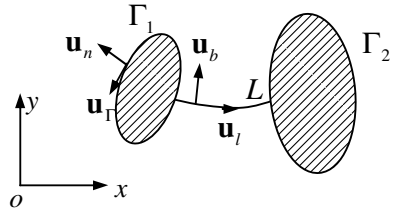


Figure 4.4 Parameters for TEM transmission line.

charge per unit length on conductor Γ_1 is

$$Q = \int_{\Gamma_1} \varepsilon' \mathbf{u}_n \cdot \mathbf{E}\, d\Gamma.$$

The capacitance C per unit length is then given by the ratio of the total charge Q and the voltage V across the two conductors

$$C = \frac{Q}{V} = \frac{\int_{\Gamma_1} \varepsilon' \mathbf{u}_n \cdot \mathbf{E}\, d\Gamma}{\int_L \mathbf{E} \cdot \mathbf{u}_l\, dl}. \tag{4.8}$$

The total current on Γ_1 is given by the line integral of the surface current density $\mathbf{J}_s = \mathbf{u}_n \times \mathbf{H}$ as follows

$$\begin{aligned} I = \int_{\Gamma_1} \mathbf{J}_s \cdot \mathbf{u}_z\, d\Gamma &= \int_{\Gamma_1} \mathbf{H} \cdot \mathbf{u}_\Gamma\, d\Gamma = \int_{\Gamma_1} \frac{1}{\eta} \mathbf{u}_z \times \mathbf{E} \cdot \mathbf{u}_\Gamma\, d\Gamma \\ &= \int_{\Gamma_1} \frac{1}{\eta} \mathbf{u}_n \cdot \mathbf{E}\, d\Gamma = \frac{Q}{\eta \varepsilon'}, \end{aligned}$$

where the wave impedance of the medium can be approximated by

$$\eta = \left(\frac{\mu}{\varepsilon - j\sigma_m/\omega} \right)^{1/2} = \left(\frac{\mu}{\varepsilon' - j\varepsilon'' - j\sigma_m/\omega} \right)^{1/2} \approx \sqrt{\frac{\mu}{\varepsilon'}}. \tag{4.9}$$

The characteristic impedance of the line is given by

$$Z_0 = \frac{V}{I} = \frac{\varepsilon' \eta}{C}. \tag{4.10}$$

The flux per unit length is (see Figure 4.4)

$$\psi = \int_L \mu \mathbf{H} \cdot \mathbf{u}_b\, dl = \int_L \mu \frac{1}{\eta} \mathbf{u}_z \times \mathbf{E} \cdot \mathbf{u}_b\, dl = \int_L \mu \frac{1}{\eta} \mathbf{E} \cdot \mathbf{u}_l\, dl = \mu \frac{V}{\eta}.$$

So the inductance per unit length is

$$L = \frac{\psi}{I} = \frac{\mu}{\eta}\frac{V}{I} = \frac{\mu}{\eta}Z_0. \tag{4.11}$$

It follows from (4.10) and (4.11) that

$$Z_0 = \sqrt{\frac{L}{C}}, \quad \mu\varepsilon' = \sqrt{LC}. \tag{4.12}$$

The total shunt current I_s consists of a displacement current I_d and a conduction current I_c. The current leaving conductor Γ_1 per unit length is

$$I_s = I_c + I_d = (\sigma_m + \omega\varepsilon'')\int_{\Gamma_1} \mathbf{E}\cdot\mathbf{u}_n\, d\Gamma + j\omega\varepsilon'\int_{\Gamma_1} \mathbf{E}\cdot\mathbf{u}_n\, d\Gamma,$$

where the first integral on the right gives the conduction current and the second gives the displacement current. The total shunt admittance is given by

$$Y = G + j\omega C$$

where

$$G = \frac{I_c}{V} = \frac{I_c}{I_d}\frac{I_d}{V} = \frac{\sigma_m + \omega\varepsilon''}{\varepsilon'}C, \quad C = \frac{I_d}{j\omega V} = \frac{\varepsilon'}{V}\int_{\Gamma_1} \mathbf{E}\cdot\mathbf{u}_n\, d\Gamma.$$

The series resistance R per unit length caused by the finite conductivity σ of the conductors can be found from the following relation

$$\frac{1}{2}|I|^2 R = \frac{1}{2}\mathrm{Re}\, Z_s \int_{\Gamma_1+\Gamma_2} |\mathbf{J}_s|^2 d\Gamma = \frac{1}{2}\mathrm{Re}\, Z_s \int_{\Gamma_1+\Gamma_2} |\mathbf{H}|^2 d\Gamma, \tag{4.13}$$

where $Z_s = (1+j)/\sigma\delta_s$ is the surface impedance and δ_s is the skin depth.

The finite conductivity of the conductors can increase the series inductance of the line by a small amount L_{int}, called **internal inductance** (see Chapter 7), due to the penetration of the magnetic field into the conductors. The inductance L_{int} can be evaluated from the inductive part of the surface impedance Z_s. The magnetic energy stored in $X_s = \mathrm{Im}\, Z_s$ is

$$\begin{aligned} W_m &= \frac{\mathrm{Im}\, Z_s}{4\omega}\int_{\Gamma_1+\Gamma_2} |\mathbf{J}_s|^2 d\Gamma \\ &= \frac{\mathrm{Im}\, Z_s}{4\omega}\int_{\Gamma_1+\Gamma_2} |\mathbf{H}|^2 d\Gamma = \frac{\mathrm{Im}\, Z_s}{4\omega}\frac{R|I|^2}{\mathrm{Re}\, Z_s} = \frac{R|I|^2}{4\omega}, \end{aligned}$$

where we have used (4.13). We define the inductance L_{int} by

$$W_m = \frac{1}{4} L_{\text{int}} |I|^2,$$

and this gives

$$\omega L_{\text{int}} = R. \tag{4.14}$$

In practice, we have $R \ll \omega L$, which implies $L_{\text{int}} \ll L$. □

Example 4.2 (Coaxial line parameters)**:** For a coaxial transmission line of inner radius a and outer radius b, the fields for the TEM mode propagating in $+z$ direction are given by

$$\mathbf{E} = \frac{V}{\rho \ln(b/a)} \mathbf{u}_\rho, \quad \mathbf{H} = \frac{V}{\eta \rho \ln(b/a)} \mathbf{u}_\varphi. \tag{4.15}$$

where $V = V_0 e^{-j\beta z}$ is the voltage wave and η is given by (4.9). The charge on the inner conductor is

$$Q = \varepsilon' \int_0^{2\pi} \mathbf{u}_\rho \cdot \mathbf{E} a \, d\varphi = \frac{2\pi \varepsilon' V}{\ln(b/a)}.$$

The capacitance per unit length is thus given by

$$C = \frac{Q}{V} = \frac{2\pi \varepsilon'}{\ln(b/a)}. \tag{4.16}$$

The current is

$$I = \int_0^{2\pi} \mathbf{H} \cdot \mathbf{u}_\varphi a \, d\varphi = \frac{2\pi V}{\eta \ln(b/a)}.$$

Hence the characteristic impedance is

$$Z_0 = \frac{V}{I} = \frac{\eta}{2\pi} \ln \frac{b}{a}. \tag{4.17}$$

The flux per unit length is

$$\psi = \mu \int_a^b \mathbf{H} \cdot \mathbf{u}_\varphi \, d\rho = \mu \frac{V}{\eta}.$$

Thus the inductance per unit length is

$$L = \frac{\psi}{I} = \frac{\mu}{2\pi} \ln \frac{b}{a}. \tag{4.18}$$

It follows from (4.16)–(4.18) that

$$Z_0 = \sqrt{\frac{L}{C}}, \quad \mu\varepsilon' = \sqrt{LC}.$$

The shunt conductance is

$$G = \frac{2\pi(\sigma_m + \omega\varepsilon'')}{\ln(b/a)}. \tag{4.19}$$

The series resistance R per unit length is given by

$$R = \frac{\operatorname{Re} Z_s}{|I|^2} \int\limits_{\Gamma_1+\Gamma_2} |\mathbf{H}|^2 d\Gamma = \frac{\operatorname{Re} Z_s}{2\pi} \frac{a+b}{ab}. \quad \square \tag{4.20}$$

4.1.2 Smith Chart

Smith chart is a graphical representation of the input impedance of a length of transmission line as given by (4.4) and was developed by Phillip H. Smith (1905–1987) in 1939. It follows from (4.4) that the normalized input impedance may be written as

$$z = \frac{Z}{Z_0} = r + jx = \frac{1+\Gamma}{1-\Gamma}. \tag{4.21}$$

Substituting $\Gamma = \Gamma_r + j\Gamma_i$ into the above equation, we may obtain

$$r = \frac{1-\Gamma_r^2-\Gamma_i^2}{(1-\Gamma_r)^2+\Gamma_i^2}, \quad x = \frac{2\Gamma_i}{(1-\Gamma_r)^2+\Gamma_i^2}.$$

These equations can be rearranged as

$$\left(\Gamma_r - \frac{r}{1+r}\right)^2 + \Gamma_i^2 = \left(\frac{1}{1+r}\right)^2, \tag{4.22}$$

$$(\Gamma_r - 1)^2 + \left(\Gamma_i - \frac{1}{x}\right)^2 = \left(\frac{1}{x}\right)^2. \tag{4.23}$$

Equations (4.22) and (4.23) represent two families of circles in the reflection-coefficient plane for different values of r and x, respectively called **resistant circles** and **reactance circles**, which are illustrated in Figure 4.5.

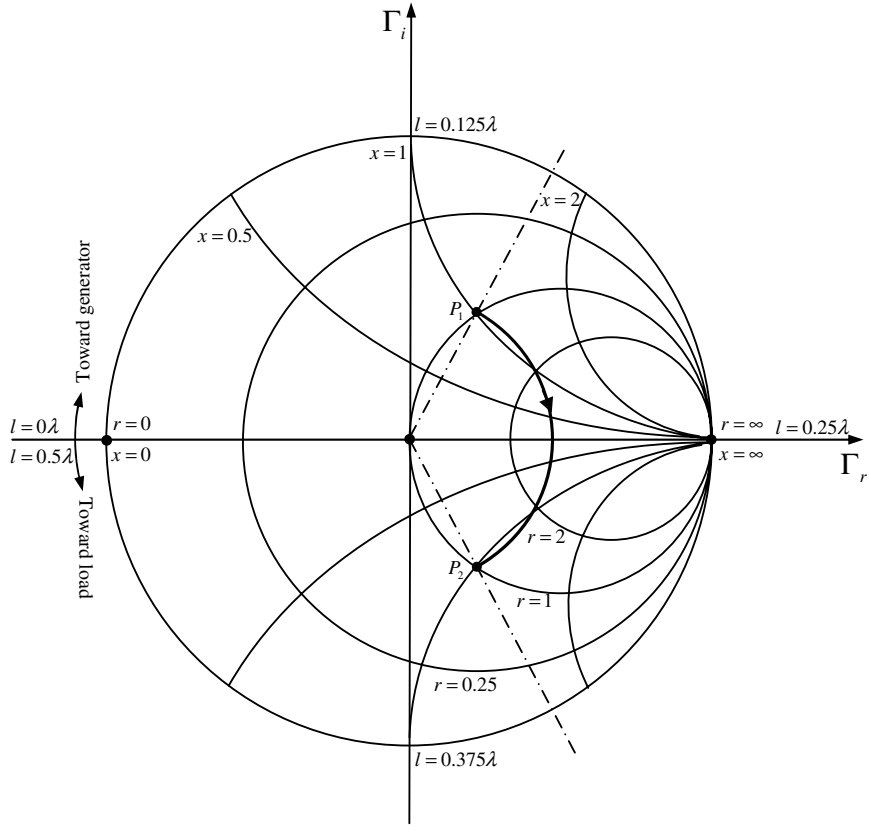


Figure 4.5 Smith chart.

For convenience in using the chart, Smith chart has scales around its periphery giving the angular rotation $2\beta l = 4\pi l/\lambda$ in terms of wavelength. Moving away from the load (toward the generator) corresponds to going around the chart in a clockwise direction. Some important properties of Smith chart are summarized below:

(1) The matching point is the center of impedance chart where $\Gamma = 0$. The open (short) circuit point corresponds to $(1, 0)$ [or $(-1, 0)$] in the Smith chart.
(2) The impedance points in the upper (or lower) half circle are inductive (or capacitive).

(3) The circle $r = 1$ passes through the matching point and divides the chart into two regions $0 \leq r < 1$ and $r > 1$ and the latter lies inside the circle $r = 1$.
(4) The circle $x = 1$ (or $x = -1$) divides the upper (or lower) half circle into $0 < x < 1$ (or $-1 < x < 0$) and $x > 1$ (or $x < -1$).
(5) For any impedance point $z_1 = r_1 + jx_1$, the admittance $g_1 + jb_1 = 1/(r_1 + jx_1)$ may be found from the value of impedance at a point $z_2 = r_2 + jx_2$, which is diametrically across from the impedance point z_1, provided that resistance r_2 and reactance x_2 are respectively interpreted as conductance g_1 and susceptance b_1.

The last property indicates that the same Smith chart can be used for both impedance and admittance calculations.

Example 4.3: Let a transmission line be terminated in a load impedance $z_1 = 1 + j$, labeled P_1 in Figure 4.5. The input impedance at a distance $l = 0.176\lambda$ away from the load can be found by rotating an angle $2\beta l = 4\pi \times 0.176$ in a clockwise direction from P_1, along a constant-radius circle through P_1. The new value of impedance is $z_2 = 1 - j$, labeled P_2 in Figure 4.5. □

4.2 Network Parameters

When the voltages and currents are defined at the reference planes of a microwave circuit, relations exist between the voltages and currents. For a linear microwave circuit, these relations are characterized by impedance or admittance matrices. In microwave engineering, the concept of power is more fundamental than the concepts of voltage and current since the latter are not easily measurable at microwave frequencies. For this reason, the scattering parameters originated in the theory of transmission lines are often introduced and are defined in such a way that the power relationship in the circuit can be expressed in a simple and straightforward manner. Scattering parameters exist for all linear passive time-invariant systems.

4.2.1 One-Port Network

Let us consider a one-port network with input impedance Z as shown in Figure 4.6. The one-port network is connected to a voltage source V_s with source impedance Z_s. The **incident voltage** and the **incident current** are defined as the terminal voltage and current when the one-port network

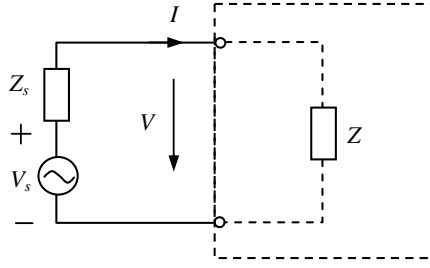


Figure 4.6 One-port network.

is conjugately matched to the source (i.e., $Z = \bar{Z}_s$)

$$V^+ = \frac{V_s \bar{Z}_s}{Z_s + \bar{Z}_s} = \frac{V_s \bar{Z}_s}{2\mathrm{Re}\, Z_s}, \quad I^+ = \frac{V_s}{Z_s + \bar{Z}_s} = \frac{V_s}{2\mathrm{Re}\, Z_s}.$$

So we have $V^+ = \bar{Z}_s I^+$. In this case, the load Z receives the maximum available power from the source, denoted P_A

$$P_A = \frac{1}{2}\mathrm{Re}(V\bar{I}) = \frac{|V_s|^2}{8\mathrm{Re}\, Z_s} = \frac{|V^+|^2 \mathrm{Re}\, Z_s}{2|\bar{Z}_s|^2}.$$

The incident voltage and current are determined by the source only. The source impedance Z_s is called the **reference impedance** of the network. In general, the input impedance Z may not be conjugately matched to the source. The **reflected voltage** and the **reflected current** are then defined by

$$V^- = V - V^+, \quad -I^- = I - I^+.$$

The minus sign in front of I^- implies that the reference direction of I^- is opposite the reference direction of I^+. The **normalized incident voltage wave** a and the **normalized reflected voltage wave** b are defined by

$$a = \frac{V^+\sqrt{\mathrm{Re}\, Z_s}}{\bar{Z}_s}, \quad b = \frac{V^-\sqrt{\mathrm{Re}\, Z_s}}{Z_s},$$

which can also be expressed as

$$a = I^+\sqrt{\mathrm{Re}\, Z_s}, \quad b = I^-\sqrt{\mathrm{Re}\, Z_s}.$$

Note that

$$P_A = \frac{1}{2}|a|^2. \tag{4.24}$$

The terminal voltage and current are thus given by

$$V = V^+ + V^- = \frac{1}{\sqrt{\text{Re}\, Z_s}}(\bar{Z}_s a + Z_s b),$$
$$I = I^+ - I^- = \frac{1}{\sqrt{\text{Re}\, Z_s}}(a - b),$$

from which we obtain

$$a = \frac{V + Z_s I}{2\sqrt{\text{Re}\, Z_s}}, \quad b = \frac{V - \bar{Z}_s I}{2\sqrt{\text{Re}\, Z_s}}.$$

The voltage reflection coefficient and current reflection coefficient are

$$\Gamma_V = \frac{V^-}{V^+} = \frac{Z_s(Z - \bar{Z}_s)}{\bar{Z}_s(Z + \bar{Z}_s)}, \quad \Gamma_I = \frac{I^-}{I^+} = \frac{Z - \bar{Z}_s}{Z + \bar{Z}_s}.$$

In general, Γ_V is not equal to Γ_I. In microwave engineering, the reference impedance Z_s is usually assumed to be real, and so we have $\Gamma_V = \Gamma_I$. The ratio of the normalized reflection wave and the normalized incident wave is the reflection coefficient

$$\Gamma = \frac{b}{a} = \frac{Z - \bar{Z}_s}{Z + \bar{Z}_s} = \Gamma_I.$$

Remark 4.2: In microwave engineering, the load is often connected to the source through a waveguide (Figure 4.7). In this case, we use modal voltage and modal current at the input terminal, which satisfy the transmission line equation as discussed in Chapter 2

$$\frac{dV}{dz} = -j\beta I(z), \quad \frac{dI}{dz} = -j\frac{\beta}{Z_w}V(z), \tag{4.25}$$

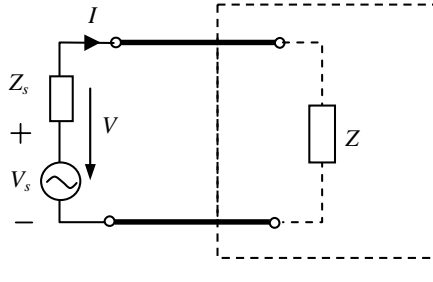


Figure 4.7 One-port microwave network.

where β is the propagation constant and Z_w is the wave impedance. The solutions of (4.25) can be expressed as

$$\begin{aligned} V(z) &= V^+ + V^- = V^+(1+\Gamma), \\ I(z) &= I^+ - I^- = I^+(1-\Gamma), \end{aligned} \tag{4.26}$$

where the superscript "+" denotes the incident wave and the subscript "−" denotes the reflected wave

$$V^+ = Ae^{-j\beta z}, \quad I^+ = \frac{A}{Z_w}e^{-j\beta z},$$

$$V^- = Be^{j\beta z}, \quad I^- = \frac{B}{Z_w}e^{j\beta z},$$

and Γ is the reflection coefficient defined by

$$\Gamma = \frac{V^-}{V^+}.$$

The incident and reflected power are respectively given by

$$P^+ = \frac{1}{2}\mathrm{Re}(V^+\bar{I}^+), \;\; P^- = \frac{1}{2}\mathrm{Re}(V^-\bar{I}^-).$$

The input power can be written as

$$P_{\mathrm{in}} = \frac{1}{2}\mathrm{Re}(V\bar{I}) = P^+(1-|\Gamma|^2).$$

The input impedance is

$$Z = \frac{V}{I} = \frac{V^+ + V^-}{I^+ - I^-} = \frac{V^+}{I^+}\frac{1+V^-/V^+}{1-I^-/I^+} = Z_w\frac{1+\Gamma}{1-\Gamma}. \tag{4.27}$$

We can introduce the normalized voltage and current

$$\begin{aligned} v^+ &= V^+/\sqrt{Z_0}, \quad v^- = V^-/\sqrt{Z_0}, \\ i^+ &= I^+\sqrt{Z_0}, \quad i^- = I^-\sqrt{Z_0}, \end{aligned}$$

where Z_0 is a reference impedance (real), and can be chosen arbitrarily. Thus the normalized input impedance is

$$z = \frac{v}{i} = \frac{v^+ + v^-}{i^+ - i^-} = \frac{v^+}{i^+}\frac{1+v^-/v^+}{1-i^-/i^+} = \frac{Z_w}{Z_0}\frac{1+\Gamma}{1-\Gamma}. \tag{4.28}$$

The normalized incident voltage wave and the normalized reflected voltage wave are then defined by

$$a = v^+, \quad b = v^-. \tag{4.29}$$

We have

$$a = \frac{1}{2}(v + i) = \frac{1}{2\sqrt{Z_0}}(V + Z_0 I),$$
$$b = \frac{1}{2}(v - i) = \frac{1}{2\sqrt{Z_0}}(V - Z_0 I).$$

The reference impedance Z_0 may be chosen as the wave impedance Z_w of the waveguide. □

Let us now consider two one-port networks, one with input impedance Z' and the other with input impedance Z''. The two networks are said to be **dual** if the product of the two input impedances are a real constant, denoted C^2, and are independent of frequency

$$Z'Z'' = C^2.$$

The series circuit and the parallel circuit shown in Figure 4.8 are dual if $Z = Y$. This condition can be met if

$$Z_i = Y_i \quad (i = 1, 2, \ldots, n).$$

Two different ladder circuits are shown in Figure 4.9. The input impedance of the circuit in Figure 4.9(a) may be written as a continued fraction

$$Z = Z_1 + \cfrac{1}{Y_2 + \cfrac{1}{Z_3 + \cfrac{1}{Y_4 + \ddots \\ \qquad\qquad\qquad\qquad Z_{n-1} + \cfrac{1}{Y_n}}}}$$

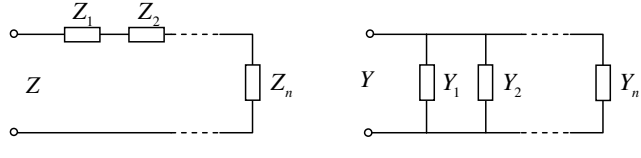


Figure 4.8 Duality of series and parallel circuit.

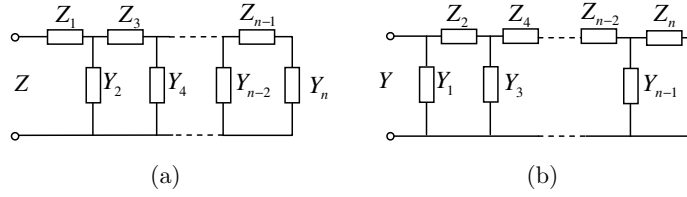


Figure 4.9 Duality of ladder circuits.

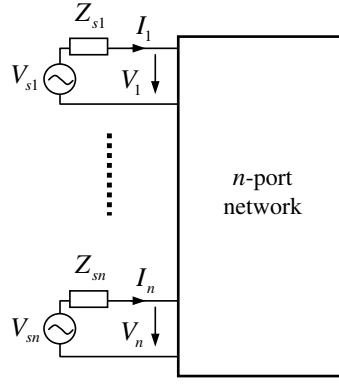


Figure 4.10 n-port network.

The input admittance of the ladder circuit in Figure 4.9(b) is

$$Y = Y_1 + \cfrac{1}{Z_2 + \cfrac{1}{Y_3 + \cfrac{1}{Z_4 + \ddots \quad \cdot \\ \qquad\qquad\qquad\qquad Y_{n-1} + \cfrac{1}{Z_n}}}}$$

Therefore, the two ladder circuits are dual ($Z = Y$) if

$$Z_{2i} = Y_{2i}, \quad Z_{2i-1} = Y_{2i-1} \quad (i = 1, 2, \ldots).$$

4.2.2 Multi-Port Network

For an n-port linear network with port number $i = 1, 2, \ldots, n$ shown in Figure 4.10, the terminal voltages and currents are related by

$$[V] = [Z][I], \tag{4.30}$$

where

$$[V] = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{bmatrix}, \quad [I] = \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_n \end{bmatrix}, \quad [Z] = \begin{bmatrix} Z_{11} & Z_{12} & \cdots & Z_{1n} \\ Z_{21} & Z_{22} & \cdots & Z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n1} & Z_{n2} & \cdots & Z_{nn} \end{bmatrix},$$

and $Z_{ij} (i, j = 1, 2, \ldots, n)$ are called **impedance parameters**. It follows from (4.30) that

$$Z_{ii} = \left.\frac{V_i}{I_i}\right|_{I_l=0, l\neq i}, \quad Z_{ij} = \left.\frac{V_i}{I_j}\right|_{I_l=0, l\neq j}.$$

Hence the impedance parameters are also called **open circuit parameters**.

Similarly, we can write

$$[I] = [Y][V], \tag{4.31}$$

where

$$[Y] = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1n} \\ Y_{21} & Y_{22} & \cdots & Y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nn} \end{bmatrix},$$

and Y_{ij} $(i, j = 1, 2, \ldots, n)$ are called **admittance parameters**. It follows from (4.30) that

$$Y_{ii} = \left.\frac{I_i}{V_i}\right|_{V_l=0, l\neq i}, \quad Y_{ij} = \left.\frac{I_i}{V_j}\right|_{V_l=0, l\neq j}.$$

Hence the admittance parameters are also called **short circuit parameters**.

For the analysis of a microwave network consisting of a cascade connection of two-port networks, it will be convenient to introduce the

transmission or ABCD matrix, defined by

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix} \tag{4.32}$$

where

$$A = \left.\frac{V_1}{V_2}\right|_{I_2=0}, \quad B = \left.\frac{V_1}{I_2}\right|_{V_2=0}, \quad C = \left.\frac{I_1}{V_2}\right|_{I_2=0}, \quad D = \left.\frac{I_1}{I_2}\right|_{V_2=0}.$$

We may introduce the normalized incident wave and reflected wave at each port:

$$a_i = \frac{V_i + Z_{si} I_i}{2\sqrt{\mathrm{Re}\, Z_{si}}}, \quad b_i = \frac{V_i - \bar{Z}_{si} I_i}{2\sqrt{\mathrm{Re}\, Z_{si}}}. \tag{4.33}$$

For a linear network, the normalized reflected wave must be linearly related to the normalized incident wave:

$$[b] = [S][a], \tag{4.34}$$

where

$$[b] = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}, \quad [a] = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}, \quad [S] = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1n} \\ S_{21} & S_{22} & \cdots & S_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n1} & S_{n2} & \cdots & S_{nn} \end{bmatrix}$$

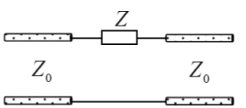
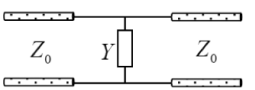
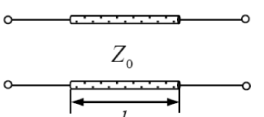
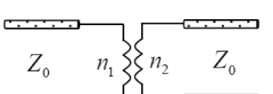
and S_{ij} $(i, j = 1, 2, \ldots, n)$ are called **scattering parameters**. From (4.34), we obtain

$$S_{ii} = \left.\frac{b_i}{a_i}\right|_{a_l=0, l\neq i}, \quad S_{ij} = \left.\frac{b_i}{a_j}\right|_{a_l=0, l\neq j}.$$

The network parameters of some simple two-port networks are listed in Table 4.2.

Remark 4.3: For an n-port microwave network, the reference plane of each port is assumed to be in the single-mode region of the transmission

Table 4.2 Network parameters of simple circuits

	Circuits	Notations	S parameters	ABCD parameters
Series impedance		$z = \dfrac{Z}{Z_0}$	$\begin{bmatrix} \dfrac{z}{2+z} & \dfrac{2}{2+z} \\ \dfrac{2}{2+z} & \dfrac{z}{2+z} \end{bmatrix}$	$\begin{bmatrix} 1 & Z \\ 0 & 1 \end{bmatrix}$
Shunt impedance		$y = \dfrac{Y}{Y_0}$	$\begin{bmatrix} \dfrac{-y}{2+y} & \dfrac{2}{2+y} \\ \dfrac{2}{2+y} & \dfrac{-y}{2+y} \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ Y & 1 \end{bmatrix}$
Transmission line		$\theta = \dfrac{2\pi l}{\lambda_g}$	$\begin{bmatrix} 0 & e^{-j\theta} \\ e^{-j\theta} & 0 \end{bmatrix}$	$\begin{bmatrix} \cos\theta & jZ_0 \sin\theta \\ j\dfrac{\sin\theta}{Z_0} & \cos\theta \end{bmatrix}$
Ideal transformer		$n = \dfrac{n_1}{n_2}$	$\begin{bmatrix} -\dfrac{1-n^2}{1+n^2} & \dfrac{2n}{1+n^2} \\ \dfrac{2n}{1+n^2} & \dfrac{1-n^2}{1+n^2} \end{bmatrix}$	$\begin{bmatrix} n & 0 \\ 0 & \dfrac{1}{n} \end{bmatrix}$

line with real characteristic impedance Z_{0i} $(i = 1, 2, \ldots, n)$. In this case, the normalized incident wave and reflected wave at each port are defined by

$$a_i = \frac{V_i + Z_{0i} I_i}{2\sqrt{Z_{0i}}}, \quad b_i = \frac{V_i - Z_{0i} I_i}{2\sqrt{Z_{0i}}}. \tag{4.35}$$

Thus

$$V_i = \sqrt{Z_{0i}}(a_i + b_i), \quad I_i = \frac{1}{\sqrt{Z_{0i}}}(a_i - b_i). \tag{4.36}$$

The normalized incident wave and reflected wave defined by (4.33) are a generalized version of (4.35) for an arbitrary circuit. □

Example 4.4 (Lossless condition)**:** Consider an n-port network, the power delivered to the network is

$$P = \frac{1}{2}\mathrm{Re}[V]^T[\bar{I}] = \frac{1}{2}\big([a]^T[\bar{a}] - [b]^T[\bar{b}]\big) = \frac{1}{2}[a]^T\big([1] - [S]^T[\bar{S}]\big)[\bar{a}].$$

If the network is lossless, then $P = 0$. This gives the **lossless condition**

$$[1] - [S]^T[\bar{S}] = 0, \tag{4.37}$$

where [1] denotes the identity matrix. □

4.2.3 Foster Reactance Theorem

Let us consider a lossless one-port network consisting of a feeding waveguide and a finite system, outside of which the electromagnetic fields are assumed to be zero. Let the system be enclosed by a surface S which cuts the feeding line perpendicular to its axis. The intersection is denoted by Ω, as illustrated in Figure 4.11. We may introduce the complex frequency $s = \alpha + j\omega$ so that

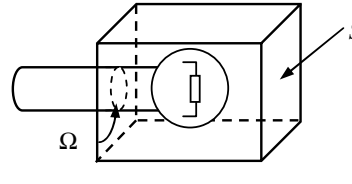


Figure 4.11 One-port network.

the Poynting theorem in complex frequency domain can be written as

$$\nabla \cdot \left[\frac{1}{2}\mathbf{E}(\mathbf{r},s) \times \bar{\mathbf{H}}(\mathbf{r},s)\right] = -\frac{1}{2}\alpha\big[\mu|\mathbf{H}(\mathbf{r},s)|^2 + \varepsilon|\mathbf{E}(\mathbf{r},s)|^2\big] - j\frac{1}{2}\omega\big[\mu|\mathbf{H}(\mathbf{r},s)|^2 - \varepsilon|\mathbf{E}(\mathbf{r},s)|^2\big]. \tag{4.38}$$

Taking the integration of (4.38) over the connected region V bounded by S, we obtain

$$\int_S \frac{1}{2}\big[\mathbf{E}(\mathbf{r},s) \times \bar{\mathbf{H}}(\mathbf{r},s)\big] \cdot \mathbf{u}_n \, dS = -2\alpha\big[W_m(s) + W_e(s)\big] - 2j\omega\big[W_m(s) - W_e(s)\big], \tag{4.39}$$

where

$$W_m(s) = \frac{1}{4}\int_V \mu|\mathbf{H}(\mathbf{r},s)|^2 dV(\mathbf{r}), \quad W_e(s) = \frac{1}{4}\int_V \varepsilon|\mathbf{E}(\mathbf{r},s)|^2 dV(\mathbf{r}).$$

If S is big enough, the fields vanish on S except on the terminal cross-section Ω. Thus for a single-mode feeding waveguide, we have

$$\frac{1}{2}V(s)\bar{I}(s) = 2\alpha[W_m(s) + W_e(s)] + 2j\omega[W_m(s) - W_e(s)]. \tag{4.40}$$

The impedance in the complex frequency plane can then be expressed as

$$Z(s) = \frac{4\alpha}{|I(s)|^2}[W_m(s) + W_e(s)] + \frac{4j\omega}{|I(s)|^2}[W_m(s) - W_e(s)]. \tag{4.41}$$

This can be rewritten as

$$Z(s) = \frac{4sW_m(s)}{|I(s)|^2} + \frac{4\bar{s}W_e(s)}{|I(s)|^2}. \tag{4.42}$$

We may introduce a new quantity $W'_e(s) = |s|^2 W_e(s)$ and replace all the complex conjugate $\bar{s}$ with $-s$ so that (4.42) becomes analytic and can be written as

$$Z(s) = \frac{4sW_m(s)}{I(s)I(-s)} + \frac{4s^{-1}W'_e(s)}{I(s)I(-s)}. \tag{4.43}$$

If α is assumed to be small, a Taylor series expansion may be introduced. For an arbitrary function $\mathbf{A}(s)$, we have

$$\mathbf{A}(s) \cdot \mathbf{A}(-s) = |\mathbf{A}(j\omega)|^2 + j\alpha T(\omega) + o(\alpha),$$

where $T(\omega)$ is a real function of ω. Using the following decomposition

$$Z(s) = R(\alpha, \omega) + jX(\alpha, \omega), \tag{4.44}$$

we may find that

$$R(\alpha, \omega) = \frac{4\alpha}{|I|^2}(W_m + W_e), \quad X(\alpha, \omega) = \frac{4\omega(W_m - W_e)}{|I|^2}, \tag{4.45}$$

where the energies and current are all calculated at $\alpha = 0$. From Cauchy–Riemann conditions, we obtain

$$\left.\frac{\partial X}{\partial \omega}\right|_{\alpha=0} = \frac{4}{|I|^2}(W_m + W_e) > 0. \tag{4.46}$$

This is the Foster reactance theorem for a lossless one-port system. The following conclusions may be drawn from (4.46):

(1) $X(\omega)$ is a monotonically increasing with frequency.
(2) $X(\omega)$ is an odd function of frequency.
(3) The poles and zeroes of $X(\omega)$ must alternate with increasing frequency. After passing through a pole, the function $X(\omega)$ will be negative and then pass through zero before reaching the next pole.
(4) The poles and zeros of $X(\omega)$ are symmetrical about the origin.

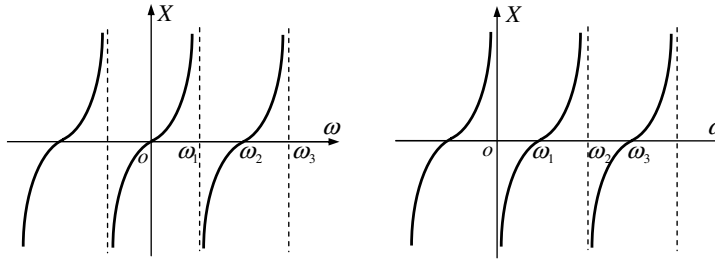
These properties are illustrated in Figure 4.12. There are four possible cases for the reactance function

(1) $X(0) = -\infty,\ X(\infty) = +\infty$.
(2) $X(0) = -\infty,\ X(\infty) = 0$.
(3) $X(0) = 0,\ X(\infty) = +\infty$.
(4) $X(0) = 0,\ X(\infty) = 0$.

As a result, the reactance function may have four different forms:

$$X(\omega) = H\frac{(\omega^2 - \omega_1^2)(\omega^2 - \omega_3^2)\dots(\omega^2 - \omega_n^2)}{\omega(\omega^2 - \omega_2^2)(\omega^2 - \omega_4^2)\dots(\omega^2 - \omega_{n-1}^2)},$$

$$X(\omega) = H\frac{(\omega^2 - \omega_1^2)(\omega^2 - \omega_3^2)\dots(\omega^2 - \omega_{n-1}^2)}{\omega(\omega^2 - \omega_2^2)(\omega^2 - \omega_4^2)\dots(\omega^2 - \omega_n^2)},$$



(a) (b)

Figure 4.12 Reactance curves.

$$X(\omega) = H\frac{\omega(\omega^2-\omega_2^2)(\omega^2-\omega_4^2)\dots(\omega^2-\omega_n^2)}{(\omega^2-\omega_1^2)(\omega^2-\omega_3^2)\dots(\omega^2-\omega_{n-1}^2)},$$

$$X(\omega) = H\frac{\omega(\omega^2-\omega_2^2)(\omega^2-\omega_4^2)\dots(\omega^2-\omega_{n-1}^2)}{(\omega^2-\omega_1^2)(\omega^2-\omega_3^2)\dots(\omega^2-\omega_n^2)}.$$

The input impedance of the lossless network is $Z(j\omega) = jX(\omega)$. Introducing $s = j\omega$, we have

$$Z(s) = H\frac{(s^2-s_1^2)(s^2-s_3^2)\dots(s^2-s_n^2)}{s(s^2-s_2^2)(s^2-s_4^2)\dots(s^2-s_{n-1}^2)},$$

$$Z(s) = H\frac{(s^2-s_1^2)(s^2-s_3^2)\dots(s^2-s_{n-1}^2)}{s(s^2-s_2^2)(s^2-s_4^2)\dots(s^2-s_n^2)},$$

$$Z(s) = H\frac{s(s^2-s_2^2)(s^2-s_4^2)\dots(s^2-s_n^2)}{(s^2-s_1^2)(s^2-s_3^2)\dots(s^2-s_{n-1}^2)},$$

$$Z(s) = H\frac{s(s^2-s_2^2)(s^2-s_4^2)\dots(s^2-s_{n-1}^2)}{(s^2-s_1^2)(s^2-s_3^2)\dots(s^2-s_n^2)}.$$

If we let $s = \sigma + j\omega$, the above expressions are then extended to the complex frequency domain and can be written as

$$Z(s) = \frac{a_{n+1}s^{n+1} + a_{n-1}s^{n-1} + \dots + a_0}{a_n s^n + a_{n-2}s^{n-2} + \dots + a_1 s},$$

$$Z(s) = \frac{a_n s^n + a_{n-2}s^{n-2} + \dots + a_0}{a_{n+1}s^{n+1} + a_{n-1}s^{n-1} + \dots + a_1 s},$$

$$Z(s) = \frac{a_{n+1}s^{n+1} + a_{n-1}s^{n-1} + \cdots + a_1 s}{a_n s^n + a_{n-2}s^{n-2} + \cdots + a_0},$$

$$Z(s) = \frac{a_{n+1}s^n + a_{n-2}s^{n-2} + \cdots + a_1 s}{a_{n+1}s^{n+1} + a_{n-1}s^{n-1} + \cdots + a_0}.$$

Remark 4.4: It can be shown that the driving point immittance (impedance or admittance) of a linear lumped one-port network is a rational function in complex frequency domain, which is defined as the ratio of two polynomials. The driving point immittance of a passive one-port network is a positive real function. □

Remark 4.5: A **positive real function** $Z(s)$ is defined as an analytical function that satisfies the following conditions:

(1) $Z(s)$ is analytic in the open right-half of the s-plane.
(2) $Z(\bar{s}) = \bar{Z}(s)$ for all s in the open right-half of the s-plane.
(3) $\mathrm{Re}\, Z(s) \geq 0$ whenever $\mathrm{Re}\, s \geq 0$. □

Remark 4.6: A rational function $Z(s)$ is positive real if and only if it satisfies the following conditions:

(1) $Z(s)$ is real whenever s is real.
(2) $Z(s)$ has no poles in the open right-half of s-plane.
(3) If $Z(s)$ has poles on $j\omega$-axis, they are simple and the residues at these poles are real and positive.
(4) $\mathrm{Re}\, Z(s) \geq 0$ for all ω, except at the poles. □

4.3 Impedance Matching Circuits

Impedance matching circuits are often used to maximize the power transfer between a signal source and a load or to minimize reflections from the load. The concept of impedance matching has been widely used in various situations where optimum power transmission is needed between a source and a load.

4.3.1 Basic Concept of Match

Power transfer is maximized when source is conjugately matched to the load. In case of resistive terminations, the source resistance must be equal to the load resistance for maximum power transfer. In practice, terminations

generally represent complex impedances, and their real parts may not be equal. In this case, an impedance matching circuit is required to eliminate the mismatch. When the real parts are equal, the matching task is to resonate the unwanted reactance or susceptance at operating frequency. Perfect match (zero reflection coefficient) can only be achieved at selected single frequencies. Matching a real (resistive) source to a complex load represents two problems:

(1) The imaginary part of the load must be tuned out.
(2) The real parts must be adjusted to have equal values.

4.3.1.1 *Impedance Matching for Pure Resistances*

If both terminations are resistive but unequal, an impedance transformer is needed to assure maximum power transfer (zero reflection). This can be achieved by an L-network design. A properly chosen two-element LC section can always match two unequal resistive terminations. The series element is always placed on the low impedance side, and a parallel element is used next to the high-impedance termination, as shown in Figure 4.13. Let us consider the case depicted in Figure 4.13(a). Under the condition of conjugate match, we have

$$R_s - jX_s = \frac{jR_lX_p}{R_l + jX_p},$$

which implies

$$\sqrt{R_s^2 + X_s^2} = \frac{R_lX_p}{\sqrt{R_l^2 + X_p^2}}. \tag{4.47}$$

We introduce the quality factors

$$Q_s = \frac{X_s}{R_s}, \quad Q_p = \frac{R_l}{X_p},$$

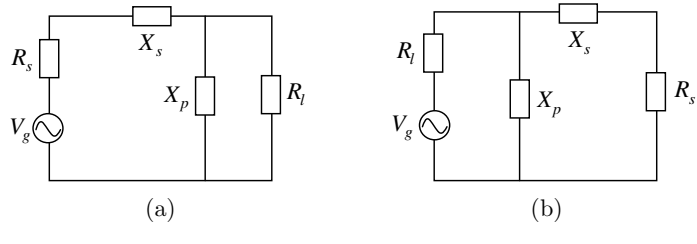


Figure 4.13 Matching for pure resistances ($R_s < R_l$).

and both quality factors should be equal under conjugate match. As a result, (4.47) may be written as

$$1 + Q^2 = \frac{R_l}{R_s}, \tag{4.48}$$

where $Q = Q_s = Q_p$. Similar discussions hold for the case in Figure 4.13(b). The design of the L-network can be summarized as follows:

Step 1: Add a shunt reactance (capacitor or inductor) to the larger termination, such that

$$X_p = \frac{R_l}{Q}, \quad Q = \sqrt{\frac{R_l}{R_s} - 1}. \tag{4.49}$$

Step 2: Add a series reactance (opposite kind of what is selected in Step 1), to the smaller termination, such that

$$X_s = R_s Q. \tag{4.50}$$

Step 3: Compute the matching element values:

$$C = \frac{1}{2\pi f X_{p \text{ or } s}}, \quad L = \frac{X_{p \text{ or } s}}{2\pi f}. \tag{4.51}$$

Example 4.5: Design an L-section that matches a 500 Ω resistive load to a 50 Ω transmission line at 500 MHz.

Solution: From (4.48)–(4.50), we obtain

$$Q = \sqrt{\frac{R_l}{R_s} - 1} = \sqrt{\frac{500}{50} - 1} = 3,$$

$$X_p = \frac{R_l}{Q} = \frac{500}{3} = 166.7\,\Omega,$$

$$X_s = R_s Q = 50 \times 3 = 150\,\Omega.$$

The reactance X_p is connected in parallel with the load while X_s is in series with the transmission line. If X_p is assumed to be inductive, then X_s must be capacitive. In this case, the component values can be calculated through (4.51) as follows

$$L_p = \frac{X_p}{2\pi f} = \frac{166.7}{2\pi \times 500 \times 10^6} \approx 53\,\text{nH},$$

$$C_s = \frac{1}{2\pi f X_s} = \frac{1}{2\pi \times 500 \times 10^6 \times 150} = 2\,\text{pF}.$$

The component values can be determined similarly if X_p is assumed to be capacitive and X_s inductive. □

4.3.1.2 *Impedance Matching for Complex Loads*

For two complex impedances, the matching network between them may be first designed for two pure resistances (the real parts of the two complex impedances). The imaginary parts can be taken into account by the following two basic steps:

Step 1 (Absorption): Reactances of the complex impedances are absorbed into the impedance-matching network, up to the maximums, that are equal to the matching component values.

Step 2 (Resonance): Beyond the limits of maximum absorption, the reactances may be resonated with an equal and opposite reactance at the frequency of interest.

Example 4.6: Design an L-section that matches a $10 + j15\,\Omega$ load to a $50\,\Omega$ transmission line at 500 MHz.

Solution: We first consider the matching between the real part of the load and the transmission line. From (4.48)–(4.50), we obtain

$$Q = \sqrt{\frac{R_l}{R_s} - 1} = \sqrt{\frac{50}{10} - 1} = 2,$$

$$X_p = \frac{R_l}{Q} = \frac{50}{2} = 25\,\Omega,$$

$$X_s = R_s Q = 10 \times 2 = 20\,\Omega.$$

Therefore, the reactance X_p is connected in parallel with the transmission line and X_s is in series with the load. If X_p is assumed to be capacitive $25\,\Omega$, X_s must be inductive $20\,\Omega$. The inductive reactance $15\,\Omega$ of the load can be absorbed into X_s, and another series inductor for the remaining $X_s' = 5\,\Omega$ is needed and its value is determined by

$$L_s = \frac{X_s'}{2\pi f} = \frac{5}{2\pi \times 500 \times 10^6} \approx 1.6\,\text{nH}.$$

The parallel capacitance is given by

$$C_p = \frac{1}{2\pi f X_p} = \frac{1}{2\pi \times 500 \times 10^6 \times 25} = 12.7\,\text{pF}.$$

If X_p is assumed to be inductive $25\,\Omega$, X_s must be capacitive $20\,\Omega$. Since the load has a $15\,\Omega$ inductive reactance, an additional $15\,\Omega$ capacitive reactance must be introduced. Therefore, the component values are determined by

$$C_s = \frac{1}{2\pi f X_s} = \frac{1}{2\pi \times 500 \times 10^6 \times (20+15)} = 9.1\,\text{pF},$$

$$L_p = \frac{X_p}{2\pi f} = \frac{25}{2\pi \times 500 \times 10^6} \approx 8\,\text{nH}. \quad \square$$

4.3.2 Quarter-Wave Impedance Transformer

A **quarter-wave impedance transformer** (often written as **$\lambda/4$ impedance transformer**) consists of a length of transmission line or waveguide, which is one-quarter of a wavelength long and terminated in a real load impedance, and can be used to match a real load to a transmission line. Consider the circuit depicted in Figure 4.14, where a transmission line of length l at design frequency with characteristic impedance Z_1 (the transformer) is inserted between a real load Z_L and a transmission line of characteristic impedance Z_0. The input impedance looking into the impedance transformer is

$$Z = Z_1 \frac{Z_L + jZ_1 \tan \beta l}{Z_1 + jZ_L \tan \beta l}. \tag{4.52}$$

When $l = \lambda/4$ (or $\beta l = \pi/2$) and the matching condition $Z = Z_0$ are applied, we have

$$Z_1^2 = Z_0 Z_L, \tag{4.53}$$

which determines the characteristic impedance of the transformer. Using (4.52) and (4.53), the amplitude of the reflection coefficient at the input of

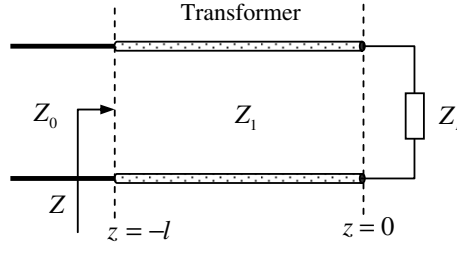


Figure 4.14 Impedance transformer.

the transformer may be written as

$$|\Gamma| = \left|\frac{Z - Z_0}{Z + Z_0}\right| = \frac{|Z_L - Z_0|}{\sqrt{(Z_L + Z_0)^2 + 4Z_1^2 \tan^2 \beta l}}. \tag{4.54}$$

When βl approaches $\pi/2$, this may be approximated by

$$|\Gamma| \approx \frac{|Z_L - Z_0|}{2Z_1} |\cos \beta l|. \tag{4.55}$$

Remark 4.1: It is noted that the reflection coefficient at the input of an impedance transformer of length l may be expressed as

$$\Gamma = \frac{\Gamma_1 + \Gamma_2 e^{-j2\beta l}}{1 + \Gamma_1 \Gamma_2 e^{-j2\beta l}}, \tag{4.56}$$

where

$$\Gamma_1 = \frac{Z_1 - Z_0}{Z_1 + Z_0}, \quad \Gamma_2 = \frac{Z_L - Z_1}{Z_L + Z_1},$$

are called **partial reflection coefficients** at junctions $z = -l$ and $z = 0$. If $|\Gamma_1 \Gamma_2| \ll 1$, (4.56) can be approximated by

$$\Gamma \approx \Gamma_1 + \Gamma_2 e^{-j2\beta l}, \tag{4.57}$$

where the factor $e^{-j2\beta l}$ denotes the phase delay when the incident wave travels up and down the line. □

4.3.3 Tapered Line Transformer

A uniform section of transmission line may be flared out to form a tapered transmission line transformer as illustrated in Figure 4.15. The tapered transmission line may be considered being made up of a number of incremental sections of length Δz. The incremental change of the characteristic impedance from one section to the next is assumed to be ΔZ. The partial incremental reflection coefficient at the step z is then given by

$$d\Gamma = \frac{Z + dZ - Z}{Z + dZ + Z} \approx \frac{dZ}{2Z} = \frac{1}{2} \frac{d}{dz} \left(\ln \frac{Z}{Z_0}\right) dz.$$

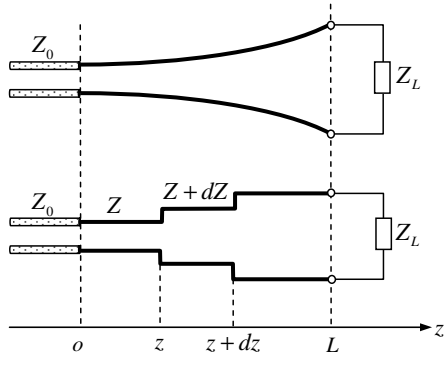


Figure 4.15 Tapered transformer approximated by multiple uniform transmission lines.

Making use of (4.57), the total reflection coefficient at $z = 0$ can be represented by the sum of all the partial reflections with proper phase shifts

$$\Gamma = \frac{1}{2}\int_0^L e^{-j2\beta z}\frac{d}{dz}\left(\ln\frac{Z}{Z_0}\right)dz. \tag{4.58}$$

4.4 Passive Components

A **passive component** refers to a component that consumes but does not produce energy. An electronic circuit consisting entirely of passive components is called a **passive circuit**. In this section, we discuss the basic operating principles of several commonly used passive devices.

4.4.1 Electronically Controlled Phase Shifters

A phase shifter is a microwave network which provides a controllable shift in the phase angle of the RF signal transmitted through it. Ideal phase shifters should be perfectly matched to the input and output lines and should provide low insertion loss in all phase states. Phase shifters can be controlled electrically, magnetically or mechanically. Electronically-controlled phase shifters can be analog or digital. Analog phase shifters provide a continuously variable phase, which can be realized with varactor diodes that change capacitance with voltage or other nonlinear dielectrics. Digital phase shifters provide a discrete set of phase states, which can be realized by PIN diodes that switch circuit elements in and out of the transmission path. Each switching operation adds or subtracts a finite

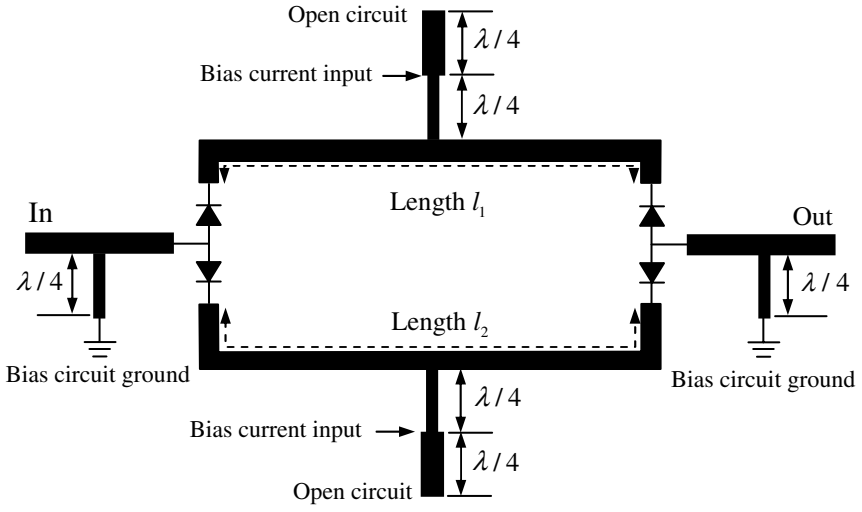


Figure 4.16 A switched-line phase shifter.

phase-shift increment. The digital phase shifters using PIN diodes have the advantages of being small size, high speed and easy to be integrated with planar circuits, and are widely used in phased arrays.

Figure 4.16 shows a simple design for the phase shifter using PIN diodes, which switch one of two alternate transmission lines of different length into the transmission path. The bias currents are applied to the circuit at the midpoint of a half-wave open-circuited stub. The quarter-wave transformer of low impedance transforms the open-circuit impedance to a short circuit or low impedance at the midpoint, which is transformed to high impedance by a quarter-wave transformer of high impedance to reduce the influence on the main transmission line. The short-circuited high-impedance quarter-wave transformers connected to the input and output lines are used as DC return path for the bias current. The differential phase shift between the two transmission-line sections of lengths l_1 and l_2 is

$$\Delta\varphi = \beta(l_2 - l_1),$$

where β is the propagation constant. The phase shift $\Delta\varphi$ depends on the frequency since β is a function of frequency. For a TEM transmission line, the phase velocity $v_p = \omega/\beta$ is independent of frequency. This implies that the differential time delay $\Delta\tau = (l_2 - l_1)/v_p$ is a constant, which is useful for reducing signal distortion in broadband systems. By use of a cascade

connection of several phase shifters, any phase shift between 0° and 180° can be obtained with a resolution equal to the smallest differential phase shift.

4.4.2 Attenuators

An **attenuator** is an electronic device that reduces the power of a signal by a predetermined ratio to prevent overloading or provide isolation without appreciably distorting its waveform. Attenuators are usually passive devices made from simple voltage divider networks. Fixed attenuators are used when attenuation is constant. Variable attenuators are formed by switching between different resistances and the variability can be in steps or continuous, obtained either manually or programmably. In measuring signals, attenuators are used to lower the amplitude of the signal a known amount to enable measurements, or to protect the measuring device from signal levels that might damage it.

The attenuator is a two-port device as described in Figure 4.17. The **attenuation** is defined by

$$A(dB) = 10 \log \frac{P_{\text{in}}}{P_{\text{out}}}. \tag{4.59}$$

Example 4.7 (Fixed power attenuator)**:** The fixed power attenuators may use a Tee- or Pi-network of resistors as indicated in Figure 4.18. Let $\alpha = 10^{-A(dB)/10}$. Once the attenuation A is specified, the component values can be determined as follows.

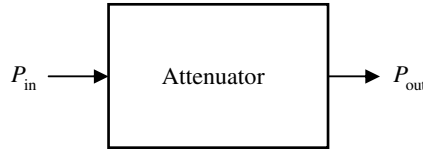


Figure 4.17 Attenuator.

Figure 4.18 Tee- and Pi configurations for attenuators.

Tee-network:

$$R_p = \frac{2\sqrt{\alpha Z_1 Z_2}}{|\alpha - 1|}, \quad R_{s1} = Z_1 \frac{\alpha + 1}{|\alpha - 1|} - R_p, \quad R_{s2} = Z_2 \frac{\alpha + 1}{|\alpha - 1|} - R_p$$

Pi-network:

$$R_s = \frac{|\alpha - 1|\sqrt{Z_1 Z_2}}{2\sqrt{\alpha}}, \quad R_{p1} = \left(\frac{1}{Z_1} \frac{\alpha + 1}{|\alpha - 1|} - \frac{1}{R_s} \right)^{-1},$$

$$R_{p2} = \left(\frac{1}{Z_2} \frac{\alpha + 1}{|\alpha - 1|} - \frac{1}{R_s} \right)^{-1}. \quad \square$$

4.4.3 Power Dividers and Combiners

A power divider is used to divide an input signal into two or more signals of lesser power and may be described as a three-port network as illustrated in Figure 4.19(a). Figure 4.19(b) shows an equal-split resistive power divider.

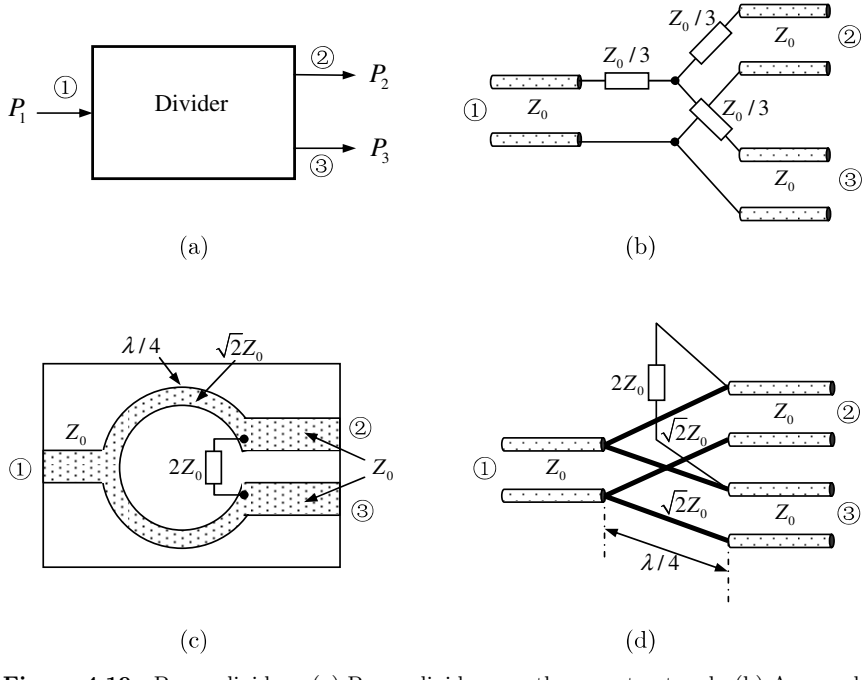


Figure 4.19 Power dividers. (a) Power divider as a three-port network. (b) An equal-split resistive power divider. (c) An equal-split microstrip Wilkinson power divider. (d) Equivalent circuit for Wilkinson power divider.

It is noted that

$$P_2 = P_3 = \frac{1}{4}P_1.$$

Therefore, half of the input power to the resistive divider is dissipated in the resistors.

The Wilkinson power divider, first proposed by Ernest J. Wilkinson (1960), can achieve high degree of isolation between the output ports while maintaining a matched condition on all ports. It uses quarter-wave transformers or lumped circuit elements (inductors and capacitors). Figure 4.19(c) shows an equal-split Wilkinson microstrip power divider. Its equivalent circuit is depicted in Figure 4.19(d). An ideal Wilkinson power divider would yield

$$P_2 = P_3 = \frac{1}{2}P_1.$$

A unique feature of the Wilkinson power divider is the use of resistor connected between the output ports. The resistor does not consume any power if there is no current in the resistor, which can be accomplished if the output ports are properly loaded.

We simply note that a power combiner is the reverse of a divider, and both use exactly the same circuits.

4.4.4 Directional Couplers

Directional couplers are most frequently constructed from two coupled transmission lines set close enough together such that energy passing through one is coupled to the other. They are often used to provide a signal sample for measurement or monitoring, and can be designed by using hollow waveguides, microstrip line or strip line. Directional couplers are four-port networks. Figure 4.20 shows two commonly used generic symbols for them. The waves add in phase at the coupled port and are cancelled at the isolated port. The directional couplers are characterized by **coupling factor**, **directivity**, **isolation** and **insertion loss**, which are defined respectively by

$$\mathrm{C} = 10\log\frac{P_1}{P_3},$$

$$\mathrm{D} = 10\log\frac{P_3}{P_4},$$

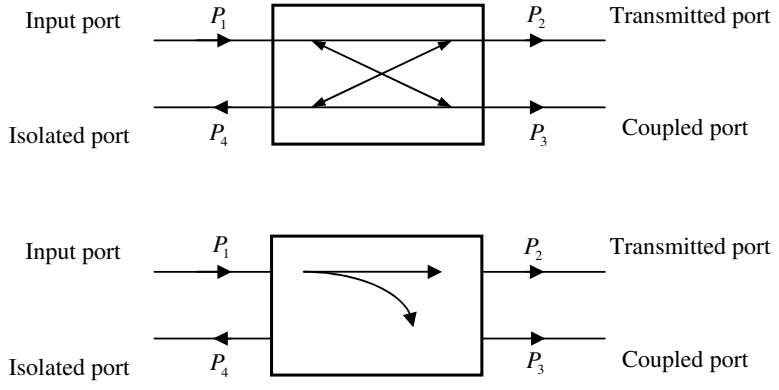


Figure 4.20 Two commonly used generic symbols for directional couplers.

$$\mathrm{I} = 10 \log \frac{P_1}{P_4},$$

$$\mathrm{IL} = 10 \log \frac{P_1}{P_2}.$$

Note that all these quantities are defined as positive in dB. Microwave engineers often present these quantities as negative numbers. The coupling factor is a primary property of directional coupler and is related to other quantities as follows

$$\mathrm{I} = \mathrm{D} + \mathrm{C} \ (\mathrm{dB}).$$

A perfectly matched directional coupler is characterized by scattering matrix in the following form

$$[S] = \begin{bmatrix} 0 & \alpha & \beta & 0 \\ \alpha & 0 & 0 & -\beta \\ \beta & 0 & 0 & \alpha \\ 0 & -\beta & \alpha & 0 \end{bmatrix},$$

where α and β are either real or imaginary and satisfy

$$|\alpha|^2 + |\beta|^2 = 1.$$

4.4.4.1 *Hole Couplers*

A hollow waveguide may be coupled to another through holes. Figure 4.21 shows two waveguides coupled together by a small hole (Bethe hole coupler) centered at $z = 0$, denoted by S_a. The power is assumed to be

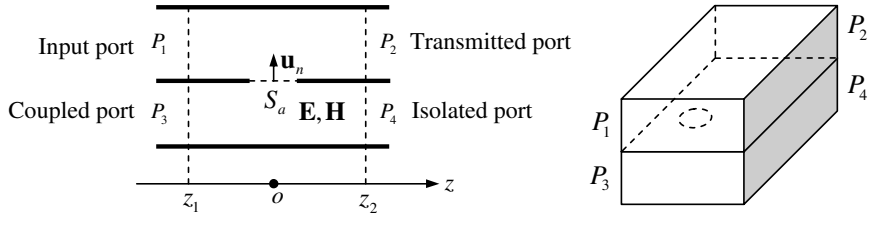


Figure 4.21 A single-hole coupler.

incident at Port 1 of the upper waveguide and is coupled to the lower waveguide through the hole. A general theory for small-hole coupling has been presented in Section 2.4.5. The theory states that the small hole is equivalent to a combination of radiating electric and magnetic dipoles. Choose two cross-sectional planes $z = z_1$ and $z = z_2$ so that the hole lies between z_1 and z_2, and only the dominant modes are present in the regions $z < z_1$ and $z > z_2$. The fields $\mathbf{E}, \mathbf{H}$ in the lower waveguide may be represented in terms of the dominant vector modal function as follows

$$\begin{aligned} &\mathbf{E} = A_1\mathbf{E}_1^+, \quad \mathbf{H} = A_1\mathbf{H}_1^+, \quad z \geq z_2, \\ &\mathbf{E} = B_1\mathbf{E}_1^-, \quad \mathbf{H} = B_1\mathbf{H}_1^-, \quad z \leq z_1, \end{aligned} \tag{4.60}$$

where

$$\begin{aligned} &\mathbf{E}_1^+ = (\mathbf{e}_1 + \mathbf{u}_z e_{z1})e^{-j\beta_1 z}, \quad \mathbf{H}_1^+ = (\mathbf{h}_1 + \mathbf{u}_z h_{z1})e^{-j\beta_1 z}, \\ &\mathbf{E}_1^- = (\mathbf{e}_1 - \mathbf{u}_z e_{z1})e^{j\beta_1 z}, \quad \mathbf{H}_1^- = (-\mathbf{h}_1 + \mathbf{u}_z h_{z1})e^{j\beta_1 z}, \end{aligned} \tag{4.61}$$

with

$$\mathbf{h}_1 = \frac{\mathbf{u}_z \times \mathbf{e}_1}{Z_{w1}}, \quad e_{z1} = \frac{\nabla \cdot \mathbf{e}_1}{j\beta_1}, \quad h_{z1}\mathbf{u}_z = -\frac{\nabla \times \mathbf{e}_1}{j\beta_1 Z_{w1}}.$$

The coefficients in (4.60) may be determined by the electric and magnetic dipoles

$$\begin{aligned} A_1 &= -\frac{Z_{w1}}{2}\left[-\mathbf{E}_1^-(0,0)\cdot j\omega\mathbf{p} + \mathbf{H}_1^-(0,0)\cdot j\omega\mu_0\mathbf{m}\right], \\ B_1 &= -\frac{Z_{w1}}{2}\left[-\mathbf{E}_1^+(0,0)\cdot j\omega\mathbf{p} + \mathbf{H}_1^+(0,0)\cdot j\omega\mu_0\mathbf{m}\right]. \end{aligned} \tag{4.62}$$

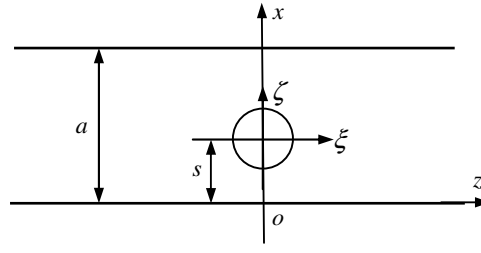


Figure 4.22 A small circular hole in broad wall of rectangular waveguide.

Let us consider two identical rectangular waveguides coupled by a small circular hole of radius a_0 in the common broad wall between the two guides as shown in Figure 4.22. For TE_{10} mode, we have

$$\mathbf{e}_1 = -\mathbf{u}_y \sqrt{\frac{2}{ab}} \sin \frac{\pi}{a} x,$$

$$\mathbf{h}_1 = \mathbf{u}_x \frac{1}{Z_{w1}} \sqrt{\frac{2}{ab}} \sin \frac{\pi}{a} x,$$

$$h_{z1}\mathbf{u}_z = \mathbf{u}_z \frac{1}{j\beta_1 Z_{w1}} \frac{\pi}{a} \sqrt{\frac{2}{ab}} \cos \frac{\pi}{a} x.$$

Assume that the incident dominant TE_{10} mode at Port 1 is of unit amplitude. Then the electric and magnetic dipole moments may be expressed as (see Section 7.3.2)

$$\begin{aligned} \mathbf{p} &= \mathbf{u}_y \frac{4a_0^3}{3} \varepsilon_0 \sqrt{\frac{2}{ab}} \sin \frac{\pi}{a} s, \\ \mathbf{m} &= \frac{8}{3} a_0^3 \left[\mathbf{u}_x \frac{1}{Z_{w1}} \sqrt{\frac{2}{ab}} \sin \frac{\pi}{a} s + \mathbf{u}_z \frac{1}{j\beta_1 Z_{w1}} \frac{\pi}{a} \sqrt{\frac{2}{ab}} \cos \frac{\pi}{a} s \right]. \end{aligned} \tag{4.63}$$

Substituting (4.61) and (4.63) into (4.62), we obtain

$$A_1 = -j\omega Z_{w1} \frac{4a_0^3}{3ab} \varepsilon_0 \sin^2 \frac{\pi}{a} s + j\omega \mu_0 \frac{8a_0^3}{3abZ_{w1}} \left[\sin^2 \frac{\pi}{a} s + \frac{1}{\beta_1^2} \left(\frac{\pi}{a} \right)^2 \cos^2 \frac{\pi}{a} s \right], \tag{4.64}$$

$$B_1 = -j\omega Z_{w1} \frac{4a_0^3}{3ab} \varepsilon_0 \sin^2 \frac{\pi}{a} s - j\omega \mu_0 \frac{8a_0^3}{3abZ_{w1}} \left[\sin^2 \frac{\pi}{a} s - \frac{1}{\beta_1^2} \left(\frac{\pi}{a} \right)^2 \cos^2 \frac{\pi}{a} s \right]. \tag{4.65}$$

Therefore, the wave excited toward Port 3 is generally different from that excited toward Port 4. The hole couplers can be divided as forward-wave and backward-wave couplers. For example, a backward-wave coupler is obtained by setting the power delivered to the isolated Port 4 to zero by letting $A_1 = 0$

$$\left(-\varepsilon_0 + \frac{2\mu_0}{Z_{w1}^2}\right)\sin^2\frac{\pi}{a}s + \frac{\pi^2}{\beta_1^2 a^2}\frac{2\mu_0}{Z_{w1}^2}\cos^2\frac{\pi}{a}s = 0. \tag{4.66}$$

This may reduce to

$$\sin\frac{\pi}{a}s = \pi\sqrt{\frac{2}{(2\pi)^2 - k^2a^2}}. \tag{4.67}$$

The coupling factor and directivity are respectively given by

$$\mathrm{C} = 20\log\left|\frac{1}{B_1}\right|, \quad D = 20\log\left|\frac{B_1}{A_1}\right|.$$

A Bethe hole coupler can be designed by first determining the hole position through (4.67). The hole radius may be determined by the requirement of coupling factor.

In the forward-wave coupler, the waves in both waveguides have the same energy flow directions. The forward coupler may be realized by introducing multiple holes, which are spaced a quarter wavelength apart so that the reverse wave cancels out, as indicated in Figure 4.23.

4.4.4.2 *Branch-Line Coupler*

The branch-line coupler consists of two parallel transmission lines physically coupled together with two or more branch lines between them. The branch

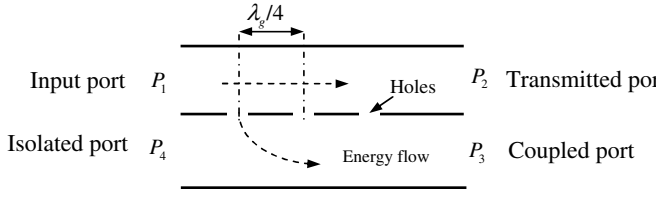


Figure 4.23 Multihole couplers.

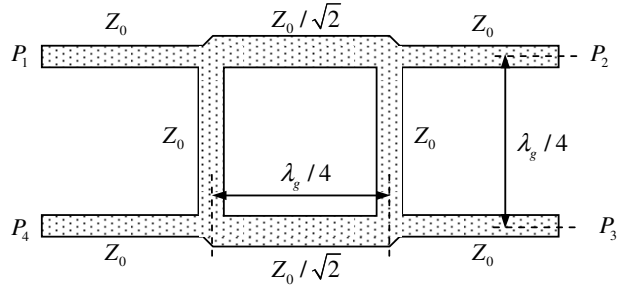


Figure 4.24 Branch-line coupler.

lines are spaced a quarter wavelengths apart and can be used for 3 dB hybrids (Figure 4.24). The scattering matrix for the branch-line coupler is

$$[S] = \frac{-1}{\sqrt{2}} \begin{bmatrix} 0 & j & 1 & 0 \\ j & 0 & 0 & 1 \\ 1 & 0 & 0 & j \\ 0 & 1 & j & 0 \end{bmatrix}.$$

The input power at Port 1 is evenly divided between Port 2 and Port 3 with a 90° phase shift. There is no output power at Port 4.

4.4.5 Filters

The study of microwave filters began in late 1920s (Mason and Sykes, 1937). Many microwave filters were developed during World War II due to the needs in radar and electronic counter-measures. A microwave filter is a two-port network and is used to reject unwanted frequency components of a signal in the stopband and enhance wanted ones within the passband. Some typical ideal filter response curves are shown in Figure 4.25. An ideal filter would have zero attenuation in the passband and infinite attenuation in the stopband.

4.4.5.1 *Insertion Loss*

The **insertion loss** of a network between a source and load is defined by (Figure 4.26)

$$\text{IL} = 10 \log \frac{P_A}{P_L} \text{ (dB)}, \tag{4.68}$$

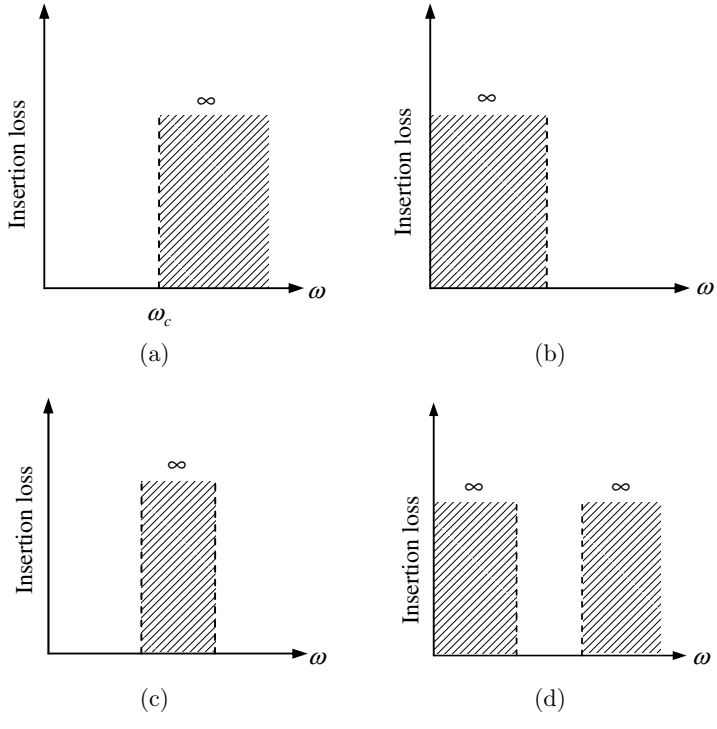


Figure 4.25 Typical filter response curves. (a) Low-pass filter. (b) High-pass filter. (c) Band rejection filter. (d) Bandpass filter.

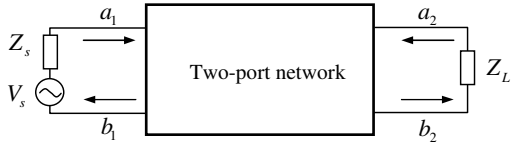


Figure 4.26 Insertion loss.

where P_A is the available power from the source; P_L is the power received by the load. If the load is matched, (4.68) can be written as

$$\text{IL} = 10 \log \frac{|a_1|^2/2}{|b_2|^2/2} = 10 \log \frac{1}{|S_{21}|^2}. \tag{4.69}$$

If the network is lossless, this can be expressed as

$$\text{IL} = 10 \log \frac{1}{1 - |\Gamma|^2}, \tag{4.70}$$

where Γ is the reflection coefficient at the input of the network. Since

$$|\Gamma|^2 = \Gamma(j\omega)\bar{\Gamma}(j\omega) = \Gamma(j\omega)\Gamma(-j\omega),$$

both $|S_{21}|^2$ and $|\Gamma|^2$ are functions of ω^2. From the relationship between the reflection coefficient and the input impedance we may find that

$$|\Gamma|^2 = \Gamma(j\omega)\Gamma(-j\omega) = \frac{A(\omega^2)}{B(\omega^2)},$$

where A and B are polynomials of same order and are functions of ω^2. Let $s = j\omega$. We have

$$\Gamma(s)\Gamma(-s) = \frac{G(-s^2)}{H(-s^2)} = \frac{s^{2n} + G_{n-1}s^{2n-2} + \cdots + G_1 s^2 + G_0}{s^{2n} + H_{n-1}s^{2n-2} + \cdots + H_1 s^2 + H_0}. \quad (4.71)$$

Note that the coefficients G_i $(i = 0, 1, \ldots, n)$ and H_i $(i = 0, 1, \ldots, n)$ are real. If we let $s = \sigma + j\omega$, the above expression can be extended to the complex frequency domain. By factorization, we have

$$\Gamma(s)\Gamma(-s) = \frac{(s - s_{G1})(s - s_{G2})\ldots(s - s_{Gn})(s - s'_{G1})(s - s'_{G2})\ldots(s - s'_{Gn})}{(s - s_{H1})(s - s_{H2})\ldots(s - s_{Hn})(s - s'_{H1})(s - s'_{H2})\ldots(s - s'_{Hn})}, \quad (4.72)$$

where $\{s_{Gi} | i = 1, 2, \ldots, n\}$ and $\{s_{Hi} | i = 1, 2, \ldots, n\}$ are respectively the roots of the numerator and denominator of (4.71) on the left half of the complex frequency plane; $\{s'_{Gi} | i = 1, 2, \ldots, n\}$ and $\{s'_{Hi} | i = 1, 2, \ldots, n\}$ are respectively the roots of the numerator and denominator of (4.71) on the right half of the complex frequency plane. Equation (4.72) can be written as

$$\Gamma(s)\Gamma(-s) = \frac{P(s)P'(s)}{Q(s)Q'(s)}, \quad (4.73)$$

where P and Q are the polynomials that have roots on the left half of the complex frequency plane; P' and Q' are the polynomials that have roots on the right half of the complex frequency plane. Thus, we may choose

$$\Gamma(s) = \pm\frac{P(s)}{Q(s)}. \quad (4.74)$$

The normalized input impedance can be written as

$$z(s) = \frac{1+\Gamma(s)}{1-\Gamma(s)} = \frac{Q(s)\pm P(s)}{Q(s)\mp P(s)}.$$

Example 4.8: For

$$|\Gamma|^2 = \frac{\omega^6}{\omega^6+1},$$

we may write

$$\Gamma(s)\Gamma(-s) = \frac{-s^6}{1-s^6} = \frac{P(s)P'(s)}{Q(s)Q'(s)}.$$

The roots of the equation $-s^6 = 0$ are all zero. We simply choose $P(s) = s^3$. The equation $1 - s^6 = 0$ has six roots

$$s_n = e^{jn\frac{\pi}{3}} \quad (n = 1, 2, 3, 4, 5, 6),$$

where s_2, s_3, and s_4 are on the left half of the complex frequency plane. Hence $Q(s)$ can be chosen as

$$Q(s) = (s-s_2)(s-s_3)(s-s_4) = s^3 + 2s^2 + 2s + 1.$$

The normalized input impedance is thus given by

$$z(s) = \frac{Q(s)+P(s)}{Q(s)-P(s)} = \frac{2s^3+2s^2+2s+1}{2s^2+2s+1}.$$

This can be written as

$$z(s) = s + \frac{1}{2s + \frac{1}{s+\frac{1}{1}}}.$$

The normalized input impedance can be realized by a ladder network as shown in Figure 4.27. □

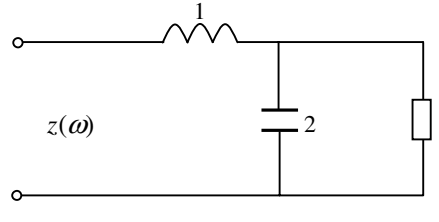


Figure 4.27 Synthesized ladder network.

4.4.5.2 *Low-Pass Filter Prototypes*

Prototype filter is used as a template, and all other filters can be derived from it by applying a scaling factor to the components of the prototype. Most commonly, the prototype filter is a low-pass filter. The response of an ideal low-pass filter is shown in Figure 4.25(a), which is not practical. But it can be approximated by various practical responses. A general expression of the insertion loss for the low-pass prototypes can be expressed as

$$\mathrm{IL} = 10 \log \left[1 + P_n^2(x)\right], \tag{4.75}$$

where $x = \omega/\omega_c$ (called normalized frequency) and ω_c is referred to as cut-off frequency. The function $P_n(x)$ may take various forms depending on the specified responses. The low-pass prototypes can be realized by the ladder networks shown in Figure 4.28. The element values g_i $(i = 1, 2, \ldots, n)$ are normalized inductance for series inductors or capacitance for shunt capacitors and they alternate between series and shunt connections. The element values g_0 and g_{n+1} are normalized resistance or conductance for the source and load respectively. The element g_0 is resistance (or conductance) if g_1 is capacitance (or inductance). The element g_{n+1} is resistance (or conductance) if g_n is a shunt capacitor (or series inductor).

The two circuits shown in Figure 4.28 are dual of each other. Once the element values for the prototypes are known, the practical element values

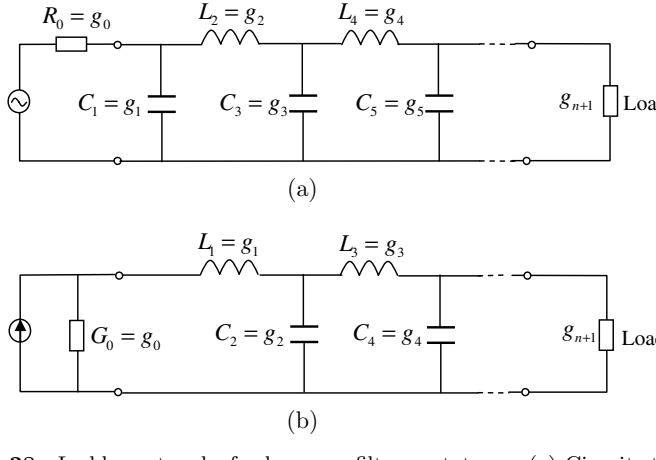


Figure 4.28 Ladder networks for low-pass filter prototypes. (a) Circuit starting with a shunt capacitor. (b) Circuit starting with a series inductor.

can be found by scaling with respect to the source resistance Z_s and the cut-off frequency ω_c as follows

(1) Resistance: $R_0 = g_0 Z_s, R_{n+1} = g_{n+1} Z_s$.
(2) Conductance: $G_0 = \frac{g_0}{Z_s}, G_{n+1} = \frac{g_{n+1}}{Z_s}$.
(3) Inductance: $L_i = \frac{Z_s g_i}{\omega_c}$.
(4) Capacitance: $C_i = \frac{g_i}{Z_s \omega_c}$.

1. Maximally Flat Response (Butterworth Response)

The Butterworth response (Figure 4.29), proposed by Butterworth in 1930, is defined by

$$\mathrm{IL} = 10 \log(1 + \alpha x^{2n}). \tag{4.76}$$

The passband is from $\omega = 0$ to $\omega = \omega_c$. In the passband the insertion loss increases very slowly and is flat. If the insertion loss is required to be 3 dB at the band edge ($x = 1$) we have $\alpha = 1$. In this case, (4.76) becomes

$$\mathrm{IL} = 10 \log(1 + x^{2n}). \tag{4.77}$$

For $x > 1$, the insertion loss increases rapidly with x. If an insertion loss IL_s is required at x_s in the stopband, the minimum number n (the number of reactive components in the prototypes or the order of the filter) is thus given by

$$n = \frac{\lg[10^{(\mathrm{IL}_s/10)} - 1]}{2 \lg x_s}. \tag{4.78}$$

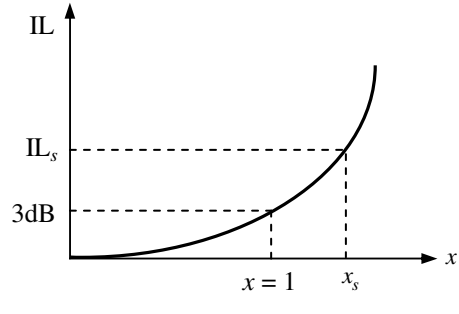


Figure 4.29 Butterworth response.

Equation (4.77) can be extended to the complex frequency plane by letting $s = jx$

$$\mathrm{IL} = 10\log[1 + (-s^2)^n]. \tag{4.79}$$

This implies

$$|\Gamma|^2 = \frac{(-s^2)^n}{1 + (-s^2)^n}. \tag{4.80}$$

It follows that

$$P(s) = s^n, \quad Q(s) = \prod_{i=1}^{n} (s - s_i).$$

where s_i $(i = 1, 2, \ldots, n)$ are the roots of the equation $1 + (-s^2)^n = 0$, which are on the left half of the complex frequency plane

$$s_i = e^{j\frac{2i+n-1}{2n}\pi} \quad (i = 1, 2, \ldots, n). \tag{4.81}$$

The normalized input impedance is thus given by

$$z(s) = \frac{Q(s) - P(s)}{Q(s) + P(s)} = \frac{a_{n-1}s^{n-1} + a_{n-2}s^{n-2} + \cdots a_1 s + a_0}{2s^n + a_{n-1}s^{n-1} + a_{n-2}s^{n-2} + \cdots a_1 s + a_0}. \tag{4.82}$$

where

$$\begin{aligned}
a_{n-1} &= -(s_1 + s_2 + \cdots + s_n), \\
a_{n-2} &= s_1 s_2 + s_2 s_3 + \cdots, \\
a_{n-3} &= s_1 s_2 s_3 + s_1 s_2 s_4 + \cdots, \\
&\vdots \\
a_0 &= (-1)^n s_1 s_2 s_3 \cdots s_n.
\end{aligned}$$

The normalized input impedance can be realized by the ladder networks shown in Figure 4.28. It can be shown that the element values are

$$g_i = 2\sin(2i - 1)\frac{\pi}{2n} \quad (i = 1, 2, \ldots, n). \tag{4.83}$$

Table 4.3 Element values for Butterworth prototypes ($g_0 = 1$, $\omega_c = 1$)

n	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
1	2.000	1.000							
2	1.414	1.414	1.000						
3	1.000	2.000	1.000	1.000					
4	0.765	1.848	1.848	0.765	1.000				
5	0.618	1.618	2.000	1.618	0.618	1.000			
6	0.518	1.414	1.932	1.932	1.414	0.518	1.000		
7	0.445	1.247	1.802	2.000	1.802	1.247	0.445	1.000	
8	0.390	1.111	1.663	1.962	1.962	1.663	1.111	0.390	1.000

The element values from (4.83) are listed in Table 4.3 for $n = 1$–8. Note that the load is always unity.

Example 4.9: Design a Butterworth low-pass filter with a 3 dB cut-off frequency $f_c = 1\,\text{GHz}$. The source resistance is $Z_s = 50\,\Omega$. The insertion loss at 1.5 GHz is at least 15 dB.

Solution: It follows from (4.78) that the minimum order of the filter is

$$n = \frac{\lg(10^{1.5} - 1)}{2\lg 1.5} = 4.2. \tag{4.84}$$

So we choose $n = 5$. This determines the prototype element values g_i ($i = 1, 2, 3, 4, 5$), which can be found from Table 4.3. If Figure 4.28(a) is used as the prototype, the filter circuit element values can be determined by scaling as follows

$$R_0 = g_0 Z_s = 1 \times 50 = 50\,\Omega, \qquad R_6 = g_6 Z_s = 1 \times 50 = 50\,\Omega,$$

$$L_2 = \frac{Z_s g_2}{\omega_c} = \frac{50 \times 1.618}{2\pi \times 10^9} = 12.87\,\text{nH}, \qquad L_4 = \frac{Z_s g_4}{\omega_c} = \frac{50 \times 1.618}{2\pi \times 10^9} = 12.87\,\text{nH},$$

$$C_1 = \frac{g_1}{Z_s \omega_c} = \frac{0.618}{50 \times 2\pi \times 10^9} = 1.98\,\text{pF}, \quad C_3 = \frac{g_3}{Z_s \omega_c},$$

$$C_5 = \frac{g_5}{Z_s \omega_c} = \frac{0.618}{50 \times 2\pi \times 10^9} = 1.98\,\text{pF}. \qquad = \frac{2}{50 \times 2\pi \times 10^9} = 6.37\,\text{pF},$$

Figure 4.30 shows the final filter circuit, which starts with a shunt capacitor and ends with a shunt capacitor. □

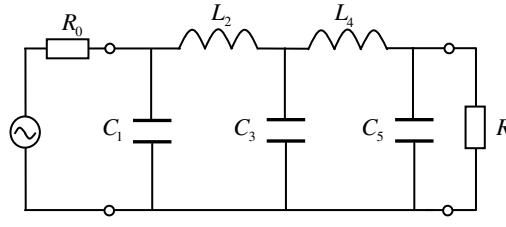


Figure 4.30 Low-pass filter for Example 4.9.

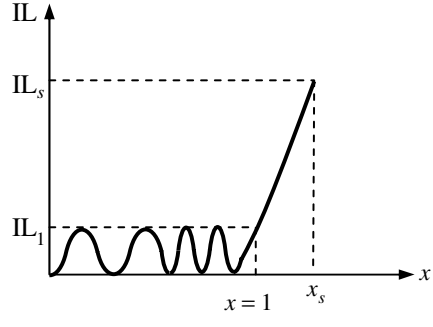


Figure 4.31 Chebyshev response.

2. Equal-Ripple Low-Pass Filter Response (Chebyshev Response)

The Chebyshev response (Figure 4.31) is defined by

$$\mathrm{IL} = 10\log\bigl[1 + \alpha T_n^2(x)\bigr], \tag{4.85}$$

where $T_n(x)$ is the nth order Chebyshev polynomial, defined by

$$T_n(x) = \begin{cases} \cos(n\cos^{-1}x), & 0 \le x \le 1 \\ \cosh(n\cosh^{-1}x), & x > 1 \end{cases}. \tag{4.86}$$

Equivalently, the Chebyshev polynomial can be defined by

$$\begin{aligned} &T_1(x) = x, \quad T_2(x) = 2x^2 - 1, \\ &T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x), \quad n > 2. \end{aligned} \tag{4.87}$$

Since $T_n(x)$ oscillates between ± 1 in $|x| < 1$, the insertion loss have ripples of amplitude $\mathrm{IL}_1 = 10\log(1+\alpha)$, which is the maximum insertion loss in the

passband from $\omega = 0$ to $\omega = \omega_c$. The passband ripple level is determined by α

$$\alpha = 10^{\mathrm{IL}_1/10} - 1. \tag{4.88}$$

In the region $x > 1$, the insertion loss increases rapidly with x. Similarly if an insertion loss IL_s is required at x_s in the stopband, the minimum number n of the reactive components is given by

$$n = \frac{\cosh^{-1}\sqrt{(10^{\mathrm{IL}_s/10} - 1)/\alpha}}{\cosh^{-1} x_s}. \tag{4.89}$$

Equation (4.85) can be extended to the complex frequency plane by letting $s = jx$

$$\mathrm{IL} = 10\log\bigl[1 + \alpha T_n^2(-js)\bigr]. \tag{4.90}$$

It follows that

$$|\Gamma|^2 = \frac{\alpha T_n^2(-js)}{1 + \alpha T_n^2(-js)}. \tag{4.91}$$

Ignoring the details of derivation, one may find that

$$P(s) = \prod_{i=1}^{n}(s - s_i'), \quad Q(s) = \prod_{i=1}^{n}(s - s_i).$$

Here s_i' and s_i $(i = 1, 2, \ldots, n)$ are respectively the roots of the equation $\alpha T_n^2(-js) = 0$ and equation $1 + \alpha T_n^2(-js) = 0$ on the left half of the complex frequency plane

$$\begin{cases} s_i' = j\cos\dfrac{(2i-1)\pi}{2n} \\ s_i = -\sinh\theta_2 \sin\dfrac{(2i-1)\pi}{2n} + j\cosh\theta_2\cos\dfrac{(2i-1)\pi}{2n} \end{cases}, \quad i = 1, 2, \ldots, n \tag{4.92}$$

with

$$\sinh\theta_2 = \frac{1}{2}(\chi^{1/n} - \chi^{-1/n}), \quad \cosh\theta_2 = \frac{1}{2}(\chi^{1/n} + \chi^{-1/n}),$$

$$\chi = \sqrt{1 + \frac{1}{\alpha}} + \frac{1}{\sqrt{\alpha}}.$$

The normalized input impedance is thus given by

$$z(s) = \frac{Q(s) - P(s)}{Q(s) + P(s)}. \tag{4.93}$$

From the input impedance, the normalized element values of the ladder networks can be determined as follows

$$\begin{aligned} g_1 &= \frac{2a_1}{\gamma}, \\ g_i &= \frac{4a_{i-1}a_i}{b_{i-1}g_{i-1}} \quad (i = 2, 3, \ldots, n), \\ g_{n+1} &= \begin{cases} 1 \ (n \text{ is odd}) \\ \tanh^2(\beta/4) \ (n \text{ is even}) \end{cases}, \end{aligned} \tag{4.94}$$

where

$$\begin{aligned} \beta &= \ln\left[\coth\left(\frac{\mathrm{IL}_1}{17.37}\right)\right], \\ \gamma &= \sinh\left(\frac{\beta}{2n}\right), \\ a_i &= \sin\left(\frac{2i-1}{2n}\pi\right) \quad (i = 1, 2, \ldots, n), \\ b_i &= \gamma^2 + \sin^2\left(\frac{i\pi}{n}\right) \quad (i = 1, 2, \ldots, n). \end{aligned}$$

For convenience, the element values from (4.94) are listed in Table 4.4 for $n = 1$–8.

Example 4.10: Design a Chebyshev low-pass filter with a bandwidth $f_c = 1.9\,\mathrm{GHz}$. The ripple level is $0.1\,\mathrm{dB}$ and the insertion loss at $2.5\,\mathrm{GHz}$ is at least $30\,\mathrm{dB}$. The source impedance is $Z_s = 50\,\Omega$.

Solution: From (4.88) and (4.89), we may find that the minimum order of the filter is $n = 7.8$. So we choose $n = 8$. The prototype element values can be found from Table 4.4. If Figure 4.28(b) is used as the prototype, the filter circuit element values can be determined by scaling as follows

$$\begin{aligned} G_0 &= \frac{g_0}{Z_s} = \frac{1}{50} = 0.02\,\mho, \\ R_9 &= g_9 Z_s = 1.355 \times 50 = 67.75\,\Omega, \end{aligned}$$

Table 4.4 Element values for Chebyshev prototypes ($g_0 = 1,\ \omega_c = 1$)

n	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
				$\mathrm{IL}_1 = 0.1\,\mathrm{dB}$					
1	0.305	1.000							
2	0.843	0.622	1.355						
3	1.032	1.147	1.032	1.000					
4	1.109	1.306	1.770	0.818	1.355				
5	1.147	1.371	1.975	1.371	1.147	1.000			
6	1.168	1.404	2.056	1.517	1.903	0.862	1.355		
7	1.181	1.423	2.097	1.573	2.097	1.423	1.181	1.000	
8	1.190	1.435	2.120	1.601	2.170	1.564	1.944	0.87 8	1.355
				$\mathrm{IL}_1 = 0.5\,\mathrm{dB}$					
1	0.699	1.000							
2	1.403	0.707	1.984						
3	1.596	1.097	1.596	1.000					
4	1.670	1.193	2.366	0.842	1.984				
5	1.706	1.230	2.541	1.230	1.706	1.000			
6	1.725	1.248	2.606	1.314	2.476	0.870	1.984		
7	1.737	1.258	2.638	1.344	2.638	1.258	1.737	1.000	
8	1.745	1.265	2.656	1.359	2.696	1.339	2.509	0.880	1.984

$$L_1 = \frac{Z_s g_1}{\omega_c} = \frac{50 \times 1.19}{2\pi \times 1.9 \times 10^9} = 4.98\,\mathrm{nH},$$

$$L_3 = \frac{Z_s g_3}{\omega_c} = \frac{50 \times 2.12}{2\pi \times 1.9 \times 10^9} = 8.88\,\mathrm{nH},$$

$$L_5 = \frac{Z_s g_5}{\omega_c} = \frac{50 \times 2.17}{2\pi \times 1.9 \times 10^9} = 9.09\,\mathrm{nH},$$

$$L_7 = \frac{Z_s g_7}{\omega_c} = \frac{50 \times 1.944}{2\pi \times 1.9 \times 10^9} = 8.14\,\mathrm{nH},$$

$$C_2 = \frac{g_2}{Z_s \omega_c} = \frac{1.435}{50 \times 2\pi \times 1.9 \times 10^9} = 2.40\,\mathrm{pF},$$

$$C_4 = \frac{g_4}{Z_s \omega_c} = \frac{1.601}{50 \times 2\pi \times 1.9 \times 10^9} = 2.68\,\mathrm{pF},$$

$$C_6 = \frac{g_6}{Z_s \omega_c} = \frac{1.564}{50 \times 2\pi \times 1.9 \times 10^9} = 2.61\,\mathrm{pF},$$

$$C_8 = \frac{g_8}{Z_s \omega_c} = \frac{0.878}{50 \times 2\pi \times 1.9 \times 10^9} = 1.47\,\mathrm{pF}.$$

Note that the final filter circuit begins with a series inductor and ends with a shunt capacitor.

4.4.5.3 *Frequency Transformations*

Filters are often required to operate at many other different frequency bands (other than the low-pass filters previously studied), and they can be derived from the prototype filters by applying a transformation to achieve high-pass, bandpass and bandstop characteristics.

1. Low-Pass to High-Pass Transformation

We use x to denote the frequency for the low-pass prototype filter, and ω for the high-pass filter, and their frequency responses of insertion loss are shown in Figure 4.32. It can be seen that the response of the low-pass prototype in the second quadrant is similar to the high-pass response. A transformation from low-pass to high-pass may be constructed by requiring that the insertion losses for both filters are the same at the following three frequency points

$$x = -\infty \leftrightarrow \omega = 0,$$
$$x = -1 \leftrightarrow \omega = \omega_1,$$
$$x = 0 \leftrightarrow \omega = \infty.$$

The transformation is thus given by

$$x = -\frac{\omega_1}{\omega}. \tag{4.95}$$

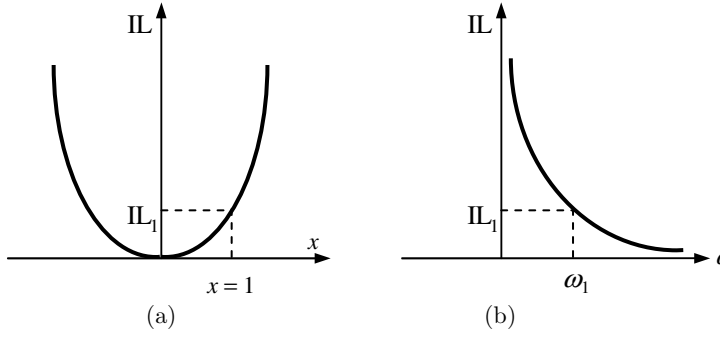


Figure 4.32 Low-pass to high-pass transformation.

Applying (4.95) to the series reactances jxL_i and shunt susceptances jxC_i of the prototype filter, we obtain

$$jxL_i = -j\frac{\omega_1}{\omega}L_i = \frac{1}{j\omega C_i'},$$
$$jxC_i = -j\frac{\omega_1}{\omega}C_i = \frac{1}{j\omega L_i'},$$

where $C_i' = \frac{1}{\omega_1 L_i}$ and $L_i' = \frac{1}{\omega_1 C_i}$ are the normalized capacitance and inductance in the high-pass filter. The above analysis indicates that the inductances and capacitances in the prototype filter have been replaced by capacitances and inductances respectively, as illustrated in Figure 4.33. If the source resistance is Z_s, the component values can be obtained by scaling as follows

$$C_i' = \frac{1}{Z_s\omega_1 L_i}, \quad L_i' = \frac{Z_s}{\omega_1 C_i}. \tag{4.96}$$

2. Low-Pass to Bandpass Transformation

A transformation from low-pass to bandpass may be constructed by requiring that the insertion losses for both filters are the same at the

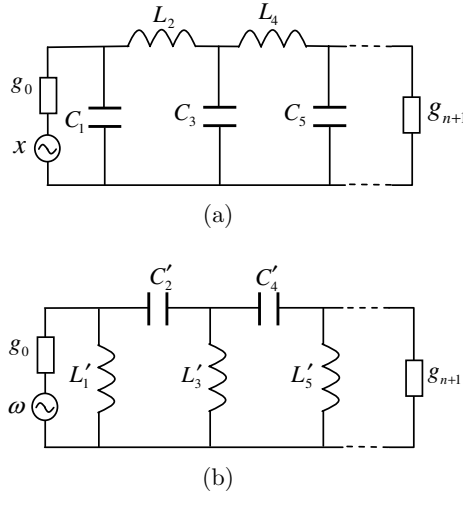


Figure 4.33 Transformation from low-pass prototype filter to high-pass filter. (a) Low-pass filter. (b) High-pass filter.

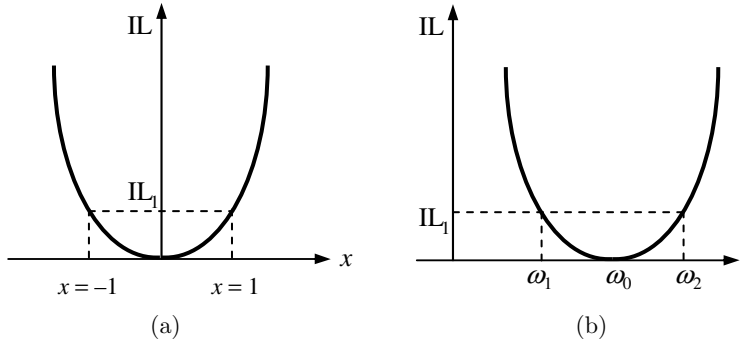


Figure 4.34 Low-pass to bandpass transformation.

following five frequency points (Figure 4.34)

$$x = -\infty \leftrightarrow \omega = 0,$$
$$x = -1 \leftrightarrow \omega = \omega_1,$$
$$x = 0 \leftrightarrow \omega = \omega_0,$$
$$x = 1 \leftrightarrow \omega = \omega_2,$$
$$x = \infty \leftrightarrow \omega = \infty.$$

The transformation is then given by

$$x = \frac{1}{B_f}\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right), \tag{4.97}$$

where $B_f = \frac{\omega_2 - \omega_1}{\omega_0}$ is the fractional bandwidth of the passband. Applying (4.97) to the series reactances jxL_i and shunt susceptances jxC_i of the prototype filter, we obtain

$$jxL_i = j\frac{1}{B_f}\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right)L_i = j\left(\omega L'_{si} - \frac{1}{\omega C'_{si}}\right),$$
$$jxC_i = j\frac{1}{B_f}\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right)C_i = j\left(\omega C'_{pi} - \frac{1}{\omega L'_{pi}}\right),$$

where

$$L'_{si} = \frac{L_i}{B_f\omega_0}, \quad C'_{si} = \frac{B_f}{\omega_0 L_i}, \quad C'_{pi} = \frac{C_i}{B_f\omega_0}, \quad L'_{pi} = \frac{B_f}{\omega_0 C_i}$$

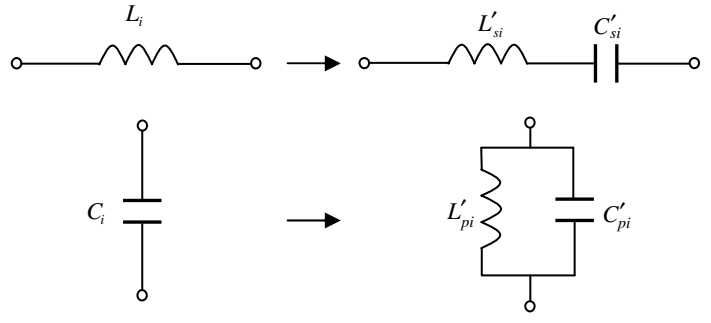


Figure 4.35 Transformation from low-pass prototype filter to bandpass filter.

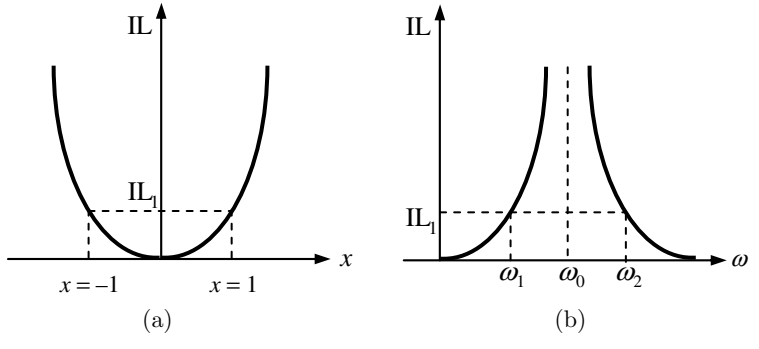


Figure 4.36 Low-pass to bandpass transformation.

are the normalized capacitances and inductances in the bandpass filter. Thus the inductances and capacitances in the prototype filter have been replaced by a series LC circuit and a shunt LC circuit respectively, as illustrated in Figure 4.35.

3. Low-Pass to Bandstop Transformation

A transformation from low-pass to bandstop may be constructed by requiring that the insertion losses for both filters are the same at the following frequency points (Figure 4.36)

$$x = \pm\infty \leftrightarrow \omega = \omega_0,$$

$$x = -1 \leftrightarrow \omega = \omega_2,$$

$$x = 0 \leftrightarrow \omega = 0, \infty,$$

$$x = 1 \leftrightarrow \omega = \omega_1.$$

The transformation is found to be

$$\frac{1}{x} = -\frac{1}{B_f}\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right). \tag{4.98}$$

Similarly, we have

$$jxL_i = -jB_f\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right)^{-1} L_i = \left(\frac{1}{j\omega L'_{si}} + j\omega C'_{si}\right)^{-1},$$

$$jxC_i = -jB_f\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right)^{-1} C_i = \left(\frac{1}{j\omega C'_{pi}} + j\omega L'_{pi}\right)^{-1},$$

where

$$L'_{si} = \frac{B_f L_i}{\omega_0}, \quad C'_{si} = \frac{1}{B_f \omega_0 L_i}, \quad C'_{pi} = \frac{B_f C_i}{\omega_0}, \quad L'_{pi} = \frac{1}{B_f \omega_0 C_i}$$

are the normalized capacitances and inductances in the bandstop filter. In order to obtain the bandstop filter, we only need to replace the inductances and capacitances in the prototype filter by series LC circuits as illustrated in Figure 4.37.

4.4.5.4 *Filter Implementation*

The previous lumped-element filters can be realized by the distributed-elements. Paul Richards proposed the commensurate line theory in 1948 (Richards, 1948), which can be used to replace the lumped elements

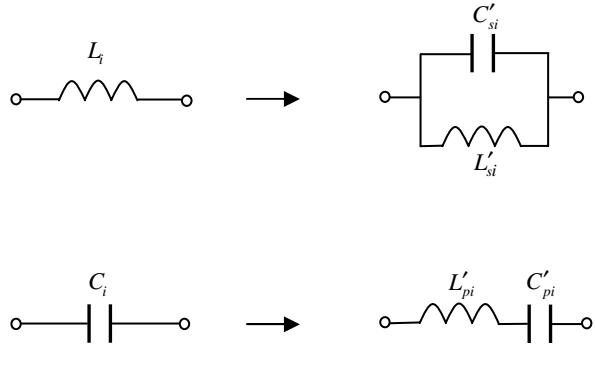


Figure 4.37 Transformation from low-pass prototype filter to bandstop filter.

by transmission line sections of same length with different characteristic impedances. Richards' theory allows a lumped-element design to be transformed directly into a distributed element design by using a simple transformation, called **Richards transformation**. The distributed element design resulted from Richards transformation includes series connected elements, which are generally difficult to implement. K. Kuroda solved this problem by introducing a set of transformations known as **Kuroda's identities**, published in Kuroda's PhD thesis in Japanese in 1955, to eliminate the series elements. The Richards transformation is defined by

$$x = \tan \beta l = \tan\left(\frac{2\pi}{\lambda} l\right) = \tan\left(\frac{\omega}{v_p} l\right), \tag{4.99}$$

where β is the propagation constant, λ is the wavelength, v_p is the phase velocity and l is the length of the transmission line. If x is used as the frequency variable, the reactance of an inductor and the susceptance of a capacitor can thus be written as

$$jxL = jL \tan \beta l, \quad jxC = jC \tan \beta l. \tag{4.100}$$

Therefore, an inductor (or a capacitor) can be replaced with a short (or open)-circuited stub of length βl and characteristic impedance L (or $1/C$). Let x be the normalized frequency for a low-pass filter prototype. At cut-off frequency ω_c, (4.99) becomes

$$x = 1 = \tan \beta l,$$

which implies $l = \lambda/8$, where λ is the wavelength at the cut-off frequency ω_c. Therefore, all the inductors and capacitors in the low-pass filter prototypes can be replaced with short-circuited and open-circuited stubs of the same length, which is $\lambda/8$ at ω_c. These lines are called **commensurate lines**.

Example 4.11: The low-pass filter prototype shown in Figure 4.38(a) can be transformed to the distributed element circuit shown in Figure 4.38(b) by using Richards transformation, where the series and shunt stubs have the same length $\lambda/8$ at ω_c. □

It is noted that the series stubs in Figure 4.38(b) are difficult to implement in practice. In addition the distance between the two series stubs is zero, which is also difficult to implement. To solve these problems one can use Kuroda's identities. One of the Kuroda's identities is shown in Figure 4.39, where the two circuits can be easily shown to be equivalent by using the ABCD parameters for transmission line stubs listed in Table 4.2.

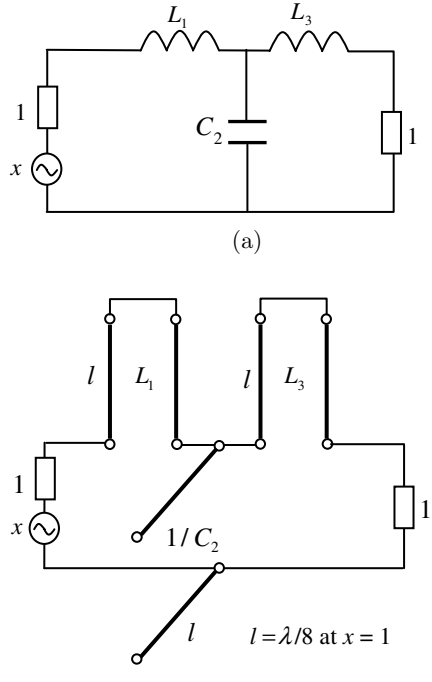


Figure 4.38 Richards transformation. (a) Low-pass filter prototype. (b) The equivalent distributed element circuit.

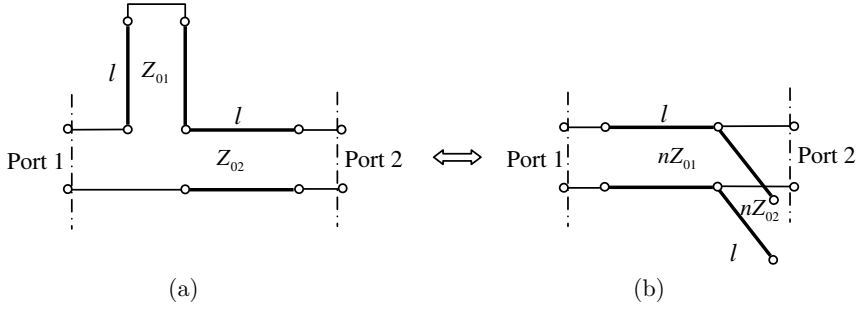
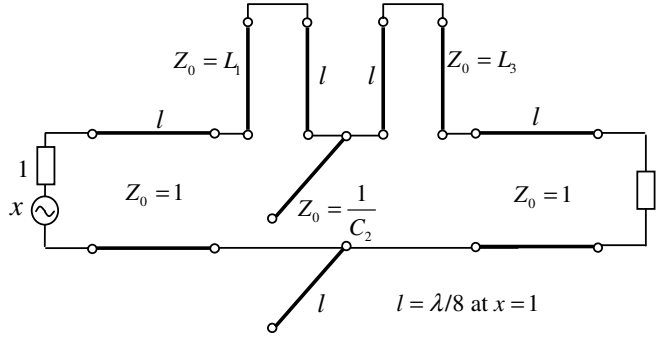
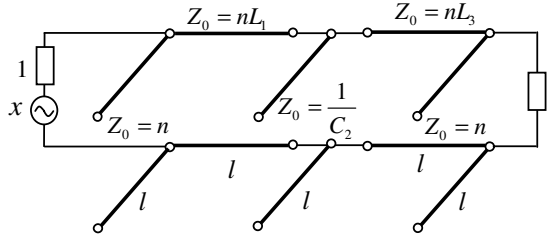


Figure 4.39 Kuroda identity, $n = 1 + Z_{02}/Z_{01}$.

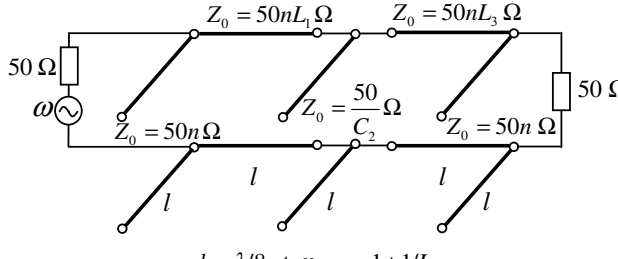
Now we can separate the series transmission line stubs in Figure 4.38(b) by first adding two transmission lines of length $\lambda/8$ at ω_c and characteristic impedance $Z_0 = 1$ as illustrated in Figure 4.40(a). This process would not affect the performance of the filter as the two transmission lines added



(a)

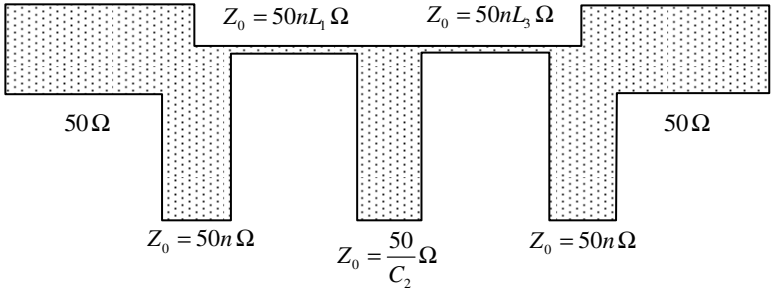


(b)



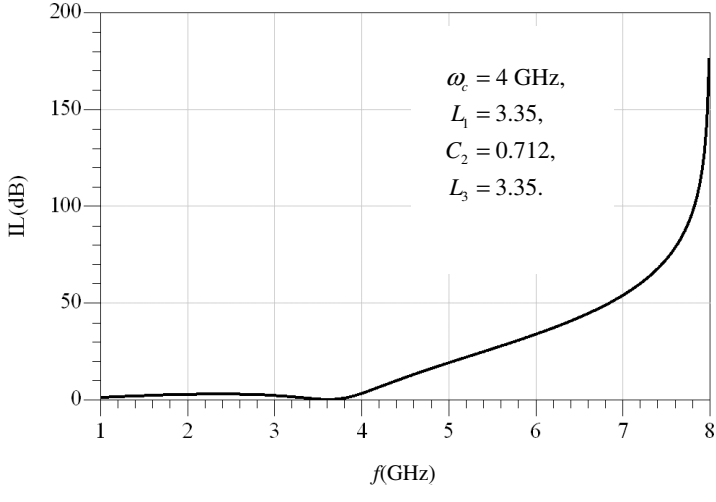
(c)

Figure 4.40 Circuit transformation by Kuroda identity. (a) Adding transmission lines. (b) Applying Kuroda identity. (c) Final circuit after scaling. (d) Layout in microstrip. (e) Simulated insertion loss using ADS.



$\omega_c = 4\,\text{GHz}, L_1 = 3.35, C_2 = 0.712, L_3 = 3.35.$

(d)



(e)

Figure 4.40 (*Continued*)

are matched to the load and the source. The next step is to apply the Kuroda's transformation shown in Figure 4.39 to the distributed circuit in Figure 4.40(a), yielding the distributed circuit shown in Figure 4.40(b). If the system uses $50\,\Omega$ as the reference impedance, the final circuit can be obtained by scaling as shown in Figure 4.40(c), which can be easily fabricated using microstrip. A typical layout is shown in Figure 4.40(d) and the simulated insertion loss using ADS is shown in Figure 4.40(e).

4.5 Active Components

RF systems require some active components that are not passive to achieve certain functions such as amplification, detection, frequency shift and signal generation. Diodes, transistors and tubes are the building blocks of many active components including amplifiers, oscillators, mixers, and detectors. For this reason, an understanding of the basic principles of these components is needed in order to properly bias the transistor or diode to the required operating point. From the practical point of view, the diode or transistors can be characterized by their terminal properties, usually the measured or manufacturer's given two-port parameters, which can be used as our starting point to design various active components.

Microwave components are usually assembled either as hybrid microwave integrated circuits (MIC) or as monolithic microwave integrated circuits (MMIC). In hybrid construction, the transmission lines and matching networks are usually realized as microstrip circuit elements on a substrate, and the discrete components such as chip capacitors, resistors and transistors are connected in place by soldering or using wire-bonding techniques. In a monolithic construction, all active devices and passive circuit elements are fabricated in a single semiconductor crystal, and the overall design and mask making is facilitated by the use of computer-aided design tools.

4.5.1 Amplifiers

In many applications, a signal must be amplified to a useful level in order to perform desired operation. In the amplifier design, the noise performance is one of the important factors to be considered. Other system requirements include gain, bandwidth, and input and output voltage standing wave ratio (VSWR).

At microwave frequencies, the scattering parameters of the transistor can be measured by placing the transistor into a test circuit with appropriate bias. The measured scattering parameters may vary with bias conditions, temperature, and from transistor to transistor. In practice, the design should leave some margin for the variations of scattering parameters.

4.5.1.1 *Power Gains for Two-Port Network*

A general two-port network is shown in Figure 4.41. The **transducer power gain** is defined as the ratio of power dissipated in the load Z_L

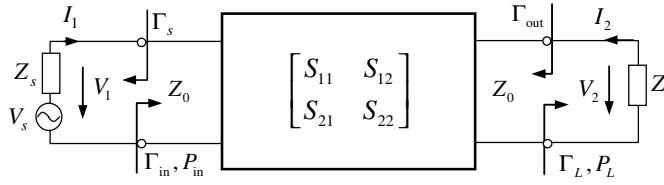


Figure 4.41 Two-port network.

to the available power from the source

$$G_T = \frac{P_L}{P_A}, \tag{4.101}$$

which depends on both Z_s and Z_L. The **power gain** is defined as the ratio of power dissipated in the load Z_L to the input power to the network

$$G_P = \frac{P_L}{P_{\text{in}}}, \tag{4.102}$$

which is independent of the source impedance Z_s. The **available power gain** is defined as the ratio of the available power from the network, denoted by P'_A, and available power from the source

$$G_A = \frac{P'_A}{P_A}, \tag{4.103}$$

which is independent of the load impedance Z_L.

The input reflection coefficient is

$$\Gamma_{\text{in}} = \frac{b_1}{a_1} = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L} = \frac{Z_{\text{in}} - Z_0}{Z_{\text{in}} + Z_0}, \tag{4.104}$$

where Z_{in} is the input impedance and $\Gamma_L = a_2/b_2$ is the reflection coefficient of the load

$$\Gamma_L = \frac{Z_L - Z_0}{Z_L + Z_0}. \tag{4.105}$$

The output reflection coefficient is

$$\Gamma_{\text{out}} = \frac{b_2}{a_2} = S_{22} + \frac{S_{12}S_{21}\Gamma_s}{1 - S_{11}\Gamma_s} = \frac{Z_{\text{out}} - Z_0}{Z_{\text{out}} + Z_0}, \tag{4.106}$$

where Z_{out} is the output impedance and $\Gamma_s = a_1/b_1$ is the reflection coefficient of the source

$$\Gamma_s = \frac{Z_s - Z_0}{Z_s + Z_0}. \tag{4.107}$$

Note that

$$a_1 = \frac{V_1 + Z_0 I_1}{2\sqrt{Z_0}} = \frac{V_s}{2\sqrt{Z_0}} \frac{Z_{\text{in}} + Z_0}{Z_s + Z_{\text{in}}} = \frac{V_s(1-\Gamma_s)}{2\sqrt{Z_0}(1-\Gamma_{\text{in}}\Gamma_s)},$$
$$b_1 = \frac{V_1 - Z_0 I_1}{2\sqrt{Z_0}} = \frac{V_s}{2\sqrt{Z_0}} \frac{Z_{\text{in}} - Z_0}{Z_s + Z_{\text{in}}} = \frac{V_s(1-\Gamma_s)\Gamma_{\text{in}}}{2\sqrt{Z_0}(1-\Gamma_{\text{in}}\Gamma_s)},$$
$$b_2 = \frac{S_{21}a_1}{1-\Gamma_L S_{22}} = \frac{V_s(1-\Gamma_s)}{2\sqrt{Z_0}(1-\Gamma_{\text{in}}\Gamma_s)} \frac{S_{21}}{1-\Gamma_L S_{22}}.$$

The input power to the network is thus given by

$$P_{\text{in}} = \frac{1}{2}\text{Re}\, V_1 I_1 = \frac{1}{2}(|a_1|^2 - |b_1|^2) = \frac{1}{2}|a_1|^2(1-|\Gamma_{\text{in}}|^2)$$
$$= \frac{|V_s|^2|1-\Gamma_s|^2}{8Z_0|1-\Gamma_{\text{in}}\Gamma_s|^2}(1-|\Gamma_{\text{in}}|^2). \tag{4.108}$$

The power absorbed by the load Z_L is

$$P_L = -\frac{1}{2}\text{Re}\, V_2 I_2 = \frac{1}{2}(|b_2|^2 - |a_2|^2) = \frac{1}{2}|b_2|^2\big(1-|\Gamma_L|^2\big)$$
$$= |S_{21}|^2 \frac{|V_s|^2|1-\Gamma_s|^2}{8Z_0|1-\Gamma_{\text{in}}\Gamma_s|^2} \frac{\big(1-|\Gamma_L|^2\big)}{|1-\Gamma_L S_{22}|^2}. \tag{4.109}$$

The available power from the source can be determined from P_{in} by letting $Z_s = \bar{Z}_{\text{in}}$ or $\Gamma_{\text{in}} = \bar{\Gamma}_s$

$$P_A = \frac{|V_s|^2|1-\Gamma_s|^2}{8Z_0(1-|\Gamma_s|^2)}. \tag{4.110}$$

The available power from the network can be determined from P_L by letting $Z_{\text{out}} = \bar{Z}_L$ or $\Gamma_{\text{out}} = \bar{\Gamma}_L$

$$P'_A = \frac{|V_s|^2|1-\Gamma_s|^2}{8Z_0|1-\Gamma_{\text{in}}\Gamma_s|^2_{\Gamma_L=\bar{\Gamma}_{\text{out}}}} \frac{|S_{21}|^2}{|1-\bar{\Gamma}_{\text{out}}S_{22}|^2}\big(1-|\Gamma_{\text{out}}|^2\big). \tag{4.111}$$

It follows from (4.104) and (4.106) that

$$1-\Gamma_{\text{in}}\Gamma_s = (1-S_{11}\Gamma_s)\frac{1-\Gamma_{\text{out}}\Gamma_L}{1-S_{22}\Gamma_L}.$$

Thus, we have

$$P'_A = \frac{|V_s|^2|1-\Gamma_s|^2}{8Z_0\big(1-|\Gamma_{\text{out}}|^2\big)}\frac{|S_{21}|^2}{|1-\Gamma_s S_{11}|^2}. \tag{4.112}$$

Making use of the preceding results, the transducer power gain is:

$$G_T = \frac{P_L}{P_A} = |S_{21}|^2\frac{\big(1-|\Gamma_s|^2\big)}{|1-\Gamma_{\text{in}}\Gamma_s|^2}\frac{\big(1-|\Gamma_L|^2\big)}{|1-\Gamma_L S_{22}|^2}. \tag{4.113}$$

Note that the transducer power gain reduces to

$$G_T = \frac{P_L}{P_A} = |S_{21}|^2, \tag{4.114}$$

when $\Gamma_s = \Gamma_L = 0$. The power gain is

$$G_P = \frac{P_L}{P_{\text{in}}} = \frac{|S_{21}|^2(1-|\Gamma_L|^2)}{\big(1-|\Gamma_{\text{in}}|^2\big)|1-\Gamma_L S_{22}|^2}. \tag{4.115}$$

The available power gain is

$$G_A = \frac{P'_A}{P_A} = \frac{|S_{21}|^2\big(1-|\Gamma_s|^2\big)}{|1-\Gamma_s S_{11}|^2\big(1-|\Gamma_{\text{out}}|^2\big)}. \tag{4.116}$$

4.5.1.2 *Stability Criteria*

The conditions for the stability of a two-port network indicated in Figure 4.41 require that the reflected power from the network ports is smaller than the incident power, i.e.,

$$|\Gamma_{\text{in}}| = \left|S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1-S_{22}\Gamma_L}\right| < 1, \tag{4.117}$$

$$|\Gamma_{\text{out}}| = \left|S_{22} + \frac{S_{12}S_{21}\Gamma_s}{1-S_{11}\Gamma_s}\right| < 1. \tag{4.118}$$

If the above conditions hold for all passive source and load impedances (i.e., $|\Gamma_s| < 1$ and $|\Gamma_L| < 1$), the network is said to be **unconditionally or absolutely stable**. Otherwise, the input or output impedance of the network would have a negative real part. Actually if $Z_{\text{in}} = -R_{\text{in}} + jX_{\text{in}}$,

then

$$|\Gamma_{\text{in}}| = \left|\frac{-R_{\text{in}} + jX_{\text{in}} - Z_0}{-R_{\text{in}} + jX_{\text{in}} + Z_0}\right| = \sqrt{\frac{(R_{\text{in}} + Z_0)^2 + X_{\text{in}}^2}{(R_{\text{in}} - Z_0)^2 + X_{\text{in}}^2}} > 1,$$

and the input current is

$$I = \frac{V_s}{R_s - R_{\text{in}} + j(X_s + X_{\text{in}})}.$$

If $R_s = R_{\text{in}}$ and $X_s + X_{\text{in}} = 0$, the current I becomes infinite. In this case, a self-sustained oscillation can be produced by the thermal noise in the input even if $V_s = 0$. In general, the conditions (4.117) and (4.118) only hold for a restricted set of values for $\Gamma_L$or Γ_s. In this case, the network is said to be **conditionally stable**. Note that if $\Gamma_s = \Gamma_L = 0$, (4.117) and (4.118) imply that

$$|S_{11}| < 1, \quad |S_{22}| < 1. \tag{4.119}$$

Therefore, the two inequalities in (4.119) are necessary if the network is absolutely stable. The **stability circles** are defined as the loci in the Γ_L (or Γ_s) plane for which $|\Gamma_{\text{in}}| = 1$ (or $|\Gamma_{\text{out}}| = 1$), and they define the boundaries between stable and unstable regions of Γ_L and Γ_s. The equations

$$|\Gamma_{\text{in}}| = \left|S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L}\right| = 1,$$

$$|\Gamma_{\text{out}}| = \left|S_{22} + \frac{S_{12}S_{21}\Gamma_s}{1 - S_{11}\Gamma_s}\right| = 1$$

can be written as

$$|\Gamma_L - C_L| = R_L, \tag{4.120}$$

$$|\Gamma_s - C_s| = R_s, \tag{4.121}$$

where

$$C_L = \frac{\overline{S_{22} - \Delta\bar{S}_{11}}}{|S_{22}|^2 - |\Delta|^2}, \quad R_L = \left|\frac{S_{12}S_{21}}{|S_{22}|^2 - |\Delta|^2}\right|, \tag{4.122}$$

$$C_s = \frac{\overline{S_{11} - \Delta\bar{S}_{22}}}{|S_{11}|^2 - |\Delta|^2}, \quad R_s = \left|\frac{S_{12}S_{21}}{|S_{11}|^2 - |\Delta|^2}\right|, \tag{4.123}$$

$$\Delta = S_{11}S_{22} - S_{12}S_{21}. \tag{4.124}$$

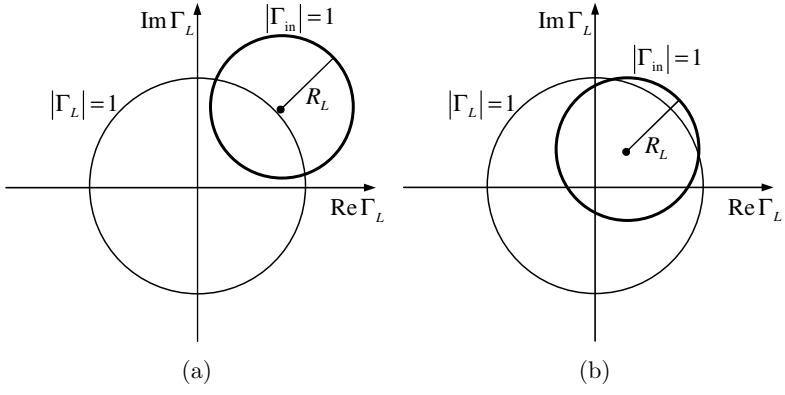


Figure 4.42 Stability circle. (a) Origin is outside stability circle. (b) Origin is inside stability circle.

Equations (4.120) and (4.121) define the output stability and input stability circle respectively. Consider the output stability circle plotted in Figure 4.42. If $\Gamma_L = 0$, we have $|\Gamma_{\text{in}}| = |S_{11}|$. If $|S_{11}| < 1$ (or $|S_{11}| > 1$), then the origin $\Gamma_L = 0$ must be in a stable region (or an unstable region). When the origin $\Gamma_L = 0$ is outside the stability circle, the region that is exterior (or interior) to the stability circle and satisfies $|\Gamma_L| < 1$ is the stable region for Γ_L if $|S_{11}| < 1$ (or $|S_{11}| > 1$). When the origin $\Gamma_L = 0$ is inside the stability circle, the region that is interior (or exterior) to the stability circle and satisfies $|\Gamma_L| < 1$ is the stable region for Γ_L if $|S_{11}| < 1$ (or $|S_{11}| > 1$). Similar discussions can be carried out for the input stability circle.

If the network is absolutely stable, the stability circles must be completely outside the circle $|\Gamma_L| = 1$. The necessary and sufficient conditions for the network to be absolutely stable are given by

$$K = \frac{1 - |S_{11}|^2 - |S_{22}|^2 + |\Delta|^2}{2|S_{12}S_{21}|} > 1, \quad |\Delta| < 1. \tag{4.125}$$

A single sufficient condition for absolute stability is available (Edwards and Sinksy, 1992). If the following condition is met

$$\mu = \frac{1 - |S_{11}|^2}{|S_{22} - \Delta\bar{S}_{11}| + |S_{12}S_{21}|} > 1, \tag{4.126}$$

the network is absolutely stable. Furthermore, the larger the value of μ, the greater is the stability.

4.5.1.3 *Noise Theory for Two-Port Network*

Thermal noise, also known as Johnson–Nyquist noise, after John B. Johnson (Swedish-born American electrical engineer and physicist, 1887–1970) and Harry T. Nyquist (Swedish-born American electrical engineer and physicist, 1889–1976), is a random process generated by the thermal agitation of the charge carriers (usually the electrons) inside an electrical conductor at equilibrium, which is independent of applied voltage. For example, the electrons in a resistor will have a random motion due to the thermal agitation, and thus produce a random voltage across the resistor terminals.

1. Power Spectral Density

A noisy resistor may be modeled as a noise-free resistor in series with a noise voltage generator $e_n(t)$ (Thévenin equivalent circuit) or in shunt with a noise current source $i_n(t)$ (Norton equivalent circuit) as indicated in Figure 4.43.

Thermal noise is usually considered as a stationary ergodic random process for which ensemble averages are equal to time averages (see Section 8.1). The time average of the noise voltage of a resistor is defined by

$$\overline{\overline{e_n(t)}} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} e_n(t)dt, \tag{4.127}$$

which is assumed to be zero. The correlation function of the noise voltage is defined by

$$R_n(\tau) = \overline{\overline{e_n(t+\tau)e_n(t)}} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} e_n(t+\tau)e_n(t)dt. \tag{4.128}$$

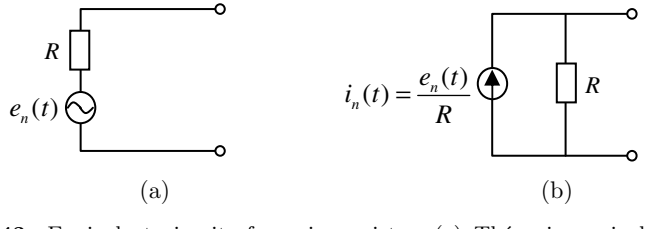


Figure 4.43 Equivalent circuits for noise resistor. (a) Thévenin equivalent circuit. (b) Norton equivalent circuit.

Note that $R(0) = \overline{\overline{e_n^2(t)}}$ represents the average noise power dissipated in a resistor of $1\,\Omega$ and is considered to have the dimension of power. The **power spectral density** of the noise voltage is defined as the Fourier transform of the correlation function

$$S_n(\omega) = \int_{-\infty}^{\infty} R_n(\tau) e^{-j\omega\tau} d\tau. \tag{4.129}$$

Thus

$$R_n(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_n(\omega) e^{j\omega\tau} d\omega. \tag{4.130}$$

Equations (4.129) and (4.130) are called **Wiener–Khintchine relations**.

Thermal noise is approximately white, i.e., the power spectral density is nearly constant throughout the frequency spectrum. The power spectral density function is an even function of ω. Thus we may write

$$R_n(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_n(\omega) d\omega = \frac{1}{2\pi} \int_{0}^{\infty} S_p(\omega) d\omega, \tag{4.131}$$

where $S_p(\omega) = 2S_n(\omega)(\omega > 0)$ is one-sided power-spectral density. For thermal noise in a resistor, the power spectral densities for the noise voltage and current are respectively given by **Nyquist's formulae** (Nyquist, 1928)

$$S_e(\omega) = 4kTR, \quad \omega > 0, \tag{4.132}$$

$$S_i(\omega) = \frac{4kT}{R}, \quad \omega > 0, \tag{4.133}$$

where $k = 1.38 \times 10^{-23}\,\mathrm{J/K}$ is the Boltzmann's constant and T is the absolute temperature of the resistor R.

2. Filtered Noise

Consider a two-port network connected to a voltage source V and a current source I as shown in Figure 4.44. The input power produced by voltage source and current source are respectively given by

$$P_{\text{in}}^V = \frac{1}{2} \left| \frac{V}{Z_s + Z_{\text{in}}} \right|^2 R_{\text{in}} = \frac{1}{2} |V|^2 \frac{M}{4R_s},$$

$$P_{\text{in}}^I = \frac{1}{2} \left| \frac{IZ_s}{Z_s + Z_{\text{in}}} \right|^2 R_{\text{in}} = \frac{1}{2} |I|^2 \frac{M}{4G_s},$$

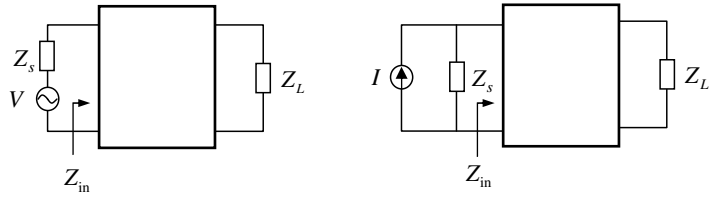


Figure 4.44 A two-port network connected to voltage and current source.

where

$$M = \frac{4R_s R_{\text{in}}}{|Z_s + Z_{\text{in}}|^2} \tag{4.134}$$

is the impedance-mismatch factor, and $R_{\text{in}} = \text{Re}\, Z_{\text{in}}$, $R_s = \text{Re}\, Z_s$. Note that $M/4R_s$ and $M/4G_s$ are power transfer functions. When both the voltage and current source are present, the input power will be

$$P_{\text{in}} = \frac{1}{2}\left|\frac{V + IZ_s}{Z_s + Z_{\text{in}}}\right|^2 R_{\text{in}} = \frac{|V|^2}{2}\frac{M}{4R_s} + \frac{|I|^2}{2}\frac{M}{4G_s} + \text{Re}\left(\frac{V\bar{I}}{2}\frac{2\bar{Z}_s R_{\text{in}}}{|Z_s + Z_{\text{in}}|^2}\right). \tag{4.135}$$

The last term stands for the interaction between the two sources.

Now we consider the situation when the voltage generator V and current generator I are respectively replaced by the noise voltage source $e_n(t)$ and noise current source $i_n(t)$. The cross-correlation between the voltage source and the current source is defined by

$$R_x(\tau) = \overline{\overline{e_n(t+\tau) i_n(t)}} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} e_n(t+\tau) i_n(t) dt. \tag{4.136}$$

The cross-power spectral density is the Fourier transform of the cross-correlation

$$S_x(\omega) = \int_{-\infty}^{\infty} R_x(\tau) e^{-j\omega\tau} d\tau = S_{xr}(\omega) + jS_{xi}(\omega).$$

Note that $S_{xr}(\omega)$ is an even function of ω and $S_{xi}(\omega)$ is an odd function of ω and $S_x(-\omega) = \bar{S}_x(\omega)$. The input noise power spectral density $S_{\text{in}}(\omega)$ can be obtained from (4.135) by replacing $|V|^2/2$ with $S_e(\omega)$, $|I|^2/2$ with

$S_i(\omega)$, $V\bar{I}/2$ with $S_x(\omega)$ as follows

$$S_{\text{in}}(\omega) = S_e(\omega)\frac{M}{4R_s} + S_i(\omega)\frac{M}{4G_s} + \frac{S_{xr}(\omega)R_s + S_{xi}X_s}{R_s}M, \quad \omega > 0. \quad (4.137)$$

Note that both $S_{xr}R_s$ and $S_{xi}X_s$ are even functions of ω, which brings an extra factor of 2 in the last term when only positive frequency is considered. The power spectral density delivered to the load Z_L is

$$\begin{aligned} S_{\text{out}}(\omega) &= G_P(\omega)S_{\text{in}}(\omega) \\ &= G_P(\omega)\left[S_e(\omega)\frac{M}{4R_s} + S_i(\omega)\frac{M}{4G_s}\right. \\ &\quad \left. + \frac{S_{xr}(\omega)R_s + S_{xi}X_s}{R_s}M\right], \quad \omega > 0. \end{aligned} \quad (4.138)$$

The total output noise power delivered to the load is given by

$$P_{n,\text{out}} = \frac{1}{2\pi}\int_0^\infty S_{\text{out}}(\omega)d\omega.$$

The internal noise sources in a noisy two-port network can be replaced by a series noise voltage generator and a shunt noise current generator at the input of the network as shown in Figure 4.45. For thermal noise sources, we may write

$$\begin{aligned} &S_e(\omega) = 4kTR_e, \quad \text{for } e_n(t), \\ &S_i(\omega) = 4kTG_i, \quad \text{for } i_n(t), \\ &2[S_{xr}(\omega) + jS_{xi}(\omega)] = 4kT(\gamma_r + j\gamma_i), \end{aligned}$$

where R_e, G_i and $\gamma_r + j\gamma_i$ are equivalent noise resistance, noise conductance, and noise impedance respectively. The noise power spectral density $S_{\text{in}}(\omega)$ input to the noise-free network include the contributions from R_s, R_e, and

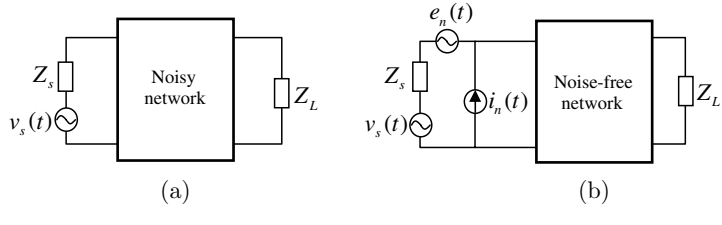


Figure 4.45 A noisy two-port network and equivalent input noise sources.

G_i, and can be obtained from (4.137) as follows

$$S_{\text{in}}(\omega) = kTM + kTM\frac{R_e}{R_s} + kTM\frac{G_i}{G_s} + 2kT\frac{R_s\gamma_r + X_s\gamma_i}{R_s}M, \quad \omega > 0. \tag{4.139}$$

The first term on the right-hand side denotes the input thermal noise power spectral density from the source resistance R_s.

3. Noise Figure

The **noise figure** of a two-port network as shown in Figure 4.45(a) is defined as the ratio of signal-to-noise at input and output:

$$F = \frac{S_{\text{in}}/N_{\text{in}}}{S_{\text{out}}/N_{\text{out}}} \geq 1, \tag{4.140}$$

where S_{in} (or N_{in}) and S_{out} (or N_{out}) are the signal (noise) power at input and output respectively. Equation (4.140) can be rewritten as

$$F = \frac{S_{\text{in}}/N_{\text{in}}}{S_{\text{out}}/N_{\text{out}}} = \frac{S_{\text{in}}/N_{\text{in}}}{G_pS_{\text{in}}/G_p(N_{\text{in}} + N_{\text{internal}})} = \frac{N_{\text{in}} + N_{\text{internal}}}{N_{\text{in}}}. \tag{4.141}$$

Hence the noise figure can be obtained by dividing (4.139) by kTM as follows

$$F = 1 + \frac{R_e}{R_s} + \frac{G_i}{G_s} + 2\frac{R_s\gamma_r + X_s\gamma_i}{R_s}, \quad \omega > 0. \tag{4.142}$$

The noise figure can be minimized by optimizing the source impedance through $\partial F/\partial R_s = \partial F/\partial X_s = 0$. The optimized source impedance, denoted Z_{opt}, is given by

$$Z_{\text{opt}} = \sqrt{\frac{R_e}{G_i} - \frac{\gamma_i^2}{G_i^2}} - j\frac{\gamma_i}{G_i}. \tag{4.143}$$

The noise figure (4.142) can then be written as

$$F = F_{\text{min}} + \frac{G_i}{R_s}|Z_s - Z_{\text{opt}}|^2, \quad \omega > 0, \tag{4.144}$$

where F_{min} is the minimized noise figure when (4.143) is introduced into (4.142). Equation (4.144) can also be written as

$$\begin{aligned} F - F_{\text{min}} &= 4G_iZ_0\frac{|\Gamma_s - \Gamma_{\text{opt}}|^2}{|1 - \Gamma_{\text{opt}}|^2\left(1 - |\Gamma_s|^2\right)} \\ &= 4\frac{R_e}{Z_0}\frac{|\Gamma_s - \Gamma_{\text{opt}}|^2}{|1 + \Gamma_{\text{opt}}|^2\left(1 - |\Gamma_s|^2\right)}, \quad \omega > 0. \end{aligned} \tag{4.145}$$

where

$$\Gamma_s = \frac{Z_s - Z_0}{Z_s + Z_0}, \quad \Gamma_{\text{opt}} = \frac{Z_{\text{opt}} - Z_0}{Z_{\text{opt}} + Z_0}.$$

We may further rewrite (4.145) as

$$|\Gamma_s - C_{NF}| = R_{NF}, \tag{4.146}$$

where

$$C_{NF} = \frac{\Gamma_{\text{opt}}}{N+1}, \quad R_{NF} = \frac{\sqrt{N(N+1-|\Gamma_{\text{opt}}|^2)}}{N+1},$$

$$N = \frac{|\Gamma_s - \Gamma_{\text{opt}}|^2}{1-|\Gamma_s|^2} = \frac{F - F_{\min}}{4G_i Z_0}|1-\Gamma_{\text{opt}}|^2 = \frac{F-F_{\min}}{4R_e/Z_0}|1+\Gamma_{\text{opt}}|^2, \tag{4.147}$$

which is a constant for a given noise figure. Equation (4.146) defines a set of circles in the source reflection coefficient plane Γ_s, called **constant noise figure circles**. When $N = 0$, the circle degenerates to a single point at Γ_{opt} giving the minimum noise figure $F_{\min}$.

4.5.1.4 *Amplifier Design*

A general microwave amplifier circuit is shown in Figure 4.46, where the input and output matching network transform the input and output impedance Z_0 to the source and load impedance Z_s and Z_L; the transistor is characterized by the scattering matrix. The transducer power gain (4.113) can be written as

$$G_T = G_s G_0 G_L, \tag{4.148}$$

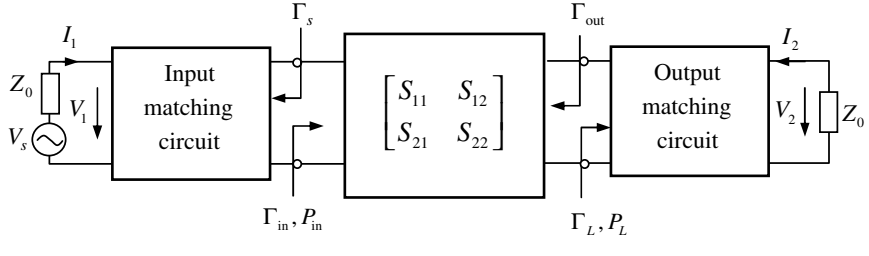


Figure 4.46 Amplifier circuit.

where

$$G_s = \frac{(1-|\Gamma_s|^2)}{|1-\Gamma_{\text{in}}\Gamma_s|^2}, \quad G_0 = |S_{21}|^2, \quad G_L = \frac{(1-|\Gamma_L|^2)}{|1-\Gamma_L S_{22}|^2}, \tag{4.149}$$

respectively represent the gains due to the impedance matching on the source side, the transistor characterized by the scattering matrix, and the impedance matching on the load side.

1. Unilateral Design

In amplifier design, the first step is to determine the stable region of Γ_s and Γ_L. Then the input and output matching circuits can be designed. In many situations, the transistor may be considered as unilateral, i.e., $S_{12} \approx 0$. In this case, we have $\Gamma_{\text{in}} = S_{11}$ and $\Gamma_{\text{out}} = S_{22}$. If both input and output are conjugately matched

$$\Gamma_s = \bar{\Gamma}_{\text{in}} = \bar{S}_{11}, \quad \Gamma_L = \bar{\Gamma}_{\text{out}} = \bar{S}_{22},$$

the gains for the input and output matching circuits will be maximized. Thus

$$\max G_s = \frac{1}{1-|S_{11}|^2}, \quad G_0 = |S_{21}|^2, \quad \max G_L = \frac{1}{1-|S_{22}|^2}. \tag{4.150}$$

We now introduce the normalized gain factors g_s and g_L

$$\begin{aligned} g_s &= \frac{G_s}{\max G_s} = \frac{\left(1-|\Gamma_s|^2\right)}{|1-S_{11}\Gamma_s|^2}\left(1-|S_{11}|^2\right), \\ g_L &= \frac{G_L}{\max G_L} = \frac{\left(1-|\Gamma_L|^2\right)}{|1-S_{22}\Gamma_L|^2}\left(1-|S_{22}|^2\right). \end{aligned} \tag{4.151}$$

Rearranging gives

$$|\Gamma_s - C_s| = R_s, \tag{4.152}$$

$$|\Gamma_L - C_L| = R_L, \tag{4.153}$$

where

$$\begin{aligned} C_s &= \frac{g_s \bar{S}_{11}}{1-(1-g_s)|S_{11}|^2}, \quad R_s = \frac{\sqrt{1-g_s}\left(1-|S_{11}|^2\right)}{1-(1-g_s)|S_{11}|^2}, \\ C_L &= \frac{g_L \bar{S}_{22}}{1-(1-g_L)|S_{22}|^2}, \quad R_L = \frac{\sqrt{1-g_L}\left(1-|S_{22}|^2\right)}{1-(1-g_L)|S_{22}|^2}. \end{aligned} \tag{4.154}$$

Equations (4.152) and (4.153) represent two **constant gain circles** in Γ_s- or Γ_L-plane.

Example 4.12: A FET transistor has the following scattering matrix at 4 GHz:

$$S_{11} = 0.75\angle -120^\circ, \quad S_{12} = 0,$$
$$S_{21} = 2.5\angle 80^\circ, \quad S_{22} = 0.6\angle -70^\circ.$$

We wish to design an amplifier to have a gain of 11 dB at 4 GHz.

Solution: From (4.125), we may find that $K > 1$ and $|\Delta| < 1$. The transistor is thus unilateral and absolutely stable. From (4.150), the maximum matching gains are

$$\max G_s = \frac{1}{1 - |S_{11}|^2} = 3.6\,\text{dB},$$
$$\max G_L = \frac{1}{1 - |S_{22}|^2} = 1.9\,\text{dB}.$$

The gain of the mismatched transistor is

$$G_0 = |S_{21}|^2 = 8\,\text{dB}.$$

Thus the maximum transducer gain is

$$\max G_T = \max G_s + G_0 + \max G_L = 13.47\,\text{dB},$$

which is 2.5 dB higher than the desired value of 11 dB. Since G_0 is 8 dB, the remaining 3 dB can be obtained through G_s and G_L. We choose $G_s = 2$ dB and $G_L = 1$ dB, which determine two constant gain circles on Smith chart with

$$C_s = 0.627\angle 120^\circ, \quad R_s = 0.294,$$
$$C_L = 0.520\angle 70^\circ, \quad R_L = 0.303.$$

determined by (4.154). The reflection coefficients Γ_s and Γ_L are chosen along the constant gain circles as shown in Figure 4.47 to minimize the distance from the center of the chart, where $\Gamma_s = 0.33\angle 120^\circ$, $\Gamma_L = 0.22\angle 70^\circ$. The input and output matching networks can be designed according to the reflection coefficients and the final amplifier circuit is shown in Figure 4.48 (Pozar, 1998). □

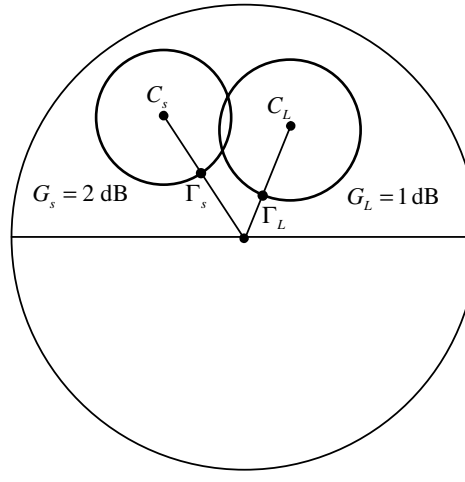


Figure 4.47 Constant gain circles on Smith chart.

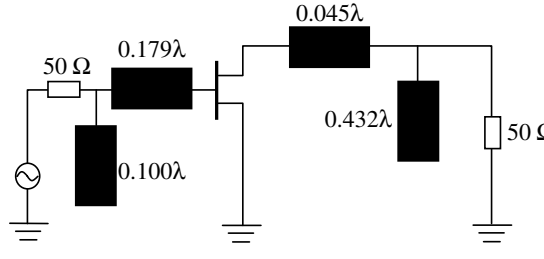


Figure 4.48 Amplifier design.

2. Conjugately Matched Amplifier Design

When the input and output of the transistor are conjugately matched, i.e.,

$$\Gamma_{\text{in}} = \bar{\Gamma}_s, \quad \Gamma_{\text{out}} = \bar{\Gamma}_L, \tag{4.155}$$

the power transfer from the input matching network to the output matching network will be maximized. From (4.113), we obtain the maximized transducer power gain

$$\max G_T = \frac{1}{1-|\Gamma_s|^2}|S_{21}|^2 \frac{\left(1-|\Gamma_L|^2\right)}{|1-\Gamma_L S_{22}|^2}. \tag{4.156}$$

It follow from (4.155), (4.104) and (4.106) that

$$\bar{\Gamma}_s = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L}, \tag{4.157}$$

$$\bar{\Gamma}_L = S_{22} + \frac{S_{12}S_{21}\Gamma_s}{1 - S_{11}\Gamma_s}. \tag{4.158}$$

These equations may be solved to yield the reflection coefficients

$$\Gamma_s = \frac{A_1 \pm \sqrt{A_1^2 - 4|B_1|^2}}{2B_1}, \quad \Gamma_L = \frac{A_2 \pm \sqrt{A_2^2 - 4|B_2|^2}}{2B_2}, \tag{4.159}$$

where

$$\begin{aligned} A_1 &= 1 + |S_{11}|^2 - |S_{22}|^2 - |\Delta|^2, \\ A_2 &= 1 + |S_{22}|^2 - |S_{11}|^2 - |\Delta|^2, \\ B_1 &= S_{11} - \Delta\bar{S}_{22}, \\ B_2 &= S_{22} - \Delta\bar{S}_{11}. \end{aligned}$$

The minus sign is used when $A_i > 0$ and plus sign is used when $A_i < 0$. The input and output matching network can be determined from (4.159) by using Smith chart. Using the conjugate impedance matching conditions (4.155), the power gain (4.115) for an absolutely stable transistor can be written as (e.g., Collin, 2001)

$$\max G_P = \frac{|S_{21}|}{|S_{12}|}\left(K - \sqrt{K^2 - 1}\right), \tag{4.160}$$

where K is given by (4.125).

Example 4.13: The scattering parameters of a FET have the following values at 4 GHz:

$$S_{11} = 0.72\angle{-116}^\circ, \quad S_{12} = 0.03\angle{57}^\circ, \quad S_{21} = 2.60\angle{76}^\circ, \quad S_{22} = 0.73\angle{-54}^\circ.$$

We wish to design an amplifier for maximum gain at 4 GHz.

Solution: It is easy to verify that $K = 1.195 > 1$ and $|\Delta| = 0.48 < 1$. So the transistor is absolutely stable. When the transistor is conjugately

matched, the input and output reflection coefficients are given by (4.159): $\Gamma_s = 0.872\angle 123°, \Gamma_L = 0.876\angle 61°$. These values can be used to design the input and output matching networks. The maximum transducer gain may be determine from (4.156) as $\max G_T = 16.7\,\text{dB}$. □

3. Bilateral Design

If the transistor is bilateral (i.e., $S_{12} \neq 0$), we need to adopt the power gain approach to simplify the design procedure. The power gain (4.115) can be rewritten as

$$G_P = |S_{21}|^2 g_p, \tag{4.161}$$

where

$$g_p = \frac{\left(1 - |\Gamma_L|^2\right)}{\left(1 - |\Gamma_{\text{in}}|^2\right)|1 - \Gamma_L S_{22}|^2}$$

is the normalized gain. Introducing (4.104) into the above expression, we obtain

$$g_p = \frac{1 - |\Gamma_L|^2}{|1 - \Gamma_L S_{22}|^2 - |S_{11} - \Gamma_L \Delta|^2}.$$

This can be rearranged as

$$|\Gamma_L - C_p| = R_p \tag{4.162}$$

where

$$C_p = \frac{(\bar{S}_{22} - \bar{\Delta} S_{11}) g_p}{1 + \left(|S_{22}|^2 - |\Delta|^2\right) g_p}, \quad R_p = \frac{\sqrt{1 - 2K g_p |S_{12} S_{21}| + g_p^2 |S_{12} S_{21}|^2}}{|1 + \left(|S_{22}|^2 - |\Delta|^2\right) g_p|} \tag{4.163}$$

with K given by (4.125). Equation (4.162) represents a set of constant power gain circles. The boundary of Smith chart coincides with the $g_p = 0$ circle. When $R_p = 0$, we have

$$g_p = \frac{K \pm \sqrt{K^2 - 1}}{|S_{12} S_{21}|}. \tag{4.164}$$

For passive load (i.e., $|\Gamma_L| < 1$), the plus sign should be ignored. In this case, (4.164) gives the maximum gain previously obtained in (4.160).

In a similar way, the available power gain (4.116) can be rearranged as constant available power gain circles:

$$|\Gamma_s - C_a| = R_a \tag{4.165}$$

where

$$C_a = \frac{(\bar{S}_{11} - \bar{\Delta}S_{22})g_a}{1 + \big(|S_{11}|^2 - |\Delta|^2\big)g_a}, \quad R_a = \frac{\sqrt{1 - 2Kg_a|S_{12}S_{21}| + g_a^2|S_{12}S_{21}|^2}}{|1 + \big(|S_{11}|^2 - |\Delta|^2\big)g_a|}, \tag{4.166}$$

where $g_a = G_A/|S_{21}|^2$. Both the constant power gain circles (4.162) and the constant available power gain circles (4.165) can be used to design the amplifier. For convenience, the former may be applied for the situation where the input is required to be conjugately matched, and the latter for the situation where the output is required to be conjugately matched.

Example 4.14: The scattering parameters of a FET have the following values at 2.4 GHz:

$$S_{11} = 0.3\angle 30^\circ, \quad S_{12} = 0.2\angle -60^\circ, \quad S_{21} = 2.5\angle -80^\circ, \quad S_{22} = 0.2\angle -15^\circ.$$

The input of the amplifier is assumed to be conjugately matched. We wish to design an amplifier that provides a power gain of $G_p = 8$ dB at 2.4 GHz.

Solution: It is easy to verify that the transistor is absolutely stable. We use the constant power gain circle to design the amplifier. The normalized gain is

$$g_p = \frac{G_p}{|S_{21}|^2} = 1.0096.$$

The corresponding circle parameters are

$$C_p = 0.11\angle 69^\circ, \quad R_p = 0.35.$$

The constant power gain circle is plotted on Smith chart as shown in Figure 4.49. There are multiple choices for the reflection coefficient Γ_L. For simplicity in designing the output matching circuit, we choose Γ_L as the intersection point of the constant gain circle and the constant resistance

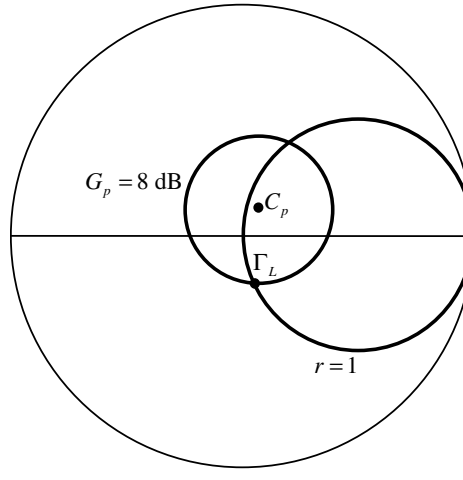


Figure 4.49 Constant power gain circle on Smith chart.

$r = 1$ circle, which is $\Gamma_L = 0.26\angle -75°$. The input reflection coefficient is then determined by (4.104), yielding $\Gamma_{\text{in}} = 0.277\angle 55.6°$. Thus the source reflection coefficient is $\Gamma_s = \bar{\Gamma}_{\text{in}} = 0.277\angle -55.6°$. □

4. Low Noise Amplifier Design

In a wireless system, the receiving antenna not only picks up useful signal energy but also a certain amount of noise-like radiation from other sources. The received signal, along with some noise is very weak and must be amplified to a level to be useful. The first amplifier stage must thus be designed for minimum noise.

Equation (4.145) indicates that there is an optimum source impedance or source reflection coefficient that will result in the lowest noise figure. Generally, it is impossible to achieve the lowest noise figure and maximum gain at the same time, and hence a compromise between them must be made, which is usually done by constant gain circles and constant noise figure circles.

Example 4.15: A transistor has the following parameters at 1 GHz:

$$S_{11} = 0.707\angle -155°, \quad S_{12} = 0, \quad S_{21} = 5.0\angle 180°, \quad S_{22} = 0.51\angle -20°,$$
$$F_{\min} = 3\,\text{dB}, \quad R_e = 4\,\Omega, \quad \Gamma_{\text{opt}} = 0.45\angle 180°.$$

Design an amplifier to have a power gain of 16 dB and a noise figure of less than 3.5 dB.

Solution: It is easy to verify that the transistor is absolutely stable. From (4.150), we obtain

$$\max G_s = 3\,\text{dB}, \quad G_0 = 13.98\,\text{dB}, \quad \max G_L = 1.31\,\text{dB}.$$

Thus the maximum transducer gain is

$$\max G_T = \max G_s + G_0 + \max G_L = 18.29\,\text{dB}.$$

So we have 2.29 dB more than is required by the specifications. Since $G_0 \approx 14\,\text{dB}$, we may choose $G_s = 1.22\,\text{dB}$ and $G_L = 0.78\,\text{dB}$ for an overall gain of 16 dB. The constant gain circle parameters for G_s and G_L are found from (4.154) as follows

$$C_s = 0.56\angle 155^\circ, \quad R_s = 0.35, \quad C_L = 0.47\angle 20^\circ, \quad R_L = 0.26.$$

The 3.5 dB noise figure circle parameters may be determined from (4.147)

$$C_{NF} = 0.37\angle 180^\circ, \quad R_{NF} = 0.40.$$

To meet the specifications, the input and output reflection coefficients may be respectively chosen as A and B as indicated in Figure 4.50. □

4.5.2 Oscillators

An **oscillator** converts direct current (DC) from a power supply to an alternating current signal. The oscillators can be classified as the feedback oscillator and negative resistance oscillator. The vacuum tube feedback

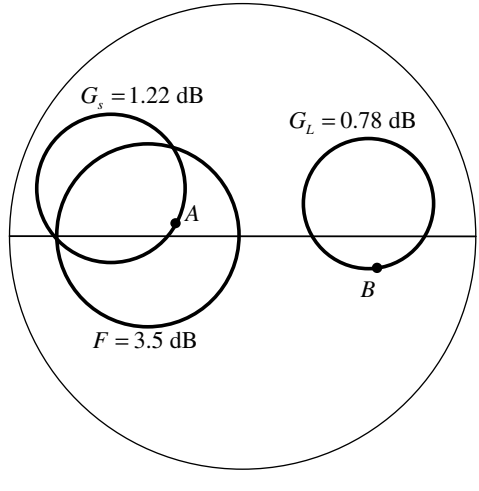


Figure 4.50 Constant noise figure circle and constant gain circle.

oscillator was invented around 1912 by a number of researchers including Edwin H. Armstrong (American electrical engineer, 1890–1954) Alexander Meissner (Austrian engineer and physicist, 1883–1958), Irving Langmuir (American chemist and physicist, 1881–1957), and Lee De Forest (1873–1961). The most common form of the feedback oscillator is a two-port amplifying active element (e.g., a transistor) connected in a feedback loop with its output fed back into its input through a frequency selective filter to boost the amplification. When the power supply to the amplifying active element is first switched on, noise in the circuit provides a signal to get oscillations started. The noise travels around the loop and is amplified and filtered until it becomes a sine wave at a single frequency.

Oscillators can also be built using one-port devices with negative resistance such as magnetron tubes, tunnel diodes and Gunn diodes. In negative resistance oscillators, a resonant circuit is connected across a device with negative resistance. A resonant circuit by itself is almost an oscillator as it can store energy in the form of oscillations if excited. However, the internal resistance in the resonant circuit will dissipate energy and cause the oscillations to decline to zero. To sustain the oscillation, negative resistance can be introduced to cancel the internal resistance in the resonant circuit, forming a resonator with no damping. Negative resistance oscillators are often used at high frequencies in the microwave range and above, since at these frequencies feedback oscillators perform poorly due to excessive phase shift in the feedback path.

The negative resistance oscillator model is not limited to one-port devices like diodes. The feedback oscillator circuits with two-port amplifying devices such as transistors and tubes also have negative resistance. At high frequencies, transistors and FETs do not need a feedback loop. With certain loads applied to one port, the transistors and FETs can become unstable at the other port and show negative resistance due to internal feedback. For this reason, high frequency oscillators in general are designed using negative resistance techniques.

4.5.2.1 *Feedback Oscillators*

A feedback oscillator is a simple linear feedback system as depicted in Figure 4.51, where $A(\omega)$ is the gain (or transfer function) of the amplifying active element, and $\beta(\omega)$ is the transfer function of the feedback path. In frequency domain, the output y of the system is related to the input x by

$$y = A(\omega)[x + \beta(\omega)y].$$

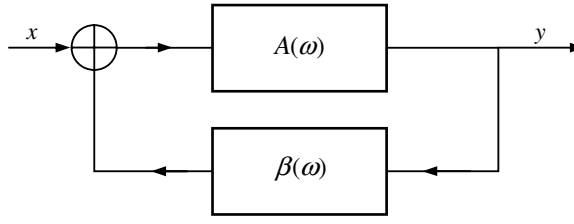


Figure 4.51 Feedback oscillator.

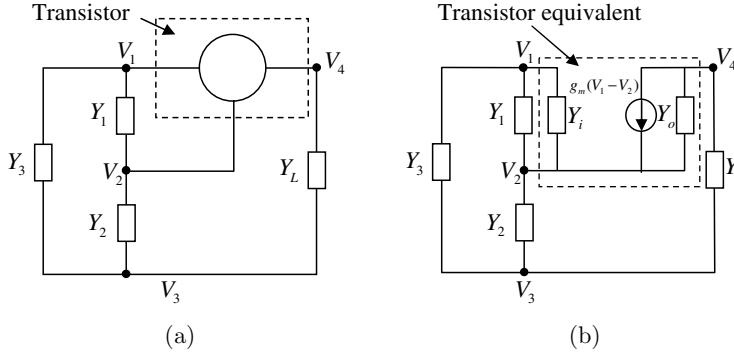


Figure 4.52 A oscillator and its equivalent circuit.

The transfer function of the system is thus given by

$$H(\omega) = \frac{y}{x} = \frac{A(\omega)}{1 - A(\omega)\beta(\omega)}. \tag{4.167}$$

The product $A(\omega)\beta(\omega)$ is referred to as **loop gain** around the feedback loop. When the condition $A(\omega)\beta(\omega) = 1$ is met, $H(\omega)$ becomes infinite. This implies that the system has a non-zero output y even if the input x is zero and thus the system oscillates. The condition $A(\omega)\beta(\omega) = 1$ is known as **Barkhausen stability criterion**, named after German physicist Heinrich G. Barkhausen (1881–1956). It is noted that Barkhausen's criterion is a necessary condition for oscillation but not a sufficient condition and it applies to linear circuits with a feedback loop and cannot be applied to one port negative resistance active elements like tunnel diode oscillators.

A general oscillator circuit using a transistor is shown in Figure 4.52(a). The small-signal equivalent circuit of the transistor is shown in Figure 4.52(b), where g_m is the transconductance; Y_i and Y_o are the input

and output admittances, respectively. From the circuit theory, the following algebraic equation for the node voltages can be obtained as follows:

$$\begin{bmatrix} (Y_1+Y_3+Y_i) & -(Y_1+Y_i) & -Y_3 & 0 \\ -(Y_1+Y_i+g_m) & (Y_1+Y_2+Y_i+Y_o+g_m) & -Y_2 & -Y_o \\ -Y_3 & -Y_2 & (Y_2+Y_3+Y_L) & -Y_L \\ g_m & -(Y_o+g_m) & -Y_L & (Y_o+Y_L) \end{bmatrix} \times \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} = 0. \tag{4.168}$$

The existence of a nontrivial solution to the equation requires that the determinant of the coefficient matrix vanishes. This condition determines the resonant frequency of the oscillator and places restrictions on the nature of the circuit components. The well-known Hartley and Colpitts oscillators are special cases where $Y_L = \infty$, as shown in Figure 4.53.

4.5.2.2 *Negative Resistance Oscillators*

A negative resistance oscillator is shown in Figure 4.54, where $Z_{\text{in}} = R_{\text{in}} + jX_{\text{in}}$ is the input impedance of an active device and $Z_L = R_L + jX_L$ is the load impedance. Apparently, we have

$$(Z_{\text{in}} + Z_L)I = 0.$$

The existence of a nontrivial solution of the above equation leads to

$$R_{\text{in}} + R_L = 0, \quad X_{\text{in}} + X_L = 0, \tag{4.169}$$

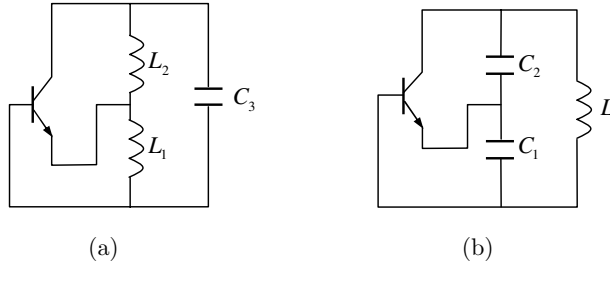


Figure 4.53 Typical oscillators. (a) Hartley oscillator. (b) Colpitts oscillator.

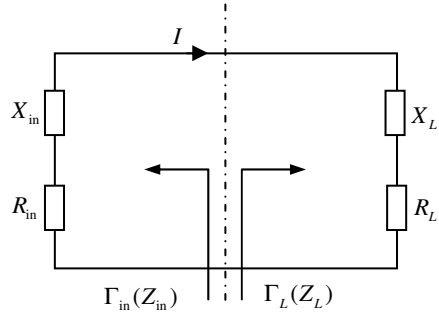


Figure 4.54 Negative resistance oscillator.

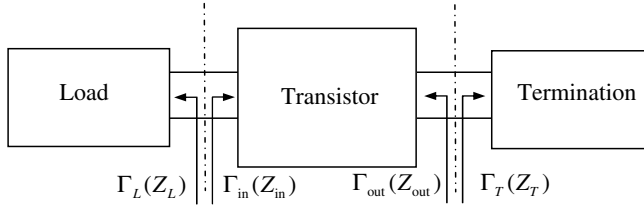


Figure 4.55 Two-port transistor oscillator.

or equivalently

$$\Gamma_{\text{in}} = \frac{Z_{\text{in}} - Z_0}{Z_{\text{in}} + Z_0} = \frac{Z_L + Z_0}{Z_L - Z_0} = \frac{1}{\Gamma_L}. \tag{4.170}$$

The first equation of (4.169) indicates that R_{in} must be negative for a passive load $R_L > 0$. This implies that the active device is an energy source. The second equation of (4.169) determines the frequency of oscillation.

Figure 4.55 shows a two-port transistor oscillator. Unlike the amplifier design, the transistor used in an oscillator must be unstable. The reflection coefficient Γ_T is selected to produce a large negative resistance at the input to the transistor. As the oscillator power builds up, R_{in} will become less negative. The load impedance Z_L must be chosen to match the input impedance Z_{in} so that $R_L + R_{\text{in}} < 0$. In practice, we may choose

$$R_L = -\frac{R_{\text{in}}}{3}, \quad X_L = -X_{\text{in}}.$$

It is noted that whenever $\Gamma_{\text{in}}\Gamma_L = 1$, we have $\Gamma_{\text{out}}\Gamma_T = 1$ and vice versa.

Example 4.16: The scattering parameters of a transistor at 4 GHz are given by ($Z_0 = 50\,\Omega$)

$$S_{11} = 2.18\angle -35^\circ, \quad S_{12} = 1.26\angle 18^\circ, \quad S_{21} = 2.75\angle 96^\circ, \quad S_{22} = 0.52\angle 155^\circ.$$

Design an oscillator at 4 GHz.

Solution: From (4.125), we may find that

$$K = 0.21, \quad \Delta = 2.34\angle -68.9.$$

Therefore, the transistor is potentially not stable and can be used for an oscillator design. From (4.122), the output stability circle parameters are found to be

$$C_T = 1.08\angle 33^\circ, \quad R_T = 0.665.$$

Since $|S_{11}| > 1$, the stable region is inside the output stability circle. The choice of Γ_T should make $|\Gamma_{\text{in}}|$ large. After a few trials we can select $\Gamma_T = 0.59\angle -104^\circ$ (see Figure 4.56). The input reflection coefficient Γ_{in} can be calculated from Γ_T as follows

$$\Gamma_{\text{in}} = S_{11} + \frac{S_{12}S_{21}\Gamma_T}{1 - S_{22}\Gamma_T} = 3.96\angle -2.4^\circ.$$

This gives

$$Z_{\text{in}} = Z_0 \frac{1 + \Gamma_{\text{in}}}{1 - \Gamma_{\text{in}}} = -84 - j1.9\,\Omega$$

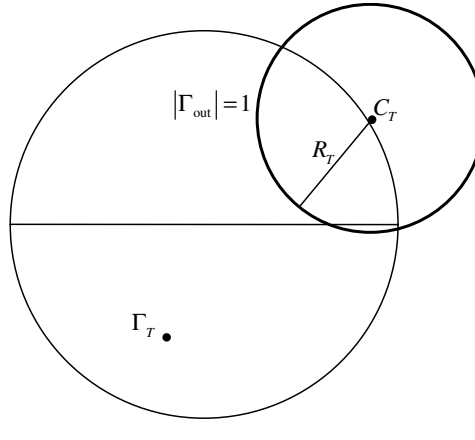


Figure 4.56 Output stability circle.

The load impedance can thus be chosen as

$$Z_L = -\frac{R_{\text{in}}}{3} - jX_{\text{in}} = 28 + j1.9\,\Omega. \quad \square$$

4.5.2.3 *Dielectric Resonator Oscillators*

The dielectric resonators have many advantages over the conventional metal cavity resonators, which include smaller size, excellent temperature stability, lower cost, higher Q, compact and ease of manufacturing and integration. A dielectric resonator oscillator is shown in Figure 4.57, where the dielectric resonator DR is placed in close proximity to a microstrip line. The magnitude of the reflection coefficient Γ_L is controlled by the coupling (or the spacing d) between the resonator and the microstrip line. The phase angle of Γ_L is controlled by the length l_r of the microstrip line. The output circuit is a standard stub-matched circuit used to transform the load impedance to the required value for the transistor. The series reactance jX in the common source lead is used as a feedback to make the equivalent transistor (i.e., the transistor and the series reactance combined) more unstable.

The dielectric resonator DR near a microstrip line is equivalent to a series impedance as indicated in Figure 4.58. The equivalent resonator impedance Z-coupled to the microstrip line is given by

$$Z = \frac{n^2 R}{1 + j2Q(\omega - \omega_0)/\omega_0}, \tag{4.171}$$

where $\omega_0 = 1/\sqrt{LC}$ is the resonant frequency, $Q = R/\omega_0 L$ is the unloaded resonator Q.

4.5.3 Mixers

The block diagram of a superheterodyne receiver is shown in Figure 4.59. A low level RF signal from the antenna is first amplified by a low noise

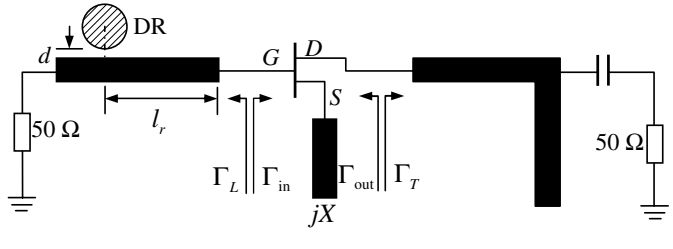


Figure 4.57 Dielectric resonator oscillator.

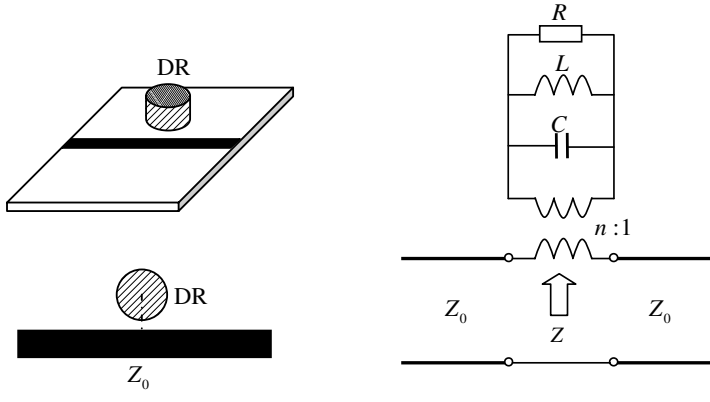


Figure 4.58 (a) Dielectric resonator coupled to microstrip line. (b) Equivalent circuit.

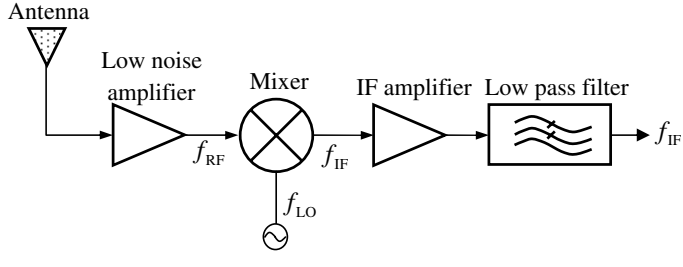


Figure 4.59 A superheterodyne receiver.

amplifier (optional), and then is mixed with a local oscillator (LO) signal to produce an intermediate frequency (IF), usually between 10 and 100 MHz. The mixer is a nonlinear device such as a diode or a transistor, and it will produce signals at intermediate frequency $f_{\mathrm{RF}} - f_{\mathrm{LO}}$ and other harmonic frequencies $mf_{\mathrm{RF}} \pm nf_{\mathrm{LO}}$, where m and n are integers.

4.5.3.1 *Characteristics of Diode*

The V–I characteristic of a typical diode is given by

$$i_d = I_s(e^{\alpha v_d} - 1) \tag{4.172}$$

where $\alpha = q/nkT$, q is the charge of electron, k is Boltzmann's constant, T is temperature in Kelvin, n is the ideality factor ranging between 1 and 2, and I_s is the saturation current typically between 10^{-6} and 10^{-15} A. At

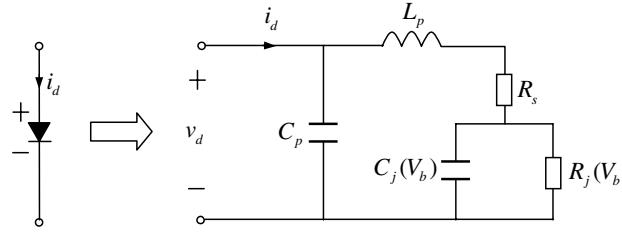


Figure 4.60 Equivalent circuit for the diode.

room temperature ($T = 290\,\mathrm{K}$), we have $\alpha \approx 1/25$ (mV). Assume that V_b is a DC bias voltage and v is a small AC signal. Letting $v_d = V_b + v$ in (4.172), we may have the following Taylor series expansion about V_b (called **small-signal approximation**)

$$i_d = I_b + g_d v + g_d' \frac{v^2}{2} \cdots , \tag{4.173}$$

where

$$I_b = I_s(e^{\alpha V_b} - 1), \quad g_d = \alpha I_s e^{\alpha V_b}, \quad g_d' = \alpha g_d.$$

The parameter g_d is called the **dynamic conductance** of the diode and its inverse is called **junction resistance** of the diode denoted by R_j. A small signal equivalent circuit for the diode is depicted in Figure 4.60, where C_j and R_j are junction capacitance and resistance respectively and both depend on bias voltage, L_p and C_p are the lead inductance and the packaging capacitance between the leads respectively, and the series resistance R_s is the bulk resistance.

4.5.3.2 *Mixer Designs*

We assume that the AC voltage v consists of a local-oscillator signal $v_l = V_l \cos\omega_l t$, an RF signal $v_r = V_r \cos\omega_r t$, and an IF signal $v_i = V_i \cos(\omega_i t + \varphi)$ at the IF frequency $\omega_i = \omega_r - \omega_l$. Then we have

$$\begin{aligned} i_d &= I_b + g_d(v_r + v_l - v_i) + \frac{g_d'}{2}(v_r + v_l - v_i)^2 + \cdots \\ &= I_b + g_d(v_r + v_l - v_i) + \frac{g_d'}{2}(v_r^2 + v_l^2 + v_i^2 + 2v_l v_r - 2v_r v_i - 2v_l v_i) + \cdots \end{aligned} \tag{4.174}$$

In practice, the amplitude of the local-oscillator signal is much larger than the RF and IF signals. For this reason, the higher order powers of the RF

and IF signals can be ignored. Retaining the terms that are linear in v_r and v_i gives

$$i_d = I_b + g_d(v_r + v_l - v_i) + \frac{g_d'}{2}(v_l^2 + 2v_l v_r - 2v_l v_i). \tag{4.175}$$

Note that

$$\begin{aligned} v_l^2 &= \frac{1}{2}V_l^2 + \frac{1}{2}v_l^2 \cos 2\omega_l t, \\ 2v_l v_r &= V_l V_r[\cos(\omega_r - \omega_l)t + \cos(\omega_l + \omega_r)t], \\ -2v_l v_i &= -V_l V_i\{\cos[(\omega_l - \omega_i)t - \varphi] + \cos[(\omega_l + \omega_i)t + \varphi]\}. \end{aligned}$$

The frequency components $\omega_r + \omega_l$ and $\omega_r - \omega_l$ are respectively called **upper sideband** and **lower sideband** and are used in a mixer for up conversion in a transmitter or down conversion in a heterodyne receiver. The total IF current is thus given by

$$i_i = -g_d v_i + \frac{g_d'}{2} V_l V_r \cos(\omega_r - \omega_l)t. \tag{4.176}$$

An important performance index is the **conversion loss** defined by

$$L_C = 10 \log \frac{\text{available input RF power}}{\text{IF output power}}. \tag{4.177}$$

Typical values for conversion loss for a single-diode mixer are 4 to 7 dB. Mixers may be classified by their topology. A single-diode mixer is shown in Figure 4.61. It is noted that there is no isolation between the RF and LO ports, and the LO signal may interfere with the reception of the RF signal or may be radiated out from the receiving antenna. The poor isolation between

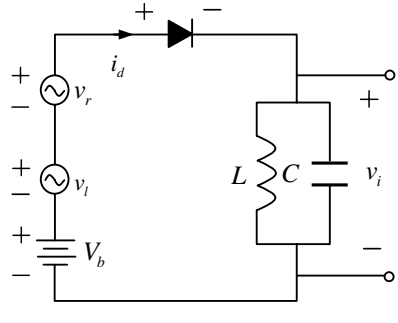


Figure 4.61 Single-diode mixer.

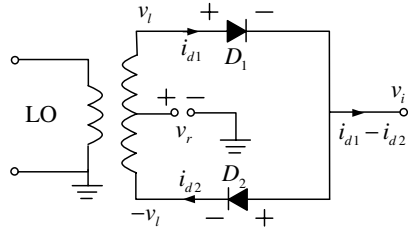


Figure 4.62 A single balanced mixer.

the IF port and the RF and LO port may lead to the noise interference and the generation of many spurious signals. These shortcomings of the single-diode mixer can be alleviated by a balanced mixer.

The basic circuit of a single balanced mixer is shown in Figure 4.62. The voltages across the diode D_1 and diode D_2 are $v_l + v_r - v_i$ and $v_l - v_r + v_i$, respectively. Thus we have

$$i_{d1} = I_b + g_d(v_r + v_l - v_i) + \frac{g_d'}{2}(v_r^2 + v_l^2 + v_i^2 + 2v_l v_r - 2v_r v_i - 2v_l v_i),$$

$$i_{d2} = I_b + g_d(v_l - v_r + v_i) + \frac{g_d'}{2}(v_r^2 + v_l^2 + v_i^2 - 2v_l v_r - 2v_r v_i + 2v_l v_i).$$

The input current into the IF low-pass filter is

$$i_{d1} - i_{d2} = 2g_d(v_r - v_i) + 2g_d'(v_l v_r - v_l v_i).$$

The IF current is given by

$$i_i = -2g_d v_i + g_d' V_l V_r \cos(\omega_r - \omega_l)t.$$

It can be seen from the above equations that there is no local-oscillator voltage at the IF port, which implies that the LO and IF ports are isolated. Also the LO and RF ports are isolated as can be seen from Figure 4.62.

Everything should be made as simple as possible, but not simpler.

—Albert Einstein

This page intentionally left blank

Chapter 5

Antennas

> It is through science that we prove, but through intuition that we discover.
>
> —Jules Henri Poincaré (French mathematician, 1854–1912)

Antennas are essential components in all wireless systems. Understanding the radiation by antennas and propagation of electromagnetic waves in space is important in radio frequency engineering. An **antenna** is a device which converts a guided wave into a radio wave in free space, and vice versa. In transmitting mode, a radio transmitter supplies RF power to the antenna terminals through a transmission line, and the antenna radiates the energy into space, forming a specific energy distribution pattern. In receiving mode, the antenna intercepts the power from an incident electromagnetic wave, generating a voltage at its terminal for further processing.

Typically an antenna consists of an arrangement of conductors (scatterers), which are connected to the receiver or transmitter through a transmission line or waveguide. In 1886, Hertz invented the first wire antennas (a dipole and a loop) to confirm Maxwell's theory and the existence of electromagnetic waves. In 1897, H. C. Pocklington (1870–1952) derived an integral equation for the current on a straight wire, which marked the start of antenna theory and analysis. Modern antenna theory was started during the World War II and a number of classical antennas were introduced during that time (Silver, 1949). The sources of radiation fields are the current distributions, including both conduction current and displacement current. The antenna can thus be classified as conduction-current type and displacement-current type. For the conduction-current antenna, the source of radiation is conduction current on a metallic radiator surface. Linear antenna, loop, helix and spiral antenna are of the conduction-current type, and they are typically for lower frequency,

lower gain, and wide beamwidth applications. For the displacement-current antenna, the source of the radiation is the electromagnetic fields at the antenna aperture or on the antenna surface. Horn antenna, slot antenna, aperture antenna, parabolic reflector, dielectric rod antenna belong to this type, and they are usually for higher frequency, higher gain, and narrow beamwidth applications. The antennas can also be categorized into four basic types: electrically small antenna, resonant antenna, broadband antenna, and aperture antenna. For the small antenna, its maximum extent is much less than a wavelength and it has low directivity, low radiation resistance, low radiation efficiency and high input reactance. Both the resonant antenna and broadband antenna have real input impedance but the bandwidth is narrow for the former and very wide for the latter. The aperture antenna has very high gain and moderate bandwidth. The radiation patterns of an antenna can be omni-directional or directional, depending on the antenna applications. Some important inventions in antenna history are listed in Table 5.1.

The most important parameters for characterizing antenna are gain, efficiency, input impedance, bandwidth, radiation pattern, beamwidth, sidelobes, front-to-back ratio, and polarization. There are trade-offs between these antenna parameters. To satisfy one parameter requirement, one may have to sacrifice one or more other parameter levels. Most of the antenna parameters are subject to certain limitations, which can be understood

Table 5.1 Typical antenna types and their inventors

Antenna types	Inventers
Dipole antenna and loop antenna	Invented by Hertz in 1886.
Yagi–Uda antenna (array)	Invented by Shintaro Uda and Hidetsugu Yagi in 1926.
Log-periodic antenna	Invented by R. H. DuHamel and Dwight E. Isbell in 1957.
Horn antennas	Invented by Jagadish Chandra Bose (1858–1937) in 1897; first experimental research by Gorge Clark Southworth (1890–1972) and Wilmer Lanier Barrow (1903–1975) in 1936; theoretical analysis by Barrow and Lan Jen Chu (1913–1973) in 1939; corrugated horn invented by A. F. Kay in 1962.
Parabolic reflector antennas	Invented by Hertz in 1888; Cassegrain antenna was developed in Japan in 1963 by NTT, KDDI and Mitsubishi Electric.
Microstrip antennas	Invented by Robert E. Munson in 1972.

by spherical wavefunction expansion of the fields produced by antenna. The propagating modes supported by an antenna depend on the size of the smallest circumscribing sphere enclosing the antenna. The bigger the antenna size (the size of the sphere), the more propagating modes the antenna will generate. When the antenna is very small, no propagating modes can exist and all the spherical modes are rapidly cut-off. As a result, the stored energy around the antenna becomes very large and the radiation power becomes very small, and the antenna has a high quality factor.

A more useful performance index for describing antenna is the product of antenna gain and bandwidth for they must be maximized simultaneously in most applications. It can be shown that antenna fractional bandwidth is reciprocal to antenna quality factor. Thus, the product of antenna gain and bandwidth can be expressed as the ratio of antenna gain over antenna quality factor. The maximum possible product of antenna gain and bandwidth is an upper bound of the antenna performance, which can be used to determine the antenna size required to achieve a specified antenna performance.

5.1 From Transmission Lines to Antennas

Many practical antennas are fed by a waveguide. For this reason, most antennas may be constructed as a transition structure or an interface that serves as a bridge between the waveguide and free space. It is noted that a simple open-ended waveguide (such as an open transmission line) cannot be used as an antenna. At the open end, the impedance suddenly changes from the wave impedance in the waveguide to the impedance of free space, and a significant portion of energy may be reflected back to the source to form a standing wave in the waveguide due to the mismatch of impedance. Moreover, the diameter of the waveguide cross section is normally less than one wavelength, making it difficult to get a directive pattern as a result of diffraction.

To avoid the abrupt change of the impedance at the open end, the ends of the waveguide may be flared out to form a taper transition structure (such as a horn), along which the impedance changes gradually. The transition structure would minimize the reflected energy and guide the energy into free space. In order to achieve a specific energy distribution pattern in free space, the transition structure must be properly designed to form an antenna.

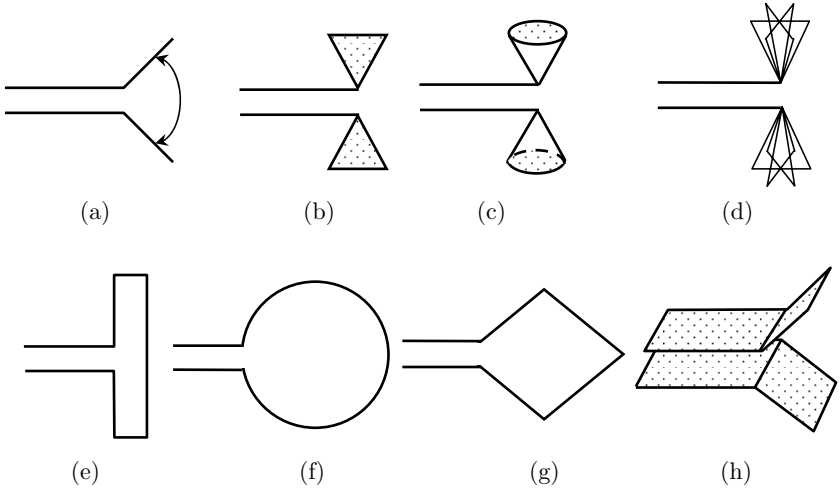


Figure 5.1 Antennas transformed from two-wire lines. (a) Flared dipole. (b) Bow–tie. (c) Biconical. (d) Double wire cones. (e) Folded dipole. (f) Circular loop. (g) Rhombic loop. (h) Planar horn.

5.1.1 Antennas Transformed from Two-Wire Lines

A two-wire transmission line may be flared out to form a dipole antenna; the two arms of the dipole can be transformed to form a bow–tie antenna, and then a solid biconical antenna; the metal surface of the biconical may be simulated by metal wires to form a double wire cones; the two ends of the dipole antenna can be joined together to form a folded dipole, a circular loop, and a rhombic loop etc.; the two wires may be altered to two plates to form a planar horn antenna. All these transformations have been illustrated in Figure 5.1.

5.1.2 Antennas Transformed from Coaxial Cables

The outer conductor of a coaxial cable may be flared out to form a monopole antenna; the center conductor of the monopole may be deformed into a helix or other bended wire structures; the center conductor of the monopole can also be transformed to form a triangular sheet, and then a cone; the end of the center conductor of the monopole may be loaded with different wire structures or solid metal plates to form so-called top loaded antenna; the center conductor of the monopole can be bended to form a low profile L-shaped antenna. These transformation processes are illustrated in Figure 5.2.

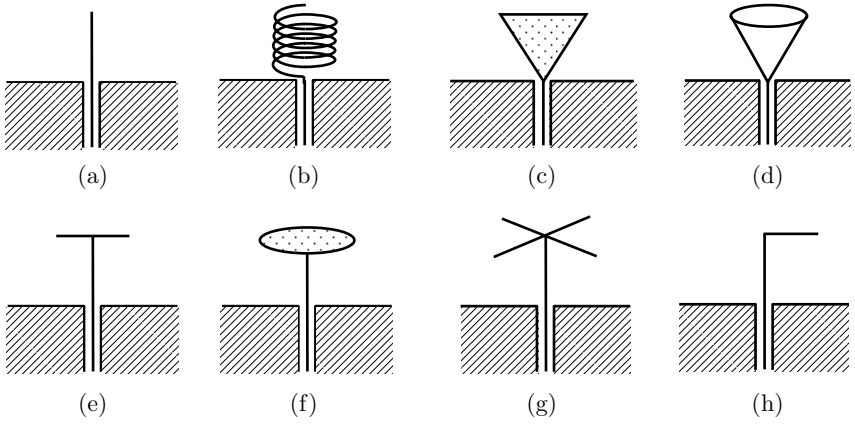


Figure 5.2 Antennas transformed from coaxial cables. (a) Monopole. (b) Helix. (c) Triangular sheet. (d) Unicone. (e)–(h) Top loaded monopoles.

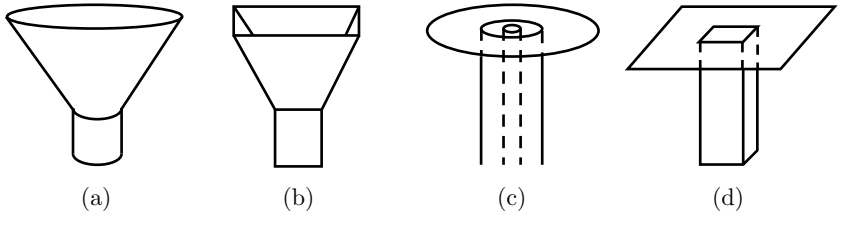


Figure 5.3 Antennas transformed from hollow waveguides. (a) Circular horn. (b) Rectangular horn. (c) Coaxial aperture. (d) Rectangular aperture.

5.1.3 Antennas Transformed from Waveguides

As shown in Figure 5.3, a circular horn and a rectangular horn can be formed by opened-out circular waveguide and rectangular waveguide respectively; the outer conductor of a coaxial waveguide can be flared out to form a coaxial aperture antenna; the rectangular waveguide can be flared out to form a rectangular aperture antenna.

5.2 Antenna Parameters

An arbitrary transmitting antenna system and a receiving antenna system are shown in Figure 5.4. The power incident to the matching network is denoted by P_m; the power accepted by the antenna is denoted by P_a; and the power radiated by the antenna is denoted by P_{rad}. Due to the mismatch, portion of the power P_m is reflected back to the source by the matching

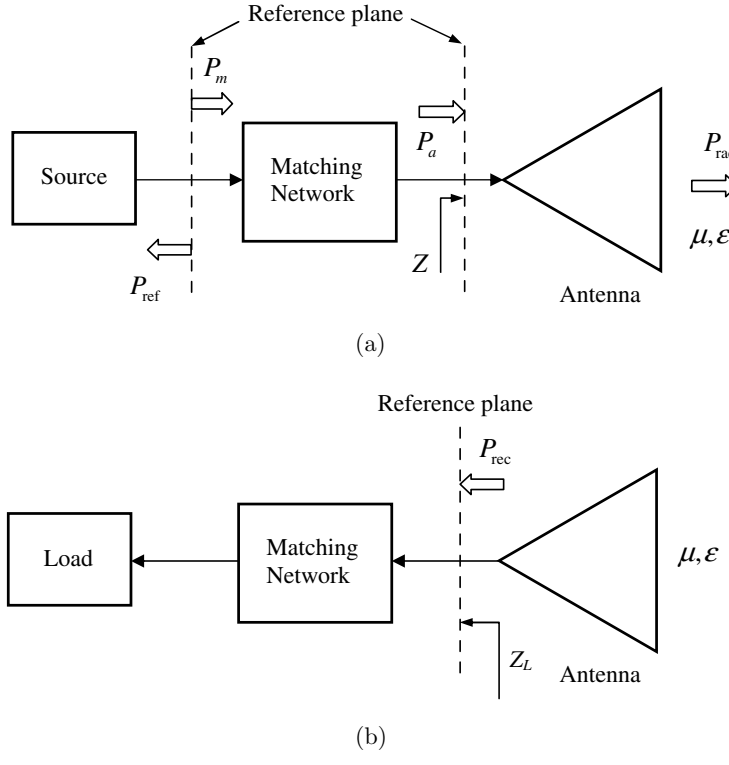


Figure 5.4 (a) Transmitting antenna. (b) Receiving antenna.

network, which is denoted by P_{ref}. The power accepted by the antenna can be expressed as

$$P_a = \frac{1}{2}\text{Re}\, V\overline{I} = P_m - P_{\text{ref}} - P_{\text{loss}}^{\text{match}},$$

where V and I are the modal voltage and modal current for the dominant mode in the feeding waveguide respectively and they are calculated at the reference plane, and $P_{\text{loss}}^{\text{match}}$ is the power loss in the matching network. Let $\mathbf{E}$ and $\mathbf{H}$ respectively denote the electric field and magnetic field generated by the antenna. The radiated power of the antenna can be represented by

$$P_{\text{rad}} = \frac{1}{2}\int_S \text{Re}\big(\mathbf{E} \times \bar{\mathbf{H}}\big) \cdot \mathbf{u}_n \, dS, \tag{5.1}$$

where S is an arbitrary surface enclosing the antenna.

Antenna performances depend on the antenna geometry as well as how the antenna is used. In mobile devices, the antenna position keeps changing

as the subscriber travels around, and reasonable antenna performances are expected for all different positions. The antenna design is thus based on those positions that are used most often.

5.2.1 Field Regions

The space around an electromagnetic radiator can be divided into **reactive near-field region, radiating near-field region** and **far-field region**. The reactive near-field and the far-field region are respectively defined by $r < R_1 = 0.62\sqrt{D^3/\lambda}$ and $r > R_2 = 2D^2/\lambda$. Here r is the distance from the radiator, D is the largest dimension of the radiator, and λ is the wavelength. The radiating near-field region is defined by $R_1 < r < R_2$, as illustrated in Figure 5.5.

Antennas are used for wireless communication and they are usually located in the far-field region of each other. Many antenna parameters are also determined in the far-field region. For this reason, the far-field region is the most important for most applications. In this region, the radiation pattern (the angular field distribution) does not change shape with distance; the electric field **E** and magnetic field **H** are orthogonal to each other and are in phase, and they both fall off with distance as $1/r$; the Poynting vector only has a radial component. The far-field region is also called **Fraunhofer region**, named after Joseph von Fraunhofer (German optician, 1787–1826). In the reactive near-field region, the relationship

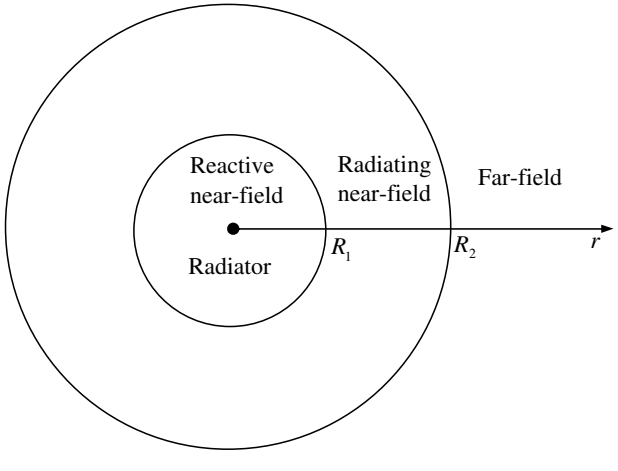


Figure 5.5 Field regions of radiator.

between the electric field **E** and magnetic field **H** is very complicated and the fields change rapidly with the distance. In this region, the Poynting vector contains both radial component and transverse components. The radial component represents the radiating power (the far-field) and the transverse components are reactive. The radiating near-field region is also called **Fresnel region**, named after French physicist Augustin-Jean Fresnel (1788–1827). It is a transition region where the reactive field becomes smaller than the radiating field.

5.2.2 Radiation Patterns and Radiation Intensity

The **radiation pattern** of antenna is a mathematical function or a graphical representation of the radiation properties of the antenna as a function of space coordinates. In most cases, the radiation pattern is determined in the far-field region. Radiation properties can be power flux density, radiation density, field strength, phase or polarization. For a linearly polarized antenna, the radiation pattern is usually described by E-plane and H-plane patterns. The **E-plane** is defined as the plane containing the electric field vector and the direction of the maximum radiation and the **H-plane** is defined as the plane containing the magnetic field vector and the direction of maximum radiation. The antenna radiation pattern magnitude must be plotted relative to a recognized standard. The most common standard level is that of a perfect isotropic radiator (antenna), which would radiate energy equally in all directions.

Let $\mathbf{u}_r$ be the unit vector along a far-field observation point $\mathbf{r} = r\mathbf{u}_r$. The **radiation intensity** of an antenna in the direction $\mathbf{u}_r$ is defined as the power radiated from the antenna per unit solid angle

$$U(\mathbf{u}_r) = \frac{r^2}{2}\mathrm{Re}\big[\mathbf{E}(\mathbf{r}) \times \bar{\mathbf{H}}(\mathbf{r})\big] = \frac{r^2}{2\eta}|\mathbf{E}(\mathbf{r})|^2, \tag{5.2}$$

where $\eta = \sqrt{\mu/\varepsilon}$ is the intrinsic impedance of the medium. The radiation intensity for an isotropic radiator is $U(\mathbf{u}_r) = P_{\mathrm{rad}}/4\pi$.

A typical radiation pattern is shown in Figure 5.6. A radiation pattern can be divided into various parts, called lobes. A **major lobe** refers to the radiation lobe which contains the direction of maximum radiation. All other lobes are called **minor lobes**. A **side lobe** refers to the minor lobe adjacent to the major lobe. A **back lobe** is a minor lobe which directs energy toward the direction opposite to the major lobe. The **half power beam width** (HPBW) is the angle between the half-power (−3 dB)

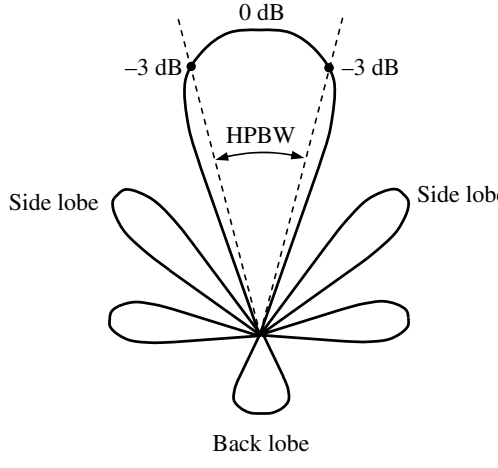


Figure 5.6 Radiation pattern.

points of the main lobe, when referenced to the peak radiated power of the main lobe.

5.2.3 Radiation Efficiency, Antenna Efficiency and Matching Network Efficiency

Not all the incident power to the antenna will be radiated to the free space. The power loss may come from the impedance mismatch that causes portion of the incident power reflected back to the transmitter, or from the imperfect conductors and dielectrics that cause portion of the incident power to be dissipated as heat. The **radiation efficiency** of the antenna is defined by

$$e_r = \frac{P_{\text{rad}}}{P_a}. \tag{5.3}$$

The radiation efficiency reflects the conduction and dielectric losses of the antenna. The **antenna efficiency** is defined by

$$e_t = \frac{P_{\text{rad}}}{P_m} = \frac{P_m - P_{\text{ref}}}{P_m} \cdot \frac{P_a}{P_m - P_{\text{ref}}} \cdot \frac{P_{\text{rad}}}{P_a} = (1 - |\Gamma|^2) e_s e_r, \tag{5.4}$$

where $e_s = P_a/(P_m - P_{\text{ref}})$ is the efficiency describing the loss in the matching network; Γ is the reflection coefficient at the input of the matching network; and

$$e_m = \frac{P_a}{P_m} = (1 - |\Gamma|^2) e_s \tag{5.5}$$

is the **matching network efficiency**. Better antenna efficiency means

(1) Better quality of communication.
(2) Better wireless coverage.
(3) Longer battery life for wireless terminals.

5.2.4 Directivity and Gain

The **directivity** of an antenna is defined as the ratio of the radiation intensity in a given direction from an antenna to the radiation intensity averaged over all directions

$$D(\mathbf{u}_r) = \frac{U(\mathbf{u}_r)}{P_{\text{rad}}/4\pi}. \tag{5.6}$$

Theoretically, there is no mathematical limit to the directivity that can be obtained from currents confined in an arbitrarily small volume. However, the high field intensities around a small antenna with a high directivity will produce high energy storage around the antenna, large power dissipation, low radiation efficiency and narrow bandwidth. For antennas with two orthogonal polarization components, we may introduce the partial directivity of an antenna for a given polarization component. In a spherical coordinate system, we may write $\mathbf{E}(\mathbf{r}) = E_\theta(\mathbf{r})\mathbf{u}_\theta + E_\varphi(\mathbf{r})\mathbf{u}_\varphi$, and

$$U(\mathbf{u}_r) = U_\theta(\mathbf{u}_r) + U_\varphi(\mathbf{u}_r),$$

where

$$U_\theta(\mathbf{u}_r) = \frac{r^2}{2\eta}|E_\theta(\mathbf{r})|^2, \quad U_\varphi(\mathbf{u}_r) = \frac{r^2}{2\eta}|E_\varphi(\mathbf{r})|^2.$$

The directivity can be written as

$$D(\mathbf{u}_r) = D_\theta(\mathbf{u}_r) + D_\varphi(\mathbf{u}_r), \tag{5.7}$$

where

$$D_\theta(\mathbf{u}_r) = \frac{U_\theta(\mathbf{u}_r)}{P_{\text{rad}}/4\pi}, \quad D_\varphi(\mathbf{u}_r) = \frac{U_\varphi(\mathbf{u}_r)}{P_{\text{rad}}/4\pi}$$

are the partial directivities for θ and φ component, respectively.

The **absolute gain** of an antenna is defined as the ratio of the radiation intensity in a given direction to the radiation intensity that would be obtained if the power accepted by the antenna were radiated isotropically.

$$G(\mathbf{u}_r) = \frac{U(\mathbf{u}_r)}{P_a/4\pi} = e_r D(\mathbf{u}_r). \tag{5.8}$$

The old definition of the gain is

$$G_{\text{old}}(\mathbf{u}_r) = \frac{U(\mathbf{u}_r)}{P_m/4\pi} = e_t D(\mathbf{u}_r). \tag{5.9}$$

This is also called **absolute gain**, which has included the effects of matching network. The gain of an antenna often refers to the maximum gain and is usually given in decibels. Similarly, we may introduce partial gains in a spherical coordinate system

$$G(\mathbf{u}_r) = G_\theta(\mathbf{u}_r) + G_\varphi(\mathbf{u}_r). \tag{5.10}$$

5.2.5 Input Impedance, Bandwidth and Antenna Quality Factor

The **input impedance** of antenna is defined as the ratio of the voltage to current at the input reference plane of the antenna. The **bandwidth** of an antenna is defined as the range of frequencies within which the performance of the antenna, with respect to some characteristics (such as the input impedance, return loss, gain, radiation efficiency, pattern, beamwidth, polarization, sidelobe level, and beam direction), conforms to a specified standard. Antenna bandwidth is an important quantity, which measures the quality of signal transmission such as signal distortion. For broadband antennas, the bandwidth is usually expressed as the ratio of the upper-to-lower frequencies of acceptable operation. For narrow band antennas, the bandwidth is expressed as a percentage of the frequency difference (upper minus lower) over the center frequency of the bandwidth (fractional bandwidth). The bandwidth can be enhanced by introducing losses, parasitic elements, loading or changing matching network.

If the impedance of antenna is not perfectly matched to that of the source, some power will be reflected back and not transmitted. This reflected power relative to incident power is called **return loss**. A figure of merit for antenna is **return loss bandwidth**, which is defined as the frequency range where return loss is below an acceptable level. For example, a return loss of -10 dB indicates 90% of the power is transmitted. At -7 dB return loss, 80% of the power is transmitted. The return loss bandwidth is closely related to antenna physical volume. Increasing the return loss bandwidth is one of the challenges in small antenna design.

According to the IEEE Standard Definitions of Terms for Antennas, the **quality factor** of a resonant antenna is defined as the ratio of 2π times energy stored in the fields excited by the antenna to the energy radiated

per cycle:

$$Q = \frac{\omega \widetilde{W}}{P_{\text{rad}}} = \frac{\omega\left(\widetilde{W}_e + \widetilde{W}_m\right)}{P_{\text{rad}}}, \tag{5.11}$$

where $\widetilde{W} = \widetilde{W}_e + \widetilde{W}_m$; $\widetilde{W}_e$ stands for the stored electric field energy and $\widetilde{W}_m$ for the stored magnetic field energy; ω is the frequency and P_{rad} is the total radiated power. In most publications, antenna Q is traditionally defined by

$$Q = \begin{cases} \dfrac{2\omega \widetilde{W}_m}{P_{\text{rad}}}, & \widetilde{W}_m > \widetilde{W}_e, \\ \dfrac{2\omega \widetilde{W}_e}{P_{\text{rad}}}, & \widetilde{W}_e > \widetilde{W}_m, \end{cases} \tag{5.12}$$

which has a conditional statement and is more difficult to handle than (5.11) in theoretical study. The quality factor defined by (5.12) applies to an antenna tuned to resonance only, while the IEEE standard definition (5.11) applies to an antenna under any conditions, at resonance or above resonance. Both definitions give the exact same values when the antenna is tuned to resonance. For this reason, the IEEE standard definition (5.11) may be considered as a generalization of the usual definition (5.12).

It will be shown later that, for a high quality factor antenna, the quality factor is reciprocal of antenna fractional bandwidth for input impedance. The antenna quality factor is a field quantity and is more convenient for theoretical study while the antenna bandwidth requires more information on the frequency behavior of the input impedance. We will use Q_{real} to indicate that all the stored energy around antenna has been included in the calculation of antenna quality factor, to distinguish it from another antenna quality factor, denoted by Q, to be introduced later, in which only the stored energy outside the circumscribing sphere of the antenna is included. Obviously, we have $Q_{\text{real}} \gg Q$.

5.2.6 Vector Effective Length, Equivalent Area and Antenna Factor

Let $\mathbf{r} = r\mathbf{u}_r$ be a far-field observation point. The electric field of the antenna in a homogeneous and isotropic medium can be expressed as [see (5.34)]

$$\mathbf{E}(\mathbf{r}) = -\frac{j\omega\mu I}{4\pi r} e^{-jkr} \mathbf{L}(\mathbf{u}_r). \tag{5.13}$$

Here I is the exciting current at the feeding plane, and $\mathbf{L}$ is called the **antenna vector effective length** defined by

$$\mathbf{L}(\mathbf{u}_r) = \frac{1}{I}\int_{V_0} \left\{\mathbf{J}(\mathbf{r}') - \left[\mathbf{J}(\mathbf{r}')\cdot\mathbf{u}_r\right]\mathbf{u}_r\right\} e^{jk\mathbf{u}_r\cdot\mathbf{r}'}\, dV(\mathbf{r}'), \tag{5.14}$$

where V_0 is the source region and $\mathbf{J}$ is the current distribution in the source region. The open circuit voltage at the antenna-feeding plane induced by an incident field $\mathbf{E}_{\text{in}}$ from the direction $-\mathbf{u}_r$ is given by (e.g., Geyi, 2010)

$$V_{\text{oc}}(\mathbf{u}_r) = -\frac{1}{I}\int_{V_0} \mathbf{E}_{\text{in}}(\mathbf{r}')\cdot\mathbf{J}(\mathbf{r}')dV(\mathbf{r}'), \tag{5.15}$$

which results from the reciprocity of transmitting and receiving antenna. The incident field may be written as

$$\mathbf{E}_{\text{in}}(\mathbf{r}) = \mathbf{E}_{\text{in}}(o)e^{jk\mathbf{u}_r\cdot\mathbf{r}},$$

where $\mathbf{E}_{\text{in}}(o)$ is the field strength at the origin (antenna position) and is perpendicular to $\mathbf{u}_r$. Thus

$$V_{\text{oc}}(\mathbf{u}_r) = -\frac{1}{I}\mathbf{E}_{\text{in}}(o)\cdot\int_{V_0}\mathbf{J}(\mathbf{r}')e^{jk\mathbf{u}_r\cdot\mathbf{r}'}dV(\mathbf{r}') = -\mathbf{E}_{\text{in}}(o)\cdot\mathbf{L}(\mathbf{u}_r).$$

This relation has been used as the definition of the vector effective length in most literature. According to the equivalent circuit for the receiving antenna as shown in Figure 5.7, the received power by the load is

$$P_{\text{rec}}(\mathbf{u}_r) = \frac{1}{2}\left|\frac{V_{\text{oc}}(\mathbf{u}_r)}{Z+Z_L}\right|^2 \text{Re}\, Z_L = \frac{1}{2}\left|\frac{\mathbf{E}_{\text{in}}(o)\cdot\mathbf{L}(\mathbf{u}_r)}{Z+Z_L}\right|^2 \text{Re}\, Z_L,$$

where Z is the antenna input impedance.

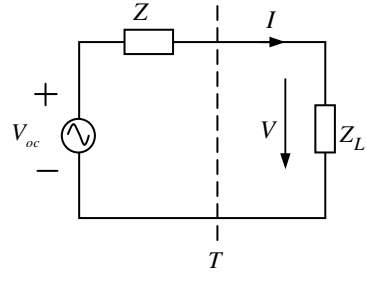


Figure 5.7 Equivalent circuit for receiving antenna.

The **antenna equivalent area** is a transverse area defined as the ratio of received power to the power flux density of the incident plane wave

$$A_e(\mathbf{u}_r) = \frac{P_{\text{rec}}(\mathbf{u}_r)}{|\mathbf{E}_{\text{in}}(o)|^2/2\eta} = \left|\frac{\mathbf{E}_{\text{in}}(o)\cdot\mathbf{L}(\mathbf{u}_r)}{Z+Z_L}\right|^2 \frac{\eta \operatorname{Re} Z_L}{|\mathbf{E}_{\text{in}}(o)|^2}. \tag{5.16}$$

If the receiving antenna is conjugately matched and there is no polarization loss, the antenna equivalent area can be simplified as

$$A_e(\mathbf{u}_r) = \frac{1}{4}\left(\frac{\eta}{\operatorname{Re} Z_L}\right)|\mathbf{L}(\mathbf{u}_r)|^2.$$

The **antenna factor** is defined as the ratio of incident electric field strength to the induced terminal voltage

$$AF(\mathbf{u}_r) = \frac{|\mathbf{E}_{\text{in}}(o)|}{|V(\mathbf{u}_r)|},$$

where $V(\mathbf{u}_r)$ stands for the induced terminal voltage at the reference plane of the receiving antenna due to the incident field. From the equivalent circuit of a receiving antenna, we obtain

$$V(\mathbf{u}_r) = \frac{Z_L}{Z_L+Z}V_{\text{oc}}(\mathbf{u}_r).$$

So the relationship between the antenna factor and vector effective length is

$$AF(\mathbf{u}_r) = \frac{|\mathbf{E}_{\text{in}}(o)|}{|\mathbf{E}_{\text{in}}(o)\cdot\mathbf{L}(\mathbf{u}_r)|}\left|1+\frac{Z}{Z_L}\right|. \tag{5.17}$$

If there is no polarization loss, this reduces to

$$AF(\mathbf{u}_r) = \left|1+\frac{Z}{Z_L}\right|\frac{1}{|\mathbf{L}(\mathbf{u}_r)|}.$$

Let S_∞ be a large closed surface, which encloses the antenna. The transmitting and receiving properties of antenna can be expressed as functions of the effective length and they are summarized in Table 5.2.

5.2.7 Polarization and Coupling

The **polarization of wave** is defined as the curve traced by the instantaneous electric field in a plane perpendicular to the propagation direction of the wave. If the direction of electric field at a point of space is constant in time, we say the electric filed at that point is **linearly polarized**. If the tip

Table 5.2 Transmitting properties of antenna

Quantity	Expression
Poynting vector	$S(\mathbf{r}) = \dfrac{1}{2\eta}\lvert\mathbf{E}(\mathbf{r})\rvert^2 = \dfrac{\eta\lvert I\rvert^2}{8r^2}\left\lvert\dfrac{\mathbf{L}(\mathbf{u}_r)}{\lambda}\right\rvert^2.$
Radiation intensity	$U(\mathbf{u}_r) = \dfrac{r^2}{2\eta}\lvert\mathbf{E}(\mathbf{r})\rvert^2 = \dfrac{\eta\lvert I\rvert^2}{8}\left\lvert\dfrac{\mathbf{L}(\mathbf{u}_r)}{\lambda}\right\rvert^2.$
Radiated power	$P_{\text{rad}} = \dfrac{\eta\lvert I\rvert^2}{8}\displaystyle\int_{S_\infty}\left\lvert\dfrac{\mathbf{L}(\mathbf{u}_r)}{\lambda}\right\rvert^2 d\Omega.$
Radiation resistance	$R_{\text{rad}} = \dfrac{2P_{\text{rad}}}{\lvert I\rvert^2} = \dfrac{\eta}{4}\displaystyle\int_{S_\infty}\left\lvert\dfrac{\mathbf{L}(\mathbf{u}_r)}{\lambda}\right\rvert^2 d\Omega.$
Directivity	$D(\mathbf{u}_r) = \dfrac{4\pi U(\mathbf{u}_r)}{P_{\text{rad}}} = \dfrac{\pi\eta}{R_{\text{rad}}}\left\lvert\dfrac{\mathbf{L}(\mathbf{u}_r)}{\lambda}\right\rvert^2.$
Gain	$G(\mathbf{u}_r) = e_r D(\mathbf{u}_r) = \dfrac{\pi\eta}{R_{\text{rad}} + R_{\text{loss}}}\left\lvert\dfrac{\mathbf{L}(\mathbf{u}_r)}{\lambda}\right\rvert^2.$
Equivalent area	$A_e(\mathbf{u}_r) = \left\lvert\dfrac{\mathbf{E}_{\text{in}}(o)\cdot\mathbf{L}(\mathbf{u}_r)}{Z + Z_L}\right\rvert^2 \dfrac{\eta\,\text{Re}\,Z_L}{\lvert\mathbf{E}_{\text{in}}(o)\rvert^2}.$
Antenna factor	$AF(\mathbf{u}_r) = \dfrac{\lvert\mathbf{E}_{\text{in}}(o)\rvert}{\lvert\mathbf{E}_{\text{in}}(o)\cdot\mathbf{L}(\mathbf{u}_r)\rvert}\left\lvert 1 + \dfrac{Z}{Z_L}\right\rvert.$

of the electric field is a circle (or ellipse) centered at the point in the course of time we say the electric field is **circularly** (**elliptically**) **polarized** at that point. Elliptically polarized field is encountered in practice very often. The **polarization of antenna** is defined as the curve traced by instantaneous electric field radiated by the antenna in a plane perpendicular to the radial direction. The radiation fields of all antennas aside from the dipoles are elliptically polarized in general, except in some preferred directions.

For perfect transmission of power between two antennas, their polarizations must match exactly. In practice the polarization mismatch loss always exists. If two antennas have no coupling, their polarizations are said to be orthogonal. The polarization mismatch loss between a circularly polarized and linearly polarized antenna is 3 dB and half power is lost. Two linear polarized antennas orientated at an angle of 45° will also have 3 dB polarization mismatch loss.

In a cellular environment, the degree of polarization match between the mobile and base station can vary considerably. In outdoor suburban environments, the polarization of the incident field would be mainly vertical while in indoors and in dense urban environments, scattering and multipath reflections can cause the incident polarization to change dramatically. Additionally, the degree of polarization match between the incident field and mobile antenna is impacted by the user. For example, how the device is held and placed changes the degree of polarization match.

Minimizing the antenna coupling is important if isolation between two signal paths is required. The coupling between two antennas can be measured using network analyzer. The coupling is the amplitude of S_{21} over the frequency range of both bands when the network analyzer is connected to these two antennas. The coupling between two antennas in the far-field is inversely related to the square of the distance between them (assume both antennas are in free space). For example, if the distance is doubled, the coupling is reduced by a factor of four (-6 dB). Antenna coupling is strongly influenced by the out-of-band impedance of the antenna. For example, if one antenna is very poorly matched at the band of another antenna and vice versa, the coupling between the two antennas might be low even if they are placed in close proximity. Unbalanced antennas are fed against the ground, and they make the ground part of the antenna. Two unbalanced antennas fed against the same ground tend to have less isolation than the balanced antennas. This problem can be improved if the antenna design can make the currents in the ground to be localized in the vicinity of the antenna.

5.2.8 Specific Absorption Rate

A variety of devices have been designed to decrease the exposure of the users to the RF energy from the antenna. **Specific Absorption Rate** (SAR) is a quantity to measure the rate at which electromagnetic energy is absorbed by lossy dielectric media with nonmagnetic dissipative properties, and is related to antenna desgin. It is defined as the power absorbed per mass of the dielectric media

$$\mathrm{SAR}(\mathbf{r}) = \frac{1}{2}\frac{\sigma(\mathbf{r})|\mathbf{E}(\mathbf{r})|^2}{\rho(\mathbf{r})} \ (\mathrm{W/kg}) \tag{5.18}$$

where $\sigma(\mathbf{r})$ is the conductivity, $\rho(\mathbf{r})$ is the mass density.

Most regions are adopting the International Commission on Non-Ionizing Radiation Protection guidelines (e.g., Europe, Japan, Korea, etc.), which define the basic limit for local exposure to be 2 W/kg averaged over a

volume of 10 g and a period of 6 min. Federal Communications Commission (FCC) has adopted a slightly stricter limit of the ANSI/IEEE standard for the uncontrolled environment, which is 1.6 W/kg averaged over a volume of 1 g and a period of 30 min.

Meeting SAR limit regulations can be a difficult challenge for some radio devices when the transmitting power is high. SAR considerations can impact how the antenna is packaged in the device. There are two key factors that affect the system SAR level. One is the current distribution of the antenna and another is the distance between the antenna and the testing probe. The use of the embedded antenna can reduce SAR level drastically because it is less obtrusive, and introduces more distance between the antenna and the testing probe. The commonly used embedded antenna such as an inverted F antenna is placed at the top or bottom of the mobile device (e.g., cell phones). This placement substantially reduces the peak SAR in the user body. Other methods of reducing SAR include the use of conductive coating on the inner surfaces of the back cover above the antenna or the use of absorbing materials.

5.3 Spherical Vector Wavefunctions

If the antenna is very small compared to the wavelength, the radiated fields will be substantially spherical. Let the antenna be enclosed by a sphere. The radiated fields outside the sphere can be expanded as a linear combination of spherical vector wavefunctions (SVWF), which was first reported by the American physicist William Webster Hansen (1909–1949) (Hansen, 1935). Consider the vector Helmholtz equation

$$\nabla \times \nabla \times \mathbf{F}(\mathbf{r}) - \nabla\nabla \cdot \mathbf{F}(\mathbf{r}) - k^2\mathbf{F}(\mathbf{r}) = 0. \tag{5.19}$$

To find its independent vector solutions, we may start with a scalar function ψ, which is a solution of Helmholtz equation:

$$(\nabla^2 + k^2)\psi = 0. \tag{5.20}$$

It can be shown that (5.19) has three independent vector solutions

$$\mathbf{L} = \nabla\psi, \quad \mathbf{M} = \nabla \times (\mathbf{r}\psi), \quad \mathbf{N} = \frac{1}{k}\nabla \times \nabla \times (\mathbf{r}\psi).$$

If $\{\psi_n\}$ is a complete set, we may expect that the corresponding vector functions $\{\mathbf{L}_n, \mathbf{M}_n, \mathbf{N}_n\}$ also form a complete set and can be used to represent an arbitrary vector wavefunction. In the spherical coordinate

system, the solution of (5.20) is

$$\psi_{nml}^{(q)}(\mathbf{r}) = h_n^{(q)}(kr)Y_{nm}^l(\theta,\varphi).$$

Here $Y_{nm}^l(\theta,\varphi) = P_n^m(\cos\theta)f_{ml}(\varphi)$ $(n=0,1,2,\ldots; m=0,1,2,\ldots,n; l=e,o)$ are the spherical harmonics; $P_n^m(\cos\theta)$ are the associated Legendre functions; $h_n^{(q)}(kr)$ $(q=1,2)$ are the spherical Hankel functions; and

$$f_{ml}(\varphi) = \begin{cases} \cos m\varphi, & l=e \\ \sin m\varphi, & l=o \end{cases}.$$

The **SVWF** are defined by

$$\begin{aligned} \mathbf{L}_{nml}^{(q)}(\mathbf{r}) &= \nabla\big[\psi_{nml}^{(q)}(\mathbf{r})\big], \\ \mathbf{M}_{nml}^{(q)}(\mathbf{r}) &= \nabla\times\big[\mathbf{r}\psi_{nml}^{(q)}(\mathbf{r})\big] = \nabla\psi_{nml}^{(q)}(\mathbf{r})\times\mathbf{r}, \\ \mathbf{N}_{nml}^{(q)}(\mathbf{r}) &= \frac{1}{k}\nabla\times\nabla\times\big[\mathbf{r}\psi_{nml}^{(q)}(\mathbf{r})\big] = \frac{1}{k}\nabla\times\mathbf{M}_{nml}^{(q)}(\mathbf{r}). \end{aligned} \tag{5.21}$$

Explicitly

$$\mathbf{M}_{nml}^{(q)}(\mathbf{r}) = \frac{h_n^{(q)}(kr)}{\sin\theta}\frac{\partial Y_{nm}^l(\theta,\varphi)}{\partial\varphi}\mathbf{u}_\theta - h_n^{(q)}(kr)\frac{\partial Y_{nm}^l(\theta,\varphi)}{\partial\theta}\mathbf{u}_\varphi,$$

$$\begin{aligned} \mathbf{N}_{nml}^{(q)}(\mathbf{r}) = {} & \frac{n(n+1)}{kr}h_n^{(q)}(kr)Y_{nm}^l(\theta,\varphi)\mathbf{u}_r + \frac{1}{kr}\frac{d\big[rh_n^{(q)}(kr)\big]}{dr}\frac{\partial Y_{nm}^l(\theta,\varphi)}{\partial\theta}\mathbf{u}_\theta \\ & + \frac{1}{kr\sin\theta}\frac{d\big[rh_n^{(q)}(kr)\big]}{dr}\frac{\partial Y_{nm}^l(\theta,\varphi)}{\partial\varphi}\mathbf{u}_\varphi. \end{aligned}$$

It can be shown that the SVWF $\{\mathbf{L}_{nml}^{(q)}, \mathbf{M}_{nml}^{(q)}, \mathbf{N}_{nml}^{(q)}\}$ form a complete set (e.g., Geyi, 2010). So the vector potential function can be expanded as follows

$$\mathbf{A} = \frac{1}{j\omega}\sum_{n,m,l,q}\Big(\alpha_{nml}^{(q)}\mathbf{M}_{nml}^{(q)} + \beta_{nml}^{(q)}\mathbf{N}_{nml}^{(q)} + \gamma_{nml}^{(q)}\mathbf{L}_{nml}^{(q)}\Big). \tag{5.22}$$

From Maxwell equations and $\mu\mathbf{H} = \nabla\times\mathbf{A}$, the electromagnetic fields can be expressed by

$$\begin{aligned} \mathbf{E} &= -\sum_{n,m,l,q}\Big(\alpha_{nml}^{(q)}\mathbf{M}_{nml}^{(q)} + \beta_{nml}^{(q)}\mathbf{N}_{nml}^{(q)}\Big), \\ \mathbf{H} &= \frac{1}{j\eta}\sum_{n,m,l,q}\Big(\alpha_{nml}^{(q)}\mathbf{N}_{nml}^{(q)} + \beta_{nml}^{(q)}\mathbf{M}_{nml}^{(q)}\Big), \end{aligned} \tag{5.23}$$

where $\eta = \sqrt{\mu/\varepsilon}$, μ and ε are medium parameters.

Example 5.1 (Spherical waveguide)**:** The free space may be considered as spherical waveguide, and the transmission direction is along the radius r in a spherical coordinate system (r, θ, ϕ) while the waveguide cross-sections are spherical surfaces. The electromagnetic fields $\mathbf{E}$ and $\mathbf{H}$ in spherical coordinates (r, θ, ϕ) can be decomposed into transverse components $\mathbf{E}_t, \mathbf{H}_t$ and radial components $\mathbf{u}_r E_r$, $\mathbf{u}_r H_r$

$$\mathbf{E} = \mathbf{E}_t + \mathbf{u}_r E_r, \quad \mathbf{H} = \mathbf{H}_t + \mathbf{u}_r H_r.$$

Similar to the waveguide theory, we may introduce the orthonormal vector basis functions

$$\mathbf{e}_{nml} = \frac{1}{N_{nm}} \nabla_{\theta\varphi} Y_{nm}^{l}(\theta, \varphi), \quad \mathbf{h}_{nml} = \mathbf{u}_r \times \mathbf{e}_{nml}.$$

where

$$N_{nm} = \sqrt{(1 + \delta_{m0}) \frac{2\pi(n+m)! n(n+1)}{(n-m)!(2n+1)}} \tag{5.24}$$

with $\delta_{m0} = \begin{cases} 1, & m = 0 \\ 0, & m \neq 0 \end{cases}$. Then the SVWF can be expressed as

$$\begin{aligned} \mathbf{M}_{nml}^{(q)} &= -\frac{N_{nm}}{kr} \tilde{h}_n^{(q)}(kr) \mathbf{h}_{nml}, \\ \mathbf{N}_{nml}^{(q)} &= \frac{N_{nm}}{kr} \dot{\tilde{h}}_n^{(q)}(kr) \mathbf{e}_{nml} + \mathbf{u}_r \gamma_{nml}^{(q)}, \end{aligned} \tag{5.25}$$

where $\tilde{h}_n^{(q)}(kr) = kr h_n^{(q)}(kr)$, $\gamma_{nml}^{(q)} = (kr)^{-1} n(n+1) h_n^{(q)}(kr) Y_{nm}^{l}(\theta, \varphi)$ and $\dot{\tilde{h}}_n^{(q)}(kr)$ is the derivative of $\tilde{h}_n^{(q)}(kr)$ with respect to its argument. Substituting (5.25) into (5.23) gives

$$\begin{aligned} r\mathbf{E}_t &= \frac{1}{k} \sum_{m,n,l} \left[N_{nm} \tilde{h}_n^{(1)}(kr) \alpha_{nml}^{(1)} + N_{nm} \tilde{h}_n^{(2)}(kr) \alpha_{nml}^{(2)} \right] \mathbf{h}_{nml} \\ &\quad - \left[N_{nm} \dot{\tilde{h}}_n^{(1)}(kr) \beta_{nml}^{(1)} + N_{nm} \dot{\tilde{h}}_n^{(2)}(kr) \beta_{nml}^{(2)} \right] \mathbf{e}_{nml}, \\ r\mathbf{H}_t &= \frac{1}{jk\eta} \sum_{m,n,l} \left[N_{nm} \dot{\tilde{h}}_n^{(1)}(kr) \alpha_{nml}^{(1)} + N_{nm} \dot{\tilde{h}}_n^{(2)}(kr) \alpha_{nml}^{(2)} \right] \mathbf{e}_{nml} \\ &\quad - \left[N_{nm} \tilde{h}_n^{(1)}(kr) \beta_{nml}^{(1)} + N_{nm} \tilde{h}_n^{(2)}(kr) \beta_{nml}^{(2)} \right] \mathbf{h}_{nml}. \end{aligned} \tag{5.26}$$

These can be rewritten as

$$\begin{aligned} r\mathbf{E}_t &= \sum_{n,m,l} \left[V_{nml}^{\mathrm{TM}}(r)\mathbf{e}_{nml} + V_{nml}^{\mathrm{TE}}(r)\mathbf{h}_{nml}\right], \\ r\mathbf{H}_t &= \sum_{n,m,l} \left[I_{nml}^{\mathrm{TM}}(r)\mathbf{h}_{nml} - I_{nml}^{\mathrm{TE}}(r)\mathbf{e}_{nml}\right]. \end{aligned} \tag{5.27}$$

Here

$$\begin{aligned} V_{nml}^{\mathrm{TE}}(r) &= V_{nml}^{\mathrm{TE}+}(r) + V_{nml}^{\mathrm{TE}-}(r), \\ I_{nml}^{\mathrm{TE}}(r) &= I_{nml}^{\mathrm{TE}+}(r) + I_{nml}^{\mathrm{TE}-}(r), \\ V_{nml}^{\mathrm{TM}}(r) &= V_{nml}^{\mathrm{TM}+}(r) + V_{nml}^{\mathrm{TM}-}(r), \\ I_{nml}^{\mathrm{TM}}(r) &= I_{nml}^{\mathrm{TM}+}(r) + I_{nml}^{\mathrm{TM}-}(r) \end{aligned}$$

are the equivalent modal voltages and currents for TE and TM modes with

$$V_{nml}^{\mathrm{TM}+}(r) = -\frac{N_{nm}\beta_{nml}^{(2)}}{k}\dot{\tilde{h}}_n^{(2)}(kr), \quad V_{nml}^{\mathrm{TM}-}(r) = -\frac{N_{nm}\beta_{nml}^{(1)}}{k}\dot{\tilde{h}}_n^{(1)}(kr),$$

$$V_{nml}^{\mathrm{TE}+}(r) = \frac{N_{nm}\alpha_{nml}^{(2)}}{k}\tilde{h}_n^{(2)}(kr), \quad V_{nml}^{\mathrm{TE}-}(r) = \frac{N_{nm}\alpha_{nml}^{(1)}}{k}\tilde{h}_n^{(1)}(kr),$$

$$I_{nml}^{\mathrm{TE}+}(r) = -\frac{N_{nm}\alpha_{nml}^{(2)}}{j\eta k}\dot{\tilde{h}}_n^{(2)}(kr), \quad I_{nml}^{\mathrm{TE}-}(r) = -\frac{N_{nm}\alpha_{nml}^{(1)}}{j\eta k}\dot{\tilde{h}}_n^{(1)}(kr),$$

$$I_{nml}^{\mathrm{TM}+}(r) = -\frac{N_{nm}\beta_{nml}^{(2)}}{j\eta k}\tilde{h}_n^{(2)}(kr), \quad I_{nml}^{\mathrm{TM}-}(r) = -\frac{N_{nm}\beta_{nml}^{(1)}}{j\eta k}\tilde{h}_n^{(1)}(kr),$$

where the superscripts + and − denote outward-going and inward-going waves, respectively. The radially directed wave impedances for TE modes and TM modes are defined by

$$\begin{aligned} Z_n^{\mathrm{TE}}(r) &= \frac{V_{nml}^{\mathrm{TE}+}(r)}{I_{nml}^{\mathrm{TE}+}(r)} = -j\eta\frac{\tilde{h}_n^{(2)}(kr)}{\dot{\tilde{h}}_n^{(2)}(kr)}, \\ Z_n^{\mathrm{TM}}(r) &= \frac{V_{nml}^{\mathrm{TM}+}(r)}{I_{nml}^{\mathrm{TM}+}(r)} = j\eta\frac{\dot{\tilde{h}}_n^{(2)}(kr)}{\tilde{h}_n^{(2)}(kr)}. \end{aligned}$$

Note that the wave impedances approach to η as $r \to \infty$. □

5.4 Generic Properties of Antennas

If the antenna current distribution is known, all the antenna performances can be determined. Some performances of antenna are very sensitive

to the antenna current distribution while some of them are relatively insensitive. Since the exact current distribution of antenna is very complicated and is not easy to obtain, people usually use approximations to find a simplified current distribution in order to predict the antenna performances that are insensitive to the current distribution, such as gain, antenna pattern, and radiation resistance. In the feeding area, approximations have to be adopted on the basis of a good understanding of how the current distribution affects the various antenna performances. The factors that affect the antenna current distribution include antenna shape, size, excitation, and the environment of the antenna. Müller has systematically studied the properties of electromagnetic radiation patterns (Müller, 1956; 1969) and a summary has been given by Colton and Kress (1983; 1998).

5.4.1 Far Fields and Scattering Matrix

Let V_0 be the volume occupied by the electric current source $\mathbf{J}$ and $\mathbf{J}_m$, as shown in Figure 5.8. The fields produced by a time-harmonic current source $\mathbf{J}$ can be expressed as

$$\mathbf{E}(\mathbf{r}) = -jk\eta \int_{V_0} G(\mathbf{r},\mathbf{r}')\mathbf{J}(\mathbf{r}')dV(\mathbf{r}') - \frac{\eta}{jk}\int_{V_0} \nabla'\cdot\mathbf{J}(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')dV(\mathbf{r}') - \int_{V_0} \mathbf{J}_m(\mathbf{r}')\times\nabla' G(\mathbf{r},\mathbf{r}')dV(\mathbf{r}'), \tag{5.28}$$

$$\mathbf{H}(\mathbf{r}) = -j\frac{k}{\eta}\int_{V_0} G(\mathbf{r},\mathbf{r}')\mathbf{J}_m(\mathbf{r}')dV(\mathbf{r}') - \frac{1}{j\eta k}\int_{V_0} \nabla'\cdot\mathbf{J}_m(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')dV(\mathbf{r}') + \int_{V_0} \mathbf{J}(\mathbf{r}')\times\nabla' G(\mathbf{r},\mathbf{r}')dV(\mathbf{r}'), \tag{5.29}$$

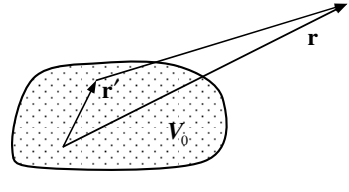


Figure 5.8 An arbitrary current source.

where $G(\mathbf{r},\mathbf{r}') = e^{-jkR}/4\pi R$ with $R = |\mathbf{r}-\mathbf{r}'|$. Making use of the Gauss theorem, we have

$$\int_{V_0} \nabla'\cdot\mathbf{J}(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')dV(\mathbf{r}') = -\int_{V_0}\big[\mathbf{J}(\mathbf{r}')\cdot\nabla'\big]\nabla' G(\mathbf{r},\mathbf{r}')dV(\mathbf{r}'),$$

and the electromagnetic fields can be rewritten as

$$\begin{aligned}\mathbf{E}(\mathbf{r}) &= -jk\eta\int_{V_0}\left(\overleftrightarrow{\mathbf{I}} + \frac{1}{k^2}\nabla\nabla\right)G(\mathbf{r},\mathbf{r}')\cdot\mathbf{J}(\mathbf{r}')dV(\mathbf{r}') \\ &\quad - \int_{V_0}\mathbf{J}_m(\mathbf{r}')\times\nabla' G(\mathbf{r},\mathbf{r}')dV(\mathbf{r}'), \qquad (5.30)\\ \mathbf{H}(\mathbf{r}) &= -j\frac{k}{\eta}\int_{V_0}\left(\overleftrightarrow{\mathbf{I}} + \frac{1}{k^2}\nabla\nabla\right)G(\mathbf{r},\mathbf{r}')\cdot\mathbf{J}_m(\mathbf{r}')dV(\mathbf{r}') \\ &\quad + \int_{V_0}\mathbf{J}(\mathbf{r}')\times\nabla' G(\mathbf{r},\mathbf{r}'dV(\mathbf{r}'), \qquad (5.31)\end{aligned}$$

where $\overleftrightarrow{\mathbf{I}}$ is the identity dyadic tensor. Let $\mathbf{u}_R = (\mathbf{r}-\mathbf{r}')/|\mathbf{r}-\mathbf{r}'|$. Then

$$\begin{aligned}\nabla' G(\mathbf{r},\mathbf{r}') &= \left(jk+\frac{1}{R}\right)G(\mathbf{r},\mathbf{r}')\mathbf{u}_R,\\ \big[\mathbf{J}(\mathbf{r}')\cdot\nabla'\big]\nabla' G(\mathbf{r},\mathbf{r}') &= G(\mathbf{r},\mathbf{r}')\left[-k^2+\frac{3}{R}\left(jk+\frac{1}{R}\right)\right]\big[\mathbf{J}(\mathbf{r}')\cdot\mathbf{u}_R\big]\mathbf{u}_R\\ &\quad - G(\mathbf{r},\mathbf{r}')\frac{\mathbf{J}(\mathbf{r}')}{R}\left(jk+\frac{1}{R}\right).\end{aligned}$$

If R is sufficiently large, we may ignore the terms higher than R^{-1}. Thus

$$\big[\mathbf{J}(\mathbf{r}')\cdot\nabla'\big]\nabla' G(\mathbf{r},\mathbf{r}') \approx -k^2 G(\mathbf{r},\mathbf{r}')\big[\mathbf{J}(\mathbf{r}')\cdot\mathbf{u}_R\big]\mathbf{u}_R. \qquad (5.32)$$

In the far-field region defined by $r \gg r'$, $kr \gg 1$, the following approximations can be made

$$\begin{aligned}R = |\mathbf{r}-\mathbf{r}'| \approx r - \mathbf{u}_r\cdot\mathbf{r}', \quad \frac{1}{|\mathbf{r}-\mathbf{r}'|} &\approx \frac{1}{r},\\ \frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}\mathbf{u}_R &\approx \frac{e^{-jkr}}{r}e^{jk\mathbf{u}_r\cdot\mathbf{r}'}\mathbf{u}_r,\end{aligned} \qquad (5.33)$$

where $\mathbf{u}_r$ is the unit vector along $\mathbf{r}$. It is readily found from (5.28), (5.29), (5.32) and (5.33) that the far-fields have the following asymptotic forms

$$\begin{aligned}\mathbf{E}(\mathbf{r}) &= \frac{e^{-jkr}}{r}\left[\mathbf{E}_\infty(\mathbf{u}_r) + O\left(\frac{1}{r}\right)\right],\\ \mathbf{H}(\mathbf{r}) &= \frac{e^{-jkr}}{r}\left[\mathbf{H}_\infty(\mathbf{u}_r) + O\left(\frac{1}{r}\right)\right].\end{aligned} \tag{5.34}$$

Here the vector fields $\mathbf{E}_\infty$ and $\mathbf{H}_\infty$ are defined on the unit sphere Ω, and are known as the **electric far-field pattern** and **magnetic far-field pattern** respectively. The far-field patterns are independent of the distance r and are given by

$$\begin{aligned}\mathbf{E}_\infty(\mathbf{u}_r) &= -\frac{jk\eta}{4\pi}\int\limits_{V_0}\left[\mathbf{J} - (\mathbf{J}\cdot\mathbf{u}_r)\mathbf{u}_r + \frac{1}{\eta}\mathbf{J}_m\times\mathbf{u}_r\right]e^{jk\mathbf{u}_r\cdot\mathbf{r}'}dV(\mathbf{r}'),\\ \mathbf{H}_\infty(\mathbf{u}_r) &= -\frac{jk}{4\pi\eta}\int\limits_{V_0}\left[\mathbf{J}_m - (\mathbf{J}_m\cdot\mathbf{u}_r)\mathbf{u}_r - \eta\mathbf{J}\times\mathbf{u}_r\right]e^{jk\mathbf{u}_r\cdot\mathbf{r}'}dV(\mathbf{r}'),\end{aligned} \tag{5.35}$$

and satisfy

$$\eta\mathbf{H}_\infty(\mathbf{u}_r) = \mathbf{u}_r\times\mathbf{E}_\infty(\mathbf{u}_r),\ \ \mathbf{u}_r\cdot\mathbf{E}_\infty(\mathbf{u}_r) = \mathbf{u}_r\cdot\mathbf{H}_\infty(\mathbf{u}_r) = 0. \tag{5.36}$$

Remark 5.1: Let S be any closed surface that encloses the source region V_0. The far-field patterns can also be expressed as

$$\begin{aligned}\mathbf{E}_\infty(\mathbf{u}_r) &= -\frac{jk\eta}{4\pi}\int\limits_{S}\left[\mathbf{J}_s - (\mathbf{J}_s\cdot\mathbf{u}_r)\mathbf{u}_r + \frac{1}{\eta}\mathbf{J}_{ms}\times\mathbf{u}_r\right]e^{jk\mathbf{u}_r\cdot\mathbf{r}'}dS(\mathbf{r}'),\\ \mathbf{H}_\infty(\mathbf{u}_r) &= -\frac{jk}{4\pi\eta}\int\limits_{S}\left[\mathbf{J}_{ms} - (\mathbf{J}_{ms}\cdot\mathbf{u}_r)\mathbf{u}_r - \eta\mathbf{J}_s\times\mathbf{u}_r\right]e^{jk\mathbf{u}_r\cdot\mathbf{r}'}dS(\mathbf{r}'),\end{aligned} \tag{5.37}$$

where $\mathbf{J}_s = \mathbf{u}_n\times\mathbf{H}$ and $\mathbf{J}_{ms} = -\mathbf{u}_n\times\mathbf{E}$ are the equivalent surface electric and magnetic current respectively. □

It follows from (5.34) and (5.36) that the far-fields satisfy the Silver–Müller **radiation condition**

$$\lim_{r\to\infty} r(\mathbf{u}_r\times\mathbf{E} - \eta\mathbf{H}) = 0.$$

If there are no magnetic sources, the Poynting vector in the far-field region can be expressed as

$$\mathbf{S} = \frac{1}{2}\mathrm{Re}(\mathbf{E} \times \bar{\mathbf{H}}) = \mathbf{u}_r \frac{k^2\eta}{32\pi^2 r^2} \left| \mathbf{u}_r \times \int_{V_0} \mathbf{J} e^{jk\mathbf{u}_r \cdot \mathbf{r}'} dV(\mathbf{r}') \right|^2 . \tag{5.38}$$

The total radiated power by the current $\mathbf{J}$ is

$$\begin{aligned} P_{\mathrm{rad}} &= \int_{S_\infty} \mathbf{S} \cdot \mathbf{u}_n \, dS(\mathbf{r}') = \int_{S_\infty} \mathbf{S} \cdot \mathbf{u}_n r^2 d\Omega \\ &= \frac{k^2\eta}{32\pi^2} \int_{S_\infty} \left| \mathbf{u}_r \times \int_{V_0} \mathbf{J} e^{jk\mathbf{u}_r \cdot \mathbf{r}'} dV(\mathbf{r}') \right|^2 d\Omega(\mathbf{r}). \end{aligned} \tag{5.39}$$

From the Poynting theorem and (5.30), the radiated power can also be calculated by the **method of induced electromotive force** (EMF)

$$\begin{aligned} P_{\mathrm{rad}} &= -\frac{1}{2}\mathrm{Re} \int_{V_0} \overline{\mathbf{J}}(\mathbf{r}') \cdot \mathbf{E}(\mathbf{r}') dV(\mathbf{r}') \\ &= \frac{k\eta}{8\pi} \int_{V_0} \int_{V_0} \overline{\mathbf{J}}(\mathbf{r}) \cdot \left(\overleftrightarrow{\mathbf{I}} + \frac{1}{k^2} \nabla\nabla \right) \frac{\sin(k|\mathbf{r} - \mathbf{r}'|)}{|\mathbf{r} - \mathbf{r}'|} \cdot \mathbf{J}(\mathbf{r}') dV(\mathbf{r}) dV(\mathbf{r}'). \end{aligned} \tag{5.40}$$

Figure 5.9 shows a typical scattering problem. An impressed source $\mathbf{J}_{\mathrm{imp}}$ generates the incident fields $(\mathbf{E}_{\mathrm{in}}, \mathbf{H}_{\mathrm{in}})$, which induce a current distribution $\mathbf{J}$ on the scatterer characterized by medium tensors $(\overleftrightarrow{\boldsymbol{\mu}}, \overleftrightarrow{\boldsymbol{\varepsilon}})$. The induced

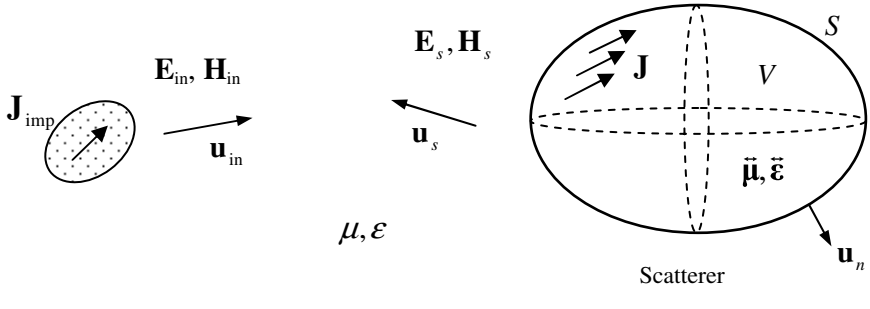


Figure 5.9 Scattering problem.

current $\mathbf{J}$ produces the scattered fields $(\mathbf{E}_s, \mathbf{H}_s)$. When the impressed source and the scatterer are in close proximity, the scattering problem essentially becomes a radiation (antenna) problem. In the far-field region of the impressed source, the incident fields in a direction $\mathbf{u}_{\text{in}}$ are a plane wave and may be expressed by

$$\mathbf{E}_{\text{in}}(\mathbf{r}) = \mathbf{p}_{\text{in}} e^{-jk\mathbf{u}_{\text{in}}\cdot\mathbf{r}}, \quad \mathbf{H}_{\text{in}}(\mathbf{r}) = \frac{1}{\eta}\mathbf{u}_{\text{in}} \times \mathbf{p}_{\text{in}} e^{-jk\mathbf{u}_{\text{in}}\cdot\mathbf{r}}, \tag{5.41}$$

where $\mathbf{p}_{\text{in}}$ is a constant vector and can be decomposed into the sum of two orthogonal components

$$\mathbf{p}_{\text{in}} = p_{\text{in}}^{(1)}\mathbf{u}_1 + p_{\text{in}}^{(2)}\mathbf{u}_2. \tag{5.42}$$

Here $\mathbf{u}_1$ and $\mathbf{u}_2$ are unit vectors, which, together with $\mathbf{u}_{\text{in}}$, form a orthogonal set of unit vectors. Assume that the scatterer is in the far-field region of the impressed source. At the far-field region of the induced current $\mathbf{J}$, the scattered fields due to incident fields (5.41) are of the form

$$\mathbf{E}_s(\mathbf{r}) = \frac{e^{-jk\mathbf{u}_s\cdot\mathbf{r}}}{r}\mathbf{E}_\infty(\mathbf{u}_s), \quad \mathbf{H}_s(\mathbf{r}) = \frac{e^{-jk\mathbf{u}_s\cdot\mathbf{r}}}{r}\mathbf{H}_\infty(\mathbf{u}_s).$$

The electric far-field pattern can be written as

$$\begin{aligned}\mathbf{E}_\infty(\mathbf{u}_s) &= p_{\text{in}}^{(1)}\mathbf{E}_\infty^{(1)}(\mathbf{u}_s) + p_{\text{in}}^{(2)}\mathbf{E}_\infty^{(2)}(\mathbf{u}_s) \\ &= \left[\mathbf{E}_\infty^{(1)}(\mathbf{u}_s)\mathbf{u}_1 + \mathbf{E}_\infty^{(2)}(\mathbf{u}_s)\mathbf{u}_2\right]\cdot\mathbf{p}_{\text{in}} = \overleftrightarrow{\mathbf{S}}(\mathbf{u}_{\text{in}}, \mathbf{u}_s)\cdot\mathbf{p}_{\text{in}},\end{aligned} \tag{5.43}$$

where $\mathbf{E}_\infty^{(1)}(\mathbf{u}_s)$ and $\mathbf{E}_\infty^{(2)}(\mathbf{u}_s)$ are respectively the scattered fields induced by the incident field $\mathbf{E}_{\text{in}}(\mathbf{r}) = \mathbf{u}_1 e^{-jk\mathbf{u}_{\text{in}}\cdot\mathbf{r}}$ and $\mathbf{E}_{\text{in}}(\mathbf{r}) = \mathbf{u}_2 e^{-jk\mathbf{u}_{\text{in}}\cdot\mathbf{r}}$, and

$$\overleftrightarrow{\mathbf{S}}(\mathbf{u}_{\text{in}}, \mathbf{u}_s) = \mathbf{E}_\infty^{(1)}(\mathbf{u}_s)\mathbf{u}_1 + \mathbf{E}_\infty^{(2)}(\mathbf{u}_s)\mathbf{u}_2 \tag{5.44}$$

is a dyad and referred to as the **scattering matrix**.

We now assume that a different plane wave

$$\mathbf{E}'_{\text{in}}(\mathbf{r}) = \mathbf{p}'_{\text{in}} e^{-jk\mathbf{u}'_{\text{in}}\cdot\mathbf{r}}, \quad \mathbf{H}'_{\text{in}}(\mathbf{r}) = \frac{1}{\eta}\mathbf{u}'_{\text{in}} \times \mathbf{p}'_{\text{in}} e^{-jk\mathbf{u}'_{\text{in}}\cdot\mathbf{r}} \tag{5.45}$$

is incident upon a scatterer which occupies the same volume as the scatterer shown in Figure 5.9, but endowed with transposed medium parameters $\overleftrightarrow{\boldsymbol{\mu}}^t$, $\overleftrightarrow{\boldsymbol{\varepsilon}}^t$. The scattered fields induced by the incident fields (5.45) are denoted

by $\mathbf{E}'_s, \mathbf{H}'_s$. From the reciprocity theorem, we may write

$$\int_S [(\mathbf{E}_{\text{in}} + \mathbf{E}_s) \times (\mathbf{H}'_{\text{in}} + \mathbf{H}'_s) - (\mathbf{E}'_{\text{in}} + \mathbf{E}'_s) \times (\mathbf{H}_{\text{in}} + \mathbf{H}_s)] \cdot \mathbf{u}_n \, dS(\mathbf{r}) = 0.$$

This can be reduced to

$$\int_S (\mathbf{E}_s \times \mathbf{H}'_{\text{in}} - \mathbf{E}'_{\text{in}} \times \mathbf{H}_s) \cdot \mathbf{u}_n \, dS(\mathbf{r})$$

$$= \int_S (\mathbf{E}'_s \times \mathbf{H}_{\text{in}} - \mathbf{E}_{\text{in}} \times \mathbf{H}'_s) \cdot \mathbf{u}_n \, dS(\mathbf{r}), \tag{5.46}$$

where we have used the following relations (reciprocity theorems)

$$\int_S (\mathbf{E}_{\text{in}} \times \mathbf{H}'_{\text{in}} - \mathbf{E}'_{\text{in}} \times \mathbf{H}_{\text{in}}) \cdot \mathbf{u}_n \, dS(\mathbf{r}) = 0,$$

$$\int_S (\mathbf{E}_s \times \mathbf{H}'_s - \mathbf{E}'_s \times \mathbf{H}_s) \cdot \mathbf{u}_n \, dS(\mathbf{r}) = 0.$$

Inserting (5.41) and (5.45) into (5.46), we obtain

$$\int_S \mathbf{p}'_{\text{in}} \cdot \left[\mathbf{u}_n \times \mathbf{H}_s - \eta^{-1} \mathbf{u}'_{\text{in}} \times (\mathbf{u}_n \times \mathbf{E}_s)\right] e^{-jk\mathbf{u}'_{\text{in}} \cdot \mathbf{r}} dS(\mathbf{r})$$

$$= \int_S \mathbf{p}_{\text{in}} \cdot \left[\mathbf{u}_n \times \mathbf{H}'_s - \eta^{-1} \mathbf{u}_{\text{in}} \times (\mathbf{u}_n \times \mathbf{E}'_s)\right] e^{-jk\mathbf{u}_{\text{in}} \cdot \mathbf{r}} dS(\mathbf{r}). \tag{5.47}$$

Making use of (5.37), this can be written as

$$\mathbf{p}'_{\text{in}} \cdot \mathbf{E}_\infty(-\mathbf{u}'_{\text{in}}) = \mathbf{p}_{\text{in}} \cdot \mathbf{E}'_\infty(-\mathbf{u}_{\text{in}}), \tag{5.48}$$

or

$$\mathbf{p}'_{\text{in}} \cdot \overleftrightarrow{\mathbf{S}}(\mathbf{u}_{\text{in}}, -\mathbf{u}'_{\text{in}}) \cdot \mathbf{p}_{\text{in}} = \mathbf{p}_{\text{in}} \cdot \overleftrightarrow{\mathbf{S}}'(\mathbf{u}'_{\text{in}}, -\mathbf{u}_{\text{in}}) \cdot \mathbf{p}'_{\text{in}}. \tag{5.49}$$

This implies

$$\overleftrightarrow{\mathbf{S}}^t(\mathbf{u}_{\text{in}}, -\mathbf{u}'_{\text{in}}) = \overleftrightarrow{\mathbf{S}}'(\mathbf{u}'_{\text{in}}, -\mathbf{u}_{\text{in}}). \tag{5.50}$$

5.4.2 Poynting Theorem and Stored Energies

The differential form of the complex Poynting theorem for time-harmonic fields in a homogeneous and isotropic medium is

$$\nabla \cdot \mathbf{S} = -\frac{1}{2} \bar{\mathbf{J}} \cdot \mathbf{E} - j2\omega(w_m - w_e), \tag{5.51}$$

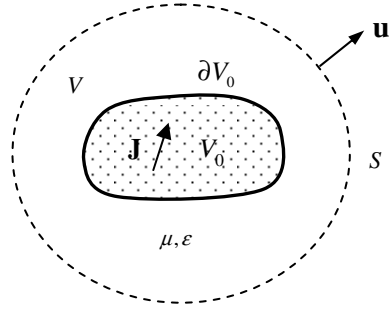


Figure 5.10 Poynting theorem.

where $\mathbf{S} = \mathbf{E} \times \bar{\mathbf{H}}/2$ is the complex Poynting vector, $w_m = \mu \mathbf{H} \cdot \bar{\mathbf{H}}/4$ and $w_e = \varepsilon \mathbf{E} \cdot \bar{\mathbf{E}}/4$ are the magnetic and electric field energy densities. Let V_0 be the volume occupied by the electric current source $\mathbf{J}$ and ∂V_0 be the surface surrounding V_0. Taking the integration of the imaginary part of (5.51) over a volume V containing V_0 as shown in Figure 5.10, we obtain

$$\mathrm{Im} \int_S \mathbf{u}_n \cdot \mathbf{S}\, dS = -\mathrm{Im} \int_{V_0} \frac{1}{2} \bar{\mathbf{J}} \cdot \mathbf{E}\, dV - 2\omega \int_V (w_m - w_e) dV, \tag{5.52}$$

where S is the boundary of V. Choosing $V = V_0$, Equation (5.52) becomes

$$\mathrm{Im} \int_{\partial V_0} \mathbf{u}_n \cdot \mathbf{S}\, dS = -\mathrm{Im} \int_{V_0} \frac{1}{2} \bar{\mathbf{J}} \cdot \mathbf{E}\, dV - 2\omega \int_{V_0} (w_m - w_e) dV. \tag{5.53}$$

If we choose $V = V_\infty$, where V_∞ is the region enclosed by a sphere S_∞ with radius r being sufficiently large so that it lies in the far-field region of the antenna system, we get

$$\mathrm{Im} \int_{S_\infty} \mathbf{u}_n \cdot \mathbf{S}\, dS = -\mathrm{Im} \int_{V_0} \frac{1}{2} \bar{\mathbf{J}} \cdot \mathbf{E}\, dV - 2\omega \int_{V_\infty} (w_m - w_e) dV. \tag{5.54}$$

Since $\mathbf{S}$ is a real vector in the far-field region, the above equation reduces to

$$-\mathrm{Im} \int_{V_0} \frac{1}{2} \bar{\mathbf{J}} \cdot \mathbf{E}\, dV = 2\omega \int_{V_\infty} (w_m - w_e) dV. \tag{5.55}$$

It follows from (5.52)–(5.54) that

$$\mathrm{Im} \int\limits_{\partial V_0} \mathbf{u}_n \cdot \mathbf{S}\, dS = 2\omega \int\limits_{V_\infty - V_0} (w_m - w_e) dV, \tag{5.56}$$

$$\mathrm{Im} \int\limits_{S} \mathbf{u}_n \cdot \mathbf{S}\, dS = 2\omega \int\limits_{V_\infty - V} (w_m - w_e) dV. \tag{5.57}$$

Taking the integration of the real part of (5.51) over the volume V containing the source region V_0, we obtain the radiated power

$$P_{\mathrm{rad}} = \mathrm{Re} \int\limits_{S} \mathbf{u}_n \cdot \mathbf{S}\, dS = -\mathrm{Re} \int\limits_{V_0} \frac{1}{2} \bar{\mathbf{J}} \cdot \mathbf{E}\, dV. \tag{5.58}$$

Equation (5.58) shows that the surface integral of the real part of the Poynting vector is independent of the surface S as long as it encloses the source region V_0. Equations (5.56) and (5.57) show that the surface integral of the imaginary part of the Poynting vector depends on the integration surface S in the near-field region (in the far-field region it becomes zero). Considering (5.52), (5.55) and (5.58) we may find that

$$\begin{aligned} \int\limits_{S} \mathbf{u}_n \cdot \mathbf{S}\, dS &= -\int\limits_{V_0} \frac{1}{2} \bar{\mathbf{J}} \cdot \mathbf{E}\, dV - j2\omega \int\limits_{V} (w_m - w_e) dV \\ &= P_{\mathrm{rad}} - j\mathrm{Im} \int\limits_{V_0} \frac{1}{2} \bar{\mathbf{J}} \cdot \mathbf{E}\, dV - j2\omega \int\limits_{V} (w_m - w_e) dV \\ &= P_{\mathrm{rad}} + j2\omega \int\limits_{V_\infty - V} (w_m - w_e) dV. \end{aligned} \tag{5.59}$$

The above relation indicates that the complex power flowing out of S is equal to the radiation power plus the reactive power outside S. This expression is the most general form of the Poynting theorem for an open system. Let $\widetilde{w}_e$ (w_e^{rad}) and $\widetilde{w}_m$ (w_m^{rad}) denote the stored (radiated) electric field and magnetic field energy densities, respectively. The stored energies are defined by (Counter, 1948; Collin and Rothschild, 1964)

$$\widetilde{w}_m = w_m - w_m^{\mathrm{rad}}, \quad \widetilde{w}_e = w_e - w_e^{\mathrm{rad}}. \tag{5.60}$$

These calculations are physically appropriate since density is a summable quantity. It is readily seen from (5.57) that w_m is equal to w_e in the far-field

zone, since the complex Poynting vector becomes real as V approaches V_∞. This observation indicates that the electric field energy and the magnetic field energy for the radiated field are identical everywhere, i.e.,

$$w_e^{\rm rad} = \frac{1}{4}\varepsilon \mathbf{E}_{\rm rad}\cdot\bar{\mathbf{E}}_{\rm rad} = \frac{1}{4}\mu \mathbf{H}_{\rm rad}\cdot\bar{\mathbf{H}}_{\rm rad} = w_m^{\rm rad}. \tag{5.61}$$

The total energy of the radiated fields is simply twice the electric or magnetic energy density of the radiated fields. Mathematically, Equation (5.61) holds everywhere. Consequently, from (5.55), (5.56) and (5.59), we obtain

$$\begin{aligned}\widetilde{W}_m - \widetilde{W}_e &= \int\limits_{V_\infty - V_0} (\widetilde{w}_m - \widetilde{w}_e)dV = \int\limits_{V_\infty - V_0} (w_m - w_e)dV \\ &= \frac{1}{2\omega}{\rm Im}\int\limits_{\partial V_0} \mathbf{S}\cdot\mathbf{u}_n\, dS. \end{aligned}\tag{5.62}$$

Here $\widetilde{W}_m$ and $\widetilde{W}_e$ stand for the total stored magnetic and electric energy in the volume surrounding the radiator. Note that the total stored energy can be expressed as

$$\begin{aligned}\widetilde{W}_e + \widetilde{W}_m &= \int\limits_{V_\infty - V_0} (w_e - w_e^{\rm rad} + w_m - w_m^{\rm rad})dV \\ &= \int\limits_{V_\infty - V_0} (w_e + w_m)dV - \int\limits_{V_\infty - V_0} (w_e^{\rm rad} + w_m^{\rm rad})dV \\ &= \int\limits_{V_\infty - V_0} (w_e + w_m)dV - \frac{r}{v}{\rm Re}\int\limits_{\partial V_0} \mathbf{S}\cdot\mathbf{u}_n\, dS, \end{aligned}\tag{5.63}$$

where r is the radius of the sphere S_∞, and v is the wave velocity. Both terms on the right-hand side of (5.63) are divergent as $r\to\infty$, but it can be shown that their difference is finite. So the stored electric and magnetic field energies may be obtained from (5.62) and (5.63) as

$$\begin{aligned}\widetilde{W}_m = \frac{1}{2}\Bigg[&\int\limits_{V_\infty - V_0} (w_e + w_m)dV - \frac{r}{v}{\rm Re}\int\limits_{\partial V_0} \mathbf{S}\cdot\mathbf{u}_n\, dS \\ &+ \frac{1}{2\omega}{\rm Im}\int\limits_{\partial V_0} \mathbf{S}\cdot\mathbf{u}_n\, dS\Bigg],\end{aligned}$$

$$\widetilde{W}_e = \frac{1}{2}\left[\int\limits_{V_\infty - V_0} (w_e + w_m)dV - \frac{r}{v}\mathrm{Re}\int\limits_{\partial V_0} \mathbf{S}\cdot\mathbf{u}_n\, dS \right.$$
$$\left. - \frac{1}{2\omega}\mathrm{Im}\int\limits_{\partial V_0} \mathbf{S}\cdot\mathbf{u}_n\, dS\right]. \tag{5.64}$$

Note that the stored energies are localized in the vicinity of antenna. Therefore (5.64) will quickly become stable (a constant independent of r) as distance r increases. In fact, (5.64) has been directly used to calculate the stored energy of antennas by using FDTD (Collardey *et al.*, 2005; 2006), and it has been demonstrated that the stored energy quickly becomes stable if r is increased to one or a few wavelengths. This fact can also be verified by the following reasoning. Let r be the radius of the sphere that encloses the sources. It will be shown later that the stored energies outside the sphere can be written as [see (5.125)]

$$\widetilde{W}_m = \sum_{n=1}^{\infty}\left(a_n^2 Q_n + b_n^2 Q_n'\right), \quad \widetilde{W}_e = \sum_{n=1}^{\infty}\left(a_n^2 Q_n' + b_n^2 Q_n\right), \tag{5.65}$$

where a_n and b_n are positive constants determined by the sources; Q_n and Q_n' are the quality factors for spherical modes [see (5.126)]. Making use of the properties of Q_n and Q_n', we have the following asymptotic behaviors [see (5.127)]

$$\widetilde{W}_m \propto \frac{1}{kr}, \quad \widetilde{W}_e \propto \frac{1}{kr}, \quad r \to \infty. \tag{5.66}$$

Therefore, the stored energies outside the sphere decrease rapidly as the radius of the sphere increases. In other words, the stored energies defined by (5.63) and (5.64) will approach to a constant value for sufficiently large r, and the stored energies exist only in the vicinity of antenna in a quasi-static form. In fact, the stored energies will become a constant if r falls into the far-field region of the antenna. This agrees with our common understanding.

5.4.3 Equivalent Circuits for Antennas

A transmitting antenna can be converted to a RLC circuit, where the resistance R is introduced to represent the radiated energy and the heat due to the loss of the medium around the antenna; the inductance L and capacitancesi C are used to represent the stored magnetic energy and

the stored electric energy around the antenna respectively. A receiving antenna can also be converted to a RLC circuit and its derivation gets more complicated.

5.4.3.1 *Equivalent Circuit for Transmitting Antennas*

We choose the source region V_0 in such a way that its surface ∂V_0 is coincident with the antenna surface (except at the antenna input terminal Ω where ∂V_0 crosses the antenna reference plane T), as shown in Figure 5.11. Applying Poynting theorem over the lossless region $V_\infty - V_0$ yields

$$\frac{1}{2}\int_{S_\infty} (\mathbf{E}\times\bar{\mathbf{H}})\cdot\mathbf{u}_n\, dS + \frac{1}{2}\int_{\partial V_0} (\mathbf{E}\times\bar{\mathbf{H}})\cdot\mathbf{u}_n\, dS = -j2\omega \int_{V_\infty - V_0} (w_m - w_e) dV. \tag{5.67}$$

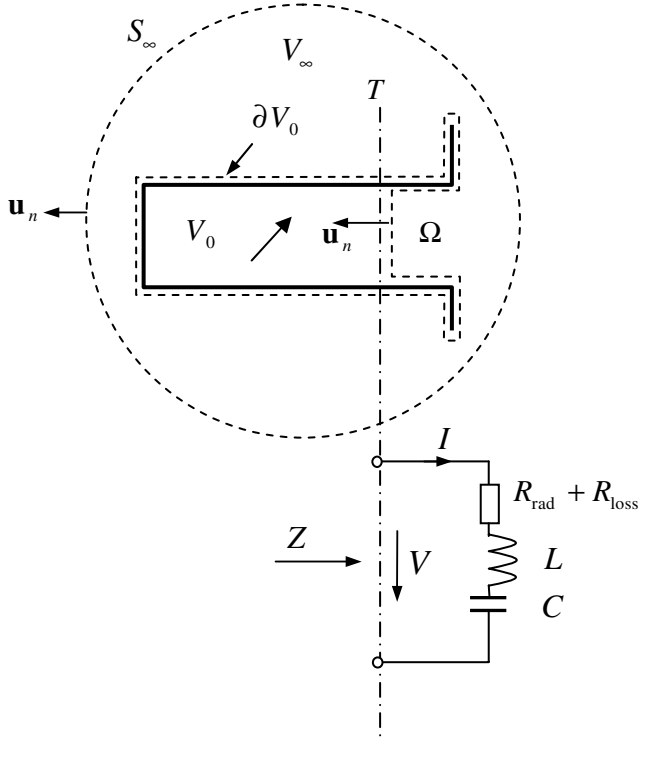


Figure 5.11 Equivalent circuit for transmitting antenna.

If we assume that the antenna surface is perfectly conducting, $\mathbf{E} \times \bar{\mathbf{H}}$ vanishes everywhere on ∂V_0 except over the input terminal Ω. We further assume that the antenna reference plane T is away from the antenna discontinuity so that the higher order modes excited by the discontinuity has negligible effects at the reference plane. For a single mode feeding waveguide, we have

$$\frac{1}{2}\int_{\partial V_0} (\mathbf{E} \times \bar{\mathbf{H}}) \cdot \mathbf{u}_n \, dS = \frac{1}{2}\int_{\Omega} (\mathbf{E} \times \bar{\mathbf{H}}) \cdot \mathbf{u}_n \, dS = -\frac{1}{2} V \bar{I}, \tag{5.68}$$

where V and I are equivalent modal voltage and current at the reference plane, respectively. Introducing (5.68) into (5.67) and using the fact that $P_{\text{rad}} = \frac{1}{2}\int_{\partial V_\infty} (\mathbf{E} \times \bar{\mathbf{H}}) \cdot \mathbf{u}_n \, dS$, we may find that

$$\frac{1}{2} V \bar{I} = P_{\text{rad}} + j2\omega \int_{V_\infty - V_0} (w_m - w_e) dV.$$

The antenna impedance Z is defined by

$$Z = \frac{V}{I} = \frac{2P_{\text{rad}}}{|I|^2} + j\frac{4\omega(W_m - W_e)}{|I|^2},$$

where W_m and W_e are the total magnetic energy and electric energy produced by the antenna respectively, and both are infinite as integration region $V_\infty - V_0$ is infinite. By use of (5.62), we have $W_m - W_e = \widetilde{W}_m - \widetilde{W}_e$. Thus

$$Z = R_{\text{rad}} + jX = R_{\text{rad}} + j\left(\omega L - \frac{1}{\omega C}\right),$$

where R_{rad} and X denote the radiation resistance, reactance respectively and their definitions are given below

$$R_{\text{rad}} = \frac{2P_{\text{rad}}}{|I|^2}, \quad X = \omega L - \frac{1}{\omega C}, \quad L = \frac{4\widetilde{W}_m}{|I|^2}, \quad C = \frac{|I|^2}{4\omega^2 \widetilde{W}_e}. \tag{5.69}$$

The equivalent circuits for the antenna are shown in Figure 5.11. It should be noted that it is the stored energy (not the total energy) of the field that we use to derive the inductance L and capacitance C. Physically, the stored energy means the electromagnetic energy that is temporarily located in the field and fully recoverable when the field is reduced to zero. The total electromagnetic energy around an antenna is infinite while the total stored

electromagnetic energy is always finite, which, on the other hand, can never be made zero as will be revealed later. The above understanding makes it possible to have a rigorous derivation of equivalent circuit for a transmitting antenna.

5.4.3.2 *Equivalent Circuit for Receiving Antennas*

An equivalent circuit for a receiving antenna is shown in Figure 5.12, where the receiving antenna has been represented by a Thévenin equivalent circuit with a voltage source $V_{\rm oc}$ and an internal impedance Z in series, and Z_L is the load connected to the receiving antenna. It is well-known that the antenna impedance Z for the receiving antenna is approximately equal to the antenna input impedance when the receiving antenna is in the transmit mode. Some conditions have to be applied for this kind of "reciprocity of impedance" to be valid. Firstly, the source of the incident field must be far away from the receiving antenna. Secondly, the equivalent source $V_{\rm oc}$ should be taken as the open circuit voltage of the receiving antenna. Such an equivalent circuit for the receiving antenna has certain limitations since the power dissipation on the internal impedance Z is difficult to interpret. Therefore an equivalent circuit for the receiving antenna, which gives a correct power balance relationship, is more appealing and useful (Geyi, 2004).

Let the fields generated from antenna 1 when antenna 2 is receiving be denoted by $\mathbf{E}$, $\mathbf{H}$. Then we may write

$$\mathbf{E} = \mathbf{E}_{\rm in} + \mathbf{E}_s, \quad \mathbf{H} = \mathbf{H}_{\rm in} + \mathbf{H}_s,$$

where $\mathbf{E}_{\rm in}, \mathbf{H}_{\rm in}$ stand for the incident fields produced by antenna 1 when antenna 2 is not present, and $\mathbf{E}_s, \mathbf{H}_s$ represent the scattered fields generated by antenna 2. We use V_{01} and V_{02} to denote the source region for antenna 1 and 2, respectively. The source regions are chosen in such a way that their boundaries, denoted by ∂V_{01} and ∂V_{02}, are coincident with the metal surface of the antennas except for a portion Ω_1 or Ω_2 where the boundaries cross the

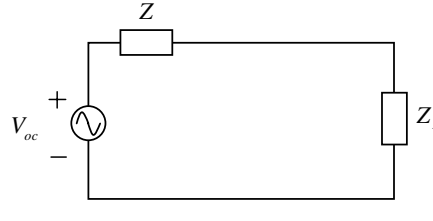


Figure 5.12 Equivalent circuit of receiving antenna.

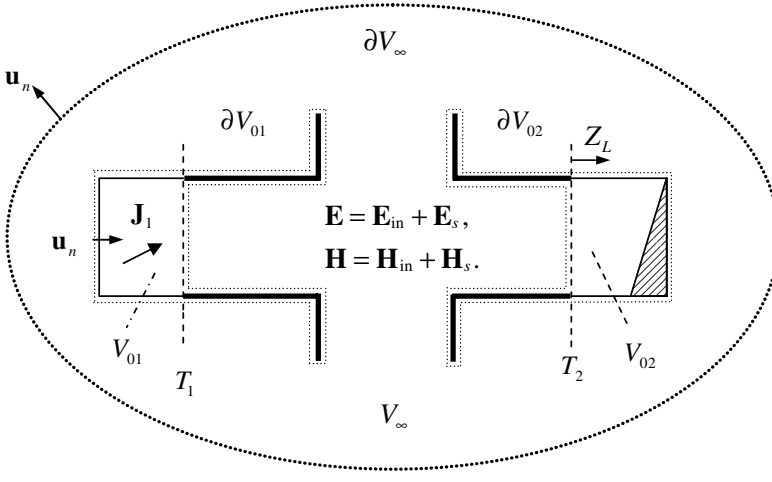


Figure 5.13 Two-antenna system.

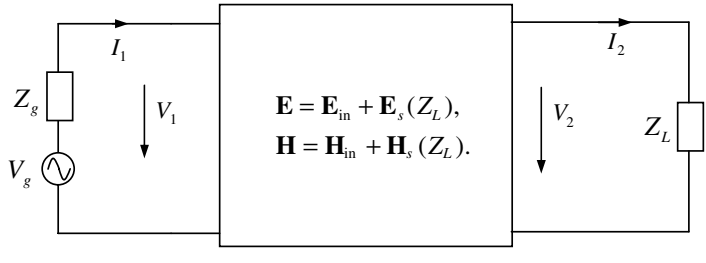


Figure 5.14 Equivalent network for two-antenna system.

antenna feeding planes T_1 or T_2. The medium around antenna is assumed to be isotropic, homogeneous, non-dispersive and lossless. This state of operation is illustrated in Figure 5.13, and Figure 5.14 is the corresponding equivalent network representation.

Taking the integration of Poynting theorem in frequency domain and using the divergence theorem over the region $V_\infty - V_{01} - V_{02}$ in Figure 5.13, we get

$$\frac{1}{2}\int_{\partial V_\infty} (\mathbf{E}\times\bar{\mathbf{H}})\cdot\mathbf{u}_n\,dS + \frac{1}{2}\int_{\partial V_{01}} (\mathbf{E}\times\bar{\mathbf{H}})\cdot\mathbf{u}_n\,dS + \frac{1}{2}\int_{\partial V_{02}} (\mathbf{E}\times\bar{\mathbf{H}})\cdot\mathbf{u}_n\,dS$$
$$= -j2\omega \int_{V_\infty - V_{01} - V_{02}} (w_m - w_e)dV. \tag{5.70}$$

If the antenna surface is perfectly conducting, $(\mathbf{E} \times \bar{\mathbf{H}}) \cdot \mathbf{u}_n$ vanishes everywhere on ∂V_{01} and ∂V_{02} except over the antenna input terminal Ω_1 and Ω_2. For a single mode feeding line, we have

$$\frac{1}{2} \int\limits_{\partial V_{01}} (\mathbf{E} \times \bar{\mathbf{H}}) \cdot \mathbf{u}_n \, dS = -\frac{V_1 \bar{I}_1}{2}, \quad \frac{1}{2} \int\limits_{\partial V_{02}} (\mathbf{E} \times \bar{\mathbf{H}}) \cdot \mathbf{u}_n \, dS = \frac{V_2 \bar{I}_2}{2}. \tag{5.71}$$

Hereafter, we will use $V_{1(\text{or }2)}$ and $I_{1(\text{or }2)}$ to represent the terminal voltage and current at the feeding plane of antenna 1 (or antenna 2) when antenna 1 is transmitting. It should be noted that all the terminal voltage and current are defined on the basis of total fields. Introducing (5.71) into (5.70), we obtain

$$\frac{V_1 \bar{I}_1}{2} - \frac{V_2 \bar{I}_2}{2} = \frac{1}{2} \int\limits_{\partial V_\infty} (\mathbf{E} \times \bar{\mathbf{H}}) \cdot \mathbf{u}_n \, dS + j2\omega \int\limits_{V_\infty - V_{01} - V_{02}} (w_m - w_e) dV.$$

The above equation can be written as

$$P_{\text{in}} = P_{\text{rad}} + P_L + j2\omega(\widetilde{W}_m - \widetilde{W}_e), \tag{5.72}$$

where $P_{\text{in}} = V_1 \bar{I}_1/2$ is the complex power transmitted by the antenna 1; $P_L = V_2 \bar{I}_2/2$ is the complex power absorbed by the load Z_L; $P_{\text{rad}} = (1/2) \int\limits_{\partial V_\infty} (\mathbf{E} \times \bar{\mathbf{H}}) \cdot \mathbf{u}_n \, dS$ is the total power radiated (escaped) away into space; $\widetilde{W}_m$ and $\widetilde{W}_e$ are the stored magnetic and electric energy respectively. In the above equation, we have used the following calculation

$$\widetilde{W}_m - \widetilde{W}_e = \int\limits_{V_\infty - V_{01} - V_{02}} (w_m - w_e) dV.$$

The power balance relation in (5.72) shows that if the power radiated away into space is zero the transmitted power from antenna 1 can be totally absorbed by the load of the receiving antenna, which is theoretically possible. An example is a two ideal planar aperture antenna system focused to each other, where the power transmission efficiency could attain 100%. When the receiving antenna is in the far-field region of the transmitting antenna the radiated power P_{rad} is generally not zero and the power absorbed by the load is just small portion of the total transmitted power. From (5.72) an equivalent circuit for the receiving antenna can

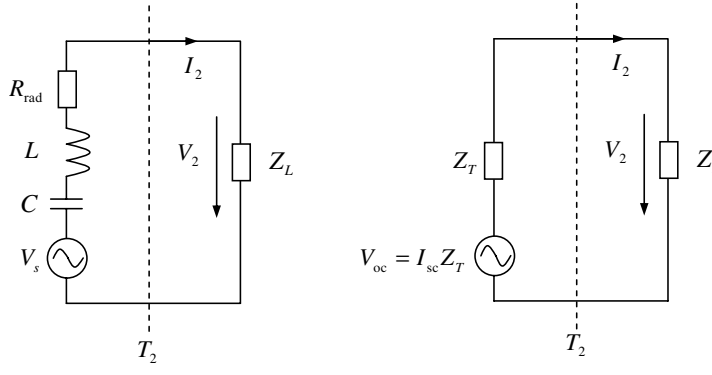


Figure 5.15 Equivalent circuits for receiving antenna.

be constructed as shown in Figure 5.15(a) by introducing the following circuit elements

$$V_s = \frac{2P_{\text{in}}}{\bar{I}_2}, \quad R_{\text{rad}} = \frac{2P_{\text{rad}}}{|I_2|^2}, \quad L = \frac{4\widetilde{W}_m}{|I_2|^2}, \quad C = \frac{|I_2|^2}{4\omega^2 \widetilde{W}_e}.$$

The physical implication of the equivalent circuit Figure 5.15(a) is very clear. The total power of the incident field, which spreads over all the space, is converted to a voltage source V_s. The energy radiated (or reradiated) away into infinity by the antenna system is represented by R_{rad}. The energy intercepted (or absorbed) by the receiving antenna, which is then dissipated in the load Z_L, is only a small part of the total incident energy when the receiving antenna is far from the transmitting antenna and it depends on the orientation of the antenna, the polarization of the incident wave and the impedance match of the receiving antenna system. If the receiving antenna is far from the transmitting antenna and its absorption cross section is finite, then the receiving antenna will never be able to catch all the incident energy.

The most well-known equivalent circuit for a receiving antenna is shown in Figure 5.15(b) and can be obtained directly from the network representation in Figure 5.14 by using the Norton theorem. This equivalent circuit can also be obtained by using a fundamental theorem derived by Collin (2003), which states that the total scattered field by a receiving

antenna can be expressed as

$$\begin{aligned}\mathbf{E}_s(Z_L) &= \mathbf{E}_s(Z_r) + \frac{I_r}{I_R}\frac{Z_L - Z_r}{Z_T + Z_L}\mathbf{E}_R,\\ \mathbf{H}_s(Z_L) &= \mathbf{H}_s(Z_r) + \frac{I_r}{I_R}\frac{Z_L - Z_r}{Z_T + Z_L}\mathbf{H}_R,\end{aligned} \tag{5.73}$$

where $\mathbf{E}_s(Z_L)$ is the scattered field by the receiving antenna that is terminated in a load impedance Z_L; $\mathbf{E}_s(Z_r)$ is the field scattered by the receiving antenna when it is terminated in a reference impedance Z_r and I_r is the current that flows into Z_r [see Figure 5.16(a)]; $\mathbf{E}_R$ is the field radiated by the receiving antenna for an input current I_R in the presence of the transmitting antenna with its source generator short-circuited ($V_g = 0$) and Z_T is the corresponding input impedance of the receiving antenna in this case [see Figure 5.16(b)]. Huygens' principle indicates that given a source inside a hypothetical surface S, there is a certain source spreading over S, which gives the same field outside S as the original source inside S.

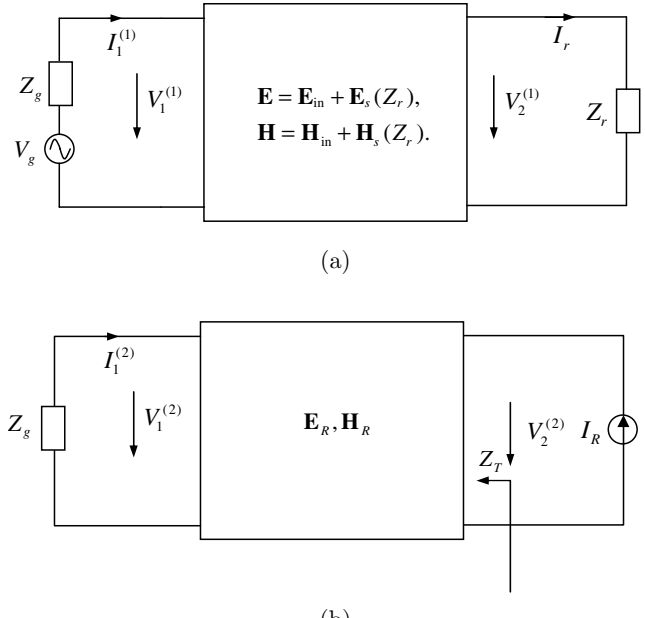


Figure 5.16 Two different field solutions.

In other words, an observer outside S will be hard to tell exactly whether the field is produced by the source inside the hypothetical surface or the source spreading over the surface. Therefore, it is not surprising that the equivalent circuits for a receiving antenna are not unique and their physical implications are difficult to interpret sometimes. The equivalent circuit in Figure 5.15(a) seems to be more convenient as its equivalent source V_s has a clear physical implication while the power generated by the equivalent source $V_{\rm oc}$ in Figure 5.15(b) is obscure. On the other hand, the equivalent circuit in Figure 5.15(b) reflects the fact that antenna impedance is reciprocal, i.e., when the transmitting antenna is far away from the receiving antenna the input impedance of the receiving antenna in the transmitting mode is equal to the impedance of the antenna in the receiving mode if the equivalent source $V_{\rm oc}$ is viewed as the excitation source.

The derivation of the relations in (5.73) is straightforward. We assume that the fields $[\mathbf{E}_s(Z_L), \mathbf{H}_s(Z_L)]$ produced by the two antenna system in a state of operation shown in Figure 5.14 can be expressed as the superposition of the two field solutions $[\mathbf{E}_s(Z_r), \mathbf{H}_s(Z_r)]$ and $(\mathbf{E}_R, \mathbf{H}_R)$ produced by the two antenna system in the states of operation shown in Figure 5.16

$$\begin{aligned} \mathbf{E}_s(Z_L) &= \mathbf{E}_s(Z_r) + \alpha \mathbf{E}_R, \\ \mathbf{H}_s(Z_L) &= \mathbf{H}_s(Z_r) + \alpha \mathbf{H}_R, \end{aligned} \tag{5.74}$$

where α is a constant to be determined. Clearly the fields $[\mathbf{E}_s(Z_L), \mathbf{H}_s(Z_L)]$ satisfy the boundary conditions on the antenna surfaces and the radiation condition at infinity since both $[\mathbf{E}_s(Z_r), \mathbf{H}_s(Z_r)]$ and $(\mathbf{E}_R, \mathbf{H}_R)$ satisfy these boundary conditions. On the feeding plane T_1, we have

$$V_1 = V_1^{(1)} + \alpha V_1^{(2)} = V_g - I_1^{(1)} Z_g - \alpha I_1^{(2)} Z_g = V_g - (I_1^{(1)} + \alpha I_1^{(2)}) Z_g.$$

According to the second expression of (5.74), we have $I_1 = I_1^{(1)} + \alpha I_1^{(2)}$. As a result, the above equation can be written as

$$V_1 = V_g - I_1 Z_g,$$

which indicates that the boundary condition on the feeding plane T_1 is satisfied. In order to satisfy the boundary condition on the feeding plane T_2, we must have

$$V_2 = I_2 Z_L = (I_r - \alpha I_R) Z_L = V_2^{(1)} + \alpha V_2^{(2)} = I_r Z_r + \alpha I_R Z_T,$$

which can be used to solve for α

$$\alpha = \frac{I_r}{I_R}\frac{Z_L - Z_r}{Z_T + Z_L}.$$

If we choose $Z_r = 0$, (5.73) becomes

$$\begin{aligned}\mathbf{E}_s(Z_L) &= \mathbf{E}_s(Z_r = 0) + \frac{I_r}{I_R}\frac{Z_L}{Z_T + Z_L}\mathbf{E}_R,\\ \mathbf{H}_s(Z_L) &= \mathbf{H}_s(Z_r = 0) + \frac{I_r}{I_R}\frac{Z_L}{Z_T + Z_L}\mathbf{H}_R.\end{aligned} \tag{5.75}$$

This is a fundamental theorem in antenna scattering and has been derived by many authors (Aharoni, 1946; King and Harrison, 1944; Stevenson, 1948; Collin, 1968).

If we choose $Z_r = Z_c$, where Z_c represents the characteristic impedance of the transmission line that connects the receiving antenna to its load termination, (5.73) becomes

$$\begin{aligned}\mathbf{E}_s(Z_L) &= \mathbf{E}_s(Z_c) + \frac{I_r}{I_R}\frac{Z_L - Z_c}{Z_T + Z_L}\mathbf{E}_R,\\ \mathbf{H}_s(Z_L) &= \mathbf{H}_s(Z_c) + \frac{I_r}{I_R}\frac{Z_L - Z_c}{Z_T + Z_L}\mathbf{H}_R.\end{aligned} \tag{5.76}$$

If the load impedance is equal to the characteristic impedance, the scattered fields are given by only the first term on the right-hand side of (5.76). In this case, the antenna does not radiate and the term $\mathbf{E}_s(Z_c)$ represents the scattering from the antenna structure with no radiation mode component present, which may be called the **intrinsic scattered field** as proposed by Collin.

If we choose $Z_r = \bar{Z}_T$, (5.73) becomes

$$\begin{aligned}\mathbf{E}_s(Z_L) &= \mathbf{E}_s(\bar{Z}_T) + \frac{I_r}{I_R}\frac{Z_L - \bar{Z}_T}{Z_L + Z_T}\mathbf{E}_R,\\ \mathbf{H}_s(Z_L) &= \mathbf{H}_s(\bar{Z}_T) + \frac{I_r}{I_R}\frac{Z_L - \bar{Z}_T}{Z_L + Z_T}\mathbf{H}_R.\end{aligned} \tag{5.77}$$

If the antenna load impedance is equal to the complex conjugate of the antenna input impedance the scattered fields are given by only the first term on the right-hand side of (5.77), which are called **structural scattered fields** as proposed by Green (1963). In general, the scattered field from an

antenna will have a radiation pattern that can be very different from the radiation pattern when the antenna is used to transmit.

5.4.4 Foster Reactance Theorem for Lossless Antennas

The Foster theorems, named after the American mathematician Ronald Martin Foster (1896–1998), state that the slope of the reactance curve or susceptance curve as a function of frequency is always positive for a lossless circuit. Although the Foster theorems are typically stated for a lossless network in textbooks, they can be generalized to a lossy network in numerous situations. For example, the Foster reactance theorem holds for a simple series RLC circuit or any lossy network that consists of a resistor connected in series to a lossless network. From the viewpoint of circuit theory, an **ideal antenna**, defined as an antenna without Ohmic loss, is a one-port lossy network with radiation loss only. By using the complex frequency domain approach, the Foster theorems can be shown to hold for an ideal antenna (Geyi *et al.*, 2000; Geyi, 2007a).

To prove that the Foster theorem holds for an ideal antenna, we introduce the complex frequency $s = \alpha + j\omega$ and all calculations are confined to the complex frequency plane. For clarity, all quantities in the complex frequency domain are denoted by the same symbols in real frequency domain but explicitly showing the dependence on s. Taking the Laplace transform of the time-domain Maxwell's equations in a lossless medium yields

$$\nabla \times \mathbf{E}(\mathbf{r}, s) = -s\mu\mathbf{H}(\mathbf{r}, s), \quad \nabla \times \mathbf{H}(\mathbf{r}, s) = s\varepsilon\mathbf{E}(\mathbf{r}, s). \tag{5.78}$$

The frequency-domain quantities can be recovered by letting $\alpha = 0$ in (5.78). From (5.78) a relation similar to (5.51) can be obtained in the region outside V_0

$$\nabla \cdot \left[\frac{1}{2}\mathbf{E}(\mathbf{r}, s) \times \bar{\mathbf{H}}(\mathbf{r}, s)\right] = -\frac{1}{2}\alpha\left[\mu|\mathbf{H}(\mathbf{r}, s)|^2 + \varepsilon|\mathbf{E}(\mathbf{r}, s)|^2\right] - j\frac{1}{2}\omega\left[\mu|\mathbf{H}(\mathbf{r}, s)|^2 - \varepsilon|\mathbf{E}(\mathbf{r}, s)|^2\right]. \tag{5.79}$$

Taking the integration of (5.79) over the connected region $V_\infty - V_0$, as shown in Figure 5.11, gives

$$\int_{\partial V_0 + S_\infty} \frac{1}{2}[\mathbf{E}(\mathbf{r}, s) \times \bar{\mathbf{H}}(\mathbf{r}, s)] \cdot \mathbf{u}_n \, dS = -2\alpha[W_m(s) + W_e(s)] - 2j\omega[W_m(s) - W_e(s)], \tag{5.80}$$

where

$$W_m(s) = \frac{1}{4} \int\limits_{V_\infty - V_0} \mu |\mathbf{H}(\mathbf{r}, s)|^2 dV(\mathbf{r}),$$

$$W_e(s) = \frac{1}{4} \int\limits_{V_\infty - V_0} \varepsilon |\mathbf{E}(\mathbf{r}, s)|^2 dV(\mathbf{r}).$$

We assume again that the antenna reference plane is away from the antenna discontinuity so that the higher-order modes excited by the discontinuity have negligible effects at the reference plane. Thus for a single-mode feeding waveguide, we can make the following approximation

$$\frac{1}{2} \int\limits_{\partial V_0} [\mathbf{E}(\mathbf{r}, s) \times \bar{\mathbf{H}}(\mathbf{r}, s)] \cdot \mathbf{u}_n \, dS(\mathbf{r}) = -\frac{1}{2} V(s) \bar{I}(s). \tag{5.81}$$

Letting $P_{\text{rad}}(s) = \frac{1}{2} \int_{S_\infty} [\mathbf{E}(\mathbf{r}, s) \times \bar{\mathbf{H}}(\mathbf{r}, s)] \cdot \mathbf{u}_n \, dS(\mathbf{r})$ and substituting it into (5.80), we get

$$\frac{1}{2} V(s) \bar{I}(s) = P_{\text{rad}}(s) + 2\alpha [W_m(s) + W_e(s)] + 2j\omega [W_m(s) - W_e(s)]. \tag{5.82}$$

We now choose the observation $\mathbf{r}$ such that it is located in the far-field region of the antenna. As indicated by (5.66), the stored energies will become a constant if $r = |\mathbf{r}|$ is big enough (say, $\mathbf{r}$ is in the far-field region). Since $v = 3 \times 10^8$ m/s in free space, we have $r/v \ll 1$ for any practical antenna system with r in the far-field region. If α is sufficiently small, we can make a first order approximation $e^{-\alpha r/c} \approx 1 - \alpha r/v$, and derive directly from the Maxwell equations, defined in the complex plane, the following:

$$\mathbf{E}_{\text{rad}}(\mathbf{r}, s) \approx -\left(1 - r\frac{\alpha}{v}\right) \frac{j\omega\mu}{4\pi r} e^{-jkr} \int\limits_{V_0} [\mathbf{J}(\mathbf{r}', s) - \mathbf{J}(\mathbf{r}', s) \cdot \mathbf{u}_r] e^{-jk\mathbf{u}_r \cdot \mathbf{r}'} dV(\mathbf{r}'),$$

$$\mathbf{H}_{\text{rad}}(\mathbf{r}, s) \approx -\eta \left(1 - r\frac{\alpha}{v}\right) \frac{j\omega\varepsilon}{4\pi r} e^{-jkr} \int\limits_{V_0} \mathbf{u}_r \times \mathbf{J}(\mathbf{r}', s) e^{-jk\mathbf{u}_r \cdot \mathbf{r}'} dV(\mathbf{r}').$$

Hence

$$P_{\text{rad}}(s) = P_{\text{rad}}(\omega)(1 - r\alpha/v)^2 \approx (1 - 2r\alpha/v) P_{\text{rad}}(\omega) \tag{5.83}$$

where $P_{\text{rad}}(\omega)$, previously defined in (5.58), is the radiated power in the frequency domain, which is independent of α. Substituting (5.83) into

(5.82), we obtain

$$\frac{1}{2}V(s)\bar{I}(s) = P_{\rm rad}(\omega) + 2\alpha\left[W_m(s) + W_e(s) - \frac{r}{v}P_{\rm rad}(\omega)\right] + 2j\omega[W_m(s) - W_e(s)]. \tag{5.84}$$

The impedance and admittance in the complex frequency plane can then be expressed as

$$Z(s) = \frac{2P_{\rm rad}(\omega)}{|I(s)|^2} + \frac{4\alpha}{|I(s)|^2}\left[W_m(s) + W_e(s) - \frac{r}{v}P_{\rm rad}(\omega)\right] + \frac{4j\omega}{|I(s)|^2}[W_m(s) - W_e(s)]. \tag{5.85}$$

Similarly, we can introduce the stored energies in the complex frequency domain

$$\widetilde{W}_m(s) + \widetilde{W}_e(s) = W_m(s) + W_e(s) - \frac{r}{v}P_{\rm rad}(\omega),$$
$$\widetilde{W}_m(s) - \widetilde{W}_e(s) = W_m(s) - W_e(s),$$

and rewrite (5.85) as

$$Z(s) = \frac{2P_{\rm rad}(\omega)}{|I(s)|^2} + \frac{4s\widetilde{W}_m(s)}{|I(s)|^2} + \frac{4\bar{s}\widetilde{W}_e(s)}{|I(s)|^2}. \tag{5.86}$$

To get rid of the complex conjugation $\bar{s}$, we may introduce a new quantity $\widetilde{W}'_e(s) = |s|^2\widetilde{W}_e(s)$. Now (5.86) can be extended to an analytic function of s by replacing all complex conjugations $\bar{s}$ in $\widetilde{W}_m(s)$ and $\widetilde{W}'_e(s)$ with $-s$, and $j\omega$ (resp. $-j\omega$) by s (resp. $-s$) in $P_{\rm rad}(\omega)$. Thus (5.86) become analytic and can be written as

$$Z(s) = \frac{2P_{\rm rad}(s)}{I(s)I(-s)} + \frac{4s\widetilde{W}_m(s)}{I(s)I(-s)} + \frac{4s^{-1}\widetilde{W}'_e(s)}{I(s)I(-s)}. \tag{5.87}$$

Note that (5.86) and (5.87) are identical when $\alpha = 0$. If α is assumed to be small, a Taylor expansion may be assumed for an arbitrary analytic function $\mathbf{A}(s)$ so that

$$\mathbf{A}(s)\cdot\mathbf{A}(-s) = |\mathbf{A}(j\omega)|^2 + j\alpha T(\omega) + o(\alpha),$$

where $T(\omega)$ is a real function of ω. When this relation is used in (5.87) and use is made of the following decompositions

$$Z(s) = R(\alpha,\omega) + jX(\alpha,\omega), \tag{5.88}$$

we may find that

$$R(\alpha,\omega)=\frac{2P_{\rm rad}}{|I|^2}+\frac{4\alpha}{|I|^2}(\widetilde{W}_m+\widetilde{W}_e), \tag{5.89}$$

where the power, energies, voltage and current are all calculated at $\alpha=0$. Since $Z(s)$ is an analytic function, its real and imaginary parts satisfy the Cauchy–Riemann conditions

$$\frac{\partial R(\alpha,\omega)}{\partial\alpha}=\frac{\partial X(\alpha,\omega)}{\partial\omega},\quad \frac{\partial R(\alpha,\omega)}{\partial\omega}=-\frac{\partial X(\alpha,\omega)}{\partial\alpha}. \tag{5.90}$$

By direct calculation, we have

$$\left.\frac{\partial R(\alpha,\omega)}{\partial\alpha}\right|_{\alpha=0}=\frac{4(\widetilde{W}_m+\widetilde{W}_e)}{|I|^2}. \tag{5.91}$$

From (5.90) and (5.91), we obtain

$$\left.\frac{\partial X}{\partial\omega}\right|_{\alpha=0}=\frac{4}{|I|^2}(\widetilde{W}_m+\widetilde{W}_e)>0. \tag{5.92}$$

This is the Foster theorem for a lossless antenna system, which indicates that the slope of the reactance curve as a function of the frequency for an ideal antenna is always positive. From (5.92), we obtain the stored magnetic and electric field energies

$$\widetilde{W}_e=\frac{1}{8}|I|^2\left(\frac{\partial X}{\partial\omega}-\frac{X}{\omega}\right),\quad \widetilde{W}_m=\frac{1}{8}|I|^2\left(\frac{\partial X}{\partial\omega}+\frac{X}{\omega}\right). \tag{5.93}$$

Note that the reactance X can be written as

$$X=\frac{4\omega}{|I|^2}(\widetilde{W}_m-\widetilde{W}_e). \tag{5.94}$$

Equations (5.93) was used by Harrington to study antenna Q and bandwidth in 1968 in his book *Field Computation by Moment Methods* (Harrington, 1968) although a rigorous proof of Foster reactance theorem for a radiating system was not available at the time. Harrington believed that (5.93) is approximately valid for a high Q network in the vicinity of resonance. In a discussion with Collin, Rhodes also believed that the fact that the slope of the input reactance can be negative at some frequencies is immaterial; it is always positive at the only frequency (resonance) for which bandwidth and Q are defined (Collin, 1967; Rhodes, 1967).

Example 5.2 (Demonstration of Foster reactance theorem for lossless antennas)**:** The Foster reactance theorem is valid for an ideal metal antenna, subject to the following conditions: (a) The antenna consists of perfect conductor and the surrounding material is lossless; (b) The antenna is fed by a waveguide which is assumed to be in the state of single-mode operation, and the antenna input terminal is positioned in the single-mode region of the feeding waveguide. Equation (5.92) indicates that the slope of the reactance of the antenna must be great than zero. This result has been the main cause to provoke argument. Some typical numerical examples will now be presented to validate the Foster reactance theorem for ideal (lossless) antennas.

The reactance curves of a dipole fed by two-wire transmission line, a coaxial aperture and a monopole antenna fed by coaxial cable, and rectangular aperture antenna fed by waveguide are shown in Figures 5.17–5.20. It can be seen that Foster reactance theorem holds very well in the frequency range between the cut-off frequency f_c of the dominant mode and the cut-off frequency f_{c1} of the first higher order mode (the feeding waveguide is assumed to be in a single-mode operation).

Although a feeding waveguide is assumed in the proof of the Foster reactance theorem, the Foster reactance theorem is also valid for point-fed antennas. As pointed out in Balanis (2005), the delta-gap source modeling is the simplest and most widely used, but it is also the least accurate, especially for impedances. Theoretically, the delta-gap source is only valid

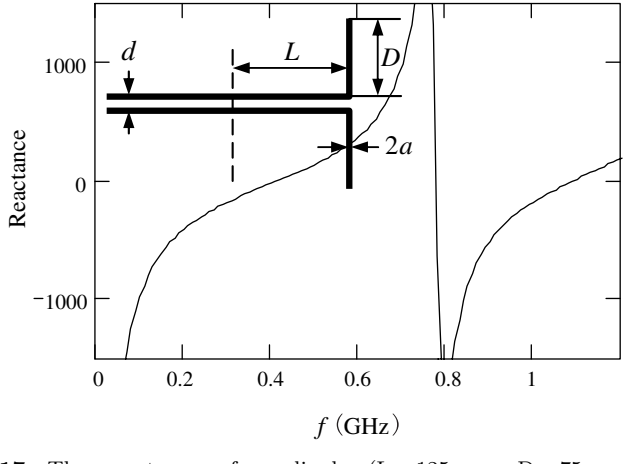


Figure 5.17 The reactance of a dipole (L = 125 mm, D = 75 mm, d = 5 mm, a = 0.5 mm).

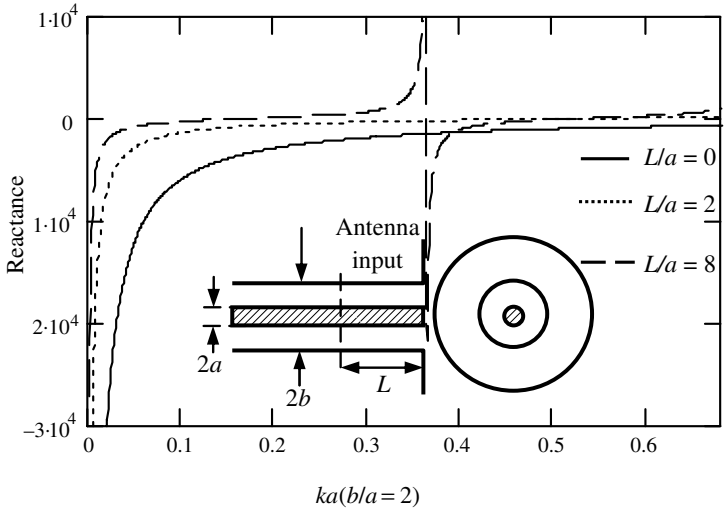


Figure 5.18 Reactance of the coaxial aperture antenna with infinite flange.

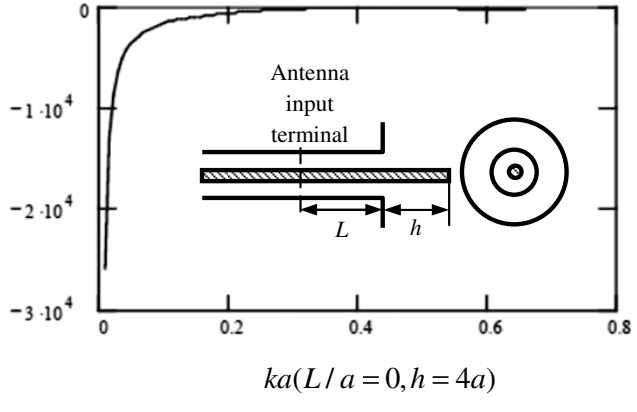


Figure 5.19 Reactance of a cylindrical monopole antenna.

for infinitely thin wire antennas. For thick wires, the delta-gap source modeling produces reasonable results for the impedance only when the calculation is limited to the low frequency range. This observation has been widely ignored.

The reactance curves of the thin dipole antenna, loop antenna and folded dipole antenna are shown in Figures 5.21–5.23. It can be seen that the Foster reactance theorem holds very well. Note that the reactance curves not only have some zeros (resonant frequencies) where the stored electric

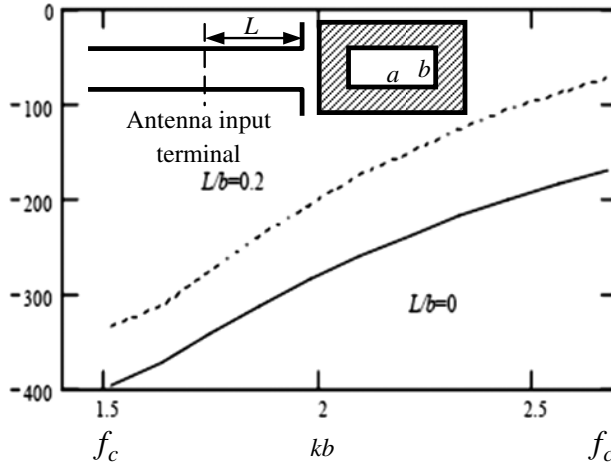


Figure 5.20 Reactance of a rectangular aperture.

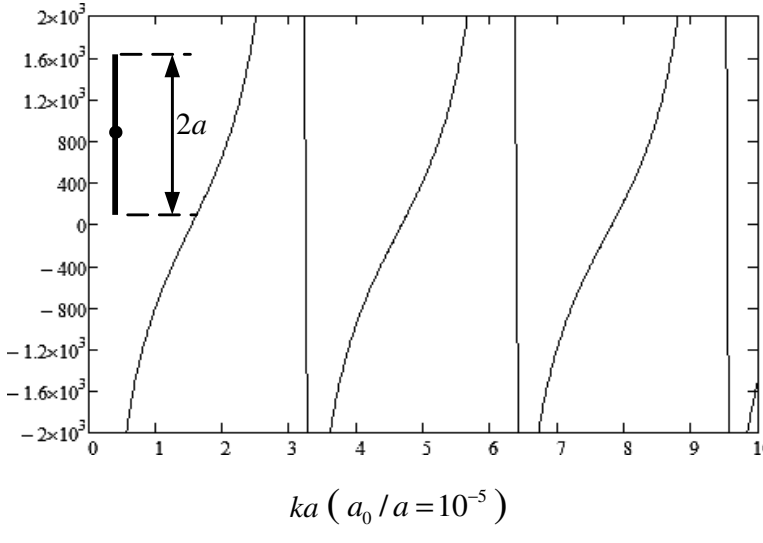


Figure 5.21 Reactance of a dipole antenna.

energy equals the stored magnetic energy, but also have some singularities where the slope of the reactance curve becomes infinite caused by the zeros of the input current. A negative slope may occur around these singularities if either the delta-gap source is inappropriately applied or heat loss is present (such as in the measurements). However, the slope of reactance is always

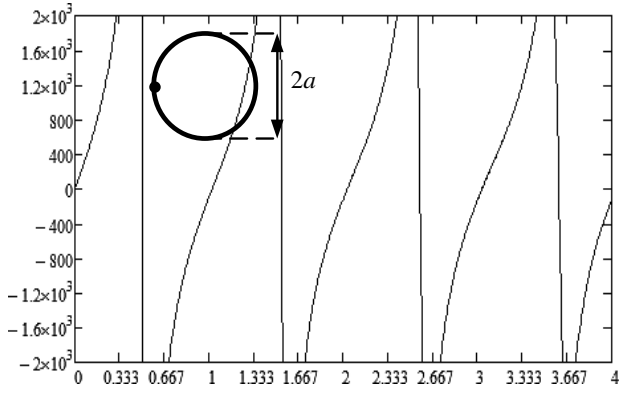


Figure 5.22 Reactance of a loop antenna.

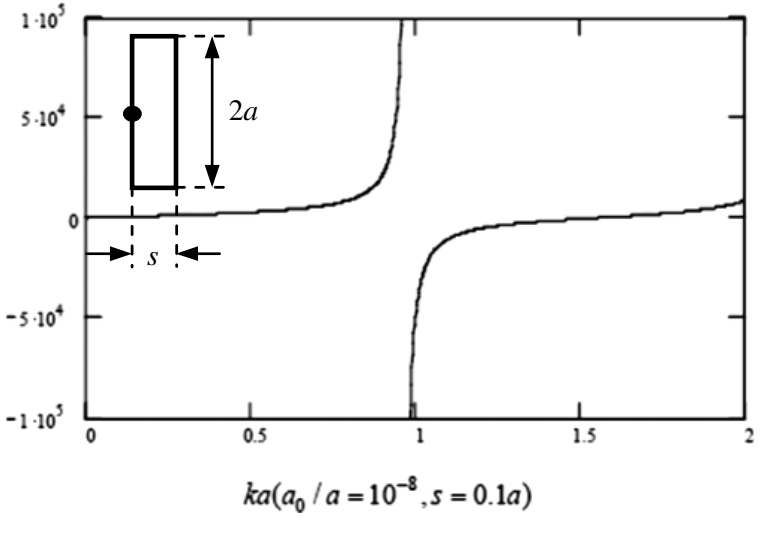


Figure 5.23 Reactance of a folded dipole.

positive in the vicinity of resonant frequency as noted by Rhodes (1967). Since the bandwidth and Q are defined at resonant frequency, the Foster reactance theorem is always approximately valid for a lossy system and can be applied to study antenna Q and bandwidth, as claimed by Harrington (1968). □

Remark 5.2: From the Poynting theorem in frequency domain, we may obtain (Geyi, 2010)

$$P^{\text{rad}} = \frac{\omega\eta v}{8\pi}\int\limits_{V_0}\int\limits_{V_0}\left[\frac{1}{v^2}\frac{\bar{\mathbf{J}}(\mathbf{r})\cdot\mathbf{J}(\mathbf{r}')}{R} - \frac{\bar{\rho}(\mathbf{r})\rho(\mathbf{r}')}{R}\right]\sin(kR)dV(\mathbf{r})dV(\mathbf{r}'), \tag{5.95}$$

$$\widetilde{W}_m - \widetilde{W}_e = \frac{\eta v}{16\pi}\int\limits_{V_0}\int\limits_{V_0}\left[\frac{1}{v^2}\frac{\bar{\mathbf{J}}(\mathbf{r})\cdot\mathbf{J}(\mathbf{r}')}{R} - \frac{\bar{\rho}(\mathbf{r})\rho(\mathbf{r}')}{R}\right]\cos(kR)dV(\mathbf{r})dV(\mathbf{r}'), \tag{5.96}$$

where V_0 stands for the source region that the induced current $\mathbf{J}$ occupies. Once the current distribution $\mathbf{J}$ and the input terminal current I are known, the stored energies and thus the Q can be determined from Equations (5.93)–(5.96). It follows from (5.94) that

$$\frac{\partial X}{\partial\omega} = \frac{4(\widetilde{W}_m - \widetilde{W}_e)}{|I|^2} + \omega\frac{\partial}{\partial\omega}\left[\frac{4(\widetilde{W}_m - \widetilde{W}_e)}{|I|^2}\right]. \tag{5.97}$$

Considering (5.96) and assuming that the normalized current distribution $\mathbf{J}(r)/I$ is independent of frequency, we have

$$\begin{aligned}\frac{\partial}{\partial\omega}\frac{\widetilde{W}_m - \widetilde{W}_e}{|I|^2} = {}& \frac{\eta v}{8\pi}\int\limits_{V_0}\int\limits_{V_0}\frac{\nabla\cdot\bar{\mathbf{J}}(\mathbf{r})\nabla\cdot\mathbf{J}(\mathbf{r}')}{\omega^3|I|^2R}\cos(kR)dV(\mathbf{r})dV(\mathbf{r}') \\ & -\frac{\eta}{16\pi}\int\limits_{V_0}\int\limits_{V_0}\left[\frac{\bar{\mathbf{J}}(\mathbf{r})\cdot\mathbf{J}(\mathbf{r}')}{v^2|I|^2} - \frac{\nabla\cdot\bar{\mathbf{J}}(\mathbf{r})\nabla\cdot\mathbf{J}(\mathbf{r}')}{\omega^2|I|^2}\right] \\ & \times\sin(kR)dV(\mathbf{r})dV(\mathbf{r}').\end{aligned} \tag{5.98}$$

Taking (5.98) into account and substituting (5.97) into (5.93), we immediately get

$$\begin{aligned}\widetilde{W}_e = {}& \frac{\eta v}{16\pi}\int\limits_{V_0}\int\limits_{V_0}\bar{\rho}(\mathbf{r})\rho(\mathbf{r}')\frac{\cos kR}{R}\,dV(\mathbf{r})dV(\mathbf{r}') \\ & +\frac{\omega\eta}{32\pi}\int\limits_{V_0}\int\limits_{V_0}\left[\bar{\rho}(\mathbf{r})\rho(\mathbf{r}') - \frac{1}{v^2}\bar{\mathbf{J}}(\mathbf{r})\cdot\mathbf{J}(\mathbf{r}')\right]\sin kR\,dV(\mathbf{r})dV(\mathbf{r}'),\end{aligned} \tag{5.99}$$

$$\widetilde{W}_m = \frac{\eta v}{16\pi}\int\limits_{V_0}\int\limits_{V_0}\frac{1}{v^2}\bar{\mathbf{J}}(\mathbf{r})\cdot\mathbf{J}(\mathbf{r}')\frac{\cos kR}{R}\,dV(\mathbf{r})dV(\mathbf{r}')$$
$$+\frac{\omega\eta}{32\pi}\int\limits_{V_0}\int\limits_{V_0}\left[\bar{\rho}(\mathbf{r})\rho(\mathbf{r}')-\frac{1}{v^2}\bar{\mathbf{J}}(\mathbf{r})\cdot\mathbf{J}(\mathbf{r}')\right]\sin kR\,dV(\mathbf{r})dV(\mathbf{r}') \tag{5.100}$$

these results, implied by the Foster reactance theorem, have been obtained by Vandenbosch in a rather involved way (Vandenbosch, 2010). In deriving them, an assumption that the normalized current distribution $\mathbf{J}(r)/I$ is independent of frequency has been applied. This assumption is, however, generally not true for electrically large antennas. To get rid of the assumption, a more rigorous approach will be introduced below. □

5.4.5 Quality Factor and Bandwidth

The evaluation of the antenna quality factor can be traced back to the classical work of Chu, who derived the theoretical value of Q for an ideal antenna enclosed in a circumscribing sphere (Chu, 1948). Chu's analysis is based on the spherical mode expansions and is only valid for an omni-directional antenna that radiates either TE or TM modes. In order to avoid the difficulty that the total electric and magnetic field energies are infinite, Chu introduced the equivalent impedance for each mode and obtained an expression of antenna Q through the calculation of stored energies in the truncated equivalent ladder circuit for the impedance. Collin and Rothschild used a method for evaluating antenna Q (Collin and Rothschild, 1964) based on the idea proposed by Counter (1948) that the total stored energy can be calculated by subtracting the radiated field energy away from the total energy in the fields. Such method has been successfully used by Fante (1969) and re-examined by McLean (1996) to study the antenna Q. All these studies only utilize the stored energies outside the circumscribing sphere of the antenna in the calculation of antenna Q, which is much smaller than Q_{real} (Collardey *et al.*, 2005; 2006).

5.4.5.1 *Evaluation of Q from Input Impedance*

From the stored energies expressed by (5.93), the antenna quality factor may be written as

$$Q_{\text{real}} = \frac{1}{2R_{\text{rad}}}\omega\frac{\partial X}{\partial\omega}, \tag{5.101}$$

if (5.11) is used; or

$$Q_{\text{real}} = \frac{1}{2R_{\text{rad}}}\left(\omega\frac{dX}{d\omega} \pm X\right) \tag{5.102}$$

where either + or − is chosen to give the higher Q, if (5.12) is used. Thus once the antenna input impedance is known, the quality factor may be determined from (5.101) or (5.102).

5.4.5.2 *Evaluation of Q from Current Distribution*

Let us consider an arbitrary current distribution $\mathbf{J}$, which occupies a finite region V_0 bounded by ∂V_0 as shown in Figure 5.10. The current distribution produces electric field $\mathbf{E}$ and magnetic field $\mathbf{H}$. Let S be a closed surface large enough to enclose the source region V_0. We introduce the complex frequency $s = \alpha + j\omega$ and all calculations are confined to the complex frequency plane. Then Equation (5.79) applies. Taking the integration of (5.79) over the region V bounded by S gives

$$-\int_{V_0}\frac{1}{2}\mathbf{E}(\mathbf{r},s)\cdot\bar{\mathbf{J}}(\mathbf{r},s)dV(\mathbf{r}) = P_{\text{rad}}(s) + 2\alpha[W_m(s)+W_e(s)] + 2j\omega[W_m(s)-W_e(s)], \tag{5.103}$$

where

$$W_m(s) = \frac{1}{4}\int_V \mu|\mathbf{H}(\mathbf{r},s)|^2 dV(\mathbf{r}), \quad W_e(s) = \frac{1}{4}\int_V \varepsilon|\mathbf{E}(\mathbf{r},s)|^2 dV(\mathbf{r}).$$

$$P_{\text{rad}}(s) = \frac{1}{2}\int_S [\mathbf{E}(\mathbf{r},s)\times\bar{\mathbf{H}}(\mathbf{r},s)]\cdot\mathbf{u}_n\, dS(\mathbf{r}).$$

Similarly (5.83) holds. Substituting it into (5.103), we obtain

$$-\int_{V_0}\frac{1}{2}\bar{\mathbf{J}}(\mathbf{r},s)\cdot\mathbf{E}(\mathbf{r},s)dV(\mathbf{r}) = P_{\text{rad}}(\omega) + 2\alpha\left[\widetilde{W}_m(s)+\widetilde{W}_e(s)\right] + 2j\omega\left[\widetilde{W}_m(s)-\widetilde{W}_e(s)\right]. \tag{5.104}$$

In the complex frequency domain, the fields produced by the current $\mathbf{J}(\mathbf{r},s)$ can be represented by

$$\mathbf{E}(\mathbf{r},s) = -\eta v\nabla\int_{V_0}\frac{\rho(\mathbf{r}',s)}{4\pi R}e^{-\frac{sR}{v}}\,dV(\mathbf{r}') - \frac{\eta}{v}s\int_{V_0}\frac{\mathbf{J}(\mathbf{r}',s)}{4\pi R}e^{-\frac{sR}{v}}\,dV(\mathbf{r}'). \tag{5.105}$$

Introducing (5.105) into (5.104), we obtain

$$-\frac{1}{2}\int\limits_{V_0}\bar{\mathbf{J}}(\mathbf{r},s)\cdot\mathbf{E}(\mathbf{r},s)dV(\mathbf{r})=\eta v\bar{s}\int\limits_{V_0}\int\limits_{V_0}\frac{\bar{\rho}(\mathbf{r},s)\rho(\mathbf{r}',s)}{8\pi R}e^{-\frac{sR}{v}}\,dV(\mathbf{r})dV(\mathbf{r}')$$
$$+\frac{\eta}{v}s\int\limits_{V_0}\int\limits_{V_0}\frac{\bar{\mathbf{J}}(\mathbf{r},s)\cdot\mathbf{J}(\mathbf{r}',s)}{8\pi R}e^{-\frac{sR}{v}}\,dV(\mathbf{r})dV(\mathbf{r}'). \tag{5.106}$$

It follows from (5.104) and (5.106) that

$$\frac{\eta v}{16\pi}\int\limits_{V_0}\int\limits_{V_0}\frac{\bar{\rho}(\mathbf{r},s)\rho(\mathbf{r}',s)}{R}\bar{s}e^{-\frac{sR}{v}}\,dV(\mathbf{r})dV(\mathbf{r}')$$
$$+\frac{\eta v}{16\pi}\int\limits_{V_0}\int\limits_{V_0}\frac{1}{v^2}\frac{\bar{\mathbf{J}}(\mathbf{r},s)\cdot\mathbf{J}(\mathbf{r}',s)}{R}se^{-\frac{sR}{v}}\,dV(\mathbf{r})dV(\mathbf{r}')$$
$$=\frac{1}{2}P_{\mathrm{rad}}(\omega)+\alpha\left[\widetilde{W}_m(s)+\widetilde{W}_e(s)\right]+j\omega\left[\widetilde{W}_m(s)-\widetilde{W}_e(s)\right]. \tag{5.107}$$

For arbitrary analytic functions $a(s)=a_r(\alpha,\omega)+ja_i(\alpha,\omega)$ and $b(s)=b_r(\alpha,\omega)+jb_i(\alpha,\omega)$ the Cauchy–Riemann conditions hold

$$\begin{aligned}&\frac{\partial a_r(\alpha,\omega)}{\partial\alpha}=\frac{\partial a_i(\alpha,\omega)}{\partial\omega},\quad \frac{\partial a_i(\alpha,\omega)}{\partial\alpha}=-\frac{\partial a_r(\alpha,\omega)}{\partial\omega};\\ &\frac{\partial b_r(\alpha,\omega)}{\partial\alpha}=\frac{\partial b_i(\alpha,\omega)}{\partial\omega},\quad \frac{\partial b_i(\alpha,\omega)}{\partial\alpha}=-\frac{\partial b_r(\alpha,\omega)}{\partial\omega}.\end{aligned} \tag{5.108}$$

These relations imply

$$\frac{\partial\bar{a}(s)}{\partial\alpha}=j\frac{\partial\bar{a}(s)}{\partial\omega},\quad \frac{\partial\bar{a}(s)}{\partial\alpha}=j\frac{\partial\bar{a}(s)}{\partial\omega}.$$

The function $\bar{a}(s)b(s)$ may be expanded into a Taylor series at $\alpha=0$

$$\bar{a}(s)b(s)\approx\bar{a}(j\omega)b(j\omega)+\alpha[\Omega_1(j\omega)+j\Omega_2(j\omega)]+o(\alpha),$$

where the Cauchy–Riemann conditions (5.108) have been used, and

$$\Omega_1(j\omega)=\mathrm{Re}\left[j\frac{\partial\bar{a}(j\omega)}{\partial\omega}b(j\omega)-j\bar{a}(j\omega)\frac{\partial b(j\omega)}{\partial\omega}\right],$$
$$\Omega_2(j\omega)=\mathrm{Im}\left[j\frac{\partial\bar{a}(j\omega)}{\partial\omega}b(j\omega)-j\bar{a}(j\omega)\frac{\partial b(j\omega)}{\partial\omega}\right].$$

For small α, we may use the following approximations

$$e^{-\frac{sR}{v}} \approx (\cos kR - j\sin kR) - \alpha\frac{R}{v}(\cos kR - j\sin kR).$$

Thus we have

$$\begin{aligned}
&\bar{a}(s)b(s)\bar{s}e^{-\frac{sR}{v}} \\
&\quad = \alpha\bar{a}(j\omega)b(j\omega)\cos kR - j\alpha\bar{a}(j\omega)b(j\omega)\sin kR \\
&\qquad -j\omega\bar{a}(j\omega)b(j\omega)\cos kR - \omega\bar{a}(j\omega)b(j\omega)\sin kR \\
&\qquad + j\omega\alpha\frac{R}{v}\bar{a}(j\omega)b(j\omega)\cos kR + \alpha\omega\frac{R}{v}\bar{a}(j\omega)b(j\omega)\sin kR \\
&\qquad - j\omega\alpha\Omega_1(j\omega)\cos kR - \alpha\omega\Omega_1(j\omega)\sin kR \\
&\qquad + \alpha\omega\Omega_2(j\omega)\cos kR - j\omega\alpha\Omega_2(j\omega)\sin kR + o(\alpha), \qquad (5.109)
\end{aligned}$$

and

$$\begin{aligned}
&\bar{a}(s)b(s)se^{-\frac{sR}{v}} \\
&\quad = \alpha\bar{a}(j\omega)b(j\omega)\cos kR - j\alpha\bar{a}(j\omega)b(j\omega)\sin kR \\
&\qquad + j\omega\bar{a}(j\omega)b(j\omega)\cos kR + \omega\bar{a}(j\omega)b(j\omega)\sin kR \\
&\qquad - j\omega\alpha\frac{R}{c}\bar{a}(j\omega)b(j\omega)\cos kR - \alpha\omega\frac{R}{v}\bar{a}(j\omega)b(j\omega)\sin kR \\
&\qquad + j\omega\alpha\Omega_1(j\omega)\cos kR + \alpha\omega\Omega_1(j\omega)\sin kR \\
&\qquad - \alpha\omega\Omega_2(j\omega)\cos kR + j\omega\alpha\Omega_2(j\omega)\sin kR + o(\alpha). \qquad (5.110)
\end{aligned}$$

It follows from (5.107), (5.109) and (5.110) that

$$\begin{aligned}
P_{\mathrm{rad}}(\omega) &= \frac{\omega\eta c}{8\pi}\int_{V_0}\int_{V_0}\left[\frac{1}{v^2}\bar{\mathbf{J}}(\mathbf{r})\cdot\mathbf{J}(\mathbf{r}') - \bar{\rho}(\mathbf{r})\rho(\mathbf{r}')\right] \\
&\qquad \times\frac{\sin kR}{R}\,dV(\mathbf{r})dV(\mathbf{r}'), \qquad (5.111) \\
\widetilde{W}_m - \widetilde{W}_e &= \frac{\eta v}{16\pi}\int_{V_0}\int_{V_0}\left[\frac{1}{v^2}\bar{\mathbf{J}}(\mathbf{r})\cdot\mathbf{J}(\mathbf{r}') - \bar{\rho}(\mathbf{r})\rho(\mathbf{r}')\right] \\
&\qquad \times\frac{\cos kR}{R}\,dV(\mathbf{r})dV(\mathbf{r}'), \qquad (5.112)
\end{aligned}$$

$$\widetilde{W}_m + \widetilde{W}_e = \frac{\eta v}{16\pi} \int\limits_{V_0} \int\limits_{V_0} \frac{\bar{\rho}(\mathbf{r})\rho(\mathbf{r}')}{R} \cos kR \, dV(\mathbf{r}) dV(\mathbf{r}')$$

$$+ \frac{\eta v}{16\pi} \int\limits_{V_0} \int\limits_{V_0} \omega \frac{R}{v} \frac{\bar{\rho}(\mathbf{r})\rho(\mathbf{r}')}{R} \sin kR \, dV(\mathbf{r}) dV(\mathbf{r}')$$

$$- \frac{\eta v}{16\pi} \int\limits_{V_0} \int\limits_{V_0} \frac{\omega \Omega_{\rho 1}(j\omega)}{R} \sin kR \, dV(\mathbf{r}) dV(\mathbf{r}')$$

$$+ \frac{\eta v}{16\pi} \int\limits_{V_0} \int\limits_{V_0} \frac{\omega \Omega_{\rho 2}(j\omega)}{R} \cos kR \, dV(\mathbf{r}) dV(\mathbf{r}')$$

$$+ \frac{\eta v}{16\pi} \int\limits_{V_0} \int\limits_{V_0} \frac{1}{v^2} \frac{\bar{\mathbf{J}}(\mathbf{r}) \cdot \mathbf{J}(\mathbf{r}')}{R} \cos kR \, dV(\mathbf{r}) dV(\mathbf{r}')$$

$$- \frac{\eta v}{16\pi} \int\limits_{V_0} \int\limits_{V_0} \frac{1}{v^2} \omega \frac{R}{v} \frac{\bar{\mathbf{J}}(\mathbf{r}) \cdot \mathbf{J}(\mathbf{r}')}{R} \sin kR \, dV(\mathbf{r}) dV(\mathbf{r}')$$

$$+ \frac{\eta v}{16\pi} \int\limits_{V_0} \int\limits_{V_0} \frac{1}{v^2} \frac{\omega \Omega_{\mathbf{J}1}(j\omega)}{R} \sin kR \, dV(\mathbf{r}) dV(\mathbf{r}')$$

$$- \frac{\eta v}{16\pi} \int\limits_{V_0} \int\limits_{V_0} \frac{1}{v^2} \frac{\omega \Omega_{\mathbf{J}2}(j\omega)}{R} \cos kR \, dV(\mathbf{r}) dV(\mathbf{r}'), \qquad (5.113)$$

where

$$\Omega_{\rho 1}(j\omega) = \mathrm{Re}\left[j \frac{\partial \bar{\rho}(\mathbf{r})}{\partial \omega} \rho(\mathbf{r}') - j \bar{\rho}(\mathbf{r}) \frac{\partial \rho(\mathbf{r}')}{\partial \omega} \right],$$

$$\Omega_{\rho 2}(j\omega) = \mathrm{Im}\left[j \frac{\partial \bar{\rho}(\mathbf{r})}{\partial \omega} \rho(\mathbf{r}') - j \bar{\rho}(\mathbf{r}) \frac{\partial \rho(\mathbf{r}')}{\partial \omega} \right],$$

$$\Omega_{\mathbf{J}1}(j\omega) = \mathrm{Re}\left[j \frac{\partial \bar{\mathbf{J}}(\mathbf{r})}{\partial \omega} \cdot \mathbf{J}(\mathbf{r}') - j \bar{\mathbf{J}}(\mathbf{r}) \cdot \frac{\partial \mathbf{J}(\mathbf{r}')}{\partial \omega} \right],$$

$$\Omega_{\mathbf{J}2}(j\omega) = \mathrm{Im}\left[j \frac{\partial \bar{\mathbf{J}}(\mathbf{r})}{\partial \omega} \cdot \mathbf{J}(\mathbf{r}') - j \bar{\mathbf{J}}(\mathbf{r}) \cdot \frac{\partial \mathbf{J}(\mathbf{r}')}{\partial \omega} \right].$$

Note that the integral

$$\int_V \int_V \left[j\frac{\partial \bar{\rho}(\mathbf{r})}{\partial \omega}\rho(\mathbf{r}') - j\bar{\rho}(\mathbf{r})\frac{\partial \rho(\mathbf{r}')}{\partial \omega} \right] \frac{\sin kR}{R}\, dV(\mathbf{r}) dV(\mathbf{r}')$$

is real. Thus (5.113) can be written as

$$\begin{aligned}
\widetilde{W}_m + \widetilde{W}_e = {} & \frac{\eta v}{16\pi} \int_V \int_V \left[\bar{\rho}(\mathbf{r})\rho(\mathbf{r}') + \frac{1}{v^2}\bar{\mathbf{J}}(\mathbf{r}) \cdot \mathbf{J}(\mathbf{r}') \right] \frac{\cos kR}{R}\, dV(\mathbf{r}) dV(\mathbf{r}') \\
& + \frac{\omega \eta}{16\pi} \int_V \int_V \left[\bar{\rho}(\mathbf{r})\rho(\mathbf{r}') - \frac{1}{c^2}\bar{\mathbf{J}}(\mathbf{r}) \cdot \mathbf{J}(\mathbf{r}') \right] \sin kR\, dV(\mathbf{r}) dV(\mathbf{r}') \\
& - \frac{\omega \eta v}{8\pi} \int_V \int_V \mathrm{Im}\left[\bar{\rho}(\mathbf{r}) \frac{\partial \rho(\mathbf{r}')}{\partial \omega} \right] \frac{\sin kR}{R}\, dV(\mathbf{r}) dV(\mathbf{r}') \\
& + \frac{\omega \eta v}{8\pi} \int_V \int_V \frac{1}{v^2} \mathrm{Im}\left[\bar{\mathbf{J}}(\mathbf{r}) \cdot \frac{\partial \mathbf{J}(\mathbf{r}')}{\partial \omega} \right] \frac{\sin kR}{R}\, dV(\mathbf{r}) dV(\mathbf{r}').
\end{aligned} \tag{5.114}$$

From (5.112) and (5.114), the stored energies can be obtained as follows

$$\begin{aligned}
\widetilde{W}_m = {} & \frac{\eta v}{16\pi} \int_{V_0} \int_{V_0} \frac{1}{v^2}\bar{\mathbf{J}}(\mathbf{r}) \cdot \mathbf{J}(\mathbf{r}') \frac{\cos kR}{R}\, dV(\mathbf{r}) dV(\mathbf{r}') \\
& + \frac{\omega \eta}{32\pi} \int_{V_0} \int_{V_0} \left[\bar{\rho}(\mathbf{r})\rho(\mathbf{r}') - \frac{1}{v^2}\bar{\mathbf{J}}(\mathbf{r}) \cdot \mathbf{J}(\mathbf{r}') \right] \sin kR\, dV(\mathbf{r}) dV(\mathbf{r}') \\
& - \frac{\omega \eta v}{16\pi} \int_{V_0} \int_{V_0} \mathrm{Im}\left[\bar{\rho}(\mathbf{r}) \frac{\partial \rho(\mathbf{r}')}{\partial \omega} \right] \frac{\sin kR}{R}\, dV(\mathbf{r}) dV(\mathbf{r}') \\
& + \frac{\omega \eta v}{16\pi} \int_{V_0} \int_{V_0} \frac{1}{v^2} \mathrm{Im}\left[\bar{\mathbf{J}}(\mathbf{r}) \cdot \frac{\partial \mathbf{J}(\mathbf{r}')}{\partial \omega} \right] \frac{\sin kR}{R}\, dV(\mathbf{r}) dV(\mathbf{r}'),
\end{aligned} \tag{5.115}$$

$$\begin{aligned}
\widetilde{W}_e = {} & \frac{\eta v}{16\pi} \int_{V_0} \int_{V_0} \bar{\rho}(\mathbf{r})\rho(\mathbf{r}') \frac{\cos kR}{R}\, dV(\mathbf{r}) dV(\mathbf{r}') \\
& + \frac{\omega \eta}{32\pi} \int_{V_0} \int_{V_0} \left[\bar{\rho}(\mathbf{r})\rho(\mathbf{r}') - \frac{1}{v^2}\bar{\mathbf{J}}(\mathbf{r}) \cdot \mathbf{J}(\mathbf{r}') \right] \sin kR\, dV(\mathbf{r}) dV(\mathbf{r}')
\end{aligned}$$

$$-\frac{\omega\eta v}{16\pi}\int_{V_0}\int_{V_0}\mathrm{Im}\left[\bar{\rho}(\mathbf{r})\frac{\partial\rho(\mathbf{r}')}{\partial\omega}\right]\frac{\sin kR}{R}\,dV(\mathbf{r})dV(\mathbf{r}')$$

$$+\frac{\omega\eta v}{16\pi}\int_{V_0}\int_{V_0}\frac{1}{v^2}\mathrm{Im}\left[\bar{\mathbf{J}}(\mathbf{r})\cdot\frac{\partial\mathbf{J}(\mathbf{r}')}{\partial\omega}\right]\frac{\sin kR}{R}\,dV(\mathbf{r})dV(\mathbf{r}'). \tag{5.116}$$

Equations (5.115) and (5.116) are the most general expressions for the stored energies. The last two terms in (5.115) and (5.116) will be denoted by

$$\widetilde{W}_d = -\frac{\omega\eta v}{16\pi}\int_{V}\int_{V}\mathrm{Im}\left[\bar{\rho}(\mathbf{r})\frac{\partial\rho(\mathbf{r}')}{\partial\omega}\right]\frac{\sin kR}{R}\,dV(\mathbf{r})dV(\mathbf{r}')$$

$$+\frac{\omega\eta v}{16\pi}\int_{V}\int_{V}\frac{1}{v^2}\mathrm{Im}\left[\bar{\mathbf{J}}(\mathbf{r})\cdot\frac{\partial\mathbf{J}(\mathbf{r}')}{\partial\omega}\right]\frac{\sin kR}{R}\,dV(\mathbf{r})dV(\mathbf{r}'), \tag{5.117}$$

which represent frequency-derivative terms of the source distributions, and disappear in (5.99) and (5.100). The contribution of $\widetilde{W}_d$ to the stored energies could be significant, and cannot be ignored in general except for small antennas. Equations (5.115) and (5.116) reduce to (5.99) and (5.100) if either the source distributions are assumed to be independent of frequency or the source distributions are purely real (or imaginary). From (5.115) and (5.116), the antenna Q_{real} may be determined either by (5.11) or (5.12).

Remark 5.3: A method for calculating the stored energies for small antennas was proposed in Geyi (2003b). The method is based on the understanding that, for a small antenna, the total energy in Poynting theorem can be easily separated into the stored energy and radiated energy by using the low frequency expansions. The Poynting theorem in frequency domain provides an equation on the stored electric and magnetic energy while the Poynting theorem in time domain can be used as another independent equation for the stored electric and magnetic energy. By solving these equations, the stored electric and magnetic energy can be obtained as follows

$$\widetilde{W}_e = \frac{\eta v}{16\pi}\int_{V_0}\int_{V_0}\frac{1}{R}\rho(\mathbf{r})\bar{\rho}(\mathbf{r}')dV(\mathbf{r})dV(\mathbf{r}'), \tag{5.118}$$

$$\widetilde{W}_m = \frac{\eta v}{16\pi}\frac{1}{v^2}\int_{V_0}\int_{V_0}\frac{\mathbf{J}(\mathbf{r})\cdot\bar{\mathbf{J}}(\mathbf{r}')}{R}\,dV(\mathbf{r})dV(\mathbf{r}')$$

$$+\frac{\eta v}{16\pi}\frac{k^2}{2}\int_{V_0}\int_{V_0}R\rho(\mathbf{r})\bar{\rho}(\mathbf{r}')dV(\mathbf{r})dV(\mathbf{r}'). \tag{5.119}$$

It can be shown that (5.115) and (5.116) reduce to (5.118) and (5.119) for small antennas. □

5.4.5.3 *Relationship between Q and Bandwidth*

Consider a high quality factor system. Let ω_r denote one of the resonant frequencies of a single antenna system. Then, by definition, we have

$$X(\omega_r) = 0. \tag{5.120}$$

For small α, we have $X(\alpha, \omega_r) \approx X(\omega_r) = 0$ at the resonant frequency ω_r. From (5.90), we obtain

$$\left.\frac{dR}{d\omega}\right|_{\omega_r} = -\left.\frac{\partial X(\alpha, \omega_r)}{\partial \alpha}\right|_{\alpha=0} \approx 0.$$

Thus as one moves off resonance, the antenna input impedance Z can be written as

$$Z \approx R|_{\omega_r} + j(\omega - \omega_r)\left.\frac{dX}{d\omega}\right|_{\omega_r} + \cdots.$$

The frequency at which the absolute value of the input impedance is equal to $\sqrt{2}$ times its value at resonance is the half-power point. The half-power points occur when

$$R|_{\omega_r} = \left|(\omega - \omega_r)\left.\frac{dX}{d\omega}\right|_{\omega_r}\right|, \tag{5.121}$$

so that the fractional bandwidth B_f can be written

$$B_f = \frac{2|\omega - \omega_r|}{\omega_r} \approx \frac{2R|_{\omega_r}}{\omega_r |dX/d\omega|_{\omega_r}}. \tag{5.122}$$

From (5.101) or (5.102), we obtain

$$B_f = \frac{1}{Q_{\text{real}}|_{\omega_r}}. \tag{5.123}$$

Thus we have proved that the antenna fractional bandwidth is the inversion of antenna Q_{real} when $Q_{\text{real}} \gg 1$.

5.4.5.4 *Minimum Possible Antenna Quality Factor*

The study of antenna quality factor was usually based on the spherical wavefunction expansion outside the circumscribing sphere of the antenna.

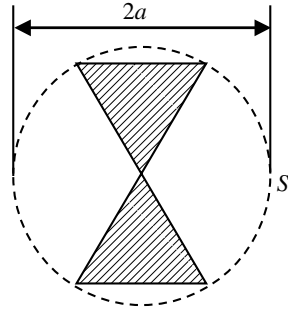


Figure 5.24 Antenna and its circumscribing sphere.

The antenna quality factor resulting from the spherical wavefunction expansion is much lower than the real value Q_{real} as the stored energy inside the circumscribing sphere has been ignored. Assume that the antenna is enclosed by the circumscribing sphere of radius a, denoted by V_a with bounding surface S_a, as illustrated in Figure 5.24. The total stored energy outside the circumscribing sphere can be evaluated through (5.63)

$$\widetilde{W}_e + \widetilde{W}_m = \int_a^\infty dr \left\{ \int_0^{2\pi} d\varphi \int_0^\pi r^2 \left(\frac{\varepsilon}{4} |\mathbf{E}|^2 + \frac{\mu}{4} |\mathbf{H}|^2 \right) \sin\theta \, d\theta \right.$$
$$\left. - \frac{1}{2v} Re \int_{\partial V_\infty} \mathbf{E} \times \bar{\mathbf{H}} \cdot \mathbf{u}_r \, dS \right\}.$$

From (5.23) and (5.26), we obtain

$$P_{\text{rad}} = \frac{1}{2} \text{Re} \int_{S_\infty} \mathbf{E} \times \bar{\mathbf{H}} \cdot \mathbf{u}_r \, dS = \frac{1}{2k^2\eta} \sum_{n,m,l} N_{nm}^2 \left(\left| \alpha_{nml}^{(2)} \right|^2 + \left| \beta_{nml}^{(2)} \right|^2 \right), \tag{5.124}$$

$$\omega \widetilde{W}_m = \frac{1}{4k^2\eta} \sum_{n,m,l} N_{nm}^2 \left(\left| \alpha_{nml}^{(2)} \right|^2 Q_n + \left| \beta_{nml}^{(2)} \right|^2 Q_n' \right),$$
$$\omega \widetilde{W}_e = \frac{1}{4k^2\eta} \sum_{n,m,l} N_{nm}^2 \left(\left| \alpha_{nml}^{(2)} \right|^2 Q_n' + \left| \beta_{nml}^{(2)} \right|^2 Q_n \right), \tag{5.125}$$

where

$$Q_n = ka - \left|h_n^{(2)}(ka)\right|^2 \left[\frac{1}{2}(ka)^3 + ka(n+1)\right] - \frac{1}{2}(ka)^3\left|h_{n+1}^{(2)}(ka)\right|^2$$

$$+ \frac{1}{2}(ka)^2(2n+3)[j_n(ka)j_{n+1}(ka) + n_n(ka)n_{n+1}(ka)] \tag{5.126}$$

$$Q_n' = ka - \frac{1}{2}(ka)^3\left[\left|h_n^{(2)}(ka)\right|^2 - j_{n-1}(ka)j_{n+1}(ka) - n_{n-1}(ka)n_{n+1}(ka)\right].$$

For the first three modes, we have

$$Q_1 = \frac{1}{ka} + \frac{1}{(ka)^3}, \quad Q_1' = \frac{1}{ka},$$

$$Q_2 = \frac{3}{ka} + \frac{6}{(ka)^3} + \frac{18}{(ka)^5}, \quad Q_2' = \frac{3}{ka} + \frac{4}{(ka)^3}, \tag{5.127}$$

$$Q_3 = \frac{6}{ka} + \frac{21}{(ka)^3} + \frac{135}{(ka)^5} + \frac{675}{(ka)^7}, \quad Q_3' = \frac{6}{ka} + \frac{15}{(ka)^3} + \frac{45}{(ka)^5}.$$

It follows from (5.11), (5.124) and (5.125) that

$$Q = \frac{\frac{1}{2}\sum\limits_{n,m,l} N_{nm}^2\left(\left|\alpha_{nml}^{(2)}\right|^2 + \left|\beta_{nml}^{(2)}\right|^2\right)(Q_n + Q_n')}{\sum\limits_{n,m,l} N_{nm}^2\left(\left|\alpha_{nml}^{(2)}\right|^2 + \left|\beta_{nml}^{(2)}\right|^2\right)}. \tag{5.128}$$

Since $Q_n > Q_n'$, $Q_{n+1} > Q_n$ and $Q_{n+1}' > Q_n'$ (Fante, 1969), we have

$$Q \geq \frac{\frac{1}{2}\sum\limits_{n,m,l} N_{nm}^2\left(\left|\alpha_{nml}^{(2)}\right|^2 + \left|\beta_{nml}^{(2)}\right|^2\right)(Q_1 + Q_1')}{\sum\limits_{n,m,l} N_{nm}^2\left(\left|\alpha_{nml}^{(2)}\right|^2 + \left|\beta_{nml}^{(2)}\right|^2\right)} = \frac{1}{2}(Q_1 + Q_1').$$

Therefore, minimum possible value for Q is given by

$$\min Q = \frac{Q_1 + Q_1'}{2} = \frac{1}{ka} + \frac{1}{2(ka)^3}. \tag{5.129}$$

The antenna will attain the lowest Q if only TE_{1m} and TM_{1m} modes are equally excited. In this case, the stored electric energy and magnetic energy outside the circumscribing sphere will be equal, and the antenna will be at resonance outside the sphere. The existence of a lower bound for antenna Q

implies that the stored energy around antenna can never be made zero. Once the maximum antenna size is given, this lower bound is then determined. For a small antenna ($ka < 1$), Equation (5.129) can be approximated by

$$\min Q \approx \frac{1}{2(ka)^3}. \tag{5.130}$$

Since Q_{real} is always greater than the Q defined by (5.128), Equation (5.129) may be considered as the minimum possible value for Q_{real}.

Remark 5.4: The same lowest bound (5.129) for Q may be obtained if the traditional definition (5.12) is used (Geyi, 2003a; 2012). □

5.4.6 Maximum Possible Product of Gain and Bandwidth

In most applications, we need to maximize antenna gain and bandwidth simultaneously. For this reason, a reasonable quantity characterizing antenna would be the product of antenna gain and bandwidth, or the ratio of antenna gain to antenna Q_{real}. The ratio of gain to Q_{real} is actually the ratio of radiation intensity over the averaged stored energy around the antenna. In order to seek the maximum possible ratio of gain over antenna quality factor, we may use Q defined by (5.128) to replace Q_{real} in the optimization process.

5.4.6.1 *Directive Antenna*

We assume that the antenna is placed in a spherical coordinate system (r, θ, φ) and enclosed by the smallest circumscribing sphere of radius a, and the spherical coordinate system is oriented in such a way that the maximum radiation is in $(\theta, \varphi) = (0, 0)$ direction. The directivity in the direction of $(\theta, \varphi) = (0, 0)$ is then given by (Geyi, 2003a)

$$\begin{aligned} G &= 4\pi r^2 \frac{\frac{1}{2}\text{Re}(\mathbf{E} \times \bar{\mathbf{H}}) \cdot \mathbf{u}_r}{P^{\text{rad}}} \\ &= \pi \frac{\left|\sum_n (n+1)nj^n\left(\beta_{n1e}^{(2)} + j\alpha_{n1o}^{(2)}\right)\right|^2 + \left|\sum_n (n+1)nj^n\left(\beta_{n1o}^{(2)} - j\alpha_{n1e}^{(2)}\right)\right|^2}{\sum_{n,m,l} N_{nm}^2\left(\left|\alpha_{nml}^{(2)}\right|^2 + \left|\beta_{nml}^{(2)}\right|^2\right)}. \end{aligned} \tag{5.131}$$

From (5.128) and (5.131), we obtain

$$\left.\frac{G}{Q}\right|_{\text{dir}} = \frac{2\pi}{\sum\limits_{n,m,l} N_{nm}^2\left(\left|\alpha_{nml}^{(2)}\right|^2+\left|\beta_{nml}^{(2)}\right|^2\right)(Q_n+Q_n')} \cdot\left[\left|\sum_{n=1}^{\infty}(n+1)nj^n\left(\beta_{n1e}^{(2)}+j\alpha_{n1o}^{(2)}\right)\right|^2 + \left|\sum_{n=1}^{\infty}(n+1)nj^n\left(\beta_{n1o}^{(2)}-j\alpha_{n1e}^{(2)}\right)\right|^2\right]. \tag{5.132}$$

Since only $\alpha_{n1l}^{(2)}$ and $\beta_{n1l}^{(2)}$ contribute to the numerator, the ratio (5.132) can be increased by setting $\alpha_{nml}^{(2)}=\beta_{nml}^{(2)}=0(m\neq 1)$. Thus we have

$$\left.\frac{G}{Q}\right|_{\text{dir}} = \frac{\left|\sum\limits_{n=1}^{\infty}(A_{\text{on}}+B_{en})\right|^2+\left|\sum\limits_{n=1}^{\infty}(A_{en}+B_{\text{on}})\right|^2}{\sum\limits_{n=1}^{\infty}\frac{(Q_n+Q_n')}{2n+1}\left[(|A_{en}|^2+|B_{en}|^2)+(|A_{\text{on}}|^2+|B_{\text{on}}|^2)\right]}, \tag{5.133}$$

where

$$\begin{cases} A_{\text{on}}=j^{n+1}n(n+1)\alpha_{n1o}^{(2)} \\ A_{en}=-j^{n+1}n(n+1)\alpha_{n1e}^{(2)} \end{cases}, \quad \begin{cases} B_{en}=j^{n}n(n+1)\beta_{n1e}^{(2)} \\ B_{\text{on}}=j^{n}n(n+1)\beta_{n1o}^{(2)} \end{cases}.$$

The denominator of (5.133) depends only on the magnitudes of A_n and B_n. If we adjust the phase of A_n and B_n such that they are in phase to maximize the numerator, the denominator will not change. Therefore (5.133) can be maximized as follows

$$\left.\frac{G}{Q}\right|_{\text{dir}} = \frac{\left[\sum\limits_{n=1}^{\infty}\left(\left|A_{\text{on}}\right|+\left|B_{en}\right|\right)\right]^2+\left[\sum\limits_{n=1}^{\infty}\left(\left|A_{en}\right|+\left|B_{\text{on}}\right|\right)\right]^2}{\sum\limits_{n=1}^{\infty}\frac{(Q_n+Q_n')}{2n+1}\left[\left(|A_{en}|^2+|B_{en}|^2\right)+\left(|A_{\text{on}}|^2+|B_{\text{on}}|^2\right)\right]}. \tag{5.134}$$

Making use of the inequality $(a+b)^2\leq 2(a^2+b^2)$, we get

$$\left.\frac{G}{Q}\right|_{\text{dir}} \leq 2\frac{\left(\sum\limits_{n=1}^{\infty}a_n\right)^2+\left(\sum\limits_{n=1}^{\infty}b_n\right)^2}{\sum\limits_{n=1}^{\infty}\frac{(Q_n+Q_n')}{2n+1}(a_n^2+b_n^2)} = 2\frac{(\boldsymbol{\zeta},\mathbf{C}_a)_E^2+(\boldsymbol{\zeta},\mathbf{C}_b)_E^2}{(\mathbf{C}_a,\mathbf{C}_a)_E+(\mathbf{C}_b,\mathbf{C}_b)_E}, \tag{5.135}$$

where

$$a_n = |A_{\text{on}}| + |B_{en}|, \quad b_n = |A_{en}| + |B_{\text{on}}|,$$
$$\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \ldots), \quad \mathbf{C}_{a(b)} = (C_{a(b)1}, C_{a(b)2}, \ldots),$$
$$\zeta_n = \sqrt{\frac{2n+1}{Q_n + Q'_n}}, \quad C_{an} = \frac{|a_n|}{\zeta_n}, \quad C_{bn} = \frac{|b_n|}{\zeta_n},$$

and both $\boldsymbol{\zeta}$ and $\mathbf{C}_{a(b)}$ are vectors in the Euclidean space consisting of all vectors of infinite dimension with the inner product and norm defined by $(\boldsymbol{\zeta}, \mathbf{C})_E = \sum_{n=1}^{\infty} \zeta_n C_n$ and $\|\boldsymbol{\zeta}\| = (\boldsymbol{\zeta}, \boldsymbol{\zeta})_E^{1/2}$ respectively. It follows from (5.135) and Schwartz inequality that

$$\left.\frac{G}{Q}\right|_{\text{dir}} \le 2\|\boldsymbol{\zeta}\|_E^2. \tag{5.136}$$

The equality holds if $\mathbf{C}_a = \mathbf{C}_b = c_1\boldsymbol{\zeta}$. Thus the upper limit of ratio of gain to Q for a directional antenna is

$$\max \left.\frac{G}{Q}\right|_{\text{dir}} = 2\|\boldsymbol{\zeta}\|_E^2 = \sum_{n=1}^{\infty} \frac{2(2n+1)}{Q_n + Q'_n}. \tag{5.137}$$

This is the maximum possible ratio of gain to Q for a directive antenna.

Remark 5.5: The same upper bound (5.137) may be obtained if the traditional definition (5.12) for Q is used (Geyi, 2003a; 2012). □

5.4.6.2 *Omni-directional Antenna*

We assume that the antenna has an omni-directional pattern and the field is independent of φ, and consider the maximum possible ratio of gain to Q in the direction of $\theta = \pi/2$. The directivity for an omni-directional antenna can be expressed as (Geyi, 2003a; 2010)

$$G = 4\pi \frac{\left|\sum_{n=1}^{\infty} j^n \beta_{n0e}^{(2)} P_n^1(0)\right|^2 + \left|\sum_{n=1}^{\infty} j^{n+1} \alpha_{n0e}^{(2)} P_n^1(0)\right|^2}{\sum_{n,m,l} N_{nm}^2 \left(\left|\alpha_{nml}^{(2)}\right|^2 + \left|\beta_{nml}^{(2)}\right|^2\right)}. \tag{5.138}$$

It follows from (5.128) and (5.138) that

$$\left.\frac{G}{Q}\right|_{\text{omn}} = 8\pi \frac{\left|\sum_{n} j^n \beta_{n0e}^{(2)} P_n^1(0)\right|^2 + \left|\sum_{n} j^{n+1} \alpha_{n0e}^{(2)} P_n^1(0)\right|^2}{\sum_{n,m,l} N_{nm}^2 \left(\left|\alpha_{nml}^{(2)}\right|^2 + \left|\beta_{nml}^{(2)}\right|^2\right)(Q_n + Q'_n)}. \tag{5.139}$$

Only $\alpha_{n0e}^{(2)}$ and $\beta_{n0e}^{(2)}$ contribute to the numerator. Therefore, the ratio (5.139) can be increased by setting $\alpha_{nml}^{(2)} = \beta_{nml}^{(2)} = 0 (m \neq 0)$, $\alpha_{n0o}^{(2)} = \beta_{n0o}^{(2)} = 0$. Let $A_n = j^{n+1}\alpha_{n0e}^{(2)}$ and $B_n = j^n \beta_{n0e}^{(2)}$, we have

$$\left.\frac{G}{Q}\right|_{\text{omn}} = 8\pi \frac{\left|\sum\limits_{n=1}^{\infty} A_n P_n^1(0)\right|^2 + \left|\sum\limits_{n=1}^{\infty} B_n P_n^1(0)\right|^2}{\sum\limits_{n=1}^{\infty} N_{n0}^2 \big(|A_n|^2 + |B_n|^2\big)(Q_n + Q_n')}. \tag{5.140}$$

Since the denominator of (5.140) depends only on the magnitude of A_n and B_n, the denominator is not changed if the phases of A_n and B_n are adjusted to maximize the ratio of gain to Q. If we choose the phases of A_n and B_n to be the negative of $P_n^1(0)$, the terms in the numerator will be added in phase. Thus

$$\left.\frac{G}{Q}\right|_{\text{omn}} = 8\pi \frac{\left(\sum\limits_{n=1}^{\infty} |A_n|\big|P_n^1(0)\big|\right)^2 + \left(\sum\limits_{n=1}^{\infty} |B_n|\big|P_n^1(0)\big|\right)^2}{\sum\limits_{n=1}^{\infty} N_{n0}^2 \big(|A_n|^2 + |B_n|^2\big)(Q_n + Q_n')}.$$

Introducing

$$a_n = |A_n|\big|P_n^1(0)\big|, \quad b_n = |B_n|\big|P_n^1(0)\big|,$$

we have

$$\begin{aligned}\left.\frac{G}{Q}\right|_{\text{omn}} &= 8\pi \frac{\left(\sum\limits_{n=1}^{\infty} a_n\right)^2 + \left(\sum\limits_{n=1}^{\infty} b_n\right)^2}{\sum\limits_{n=1}^{\infty} N_{n0}^2 (a_n^2 + b_n^2)(Q_n + Q_n')/\big|P_n^1(0)\big|^2} \\ &= 8\pi \frac{(\boldsymbol{\xi}, \mathbf{D}_a)_E^2 + (\boldsymbol{\xi}, \mathbf{D}_b)_E^2}{(\mathbf{D}_a, \mathbf{D}_a)_E + (\mathbf{D}_b, \mathbf{D}_b)_E} \leq 8\pi \|\boldsymbol{\xi}\|_E^2, \end{aligned} \tag{5.141}$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots)$, $\mathbf{D}_{a(b)} = (\mathbf{D}_{a(b)1}, \mathbf{D}_{a(b)2}, \ldots)$ with

$$\xi_n = \frac{\big|P_n^1(0)\big|}{N_{n0}\sqrt{Q_n + Q_n'}}, \quad D_{an} = \frac{|a_n|}{\xi_n}, \quad D_{bn} = \frac{|b_n|}{\xi_n}.$$

The ratio (5.141) reaches maximum if $\mathbf{D}_a = \mathbf{D}_b = c_1 \boldsymbol{\xi}$. As a result, the upper limit of the ratio of gain to Q for an omni-directional antenna is

$$\max \left.\frac{G}{Q}\right|_{\text{omn}} = 8\pi \|\boldsymbol{\xi}\|_E^2 = \sum_{n=1}^{\infty} \frac{2(2n+1)\big|P_n^1(0)\big|^2}{n(n+1)(Q_n + Q_n')}. \tag{5.142}$$

Remark 5.6: The same upper bound (5.137) may be obtained if the traditional definition (5.12) for Q is used (Geyi, 2003a; 2012). □

Remark 5.7: Chu has shown that the maximum ratio of gain to Q for an omni-directional antenna is (Chu, 1948)

$$\max \frac{G}{Q}\bigg|_{\text{omn}}^{\text{Chu}} = \sum_{n=1}^{\infty} \frac{(2n+1)\left|P_n^1(0)\right|^2}{n(n+1)Q_n^{\text{Chu}}}. \tag{5.143}$$

Here Q_n^{Chu} is the quality factor of nth TM modes and is a function of ka. Chu's theory is valid only for an omni-directional antenna that radiates either TE or TM modes, and is based on the equivalent ladder network representation of the wave impedance of each mode and the stored energies in some elements have been neglected. Hence the Chu's limit just holds approximately. Also note that the upper limit (5.142) can be twice as much as Chu's limit (5.143) if ka is small. □

5.4.6.3 *Best Possible Antenna Performance*

Since the antenna fractional bandwidth B_f is reciprocal to antenna Q_{real} if Q_{real} is not very small, the product of antenna gain and bandwidth can be expressed as $GB_f \approx G/Q_{\text{real}}$. The antenna quality factor used in (5.137) and (5.142) does not include the stored energies inside the circumscribing sphere of the antenna, it is thus smaller than the real antenna Q_{real}. It follows from (5.137) and (5.142) that the products of gain and bandwidth for an arbitrary antenna of dimension $2a$ are bounded by

$$\begin{aligned} GB_f|_{\text{dir}} &\le \max\, GB_f|_{\text{dir}} = \sum_{n=1}^{\infty} \frac{2(2n+1)}{Q_n(ka)+Q_n'(ka)}, \\ GB_f|_{\text{omn}} &\le \max\, GB_f|_{\text{omn}} = \sum_{n=1}^{\infty} \frac{2(2n+1)\left|P_n^1(0)\right|^2}{n(n+1)\left[Q_n(ka)+Q_n'(ka)\right]}. \end{aligned} \tag{5.144}$$

The first expression applies for directional antennas, and the second one for the omni-directional antennas. It should be notified that the right-hand sides of (5.144) are finite numbers. From (5.129), the fractional bandwidth of an arbitrary antenna of dimension $2a$ has an upper limit too

$$B_f \le \max B_f = \frac{2(ka)^3}{2(ka)^2+1}. \tag{5.145}$$

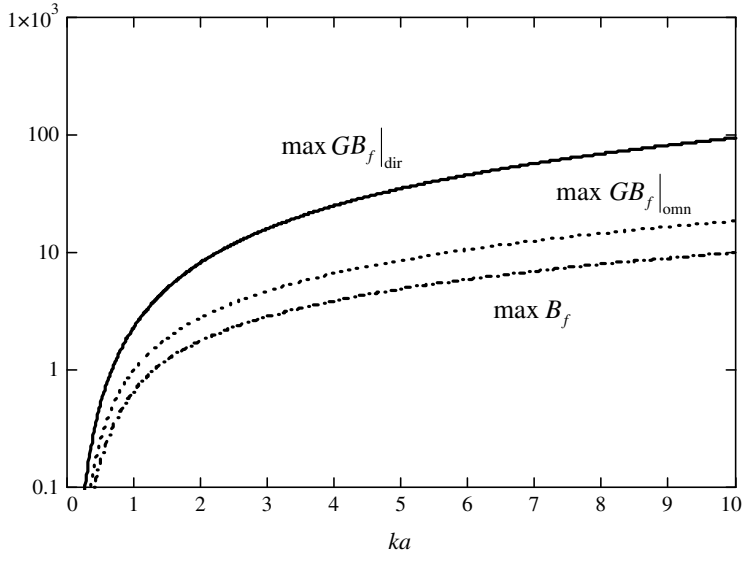


Figure 5.25 Upper bounds of antenna performances.

Equations (5.144) indicate that one can sacrifice the bandwidth to enhance the gain. If the bandwidth is rendered very small, a high gain antenna can be achieved. One can also sacrifice the gain to improve the bandwidth. But the improvement will be limited as the bandwidth itself is bounded by the right-hand side of (5.145).

The upper bounds $\max GB_f|_{\text{dir}}$, $\max GB_f|_{\text{omn}}$ and $\max B_f$ are all monotonically increasing functions of ka as shown in Figure 5.25. It can be seen that $\max GB_f|_{\text{dir}}$ is always higher than $\max GB_f|_{\text{omn}}$. The rate of increase of these upper bounds for small ka is much higher than that for large ka, which implies that a little increase in the size of the small antennas will notably improve their performances. For the small antennas with $ka < 1$, only the first terms of the infinite series in (5.144) are significant. Thus we may write

$$\begin{aligned} \max GB_f|_{\text{dir}} &\approx \frac{6}{Q_1 + Q_1'} = \frac{6(ka)^3}{2(ka)^2 + 1}, \\ \max GB_f|_{\text{omn}} &\approx \frac{3}{Q_1 + Q_1'} = \frac{3(ka)^3}{2(ka)^2 + 1}. \end{aligned} \tag{5.146}$$

The right-hand sides of (5.146) are the best possible antenna performances that a small antenna of maximum dimension $2a$ can achieve. They set up

a target that can be approached by various methods and have been proven to be very useful for small antenna design for which try and error method is often used.

5.5 Wire Antennas

Wire antennas are fundamental to understanding the radiation mechanisms. This can be illustrated by the evolution of the handset antenna design from a monopole to the planar inverted-F antenna. The monopole antennas (see Figure 5.26) are the first type of antennas recognized for radio communication devices. They are easy to design, light weight, and have omni-directional radiation pattern in the horizontal plane. However, since the physical length of a monopole antenna is quarter of its wavelength at the operating frequency, this antenna is relatively very long. Therefore, monopole antennas are usually used as external antennas. As the size of handheld devices was decreasing, the inverted-L antenna (ILA) was found to be a promising alternative to replace the external monopole antenna. The ILA is an end-fed short monopole with a horizontal wire element placed on top that acts as a capacitive load (see Figure 5.27). The design of the ILA has a simple layout making it cost efficient to manufacture. Although the radiation properties of the ILA have advantages over those of the monopole antenna by radiating in both polarizations due to the horizontal arm, its input impedance is similar to that of the short monopole: low resistance and high reactance. This prompted antenna designers to search for an antenna with nearly resistive load thus provides reduced mismatch loss. For this purpose, the inverted-F antenna (IFA) was introduced (see Figure 5.28) (King *et al.*, 1960), which adds a second inverted-L section

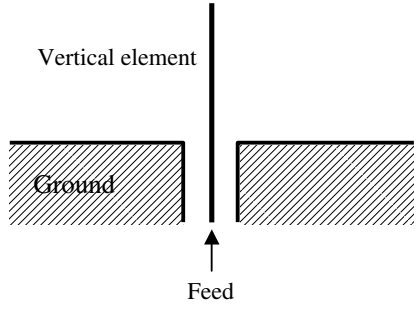


Figure 5.26 Fundamental structure of the monopole antenna.

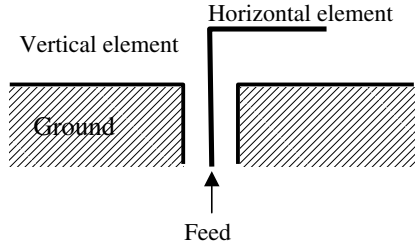


Figure 5.27 ILA-modified from the monopole antenna.

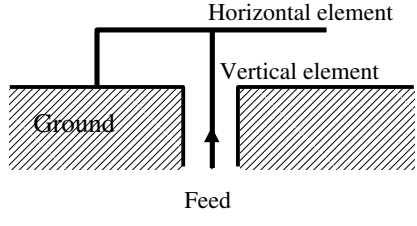


Figure 5.28 IFA-modified from the ILA.

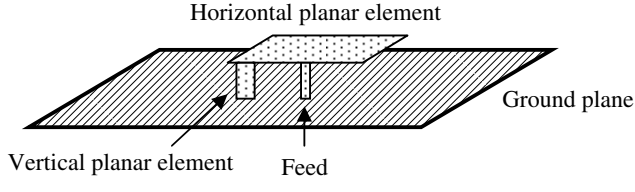


Figure 5.29 Basic layout of the PIFA-modified from the IFA.

to the end of an ILA. The additional inverted-L segment introduces a convenient tuning option to the original ILA and greatly improves the antenna usability. Even with the improvement in the match of the IFA over the ILA, both these antennas have inherently narrow bandwidths. To obtain broad bandwidth characteristics, antenna designers transformed the horizontal element from a wire to a plate (see Figure 5.29), and the planar inverted-F antenna (PIFA) was introduced (Taga and Tsunekawa, 1987). The PIFA is widely used in nowadays mobile handheld devices. It is a self-resonating antenna with purely resistive impedance at the frequency of operation. This makes it a practical candidate for mobile handheld design since it does not require a matching circuit between the antenna and the

load reducing both cost and losses. Despite the relative simple design of the ILA, IFA, and the PIFA, the optimal design of any of these antennas is not unique. Numerous designs have been reported in the literature. Many of them suggest approaches to further improve the bandwidth and the performance of these antennas.

The evolution from a monopole to the PIFA indicates that the essential component of a handset antenna is the "wire". The patch(s), slot(s), and stub(s) are only used to compensate for the mismatch and improve the radiation characteristics. Notice that at the megahertz frequency range, the current flowing on the surface of a conductor no longer has a uniform distribution due to the skin effect. Instead it is confined to a relatively small area, and the effective area of the conductor is smaller than the actual dimension. For example, by simulating a basic PIFA and examining the current distribution on its surface at the frequency of operation, one can see that the current distribution is concentrated at the edge(s) of the antenna. For this reason, the length of these edge(s) where the current is concentrated is the major parameter that tunes the antenna to the desired frequency. The remainders of the conductor plate(s) forming the patch(s) of the antenna are not essential in tuning the antenna but are rather to improve the antenna characteristics. In fact, removing these parts would affect the matching of the antenna and would not detune it much. From this intuition, many antennas may be represented by the fundamental wires responsible for its tuning at the frequencies of operation, and these become the backbone of the final design.

5.5.1 Asymptotic Solutions for Wire Antennas

The wire structures have been extensively investigated by a number of authors (e.g., Schelkunoff, 1952; King, 1956). When the radius of the wire model for an antenna is very thin, it is possible to find an analytical solution for the current distribution on the wire, which contains useful information on the radiation properties of the original metal antenna, and thus provides guidelines for practical antenna design (Geyi *et al.*, 2008b). Let us consider a thin wire illuminated by an incident field $\mathbf{E}_{\text{in}}$. We assume that the wire is a curved circular cylinder of radius a_0 and a curvilinear l-axis (l stands for arc length) runs along the axis of the circular cylinder as shown in Figure 5.30. The scattered field due to the current in the wire is

$$\mathbf{E}_s(\mathbf{r}) = -j\omega\mathbf{A}(\mathbf{r}) - \nabla\phi(\mathbf{r}),$$

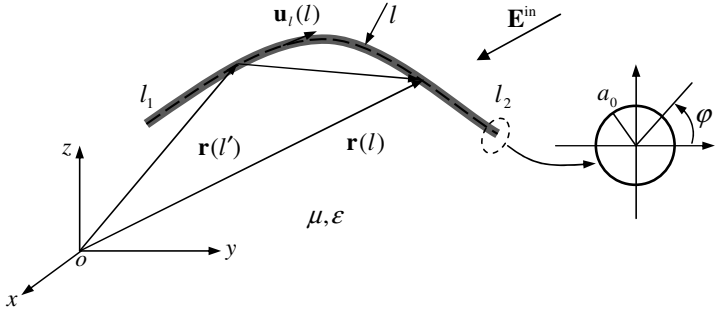


Figure 5.30 An arbitrary thin wire illuminated by an incident field.

where $\mathbf{A}(\mathbf{r})$ is the vector potential and ϕ is the scalar potential. On the surface of the thin wire the total electric field must vanish, and we have

$$\mathbf{E}_{\text{in}}(\mathbf{r}) = -\mathbf{E}_s(\mathbf{r}) = j\omega\mathbf{A}(\mathbf{r}) + \nabla\phi(\mathbf{r}). \tag{5.147}$$

Let $\mathbf{u}_l(l)$ be the unit tangent vector along l-axis. Multiplying both sides of (5.147) by $\mathbf{u}_l(l)$ leads to

$$\mathbf{E}_{\text{in}}(\mathbf{r}) \cdot \mathbf{u}_l(l) = j\omega\mathbf{A}(\mathbf{r}) \cdot \mathbf{u}_l(l) + \frac{d\phi(\mathbf{r})}{dl}. \tag{5.148}$$

The vector potential $\mathbf{A}(\mathbf{r})$ on the surface of the thin wire due to a current distribution $I(l)$ is given by

$$\mathbf{A}(\mathbf{r}) = \frac{\mu}{2\pi}\int_0^{2\pi} d\varphi' \int_{l_1}^{l_2} I(l')\mathbf{u}_l(l')\frac{e^{-jkR}}{4\pi R}dl',$$

where $R = |\mathbf{r}(l) - \mathbf{r}(l')|$. Since the integrand is singular at $l' = l$, we may rewrite the above as

$$\begin{aligned}\mathbf{A}(\mathbf{r}) = {} & \frac{\mu}{2\pi}\int_0^{2\pi} d\varphi' \int_{l_1}^{l-\tau} I(l')\mathbf{u}_l(l')\frac{e^{-jkR}}{4\pi R}dl' + \frac{\mu}{2\pi}\int_0^{2\pi} d\varphi' \int_{l-\tau}^{l+\tau} I(l')\mathbf{u}_l(l')\frac{e^{-jkR}}{4\pi R}dl' \\ & + \frac{\mu}{2\pi}\int_0^{2\pi} d\varphi' \int_{l+\tau}^{l_2} I(l')\mathbf{u}_l(l')\frac{e^{-jkR}}{4\pi R}dl',\end{aligned} \tag{5.149}$$

where τ is a small positive number. The second term on the right-hand side can be written as

$$\frac{\mu}{2\pi}\int_0^{2\pi} d\varphi' \int_{l-\tau}^{l+\tau} I(l')\mathbf{u}_l(l')\frac{e^{-jkR}}{4\pi R}dl'$$

$$= \frac{\mu}{2\pi}\mathbf{u}_l(l)I(l)\int_0^{2\pi} d\varphi' \int_{l-\tau}^{l+\tau} \frac{1}{4\pi R}dl' + \frac{\mu}{2\pi}\mathbf{u}_l(l)I(l)\int_0^{2\pi} d\varphi' \int_{l-\tau}^{l+\tau} \frac{\cos kR - 1}{4\pi R}dl'$$

$$- j\frac{\mu}{2\pi}\mathbf{u}_l(l)I(l)\int_0^{2\pi} d\varphi' \int_{l-\tau}^{l+\tau} \frac{\sin kR}{4\pi R}dl', \tag{5.150}$$

where

$$R = |\mathbf{r} - \mathbf{r}'| \approx [(l-l')^2 + \alpha^2]^{1/2}, \quad \alpha^2 = 4a_0^2 \sin^2 \frac{\varphi - \varphi'}{2},$$

if τ is small. Making use of the following asymptotic calculations for small τ

$$\int_0^{2\pi} d\varphi' \int_{l-\tau}^{l+\tau} \frac{1}{4\pi R}dl' = \ln 2\tau - \ln a_0, \tag{5.151}$$

$$\int_0^{2\pi} d\varphi' \int_{l-\tau}^{l+\tau} \frac{\cos kR - 1}{4\pi R}dl' = \mathrm{Ci}(k\tau) - \ln k\tau - \gamma, \tag{5.152}$$

$$\int_0^{2\pi} d\varphi' \int_{l-\tau}^{l+\tau} \frac{\sin kR}{4\pi R}dl' = \int_0^{\tau} \frac{\sin ku}{u}du. \tag{5.153}$$

Equation (5.150) can be written as

$$\frac{\mu}{2\pi}\int_0^{2\pi} d\varphi' \int_{l-\tau}^{l+\tau} I(l')\mathbf{u}_l(l')\frac{e^{-jkR}}{4\pi R}dl' = \frac{\mu}{2\pi}\mathbf{u}_l(l)I(l)(\ln 2\tau - \ln a_0)$$

$$+ \frac{\mu}{2\pi}\mathbf{u}_l(l)I(l)[\mathrm{Ci}(k\tau) - \ln k\tau - \gamma] - j\frac{\mu}{2\pi}\mathbf{u}_l(l)I(l)\int_0^{\tau} \frac{\sin ku}{u}du$$

$$= -\frac{\mu}{2\pi}\mathbf{u}_l(l)I(l)\ln ka_0 + \text{finite numbers}, \tag{5.154}$$

where Ci and γ are **cosine integral** and **Euler constant** respectively:

$$\mathrm{Ci}(x) = -\int_x^\infty \frac{\cos u}{u} du, \quad \gamma = 0.5772.$$

As $a_0 \to 0$, the first and third term on the right-hand side of (5.149) are finite numbers. As an asymptotic approximation, we thus have

$$\mathbf{A}(\mathbf{r}) = -\frac{\mu}{2\pi}\mathbf{u}_l(l) I(l) \ln k a_0. \tag{5.155}$$

From the Lorentz gauge condition $\nabla \cdot \mathbf{A} + j\omega\mu\varepsilon\varphi = 0$, we may find that

$$\frac{d\mathbf{A}(\mathbf{r}) \cdot \mathbf{u}_l(l)}{dl} + j\omega\mu\varepsilon\phi = 0. \tag{5.156}$$

It follows from (5.148), (5.155) and (5.156) that

$$\begin{aligned} &\frac{d\phi(l)}{dl} + j\omega L_0 I(l) = \mathbf{E}_{\mathrm{in}}(\mathbf{r}) \cdot \mathbf{u}_l(l), \\ &\frac{dI(l)}{dl} + j\omega C_0 \phi(l) = 0, \end{aligned} \tag{5.157}$$

where

$$L_0 = -\frac{\mu}{2\pi} \ln k a_0, \quad C_0 = \frac{\mu\varepsilon}{L_0}. \tag{5.158}$$

From (5.157) we obtain

$$\frac{d^2 I(l)}{dl^2} + k^2 I(l) = -j\omega C_0 \mathbf{E}_{\mathrm{in}}(\mathbf{r}) \cdot \mathbf{u}_l(l) \tag{5.159}$$

where $k = \omega\sqrt{\mu\varepsilon}$. If the thin wire is excited by a localized incident voltage source at $l = l'$, the source term $\mathbf{E}_{\mathrm{in}} \cdot \mathbf{u}_l(l)$ in (5.159) can be replaced by a delta function for the thin wire

$$\frac{d^2 I(l)}{dl^2} + k^2 I(l) = -j\omega C_0 V_s \delta(l - l') \tag{5.160}$$

where V_s is the amplitude of the delta voltage source. The current distributions for various wire structures can be determined from (5.160).

All other antenna properties, such as gain and radiation pattern, can be derived from the current distributions.

5.5.2 Dipole Antenna

The dipole antenna consists of two arms or poles and was first invented by Hertz around 1886 in his pioneering experiments with radio waves. It is one of the most commonly used types of RF antennas and is of fundamental importance. The dipole antenna can be used on its own or incorporated into other antenna designs as a radiating element. Figure 5.31 shows a typical configuration of dipole antenna excited by a delta gap at $l = l'$. The general solution of (5.160) for the dipole antenna may be assumed to be

$$I(l) = \begin{cases} C_1 \cos kl + C_2 \sin kl, & 0 < l < l' \\ C_3 \cos kl + C_4 \sin kl, & l' < l < L, \end{cases} \tag{5.161}$$

where C_i $(i = 1, 2, 3, 4)$ are constants and can be determined by applying the following conditions

$$I(0) = I(L) = 0, \tag{5.162}$$

$$\left.\frac{dI}{dl}\right|_{l=l'+} - \left.\frac{dI}{dl}\right|_{l=l'-} = -j\omega C_0 V_s. \tag{5.163}$$

It is readily to find that the current distribution for the dipole is given by

$$I(l) = -\frac{j\pi V_s}{\eta \sin kL \ln ka_0}[-\cos k(L - |l - l'|) + \cos k(L - l - l')]. \tag{5.164}$$

For a straight dipole excited at the center, (5.164) reduces to

$$I(l) = \begin{cases} \dfrac{-j\pi V_s}{\eta \cos\left(\frac{kL}{2}\right) \ln ka_0} \sin k(L - l), & l > L/2 \\ \dfrac{-j\pi V_s}{\eta \cos\left(\frac{kL}{2}\right) \ln ka_0} \sin kl, & l < L/2 \end{cases}. \tag{5.165}$$

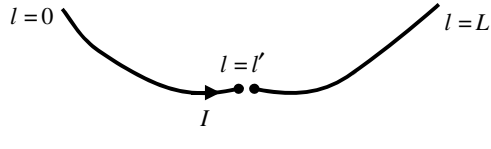


Figure 5.31 An arbitrary dipole antenna excited by a delta gap.

Introducing a new variable $z = l - L/2$, we have

$$I(l) = I_d \sin k\left(\frac{L}{2} - |z|\right), \quad |z| < \frac{L}{2}, \tag{5.166}$$

where $I_d = \frac{-j\pi V_s}{\eta \cos(kL/2) \ln ka_0}$. For a small dipole, (5.166) reduces to

$$I(l) \approx I_d k\left(\frac{L}{2} - |z|\right), \quad |z| < \frac{L}{2}. \tag{5.167}$$

Equations (5.166) and (5.167) are the well-known results for dipole antennas.

Example 5.3 (Fields from center-fed straight dipole antenna)**:** Considering the symmetry of the dipole shown in Figure 5.32. The near fields in the cylindrical system may be obtained from (5.166) as follows

$$\begin{aligned}\mathbf{H} &= \frac{1}{\mu}\nabla \times \mathbf{A} = -\mathbf{u}_\varphi \frac{1}{\mu}\frac{\partial A_z}{\partial \rho} \\ &= -\mathbf{u}_\varphi \frac{I_d}{j4\pi y}\left[e^{-jkR_1} + e^{-jkR_2} - 2\cos\left(\frac{kL}{2}\right)e^{-jkr}\right], \\ \mathbf{E} &= \mathbf{u}_\rho E_\rho + \mathbf{u}_z E_z = \frac{1}{j\omega\varepsilon}\nabla \times \mathbf{H},\end{aligned}$$

with

$$E_\rho = j\frac{\eta I_d}{4\pi y}\left[\left(z - \frac{L}{2}\right)\frac{e^{-jkR_1}}{R_1} + \left(z + \frac{L}{2}\right)\frac{e^{-jkR_2}}{R_2} - 2z\cos\left(\frac{kL}{2}\right)\frac{e^{-jkr}}{r}\right],$$

$$E_z = -j\frac{\eta I_d}{4\pi}\left[\frac{e^{-jkR_1}}{R_1} + \frac{e^{-jkR_2}}{R_2} - 2\cos\left(\frac{kL}{2}\right)\frac{e^{-jkr}}{r}\right],$$

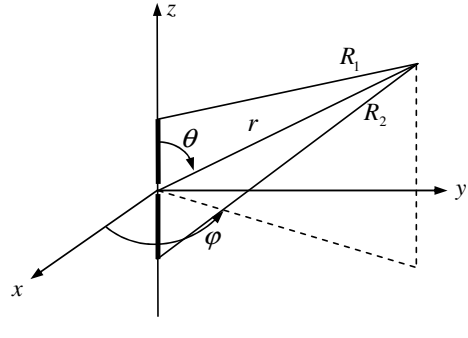


Figure 5.32 Dipole analysis.

where

$$r = \sqrt{x^2 + y^2 + z^2},$$
$$R_1 = \sqrt{x^2 + y^2 + (z - L/2)^2},$$
$$R_2 = \sqrt{x^2 + y^2 + (z + L/2)^2}.$$

The far-fields of the center-fed dipole antenna in the spherical coordinate system can be determined from (5.34) and (5.166) as follows

$$\begin{aligned} E_\theta &= j\eta \frac{I_d e^{-jkr}}{2\pi r} \frac{\cos\left(\frac{kL}{2}\cos\theta\right) - \cos\left(\frac{kL}{2}\right)}{\sin\theta}, \\ H_\varphi &= j \frac{I_d e^{-jkr}}{2\pi r} \frac{\cos\left(\frac{kL}{2}\cos\theta\right) - \cos\left(\frac{kL}{2}\right)}{\sin\theta}. \end{aligned} \tag{5.168}$$

The radiation intensity is

$$U = \eta \frac{|I_d|^2}{8\pi^2} \left[\frac{\cos\left(\frac{kL}{2}\cos\theta\right) - \cos\left(\frac{kL}{2}\right)}{\sin\theta} \right]^2.$$

The radiated power is

$$P_{\text{rad}} = \eta \frac{|I_d|^2}{4\pi} \left\{ \gamma + \ln(kL) - \text{C}_\text{i}(kL) + \frac{1}{2}\sin(kL)[\text{S}_\text{i}(2kL) - 2\text{S}_\text{i}(kL)] \right.$$
$$\left. + \frac{1}{2}\cos(kL)\left[C + \ln(kL/2) + \text{C}_\text{i}(2kL) - 2\text{C}_\text{i}(kL)\right] \right\}$$

where C_i is the cosine integral, and S_i is the **sine integral** defined by

$$\text{S}_\text{i}(x) = \int_0^x \frac{\cos y}{y}\, dy.$$

The input impedance may be evaluated by the induced EMF method

$$Z = -\frac{1}{\left|I_d \sin\left(\frac{kL}{2}\right)\right|^2} \int_{-L/2}^{L/2} \bar{I}_d \sin k\left(\frac{L}{2} - |z'|\right) E_z(\rho = a_0, z')dz = R + jX,$$

where

$$R = \frac{\eta}{2\pi \sin^2\left(\frac{kL}{2}\right)} \left\{ \gamma + \ln(kL) - \text{C}_\text{i}(kL) + \frac{1}{2}\sin(kL)[\text{S}_\text{i}(2kL) - 2\text{S}_\text{i}(kL)] \right.$$
$$\left. + \frac{1}{2}\cos(kL)[\gamma + \ln(kL/2) + \text{C}_\text{i}(2kL) - 2\text{C}_\text{i}(kL)] \right\},$$

$$X = \frac{\eta}{4\pi \sin^2(kL/2)} \left\{ 2\mathrm{S_i}(kL) + \cos(kL)[2\mathrm{S_i}(kL) - \mathrm{S_i}(2kL)] \right.$$
$$\left. - \sin(kL) \left[2\mathrm{C_i}(kL) - \mathrm{C_i}(2kL) - \mathrm{C_i}\left(2k\frac{a_0^2}{L} \right) \right] \right\}.$$

Note that the impedance is defined by the ratio of the voltage over the current at the input terminal. □

5.5.3 Loop Antenna

A loop antenna excited by a delta gap at $l = l'$ is shown in Figure 5.33. In this case, the boundary condition $I(0) = I(L)$ must be imposed. The general solution of (5.160) can be written as

$$I(l) = \begin{cases} C_1 \cos kl + C_2 \sin kl, & 0 < l < l' \\ C_3 \cos kl + C_4 \sin kl, & l' < l < L \end{cases}.$$

Making use of the facts that the current and its derivative must be continuous at $l = 0$ and (5.163), we may find the current distribution for the thin loop

$$I(l) = \frac{j\pi V_s}{\eta \ln ka_0} \frac{\cos k\left(\frac{L}{2} - |l - l'|\right)}{\sin\left(\frac{kl}{2}\right)},$$

where we have used $\frac{\omega C_0}{k} = -\frac{2\pi}{\eta \ln ka_0}$. Without loss of generality, we may set $l' = L/2$ to get

$$I(l) = \frac{j\pi V_s}{\eta \ln ka_0 \sin(kL/2)} \begin{cases} \cos kl, & l < \dfrac{L}{2} \\ \cos k(L - l), & l > \dfrac{L}{2} \end{cases}.$$

Making a substitution $s = l - L/2$, we have

$$I(s) = I_l \cos k\left(\frac{L}{2} - |s|\right), \quad |s| < \frac{L}{2}, \tag{5.169}$$

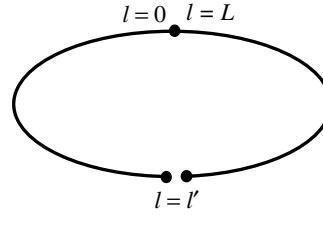


Figure 5.33 An arbitrary loop antenna excited by a delta gap.

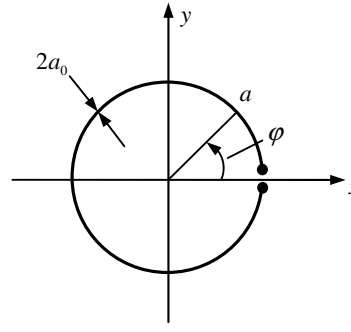


Figure 5.34 A circular loop antenna.

where $I_l = \frac{j\pi V_s}{\eta \ln ka_0 \sin(kL/2)}$. For a small loop, the current distribution may be considered to be a constant

$$I(s) \approx I_1, \quad |s| < \frac{L}{2}.$$

Equation (5.169) can be used to investigate a number of loop antennas, such as circular, triangular and rectangular loops.

Example 5.4 (Circular loop antenna)**:** Consider a circular loop antenna shown in Figure 5.34. The wire radius and the loop radius are respectively denoted by a_0 and a. The loop is excited by a feed at $\varphi = 0$, and the impressed field is assumed to be a delta voltage source

$$E_{\text{in}} = \frac{V_s}{a}\delta(\varphi).$$

An analytical solution for the current distribution on the circular wire loop is found to be (Iizuka *et al.*, 1966; Storer, 1956)

$$I(\varphi) = \frac{V_s}{Z_{00}} + 2V_s \sum_{n=1}^{\infty} \frac{\cos n\varphi}{Z_{nn}}, \tag{5.170}$$

where

$$Z_{nn} = j\pi\eta ka \left[\frac{1}{2}K_{n-1} + \frac{1}{2}K_{n+1} - \left(\frac{n}{ka}\right)^2 K_n\right],$$

$$K_0 = \frac{1}{\pi}\ln\frac{8}{(a_0/a)} - \frac{1}{2}\int_0^{2ka} [\Omega_0(x) + jJ_0(x)]dx,$$

$$K_{n+1} = K_n + \Omega_{2n+1}(ka) + jJ_{2n+1}(ka),$$

and

$$\Omega_n(x) = \frac{1}{\pi}\int_0^{\pi} \sin(x\sin\theta - n\theta)d\theta$$

is **Lommel–Weber function**. Numerical results indicate that the analytical solution (5.170) agrees very well with the asymptotic solution (5.169) for the thin circular wire loop. The input admittance of the loop antenna is given by

$$Y = \frac{I(0)}{V_s} = \frac{1}{Z_{00}} + 2\sum_{n=1}^{\infty}\frac{1}{Z_{nn}}. \tag{5.171}$$

The φ-component of the radiated fields can be written as

$$E_\varphi = -j\omega A_\varphi - \frac{1}{a}\frac{\partial\phi}{\partial\varphi}$$

with

$$A_\varphi = \mu\int_0^{2\pi} I(\varphi')\cos(\varphi-\varphi')\frac{e^{-jkR}}{4\pi R}ad\varphi',$$

$$\phi = \frac{-1}{j\omega\varepsilon}\int_0^{2\pi}\frac{dI(\varphi')}{ad\phi'}\cos(\varphi-\varphi')\frac{e^{-jkR}}{4\pi R}ad\varphi',$$

where

$$R = \sqrt{r^2 + a^2 - 2ar\sin\theta\cos(\varphi-\varphi')}.$$

For small loop $a \ll 1$, the function $f(a)$ defined by

$$f(a) = \frac{e^{-jk\sqrt{r^2+a^2-2ar\sin\theta\cos(\varphi-\varphi')}}}{\sqrt{r^2+a^2-2ar\sin\theta\cos(\varphi-\varphi')}}$$

may be expanded as a Taylor series

$$f(a) = f(0) + f'(0) + \frac{1}{2!}f''(0)a^2 + \cdots,$$

where

$$f(0) = \frac{e^{-jkr}}{r}, \quad f'(0) = \left(\frac{jk}{r} + \frac{1}{r^2}\right)e^{-jkr}\sin\theta\cos\varphi'.$$

If the current is assumed to be constant $I(\varphi) = I_l$, the field components are found to be

$$E_\varphi = \eta\frac{I_l(ka)^2\sin\theta}{4r}\left(1 + \frac{1}{jkr}\right)e^{-jkr},$$

$$H_r = j\frac{I_l ka^2 \cos\theta}{2r^2}\left(1+\frac{1}{jkr}\right)e^{-jkr},$$

$$H_\theta = -\frac{I_l (ka)^2 \sin\theta}{4r}\left(1+\frac{1}{jkr}-\frac{1}{(kr)^2}\right)e^{-jkr}.$$

Other field components are zero. □

5.6 Slot Antennas

The slot antenna was invented in 1938 by English engineer Alan Blumlein (1903–1942). A **slot antenna** consists of a metal surface with a hole or slot cut out. When the slot is driven by an incident field, it radiates electromagnetic waves in similar way to a dipole antenna. Slot antennas are often used at UHF and microwave frequencies and are widely used in radar, cell phone base stations, and can be best understood by Babinet's principle.

5.6.1 Babinet's Principle

In optics, Babinet's principle states that the sum between the field behind a screen with an opening and the field of a complementary structure is equal to the field when there is no screen. An extension of Babinet's principle was introduced by Booker (1946). Assume that an electric current $\mathbf{J}$ generates the fields $\mathbf{E}_{\text{in}}, \mathbf{H}_{\text{in}}$ in an unbounded medium of intrinsic impedance $\sqrt{\mu/\varepsilon}$ [Figure 5.35(a)]. The extended Babinet's principle can be expressed as

$$\mathbf{E}_{\text{in}} = \mathbf{E}_e + \mathbf{E}_m, \quad \mathbf{H}_{\text{in}} = \mathbf{H}_e + \mathbf{H}_m, \tag{5.172}$$

where $\mathbf{E}_e, \mathbf{H}_e$ are the fields produced by the current source $\mathbf{J}$ in the presence of an infinite conducting screen with an opening S_a in the same medium [see Figure 5.35(b)], and $\mathbf{E}_m, \mathbf{H}_m$ are the fields produced by the current source $\mathbf{J}$ in the presence of a thin magnetic conductor S_a in the same medium [see Figure 5.35(c)]. According to the duality, the problem shown in Figure 5.35(c) is equivalent to the problem shown in Figure 5.35(d) by replacing the magnetic conductor S_a with electric conductor S_a, $\mathbf{J}$ with $\mathbf{J}_m$, $\mathbf{E}_m$ with $\mathbf{H}_d$, $\mathbf{H}_m$ with $-\mathbf{E}_d$, μ with ε and ε with μ. As a result, (5.172) can be written as

$$\mathbf{E}_{\text{in}} = \mathbf{E}_e + \mathbf{H}_d, \quad \mathbf{H}_{\text{in}} = \mathbf{H}_e - \mathbf{E}_d. \tag{5.173}$$

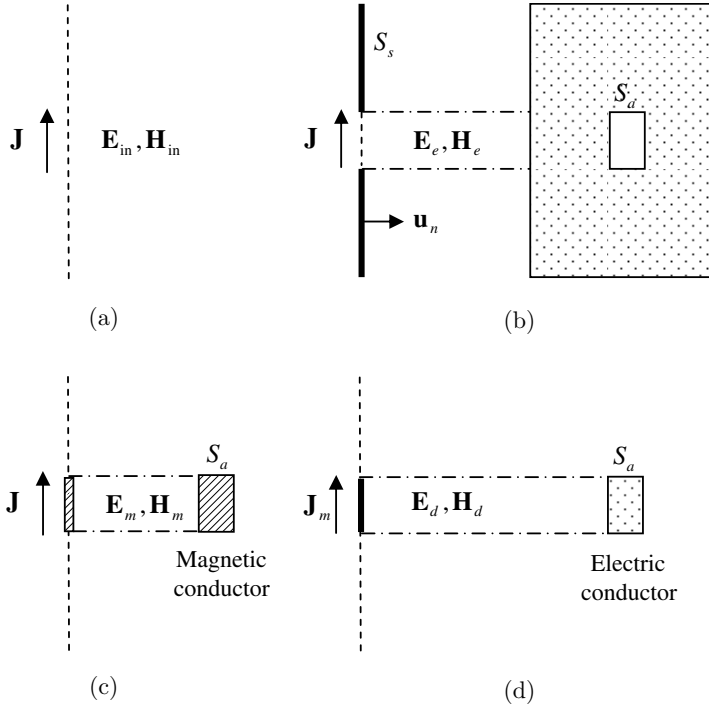


Figure 5.35 Babinet's principle.

It is noted that the intrinsic impedance of the medium in Figure 5.35(d) is $\sqrt{\varepsilon/\mu}$. The fields $\mathbf{E}_d, \mathbf{H}_d$ satisfy

$$\begin{aligned} \nabla \times \mathbf{E}_d(\mathbf{r}) &= -j\omega\varepsilon \mathbf{H}_d(\mathbf{r}) - \mathbf{J}_m(\mathbf{r}), \\ \nabla \times \mathbf{H}_d(\mathbf{r}) &= j\omega\mu \mathbf{E}_d(\mathbf{r}). \end{aligned} \tag{5.174}$$

Let $\eta = \sqrt{\mu/\varepsilon}$. The above equations can be written as

$$\begin{aligned} \nabla \times \mathbf{E}_{d1}(\mathbf{r}) &= -j\omega\mu \mathbf{H}_{d1}(\mathbf{r}) - \eta \mathbf{J}_m(\mathbf{r}), \\ \nabla \times \mathbf{H}_{d1}(\mathbf{r}) &= j\omega\varepsilon \mathbf{E}_{d1}(\mathbf{r}), \end{aligned} \tag{5.175}$$

where

$$\mathbf{E}_{d1}(\mathbf{r}) = \eta \mathbf{E}_d(\mathbf{r}), \quad \mathbf{H}_{d1}(\mathbf{r}) = \eta^{-1} \mathbf{H}_d(\mathbf{r}).$$

Equations (5.175) describe a complementary problem shown in Figure 5.36, where the intrinsic impedance of the medium is η instead of η^{-1} as in Figure 5.35(d). Thus we may write (5.173) as

$$\mathbf{E}_{\text{in}} = \mathbf{E}_e + \eta \mathbf{H}_{d1}, \quad \mathbf{H}_{\text{in}} = \mathbf{H}_e - \eta^{-1} \mathbf{E}_{d1}. \tag{5.176}$$

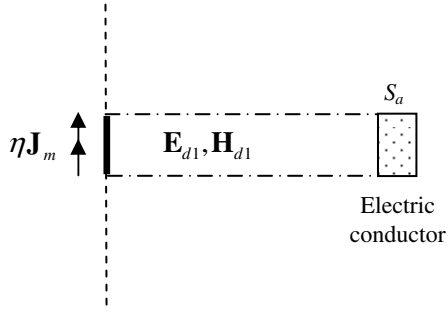


Figure 5.36 Complementary problem in a medium with intrinsic impedance η.

When the conducting screen with an opening S_a in Figure 5.35(b) and the electric conductor S_a in Figure 5.35(d) are combined, they form a solid screen. For this reason, they are called **complementary structures**. The Babinet's principle can be easily demonstrated as follows. For the problem shown Figure 5.35(b), the incident fields are $\mathbf{E}_{\text{in}}, \mathbf{H}_{\text{in}}$ and we have

$$\begin{cases} \mathbf{u}_n \times \mathbf{E}_e = 0, & \text{on } S_s \\ \mathbf{u}_n \times \mathbf{H}_e = \mathbf{u}_n \times \mathbf{H}_{\text{in}}, & \text{on } S_a \end{cases}. \tag{5.177}$$

The second equation comes from the fact that the induced current on the conducting screen due to the incident fields does not generates a tangential magnetic field at the aperture S_a on the same screen plane. Similarly for the problem shown in Figure 5.35(c), we have

$$\begin{cases} \mathbf{u}_n \times \mathbf{E}_m = \mathbf{u}_n \times \mathbf{E}_{\text{in}}, & \text{on } S_s \\ \mathbf{u}_n \times \mathbf{H}_m = 0, & \text{on } S_a \end{cases}. \tag{5.178}$$

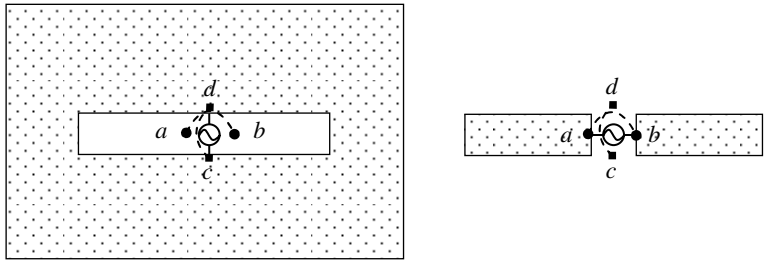
Adding (5.177) and (5.178) yields

$$\begin{cases} \mathbf{u}_n \times (\mathbf{E}_e + \mathbf{E}_m) = \mathbf{u}_n \times \mathbf{E}_{\text{in}}, & \text{on } S_s \\ \mathbf{u}_n \times (\mathbf{H}_e + \mathbf{H}_m) = \mathbf{u}_n \times \mathbf{H}_{\text{in}}, & \text{on } S_a \end{cases}. \tag{5.179}$$

From uniqueness theorem and (5.179), we immediately obtain (5.172).

5.6.2 Impedance of Slot Antennas

Figure 5.37 shows a slot antenna in an infinite conducting plane and its complementary dipole antenna, both excited by a delta voltage source. The voltage across the dipole feed is the line integral of the electric field over



(a) (b)

Figure 5.37 (a) Slot antenna. (b) Complementary dipole antenna.

the arc path ab:

$$V = \int_{ab} \mathbf{E}_{d1} \cdot \mathbf{u}_l \, dl.$$

The current at the feeding terminal of the dipole is the line integral of the magnetic field over the arc path cd

$$I = -2 \int_{cd} \mathbf{H}_{d1} \cdot \mathbf{u}_l \, dl,$$

where the factor 2 is due to that only one-half of the closed line integral is taken. The input impedance of the dipole is given by

$$Z_d = \frac{V}{I} = \frac{\int_{ab} \mathbf{E}_{d1} \cdot \mathbf{u}_l \, dl}{-2 \int_{cd} \mathbf{H}_{d1} \cdot \mathbf{u}_l \, dl}. \tag{5.180}$$

Similarly the voltage across the feed of the slot is the line integral of the electric field over the arc path cd

$$V = -\int_{cd} \mathbf{E}_e \cdot \mathbf{u}_l \, dl.$$

The current at the feed of the slot antenna is the line integral of the magnetic field over the arc path ab

$$I = -2 \int_{ab} \mathbf{H}_e \cdot \mathbf{u}_l \, dl.$$

The input impedance of the slot is then given by

$$Z_s = \frac{V}{I} = \frac{-\int_{cd} \mathbf{E}_e \cdot \mathbf{u}_l \, dl}{-2\int_{ab} \mathbf{H}_e \cdot \mathbf{u}_l \, dl}. \tag{5.181}$$

We now assume that the incident fields are highly localized at the feeding point. From Babinet's principle, we have

$$\mathbf{E}_{\text{in}} = \mathbf{E}_e + \eta \mathbf{H}_{d1} = 0, \quad \mathbf{H}_{\text{in}} = \mathbf{H}_e - \eta^{-1}\mathbf{E}_{d1} = 0 \tag{5.182}$$

outside the feeding point. Making use of (5.182) we obtain

$$Z_s Z_d = \frac{-\int_{ab} \mathbf{E}_{d1} \cdot \mathbf{u}_l \, dl}{-2\int_{cd} \mathbf{H}_{d1} \cdot \mathbf{u}_l \, dl} \times \frac{\int_{cd} \mathbf{E}_e \cdot \mathbf{u}_l \, dl}{-2\int_{ab} \mathbf{H}_e \cdot \mathbf{u}_l \, dl} = \frac{1}{4}\eta^2. \tag{5.183}$$

This is an important relationship in antenna theory. When the shape of the complementary dipole antenna is identical to its complementary conducting screen, we have $Z_s = Z_d$. In this case, frequency-independent input impedance can be achieved

$$Z_d = \frac{1}{2}\eta. \tag{5.184}$$

The antenna that satisfies (5.184) is called **self-complementary antenna** (Mushiake, 1996).

5.7 Aperture Antennas

Some typical aperture antennas are shown in Figure 5.3. An aperture antenna consisting of perfect conductors can be characterized by a generic antenna model shown in Figure 5.38. In this model, the antenna is assumed to include all possible sources, and it may be in transmitting mode, receiving mode or in a mode that the antenna transmits and receives at the same time (e.g., antenna is in transmitting mode but interfered by an arbitrary incident field from the outside of antenna). The source region V_0 of the antenna is chosen in such a way that its boundary ∂V_0 is coincident with the antenna surface, which is assumed to be a perfect conductor (except for cross sectional portion Ω where ∂V_0 crosses the antenna terminal). Let ∂V_∞ be a large surface that encloses the whole antenna system. From the representation theorem for electromagnetic fields,

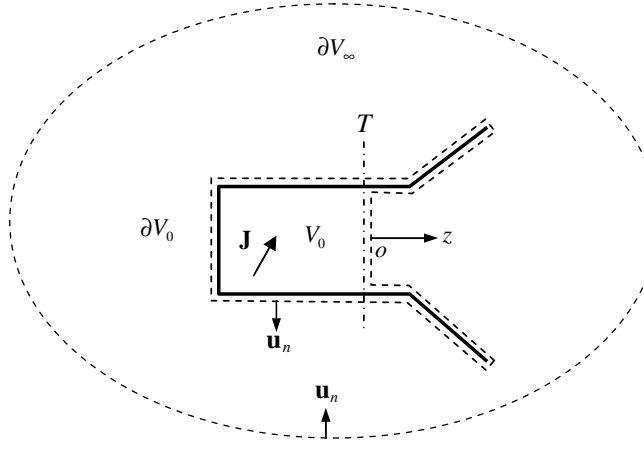


Figure 5.38 An arbitrary aperture antenna.

the total magnetic field in the region bounded by ∂V_0 and ∂V_∞ can then be expressed as

$$\mathbf{E}(\mathbf{r}) = -jk\eta \int_{\partial V_0} G(\mathbf{r},\mathbf{r}')\mathbf{J}_s(\mathbf{r}')dS(\mathbf{r}') - \int_{\partial V_0} \mathbf{J}_{ms}(\mathbf{r}') \times \nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$-\frac{\eta}{jk} \int_{\partial V_0} \nabla_s \cdot \mathbf{J}_s(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}') + \mathbf{E}_{\text{in}}^{\text{ext}}(\mathbf{r}),$$

$$\mathbf{H}(\mathbf{r}) = -j\frac{k}{\eta} \int_{\partial V_0} G(\mathbf{r},\mathbf{r}')\mathbf{J}_{ms}(\mathbf{r}')dS(\mathbf{r}') + \int_{\partial V_0} \mathbf{J}_s(\mathbf{r}') \times \nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$-\frac{1}{jk\eta} \int_{\partial V_0} \nabla_s \cdot \mathbf{J}_{ms}(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}') + \mathbf{H}_{\text{in}}^{\text{ext}}(\mathbf{r}),$$

where $\eta = \sqrt{\mu/\varepsilon}$, $\mathbf{J}_s = \mathbf{u}_n \times \mathbf{H}$, $\mathbf{J}_{ms} = -\mathbf{u}_n \times \mathbf{E}$, and $G(\mathbf{r},\mathbf{r}') = e^{-jk|\mathbf{r}-\mathbf{r}'|}/4\pi|\mathbf{r}-\mathbf{r}'|$ is the Green's function in free space, and

$$\mathbf{E}_{\text{in}}^{\text{ext}}(\mathbf{r}) = -\int_{\partial V_\infty} jk\eta G(\mathbf{r},\mathbf{r}')\mathbf{J}_s(\mathbf{r}')dS(\mathbf{r}') - \int_{\partial V_\infty} \mathbf{J}_{ms}(\mathbf{r}') \times \nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$-\frac{\eta}{jk} \int_{\partial V_\infty} \nabla_s \cdot \mathbf{J}_s(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}'),$$

$$\mathbf{H}_{\text{in}}^{\text{ext}}(\mathbf{r}) = -j\frac{k}{\eta}\int\limits_{\partial V_\infty} G(\mathbf{r},\mathbf{r}')\mathbf{J}_{ms}(\mathbf{r}')dS(\mathbf{r}') + \int\limits_{\partial V_\infty} \mathbf{J}_s(\mathbf{r}')\times\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$-\frac{1}{jk\eta}\int\limits_{\partial V_\infty} \nabla_s\cdot\mathbf{J}_{ms}(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$

represent the external incident fields. By letting the observation point $\mathbf{r}$ approach the boundary of the source region ∂V_0 from the interior of $\partial V_0 + \partial V_\infty$ and using the jump relations, we have

$$\mathbf{E}(\mathbf{r}) = -jk\eta\int\limits_{\partial V_0} G(\mathbf{r},\mathbf{r}')\mathbf{J}_s(\mathbf{r}')dS(\mathbf{r}') - \int\limits_{\partial V_0}\mathbf{J}_{ms}(\mathbf{r}')\times\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$-\frac{\eta}{jk}\int\limits_{\partial V_0}\nabla_s\cdot\mathbf{J}_s(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}') + \mathbf{E}_{\text{in}}^{\text{ext}}(\mathbf{r})$$
$$-\frac{1}{2}\mathbf{J}_{ms}(\mathbf{r})\times\mathbf{u}_n(\mathbf{r}) - \frac{\eta}{2jk}\mathbf{u}_n(\mathbf{r})\nabla_s\cdot\mathbf{J}_s(\mathbf{r}),$$
$$\mathbf{H}(\mathbf{r}) = -j\frac{k}{\eta}\int\limits_{\partial V_0} G(\mathbf{r},\mathbf{r}')\mathbf{J}_{ms}(\mathbf{r}')dS(\mathbf{r}') + \int\limits_{\partial V_0}\mathbf{J}_s(\mathbf{r}')\times\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$-\frac{1}{jk\eta}\int\limits_{\partial V_0}\nabla_s\cdot\mathbf{J}_{ms}(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}') + \mathbf{H}_{\text{in}}^{\text{ext}}(\mathbf{r})$$
$$+\frac{1}{2}\mathbf{J}_s(\mathbf{r})\times\mathbf{u}_n(\mathbf{r}) - \frac{1}{j2k\eta}\mathbf{u}_n(\mathbf{r})\nabla_s\cdot\mathbf{J}_{ms}(\mathbf{r}).$$

Multiplying both sides by $\mathbf{u}_n$, we get

$$-\frac{1}{2}\mathbf{J}_{ms}(\mathbf{r}) = -jk\eta\mathbf{u}_n(\mathbf{r})\times\int\limits_{\partial V_0} G(\mathbf{r},\mathbf{r}')\mathbf{J}_s(\mathbf{r}')dS(\mathbf{r}') - \mathbf{u}_n(\mathbf{r})$$
$$\times\int\limits_{\partial V_0}\mathbf{J}_{ms}(\mathbf{r}')\times\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}') - \frac{\eta}{jk}\mathbf{u}_n(\mathbf{r})$$
$$\times\int\limits_{\partial V_0}\nabla_s\cdot\mathbf{J}_s(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}') + \mathbf{u}_n(\mathbf{r})\times\mathbf{E}_{\text{in}}^{\text{ext}}(\mathbf{r})$$

$$\frac{1}{2}\mathbf{J}_s(\mathbf{r}) = -j\frac{k}{\eta}\mathbf{u}_n(\mathbf{r}) \times \int\limits_{\partial V_0} G(\mathbf{r},\mathbf{r}')\mathbf{J}_{ms}(\mathbf{r}')dS(\mathbf{r}') + \mathbf{u}_n(\mathbf{r}) \times \int\limits_{\partial V_0} \mathbf{J}_s(\mathbf{r}') \times \nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}') - \frac{1}{jk\eta}\mathbf{u}_n(\mathbf{r}) \times \int\limits_{\partial V_0} \nabla_s \cdot \mathbf{J}_{ms}(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}') + \mathbf{u}_n(\mathbf{r}) \times \mathbf{H}_{\text{in}}^{\text{ext}}(\mathbf{r}).$$

Making use of the boundary conditions on the metal surface of the antenna, the above equations can be written as

$$\mathbf{u}_n(\mathbf{r}) \times \int\limits_{\partial V_0} \left[jk\eta G(\mathbf{r},\mathbf{r}')\mathbf{J}_s(\mathbf{r}') + \frac{\eta}{jk}\nabla_s' \cdot \mathbf{J}_s(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')\right] dS(\mathbf{r}')$$
$$= \frac{1}{2}\mathbf{J}_{ms}(\mathbf{r})U_\Omega(\mathbf{r}) + \mathbf{u}_n(\mathbf{r}) \times \left[\mathbf{E}_{\text{in}}^{\text{int}}(\mathbf{r}) + \mathbf{E}_{\text{in}}^{\text{ext}}(\mathbf{r})\right], \tag{5.185}$$

$$-\frac{1}{2}\mathbf{J}_s(\mathbf{r}) + \mathbf{u}_n(\mathbf{r}) \times \int\limits_{\partial V_0} \mathbf{J}_s(\mathbf{r}') \times \nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}')$$
$$= -\mathbf{u}_n(\mathbf{r}) \times \left[\mathbf{H}_{\text{in}}^{\text{int}}(\mathbf{r}) + \mathbf{H}_{\text{in}}^{\text{ext}}(\mathbf{r})\right], \tag{5.186}$$

where $U_\Omega(\mathbf{r}) = 1$ for $\mathbf{r} \in \Omega$ or $U_\Omega(\mathbf{r}) = 0$ for $\mathbf{r} \notin \Omega$ and

$$\mathbf{E}_{\text{in}}^{\text{int}}(\mathbf{r}) = -\int\limits_{\Omega} \mathbf{J}_{ms}(\mathbf{r}') \times \nabla' G(\mathbf{r},\mathbf{r}')d\Omega(\mathbf{r}'), \tag{5.187}$$

$$\mathbf{H}_{\text{in}}^{\text{int}}(\mathbf{r}) = -j\frac{k}{\eta}\int\limits_{\Omega} G(\mathbf{r},\mathbf{r}')\mathbf{J}_{ms}(\mathbf{r}')d\Omega(\mathbf{r}') - \frac{1}{jk\eta}\int\limits_{\Omega} \nabla_s' \cdot \mathbf{J}_{ms}(\mathbf{r}')\nabla' G(\mathbf{r},\mathbf{r}')d\Omega(\mathbf{r}'). \tag{5.188}$$

The fields $\mathbf{E}_{\text{in}}^{\text{int}}(\mathbf{r})$ and $\mathbf{H}_{\text{in}}^{\text{int}}(\mathbf{r})$ are determined by the equivalent surface magnetic current $\mathbf{J}_m = -\mathbf{u}_z \times \mathbf{E}$ on the antenna input plane. To determine the equivalent magnetic current on the input plane, we must use the excitation conditions. The fields in the feeding waveguide may be

expressed as

$$-\mathbf{u}_z \times \mathbf{E}(\mathbf{r}) = -\sum_{n=1}^{\infty} \mathbf{u}_z \times \mathbf{e}_n(\mathbf{r}) V_n(z), \quad \mathbf{u}_z \times \mathbf{H}(\mathbf{r}) = -\sum_{n=1}^{\infty} \mathbf{e}_n(\mathbf{r}) I_n(z), \tag{5.189}$$

where

$$V_n(z) = A_n e^{-j\beta_n z} + B_n e^{j\beta_n z},$$
$$I_n(z) = Z_{wn}^{-1}(A_n e^{-j\beta_n z} - B_n e^{j\beta_n z}),$$
$$\beta_n = \begin{cases} k, & \text{TEM mode} \\ \sqrt{k^2 - k_{cn}^2}, & \text{TE or TM mode} \end{cases},$$
$$Z_{wn} = \begin{cases} \eta, & \text{TEM mode} \\ \eta k/\beta_n, & \text{TE mode} \\ \eta\beta_n/k, & \text{TM mode} \end{cases}.$$

We assume that the feeding waveguide of antenna is in a single-mode operation. Therefore, the modal voltage and current may be written as

$$V_1(z) = \delta e^{-j\beta_1 z} + B_1 e^{j\beta_1 z}, \quad V_n(z) = B_n e^{j\beta_n z} (n \geq 2),$$
$$I_1(z) = (\delta e^{-j\beta_1 z} - B_1 e^{j\beta_1 z}) Z_{w1}^{-1}, \quad I_n(z) = -B_n e^{j\beta_n z} Z_{wn}^{-1} (n \geq 2),$$

where δ is unit for a transmitting antenna excited by dominant mode of unit amplitude and is zero for a receiving antenna. Thus on the input plane ($z = 0$), (5.189) may be written as

$$\mathbf{J}_{ms}(\mathbf{r}) = -\mathbf{u}_z \times \mathbf{e}_1(\mathbf{r})(\delta + B_1) - \sum_{n=2}^{\infty} \mathbf{u}_z \times \mathbf{e}_n(\mathbf{r}) B_n,$$

$$\mathbf{J}_s(\mathbf{r}) = -\mathbf{e}_1(\mathbf{r})(\delta - B_1) Z_{w1}^{-1} + \sum_{n=2}^{\infty} \mathbf{e}_n(\mathbf{r}) B_n Z_{wn}^{-1}.$$

The expansion coefficients can be determined by the second equation of the above equations

$$B_1 = \delta + Z_{w1} \int_{\Omega} \mathbf{J}_s \cdot \mathbf{e}_1 \, d\Omega, \quad B_n = Z_{wn} \int_{\Omega} \mathbf{J}_s \cdot \mathbf{e}_n \, d\Omega.$$

The impedance of the antenna is given by

$$Z_{\text{in}} = \frac{\delta + B_1}{\delta - B_1} Z_{w1}.$$

Hence the equivalent magnetic current on the reference plane may be expressed by

$$\mathbf{J}_{ms}(\mathbf{r}) = -2\delta \mathbf{u}_z \times \mathbf{e}_1(\mathbf{r}) - \sum_{n=1}^{\infty} \mathbf{u}_z \times \mathbf{e}_n(\mathbf{r}) Z_{wn} \int_{\Omega} \mathbf{J}_s(\mathbf{r}') \cdot \mathbf{e}_n(\mathbf{r}') d\Omega(\mathbf{r}'), \tag{5.190}$$

Inserting this into (5.187) and (5.188), we obtain

$$\mathbf{E}_{\text{in}}^{\text{int}}(\mathbf{r}) = 2\delta \mathbf{G}_{e1}(\mathbf{r}) + \sum_{n=1}^{\infty} Z_{wn} \mathbf{G}_{en}(\mathbf{r}) \int_{\Omega} \mathbf{J}_s(\mathbf{r}') \cdot \mathbf{e}_n(\mathbf{r}') d\Omega(\mathbf{r}'), \tag{5.191}$$

$$\mathbf{H}_{\text{in}}^{\text{int}}(\mathbf{r}) = 2\delta \mathbf{G}_{h1}(\mathbf{r}) + \sum_{n=1}^{\infty} Z_{wn} \mathbf{G}_{hn}(\mathbf{r}) \int_{\Omega} \mathbf{J}_s(\mathbf{r}') \cdot \mathbf{e}_n(\mathbf{r}') d\Omega(\mathbf{r}'), \tag{5.192}$$

where $\mathbf{G}_{en}$ and $\mathbf{G}_{hn}$ are defined by

$$\mathbf{G}_{en}(\mathbf{r}) = \int_{\Omega} [\mathbf{u}_z \times \mathbf{e}_n(\mathbf{r}')] \times \nabla' G(\mathbf{r}, \mathbf{r}') d\Omega(\mathbf{r}'),$$

$$\mathbf{G}_{hn}(\mathbf{r}) = \frac{jk}{\eta} \left\{ \int_{\Omega} G(\mathbf{r}, \mathbf{r}') \mathbf{u}_z \times \mathbf{e}_n(\mathbf{r}') d\Omega(\mathbf{r}') - \frac{1}{k^2} \int_{\Omega} \nabla_s' \cdot [\mathbf{u}_z \times \mathbf{e}_n(\mathbf{r}')] \nabla' G(\mathbf{r}, \mathbf{r}') d\Omega(\mathbf{r}') \right\}.$$

Thus Equations (5.185) and (5.186) can be written as

$$\begin{aligned} &\mathbf{u}_n(\mathbf{r}) \times \int_{\partial V_0} \left[jk\eta G(\mathbf{r}, \mathbf{r}') \mathbf{J}_s(\mathbf{r}') + \frac{\eta}{jk} \nabla_s' \cdot \mathbf{J}_s(\mathbf{r}') \nabla' G(\mathbf{r}, \mathbf{r}') \right] dS(\mathbf{r}') \\ &\quad + \frac{1}{2} U_{\Omega}(\mathbf{r}) \sum_{n=1}^{\infty} \mathbf{u}_z \times \mathbf{e}_{tn}(\mathbf{r}) Z_{wn} \int_{\Omega} \mathbf{J}_s(\mathbf{r}') \cdot \mathbf{e}_n(\mathbf{r}') d\Omega(\mathbf{r}') \\ &\quad - \mathbf{u}_n(\mathbf{r}) \times \sum_{n=1}^{\infty} Z_{wn} \mathbf{G}_{en}(\mathbf{r}) \int_{\Omega} \mathbf{J}_s(\mathbf{r}') \cdot \mathbf{e}_n(\mathbf{r}') d\Omega(\mathbf{r}') = \mathbf{F}_e(\mathbf{r}), \end{aligned} \tag{5.193}$$

$$\begin{aligned} &-\frac{1}{2} \mathbf{J}_s(\mathbf{r}) + \mathbf{u}_n(\mathbf{r}) \times \int_{\partial V_0} \mathbf{J}_s(\mathbf{r}') \times \nabla' G(\mathbf{r}, \mathbf{r}') dS(\mathbf{r}') \\ &\quad + \mathbf{u}_n(\mathbf{r}) \times \sum_{n=1}^{\infty} Z_{wn} \mathbf{G}_{hn}(\mathbf{r}) \int_{A} \mathbf{J}_s(\mathbf{r}') \cdot \mathbf{e}_n(\mathbf{r}') d\Omega(\mathbf{r}') = \mathbf{F}_h(\mathbf{r}), \end{aligned} \tag{5.194}$$

where

$$\mathbf{F}_e(\mathbf{r}) = \delta[-\mathbf{u}_z \times \mathbf{e}_1(\mathbf{r})U_\Omega(\mathbf{r}) + 2\mathbf{u}_n(\mathbf{r}) \times \mathbf{G}_{e1}(\mathbf{r})] + \mathbf{u}_n(\mathbf{r}) \times \mathbf{E}_{\text{in}}^{\text{ext}}(\mathbf{r}),$$
$$\mathbf{F}_h(\mathbf{r}) = -2\delta\mathbf{u}_n(\mathbf{r}) \times \mathbf{G}_{h1}(\mathbf{r}) - \mathbf{u}_n(\mathbf{r}) \times \mathbf{H}_{\text{in}}^{\text{ext}}(\mathbf{r}).$$

Equation (5.193) is called **electric field integral equation** (EFIE), and (5.194) is called **magnetic field integral equation** (MFIE). Both can be used to find the current distribution on the antenna surface. Once the current distribution is known, all other properties of the antenna can then be determined (Geyi, 2006b).

5.8 Microstrip Patch Antennas

A microstrip patch antenna consists of a metallic patch bonded to an insulating dielectric substrate with a metal layer (ground) bonded to the opposite side of the substrate, as depicted in Figure 5.39 for a rectangular patch. The metallic patch can take any shapes, such as rectangular, triangular, circular, disk sector, elliptical, annular ring and square ring. The main advantages of microstrip patch antennas are that they are low profile, low cost and light weight; they can be shaped to conform to curved surfaces, and are easy to integrate with other circuits and form large arrays; and they allow both linear and circular polarizations. The microstrip patch antennas also have some disadvantages such as low gain, low efficiency, low power-handling capability and narrow bandwidth. Typical dimensions for rectangular patches are

$$\frac{\lambda}{3} < a < \frac{\lambda}{2}, \quad 0.003\lambda < h < 0.05\lambda.$$

The dielectric constants of the substrate are in the range of $2.2 < \varepsilon_r < 12$.

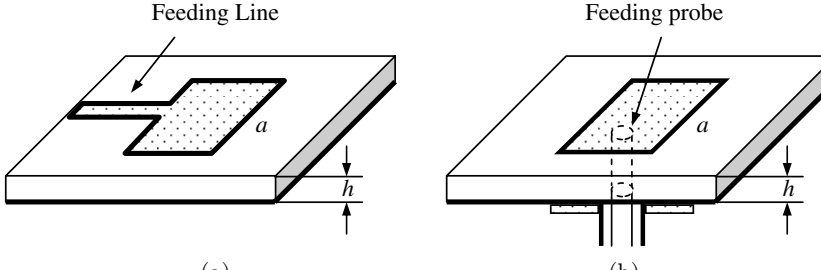


Figure 5.39 Microstrip antennas. (a) Fed by a microstrip. (b) Fed by a coaxial probe.

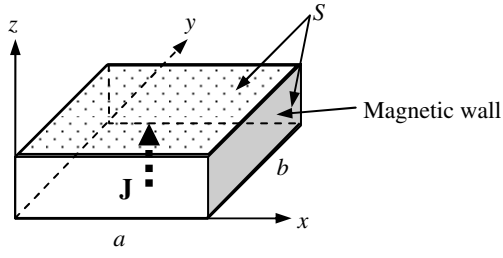


Figure 5.40 Cavity model.

The microstrip patch antennas can be fed by a microstrip line [Figure 5.39(a)] or by a coaxial line with an inner conductor terminated on the patch [Figure 5.39(b)]. In both cases, the induced source of the microstrip patch antennas can be represented by a current distribution $\mathbf{J} = \mathbf{u}_z J_z$, which is independent of z due to $h << \lambda$. This implies that the charge distribution $\rho = 0$. A magnetic side wall may be introduced along the perimeter of the patch to simulate the open circuit, as illustrated in Figure 5.40. Let V denote the region bounded by a closed surface S which consists of the lower surface of the top patch, the upper surface of the bottom ground plane and the side wall. In the interior of the region V, the z-component of the electric field satisfies

$$(\nabla^2 + k^2)E_z(x, y) = j\omega\mu J_z(x, y), \tag{5.195}$$

where $k = \omega\sqrt{\mu\varepsilon}$ is the wavenumber in the dielectric substrate. The above equation can be solved by using the orthonormal set of the eigenfunctions e_{zmn} of the corresponding homogeneous equation, which satisfy

$$\begin{cases} (\nabla^2 + k_{mn}^2)e_{zmn}(x, y) = 0, & (x, y) \in P \\ \dfrac{\partial e_{zmn}}{\partial n} = 0, & (x, y) \in \partial P \end{cases} \tag{5.196}$$

and

$$\int_P e_{zmn}(x, y)e_{zm'n'}(x, y)dx\,dy = \delta_{mm'}\delta_{nn'},$$

where P denotes the patch area and ∂P its boundary. Thus the electric field may be expanded as follows

$$E_z(x, y) = \sum_{m,n} a_{mn}e_{zmn}(x, y).$$

Substituting the expansion into (5.195) and using the orthogonal property of the eigenfunctions, we obtain

$$E_z(x,y) = j\omega\mu \sum_{m,n} \frac{e_{zmn}(x,y)}{k^2 - k_{mn}^2} \int_P J_z(x',y') e_{zmn}(x',y') dx'dy'. \quad (5.197)$$

The magnetic field may be found by

$$\mathbf{H}(x,y) = \frac{1}{jk\eta} \mathbf{u}_z \times \nabla E_z(x,y).$$

Once the electromagnetic fields are known, the equivalent magnetic current on S can be determined by $\mathbf{J}_m = -\mathbf{u}_n \times \mathbf{E}$, where $\mathbf{u}_n$ is the outward normal of S. Apparently, $\mathbf{J}_{ms}$ vanishes on the top patch and ground. Ignoring the sources outside S (i.e., the currents on the upper surface of the patch and the ground, the bound sources in the dielectric substrate), the radiated fields by the microstrip patch antenna can be determined from the equivalent magnetic current on the side wall. For a thin circular probe located at (x_0, y_0), we may assume that

$$J_z(x,y) = \frac{I}{2\pi a_0} \delta(x - x_0)\delta(y - y_0),$$

where a_0 is the radius of the probe. The voltage drop along the probe is then given by

$$V = \int_0^h \mathbf{E}(x_0,y_0) \cdot \mathbf{u}_z dz = hE_z(x_0,y_0) = \frac{jk\eta Ih}{2\pi a_0} \sum_{m,n} \frac{e_{zmn}^2(x_0,y_0)}{k^2 - k_{mn}^2}. \quad (5.198)$$

The input impedance is found to be

$$Z = \frac{V}{-I} = -\frac{jk\eta h}{2\pi a_0} \sum_{m,n} \frac{e_{zmn}^2(x_0,y_0)}{k^2 - k_{mn}^2}. \quad (5.199)$$

Example 5.5: For a rectangular patch shown in Figure 5.40, we have

$$e_{zmn}(x,y) = N_{mn} \cos\frac{m\pi}{a} x \cos\frac{n\pi}{a} y, \quad (5.200)$$

where

$$N_{mn} = \frac{C_{mn}}{\sqrt{hab}}, \quad C_{mn} = \begin{cases} 1, & m = n = 0 \\ \sqrt{2}, & m = 0 \text{ or } n = 0 \\ 2, & m \neq 0, \quad n \neq 0 \end{cases}.$$

□

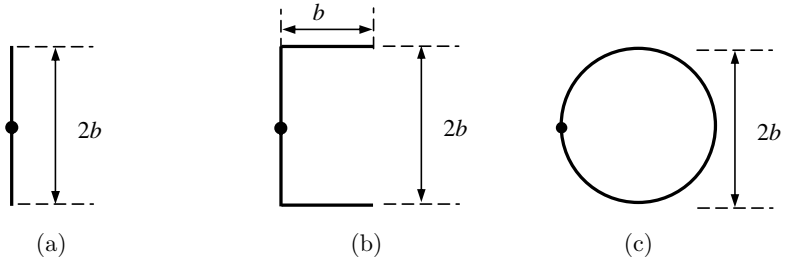


Figure 5.41 Enhancement of bandwidth. (a) Dipole. (b) Folded dipole. (c) Loop.

5.9 Broadband Antennas

It has been shown that the antenna fractional bandwidth is approximately the inverse of the antenna Q_{real} [see (5.123)]. To enhance the antenna bandwidth, we need to reduce the antenna Q_{real}, which can be achieved by letting the metal antenna occupy the space as efficiently as possible. For the wire antenna, bending the wires is an efficient way to enhance the bandwidth. To demonstrate this point, let us consider a dipole antenna, a folded dipole antenna, and a circular loop antenna shown in Figure 5.41. Roughly, all three antennas have the same maximum dimension $2b$ with wire radius a_0. The fractional bandwidths for the dipole, folded dipole and loop can be determined from (5.123) and are (Geyi, 2003b)

$$B_{\text{dipole}} = \frac{(kb)^3}{6\ln(b/a)}, \quad B_{\text{folded dipole}} = \frac{2(kb)^3}{6\ln(b/a)}, \quad B_{\text{loop}} = \frac{\pi(kb)^3}{6\ln(b/a)}.$$

respectively. Thus we have $B_{\text{dipole}} < B_{\text{folded dipole}} < B_{\text{loop}}$. The above examples are a simple illustration that properly bending the wires can enhance the antenna bandwidth.

5.9.1 Biconical Antenna

A biconical antenna may be considered as a finite section of biconical transmission line. For this reason, we start with the biconical transmission line shown in Figure 5.42. In spherical coordinate system, Maxwell equations in source-free region can be written as

$$\frac{\partial}{\partial\theta}(\sin\theta E_\varphi) - \frac{\partial E_\theta}{\partial\varphi} = -j\omega\mu r\sin\theta H_r,$$

$$\frac{\partial}{\partial\theta}(\sin\theta H_\varphi) - \frac{\partial H_\theta}{\partial\varphi} = j\omega\varepsilon r\sin\theta E_r,$$

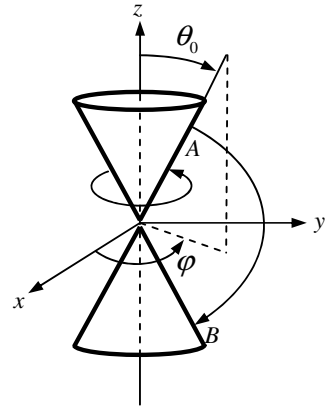


Figure 5.42 Biconical transmission line.

$$\begin{aligned}
\frac{\partial E_r}{\partial \varphi} - \sin\theta \frac{\partial (rE_\varphi)}{\partial r} &= -j\omega\mu r \sin\theta H_\theta, \\
\frac{\partial H_r}{\partial \varphi} - \sin\theta \frac{\partial (rH_\varphi)}{\partial r} &= j\omega\varepsilon r \sin\theta E_\theta, \\
\frac{\partial (rE_\theta)}{\partial r} - \frac{\partial E_r}{\partial \theta} &= -j\omega\mu r H_\varphi, \\
\frac{\partial (rH_\theta)}{\partial r} - \frac{\partial H_r}{\partial \theta} &= j\omega\varepsilon r E_\varphi,
\end{aligned} \tag{5.201}$$

with the divergence equations

$$\begin{aligned}
\sin\theta \frac{\partial}{\partial r}(r^2 E_r) + r\frac{\partial}{\partial \theta}(\sin\theta E_\theta) + r\frac{\partial E_\varphi}{\partial \varphi} &= 0, \\
\sin\theta \frac{\partial}{\partial r}(r^2 H_r) + r\frac{\partial}{\partial \theta}(\sin\theta H_\theta) + r\frac{\partial H_\varphi}{\partial \varphi} &= 0.
\end{aligned}$$

We assume that the fields are independent of φ coordinate. In this case, the Maxwell equations break up into two sets of three equations each. One set represents a TM wave and contains $\{E_r, E_\theta, H_\varphi\}$, and the other set represents a TE wave and contains $\{H_r, H_\theta, E_\varphi\}$

TM:

$$\begin{aligned}
\frac{\partial}{\partial \theta}(\sin\theta H_\varphi) &= j\omega\varepsilon r \sin\theta E_r, \\
-\sin\theta \frac{\partial (rH_\varphi)}{\partial r} &= j\omega\varepsilon r \sin\theta E_\theta,
\end{aligned}$$

$$\frac{\partial(rE_\theta)}{\partial r} - \frac{\partial E_r}{\partial \theta} = -j\omega\mu r H_\varphi, \tag{5.202}$$

TE:

$$\begin{aligned} &\frac{\partial}{\partial\theta}(\sin\theta E_\varphi) = -j\omega\mu r\sin\theta H_r, \\ &-\sin\theta\frac{\partial(rE_\varphi)}{\partial r} = -j\omega\mu r\sin\theta H_\theta, \\ &\frac{\partial(rH_\theta)}{\partial r} - \frac{\partial H_r}{\partial\theta} = j\omega\varepsilon r E_\varphi. \end{aligned} \tag{5.203}$$

On the conical boundaries, E_r must vanish. This is possible when E_r vanishes outside the biconical region. Thus (5.202) reduces to

$$\begin{aligned} &\frac{\partial(rE_\theta)}{\partial r} = -j\omega\mu(rH_\varphi), \\ &\frac{\partial(rH_\varphi)}{\partial r} = -j\omega\varepsilon(rE_\theta), \\ &\frac{\partial}{\partial\theta}(\sin\theta H_\varphi) = 0. \end{aligned} \tag{5.204}$$

The above equations imply both E_θ and H_φ are proportional to $1/\sin\theta$, and we may write

$$E_\theta = \frac{1}{r\sin\theta}(Ae^{-jkr} + Be^{jkr}), \quad H_\varphi = \frac{1}{\eta r\sin\theta}(Ae^{-jkr} - Be^{jkr}).$$

The voltage drop from the position A: (r, θ_0, φ) of the upper cone to B: $(r, \pi - \theta_0, \varphi)$ of the lower cone is given by

$$V(r) = \int_{AB} \mathbf{E}\cdot\mathbf{u}_l\, dl = \int_{\theta_0}^{\pi-\theta_0} E_\theta r\, d\theta = V^+e^{-jkr} + V^-e^{jkr}, \tag{5.205}$$

where

$$V^+ = 2A\ln\cot\frac{\theta_0}{2}, \quad V^- = 2B\ln\cot\frac{\theta_0}{2}.$$

The current density on the upper cone is

$$\begin{aligned} \mathbf{J}_s(r,\theta_0,\varphi) &= \mathbf{u}_\theta \times \mathbf{H}(r,\theta_0,\varphi) = \mathbf{u}_r H_\varphi(r,\theta_0,\varphi) \\ &= \frac{\mathbf{u}_r}{r\eta\sin\theta}(Ae^{-jkr} - Be^{jkr}). \end{aligned}$$

The total current on the upper cone is then given by the contour integral along the boundary C of the cross section of the upper cone at $z = r\sin\theta_0$:

$$I(r) = \int_C \mathbf{J}_s \cdot \mathbf{u}_r \, dl = \int_0^{2\pi} \frac{1}{r\eta \sin\theta_0}(Ae^{-jkr} - Be^{jkr}) r \sin\theta_0 \, d\varphi$$
$$= \frac{2\pi}{\eta}(Ae^{-jkr} - Be^{jkr}) = \frac{1}{Z_c}(V^+ e^{-jkr} - V^- e^{jkr}), \qquad (5.206)$$

where

$$Z_c = \frac{\eta}{\pi} \ln \cot \frac{\theta_0}{2} \qquad (5.207)$$

is the characteristic impedance of the biconical line. The impedance of the biconical line at r is

$$Z(r) = \frac{V(r)}{I(r)} = \frac{Ae^{-jkr} + Be^{jkr}}{\frac{1}{Z_c}(Ae^{-jkr} - Be^{jkr})} = Z_c \frac{1 + \Gamma e^{j2kr}}{1 - \Gamma e^{j2kr}} \qquad (5.208)$$

where $\Gamma = B/A$. The input impedance (5.208) may be rewritten as

$$Z(r) = Z_c \frac{Z(L) + jZ_c \tan k(L - r)}{Z_c + jZ(L) \tan k(L - r)}, \qquad (5.209)$$

where $Z(L)$ is the impedance at $r = L$

$$Z(L) = Z_c \frac{1 + \Gamma e^{j2kL}}{1 - \Gamma e^{j2kL}}. \qquad (5.210)$$

The infinite biconical transmission line may be truncated at $r = L$ to form a finite biconical antenna. In this case, we have $Z(L) = \infty$ and

$$Z(r) = -jZ_c \cot k(L - r). \qquad (5.211)$$

The total current $I(r)$ must vanish at $r = L$, which yields

$$I(r) = \frac{j2V^+}{Z_c} e^{-jkL} \sin k(L - r). \qquad (5.212)$$

The solid biconical antenna is massive. In practice, various variations of biconical antennas have been proposed, which include bow-tie antenna and double wire cones as indicated in Figure 5.1.

5.9.2 Helical Antenna

A **helical antenna** consists of a conducting wire wound in the form of a helix, and it was first invented by John D. Kraus in 1946 (Kraus, 2003). Helical antennas are often mounted over a ground plane and they can operate in one of two principal modes: normal mode or axial mode. In the normal-mode operation, the dimensions of the helix are small compared with the wavelength. The helix behaves like a short monopole and the radiation pattern is omni-directional with maximum radiation normal to the axis. The antenna is linearly polarized in the direction of the helix axis. In the axial-mode operation, the dimensions of the helix are comparable to a wavelength. The antenna functions as a directional antenna with maximum radiation along the helix axis away from the ground. In this case, the antenna is circularly polarized.

The helix geometry is described by its radius a, circumference $2\pi a$, spacing between turns l, pitch angle α, length of one turn L, number of turns N, axial length Nl, radius of helix wire a_0, as illustrated in Figure 5.43. The spacing l, the circumference $2\pi a$ and the turn length L may form a triangle. The angle facing the spacing is the pitch angle.

5.9.2.1 *Normal Mode*

When the helix is in normal mode, the helix may be approximated by N small loops of radius a and N short dipoles of length l joined together in series, as illustrated in Figure 5.44. The far-field produced by the short dipole is

$$E_\theta = j\eta \frac{kI_d l}{4\pi r} \sin\theta e^{-jkr}, \tag{5.213}$$

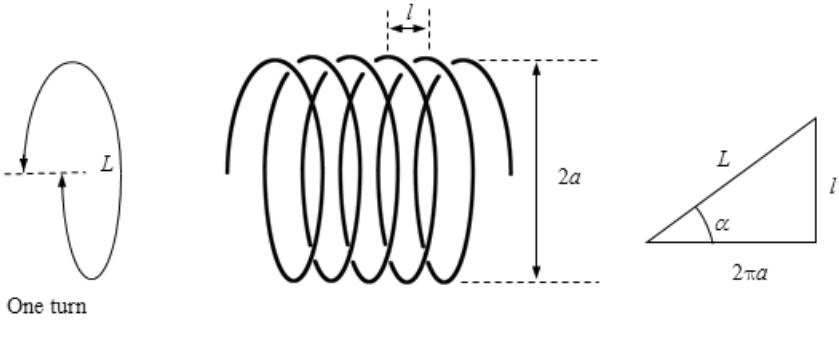


Figure 5.43 Geometric parameters of helix.

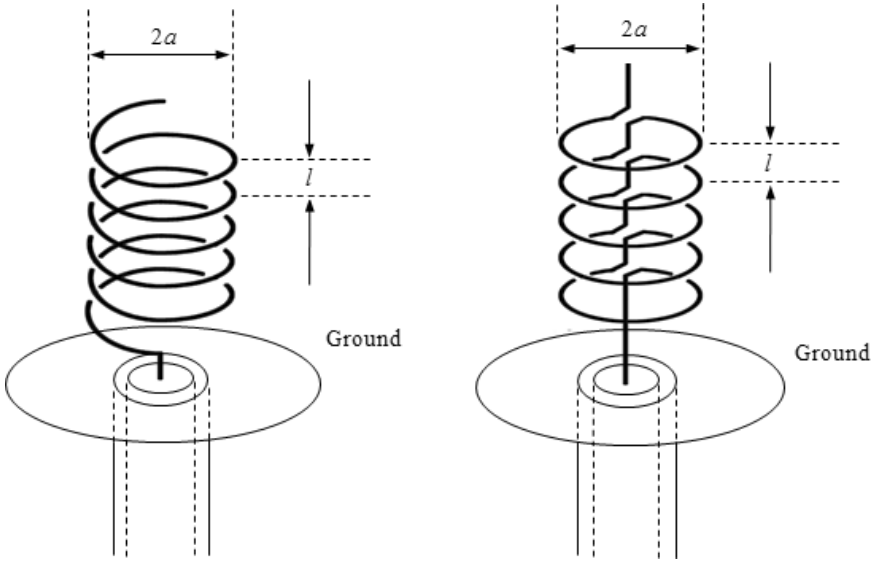


Figure 5.44 Helix antenna approximated by loops and dipoles connected in series.

where I_d is the current along the dipole and is assumed to be constant. The far-field produced by the loop is

$$E_\varphi = \eta \frac{I_l (ka)^2}{4r} \sin\theta e^{-jkr}, \tag{5.214}$$

where I_l is the current along the loop and is assumed to be constant. The **axial ratio** (AR) is

$$\mathrm{AR} = \frac{|E_\theta|}{|E_\varphi|} = \frac{1}{2\pi} \frac{\lambda l}{\pi a^2}. \tag{5.215}$$

When $\mathrm{AR} = 0$ ($l \to 0$), we have $E_\theta = 0$ and the helix antenna reduces to a loop. When $\mathrm{AR} = \infty$ ($a \to 0$), the helix antenna is essentially a dipole. When $\mathrm{AR} = 1$, the radiated field is circularly polarized and in this case, we have

$$\frac{\lambda l}{\pi a^2} = 2\pi.$$

This implies

$$\tan\alpha = \frac{l}{2\pi a} = \pi \frac{a}{\lambda}.$$

5.9.2.2 *Axial Mode*

In the axial-mode operation, the maximum radiation is along the helix axis. To achieve the circular polarization, the parameters of the helix must satisfy

$$\frac{3}{4} < \frac{2\pi a}{\lambda} < \frac{4}{3}, \quad l \approx \frac{\lambda}{4}.$$

5.9.3 Frequency-Independent Antennas

The concept of frequency-independent antenna was first proposed by V. H. Rumsey (1957), and it evolved from the observation that the pattern and impedance characteristics of an antenna depend on its dimensions measured in wavelengths. Antennas having similar geometric structures will then retain the same radiation characteristics if a frequency change does not change the ratio of antenna dimensions to wavelength. It was found that the pattern and impedance of an antenna are independent of frequency if its shape is specified entirely in terms of angles without specifying any characteristic length dimensions. Rumsey's work has been summarized and simplified by Elliot (2003). Assume that the antenna consists of perfect conductors and is surrounded by an infinite homogeneous and isotropic medium. The two terminals of the antenna are assumed to be indefinitely close to the origin and are symmetrically arranged along the $\theta = 0°, \theta = 180°$ axis. Let the surface of this antenna be described by

$$r = F(\theta, \varphi). \tag{5.216}$$

There may be several branches to the function $F(\theta, \varphi)$, corresponding to inner and outer surfaces. Suppose that this antenna is scaled to a new frequency, being K times lower than the original frequency. The antenna must then be made K times bigger, resulting in a surface

$$r = KF(\theta, \varphi). \tag{5.217}$$

One can now ask that the new surface and the old one are congruent. A necessary condition for this to occur is that both surfaces are infinite. In fact, congruence can only be established through a rotation in φ (A rotation in θ is not allowed since both pairs of terminals are symmetrically arranged along the $\theta = 0°, \theta = 180°$ axis. Translation is not allowed either since the

feeding point is fixed at the origin.). Thus if the congruence is established, we must have

$$KF(\theta, \varphi) = F(\theta, \varphi + C) \tag{5.218}$$

where C is the angle through which the second antenna must be rotated in order to achieve congruence with the first. It follows from (5.218) that

$$\frac{1}{K}\frac{dK}{dC} = \frac{1}{r}\frac{\partial r}{\partial \varphi}. \tag{5.219}$$

This implies

$$r = e^{a\varphi} f(\theta), \quad a = \frac{1}{K}\frac{dK}{dC}. \tag{5.220}$$

where $f(\theta)$ is an arbitrary function. This result was first derived by Rumsey and it reveals that any point-fed antenna whose geometry is described by a function of the form (5.220) will be independent of frequency.

Example 5.6 (Planar spirals)**:** The arbitrary function $f(\theta)$ may be chosen to satisfy

$$\frac{df}{d\theta} = A\delta\Big(\theta - \frac{\pi}{2}\Big),$$

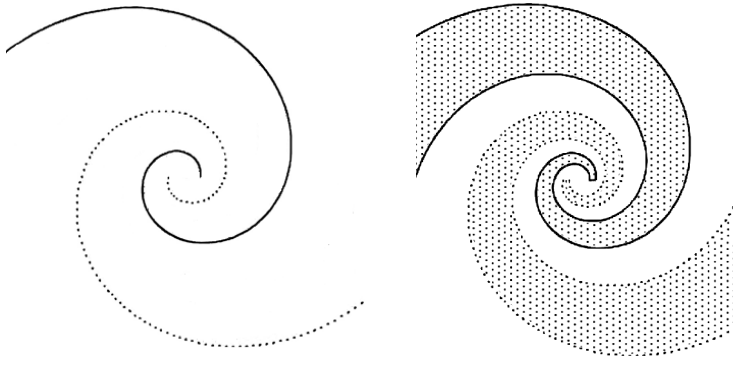
where A is an arbitrary positive number. Then (5.220) can be written as

$$r = \begin{cases} Ae^{a\varphi}, & \theta = \dfrac{\pi}{2} \\ 0, & \theta \neq \dfrac{\pi}{2} \end{cases}. \tag{5.221}$$

Using the polar coordinate system on the (x, y)-plane, this can be written as

$$\rho = \rho_0 e^{a(\varphi - \varphi_0)} \tag{5.222}$$

where we have set $A = \rho_0 e^{-a\varphi_0}$. Equation (5.222) describes a planar spiral in (x, y)-plane. Since the parameter A is arbitrary, we may fix ρ_0 and use φ_0 as a parameter. Let $\varphi_0 = 0, \pi$, the resultant two spirals are shown in Figure 5.45(a). If we allow φ_0 to takes all values from 0 to φ_1, and all values from π to $\pi + \varphi_1$, two solid spirals will be obtained, as shown in Figure 5.45(b). The planar spiral antenna is bidirectional and it radiates a broadside main beam on both sides of the plane. By wrapping a balanced two-arm spiral on the surface of a cone of revolution we can obtain a



(a) (b)

Figure 5.45 Planar spiral antennas.

unidirectional radiation pattern with a single main beam in the direction of the cone apex. □

5.10 Coupling between Two Antennas

In recent years, the use of radio communication has been growing very fast. Due to the space limitations two or more antennas may have to be placed in close proximity, such as in the handset or on antenna towers. When two antennas are in close proximity, the characteristic of each antenna will be affected by the other because of the mutual coupling between them, which will degrade the antenna performances and cause problems in communication systems.

The study of coupling between two antennas can be traced back to the early work of Carter (1932). Carter derived the expressions for self and mutual impedances (open-circuit parameters) for a radiating system by the use of reciprocity theorem and his approach has been focused on a two-linear antenna system. Since most electrical engineers are quite familiar with the circuit theory such an approach of reducing a field problem into a circuit problem is very helpful. In fact, it has been a common practice for most microwave engineers to use the circuit theory to visualize the physical process in a microwave circuit. For example, a transmitting antenna has been shown to be equivalent to a RLC resonant circuit, where R stands for the radiated and dissipated power in the antenna system, and L and C represent the stored magnetic and electric energy around the antenna respectively. From the equivalent RLC circuit the following physical picture

can be obtained. Under normal operation, the antenna is matched and tuned to resonant frequency and the exciting source of the antenna directly delivers energy to the far field as radiated energy. The stored electric and magnetic energy oscillate and change into each other in the vicinity of the antenna and play the role of controlling antenna bandwidth, and they do not contribute to the radiated field. In other words, there is no energy exchange between stored energy around the antenna and radiated energy. The energy exchange only happens when the antenna source is turned off, and in this case, the stored energy will be transformed into radiated energy or dissipated into heat. The above understanding is straightforward from the theory of a *RLC* resonant circuit, but would be hardly imagined if a pure field approach is exploited.

5.10.1 A General Approach

Consider a two-antenna system contained in a region V_∞ bounded by ∂V_∞. Let the fields generated by antenna i $(i = 1, 2)$ when antenna j $(j \neq i)$ is receiving be denoted by $\mathbf{E}_i, \mathbf{H}_i$ $(i = 1, 2)$. We use V_{0i} to denote the source region for antenna $i (i = 1, 2)$. The source region is chosen in such a way that its boundary, denoted by ∂V_{0i}, is coincident with the metal surface of the antennas except for a portion Ω_i, where the boundary crosses the antenna reference plane. We use $V_i^{(j)}$ and $I_i^{(j)} (i, j = 1, 2)$ to represent the modal voltage and modal current at the reference plane of antenna i when antenna j is transmitting. One of the states of operation is illustrated in Figure 5.46(a). Figure 5.46(b) is the corresponding equivalent network representation with

$$\begin{bmatrix} V_1^{(1)} \\ V_2^{(1)} \end{bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \begin{bmatrix} I_1^{(1)} \\ I_2^{(1)} \end{bmatrix}.$$

The coupling between the two antennas is characterized by $Z_{ij} (i, j = 1, 2; i \neq j)$, which may be determined by use of the frequency-domain reciprocity theorem

$$\int_S (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS = 0, \tag{5.223}$$

where S is assumed to be an arbitrary closed surface that does not contain any impressed sources and $\mathbf{u}_n$ is the outward unit normal. Choosing

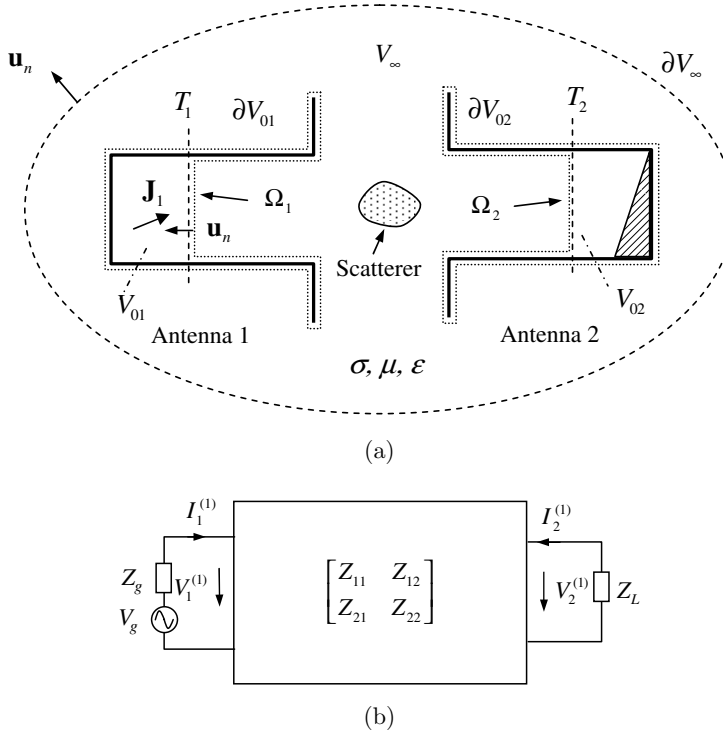


Figure 5.46 (a) A two-antenna system. (b) Equivalent network.

$S = \partial V_\infty + \partial V_{01} + \partial V_{02}$ in (5.223) yields

$$\int\limits_{\partial V_{01}} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS + \int\limits_{\partial V_{02}} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS + \int\limits_{\partial V_\infty} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS = \sum_{l=1}^{2} \left[V_l^{(2)} I_l^{(1)} - V_l^{(1)} I_l^{(2)} \right] = 0. \tag{5.224}$$

This is the well-known reciprocity theorem in network theory. If we assume that the antenna j is in the state of open circuit when antenna i $(i \neq j)$ is transmitting, the above equation reduces to $V_1^{(2)} I_1^{(1)} = V_2^{(1)} I_2^{(2)}$, or

$$Z_{12} = \left. \frac{V_1^{(2)}}{I_2^{(2)}} \right|_{I_1^{(2)}=0} = \left. \frac{V_2^{(1)}}{I_1^{(1)}} \right|_{I_2^{(1)}=0} = Z_{21}. \tag{5.225}$$

Therefore, the impedance matrix is symmetric. We now choose $S = S_1' + \partial V_{01}$ in (5.223), where S_1' is a closed surface containing antenna 1 only. Then

$$\int\limits_{\partial V_{01}} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS + \int\limits_{S_1'} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS = 0.$$

This implies

$$V_1^{(1)} I_1^{(2)} - V_1^{(2)} I_1^{(1)} = \int\limits_{S_1'} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS. \tag{5.226}$$

Similarly, we have

$$V_2^{(2)} I_2^{(1)} - V_2^{(1)} I_2^{(2)} = \int\limits_{S_2'} (\mathbf{E}_2 \times \mathbf{H}_1 - \mathbf{E}_1 \times \mathbf{H}_2) \cdot \mathbf{u}_n \, dS, \tag{5.227}$$

where S_2' is a closed surface containing antenna 2 only. The right-hand sides of (5.226) and (5.227) can be shown to be equal by choosing $S = S_1' + S_2' + \partial V_\infty$ in (5.223). When antenna 1 (or 2) is transmitting with the other antenna being open, we have

$$\begin{aligned} V_1^{(2)} I_1^{(1)} &= - \int\limits_{S_1'} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS \\ &= - \int\limits_{S_2'} (\mathbf{E}_2 \times \mathbf{H}_1 - \mathbf{E}_1 \times \mathbf{H}_2) \cdot \mathbf{u}_n \, dS = V_2^{(1)} I_2^{(2)}. \end{aligned} \tag{5.228}$$

By definition, the mutual impedance of the two-antenna system can be written as

$$\begin{aligned} Z_{12} &= \left. \frac{V_1^{(2)}}{I_2^{(2)}} \right|_{I_1^{(2)}=0} \\ &= - \frac{\int\limits_{S_1'} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS}{I_1^{(1)} I_2^{(2)}} = - \frac{\int\limits_{V_{01}} \mathbf{J}_1 \cdot \mathbf{E}_2 \, dV}{I_1^{(1)} I_2^{(2)}}, \end{aligned} \tag{5.229}$$

where use is made of the following reciprocity theorem

$$\begin{aligned} \int\limits_{V_{02}} \mathbf{J}_2 \cdot \mathbf{E}_1 \, dV &= \int\limits_{S_2'} (\mathbf{E}_2 \times \mathbf{H}_1 - \mathbf{E}_1 \times \mathbf{H}_2) \cdot \mathbf{u}_n \, dS \\ &= \int\limits_{S_1'} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS = \int\limits_{V_{01}} \mathbf{J}_1 \cdot \mathbf{E}_2 \, dV. \end{aligned} \tag{5.230}$$

Equation (5.229) may be regarded as an exact expression of Huygens' principle in a symmetrical form, and it is generally applicable to an inhomogeneous medium.

Let Z_{si} be the reference impedance for the input terminal of antenna i $(i = 1, 2)$. Introducing

$$V_j^{(i)} = \frac{\bar{Z}_{sj}}{\sqrt{\mathrm{Re}\, Z_{sj}}} a_j^{(i)} + \frac{Z_{sj}}{\sqrt{\mathrm{Re}\, Z_{sj}}} b_j^{(i)},$$
$$I_j^{(i)} = \frac{1}{\sqrt{\mathrm{Re}\, \bar{Z}_{sj}}} a_j^{(i)} - \frac{1}{\sqrt{\mathrm{Re}\, Z_{sj}}} b_j^{(i)} \quad i, j = A, B. \tag{5.231}$$

into (5.224), we obtain

$$\sum_{l=1}^{2} \left[a_l^{(1)} b_l^{(2)} - a_l^{(2)} b_l^{(1)} \right] = 0. \tag{5.232}$$

If we assume that the antenna j is matched when antenna i $(i \neq j)$ is transmitting, Equation (5.232) reduces to $a_1^{(1)} b_1^{(2)} = a_2^{(2)} b_2^{(1)}$, which gives the symmetric property of scattering matrix

$$S_{12} = \left. \frac{b_1^{(2)}}{a_2^{(2)}} \right|_{a_1^{(2)}=0,} = \left. \frac{b_2^{(1)}}{a_1^{(1)}} \right|_{a_2^{(1)}=0} = S_{21}.$$

In terms of incident and reflected power waves, Equations (5.226) and (5.227) can be written as

$$b_1^{(1)} a_1^{(2)} - b_1^{(2)} a_1^{(1)} = \frac{1}{2} \int_{S_1'} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS, \tag{5.233}$$

$$b_2^{(2)} a_2^{(1)} - b_2^{(1)} a_2^{(2)} = \frac{1}{2} \int_{S_2'} (\mathbf{E}_2 \times \mathbf{H}_1 - \mathbf{E}_1 \times \mathbf{H}_2) \cdot \mathbf{u}_n \, dS. \tag{5.234}$$

It follows from (5.230) that

$$S_{12} = \left. \frac{b_1^{(2)}}{a_2^{(2)}} \right|_{a_1^{(2)}=0} = -\frac{1}{2 a_1^{(1)} a_2^{(2)}} \int_{S_1'} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS. \tag{5.235}$$

5.10.2 Coupling between Two Antennas with Large Separation

So far the separation between antennas is arbitrary. We now assume that the antennas are located in the far-field region of each other. Determining the fields $\mathbf{E}_i, \mathbf{H}_i$ $(i = 1, 2)$ produced by the antenna i with antenna j $(j \neq i)$ in place is not an easy task. Therefore, the following simplification is made: the calculation of fields $\mathbf{E}_i$ and $\mathbf{H}_i$ is carried out with the antennas j $(j \neq i)$ removed. Physically, this assumption is equivalent to neglecting the reflections between the antennas. To derive the expressions of the impedance parameters Z_{12} when the antenna 1 and antenna 2 are far apart, two different coordinate systems for antenna 1 and antenna 2 may be used. The origins of the coordinate systems are chosen to be the geometrical center of the current distributions and the separation between antenna 1 and antenna 2 satisfies $kr_2 \gg 1$, $r_2 \gg d_2$, $r_2 \gg d_1$ where $r_2 = |\mathbf{r}_2|$ is the distance between antenna 2 and an arbitrary point of the circumscribing sphere of antenna 1 (denoted by S_1'), as shown in Figure 5.47. Let $\mathbf{r}_1'$ be a point on the circumscribing sphere of antenna 1, and $\mathbf{r}_{12} = r_{12}\mathbf{u}_{r_{12}}$, where r_{12} is the distance between the two origins and $\mathbf{u}_{r_{12}}$ is a unit vector directed from antenna 1 to antenna 2. Thus, the far field of antenna 2 at antenna 1 can be expressed as

$$\mathbf{E}_2(\mathbf{r}_2) \approx -\frac{jk\eta I_2^{(2)} e^{-jkr_2}}{4\pi r_2}\mathbf{L}_2(\mathbf{u}_{r_2}), \quad \mathbf{H}_2(\mathbf{r}_2) \approx \frac{1}{\eta}\mathbf{u}_{r_2} \times \mathbf{E}_2(\mathbf{r}_2), \tag{5.236}$$

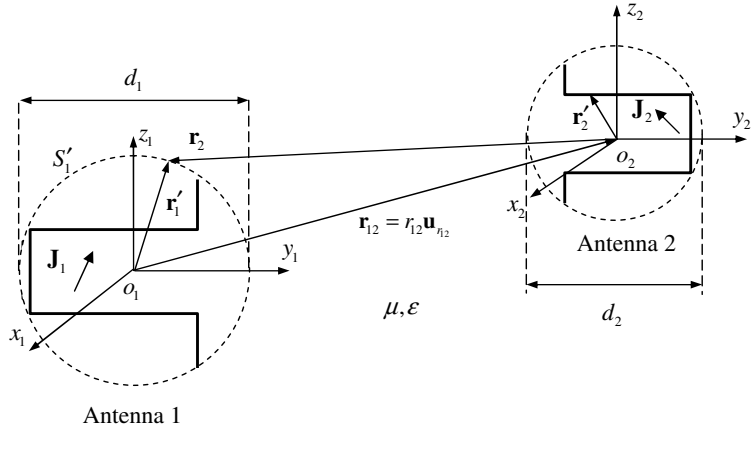


Figure 5.47 Coupling between two distant antennas.

where $\mathbf{r}_2 = \mathbf{r}_1' - \mathbf{r}_{12}$ is assumed to be a point on the sphere S_1' and

$$\mathbf{L}_2(\mathbf{u}_{r_2}) = \frac{1}{I_2^{(2)}} \int\limits_{V_{02}} [\mathbf{J}_2 - (\mathbf{J}_2 \cdot \mathbf{u}_{r_2})\mathbf{u}_{r_2}]\, e^{jk\mathbf{r}_2' \cdot \mathbf{u}_{r_2}}\, dV(\mathbf{r}_2')$$

is the antenna effective vector length. Since $\mathbf{r}_1'$ is very small compared to r_{12} in magnitude, we can make the approximation $r_2 = |\mathbf{r}_1' - \mathbf{r}_{12}| \approx r_{12} - \mathbf{u}_{r_{12}} \cdot \mathbf{r}_1'$. The field $\mathbf{E}_2$ in the coordinate system o_1 can then be represented by

$$\begin{aligned} \mathbf{E}_2(\mathbf{r}_2) &\approx -\frac{jk\eta I_2^{(2)} e^{-jkr_{12}} e^{jk\mathbf{u}_{r_{12}} \cdot \mathbf{r}_1'}}{4\pi r_{12}} \mathbf{L}_2(-\mathbf{u}_{r_{12}}), \\ \mathbf{H}_2(\mathbf{r}_2) &\approx -\frac{1}{\eta} \mathbf{u}_{r_{12}} \times \mathbf{E}_2(\mathbf{r}_2). \end{aligned} \tag{5.237}$$

Then

$$\begin{aligned} &\int\limits_{S_1'} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n\, dS \\ &\quad = \int\limits_{S_1'} \left[-\eta^{-1}\mathbf{E}_1 \times (\mathbf{u}_{r_{12}} \times \mathbf{E}_2) - \mathbf{E}_2 \times \mathbf{H}_1\right] \cdot \mathbf{u}_n\, dS \\ &\quad = \int\limits_{S_1'} \mathbf{E}_2 \cdot \left[-\eta^{-1}\mathbf{u}_{r_{12}} \times (\mathbf{E}_1 \times \mathbf{u}_n) - \mathbf{H}_1 \times \mathbf{u}_n\right] dS \\ &\quad = \int\limits_{S_1'} \mathbf{E}_2 \cdot (\mathbf{J}_{1s} - \eta^{-1}\mathbf{u}_{r_{12}} \times \mathbf{J}_{1ms}) dS, \end{aligned} \tag{5.238}$$

where $\mathbf{J}_{1s} = \mathbf{u}_n \times \mathbf{H}_1, \mathbf{J}_{1ms} = -\mathbf{u}_n \times \mathbf{E}_1$ are the equivalent electric current and magnetic current on the surface S_1' respectively. Substituting (5.237) into (5.238), we obtain

$$\begin{aligned} &\int\limits_{S_1'} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n\, dS \\ &\quad \approx -\frac{4\pi r_{12} e^{jkr_{12}}}{jk\eta} \mathbf{E}_1(\mathbf{u}_{r_{12}}) \cdot \mathbf{E}_2(-\mathbf{u}_{r_{12}}) \\ &\quad = \frac{-jk\eta I_1^{(1)} I_2^{(2)} e^{-jkr_{12}}}{4\pi r_{12}} \mathbf{L}_1(\mathbf{u}_{r_{12}}) \cdot \mathbf{L}_2(-\mathbf{u}_{r_{12}}). \end{aligned} \tag{5.239}$$

Here we have used the far-field expression of antenna 1 at antenna 2

$$\mathbf{E}_1(\mathbf{r}_{12}) = \frac{-jk\eta e^{-jkr_{12}}}{4\pi r_{12}} \int\limits_{S_1'} e^{jk\mathbf{u}_{r_{12}}\cdot\mathbf{r}_1}\left[\mathbf{J}_{1s}(\mathbf{r}_1') - \mathbf{u}_{r_{12}} \times \eta^{-1}\mathbf{J}_{1ms}(\mathbf{r}_1')\right] dS$$

$$= \frac{-jk\eta I_1^{(1)} e^{-jkr_{12}}}{4\pi r_{12}} \mathbf{L}_1(\mathbf{u}_{r_{12}}). \tag{5.240}$$

It follows from (5.239) that the mutual impedance Z_{12} is given by

$$Z_{12} = \left.\frac{V_1^{(2)}}{I_2^{(2)}}\right|_{I_1^{(2)}=0} = \frac{jk\eta e^{-jkr_{12}}}{4\pi r_{12}} \mathbf{L}_1(\mathbf{u}_{r_{12}}) \cdot \mathbf{L}_2(-\mathbf{u}_{r_{12}}). \tag{5.241}$$

Note that we have $Z_{12} = 0$ when $\mathbf{L}_1$ and $\mathbf{L}_2$ are orthogonal.

5.10.3 Power Transmission between Two Antennas

Wireless power transmission has been a research topic for years. Many applications can benefit from the research, such as microwave imaging, radar and directed energy weapons. The basic theory for the power transmission between two antennas was investigated in 1960s (Goubao and Schwering, 1961; Kay, 1960; Sherman, 1962; Borgiotti, 1966), and it has found wide applications in many fields (Brown, 1984). Theoretically, power transmission efficiency of almost 100% is attainable by increasing the sizes of the antennas. For a given power transmission efficiency over a given distance between the transmitting and receiving antenna, there exists an optimum antenna aperture distribution which can minimize the transmitting and receiving aperture sizes. To achieve the maximum transmission efficiency, the transmitting antenna must be focused at the receiving antenna. In other words, the radiated electromagnetic energy must be focused in the vicinity of the axis of the transmitting and receiving antenna apertures as it propagates.

5.10.3.1 *Power Transmission between Two General Antennas*

Let us consider the power transmission between antenna 1 and antenna 2 when antenna 1 is transmitting and antenna 2 is receiving. It follows from Figure 5.46(b) that

$$\begin{aligned} V_1^{(1)} &= Z_1^{(1)} I_1^{(1)}, \quad V_1^{(2)} = -I_1^{(2)} Z_{L1}^{(2)}, \\ V_2^{(2)} &= Z_2^{(2)} I_2^{(2)}, \quad V_2^{(1)} = -I_2^{(1)} Z_{L2}^{(1)}. \end{aligned} \tag{5.242}$$

Here $Z_i^{(i)}$ is the input impedance of antenna i when antenna i is transmitting and the other antenna is receiving, and $Z_{Li}^{(j)}$ is the load at ith terminal when antenna j is transmitting. Substituting (5.242) into (5.228), we obtain

$$I_2^{(2)} I_2^{(1)} \left[Z_2^{(2)} + Z_{L2}^{(1)} \right] = \int\limits_{S_1' \text{ or } S_2'} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS$$
$$= I_1^{(1)} I_1^{(2)} \left[Z_1^{(1)} + Z_{L1}^{(2)} \right]. \tag{5.243}$$

Multiplying (5.243) by its conjugate, we obtain

$$\left| I_2^{(2)} \right|^2 \left| I_2^{(1)} \right|^2 \left| Z_2^{(2)} + Z_{L2}^{(1)} \right|^2 = \left| \int\limits_{S_1' \text{ or } S_2'} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS \right|^2$$
$$= \left| I_1^{(1)} \right|^2 \left| I_1^{(2)} \right|^2 \left| Z_1^{(1)} + Z_{L1}^{(2)} \right|^2. \tag{5.244}$$

If the antenna 1 and antenna 2 are conjugately matched, i.e.,

$$\bar{Z}_2^{(2)} = Z_{L2}^{(1)}, \quad \bar{Z}_1^{(1)} = Z_{L1}^{(2)}.$$

Equation (5.244) can be written as

$$\frac{\frac{1}{2} \left| I_2^{(1)} \right|^2 \operatorname{Re} Z_{L2}^{(1)}}{\frac{1}{2} \left| I_1^{(1)} \right|^2 \operatorname{Re} Z_1^{(1)}} = \frac{\left| \int\limits_{S_1' \text{ or } S_2'} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS \right|^2}{4 \left| I_1^{(1)} \right|^2 \operatorname{Re} Z_1^{(1)} \left| I_2^{(2)} \right|^2 \operatorname{Re} Z_2^{(1)}}$$
$$= \frac{\frac{1}{2} \left| I_1^{(2)} \right|^2 \operatorname{Re} Z_{L1}^{(2)}}{\frac{1}{2} \left| I_2^{(2)} \right|^2 \operatorname{Re} Z_2^{(2)}}. \tag{5.245}$$

This implies

$$T_{12} = \frac{P_2^{(1)}}{P_1^{(1)}} = \frac{\left| \int\limits_{S_1' \text{ or } S_2'} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS \right|^2}{4 \operatorname{Re} \int\limits_{S_1'} (\mathbf{E}_1 \times \bar{\mathbf{H}}_1) \cdot \mathbf{u}_n \, dS \operatorname{Re} \int\limits_{S_2'} (\mathbf{E}_2 \times \bar{\mathbf{H}}_2) \cdot \mathbf{u}_n \, dS}$$
$$= \frac{P_1^{(2)}}{P_2^{(2)}} = T_{21}, \tag{5.246}$$

where $P_i^{(i)}$ is the transmit power of antenna i when other antenna is receiving and $P_j^{(i)}$ is the power received by antenna j when antenna i

is transmitting. Equation (5.246) indicates that the ratio of the power received by antenna 2 to the transmitting power of antenna 1 (known as the **power transmission efficiency**, denoted as T_{12}) is equal to the ratio of the power received by antenna 1 to the transmitting power of antenna 2 (denoted as T_{21}). It also indicates that the radiation pattern of the antenna for reception is identical with that for transmission. Equation (5.246) is the theoretical foundation for the wireless power transmission in free space, and is the starting point for optimizing the aperture distribution to achieve the maximum possible power transmission efficiency. Evidently, the power transmission efficiency gets maximized if

$$\mathbf{E}_1 = \bar{\mathbf{E}}_2, \quad \mathbf{H}_1 = -\bar{\mathbf{H}}_2 \tag{5.247}$$

hold on some closed surface that encloses either antenna 1 or 2. If the separation between antenna 1 and 2 is large enough, we may use (5.239) to obtain

$$\begin{aligned}
&\left| \int\limits_{S_1' \text{ or } S_2'} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathbf{u}_n \, dS \right|^2 \\
&\quad \approx \frac{(4\pi r_{12})^2}{\eta^2 k^2} \left| \mathbf{E}_1(\mathbf{r}_{12}) \cdot \mathbf{E}_2(-\mathbf{r}_{12}) \right|^2 \\
&\quad = \left(\frac{4\lambda}{r_{12}} \right)^2 U_1(\mathbf{u}_{r_{12}}) U_2(-\mathbf{u}_{r_{12}}) \cos\theta_{12},
\end{aligned} \tag{5.248}$$

where U_1 and U_2 are the radiation intensity of antenna 1 and 2 respectively, and θ_{12} is the angle between $\mathbf{E}_1(\mathbf{r}_{12})$ and $\mathbf{E}_2(-\mathbf{r}_{12})$. Substituting (5.248) into (5.246), we obtain the well-known **Friis transmission formula**

$$\begin{aligned}
\frac{P_2^{(1)}}{P_1^{(1)}} &= \left(\frac{\lambda}{4\pi r_{12}} \right)^2 \frac{4\pi U_1(\mathbf{u}_{r_{12}}) 4\pi U_2(-\mathbf{u}_{r_{12}}) \cos\theta_{12}}{\frac{1}{2}\mathrm{Re} \int\limits_{S_1'} (\mathbf{E}_1 \times \bar{\mathbf{H}}_1) \cdot \mathbf{u}_n \, dS \, \frac{1}{2}\mathrm{Re} \int\limits_{S_2'} (\mathbf{E}_2 \times \bar{\mathbf{H}}_2) \cdot \mathbf{u}_n \, dS} \\
&= \left(\frac{\lambda}{4\pi r_{12}} \right)^2 G_1(\mathbf{u}_{r_{12}}) G_2(-\mathbf{u}_{r_{12}}) \cos\theta_{12},
\end{aligned} \tag{5.249}$$

where G_1 and G_2 are the gains of the antenna 1 and antenna 2 respectively. Equation (5.249) may be written as

$$P_2^{(1)} = \frac{\mathrm{EIRP}}{L_s} G_2(-\mathbf{u}_{r_{12}}) \cos\theta_{12}, \tag{5.250}$$

where $L_s = (4\pi r_{12}/\lambda)^2$ is known as **free-space path loss**, and EIRP stands for the **effective isotropic radiated power** defined by EIRP $= P_1^{(1)} G_1(\mathbf{u}_{r_{12}})$. The **received isotropic power** is defined as EIRP$/L_s$, which is the power received by an isotropic antenna ($G_2 = 1$).

5.10.3.2 *Maximum Power Transmission between Two Planar Apertures*

If a two-antenna system is used to transmit electric power, the antenna geometries and their current distributions should be chosen properly in order that the electromagnetic power delivered from one antenna to the other is maximized. Let us consider the maximum power transmission between two planar apertures. The configuration of a two-planar aperture system in free space is shown in Figure 5.48, where both apertures are in an infinite conducting screen so that the tangential electric field outside the aperture is zero. When the aperture i ($i = 1, 2$) is used as a transmitting antenna, the aperture field is assumed to be

$$\mathbf{E}_i = \mathbf{u}_x E_i, \quad \mathbf{H}_i = \mathbf{u}_y \frac{1}{\eta_0} E_i,$$

where $\eta_0 = \sqrt{\mu_0/\varepsilon_0}$ is the wave impedance in free space. We will use the same notations for the aperture field distribution and the field produced by the aperture, and this will not cause any confusion. By means of equivalence theorem and image principle, the electric field produced by aperture 1 may be represented by

$$\mathbf{E}_1(\mathbf{r}) = \frac{1}{2\pi} \int_{T_1} \mathbf{u}_y \times \mathbf{u}_R \left(jk_0 + \frac{1}{|\mathbf{r} - \mathbf{r}'|} \right) e^{-jk_0|\mathbf{r}-\mathbf{r}'|} E_1(\mathbf{r}') dx' dy', \tag{5.251}$$

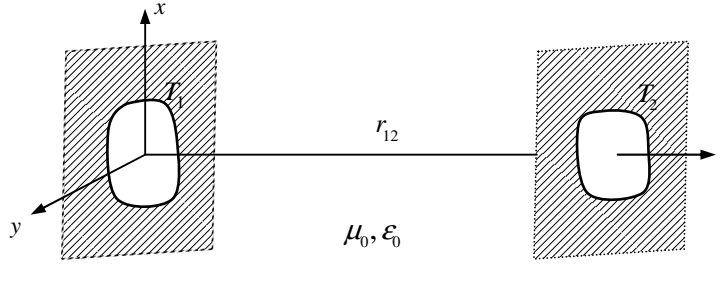


Figure 5.48 Two-planar aperture system.

where $\mathbf{u}_R = (\mathbf{r}-\mathbf{r}')/|\mathbf{r}-\mathbf{r}'|$, $k_0 = \omega\sqrt{\mu_0\varepsilon_0}$. In deriving the above expression, we have neglected the multiple scattering between the apertures. If the apertures are located in the Fresnel region of each other and the observation point $\mathbf{r}$ is on the aperture 2, the following approximations can be made

$$\mathbf{u}_y \times \mathbf{u}_R \approx \mathbf{u}_x, \quad |\mathbf{r}-\mathbf{r}'| \approx r_{12} + \frac{1}{2r_{12}}[(x-x')^2 + (y-y')^2].$$

From (5.251), we obtain

$$\mathbf{E}_1(\mathbf{r}) = \mathbf{u}_x E_1(\mathbf{r}) \approx \mathbf{u}_x \frac{je^{-jk_0 r_{12}}}{\lambda r_{12}} \int\limits_{T_1} E_1 e^{-jk_0[(x-x')^2+(y-y')^2]/2r_{12}} dx'dy'$$

$$\mathbf{H}_1(\mathbf{r}) = \mathbf{u}_y \frac{1}{\eta_0} E_1(\mathbf{r}). \tag{5.252}$$

Substituting these into (5.246) gives

$$T_{12} = \left(\frac{1}{\lambda r_{12}}\right)^2 \frac{\left|\int\limits_{T_2} \tilde{m}_1 m_2 \, dx\, dy\right|^2}{\int\limits_{T_1} |m_1|^2 \, dx\, dy \int\limits_{T_2} |m_2|^2 \, dx\, dy}, \tag{5.253}$$

where

$$\begin{aligned}
m_1(x,y) &= E_1 e^{-jk_0(x^2+y^2)/2r_{12}}, \\
m_2(x,y) &= E_2 e^{-jk_0(x^2+y^2)/2r_{12}}, \\
\tilde{m}_1(x,y) &= \int\limits_{T_1} m_1(x',y') e^{jk_0(xx'+yy')/r_{12}} \, dx'dy', \\
\tilde{m}_2(x,y) &= \int\limits_{T_2} m_2(x',y') e^{jk_0(xx'+yy')/r_{12}} \, dx'dy'.
\end{aligned}$$

Note that

$$\int\limits_{T_1} m_1 \tilde{m}_2 \, dx\, dy = \int\limits_{T_2} \tilde{m}_1 m_2 \, dx\, dy.$$

This is equivalent to $T_{12} = T_{21}$. Equation (5.253) may be written as

$$T_{12} = T_{12}^{\text{ideal}} \cdot U,$$

where

$$U = \frac{\left|\int\limits_{T_2} \tilde{m}_1 m_2 \, dx\, dy\right|^2}{\int\limits_{T_2} |\tilde{m}_1|^2 \, dx\, dy \int\limits_{T_2} |m_2|^2 \, dx\, dy}, \tag{5.254}$$

$$T_{12}^{\text{ideal}} = \frac{\text{Re} \int\limits_{T_2} (\mathbf{E}_1 \times \bar{\mathbf{H}}_1) \cdot \mathbf{u}_z \, dx \, dy}{\text{Re} \int\limits_{T_1} (\mathbf{E}_1 \times \bar{\mathbf{H}}_1) \cdot \mathbf{u}_z \, dx \, dy} = \left(\frac{1}{\lambda r_{12}} \right)^2 \frac{\int\limits_{T_2} |\tilde{m}_1|^2 \, dx \, dy}{\int\limits_{T_1} |m_1|^2 \, dx \, dy}. \tag{5.255}$$

Note that (5.255) is the power transmission efficiency between two ideal apertures. The power transmission efficiency T_{12} reaches maximum if both T_{12}^{ideal} and U are maximized. From Schwartz inequality, we have $\max U = 1$, which can be reached by letting $m_2(x, y) = c_1 \bar{\tilde{m}}_1(x, y)$, $(x, y) \in T_2$, i.e.,

$$E_2(x, y) = c_2 \bar{E}_1(x, y), \quad (x, y) \in T_2. \tag{5.256}$$

Here both c_1 and c_2 are arbitrary complex numbers. The above equation implies that the aperture distribution of antenna 2 is equal to the complex conjugate of the field produced by antenna 1 at antenna 2. We now consider the condition for maximizing T_{12}^{ideal}. Equation (5.255) can be rewritten as

$$T_{12}^{\text{ideal}} = \frac{(\hat{T} m_1, m_1)}{(m_1, m_1)},$$

where $(\cdot, \cdot)$ denotes the inner product defined by $(u, v) = \int_{T_1} u \bar{v} \, dx \, dy$ for two arbitrary functions u and v, and $\hat{T}$ is a self-adjoint operator defined by

$$\hat{T} m_1 (\xi', \varsigma') = \int\limits_{T_1} K_2(\xi, \varsigma; \xi', \varsigma') m_1(\xi, \varsigma) d\xi \, d\varsigma$$

with

$$K_2(\xi, \varsigma; \xi', \varsigma') = \left(\frac{1}{\lambda r_{12}} \right)^2 \int\limits_{T_2} e^{jk_0[(\xi - \xi')x + (\varsigma - \varsigma')y]/r_{12}} \, dx \, dy. \tag{5.257}$$

If the condition (5.256) is met, we have

$$T_{12} = T_{12}^{\text{ideal}} = \frac{(\hat{T} m_1, m_1)}{(m_1, m_1)}. \tag{5.258}$$

This is a variational expression (Rayleigh quotient), and attains an extremum when m_1 satisfies

$$\hat{T} m_1(x, y) = T_{12} m_1(x, y). \tag{5.259}$$

Therefore, the power transmission between two planar apertures is maximized if the aperture field distributions satisfy (5.256) and (5.259) simultaneously. Equation (5.259) is an eigenvalue problem and its largest

eigenvalue is the maximum possible value for the power transmission efficiency. Equation (5.259) may be used first to determine the aperture distribution of antenna 1, and the aperture distribution of antenna 2 can then be determined from (5.256).

Example 5.7: Let us consider the maximum possible power transmission between two rectangular apertures $T_1 = [-a_1, a_1] \times [-a_2, a_2]$ and $T_2 = [-b_1, b_1] \times [-b_2, b_2]$. Equation (5.257) becomes

$$\begin{aligned} K_2(\xi, \varsigma; \xi', \varsigma') &= \left(\frac{1}{\lambda r_{12}}\right)^2 \int_{-b_1}^{b_1} \int_{-b_2}^{b_2} e^{jk_0[(\xi-\xi')x+(\varsigma-\varsigma')y]/r_{12}}\, dx\, dy \\ &= \left(\frac{k_0}{\pi r_{12}}\right)^2 \frac{\sin \frac{k_0(\xi-\xi')b_1}{r_{12}} \sin \frac{k_0(\varsigma-\varsigma')b_2}{r_{12}}}{\frac{k_0(\xi-\xi')}{r_{12}} \frac{k_0(\varsigma-\varsigma')}{r_{12}}}. \end{aligned} \tag{5.260}$$

Thus (5.259) becomes

$$\left(\frac{k_0}{\pi r_{12}}\right)^2 \int_{-a_1}^{a_1} \int_{-a_2}^{a_2} \frac{\sin \frac{k_0(\xi-x)b_1}{r_{12}} \sin \frac{k_0(\varsigma-y)b_2}{r_{12}}}{\frac{k_0(\xi-x)}{r_{12}} \frac{k_0(\varsigma-y)}{r_{12}}} m_1(\xi, \varsigma) d\xi\, d\varsigma = T_{12} m_1(x, y). \tag{5.261}$$

If we assume that $m_1(x, y)$ is a separable function of x and y

$$m_1(x, y) = m_{1x}(x) m_{1y}(y).$$

Equation (5.261) reduces to

$$\begin{aligned} \int_{-a_1}^{a_1} \frac{\sin \frac{k_0 b_1}{r_{12}}(\xi - x)}{\pi(\xi - x)} m_{1x}(\xi) d\xi &= T_{12}^x m_{1x}(x), \\ \int_{-a_2}^{a_2} \frac{\sin \frac{k_0 b_2}{r_{12}}(\varsigma - y)}{\pi(\varsigma - y)} m_{1y}(\varsigma) d\varsigma &= T_{12}^y m_{1y}(y), \end{aligned} \tag{5.262}$$

with $T_{12} = T_{12}^x T_{12}^y$. Introducing the following transformations

$$\xi' = \xi/a_1, \quad x' = x/a_1, \quad \varsigma' = \varsigma/a_2, \quad y' = y/a_2,$$

(5.262) can be written as

$$\begin{aligned}\int_{-1}^{1}\frac{\sin c_1(\xi'-x')}{\pi(\xi'-x')}m_{1x}(\xi')d\xi' &= T_{12}^{x}m_{1x}(x'),\\ \int_{-1}^{1}\frac{\sin c_2(\varsigma'-y')}{\pi(\varsigma'-y')}m_{1y}(\varsigma')d\varsigma' &= T_{12}^{y}m_{1y}(y'),\end{aligned}\tag{5.263}$$

where

$$c_1=\frac{k_0a_1b_1}{r_{12}},\quad c_2=\frac{k_0a_2b_2}{r_{12}}.\tag{5.264}$$

The eigenvalue problem (5.263) also appears in signal theory and has been solved by Slepian and Pollak (1961). The largest eigenvalues are

$$T_{12}^{x}=\frac{2c_1}{\pi}\left[R_{00}^{(1)}(c_1,1)\right]^2,\quad T_{12}^{y}=\frac{2c_2}{\pi}\left[R_{00}^{(1)}(c_2,1)\right]^2.\tag{5.265}$$

where $R_{00}^{(1)}$ is the radial prolate spheroidal function. Corresponding to (5.265), the eigenfunctions are respectively given by the angular prolate spheroidal wave functions $S_{00}(c_1,x/a_1)$ and $S_{00}(c_2,y/a_2)$. Some values of T_{12}^{x} are listed in Table 5.3. Observe that the power transmisson efficiency of 100% can be achieved by increasing the parameter c_1. As a result, the maximum power transmissin efficiency and the optimal distribution for aperture T_1 are respectively given by

$$\begin{aligned}T_{12}&=\frac{2c_1}{\pi}\left[R_{00}^{(1)}(c_1,1)\right]^2\frac{2c_2}{\pi}\left[R_{00}^{(1)}(c_2,1)\right]^2,\\ E_1(x,y)&=S_{00}(c_1,x/a_1)S_{00}(c_2,y/a_2)e^{jk(x^2+y^2)/2r_{1,2}}.\end{aligned}\tag{5.266}$$

The optimal distribution for aperture T_2 is given by (5.256).

Table 5.3 Largest eigenvalue

c_1	T_{12}^{x}
0.5	0.30969
1	0.57258
2	0.88056
4	0.99589
8	1.00000

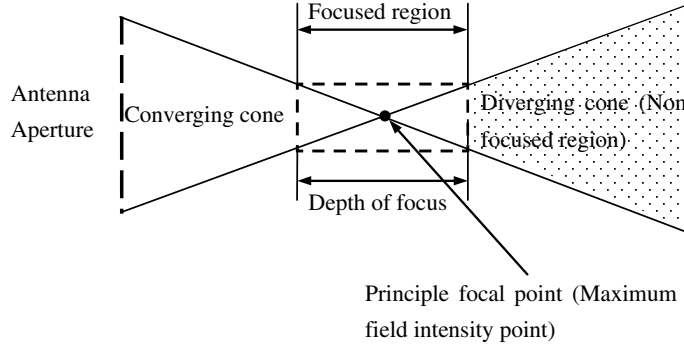


Figure 5.49 Properties of focused antenna aperture.

Equation (5.266) indicates that the transmitting aperture must be focused to the receiving aperture in order to achieve a maximum power transfer between the two apertures, and the optimization process yields a focused transmitting aperture. In addition the transverse amplitude distribution has no sidelobes. Some important properties of the focused antenna aperture are illustrated in Figure 5.49. The range between the axial $-3\,\mathrm{dB}$ points about the maximum intensity point (called principal focal point) is called the focused region and its extension is defined as the **depth of focus**.

In practice, it is difficult to realize the continuous distribution (1) with a single aperture. For this reason, an antenna array is usually adopted. □

5.10.4 Antenna Gain Measurement

Assume that antennas 1 and 2 are separated by a distance r_{12} and are located in the far-field region of each other. The Friis transmission formula indicates that if the gain of antenna 2 is known the measurement of the received power of antenna 2 actually gives the gain of antenna 1. To eliminate the properties of the testing antenna 2 we either can use a standard antenna whose gain is known, e.g., a dipole or calibrate the antenna 2 in a known field. We assume antenna 1 is transmitting and antenna 2 is receiving. We recall that antenna factor is defined as the ratio of the incident electric field at the receiving antenna to the voltage received at the antenna terminal, i.e.,

$$AF_2(\mathbf{r}_{12}) = \frac{|\mathbf{E}_1(\mathbf{r}_{12})|}{\left|V_2^{(1)}\right|}, \tag{5.267}$$

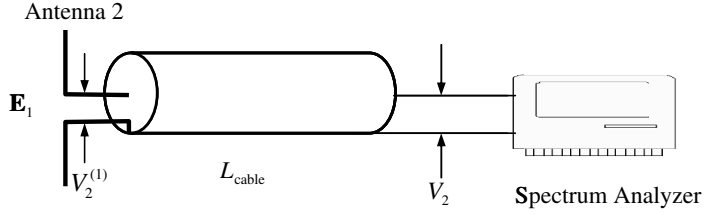


Figure 5.50 Receiving antenna.

as illustrated in Figure 5.50. Since the power density of the incident wave (Poynting vector) can be expressed as

$$S_1(\mathbf{r}_{12}) = \frac{1}{2}\frac{|\mathbf{E}_1(\mathbf{r}_{12})|^2}{\eta},$$

the received power is

$$P_2^{(1)} = S_1(\mathbf{r}_{12})A_2(\mathbf{r}_{12}) = \frac{|\mathbf{E}_1(\mathbf{r}_{12})|^2\lambda^2 G_2}{8\pi\eta}, \tag{5.268}$$

where $A_2 = \lambda^2 G_2/4\pi$ is the equivalent area of antenna 2. If the antenna under test is used as the transmitting antenna, then we have

$$|\mathbf{E}_1(\mathbf{r}_{12})| = \sqrt{2S_1(\mathbf{r}_{12})\eta} = \sqrt{\frac{P_1^{(1)}G_1}{2\pi r_{12}^2}\eta} = AF_2\left|V_2^{(1)}\right| = AF_2\,|V_2|\,L_{\text{cable}},$$

where L_{cable} is the loss of the cable connecting the antenna and the spectrum analyzer and V_2 is the voltage measured by the spectrum analyzer. So the gain for the antenna under test can be expressed as

$$G_1 = \frac{(2\pi)^2}{\eta}\frac{(AF_2\,|V_2|\,L_{\text{cable}}r_{12})^2}{P_1^{(1)}}. \tag{5.269}$$

If we use a dipole as the transmitting antenna to replace the antenna under test we will have the similar relationship as follows

$$G_{1d} = \frac{(2\pi)^2}{\eta}\frac{(AF_2|V_{2d}|L_{\text{cable}}r_{12})^2}{P_{1d}^{(1)}}. \tag{5.270}$$

It follows from (5.269) and (5.270) that

$$\frac{G_1}{G_{1d}} = \left(\frac{|V_2|}{|V_{2d}|}\right)^2\frac{P_{1d}^{(1)}}{P_1^{(1)}},$$

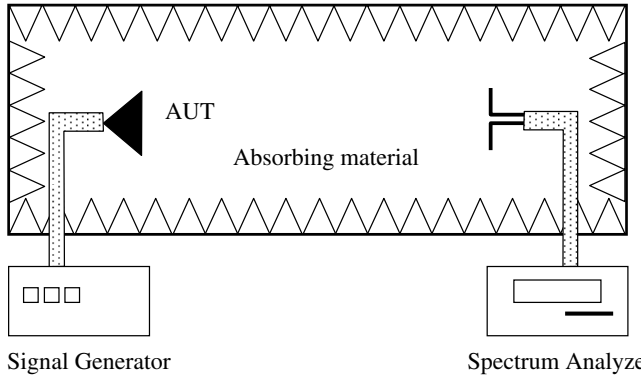


Figure 5.51 Setup of anechoic chamber.

which can be expressed in dB as

$$G_1(dBd) = V_2(dB) - V_{2d}(dB) - \left[P_1^{(1)}(dB) - P_{1d}^{(1)}(dB)\right]. \tag{5.271}$$

A typical setup of anechoic chamber for the gain measurement is shown in Figure 5.51.

5.11 Array Antennas

An **array antenna** is a group of radiators which are excited by currents with different amplitudes and phases. As a result of electromagnetic interference, the radiated fields are enhanced in the desired direction and cancelled in the non-desired direction.

5.11.1 A General Approach

Consider two rectangular coordinate systems $\mathbf{r} = (x, y, z)$ and $\tilde{\mathbf{r}}_n = (\tilde{x}_n, \tilde{y}_n, \tilde{z}_n)$, which are related by the following transformation

$$\mathbf{r} = \tilde{\mathbf{r}}_n + \mathbf{d}_n$$

as shown in Figure 5.52. The vector $\mathbf{d}_n$ starts at the origin of the coordinate system (x, y, z) and ends at the origin of the coordinate system $(\tilde{x}_n, \tilde{y}_n, \tilde{z}_n)$. The spherical coordinate systems associated with the coordinate systems $\mathbf{r}$ and $\tilde{\mathbf{r}}_n$ are denoted by (r, θ, φ) and $(\tilde{r}_n, \tilde{\theta}_n, \tilde{\varphi}_n)$ respectively. In the coordinate system $(\tilde{x}_n, \tilde{y}_n, \tilde{z}_n)$, the far-fields of an antenna are given by

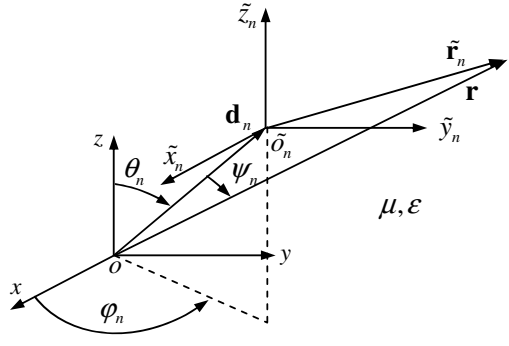


Figure 5.52 Transformation of coordinate systems.

(5.13), i.e.,

$$\mathbf{E}_n(\tilde{\mathbf{r}}_n) = -\frac{j\omega\mu I_n}{4\pi\tilde{r}_n} e^{-jk\tilde{r}_n} \mathbf{L}_n(\mathbf{u}_{\tilde{r}_n}), \tag{5.272}$$

where $\mathbf{L}_n$ is the vector effective length and $I_n = |I_n| e^{j\alpha_n}$ is the exciting current at the antenna feeding plane. Note that, as $\tilde{r}_n \to \infty$, we have

$$\mathbf{r}_n \parallel \mathbf{r}, \quad \mathbf{u}_\theta = \mathbf{u}_{\tilde{\theta}_n}, \quad \theta = \tilde{\theta}_n, \quad \varphi = \tilde{\varphi}_n. \tag{5.273}$$

Assume $|\mathbf{d}_n| \ll \tilde{r}_n$ and $\tilde{r}_n \to \infty$. Then, we can make the following approximation

$$\frac{e^{-jk\tilde{r}_n}}{\tilde{r}_n} \approx \frac{e^{-jkr}}{r} e^{jk|\mathbf{d}_n| \cos\psi_n}, \tag{5.274}$$

where

$$\cos\psi_n = \cos\theta\cos\theta_n + \sin\theta\sin\theta_n\cos(\varphi - \varphi_n). \tag{5.275}$$

Making use of (5.273) and (5.274), we have

$$\mathbf{E}_n(\mathbf{r}) = -\frac{j\omega\mu\,|I_n|\,e^{-jkr}}{4\pi r} e^{j(\alpha_n + k|\mathbf{d}_n|\cos\psi_n)} \mathbf{L}_n(\mathbf{u}_r). \tag{5.276}$$

This is the field expression in the coordinate system (x, y, z).

If we have N antennas so that (5.273) holds as $\tilde{r}_n \to \infty$ for each antenna n $(n = 1, 2, \ldots, N)$, then the total far-field can be written as

$$\mathbf{E}(\mathbf{r}) = \sum_{n=1}^{N} \mathbf{E}_n(\mathbf{r}) = -\frac{j\omega\mu}{4\pi r} e^{-jkr} \sum_{n=1}^{N} |I_n|\, \mathbf{L}_n(\mathbf{u}_r) e^{j(\alpha_n + k|\mathbf{d}_n|\cos\psi_n)}. \tag{5.277}$$

If all antennas are identical, we may write $\mathbf{L}_n(\mathbf{u}_r) = \mathbf{L}(\mathbf{u}_r)$ $(n = 1, 2, \ldots, N)$ and the above equation can be rewritten as

$$\mathbf{E}(\mathbf{r}) = \mathbf{E}_e(\mathbf{r}) \times \mathrm{AF}. \tag{5.278}$$

where

$$\mathrm{AF} = \sum_{n=1}^{N} |I_n| \, e^{j(\alpha_n + k|\mathbf{d}_n| \cos \psi_n)}, \tag{5.279}$$

is referred to as **antenna array factor** and

$$\mathbf{E}_e(\mathbf{r}) = -\frac{j\omega\mu}{4\pi r} e^{-jkr} \mathbf{L}(\mathbf{u}_r)$$

is the far-field of a single antenna normalized to the exciting current. Equation (5.278) is known as the principle of **pattern multiplication** for arrays of identical elements.

Example 5.8: Let us consider two z-directed dipoles separated by a distance d as illustrated in Figure 5.53. In this case, we have

$$\theta_1 = \frac{\pi}{2}, \quad \varphi_1 = 0, \quad \theta_2 = \frac{\pi}{2}, \quad \varphi_2 = \pi.$$

It follows from (5.275) that

$$\cos \psi_1 = \sin \theta \cos \varphi, \quad \cos \psi_1 = -\sin \theta \cos \varphi.$$

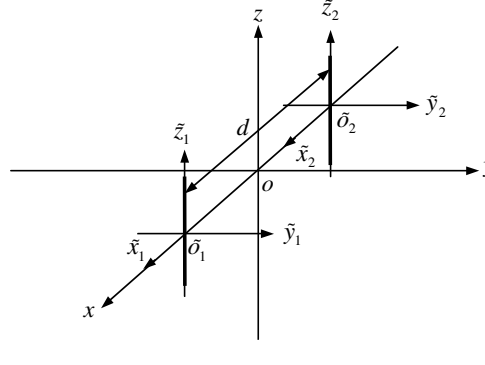


Figure 5.53 Two dipole arrays.

Assume that $I_1 = I\angle 0°, I_2 = I\angle\alpha$. Then the array factor for the two-dipole array is given by

$$\begin{aligned} \text{AF} &= Ie^{j\frac{kd}{2}\sin\theta\cos\varphi} + Ie^{j\left(\alpha-\frac{kd}{2}\sin\theta\cos\varphi\right)} \\ &= 2Ie^{j\frac{\alpha}{2}}\cos\left(\frac{\alpha}{2} - \frac{kd}{2}\sin\theta\cos\varphi\right). \quad \square \end{aligned}$$

5.11.2 Yagi–Uda Antenna

The **Yagi–Uda antenna** was first invented by Shintaro Uda in 1926 and reported in an English journal by his colleague Hidetsugu Yagi. A Yagi–Uda antenna is a directional antenna consisting of a **driven element**, typically a dipole or folded dipole, and additional parasitic elements called **reflector** or **directors**, as shown in Figure 5.54. The driven element is excited by an applied source, whose length is slightly less than half wavelength, typically 0.45λ to 0.49λ. The reflector is longer than the driven element (typically 0.5λ) and is about a quarter wavelength away from the driven element. Experience shows that using more than one reflector does not help much. The directors may have different lengths which range from 0.4λ to 0.45λ, and their separations may vary from 0.3λ to 0.4λ, as illustrated in Figure 5.54. The number of directors varies from 6 to 12. Experience has shown that the directivity can be improved by increasing the number of directors. It has been found that uniformly spaced directors of equal lengths do not make an optimum array, and analytical methods have been developed for maximizing the directivity by adjusting both the

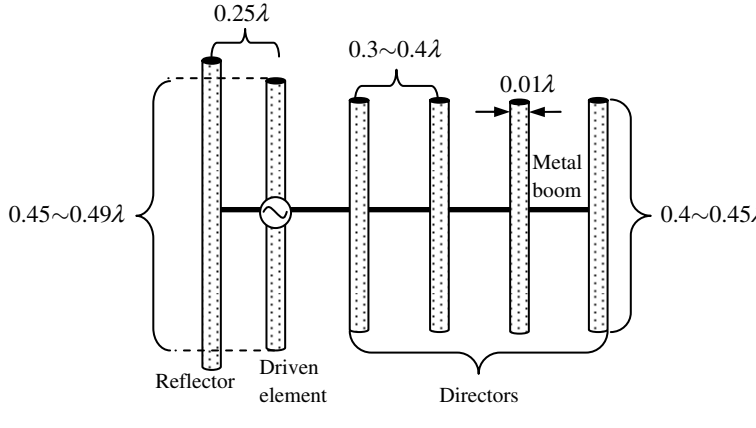


Figure 5.54 Yagi–Uda antenna array.

inter-element spacings and the lengths of the array elements (Cheng and Chen, 1973; Chen and Cheng, 1975).

Yagi–Uda antennas are directional along the axis perpendicular to the dipole from the reflector toward the driven element and the directors. By adjusting the distance between the adjacent directors it is possible to reduce the back lobe of the radiation pattern. The bandwidth of the Yagi–Uda antenna can be enhanced by increasing the length of the reflector and decreasing the length of directors at a sacrifice in gain.

5.11.3 Log Periodic Antennas

The log-periodic antenna was first introduced by DuHamel and Isbell (1957) and is illustrated in Figure 5.55, which has a toothed design cut out of sheet metal. It is assumed that the successive radii are in the common ratio

$$\frac{R_{n+1}}{R_n} = \frac{r_{n+1}}{r_n} = \tau.$$

If the shape of the original antenna is described by $r = f(\theta)$, the new antenna described by $r = Kf(\theta)$ can be made congruent to the original antenna only if K satisfies $K = \tau^m$, where m is an integer. The antenna of Figure 5.55 will have the same pattern and impedance at any two

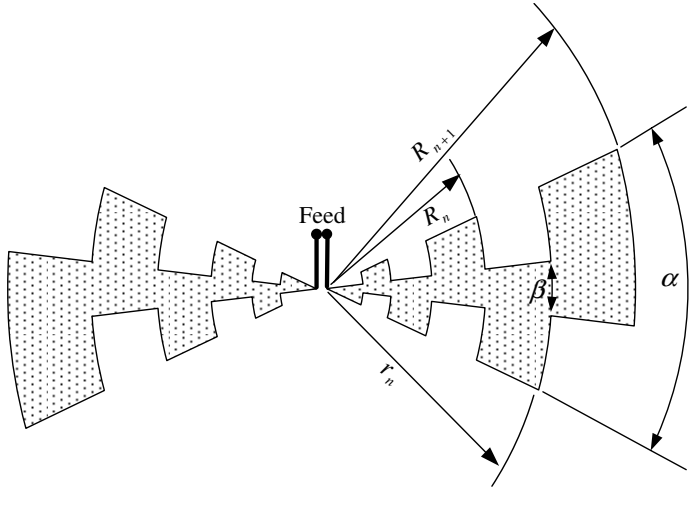


Figure 5.55 Log-periodic antenna.

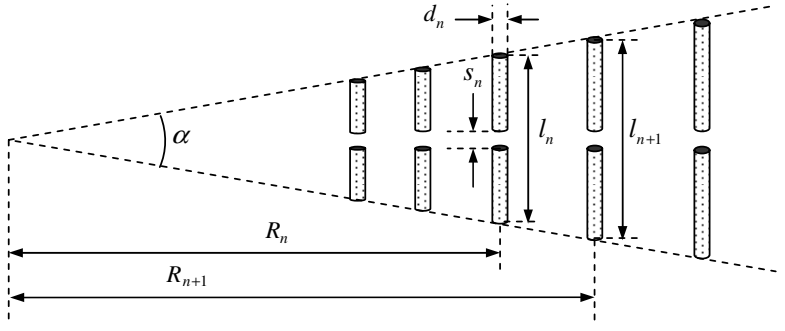


Figure 5.56 Dipole array.

frequencies f_1 and f_2 in the ratio τ: $f_1 = \tau f_2$, or

$$\log f_2 = \log f_1 + \log(1/\tau).$$

For this reason, the configuration is called a logarithmically periodic planar antenna with a period $\log(1/\tau)$ although antenna characteristics will change between the frequencies f_1 and f_2. The planar log-periodic antenna is bidirectional. It can be made unidirectional if the two halves of the antenna are folded to form a wedge-like structure. The main beam will point off the direction of the apex.

The log-periodic antenna can be realized in several different ways. The most common log periodic array consists of a number of dipole elements, diminishing in length and separation between elements from the back (big end) toward the front (small end), as shown in Figure 5.56. The element at the back of the array is a half wavelength at the lowest frequency of operation and the main beam comes from the smaller front. It has many similarities to the Yagi–Uda array. The log-periodic array is much larger than a Yagi–Uda antenna that produces the same gain. But the Yagi–Uda has a narrower bandwidth. The various dimensions of the dipole array increase logarithmically:

$$\tau = \frac{R_{n+1}}{R_n} = \frac{l_{n+1}}{l_n} = \frac{s_{n+1}}{s_n} = \frac{d_{n+1}}{d_n}.$$

5.11.4 Optimal Design of Multiple Antenna Systems

Antenna arrays can be designed with a fixed beam or a scanned beam. The scanning antenna arrays can be categorized as either switched beam system or adaptive antenna system. A **switched beam system** has a finite

number of predefined patterns or combining strategies. It detects the signal strength, choose from one of the predefined patterns, and change from one beam to another as the mobile user moves around. **An adaptive antenna system** (or **smart antenna system**) is defined as a multiple antenna system combined with signal processing in both space and time, which is capable of optimizing the radiation patterns automatically in response to the signal environment. The adaptive array system has more degrees of freedom in choosing radiation patterns, which are scenario-based and can be adjusted in real time. Compared to the switched beam system, the adaptive array system can direct the main beam toward the signal of interests and supress the radiation pattern in the direction of interference. The main benefits of the scanning arrays are summarized below:

(1) High antenna gain: Multiple antennas are combined together to increase the antenna gain in the direction of signal of interests, which results in a better coverage and longer battery life for handsets.
(2) Spatial diversity: Multiple antennas are used to minimize the fading caused by multipath propagations.
(3) Interference rejection: Multiple antennas can be combined to steer its beam toward a desired signal while steering a null toward an undesired or interfering signal, improving the signal-to-interference ratio of the received signals and thus increasing information capacity.

Figure 5.57 shows a typical arrangement of an adaptive antenna array. A phase shifter and an attenuator are installed at each element, which provide proper phase and amplitude to adjust the beam direction and the pattern shape. The amplitudes control primarily the shape of the pattern while the phases control primarily the beam direction. A digital signal process is usually deployed with the scanned antenna array to estimate the direction of arrival (DOA) of all incoming signals and the magnitude and phase of each antenna element so that a desired pattern can be achieved. This process is called **adaptive beamforming**, and many beamforming algorithms have been proposed and have been successfully applied to various signal environments (Widrow *et al.*, 1967; Balanis, 2005).

The antenna array theory involves two important aspects, the array analysis and the array synthesis. In the analysis, the radiation pattern of the antenna array is studied under the assumption that the antenna configuration and the excitation distribution in the array are known. The excitation distributions can be uniform amplitude or tapered amplitude with uniform progressive phase, etc; and the array configuration can be a

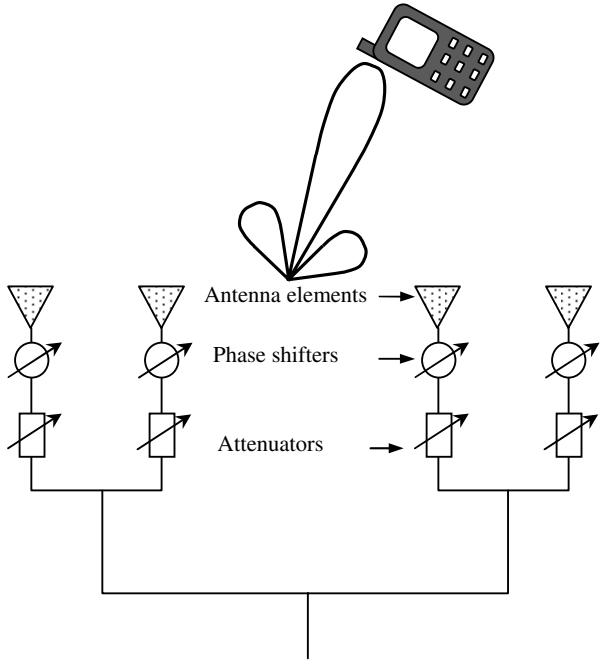


Figure 5.57 A typical arrangement of an adaptive antenna array.

straight line, a circle or a rectangle. The radiation properties of an antenna array (such as beam width, sidelobes, directivity, etc.) are determined by the number of elements, the spacing between elements, the amplitude and phase of the excitation distribution. In synthesis, one starts with the desired array pattern to find the excitation distribution for the antenna array to be designed. For most practical applications, the antenna arrays consist of identical elements which are equally spaced and similarly oriented. Once the relative physical positioning of the antenna elements is fixed, the main task of antenna array synthesis is to find the excitation of individual elements so that electromagnetic energy can be directed toward the desired direction in a specified manner. For antenna arrays with simple geometrical configurations and identical elements, the far-fields can be expressed as the product of the field of a single element at a selected reference point and the array factor. The array factor can be used to determine phase difference between elements so that the main beam can be adjusted to any desired direction to form a scanning array. In order to control the sidelobes, the array pattern may be prescribed by using polynomial approximations, such

as Chebyshev polynomials. In this section, we present a universal method for the design of antenna array, which is based on the optimization of power transmission efficiency between the antenna array to be designed and a test antenna array whose elements are placed in the desired directions so that the radiations can be optimized in these directions. The optimized distribution of excitations may be determined by an eigenvalue equation derived from the Rayleigh quotient for the power transmission efficiency. The method is applicable to any type of antenna array.

5.11.4.1 *Power Transmission between Two Antenna Arrays*

In order to beam the energy from an antenna array to different directions simultaneously, we may consider an $n_t + n_r$ antenna system shown in Figure 5.58(a), in which antennas $1 \sim n_t$ are transmitting and represent the antenna array to be designed while antennas $n_t + 1 \sim n_t + n_r$ are receiving and represent the test antennas. The test antennas are placed in the desired directions in which radiations need to be optimized. This system

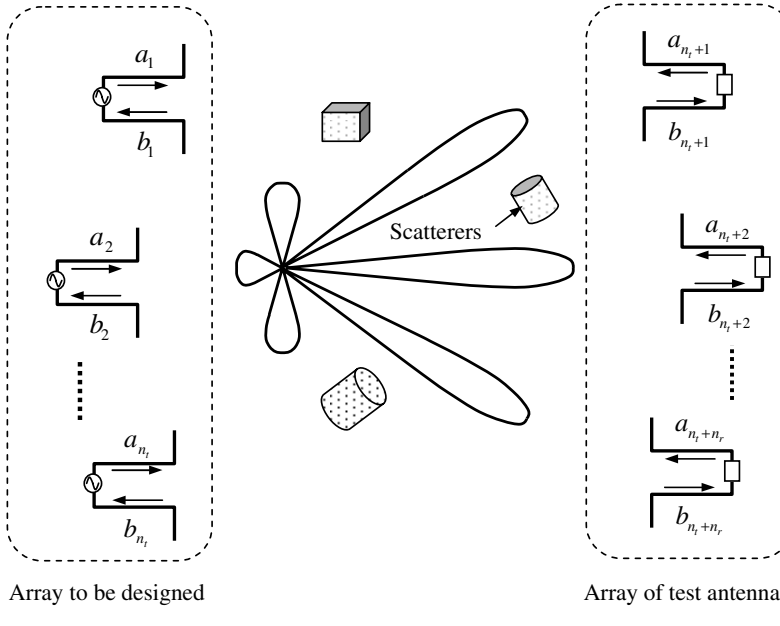


Figure 5.58 Power transmission between two antenna arrays.

can be described as an $n_t + n_r$-port network and can be characterized by scattering parameters as follows

$$\begin{bmatrix}[b_t]\\ [b_r]\end{bmatrix} = \begin{bmatrix}[S_{tt}] & [S_{tr}]\\ [S_{rt}] & [S_{rr}]\end{bmatrix}\begin{bmatrix}[a_t]\\ [a_r]\end{bmatrix}, \tag{5.280}$$

where the normalized incident and reflected waves for transmitting antenna array and test antenna array are respectively given by

$$\begin{aligned}[a_t] &= [a_1, a_2, \ldots, a_{n_t}]^T,\\ [b_t] &= [b_1, b_2, \ldots, b_{n_t}]^T,\\ [a_r] &= [a_{n_t+1}, a_{n_t+2}, \ldots, a_{n_t+n_r}]^T,\\ [b_r] &= [b_{n_t+1}, b_{n_t+2}, \ldots, b_{n_t+n_r}]^T.\end{aligned}$$

The power transmission efficiency between the two antenna arrays is defined as the ratio of the power delivered to the loads of the test antenna array to the input power to the transmitting antenna array

$$T_{\text{array}} = \frac{\frac{1}{2}(|[b_r]|^2 - |[a_r]|^2)}{\frac{1}{2}(|[a_t]|^2 - |[b_t]|^2)}. \tag{5.281}$$

Assume that the test antenna array is matched so we have $[a_r] = 0$. Making use of (5.280), (5.281) can be written as the well-known Rayleigh quotient

$$T_{\text{array}} = \frac{([A][a_t], [a_t])}{([B][a_t], [a_t])}, \tag{5.282}$$

where $(\cdot, \cdot)$ denotes the usual inner product of two column vectors; $[A]$ and $[B]$ are two matrices defined by

$$[A] = [\bar{S}_{rt}]^T S_{rt}, \quad [B] = [1] - [\bar{S}_{tt}]^T [S_{tt}].$$

If the power transmission efficiency T_{array} reaches the maximum at $[a_t]$, then we have

$$[A][a_t] = T_{\text{array}}[B][a_t]. \tag{5.283}$$

The maximum possible value of T_{array} is the largest eigenvalue of (5.283) and can be found numerically. If the whole antenna system is matched, (5.283) reduces to

$$[A][a_t] = T_{\text{array}}[a_t]. \tag{5.284}$$

Note that the surrounding materials and antenna types are assumed to be arbitrary in deriving (5.283) and (5.284). Therefore, (5.283) and (5.284)

applies to an arbitrary antenna array system. For $n_r = 1$ (a single test antenna), (5.284) has only one positive eigenvalue and the rest are zero since the rank of the matrix $[A]$ is unit in this case. Therefore, the unique non-zero eigenvalue of (5.284) gives the maximum power transmission efficiency between the antenna array and the test antenna and the corresponding eigenvector is the optimal excitation for the antenna array in the sense that the energy transmitted in the direction of the test antenna is maximized. The eigenvectors corresponding to the zero eigenvalues of (5.284) are the optimal excitation distributions for the antenna array in the sense that the energy transmitted in the direction of the test antenna is minimized and thus a null appears in the desired direction.

5.11.4.2 *Optimal Design of Antenna Arrays*

In order to illustrate the optimization method presented above, we consider a 4×4 microstrip patch antenna array operating at 2.45 GHz and built on FR-4 substrate with relative dielectric constant 4.4, loss tangent 0.02 and thickness of 3 mm. The antenna element is a rectangular microstrip patch with an inset-feed with the length and the width of the patch being 29 mm and 28 mm respectively, as shown in Figure 5.59(a). The length of the feed line is 12 mm and the width is 3 mm which is combined with the 6 mm inset to achieve 50 Ω characteristic impedance. By properly selecting the

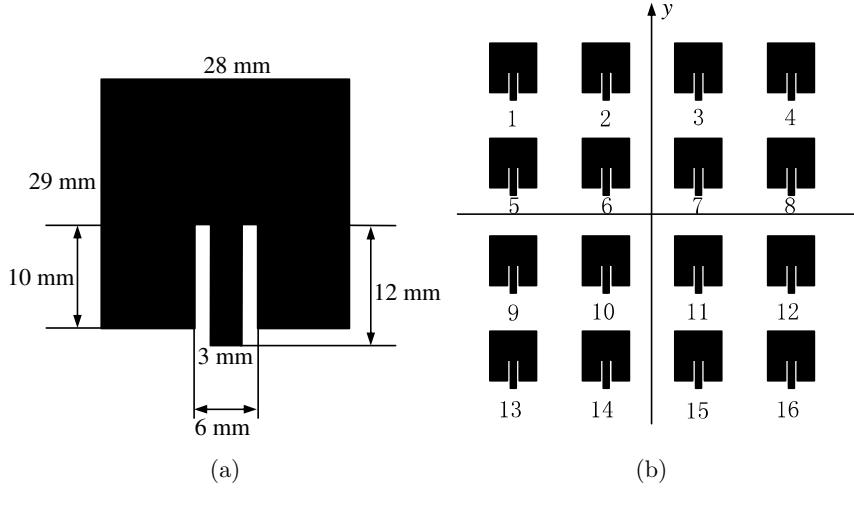


Figure 5.59 (a) Element of microstrip array. (b) Arrangement of array elements.

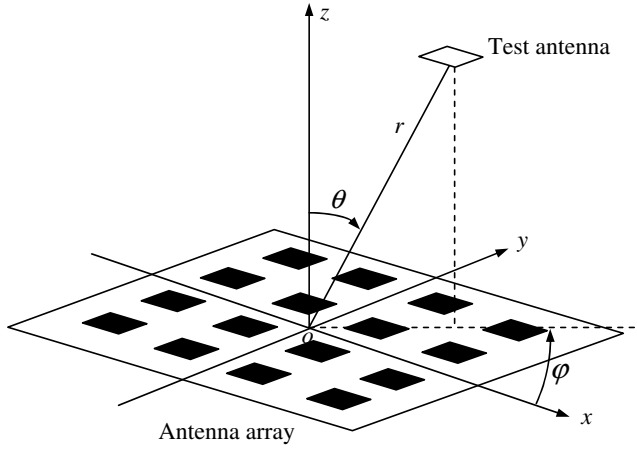


Figure 5.60 Arrangement of the 4×4 microstrip patch array the test antenna.

depth of the inset, a good matching can be achieved without additional matching elements. Due to the mutual coupling between antenna elements, the spacing between two neighboring elements must be carefully designed to avoid grating sidelobes. The distance between the neighboring elements is chosen as 55 mm for the 4×4 array. Figure 5.60 shows an arrangement of the 4×4 array with a test antenna, which is assumed to be the same as the elementary patch in the array and is perpendicular to the desired direction $\mathbf{u}_r$ (unit vector along r). In what follows, the 4×4 array will be used to explain the design procedure for various applications (Geyi, 2014).

Example 5.9 (Focused array design)**:** Focused antennas are used to focus the electromagnetic energy to a spot to reach a high power density in the radiating near field (Fresnel) region. They have wide applications in many areas such as noncontract microwave sensing, medical treatment and wireless power transmission. Consider the 4×4 array shown in Figure 5.59(b) and assume that the antenna array must be focused at $z = 100\,\text{mm}$. We may place the test antenna at $(r, \theta, \varphi) = (100\,\text{mm}, 0, 0)$. The whole system is simulated with simulation tools with one port being active and rest terminated in $50\,\Omega$, which generates the scattering parameters for the whole system. The maximum power transmission efficiency T_{array} can be obtained from (5.284) by using the computed scattering parameters. For the current configuration of the system, the maximum power transmission efficiency between the antenna array and the test antenna is found to be 30%. The corresponding optimized distribution of excitation $[a_t]$ is listed

Table 5.4 Optimized distribution of excitations

Port No.	Excitation of port	Port No.	Excitation of port
1	$0.13\angle 0°$	9	$0.22\angle -76°$
2	$0.2\angle -69°$	10	$0.38\angle -138°$
3	$0.2\angle -69°$	11	$0.38\angle -138°$
4	$0.13\angle 0°$	12	$0.22\angle -76°$
5	$0.22\angle -76°$	13	$0.13\angle 0°$
6	$0.38\angle -138°$	14	$0.2\angle -69°$
7	$0.38\angle -138°$	15	$0.2\angle -69°$
8	$0.22\angle -76°$	16	$0.13\angle 0°$

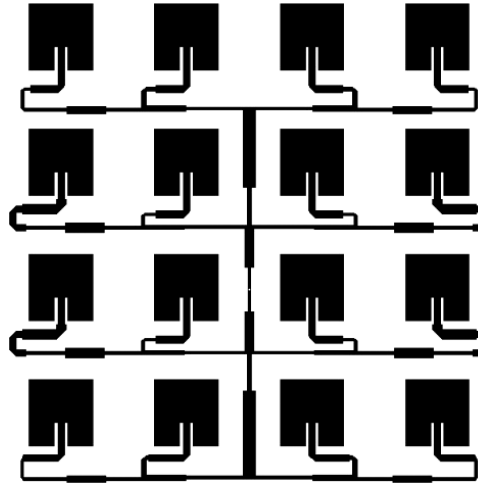


Figure 5.61 A 4×4 focused antenna array.

in Table 5.4. Note that the optimized phase distribution obeys spherical distribution. The simulation tools may be used to model the feeding network to achieve the optimized distribution of excitation at the outputs of the feeding network. The phase distribution can be realized by adjusting the length of the feeding line, and the amplitude distribution can be achieved by power dividers with different choices of width for the feeding lines. During the simulation of the feeding network, each antenna element connected to the feeding network is replaced by a $50\,\Omega$ termination. Finally, the feeding network and the antenna array are joined together and simulated as a whole to ensure that the outputs of the feeding network agree well with the optimized values. The final design of the 4×4 microstrip array is displayed in Figure 5.61. □

Table 5.5 Optimized distribution of excitations

Port No.	Excitation of port	Port No.	Excitation of port
1	$0.25\angle -46°$	9	$0.25\angle -46°$
2	$0.25\angle -46°$	10	$0.25\angle -46°$
3	$0.25\angle -46°$	11	$0.25\angle -46°$
4	$0.25\angle -46°$	12	$0.25\angle -46°$
5	$0.25\angle -46°$	13	$0.25\angle -46°$
6	$0.25\angle -46°$	14	$0.25\angle -46°$
7	$0.25\angle -46°$	15	$0.25\angle -46°$
8	$0.25\angle -46°$	16	$0.25\angle -46°$

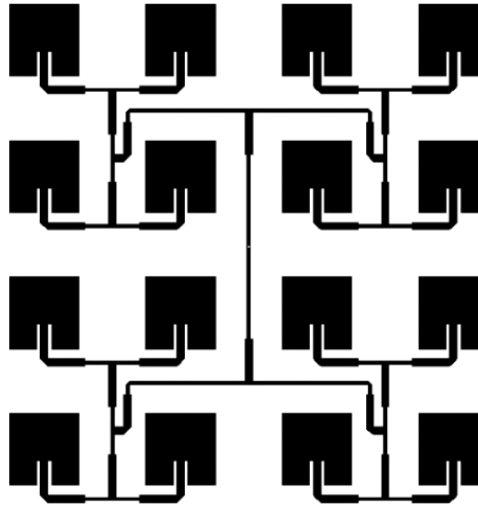


Figure 5.62 Equally excited antenna array.

Example 5.10 (Phased array design)**:** Consider the array configuration shown in Figure 5.59(b) again and assume that the desired signal is in $(0°, 0°)$ direction. In order to beam the energy to the desired direction, the test antenna is placed at $(2.5\,\text{m}, 0, 0)$, which is in the far-field region of the antenna array. The optimized excitation distribution is listed in Table 5.5 and uniform distributions for both amplitude and phase have been obtained, which agrees with our common understanding. The antenna array implemented with the feeding network is displayed in Figure 5.62.

Reexamine the 4×4 microstrip array shown in Figure 5.59(b) and assume that the desired signal is in $(60°, 90°)$ direction. To direct the energy to the desired direction, the test antenna is placed at $(2\,\text{m}, 60°, 90°)$.

Table 5.6 Optimized distribution of excitations

Port No.	Excitation of port	Port No.	Excitation of port
1	$0.2\angle 83°$	9	$0.21\angle 90°$
2	$0.25\angle -61°$	10	$0.21\angle -62°$
3	$0.28\angle 147°$	11	$0.23\angle 148°$
4	$0.3\angle 0°$	12	$0.28\angle 0°$
5	$0.21\angle 91°$	13	$0.2\angle 82°$
6	$0.21\angle -61°$	14	$0.25\angle -61°$
7	$0.23\angle 148°$	15	$0.28\angle 147°$
8	$0.28\angle 0°$	16	$0.3\angle 0°$

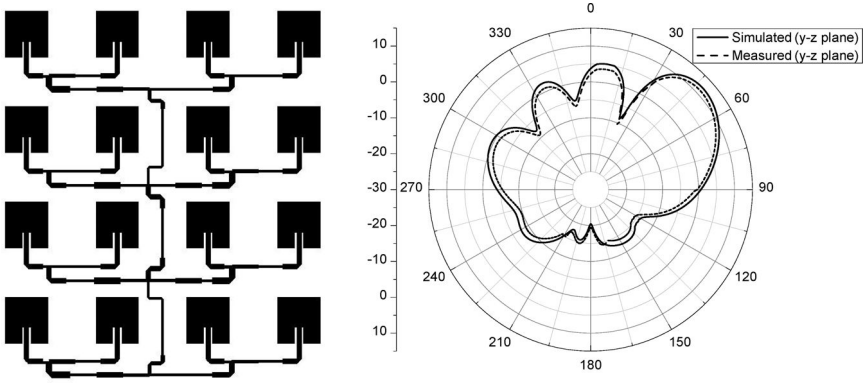


Figure 5.63 Simulated and measured radiation patterns on (y, z)-plane.

Table 5.6 gives the optimized distribution of excitations for the antenna array, and Figure 5.63 shows the corresponding radiation pattern. Apparently, the radiated energy is directed toward the desired direction. □

Example 5.11 (Multi-beam antenna design)**:** A multi-beam antenna can access multiple targets simultaneously, which is useful in many areas. The satellite communication system and the space division multiple access (SDMA) techniques all involve multi-beam antennas to either achieve seamless connection or increase the capacity of the system. In order to generate multiple beams, we may place the test antennas in the desired directions. Let us reconsider the antenna array in Figure 5.59(b) and assume that the desired signals are in the directions $(20°, 90°)$, $(60°, 90°)$, $(20°, 270°)$ and $(60°, 270°)$ directions. Four test antennas may be respectively placed at $(2\,\text{m}, 20°, 90°)$, $(2\,\text{m}, 60°, 90°)$, $(2\,\text{m}, 20°, 270°)$, and $(2\,\text{m}, 60°, 270°)$. The optimized distribution of excitations for the antenna

Table 5.7 Optimized distribution of excitations

Port No.	Excitation of port	Port No.	Excitation of port
1	$0.33\angle 179°$	9	$0.33\angle 175°$
2	$0.13\angle 175°$	10	$0.13\angle 170°$
3	$0.13\angle -2°$	11	$0.13\angle -8°$
4	$0.33\angle 0°$	12	$0.33\angle -0°$
5	$0.33\angle 175°$	13	$0.33\angle 179°$
6	$0.13\angle 169°$	14	$0.13\angle 175°$
7	$0.13\angle -8°$	15	$0.13\angle -3°$
8	$0.33\angle -4°$	16	$0.33\angle 0°$

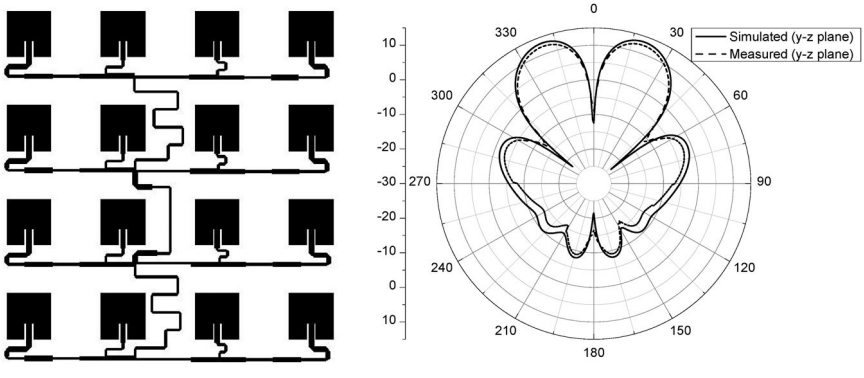


Figure 5.64 Simulated and measured radiation pattern on (y, z)-plane.

array is listed in Table 5.7. Figure 5.64 shows the corresponding radiation pattern, which consists of four major beams directed toward the desired directions. □

It is the theory which decides what we can observe.

—Albert Einstein

Chapter 6

Propagation of Radio Waves

> You make experiments and I make theories. Do you know the difference? A theory is something nobody believes, except the person who made it. An experiment is something everybody believes, except the person who made it.
>
> —Albert Einstein

When radio waves travel in free space, nothing is going to affect them. The real environment is, however, not a free space and many factors may change propagation properties of radio waves. The Earth's atmosphere (troposphere, stratosphere and ionosphere), the ground, mountains, buildings, and weather conditions all have major influences on wave propagations.

A **radio propagation model** is a mathematical formulation for the characterization of radio wave propagation as a function of frequency, distance and other conditions. A single model is usually developed to predict the behavior of propagation for all similar links under similar constraints, typically the path loss along a link or the effective coverage area of a transmitter.

When a wave strikes an obstacle it will be reflected, refracted and diffracted. **Reflection** occurs when a wave is incident upon a flat surface with large radii of curvature compared to the wavelength, and the amount of the reflection depends on the properties of the obstacle. The phenomenon of reflection obeys Snell's law and can be characterized by the **coefficient of reflection**, which is defined as the ratio of the reflected wave to the incident wave. **Refraction** of waves involves a change in the direction of the wave as it passes into the obstacle. Refraction, or bending of path of the waves, is accompanied by a change in speed and wavelength of the wave. The phenomenon of refraction obeys Snell's law and may be characterized by the **coefficient of refraction**, which is defined as the ratio of the

transmitted wave into the obstacle to the incident wave. The coefficients of reflection and refraction are functions of the medium properties, wave polarization, the angle of incidence and the wave frequency. **Diffraction** means the ability of the radio wave to turn sharp corners and bend around the obstacle, which is the result of Huygens' principle. Low frequency radio wave whose wavelength is longer than the maximum size of an obstacle can be easily propagated around the obstruction. When frequency increases the obstacle causes more and more attenuation and a shadow zone (an area where the electromagnetic fields are very weak) on the opposite side of the incidence develops. When a wave impinges upon an obstacle of small dimensions compared to the wavelength, it will be reflected in various directions. This phenomenon is referred to as **scattering**. The electromagnetic wave can also be guided by certain environments such as tunnels and corridors.

The most troublesome and frustrating problem in receiving radio signals is variation in signal strength, known as **fading**. Several conditions can produce fading, which include the change of polarization of the wave, absorption of the electromagnetic energy in the environment, e.g., by ionosphere and multipath that a radio wave may follow between the transmitter and receiver. Fading may be roughly classified into two categories: large-scale and small-scale fading. **Large-scale fading**, also called slow fading, is based on the observation made over long separation distances between the transmitter and receiver (from several hundreds to thousands of meters) and is attributed to shadowing and diffraction. The effects of diffraction paths to the path loss are multiplicative and become additive if expressed in dB. From the central limit theorem, a Gaussian random variable can be used to represent the pass-loss if the number of diffraction paths to the receiver is sufficiently large. As a result, the large-scale fading obeys a normal distribution in dB, called **log-normal distribution**, and is modeled as a log-normal random variable. Understanding large-scale fading is important to cell-site planning. **Small-scale fading** is based on the observation made over short travel distances between the transmitter and receiver (a few wavelengths) and is due to multipath. The level of received signal is the vector sum of all individual signals and they may add constructively or destructively. If the receiver is moving over a distance in the order of a wavelength or more, the received signal strength will fluctuate rapidly. In practice, the number of paths between a transmitter and receiver may be very large. According to the central limit theorem, both the in-phase and quadrature components of the received signal can be treated as a Gaussian random variable, which yields a Rayleigh (or Rician) density function for

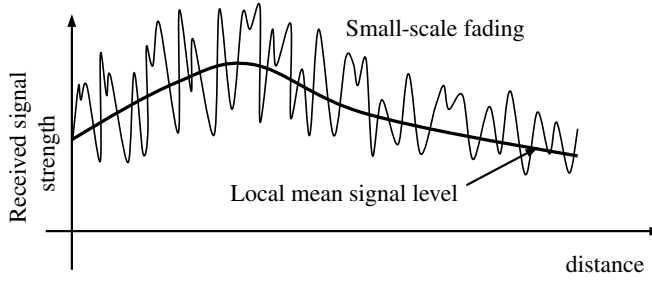


Figure 6.1 Fading over multiple distance-scales.

the amplitude and a uniform probability density function for the phase. If there is no direct path (i.e., line-of-sight) between the transmitter and receiver the fading envelope has a Rayleigh distribution and is typified by an urban or in-building environment. If there is a relatively strong direct component of signal the fades will be less deep and the fading envelope has a Rician distribution. This type of fading is likely to occur in rural environments. Understanding small-scale multipath fading is important to the design of reliable communication systems. The typical overall effect on the signal strength for a mobile receiver is illustrated in Figure 6.1. It is shown that the small-scale fading is superimposed on the local mean signal level. The mean signal level varies very slowly and has a log-normal distribution, which corresponds to the large-scale fading.

6.1 Earth's Atmosphere

The atmosphere of Earth, usually called air, is a layer of gases surrounding the Earth that is retained by Earth's gravity and has a mass of about 5×10^{18} kg. The atmosphere protects life on Earth by absorbing ultraviolet solar radiation, warming the surface through heat retention, and reducing temperature extremes between day and night.

6.1.1 Structure of Atmosphere

There is no distinct boundary between the atmosphere and outer space. The Kármán line, named after the Hungarian-American engineer and physicist Theodore von Kármán, (1881–1963), at 100 km above the Earth's surface, is often used as the boundary between the atmosphere and outer space. Figure 6.2 shows the different atmosphere layers and the variations of physical parameters of the atmosphere with the altitude. The troposphere

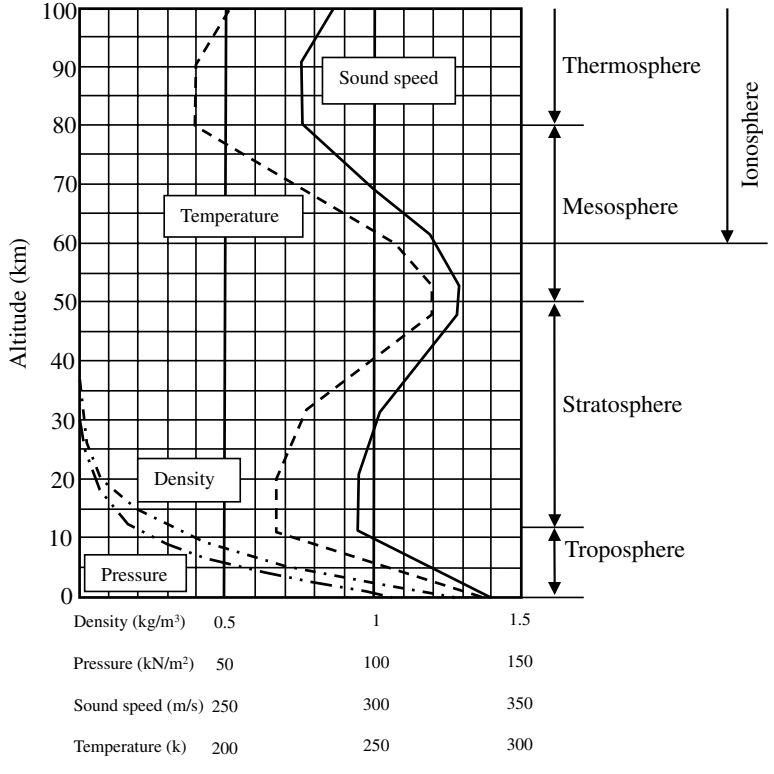


Figure 6.2 Principal layers of atmosphere.

begins at the Earth's surface and extends to between 9 km at the poles and 17 km at the equator. The troposphere contains roughly 80% of the mass of the atmosphere and 99% of its water vapor and aerosols (aerosols are extremely fine particles suspended in the atmosphere, either in the form of liquid or solid). The stratosphere extends from about 12 km to 50 km, where the horizontal mixing of gaseous components proceeds much more rapidly than in vertical mixing. The ozone layer is contained in the lower portion of the stratosphere from approximately 20 to 30 km, which contains relatively high concentrations of ozone (O_3) and absorbs most of the Sun's ultraviolet (UV) radiation. The mesosphere extends from about 50 km to about 80–85 km. It is the layer where most meteors burn up upon entering the atmosphere. The thermosphere extends from about 80 km to about 350–800 km. Within this layer, UV causes ionization, and temperatures highly depend on solar activity and can rise to 2,000°C. The outermost

layer of Earth's atmosphere is exosphere, which extends from about 500–1000 km to about 100,000 km and is usually considered a part of outer space. The exosphere is a transitional zone where the atmosphere thins out and merges with interplanetary space and no longer behaves like a gas. It mainly composed of hydrogen and helium with some nitrogen, carbon dioxide, and atomic oxygen near its base.

The ionosphere is a region of the upper atmosphere, which extends from about 60 km to 600 km, and includes the thermosphere and parts of the mesosphere and exosphere. In this region, radiation from the Sun causes the atmosphere particles to become electrically charged (photoelectric effect). Based on what wavelength of solar radiation is absorbed most frequently, the ionosphere can further be divided into three sublayers, the D, E and F layers. The D layer is the lowest in altitude, which extends approximately from 75 km to 95 km. In this region, ionization causes the absorption of the most energetic radiation and high frequency (HF) radio waves. The E layer extends approximately from 95 km to 150 km, and absorbs soft X-rays. The F layer starts around 150 km and ends at 600 km approximately. Extreme UV radiation is absorbed in this layer. Different layers of the ionosphere make long distance radio communication (beyond the horizon) possible by reflecting the radio waves back to Earth. The D and E layers reflect AM radio waves back to Earth. Radio waves with shorter wavelengths are reflected by the F region. Visible light, television and FM wavelengths are all too short to be reflected by the ionosphere.

The constituents of the atmosphere are shown in Table 6.1. By volume, dry air contains 78.09% nitrogen, 20.95% oxygen, 0.93% argon, 0.03% carbon dioxide, and small amounts of other gases. Air also contains a variable amount of water vapor, on average around 1%. The air constituents suitable for the survival of plants and animals only exist in Earth's troposphere and artificial atmospheres.

The Earth's atmosphere is characterized by a number of parameters such as pressure, temperature, humidity, the direction and speed of the winds, precipitations, evaporation, radiation, sunshine duration, horizontal visibility, electronic density, etc.

6.1.2 Weather Phenomena

Weather is the state of the atmosphere, which is driven by the differences of air pressure, temperature and moisture. Weather generally refers to day-to-day temperature and precipitation activity and is influenced by a number

Table 6.1 Constituents of the dry atmosphere by volume

Gas	Symbol	Content
Nitrogen	N_2	78.084%
Oxygen	O_2	20.947%
Argon	Ar	0.934%
Carbon dioxide	CO_2	0.033%
Neon	Ne	18.20 parts per million
Helium	He	5.20 parts per million
Methane	CH_4	2.00 parts per million
Krypton	Kr	1.10 parts per million
Sulfur dioxide	SO_2	1.00 parts per million
Hydrogen	H_2	0.50 parts per million
Nitrous oxide	N_2O	0.50 parts per million
Xenon	Xe	0.09 parts per million
Ozone	O_3	0.07 parts per million
Nitrogen dioxide	NO_2	0.02 parts per million
Iodine	I_2	0.01 parts per million
Carbon monoxide	CO	trace
Ammonia	NH_3	trace

of factors including the seasons, the altitude, the latitude, and the Earth's magnetic field. Most weather phenomena occur in the troposphere and are summarized below:

Solar radiation: Solar radiation is the electromagnetic fields emitted by the Sun, which increases the temperature at the surface of the Earth, and is expressed in W/m^2.

Evaporation: Evaporation is a phase transition from the liquid phase to gas phase that occurs at temperatures below the boiling temperature at a given pressure. Evaporation usually occurs on the surface. The energy necessary for the evaporation of water causes a decrease in the temperature.

Condensation: Condensation is the change of the physical state of matter from gaseous phase into liquid phase. Cloud condensation nuclei are small particles typically $0.2\,\mu$m, on which water vapor condenses to form cloud droplets. Condensation nuclei include sea salt crystals, mineral particles (such as dust, sand and smoke), and charged particles.

Freezing and melting: Freezing or solidification is a phase transition from the liquid phase into solid phase when the temperature of the liquid is lowered below its freezing point. At temperatures lower than $0°C$, water is solidified into snow or ice. Melting is the reverse process.

Supercooling: Supercooling or undercooling is a process of lowering the temperature of a liquid or a gas below its freezing point without becoming a solid. Water droplets in the atmosphere often remain in liquid state at temperatures significantly lower than 0°C.

Reverse sublimation: Sublimation is a phase transition directly from the solid into the gas phase without passing through an intermediate liquid phase. Reverse sublimation refers to the process that a gas turns into a solid without becoming liquid. Frost and snow are formed this way.

Wind: Wind is the horizontal movement of air caused by differences in atmospheric pressure over the Earth's surface.

Turbulence: Turbulence is an irregular motion of the air resulting from the formation of vertical currents. Turbulence may exist in the atmosphere in the form of whirlwinds with variable dimensions.

Advection: Advection refers to the transport mechanism of a substance by a fluid due to the fluid's bulk motion. This process allows humidity and water transfers between the air and the ground or the sea surface, thereby modifying the structure and composition of the lower layers of the atmosphere.

Subsidence: Subsidence is the downward vertical motion of air due to the low temperatures. As air cools, it becomes denser and moves toward the ground.

Meteors: A meteor is the visible streak of light from a meteoroid or micrometeoroid, heated and glowing from entering the Earth's atmosphere. Millions of meteors occur in the Earth's atmosphere daily.

Fog and Mist: Fog is a collection of liquid water droplets or ice crystals suspended in the air at or near the Earth's surface. Fog is distinguished from mist only by its density: Fog reduces visibility to less than 1 km, whereas mist reduces visibility to no less than 1 km.

Precipitations: As condensation intensifies, the diameter of the droplets increases. When their fall speed increases, precipitation occurs either in the form of drizzle or rain, depending on the dimensions of the droplets.

Clouds: A cloud is a visible mass of liquid droplets or frozen crystals made of water or various chemicals suspended in the atmosphere above the Earth's surface formed by the condensation of the water vapor in the atmosphere.

Auroras: An aurora is a natural light display in the sky particularly in the high latitude (Arctic and Antarctic) regions, caused by the collision of energetic charged particles with atoms in the high altitude atmosphere (thermosphere).

Electromagnetic fields may have significant influence on atmospheric pressure, precipitation and temperature. All wireless communications systems give off electromagnetic radiations and thus have an effect on the weather systems. On the other hand, meteorological phenomena have great impact on microwave propagation. For example, water in any state (liquid, solid or gas) is an obstacle to electromagnetic wave, which will absorb and scatter its energy. As a result, the electromagnetic wave is attenuated. Accurate prediction of losses due to various meteorological phenomena and other factors is important to the design of the radio systems.

6.2 Wave Propagation in Atmosphere

The types of wave propagation in atmosphere largely depend on the frequency. Waves propagating in the ionosphere and close to the Earth's surface are respectively called ionospheric waves (or sky waves) and ground waves (or surface waves). The ionosphere contains charged particles and behaves like a conductor. The ionospheric waves strike the ionosphere at an angle θ, get refracted back to the ground from the ionosphere, strike the ground, and are then reflected back toward ionosphere, and so on. The boundary of the ionosphere and the Earth's ground forms a waveguide (Figure 6.3). The effects of the ionosphere depend on ion density, the frequency of the radio wave, and the transmission angle θ. Extremely low frequency (ELF, <3 kHz) and very low frequency (VLF, 3–30 kHz) signals

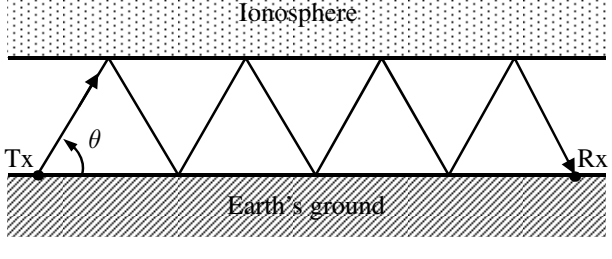


Figure 6.3 Earth–ionosphere waveguide.

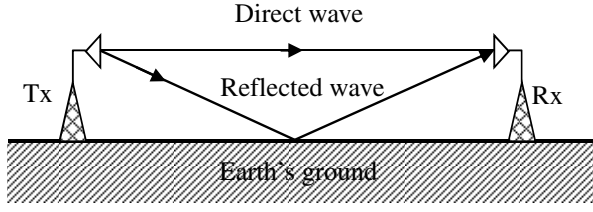


Figure 6.4 Direct wave and ground-reflected wave.

propagate efficiently in this waveguide. The effective height of the waveguide varies around the surface of the Earth due to the diurnal variations of the height of the ionospheric D-layer.

Low frequency (LF, 30–300 kHz) and medium frequency (MF, 300 kHz–3 MHz) signals can travel either as a ground wave or as an ionospheric wave but the former is the dominant mode. HF (3–30 MHz) signals often travel as an ionospheric wave but ground wave propagation is also possible in the HF band.

Very high frequency (VHF, 30–300 MHz) and ultra-high frequency (UHF, 300 MHz–3 GHz) signals only propagate as a ground wave, either via a direct wave path (line-of-sight transmission) or a reflected wave path (the wave strikes the Earth and then bounces off), as illustrated in Figure 6.4.

In the super high frequency (SHF) and extremely high frequency (EHF) band, propagation paths must include the line-of-sight path.

6.2.1 Propagation of Radio Waves over the Earth

The wave propagation is modified by the presence of the Earth and the atmosphere. The modification highly depends on the frequency of the wave, the directionality of the antenna as well as the proximity of the antenna close to the ground.

6.2.1.1 *A General Approach*

The determination of the characteristics of wave propagation over the Earth is important to the design of wireless communication systems. As a simplified model, the Earth surface is assumed to be a plane at $z = 0$. The Earth and the atmosphere are assumed to be homogeneous, respectively with medium parameters $\mu_1, \varepsilon_1, \sigma_1$ and $\mu_2, \varepsilon_2, \sigma_2$, as illustrated

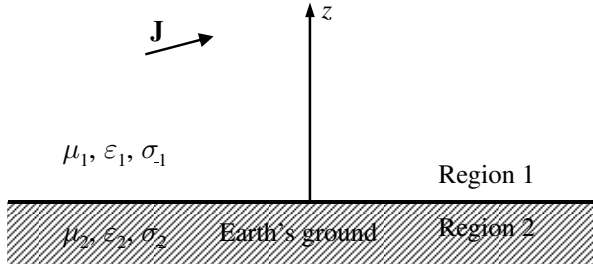


Figure 6.5 Wave propagation in homogeneous atmosphere.

in Figure 6.5. In a homogeneous medium, the electromagnetic fields can be expressed in terms of two scalar functions. In this case, we may choose the two scalar functions as the z-components of electric Hertz vector and magnetic Hertz vector

$$\mathbf{\Pi}_e = \Pi_e \mathbf{u}_z, \quad \mathbf{\Pi}_m = \Pi_m \mathbf{u}_z.$$

The electromagnetic fields can be expressed as

$$\begin{aligned}
\mathbf{E} &= \nabla \times \nabla \times \mathbf{\Pi}_e - j\omega\mu \nabla \times \mathbf{\Pi}_m \\
&= \mathbf{u}_x \left(\frac{\partial^2 \Pi_e}{\partial x \partial z} - j\omega\mu \frac{\partial \Pi_m}{\partial y} \right) + \mathbf{u}_y \left(\frac{\partial^2 \Pi_e}{\partial y \partial z} + j\omega\mu \frac{\partial \Pi_m}{\partial x} \right) \\
&\quad + \mathbf{u}_z \left(\frac{\partial^2 \Pi_e}{\partial z^2} + \omega^2 \mu \tilde{\varepsilon} \Pi_e \right), \\
\mathbf{H} &= \nabla \times \nabla \times \mathbf{\Pi}_m + j\omega\tilde{\varepsilon} \nabla \times \mathbf{\Pi}_e \\
&= \mathbf{u}_x \left(\frac{\partial^2 \Pi_m}{\partial x \partial z} + j\omega\tilde{\varepsilon} \frac{\partial \Pi_e}{\partial y} \right) + \mathbf{u}_y \left(\frac{\partial^2 \Pi_m}{\partial y \partial z} - j\omega\tilde{\varepsilon} \frac{\partial \Pi_e}{\partial x} \right) \\
&\quad + \mathbf{u}_z \left(\frac{\partial^2 \Pi_m}{\partial z^2} + \omega^2 \mu \tilde{\varepsilon} \Pi_m \right),
\end{aligned}$$

where $\tilde{\varepsilon} = \varepsilon + \sigma/j\omega$ denotes the complex permittivity. Assume that the incident source $\mathbf{J}$ is located in Region 1, which produces an incident field, denoted by Π_{e0}, Π_{m0}, when the medium in Region 1 occupies the whole space. The incident field induces a reflected field in Region 1, denoted by Π_{e1}, Π_{m1}, and a transmitted field in Region 2, denoted by Π_{e2}, Π_{m2}. The

tangential fields must be continuous at the interface $z = 0$, which leads to

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\frac{\partial \Pi_{e0}}{\partial z}+\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\frac{\partial \Pi_{e1}}{\partial z}=\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\frac{\partial \Pi_{e2}}{\partial z},$$

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\frac{\partial \Pi_{m0}}{\partial z}+\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\frac{\partial \Pi_{m1}}{\partial z}=\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\frac{\partial \Pi_{m2}}{\partial z},$$

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\mu_1(\Pi_{m0}+\Pi_{m1})=\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\mu_2\Pi_{m2},$$

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\tilde{\varepsilon}_1(\Pi_{e0}+\Pi_{e1})=\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\tilde{\varepsilon}_2\Pi_{e2}.$$

These relations can be satisfied by requiring

$$\begin{aligned}
\mu_1(\Pi_{m0}+\Pi_{m1})&=\mu_2\Pi_{m2},\\
\tilde{\varepsilon}_1(\Pi_{e0}+\Pi_{e1})&=\tilde{\varepsilon}_2\Pi_{e2},\\
\frac{\partial \Pi_{e0}}{\partial z}+\frac{\partial \Pi_{e1}}{\partial z}&=\frac{\partial \Pi_{e2}}{\partial z},\\
\frac{\partial \Pi_{m0}}{\partial z}+\frac{\partial \Pi_{m1}}{\partial z}&=\frac{\partial \Pi_{m2}}{\partial z}.
\end{aligned}\tag{6.1}$$

The potential functions Π_e and Π_m satisfy the Helmholtz equation

$$(\nabla^2+k^2)\Pi_{e,m}=0,\tag{6.2}$$

where $k=\omega\sqrt{\mu\tilde{\varepsilon}}$. The solution of the above equation can be found by using the Fourier transform. Define the Fourier transform and the inverse Fourier transform respectively as follows:

$$\begin{aligned}
\widetilde{\Pi}(\xi_1,\xi_2,z)&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Pi(x,y,z)e^{-j(\xi_1x+\xi_2y)}dx\,dy,\\
\Pi(x,y,z)&=\frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\widetilde{\Pi}(\xi_1,\xi_2,z)e^{j(\xi_1x+\xi_2y)}d\xi_1\,d\xi_2.
\end{aligned}\tag{6.3}$$

Then we have

$$\frac{d^2}{dz^2}\begin{pmatrix}\widetilde{\Pi}_{e1}\\ \widetilde{\Pi}_{m1}\end{pmatrix}+\beta_1^2\begin{pmatrix}\widetilde{\Pi}_{e1}\\ \widetilde{\Pi}_{m1}\end{pmatrix}=0,\quad \frac{d^2}{dz^2}\begin{pmatrix}\widetilde{\Pi}_{e2}\\ \widetilde{\Pi}_{m2}\end{pmatrix}+\beta_2^2\begin{pmatrix}\widetilde{\Pi}_{e2}\\ \widetilde{\Pi}_{m2}\end{pmatrix}=0,\tag{6.4}$$

where

$$\begin{aligned}
&\beta_1^2 = k_1^2 - \xi_1^2 - \xi_2^2, \quad \beta_2^2 = k_2^2 - \xi_1^2 - \xi_2^2,\\
&k_1^2 = \omega^2\mu_1\tilde{\varepsilon}_1, \quad k_2^2 = \omega^2\mu_2\tilde{\varepsilon}_2,\\
&\tilde{\varepsilon}_1 = \varepsilon_1 + \sigma_1/j\omega, \quad \tilde{\varepsilon}_2 = \varepsilon_2 + \sigma_2/j\omega.
\end{aligned}$$

According to the behavior of the fields at infinity, the solutions of (6.4) can be written as

$$\begin{aligned}
&\widetilde{\Pi}_{e1} = Ae^{-j\beta_1 z}, \quad \widetilde{\Pi}_{m1} = Be^{-j\beta_1 z},\\
&\widetilde{\Pi}_{e2} = Ce^{j\beta_2 z}, \quad \widetilde{\Pi}_{m2} = De^{j\beta_2 z},
\end{aligned} \tag{6.5}$$

where A, B, C, D are constants to be determined by the boundary conditions. Taking the Fourier transform of (6.1) gives

$$\begin{aligned}
\mu_1(\widetilde{\Pi}_{m0} + \widetilde{\Pi}_{m1}) &= \mu_2\widetilde{\Pi}_{m2},\\
\tilde{\varepsilon}_1(\widetilde{\Pi}_{e0} + \widetilde{\Pi}_{e1}) &= \tilde{\varepsilon}_2\widetilde{\Pi}_{e2},\\
\frac{\partial\widetilde{\Pi}_{e0}}{\partial z} + \frac{\partial\widetilde{\Pi}_{e1}}{\partial z} &= \frac{\partial\widetilde{\Pi}_{e2}}{\partial z},\\
\frac{\partial\widetilde{\Pi}_{m0}}{\partial z} + \frac{\partial\widetilde{\Pi}_{m1}}{\partial z} &= \frac{\partial\widetilde{\Pi}_{m2}}{\partial z}.
\end{aligned} \tag{6.6}$$

Introducing (6.5) into (6.6) yields

$$\begin{aligned}
\mu_1\Big[\widetilde{\Pi}_{m0}(0) + B\Big] &= \mu_2 D,\\
\tilde{\varepsilon}_1\Big[\widetilde{\Pi}_{e0}(0) + A\Big] &= \tilde{\varepsilon}_2 C,\\
\widetilde{\Pi}'_{e0}(0) - j\beta_1 A &= j\beta_2 C,\\
\widetilde{\Pi}'_{m0}(0) - j\beta_1 B &= j\beta_2 D,
\end{aligned} \tag{6.7}$$

where

$$\widetilde{\Pi}_{e0}(0) = \widetilde{\Pi}_{e0}\big|_{z=0}, \quad \widetilde{\Pi}_{m0}(0) = \widetilde{\Pi}_{m0}\big|_{z=0},$$

$$\widetilde{\Pi}'_{e0}(0) = \left.\frac{\partial\widetilde{\Pi}_{e0}}{\partial z}\right|_{z=0}, \quad \widetilde{\Pi}'_{m0}(0) = \left.\frac{\partial\widetilde{\Pi}_{m0}}{\partial z}\right|_{z=0}.$$

It follows from (6.7) that

$$\begin{aligned} A &= -\frac{j\tilde{\varepsilon}_2\widetilde{\Pi}'_{e0}(0) + \beta_2\tilde{\varepsilon}_1\widetilde{\Pi}_{e0}(0)}{\beta_2\tilde{\varepsilon}_1 + \beta_1\tilde{\varepsilon}_2}, \\ B &= -\frac{j\mu_2\widetilde{\Pi}'_{m0}(0) + \beta_2\mu_1\widetilde{\Pi}_{m0}(0)}{\beta_2\mu_1 + \beta_1\mu_2}, \\ C &= \frac{-j\tilde{\varepsilon}_1\widetilde{\Pi}'_{e0}(0) + \beta_1\tilde{\varepsilon}_1\widetilde{\Pi}_{e0}(0)}{\beta_2\tilde{\varepsilon}_1 + \beta_1\tilde{\varepsilon}_2}, \\ D &= \frac{-j\mu_1\widetilde{\Pi}'_{m0}(0) + \beta_1\mu_1\widetilde{\Pi}_{m0}(0)}{\beta_2\mu_1 + \beta_1\mu_2}. \end{aligned} \tag{6.8}$$

The potential functions can then be determined by

$$\begin{aligned} \Pi_{e1}(x,y,z) &= \frac{1}{4\pi^2}\int_{-\infty}^{\infty} e^{j\xi_1 x}d\xi_1 \int_{-\infty}^{\infty} Ae^{-j\beta_1 z + j\xi_2 y}d\xi_2, \\ \Pi_{m1}(x,y,z) &= \frac{1}{4\pi^2}\int_{-\infty}^{\infty} e^{j\xi_1 x}d\xi_1 \int_{-\infty}^{\infty} Be^{-j\beta_1 z + j\xi_2 y}d\xi_2, \\ \Pi_{e2}(x,y,z) &= \frac{1}{4\pi^2}\int_{-\infty}^{\infty} e^{j\xi_1 x}d\xi_1 \int_{-\infty}^{\infty} Ce^{j\beta_2 z + j\xi_2 y}d\xi_2, \\ \Pi_{m2}(x,y,z) &= \frac{1}{4\pi^2}\int_{-\infty}^{\infty} e^{j\xi_1 x}d\xi_1 \int_{-\infty}^{\infty} De^{j\beta_2 z + j\xi_2 y}d\xi_2. \end{aligned} \tag{6.9}$$

The above formulae are applicable to any incident field. Various techniques can be introduced to find the approximate solutions of (6.9) (Jones, 1964).

6.2.1.2 *Vertical Current Element over the Earth*

As a special case, we now consider the radiation of a dipole over the earth. This problem has been investigated by a number of authors (e.g., Sommerfeld, 1949; Norton, 1936; 1937). We assume that the earth is characterized by a complex dielectric constant $\tilde{\varepsilon} = \tilde{\varepsilon}_r\varepsilon_0$ with $\tilde{\varepsilon}_r = \varepsilon_r - j\frac{\sigma}{\omega\varepsilon_0}$. A z-directed current element of strength Il is placed at a height h above the surface of the earth, as illustrated in Figure 6.6, and is represented by

$$\mathbf{J}(\mathbf{r}) = \mathbf{u}_z J = \mathbf{u}_z Il\delta(x)\delta(y)\delta(z-h). \tag{6.10}$$

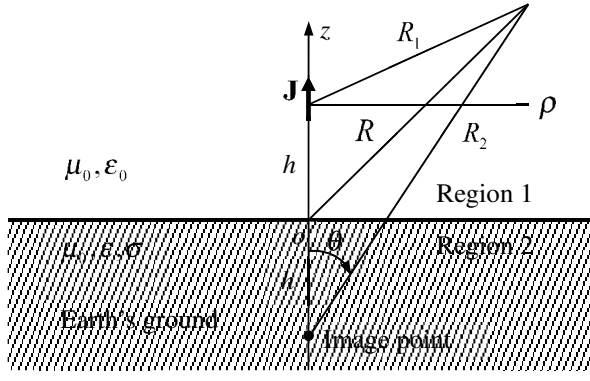


Figure 6.6 A vertical current above a flat earth.

The electric Hertz potential Π_{e0} in free space satisfies

$$(\nabla^2 + k_0^2)\Pi_{e0} = -\frac{\eta_0 Il}{jk_0}\delta(x)\delta(y)\delta(z-h), \tag{6.11}$$

where $k_0^2 = \omega^2\mu_0\varepsilon_0$, $\eta_0 = \sqrt{\mu_0/\varepsilon_0}$.

Applying the Fourier transform (6.3) to (6.11) yields

$$\frac{d^2}{dz^2}\widetilde{\Pi}_{e0} + \beta_1^2\widetilde{\Pi}_{e0} = -\frac{\eta_0 Il}{jk_0}\delta(z-h), \tag{6.12}$$

where $\beta_1^2 = k_0^2 - \xi_1^2 - \xi_2^2$. The solution of the above equation is given by (see Chapter 1)

$$\widetilde{\Pi}_{e0}(z) = \frac{\eta_0 Il}{jk_0}\frac{1}{j2\beta_1}e^{-j\beta_1|z-h|}. \tag{6.13}$$

The inverse Fourier transform is

$$\Pi_{e0}(x,y,z) = \frac{\eta_0 Il}{jk_0}\frac{e^{-jk_0R_1}}{4\pi R_1}, \tag{6.14}$$

where $R_1 = \sqrt{x^2+y^2+(z-h)^2}$ denote the distance between the dipole and the observation point. Equations (6.13) and (6.14) imply

$$\frac{e^{-jk_0R_1}}{4\pi R_1} = \frac{1}{4\pi^2}\int_{-\infty}^{\infty} e^{j\xi_1 x}d\xi_1 \int_{-\infty}^{\infty}\frac{1}{j2\beta_1}e^{-j\beta_1|z-h|}e^{j\xi_2 y}d\xi_2. \tag{6.15}$$

It follows from (6.13) that

$$\widetilde{\Pi}_{e0}(0) = \frac{1}{j\beta_1}\frac{1}{2}\frac{\eta_0 Il}{jk_0}e^{-j\beta_1 h}, \quad \widetilde{\Pi}'_{e0}(0) = \frac{1}{2}\frac{\eta_0 Il}{jk_0}e^{-j\beta_1 h}. \tag{6.16}$$

The constants A and C in (6.8) can be determined as follows

$$A = \frac{1}{j2\beta_1}\frac{\beta_1\tilde{\varepsilon} - \beta_2\varepsilon_0}{\beta_1\tilde{\varepsilon} + \beta_2\varepsilon_0}\frac{\eta_0 Il}{jk_0}e^{-j\beta_1 h},$$

$$C = \frac{-j\varepsilon_0}{\beta_2\varepsilon_0 + \beta_1\tilde{\varepsilon}}\frac{\eta_0 Il}{jk_0}e^{-j\beta_1 h}.$$

The reflected and transmitted fields Π_{e1} and Π_{e2} in (6.9) can be written as

$$\Pi_{e1}(x,y,z) = \frac{\eta_0 Il}{jk_0}\frac{1}{4\pi^2}\int_{-\infty}^{\infty} e^{j\xi_1 x}d\xi_1 \int_{-\infty}^{\infty}\frac{1}{j2\beta_1}\frac{\beta_1\tilde{\varepsilon} - \beta_2\varepsilon_0}{\beta_1\tilde{\varepsilon} + \beta_2\varepsilon_0}e^{-j\beta_1(z+h)+j\xi_2 y}d\xi_2, \tag{6.17}$$

$$\Pi_{e2}(x,y,z) = \frac{\eta_0 Il}{jk_0}\frac{1}{4\pi^2}\int_{-\infty}^{\infty} e^{j\xi_1 x}d\xi_1 \int_{-\infty}^{\infty}\frac{-j\varepsilon_0}{\beta_2\varepsilon_0 + \beta_1\tilde{\varepsilon}}e^{j\beta_2 z - j\beta_1 h + j\xi_2 y}d\xi_2. \tag{6.18}$$

By use of (6.15), (6.17) may be written as

$$\begin{aligned}\Pi_{e1}(x,y,z) &= -\frac{\eta_0 Il}{jk_0}\frac{1}{4\pi^2}\int_{-\infty}^{\infty} e^{j\xi_1 x}d\xi_1 \\ &\quad \times \int_{-\infty}^{\infty}\frac{1}{j2\beta_1}\frac{\beta_2\varepsilon_0 + \beta_1\tilde{\varepsilon} - 2\beta_1\tilde{\varepsilon}}{\beta_1\tilde{\varepsilon} + \beta_2\varepsilon_0}e^{-j\beta_1(z+h)+j\xi_2 y}d\xi_2 \\ &= -\frac{\eta_0 Il}{jk_0}\frac{e^{-jk_0 R_2}}{4\pi R_2} + \frac{\eta_0 Il}{jk_0}\frac{\tilde{\varepsilon}}{j4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{e^{j\Phi}}{\beta_1\tilde{\varepsilon} + \beta_2\varepsilon_0}d\xi_1\, d\xi_2,\end{aligned} \tag{6.19}$$

where $R_2 = \sqrt{x^2 + y^2 + (z+h)^2}$ is the distance between the image point and the observation point and $\Phi = \xi_1 x + \xi_2 y - \beta_1(z+h)$. The first term on the right-hand side of (6.19) represents the contribution from the image of the dipole.

Remark 6.1: Introducing the coordinate transformations

$$\begin{cases} \xi_1 = \tau\cos\psi \\ \xi_2 = \tau\sin\psi \end{cases}, \quad \begin{cases} x = \rho\cos\varphi \\ y = \rho\sin\varphi \end{cases},$$

we have

$$\beta_1 = \sqrt{k_0^2 - \tau^2}, \quad \beta_2 = \sqrt{k_2^2 - \tau^2}, \quad \xi_1 x + \xi_2 y = \tau\rho\cos(\psi - \varphi).$$

On account of the relation $\int_0^{2\pi} e^{j\tau\rho\cos(\psi-\varphi)}d\psi = 2\pi J_0(\tau\rho)$, (6.19) becomes

$$\Pi_{e1}(x,y,z) = -\frac{\eta_0 Il}{jk_0}\frac{e^{-jk_0R_2}}{4\pi R_2} + \frac{\eta_0 Il}{jk_0}\frac{\tilde{\varepsilon}}{j2\pi}\int_0^\infty \frac{e^{-j\beta_1(z+h)}}{\beta_1\tilde{\varepsilon} + \beta_2\varepsilon_0} J_0(\tau\rho)\tau\, d\tau. \qquad \square$$

In the spherical coordinate system (R_2, θ, φ) with image of the dipole as origin (Figure 6.6), we have

$$x = R_2\sin\theta\cos\varphi, \quad y = R_2\sin\theta\sin\varphi, \quad z + h = R_2\cos\theta.$$

Hence

$$\Phi = R_2(\xi_1\sin\theta\cos\varphi + \xi_2\sin\theta\sin\varphi - \beta_1\cos\theta).$$

The integral in (6.19) can be evaluated by the method of stationary phase (e.g., Jones, 1964). The stationary point can be determined by requiring

$$\frac{\partial\Phi}{\partial\xi_1} = \frac{\partial\Phi}{\partial\xi_2} = 0,$$

and is found to be

$$\xi_{1s} = -k_0\sin\theta\cos\varphi, \quad \xi_{2s} = -k_0\sin\theta\sin\varphi.$$

In the neighborhood of the stationary point, we may let

$$\xi_1 = \xi_{1s} + u, \quad \xi_2 = \xi_{2s} + v.$$

Then

$$\frac{1}{R_2}\Phi(u,v) = (\xi_{1s} + u)\sin\theta\cos\varphi + (\xi_{2s} + v)\sin\theta\sin\varphi - \sqrt{k_0^2 - (\xi_{1s} + u)^2 - (\xi_{2s} + v)^2}\cos\theta,$$

with

$$\frac{1}{R_2}\Phi(0,0) = -k_0,$$

$$\frac{1}{R_2}\frac{\partial\Phi^2(0,0)}{\partial u^2} = \frac{1 - \sin^2\theta\sin^2\varphi}{k_0\cos^2\theta},$$

$$\frac{1}{R_2}\frac{\partial\Phi^2(0,0)}{\partial v^2}=\frac{1-\sin^2\theta\cos^2\varphi}{k_0\cos^2\theta},$$
$$\frac{1}{R_2}\frac{\partial\Phi^2(0,0)}{\partial u\partial v}=\frac{\sin^2\theta\sin\varphi\cos\varphi}{k_0\cos^2\theta}.$$

Using Taylor's formula, we have the following approximation

$$\begin{aligned}\Phi(u,v)&\approx\Phi(0,0)+\frac{1}{2}\frac{\partial\Phi^2(0,0)}{\partial u^2}u^2+\frac{1}{2}\frac{\partial\Phi^2(0,0)}{\partial v^2}v^2+\frac{\partial\Phi^2(0,0)}{\partial u\partial v}uv\\&=-k_0R_2+\frac{R_2}{2k_0\cos^2\theta}\left[u^2(1-\sin^2\theta\sin^2\varphi)+v^2(1-\sin^2\theta\cos^2\varphi)\right.\\&\quad\left.+2uv\sin^2\theta\sin\varphi\cos\varphi\right].\end{aligned}$$

Introducing the rotation of coordinates

$$\begin{cases}u=u_1\cos\varphi-v_1\sin\varphi\\v=u_1\sin\varphi+v_1\cos\varphi\end{cases},$$

we obtain

$$\Phi(u_1,v_1)=-k_0R_2+\frac{R_2}{2k_0\cos^2\theta}\left(u_1^2+v_1^2\cos^2\theta\right).$$

Substituting this into (6.19) yields

$$\begin{aligned}\Pi_{e1}(x,y,z)&=-\frac{\eta_0 Il}{jk_0}\frac{e^{-jk_0R_2}}{4\pi R_2}+\frac{\eta_0 Il}{jk_0}\frac{1}{j4\pi^2}\frac{\tilde{\varepsilon}}{\beta_1\tilde{\varepsilon}+\beta_2\varepsilon_0}\\&\quad\times\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left[jR_2\left(-k_0+\frac{u_1^2+v_1^2\cos^2\theta}{2k_0\cos^2\theta}\right)\right]du_1dv_1,\quad(6.20)\end{aligned}$$

where β_1 and β_2 are assumed to take the values at the stationary point

$$\beta_1=\sqrt{k_0^2-\xi_{1s}^2-\xi_{2s}^2}=k_0\cos\theta,$$
$$\beta_2=k_2^2-\xi_{1s}^2-\xi_{2s}^2=k_0\sqrt{\tilde{\varepsilon}_r-\sin^2\theta}.$$

Making use of the relation $\int_{-\infty}^{\infty}e^{jax^2}dx=\sqrt{\frac{\pi}{a}}e^{j\pi/4}$, we may find that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left(jR_2\frac{u_1^2+v_1^2\cos^2\theta}{2k_0\cos^2\theta}\right)du_1dv_1=j\frac{2\pi k_0\cos\theta}{R_2}.$$

Thus (6.20) can be rewritten as

$$\Pi_{e1}(x,y,z) = -\frac{\eta_0 Il}{jk_0}\frac{e^{-jk_0R_2}}{4\pi R_2} + \frac{\eta_0 Il}{jk_0}\frac{e^{-jk_0R_2}}{2\pi R_2}\frac{\tilde{\varepsilon}k_0\cos\theta}{\beta_1\tilde{\varepsilon}+\beta_2\varepsilon_0}$$

$$= \frac{\eta_0 Il}{jk_0}\Gamma_v\frac{e^{-jk_0R_2}}{4\pi R_2}, \tag{6.21}$$

where

$$\Gamma_v = \frac{\tilde{\varepsilon}_r\cos\theta - \sqrt{\tilde{\varepsilon}_r - \sin^2\theta}}{\tilde{\varepsilon}_r\cos\theta + \sqrt{\tilde{\varepsilon}_r - \sin^2\theta}}$$

is the Fresnel reflection coefficient for a plane wave incident at the angle θ when the electric field is in the plane of incidence.

6.2.1.3 *Two-Ray Propagation Model*

In a spherical coordinate system (r,θ,φ), the far fields produced by a transmitting antenna in a homogeneous medium with medium parameters μ and ε can be expressed as

$$\mathbf{E}(\mathbf{r}) = \frac{e^{-jkr}}{r}\mathbf{E}_\infty(\mathbf{u}_r), \quad \mathbf{H}(\mathbf{r}) = \frac{e^{-jkr}}{r}\mathbf{H}_\infty(\mathbf{u}_r), \tag{6.22}$$

where $\mathbf{u}_r$ is the unit vector along the radial direction; $k = \omega\sqrt{\mu\varepsilon}$; $\mathbf{E}_\infty$ and $\mathbf{H}_\infty$ are the electric far-field pattern and magnetic far-field pattern respectively. The Poynting vector in the far-field region may be written as

$$\mathbf{S}(\mathbf{r}) = \frac{1}{2\eta}\frac{|\mathbf{E}_\infty(\mathbf{u}_r)|^2}{r^2}\mathbf{u}_r, \tag{6.23}$$

where $\eta = \sqrt{\mu/\varepsilon}$. The total radiated power from the transmitting antenna is then given by the integration of the Poynting vector over a sphere S of radius r

$$P_{\text{rad}} = \int_0^{2\pi}\int_0^{\pi}\mathbf{S}(\mathbf{r})\cdot\mathbf{u}_r r^2\sin\theta\,d\theta\,d\varphi = \frac{1}{2\eta}\int_0^{2\pi}\int_0^{\pi}|\mathbf{E}_\infty(\mathbf{u}_r)|^2\sin\theta\,d\theta\,d\varphi. \tag{6.24}$$

The directivity of the transmitting antenna can be expressed by

$$D(\theta,\varphi) = 4\pi\frac{r^2}{2\eta}\frac{|\mathbf{E}(\mathbf{r})|^2}{P_{\text{rad}}} = 4\pi\frac{|\mathbf{E}_\infty(\mathbf{u}_r)|^2}{\int_0^{2\pi}\int_0^{\pi}|\mathbf{E}_\infty(\mathbf{u}_r)|^2\sin\theta\,d\theta\,d\varphi}. \tag{6.25}$$

We may introduce the **normalized far-field pattern**, defined as the ratio of $\mathbf{E}_\infty(\mathbf{u}_r)$ to the magnitude of $\mathbf{E}_\infty(\mathbf{u}_r)$ in the direction of maximum transmission

$$\hat{\mathbf{E}}_\infty(\mathbf{u}_r) = \frac{\mathbf{E}_\infty(\mathbf{u}_r)}{\max\limits_{\mathbf{u}_r} |\mathbf{E}_\infty(\mathbf{u}_r)|}. \tag{6.26}$$

The directivity can then be rewritten as

$$D(\theta, \varphi) = 4\pi \frac{|\hat{\mathbf{E}}_\infty(\mathbf{u}_r)|^2}{\int\limits_0^{2\pi}\int\limits_0^{\pi} |\hat{\mathbf{E}}_\infty(\mathbf{u}_r)|^2 \sin\theta \, d\theta \, d\varphi}. \tag{6.27}$$

The directivity in the direction of maximum radiation is then given by

$$D_0 = \frac{4\pi}{\int\limits_0^{2\pi}\int\limits_0^{\pi} |\hat{\mathbf{E}}_\infty(\mathbf{u}_r)|^2 \sin\theta \, d\theta \, d\varphi}. \tag{6.28}$$

It follows from (6.24), (6.26) and (6.28) that

$$\max_{\mathbf{u}_r} |\mathbf{E}_\infty(\mathbf{u}_r)| = \sqrt{\frac{\eta P_{\text{rad}} D_0}{2\pi}}. \tag{6.29}$$

Therefore, we may write

$$\mathbf{E}(\mathbf{r}) = \sqrt{\frac{\eta P_{\text{rad}} D_0}{2\pi}} \frac{e^{-jkr}}{r} \hat{\mathbf{E}}_\infty(\mathbf{u}_r), \tag{6.30}$$

$$\mathbf{S}(\mathbf{r}) = \frac{P_{\text{rad}} D_0}{4\pi r^2} |\hat{\mathbf{E}}_\infty(\mathbf{u}_r)|^2 \mathbf{u}_r, \tag{6.31}$$

$$D(\theta, \varphi) = D_0 |\hat{\mathbf{E}}_\infty(\mathbf{u}_r)|^2. \tag{6.32}$$

We will use the subscripts t and r to denote the quantities related to the transmitting and receiving antenna respectively. The received power P_{rec} by a distant receiving antenna at $\mathbf{r} = (R, \theta_t, \varphi_t)$ can be expressed by (see Figure 6.7)

$$P_{\text{rec}} = |\mathbf{S}_t(R, \theta_t, \varphi_t)| A_e(\theta_r, \varphi_r), \tag{6.33}$$

where A_e is the equivalent area of the receiving antenna. If the receiving antenna is conjugately matched and there is no polarization loss, the equivalent area is given by

$$A_e(\theta_r, \varphi_r) = \frac{\lambda^2}{4\pi} D_r(\theta_r, \varphi_r).$$

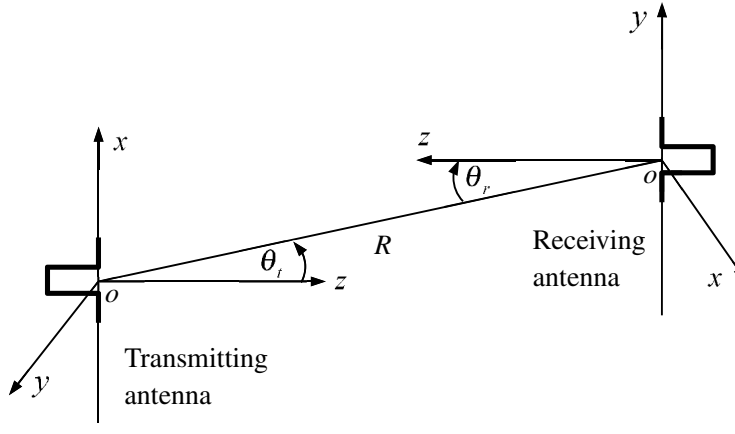


Figure 6.7 Wave propagation in free space.

Thus, we have

$$P_{\text{rec}} = |\mathbf{S}(R, \theta_t, \varphi_t)| \frac{\lambda^2}{4\pi} D_r(\theta_r, \varphi_r).$$

Making use of (6.31) and (6.32), we obtain the fundamental equation for power transmission between two antennas in free space

$$\frac{P_{\text{rec}}}{P_{\text{rad}}} = \left(\frac{\lambda}{4\pi R}\right)^2 D_{t0} D_{r0} |\hat{\mathbf{E}}_{t\infty}(\theta_t, \varphi_t)|^2 |\hat{\mathbf{E}}_{r\infty}(\theta_r, \varphi_r)|^2. \tag{6.34}$$

The above equation is the well-known Friis free-space propagation model, and has been derived in Chapter 5. It indicates that the received power falls off as the square of the separation distance between the transmitter and receiver (or decays with separation distance at a rate of 20 dB/decade). The path loss for free-space propagation model, denoted as PL, is defined as the signal attenuation measured in dB

$$PL(\text{dB}) = -10 \log \frac{P_{\text{rec}}}{P_{\text{rad}}} = -10 \log \left[\left(\frac{\lambda}{4\pi R}\right)^2 D_t D_r \right]. \tag{6.35}$$

The minus sign ensures that the path loss is a positive quantity.

The free-space propagation model (6.34) only contains a single direct path between the transmitting antenna and receiving antenna, and therefore is inaccurate for most applications related to the wave propagation near the surface of the Earth. The presence of the Earth complicates the

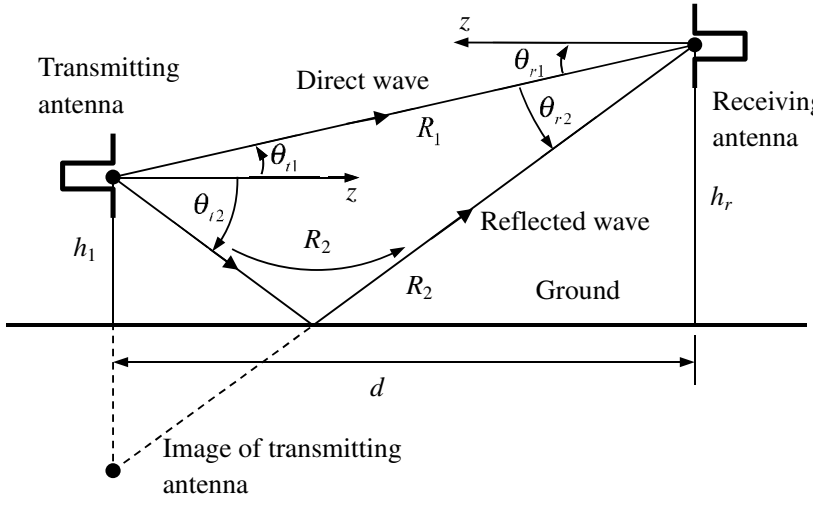


Figure 6.8 Two-ray ground reflection model.

situation in a number of ways. A two-ray ground reflection model based on geometric optics is illustrated in Figure 6.8, where a transmitting antenna and a receiving antenna are situated above a flat ground (earth) at height h_t and h_r respectively with separation d. This model is found useful in predicting the large-scale signal strength. The fields emanated from the transmitting antenna take a direct path of length R_1 and an indirect or reflected path of length R_2. Depending on the phase difference between the two paths, the two rays sum at the receiving antenna and may produce either constructive or destructive interference.

The radiating field at the receiving antenna along the direct path can be written as

$$\mathbf{E}_{td}(R_1, \theta_{t1}, \varphi_{t1}) = \sqrt{\frac{\eta P_{\text{rad}} D_{t0}}{2\pi}} \frac{e^{-jkR_1}}{R_1} \hat{\mathbf{E}}_{t\infty}(\theta_{t1}, \varphi_{t1}).$$

If the antenna heights are small compared with the separation d, the angles $\theta_{t1}, \theta_{t2}, \theta_{r1}$ and θ_{r2} are very small. In this case, we can use the approximations $\hat{\mathbf{E}}_{t\infty}(\theta_{t1}, \varphi_{t1}) = \hat{\mathbf{E}}_{t\infty}(\theta_{t2}, \varphi_{t1})$. The radiating field at the receiving antenna along the indirect path can be approximated by

$$\mathbf{E}_{tr}(R_2, \theta_{t1}, \varphi_{t1}) = \Gamma \sqrt{\frac{\eta P_{\text{rad}} D_{t0}}{2\pi}} \frac{e^{-jkR_2}}{R_2} \hat{\mathbf{E}}_{t\infty}(\theta_{t1}, \varphi_{t1}),$$

where Γ is the reflection coefficient at the ground. The total field at the receiving antenna is then given by

$$\begin{aligned}\mathbf{E}_t(R_1,\theta_{t1},\varphi_{t1}) &= \mathbf{E}_{td}(R_1,\theta_{t1},\varphi_{t1}) + \mathbf{E}_{tr}(R_2,\theta_{t1},\varphi_{t1}) \\ &= \sqrt{\frac{\eta P_{\rm rad} D_{t0}}{2\pi}}\frac{e^{-jkR_1}}{R_1}\hat{\mathbf{E}}_{t\infty}(\theta_{t1},\varphi_{t1})\left[1+\frac{R_1}{R_2}\Gamma e^{-jk(R_2-R_1)}\right] \\ &\approx \mathbf{E}_{td}(R_1,\theta_{t1},\varphi_{t1})\left[1+\Gamma e^{-jk(R_2-R_1)}\right], \end{aligned} \tag{6.36}$$

where we have made the approximation $R_1/R_2 \approx 1$. According to (6.32) and (6.33), the power received by the receiving antenna is

$$\begin{aligned} P_{\rm rec} &= \frac{R_1^2}{2\eta}\left|\mathbf{E}_t(R_1,\theta_{t1},\varphi_{t1})\right|^2 \frac{\lambda^2}{4\pi} D_{r0}\left|\hat{\mathbf{E}}_{r\infty}(\theta_{r1},\varphi_{r1})\right|^2 \\ &= P_{\rm rad} D_{t0} D_{r0}\left(\frac{\lambda}{4\pi}\right)^2 \left|\hat{\mathbf{E}}_{t\infty}(\theta_{t1},\varphi_{t1})\right|^2 \\ &\quad \times \left|\hat{\mathbf{E}}_{r\infty}(\theta_{r1},\varphi_{r1})\right|^2 \left|1+\Gamma e^{-jk(R_2-R_1)}\right|^2. \end{aligned} \tag{6.37}$$

Note that

$$\begin{aligned} R_1 &= \sqrt{d^2+(h_r-h_t)^2} \approx d + \frac{1}{2d}(h_r-h_t)^2, \\ R_2 &= \sqrt{d^2+(h_r+h_t)^2} \approx d + \frac{1}{2d}(h_r+h_t)^2. \end{aligned}$$

Thus, we have $R_2 - R_1 \approx \frac{2h_t h_r}{d}$ and (6.37) can be approximated by

$$\begin{aligned} \frac{P_{\rm rec}}{P_{\rm rad}} &= \left(\frac{\lambda}{4\pi}\right)^2 D_{t0} D_{r0}\left|\hat{\mathbf{E}}_{t\infty}(\theta_{t1},\varphi_{t1})\right|^2 \left|\hat{\mathbf{E}}_{r\infty}(\theta_{r1},\varphi_{r1})\right|^2 \\ &\quad \times \left|1+\Gamma\exp\left(-jk\frac{2h_t h_r}{d}\right)\right|^2. \end{aligned} \tag{6.38}$$

The above formula represents the loss due to the plane earth and is valid for long distances.

6.2.2 Wave Propagation in Atmosphere: Ray-Tracing Method

In studying the atmospheric refraction, it is necessary to assume that the properties of the atmosphere are inhomogeneous and vary with the height, which is called **stratified atmosphere**. The time-harmonic Maxwell equations in an isotropic inhomogeneous medium take the form

$$\begin{aligned} \nabla \times \mathbf{H}(\mathbf{r}) &= j\omega\varepsilon(\mathbf{r})\mathbf{E}(\mathbf{r}), \\ \nabla \times \mathbf{E}(\mathbf{r}) &= -j\omega\mu(\mathbf{r})\mathbf{H}(\mathbf{r}), \\ \nabla \cdot [\varepsilon(\mathbf{r})\mathbf{E}(\mathbf{r})] = 0, &\quad \nabla \cdot [\varepsilon(\mathbf{r})\mathbf{H}(\mathbf{r})] = 0. \end{aligned} \tag{6.39}$$

The **refractive index** n of the medium is defined by $n = \sqrt{\mu\varepsilon/\mu_0\varepsilon_0}$, where μ_0 and ε_0 are the permeability and permittivity in free space. The wavenumber in free space will be denoted by $k_0 = \omega\sqrt{\mu_0\varepsilon_0}$. We assume that

$$\mathbf{E} = \mathbf{E}_0(\mathbf{r})e^{-jk_0L(\mathbf{r})}, \quad \mathbf{H} = \mathbf{H}_0(\mathbf{r})e^{-jk_0L(\mathbf{r})}. \tag{6.40}$$

The function $L(\mathbf{r})$ is known as **eikonal**. The **wavefronts** are defined as the surfaces of constant phase: $L(\mathbf{r}) = \text{const}$. Substituting (6.40) into (6.39), we obtain

$$\begin{aligned} &\mathbf{H}_0(\mathbf{r}) \times \nabla L(\mathbf{r}) - \frac{\omega\varepsilon(\mathbf{r})}{k_0}\mathbf{E}_0(\mathbf{r}) = j\frac{1}{k_0}\nabla \times \mathbf{H}_0(\mathbf{r}), \\ &\mathbf{E}_0(\mathbf{r}) \times \nabla L(\mathbf{r}) + \frac{\omega\mu(\mathbf{r})}{k_0}\mathbf{H}_0(\mathbf{r}) = j\frac{1}{k_0}\nabla \times \mathbf{E}_0(\mathbf{r}), \\ &\mathbf{E}_0(\mathbf{r}) \cdot \nabla L(\mathbf{r}) = \frac{1}{jk_0}[\mathbf{E}_0(\mathbf{r}) \cdot \nabla \ln\varepsilon(\mathbf{r}) + \nabla \cdot \mathbf{E}_0(\mathbf{r})], \\ &\mathbf{H}_0(\mathbf{r}) \cdot \nabla L(\mathbf{r}) = \frac{1}{jk_0}[\mathbf{H}_0(\mathbf{r}) \cdot \nabla \ln\mu(\mathbf{r}) + \nabla \cdot \mathbf{H}_0(\mathbf{r})]. \end{aligned} \tag{6.41}$$

If the frequency is very high, k_0 becomes very large and the right-hand side of (6.41) can be equated to zero. There results

$$\begin{aligned} &\mathbf{H}_0(\mathbf{r}) \times \nabla L(\mathbf{r}) - \frac{\omega\varepsilon(\mathbf{r})}{k_0}\mathbf{E}_0(\mathbf{r}) = 0, \\ &\mathbf{E}_0(\mathbf{r}) \times \nabla L(\mathbf{r}) + \frac{\omega\mu(\mathbf{r})}{k_0}\mathbf{H}_0(\mathbf{r}) = 0, \end{aligned}$$

$$\begin{aligned}\mathbf{E}_0(\mathbf{r}) \cdot \nabla L(\mathbf{r}) &= 0,\\ \mathbf{H}_0(\mathbf{r}) \cdot \nabla L(\mathbf{r}) &= 0.\end{aligned} \tag{6.42}$$

The last two equations show that $\mathbf{E}_0$ and $\mathbf{H}_0$ are transverse to ∇L, i.e., transverse to the direction of propagation of the wavefront. From the first two equations of (6.42), it is easy to see that $\mathbf{E}_0 \cdot \mathbf{H}_0 = 0$. Therefore, the field is locally a plane wave. If $\mathbf{H}_0(\mathbf{r})$ is eliminated from the first two equations of (6.42), then

$$n^2(\mathbf{r})\mathbf{E}_0(\mathbf{r}) + [\nabla L(\mathbf{r}) \cdot \mathbf{E}_0(\mathbf{r})]\nabla L(\mathbf{r}) - [\nabla L(\mathbf{r})]^2\mathbf{E}_0(\mathbf{r}) = 0.$$

The second term is zero due to the third equation of (6.42). Thus if $\mathbf{E}_0$ is not identically zero it is necessary that

$$[\nabla L(\mathbf{r})]^2 = n^2(\mathbf{r}). \tag{6.43}$$

This is called **eikonal equation**.

Making use of the second equation of (6.42), the Poynting vector may be written as

$$\frac{1}{2}\mathrm{Re}(\mathbf{E} \times \bar{\mathbf{H}}) = \frac{1}{2}\mathrm{Re}\frac{k_0}{\omega\mu}|\mathbf{E}_0(\mathbf{r})|^2\nabla\bar{L}(\mathbf{r})e^{-jk_0(L-\bar{L})}.$$

For real $L(\mathbf{r})$, we have

$$\frac{1}{2}\mathrm{Re}(\mathbf{E} \times \bar{\mathbf{H}}) = \frac{1}{2}\frac{k_0}{\omega\mu}|\mathbf{E}_0(\mathbf{r})|^2\nabla L(\mathbf{r}).$$

So the direction of energy flow is normal to the wavefront. The curves whose tangent at each point is the direction of energy flow of the field are known as **rays**. In optics, the rays are used to model the propagation of light through an optical system, by representing the light field in terms of discrete rays. The ray optics can be used to study light reflections and refractions. Since the rays are normal to the wavefront, we may introduce a unit tangent vector to the rays

$$\mathbf{s}(\mathbf{r}) = \frac{1}{n(\mathbf{r})}\nabla L(\mathbf{r}). \tag{6.44}$$

This implies

$$\nabla L(\mathbf{r}) \cdot \nabla = n(\mathbf{r})\mathbf{s}(\mathbf{r}) \cdot \nabla = n(\mathbf{r})\frac{\partial}{\partial s}. \tag{6.45}$$

Let $\mathbf{r}$ be a point P on a ray and s be the arc length measured along the ray. Then $d\mathbf{r}/ds = \mathbf{s}$, and

$$n(\mathbf{r})\frac{d\mathbf{r}}{ds} = \nabla L(\mathbf{r}).$$

Taking the derivative with respect to s and making use of the relation $\frac{d}{ds} = \frac{d\mathbf{r}}{ds} \cdot \nabla$, we obtain

$$\frac{d}{ds}n(\mathbf{r})\frac{d\mathbf{r}}{ds} = \nabla n(\mathbf{r}). \tag{6.46}$$

This is the differential equation for the rays, called **ray equation**, which can be solved numerically with initial data to determine the rays in a region. The refractive index of the atmosphere affects the curvature of the electromagnetic wave path and gives some insight into the fading phenomenon.

Example 6.1: To apply the above theory to the stratified atmosphere, we assume that the surface of the earth may be treated as a plane and introduce a cylindrical coordinate system (ρ, φ, z) such that the z-axis is in the vertical direction and the plane $z = 0$ coincides with the surface of the earth. A point source is assumed to be located on the z-axis at height z_1. In terms of the symmetry, all quantities are independent of φ. Thus the rays are curves lying in planes passing through the z-axis. So we only need to consider one of these planes, say (x, z)-plane as illustrated in Figure 6.9.

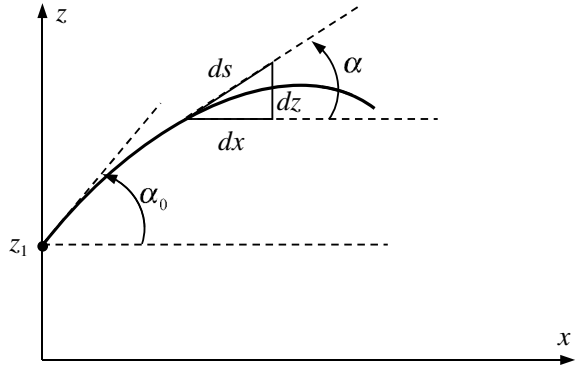


Figure 6.9 A ray for a stratified atmosphere.

We may write (6.46) as

$$\begin{aligned} \frac{d}{ds}n(z)\frac{dx}{ds} &= 0, \\ \frac{d}{ds}n(z)\frac{dz}{ds} &= \frac{dn(z)}{dz}. \end{aligned} \tag{6.47}$$

It follows from the first equation that

$$n(z)\frac{dx}{ds} = C, \tag{6.48}$$

where C is a constant characteristic of the ray and can be determined by (Figure 6.9)

$$n(z_1)\cos\alpha_0 = C. \tag{6.49}$$

Since the ray is confined in (x, z)-plane, we have

$$\frac{d\mathbf{r}}{ds}\cdot\frac{d\mathbf{r}}{ds} = \left(\frac{dx}{ds}\right)^2 + \left(\frac{dz}{ds}\right)^2 = 1.$$

It follows that

$$\frac{dz}{ds} = \frac{\pm\sqrt{n^2(z) - C^2}}{n(z)}. \tag{6.50}$$

Combining (6.48) and (6.50) yields

$$\frac{dx}{dz} = \frac{C}{\pm\sqrt{n^2(z) - C^2}}. \tag{6.51}$$

This can be used to determine the ray family. □

The behavior of the magnitude $\mathbf{E}_0$ can be determined by the Maxwell equations. Introducing (6.40) into the wave equation

$$\nabla\times\mu^{-1}\nabla\times\mathbf{E}(\mathbf{r}) - \omega^2\varepsilon\mathbf{E}(\mathbf{r}) = 0 \tag{6.52}$$

yields

$$\begin{aligned} &\frac{1}{jk_0}\left[(\nabla L\cdot\nabla\ln\mu - \nabla^2 L)\mathbf{E}_0 - 2(\nabla L\cdot\nabla)\mathbf{E}_0 - (\mathbf{E}_0\cdot\nabla\ln\mu)\nabla L\right. \\ &\quad\left. + (\nabla\cdot\mathbf{E}_0)\nabla L\right] + [(\nabla L)^2 - n^2]\mathbf{E}_0 \\ &\quad + \frac{1}{(jk_0)^2}[\nabla^2\mathbf{E}_0 + \nabla\ln\mu\times(\nabla\times\mathbf{E}_0) - \nabla(\nabla\cdot\mathbf{E}_0)] = 0. \end{aligned} \tag{6.53}$$

The second term is zero due to the eikonal equation and the third term can be ignored for large k_0. From $\nabla \cdot \mathbf{D} = 0$ we have $\nabla \cdot \mathbf{E}_0 = -\mathbf{E}_0 \cdot \nabla \ln \varepsilon$. Thus (6.53) may be written as

$$(\nabla L \cdot \nabla)\mathbf{E}_0 + \frac{1}{2}(\nabla^2 L - \nabla L \cdot \nabla \ln \mu)\mathbf{E}_0 + (\mathbf{E}_0 \cdot \nabla \ln n)\nabla L = 0. \quad (6.54)$$

This is the differential equation for the amplitude $\mathbf{E}_0$, called **transport equation**. The amplitude $\mathbf{H}_0$ of the magnetic field satisfies the similar transport equation

$$(\nabla L \cdot \nabla)\mathbf{H}_0 + \frac{1}{2}(\nabla^2 L - \nabla L \cdot \nabla \ln \varepsilon)\mathbf{H}_0 + (\mathbf{H}_0 \cdot \nabla \ln n)\nabla L = 0. \quad (6.55)$$

Taking the scalar product of (6.54) with $\bar{\mathbf{E}}_0$ and adding the resultant equation to its conjugate, we obtain

$$n\frac{d}{ds}|\mathbf{E}_0|^2 + \mu|\mathbf{E}_0|^2 \nabla \cdot \left(\frac{1}{\mu}\nabla L\right) = 0.$$

The ratio of the electric field intensity at s_2 of a ray to s_1 is then given by

$$\frac{|\mathbf{E}_0|^2_{s_2}}{|\mathbf{E}_0|^2_{s_1}} = \exp\left[-\int_{s_1}^{s_2} \frac{\mu}{n}\nabla \cdot \left(\frac{1}{\mu}\nabla L\right) ds\right]. \quad (6.56)$$

Similarly, we have

$$\frac{|\mathbf{H}_0|^2_{s_2}}{|\mathbf{H}_0|^2_{s_1}} = \exp\left[-\int_{s_1}^{s_2} \frac{\varepsilon}{n}\nabla \cdot \left(\frac{1}{\varepsilon}\nabla L\right) ds\right]. \quad (6.57)$$

In homogeneous medium, we have

$$\begin{aligned}\exp\left[-\int_{s_1}^{s_2} \frac{\mu}{n}\nabla \cdot \left(\frac{1}{\mu}\nabla L\right) ds\right] &= \exp\left[-\int_{s_1}^{s_2} \frac{\varepsilon}{n}\nabla \cdot \left(\frac{1}{\varepsilon}\nabla L\right) ds\right] \\ &= \exp\left(-\frac{1}{n}\int_{s_1}^{s_2} \nabla^2 L \, ds\right). \quad (6.58)\end{aligned}$$

Remark 6.2 (Curvilinear coordinates on a surface)**:** Let v^1 and v^2 be two parameters and S be a surface so that any point P on the surface may be represented by the position vector

$$\mathbf{r} = x(v^1, v^2)\mathbf{u}_x + y(v^1, v^2)\mathbf{u}_y + z(v^1, v^2)\mathbf{u}_z.$$

Curves along which one of the parameters remains constant are called coordinate curves. The vectors

$$\mathbf{e}_1 = \frac{\partial \mathbf{r}}{\partial v^1}, \quad \mathbf{e}_2 = \frac{\partial \mathbf{r}}{\partial v^2}$$

are linearly independent and form a basis at the point P, which is called a **local frame**. Note that the base vector $\mathbf{e}_i$ is the tangent vector along the coordinate curve v^i $(i = 1, 2)$. The metric tensor is defined by

$$g_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j = \frac{\partial x}{\partial v^i}\frac{\partial x}{\partial v^j} + \frac{\partial y}{\partial v^i}\frac{\partial y}{\partial v^j} + \frac{\partial z}{\partial v^i}\frac{\partial z}{\partial v^j}.$$

A vector function $\mathbf{A}$ at the point P may be expanded in terms of the basis $\{\mathbf{e}_1, \mathbf{e}_2\}$ at the point P

$$\mathbf{A} = \sum_{i=1}^{2} a^i \mathbf{e}_i.$$

The differential $d\mathbf{r}$ is an infinitesimal displacement from the point (v^1, v^2) to a neighboring point $(v^1 + dv^1, v^2 + dv^2)$

$$d\mathbf{r} = \sum_{i=1}^{2} \frac{\partial \mathbf{r}}{\partial v^i}\, dv^i = \sum_{i=1}^{2} \mathbf{e}_i\, dv^i.$$

The magnitude of this displacement is denoted by ds

$$ds^2 = d\mathbf{r} \cdot d\mathbf{r} = \sum_{i,j=1}^{2} \mathbf{e}_i \cdot \mathbf{e}_j dv^i dv^j = \sum_{i,j=1}^{2} g_{ij} dv^i dv^j.$$

Especially an infinitesimal displacement at (v^1, v^2) along the v^i-curve is

$$d\mathbf{r}_i = \mathbf{e}_i dv^i$$

and the magnitude of the infinitesimal displacement along the v^i-curve is

$$ds_i = \sqrt{d\mathbf{r}_i \cdot d\mathbf{r}_i} = \sqrt{g_{ii}}\, dv^i.$$

The unit vector along the normal is defined by

$$\mathbf{u}_n = \frac{\mathbf{e}_1 \times \mathbf{e}_2}{|\mathbf{e}_1 \times \mathbf{e}_2|}.$$

In general, the normals at consecutive points of a surface do not intersect. If the normals at consecutive points on a curve intersect, the curve is called a **line of curvature**. The point of intersection of consecutive normals is the **center of curvature**. It can be shown that, at any point P, there are two mutually orthogonal principal directions along which the normals at

consecutive points intersect the normal at P. Thus, there are two centers of curvature at each point P. The distances from P to the centers of curvature, counted positive in the direction of $\mathbf{u}_n$, are the two principal radii of curvature, denoted by R_1 and R_2, and their reciprocals are called the **principal curvatures**. The sum of the principal curvatures is defined as the **first curvature** J

$$J = \frac{1}{R_1} + \frac{1}{R_2}$$

and their product is defined as **Gaussian curvature** (or **second curvature**) κ

$$\kappa = \frac{1}{R_1 R_2}.$$

The coordinate curves form an orthogonal system if and only if $g_{12} = 0$.

□

We now consider the evaluation of $\nabla^2 L$ in (6.58). Consider a short section of the ray tube bounded by two closely spaced constant-phase surfaces $L = L_1$ and $L = L_1 + \Delta L_1$, as illustrated in Figure 6.10. The volume of the short section of ray tube is denoted by V and its boundary by S. The areas of the two ends of the tube are denoted by dS_1 and dS_2, and are given by

$$\begin{aligned} dS_1 &= R_1 R_2 dv^{(1)} dv^{(2)} = \frac{dv^{(1)} dv^{(2)}}{\kappa_1}, \\ dS_2 &= (R_1 + \Delta R_1)(R_2 + \Delta R_2) dv^{(1)} dv^{(2)} = \frac{dv^{(1)} dv^{(2)}}{\kappa_2}, \end{aligned} \tag{6.59}$$

where κ_1 and κ_2 are the Gaussian curvatures at constant-phase surfaces $L = L_1$ and $L = L_1 + \Delta L_1$, respectively. Then we have

$$\int_V \nabla \cdot (\kappa n \mathbf{s}) dV = \int_S \kappa n \mathbf{s} \cdot d\mathbf{S} = \kappa n \mathbf{s} \cdot \mathbf{s}\big|_{L_1 + \Delta L_1} dS_1 - \kappa n \mathbf{s} \cdot \mathbf{s}\big|_{L_1} dS_2 = 0$$

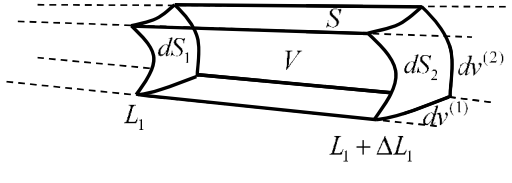


Figure 6.10 A section of ray tube.

where use is made of (6.59). The above equation implies

$$\nabla \cdot (\kappa n \mathbf{s}) = \kappa n \nabla \cdot \mathbf{s} + \mathbf{s} \cdot \nabla(\kappa n) = 0.$$

This gives

$$\nabla \cdot \mathbf{s} = -\frac{1}{\kappa n} \frac{d(\kappa n)}{ds} = -\frac{d \ln(\kappa n)}{ds}.$$

From (6.44), we obtain

$$\nabla^2 L = \nabla \cdot \nabla L = \nabla \cdot (n\mathbf{s}) = n \nabla \cdot \mathbf{s} = -n \frac{d \ln \kappa}{ds} \tag{6.60}$$

in homogeneous medium. As a result

$$\exp\left(-\frac{1}{n} \int_{s_1}^{s_2} \nabla^2 L \, ds \right) = \exp\left(\int_{s_1}^{s_2} \frac{d \ln \kappa}{ds} ds \right) = \frac{\kappa(s_2)}{\kappa(s_1)}.$$

Equations (6.56) and (6.57) can thus be written as

$$|\mathbf{E}_0|_{s_2}^2 = |\mathbf{E}_0|_{s_1}^2 \frac{\kappa(s_2)}{\kappa(s_1)}, \tag{6.61}$$

$$|\mathbf{H}_0|_{s_2}^2 = |\mathbf{H}_0|_{s_1}^2 \frac{\kappa(s_2)}{\kappa(s_1)}. \tag{6.62}$$

A detailed study about the theory and applications of geometric optics can be found in Kline and Kay (1965); Jones (1979a). The ray-tracing method can be used to predict the site-specific propagation models. One of the approaches is based on the Shooting-and-Bouncing Ray (SBR) launching algorithm. The wave propagation models predicted by the ray-tracing method, which has taken the wave reflections, diffractions and scattering into account, play an important role in the design of wireless networks.

Remark 6.3: The refractive index of air depends on pressure, temperature, and humidity. There are several models for the refractive index n of the atmosphere. Two important ones are the **exponential model** (Skolnik, 1980), which assumes $n - 1$ decreases exponentially with altitude, and the **standard model**, which assumes n varies linearly with the altitude as

$$n = n_0 + \frac{dn}{dh} h,$$

where h is the height and dn/dh is assumed to be a constant. The exponential model is better suited for the region in the higher atmosphere where $h > 1\,\text{km}$, and the standard model is mainly for the region in the lower atmosphere where $h < 1\,\text{km}$. □

6.2.3 Ionospheric Wave Propagation

The ionosphere is a region of the atmosphere that is ionized by solar radiation. The ionosphere can be mainly divided into three layers in which the electron density peaks up. These are the D, E, and F layers. During the day time, the F layer splits into two layers called F_1 and F_2 layers. The D layer disappears at night. Typical variation of electron density curve vs. height is shown in Figure 6.11, where N denotes the number of electrons per unit volume. The electrons and ions can be set in motion by electromagnetic fields. In an ionized gas, the equation of motion for a single electron of mass m and charge $-e$ with velocity $\mathbf{v}$, acted upon by an electric field $\mathbf{E}$ and magnetic field $\mathbf{B}$, is

$$m\frac{d\mathbf{v}}{dt} = -e(\mathbf{E} + \mathbf{v} \times \mathbf{B}) - m\nu\mathbf{v}, \tag{6.63}$$

where the term $-m\nu\mathbf{v}$ represents a damping force due to the collisions of electrons with the neutral molecules and ions and ν is the collision frequency. For a sinusoidal field, this becomes

$$j\omega m\mathbf{v} = -e(\mathbf{E} + \mathbf{v} \times \mathbf{B}) - m\nu\mathbf{v}. \tag{6.64}$$

Ignoring the magnetic field in (6.64), the induced current in the ionized gas is given by

$$\mathbf{J} = -eN\mathbf{v} = \frac{Ne^2}{j\omega m + m\nu}\mathbf{E}. \tag{6.65}$$

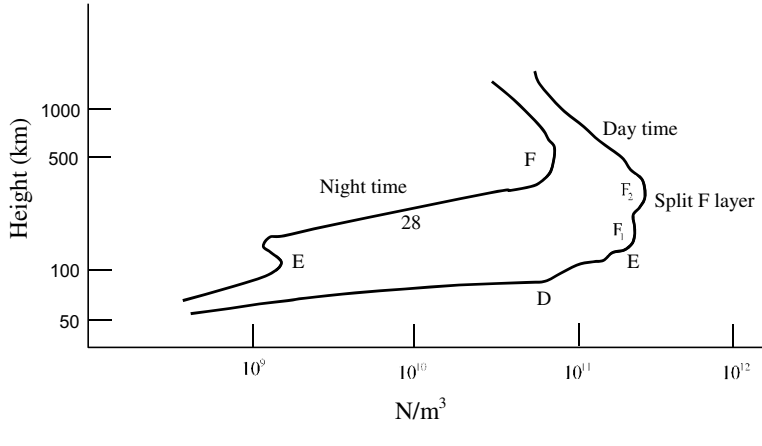


Figure 6.11 Electron density vs. height.

This implies

$$\nabla \times \mathbf{H} = j\omega\varepsilon_0\mathbf{E} + \mathbf{J} = j\omega\varepsilon_0\left[1 - \frac{Ne^2}{m\varepsilon_0\omega(\omega - j\nu)}\right]\mathbf{E}.$$

The relative dielectric constant of the ionized gas is

$$\tilde{\varepsilon}_r = 1 - \frac{Ne^2}{m\varepsilon_0\omega(\omega - j\nu)} = 1 - \frac{\omega_p^2}{\omega(\omega - j\nu)} = 1 - \frac{\omega_p^2}{\omega^2 + \nu^2} - j\frac{\omega_p^2\nu}{\omega(\omega^2 + \nu^2)}, \tag{6.66}$$

where $\omega_p = \sqrt{Ne^2/m\varepsilon_0}$ is the plasma frequency. Equation (6.66) indicates that the ionosphere is equivalent to a medium of relative dielectric constant $\varepsilon_r = 1 - \frac{\omega_p^2}{\omega^2+\nu^2}$ and conductivity $\sigma = \frac{\varepsilon_0\omega_p^2\nu}{\omega^2+\nu^2}$. Therefore, the ionosphere behaves like a dielectric at high frequencies and behaves like a conductor at low frequencies. As a first approximation, the absorption will be neglected ($\nu = 0$). In this case, we have

$$\begin{aligned} \varepsilon_r &< 1, \quad \text{for } \omega > \omega_p, \\ \varepsilon_r &= 0, \quad \text{for } \omega = \omega_p, \\ \varepsilon_r &< 0, \quad \text{for } \omega < \omega_p. \end{aligned}$$

Plane waves propagating in an ionized gas will have a propagation constant $k = \sqrt{\tilde{\varepsilon}_r}k_0$. For normal incidence of a plane wave on the ionosphere, propagation ceases when the wave reaches the height at which the electron density is high enough to make $\varepsilon_r = 0$, and the wave is then reflected back toward the Earth. For oblique incidence as illustrated in Figure 6.12, the wave will be turned around and returned to Earth if a height at which $\sqrt{\varepsilon_r} = \sin\psi_i$ exists (Figure 6.12). This can be explained as follows. According to Snell's law, the ray will follow a path such that the tangent

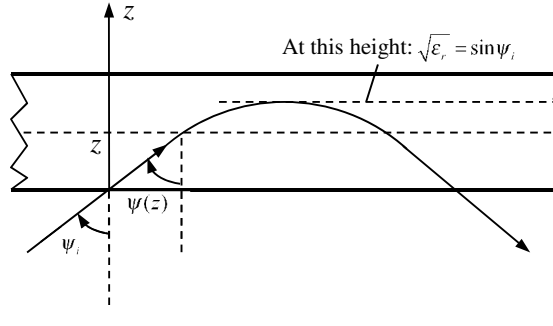


Figure 6.12 Oblique incidence upon the ionosphere.

to the ray satisfies the condition $\sqrt{\varepsilon_r}\sin\psi(z) = \sin\psi_i$, where $\psi(z)$ is the incidence angle at height z. The ray will return to the Earth if $\psi(z)$ reaches $\pi/2$.

In the proceeding discussions, the effect of Earth's magnetic field $\mathbf{B}$ has been neglected. This approximation is reasonably good at frequencies above 10 MHz but is generally not valid at frequencies below 5 MHz. We now will examine the effect of Earth's magnetic field. The Earth's magnetic field will be represented by a steady magnetic field $\mathbf{B}_0$. It follows from (6.64) and (6.65) that

$$(j\omega+\nu)\mathbf{J} + \mathbf{J}\times\frac{e}{m}\mathbf{B}_0 = \omega_p^2\varepsilon_0\mathbf{E}. \tag{6.67}$$

We now choose a rectangular coordinate system (x, y, z) so that $\mathbf{B}_0 = B_0\mathbf{u}_z$. The quantity $\omega_c = \frac{e}{m}B_0$ is referred to as the **gyrofrequency**. From (6.67), we may obtain

$$\begin{bmatrix} J_x \\ J_y \\ J_z \end{bmatrix} = \frac{\varepsilon_0\omega_p^2}{\omega_c^2 - \omega^2 + \nu^2 + 2j\omega\nu}\begin{bmatrix} j\omega+\nu & -\omega_c & 0 \\ \omega_c & j\omega+\nu & 0 \\ 0 & 0 & j\omega+\nu+\dfrac{\omega_c^2}{j\omega+\nu} \end{bmatrix}\begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix}.$$

This can be written as

$$\mathbf{J} = \overleftrightarrow{\boldsymbol{\sigma}}\cdot\mathbf{E}, \tag{6.68}$$

where $\overleftrightarrow{\boldsymbol{\sigma}}$ is a dyadic defined by

$$\overleftrightarrow{\boldsymbol{\sigma}} = \frac{\varepsilon_0\omega_p^2}{\omega_c^2 - \omega^2 + \nu^2 + 2j\omega\nu}\cdot\Bigg[(j\omega+\nu)\mathbf{u}_x\mathbf{u}_x + (j\omega+\nu)\mathbf{u}_y\mathbf{u}_y$$
$$-\omega_c(\mathbf{u}_x\mathbf{u}_y - \mathbf{u}_y\mathbf{u}_x) + \left(j\omega+\nu+\frac{\omega_c^2}{j\omega+\nu}\right)\mathbf{u}_z\mathbf{u}_z\Bigg]. \tag{6.69}$$

Hence, we may write

$$\nabla\times\mathbf{H} = j\omega\varepsilon_0\mathbf{E} + \overleftrightarrow{\boldsymbol{\sigma}}\cdot\mathbf{E} = j\omega\varepsilon_0\overleftrightarrow{\boldsymbol{\varepsilon}}_r\cdot\mathbf{E},$$

where

$$\overleftrightarrow{\boldsymbol{\varepsilon}}_r = \overleftrightarrow{\mathbf{I}} + \frac{\overleftrightarrow{\boldsymbol{\sigma}}}{j\omega\varepsilon_0}$$

and $\overleftrightarrow{\mathbf{I}} = \mathbf{u}_x\mathbf{u}_x + \mathbf{u}_y\mathbf{u}_y + \mathbf{u}_z\mathbf{u}_z$ is the unit dyadic. It is convenient to express $\overleftrightarrow{\boldsymbol{\varepsilon}}_r$ in the form

$$\overleftrightarrow{\boldsymbol{\varepsilon}}_r = \begin{bmatrix} \varepsilon_{r1} & -j\varepsilon_{r2} & 0 \\ j\varepsilon_{r2} & \varepsilon_{r1} & 0 \\ 0 & 0 & \varepsilon_{r3} \end{bmatrix}, \tag{6.70}$$

where

$$\varepsilon_{r1} = 1 - \frac{\omega_p^2(1 - j\nu/\omega)}{\omega^2 - \omega_c^2 - \nu^2 - 2j\omega\nu},$$

$$\varepsilon_{r2} = \frac{\omega_p^2(\omega_c/\omega)}{\omega^2 - \omega_c^2 - \nu^2 - 2j\omega\nu},$$

$$\varepsilon_{r3} = \varepsilon_{r1} + \frac{\omega_p^2\omega_c^2(\omega^2 - j\omega\nu)^{-1}}{\omega^2 - \omega_c^2 - \nu^2 - 2j\omega\nu}.$$

6.2.4 Tropospheric-Scatter-Propagation

Normally, microwave signals are only used for line-of-sight applications, where the receiver can be seen from the transmitter. Radio waves tend to travel in straight lines, which place a limitation on the detection range of a radar system to the objects on its horizon due to the curvature of the Earth. Tropospheric-scatter-propagation uses tropospheric scatter phenomenon to transmit microwave radio signals. When radio waves from a transmitter pass through troposphere and encounter some random irregularities or fluctuations in the index of refraction of the atmosphere, they will be scattered. A distant receiver that beams to the irregularities can pick up the signal if the transmitted power is sufficiently high. An over-the-horizon communication link is thus established as illustrated in Figure 6.13. A related system is meteor burst communications, which uses the ionized trails of meteors to improve the strength of the scattering. Tropospheric-scatter-propagation links operate in the frequency range of 200 MHz–10 GHz. At lower frequencies, the cost of building a high gain

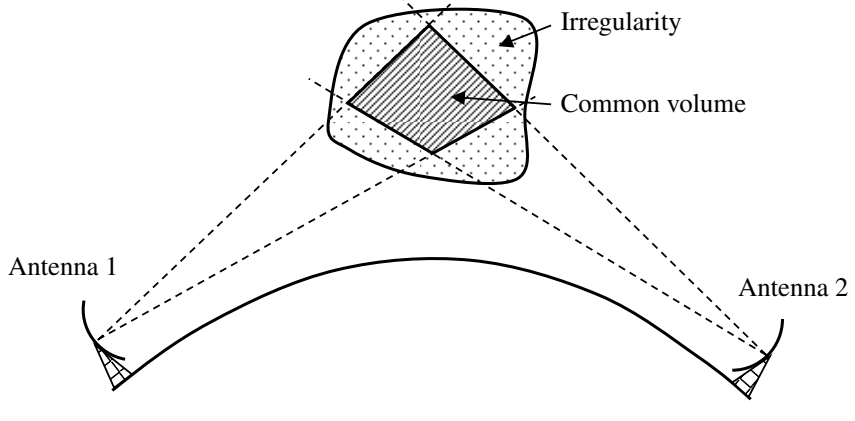


Figure 6.13 Over-the-horizon transmission.

antenna is a major concern. Operation at higher frequencies is prohibited due to the transmission loss.

We assume that the relative dielectric constant for the atmosphere is ε_r, which is very close to unity. Due to the fluctuations in temperature and pressure, the dielectric constant may change to $\varepsilon_r + \Delta\varepsilon_r$, where $\Delta\varepsilon_r$ is typically only a few parts per million. The polarization vector under the influence of a polarizing electric field $\mathbf{E}$ is

$$\mathbf{P} = \varepsilon_0\mathbf{E}(\varepsilon_r - 1 + \Delta\varepsilon_r) \approx \Delta\varepsilon_r\varepsilon_0\mathbf{E}.$$

The polarizing field $\mathbf{E}$ is assumed to be the incident field from antenna 1 characterized by a current distribution $\mathbf{J}_1(\mathbf{r}_1)$, which can be expressed as

$$\mathbf{E}_1(\mathbf{r}_1) = -\frac{jk_0\eta_0 I_1}{4\pi r_1} e^{-jk_0 r_1}\mathbf{L}_1(\mathbf{u}_{r_1}). \tag{6.71}$$

The induced polarization current in the region V_p (Figure 6.14) is

$$\mathbf{J}_p(\mathbf{r}_p) = j\omega\mathbf{P}(\mathbf{r}_p) = j\omega\Delta\varepsilon_r\varepsilon_0\mathbf{E}_1(\mathbf{r}_1) = \Delta\varepsilon_r\frac{k_0^2 I_1}{4\pi r_1} e^{-jk_0 r_1}\mathbf{L}_1(\mathbf{u}_{r_1}). \tag{6.72}$$

In terms of the reciprocity, the open-circuit voltage at the receiver induced by the incident field $\mathbf{E}_s$ generated by the polarization current $\mathbf{J}_p$ is given by

$$\begin{aligned} V_{\text{oc}}(\mathbf{u}_{r_2}) &= -\frac{1}{I_2}\int_{V_2} \mathbf{E}_s(\mathbf{r}_p)\cdot\mathbf{J}_2(\mathbf{r}_2)dV(\mathbf{r}_2) \\ &= -\frac{1}{I_2}\int_{V_p} \mathbf{E}_2(\mathbf{r}_2)\cdot\mathbf{J}_p(\mathbf{r}_p)dV(\mathbf{r}_p). \end{aligned} \tag{6.73}$$

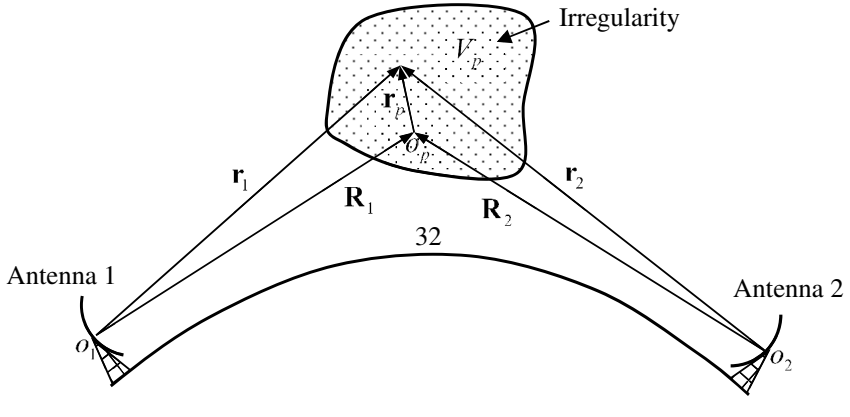


Figure 6.14 Coordinate systems used in tropospheric-scatter-propagation system.

Here $\mathbf{J}_2$ is the current distribution of antenna 2, which occupies a volume V_2 and produces the field $\mathbf{E}_2$

$$\mathbf{E}_2(\mathbf{r}_2) = -\frac{jk_0\eta_0 I_2}{4\pi r_2} e^{-jk_0 r_2} \mathbf{L}_2(\mathbf{u}_{r_2}) \tag{6.74}$$

when it is used as a transmitting antenna. Introducing (6.72) and (6.74) into (6.73), we obtain

$$V_{\text{oc}}(\mathbf{u}_{r_2}) = j\frac{k_0^3\eta_0 I_1}{16\pi^2} \int\limits_{V_p} \Delta\varepsilon_r(\mathbf{r}_p) \frac{e^{-jk_0(r_1+r_2)}}{r_1 r_2} \mathbf{L}_1(\mathbf{u}_{r_1}) \cdot \mathbf{L}_2(\mathbf{u}_{r_2}) dV(\mathbf{r}_p). \tag{6.75}$$

When the antenna 2 is conjugately matched to a load $Z_L = R_L + jX_L$, the received power is given by

$$\begin{aligned} P_{\text{rec}} = \frac{|V_{\text{oc}}|^2}{8R_L} &= \frac{1}{8R_L}\left(\frac{k_0^3\eta_0 |I_1|}{16\pi^2}\right)^2 \\ &\times \int\limits_{V_p}\int\limits_{V_p} \Delta\varepsilon_r(\mathbf{r}_p)\Delta\varepsilon_r(\mathbf{r}'_p)[\mathbf{L}_1(\mathbf{u}_{r_1}) \cdot \mathbf{L}_2(\mathbf{u}_{r_2})] \\ &\times \overline{[\mathbf{L}_1(\mathbf{u}'_{r_1}) \cdot \mathbf{L}_2(\mathbf{u}'_{r_2})]} \frac{e^{-jk_0(r_1+r_2-r'_1-r'_2)}}{r_1 r_2} dV(\mathbf{r}_p) dV(\mathbf{r}'_p). \end{aligned} \tag{6.76}$$

6.2.5 Attenuation by Rain

Radio waves propagating through atmosphere are attenuated because of the power absorption and scattering by particles encountered throughout the propagation path in the atmosphere. Both the absorption and scattering effects are especially prevalent at frequencies above 11 GHz, and are mainly affected by the dimensions of the particles and their electrical properties. The scattering loss is usually small compared to absorption loss. Attenuation due to rain droplets depends on frequency and the rainfall intensity, or rain rate R_0, presented in units of mm/h. The study of absorption and scattering by rain may be started with a single raindrop. In general, the rain droplets take on an oblate spheroidal shape under the influence of aerodynamic forces and pressure forces as they fall. When the frequency is not very high (wavelength is greater than 3 cm), the rain droplet can be approximated by a dielectric sphere with a complex dielectric constant $\tilde{\varepsilon} = \tilde{\varepsilon}_r\varepsilon_0$ with $\tilde{\varepsilon}_r = \tilde{\varepsilon}'_r - j\tilde{\varepsilon}''_r = \varepsilon_r - j\frac{\sigma}{\omega\varepsilon_0}$. We now consider the scattering of a spherical dielectric sphere of radius a illuminated by an

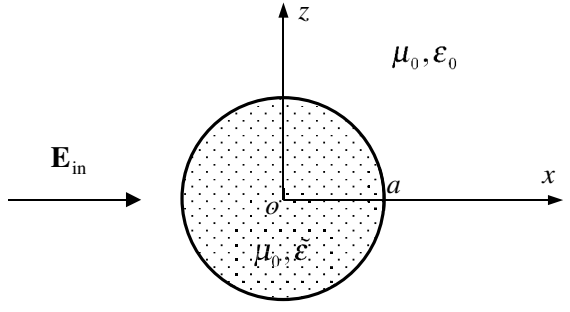


Figure 6.15 Dielectric sphere illuminated by a plane wave.

incident plane wave propagating in positive x-direction (Figure 6.15)

$$\mathbf{E}_{in} = \mathbf{u}_z E_0 e^{-jk_0 x},$$

where $k_0 = \omega\sqrt{\mu_0 \varepsilon_0}$. Since the dielectric sphere is small relative to the wavelength, the polarization produced inside the sphere can be assumed to be the same as would be produced inside the sphere by a uniform static electric field. This boundary value problem has been solved in many textbooks (e.g., Bladel, 2007). The polarization vector $\mathbf{P}$ per unit volume in the dielectric sphere is given by

$$\mathbf{P} = 3\frac{\tilde{\varepsilon}_r - 1}{\tilde{\varepsilon}_r + 2}\varepsilon_0 E_0 \mathbf{u}_z. \tag{6.77}$$

The total dipole moment of the sphere is the integral of the above expression over the spherical volume

$$\mathbf{P}_0 = P_0 \mathbf{u}_z = \frac{4}{3}\pi a^3 \mathbf{P} = 4\pi a^3 \frac{\tilde{\varepsilon}_r - 1}{\tilde{\varepsilon}_r + 2}\varepsilon_0 E_0 \mathbf{u}_z. \tag{6.78}$$

Since the sphere is very small relative to the wavelength, it is equivalent to a small electric dipole of strength $\mathbf{P}_0$. The scattered field from the sphere can thus be obtained from the field produced by the equivalent dipole. The scattered field in the far-field region is then given by [see (7.40)]

$$\mathbf{E}_s(\mathbf{r}) = -\omega k_0 \eta_0 P_0 \sin\theta \frac{e^{-jk_0 r}}{4\pi r}\mathbf{u}_\theta. \tag{6.79}$$

The total scattered power is

$$P_s = \frac{1}{2\eta_0}\int_0^{2\pi}\int_0^{\pi} |\mathbf{E}_s|^2 r^2 \sin\theta \, d\theta \, d\varphi = \frac{4}{3}\pi a^3 (k_0 a)^4 \frac{|E_0|^2}{\eta_0}\left|\frac{\tilde{\varepsilon}_r - 1}{\tilde{\varepsilon}_r + 2}\right|^2. \tag{6.80}$$

The **scattering cross-section** σ_s is defined as the ratio of the total scattered power over the incident power density

$$\sigma_s = \frac{P_s}{|\mathbf{E}_s|/2\eta_0} = \frac{8}{3}\pi a^2 (k_0 a)^4 \left|\frac{\tilde{\varepsilon}_r - 1}{\tilde{\varepsilon}_r + 2}\right|^2. \tag{6.81}$$

The power P_a absorbed by the dielectric sphere can be obtained by the polarization current density $\mathbf{J} = j\omega\mathbf{P}$ and the electric field $\mathbf{E}$ inside the sphere as follows

$$P_a = \frac{1}{2}\mathrm{Re}\int_0^a \int_0^{2\pi} \int_0^{\pi} \mathbf{E}\cdot\bar{\mathbf{J}} r^2 \sin\theta\, dr\, d\theta\, d\varphi = \frac{2}{3}\pi a^3 \mathrm{Re}\,\mathbf{E}\cdot\bar{\mathbf{J}}$$

$$= 6\pi a^3 \frac{k_0}{\eta_0}\left|\frac{\tilde{\varepsilon}_r - 1}{\tilde{\varepsilon}_r + 2}\right|^2 \frac{\tilde{\varepsilon}_r'' |E_0|^2}{(\tilde{\varepsilon}_r' - 1)^2 + (\tilde{\varepsilon}_r'')^2}. \tag{6.82}$$

The **absorption cross-section** σ_a is defined as the absorbed power divided by the incident power density

$$\sigma_a = \frac{P_a}{|\mathbf{E}_{\mathrm{in}}|^2/2\eta_0} = 12\pi k_0 a^3 \left|\frac{\tilde{\varepsilon}_r - 1}{\tilde{\varepsilon}_r + 2}\right|^2 \frac{\tilde{\varepsilon}_r''}{(\tilde{\varepsilon}_r' - 1)^2 + (\tilde{\varepsilon}_r'')^2}. \tag{6.83}$$

For small dielectric sphere, we usually have $\sigma_a > \sigma_s$. The **extinction cross-section** σ_e is defined as the total power removed from the incident field due to the scattering and absorption divided by the incident power density. Therefore, it is the sum of scattering and absorption cross sections

$$\sigma_e = \sigma_s + \sigma_a. \tag{6.84}$$

A radio wave going through rain encounters a large number of water droplets with different radii. The size distribution of droplets is usually written in the form $N(a)da$ and it represents the number of droplets with radii in the interval $[a, a + da]$ per unit volume. The power removed from a wave propagating in the z-direction with power density $p = |\mathbf{E}|^2/2\eta_0$ by the raindrops in a volume element of unit cross-sectional area and thickness dz can be written as

$$-dp = pdz \int_0^{\infty} \sigma_e N(a) da. \tag{6.85}$$

This can be rewritten as

$$\frac{dp}{dz} = -A(z)p, \tag{6.86}$$

Table 6.2 Numerical values for N_0 and Λ

Authors	Rain types	N_0 ($\text{mm}^{-1}/\text{m}^3$)	Λ (mm^{-1})
Marshall and Palmer (1948)	Any intensity of rainfall	8000	$8.2R_0^{-0.21}$
	Drizzle	30,000	$10.14R_0^{-0.21}$
Joss *et al.* (1968)	Continuous rain	7000	$8.2R_0^{-0.21}$
	Convective rain	1400	$6R_0^{-0.21}$

where $A(z) = \int_0^\infty \sigma_e N(a)da$ depends on z since the size distribution $N(a)$ may vary along the propagation path due to the non-uniform rain. The size distribution is a function of rain rate R_0 and has been investigated by a number of authors. The best known empirical expressions for the size distribution were proposed by Marshall and Palmer (1948) and Joss *et al.* (1968):

$$N(a) = N_0 e^{-\Lambda a}, \tag{6.87}$$

where N_0 and Λ are experimentally determined constants. Table 6.2 shows the numerical values of N_0 and Λ for different types of rainfall.

6.3 Statistical Models for Mobile Radio Channels

In a mobile environment, the transmission medium is very lossy and dispersive and suffers extreme random fades due to multiple scattering and the absence of a direct line-of-sight path between the base station and the mobile terminal. Fades of 40 dB or more below the mean level are common, with successive minima occurring about every half wavelength of the carrier frequency. In addition, the mobile terminal whose location is unknown introduces Doppler shifts, named after the Austrian physicist Christian Doppler (1803–1853). For this reason, the conventional antenna designs have to be tailored to the statistical nature of the environment (Jakes, 1994).

Path loss is the reduction in power density of an electromagnetic wave when it propagates through space. The path loss may be due to many effects, such as free-space loss, reflection, refraction, diffraction, absorption, terrain contours, and the distance between the transmitter and the receiver. The prediction of the path loss information is very important in designing a wireless communication system. The pass-loss models can be roughly divided into statistical and deterministic models. The statistical models

are derived from extensive field measurements and statistical analysis, and are valid for similar environments where the measurements were carried out. The deterministic models are usually based on the numerical methods, which take all the effects into consideration and involve much more data computational power.

6.3.1 Near-Earth Large-Scale Models

Channel impairments are caused by several inter-related mechanisms, which are path loss, blockage (i.e., attenuation, radio wave in the transmission path may be partially blocked or absorbed by some feature of the environment), fast fading, shadowing, random FM (relative Tx/Rx motion) and delay spread (multiple signals arrive with a slight additional delay which spreads the received signal and causes each symbol to overlap with proceeding and following symbols, producing intersymbol interference). The propagation characteristics are divided into large scale (path loss plus shadowing) and small scale. We consider large-scale models in this section.

Because of obstructions such as different types of terrains, landscapes for outdoor environment or different building structure, layout for indoor environment, propagation losses can be significantly higher than in free space. In practice, the path loss or **propagation model** for both indoor and outdoor applications should be modified as

$$L_s = \left(\frac{4\pi R}{\lambda}\right)^2 \times \text{Correction factors},$$

or in dB we have

$$L_s \text{ (dB)} = 20\log f(\text{MHz}) + 20\log R \text{ (km)} + 32.44 + \text{Correction terms}.$$

The correction terms depend on the propagation environments such as urban, suburban, open area, mountains, hills, lakes, buildings and layout of the buildings and streets etc., which are generally very complicated. So the correction terms are usually based on measured data and most of the propagation models are of semi-empirical type. The predictions from these models may have a large deviation from the actually measured data and special correction factors have to be introduced to account for significant features of the environment. It should be noted that the correction factors to be used also depend on the carrier frequency. As frequency changes some factors might become increasingly important and others may become

negligible. A rule-of-thumb is that radio waves are affected strongly by objects with physical dimensions comparable to their wavelength. The best-known outdoor models are Okumura, Hata and Cost-231 models.

6.3.1.1 *Okumura Model*

The **Okumura model** is for urban and suburban areas, which is a radio propagation model that was built using the data collected in the city of Tokyo, Japan. The model is ideal for cities with many urban structures but not many tall blocking structures. The model is served as a base for the Hata Model to be introduced later, and is given by

$$\begin{aligned} L_s\ (\text{dB}) = {} & 20\log f(\text{MHz}) + 20\log R\ (\text{km}) + 32.44 \\ & + A(f,R) - G(h_b) - G(h_m) - G_{\text{area}}, \end{aligned} \tag{6.88}$$

where $A(f,R)$ is the median attenuation relative to free space; $G(h_b) = 20\log(h_b/200)$ is base station antenna height gain factor, and h_b is the base station antenna height; $G(h_m) = 10\log(h_m/3)$ is mobile antenna height gain factor, and h_m is the mobile station antenna height; G_{area} is the area gain. The correction factors $A(f,R)$ and G_{area} are in the form of a set of curves (Okumura *et al.*, 1968). The applicable ranges for this model are

$$\begin{gathered} 100\,\text{MHz} < f < 3000\,\text{MHz}, \\ 1\,\text{km} < R < 100\,\text{km}, \\ 20\,\text{m} < h_b < 1000\,\text{m}, \\ 1\,\text{m} < h_m < 10\,\text{m}. \end{gathered}$$

Okumura's model is considered to be among the simplest and best in terms of accuracy in path loss prediction for mature cellular and land mobile radio systems in cluttered environments. Common standard deviations between predicted and measured path loss values are around 10 dB to 14 dB.

6.3.1.2 *Hata Model*

The **Hata Model** is for urban areas. It is also known as the Okumura–Hata model for being a developed version of the Okumura Model, and is the most widely used propagation model for predicting the behavior of cellular transmissions in built up areas. This model incorporates the graphical information from Okumura model and develops it further to include the effects of diffraction, reflection and scattering caused by city structures.

The Hata model is

$$L_s\ (\text{dB}) = 26.16 \log f(\text{MHz}) - 13.82 \log h_b + 69.55 + [44.9 - 6.55 \log h_b] \log R - A(h_m), \tag{6.89}$$

where $A(h_m)$ is a correction factor for the city size,

$$A(h_m) = \begin{cases} [1.1 \log f - 0.7]h_m - [1.56 \log f - 0.8], \\ \qquad \text{(small and medium city)} \\ 8.9[\log(1.54h_m)]^2 - 1.1, \quad \text{(large city, } f < 300\,\text{MHz)} \\ 3.2[\log(11.75h_m)]^2 - 4.97, \quad \text{(large city, } f > 300\,\text{MHz)} \end{cases}.$$

The applicable ranges for this model are

$$100\,\text{MHz} < f < 1500\,\text{MHz},$$
$$1\,\text{km} < R < 100\,\text{km},$$
$$20\,\text{m} < h_b < 1000\,\text{m},$$
$$1\,\text{m} < h_m < 10\,\text{m}.$$

6.3.1.3 *COST-231 Model*

The COST-231 model is for urban area that extends the urban Hata Model to cover a more elaborated range of frequencies (COST is a European Union Forum for cooperative scientific research) and it is given by

$$L_s(\text{dB}) = 46.3 + 33.9 \log f(\text{MHz}) - 13.82 \log h_b + [44.9 - 6.55 \log h_b] \log R - A(h_m) + C_M, \tag{6.90}$$

where $C_M = 0$ for suburban and mid-size city and $C_M = 3$ for metropolitan areas. The applicable ranges for this model are

$$1500\,\text{MHz} < f < 2000\,\text{MHz},$$
$$1\,\text{km} < R < 20\,\text{km},$$
$$30\,\text{m} < h_b < 200\,\text{m},$$
$$1\,\text{m} < h_m < 10\,\text{m}.$$

The above three models are suitable for large outdoor cells. To increase the capacity and lower the power, street level cells may be used, where the antennas are low and cells are small. In these cases, the propagation model should be based on more site-specific information such as street width, street orientation, rooftop height, etc.

6.3.1.4 *Log-Distance Model*

For indoor environment, the propagation loss can be higher because of a combination of attenuation by walls and ceilings, and blockage due to equipment, furniture, and even people. Experience has shown that line-of-sight propagation holds only for about the first 20 feet. Beyond 20 feet, the propagation losses increase at up to 30 dB per 100 feet in dense office environments. This is a good rule-of-thumb although it overstates path loss in most cases. The best-known models for indoor applications are log-distance path loss expressed by

$$L_s = L_s(R_0) + 10N \log\left(\frac{R}{R_0}\right),$$

where $L_s(R_0)$ is the path loss at a reference point R_0 (usually 1 meter) and N is the path loss exponent which varies according to the environment, as shown in Table 6.3. This model is not restricted to the indoor applications but it finds more useful in the indoor environment. Although it is not very accurate it is very suitable for rough quick calculations.

The complexity and numerous parameters involved in determining the path loss in an indoor environment makes it hard to have a simple statistical model with small error variance. So the site-specific modeling will eventually prevail.

6.3.2 Small-Scale Fading

Small-scale fading, or simply **fading**, is used to describe the rapid fluctuation of the amplitude of radio signal over a short period of time or travel distance, so that the large-scale path loss effects may be ignored. Signal fading occurs when waves travel along different paths (called **multipath waves**) and interfere destructively with each other when they reach a receiving antenna as shown in Figure 6.16. The term **multipath** applies when there is more than one path that the radio wave can travel from the

Table 6.3 Path loss

Environment	Path loss exponent N
Free space	2
Urban area cellular radio	2.7 to 3.5
Shadowed urban cellular radio	3 to 5
Indoor line-of-sight	1.6 to 1.8
Obstructed indoor	4 to 6
Obstructed in-factories	2 to 3

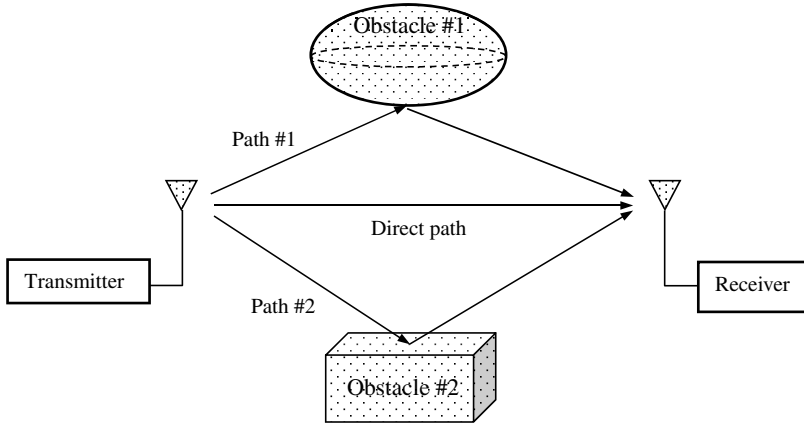


Figure 6.16 Multipath.

transmitter to the receiver. In this sense all radio channels are multipath channels. Multipath is useful because it allows radio waves to bend around corners to reach behind hills and buildings, into parking garages and tunnels.

Many physical factors in the radio propagation channel influence small-scale fading. These are summarized below:

(1) Multipath propagation — The random phase and amplitudes of the different multipath waves that arrive at the receiver cause fluctuations in signal strength, thereby inducing small-scale fading, signal distortion, or both.
(2) Speed of the mobile — The relative motion between the base station and the mobile results in random frequency modulation due to different Doppler shifts on each of the multipath waves. Doppler shifts will be positive or negative depending on whether the mobile receiver is moving toward or away from the base station.
(3) Speed of surrounding objects — The motion of objects in the radio channel will introduce a time varying Doppler shift on multipath waves. If the objects move at a greater rate than the mobile, then this effect will dominate the small-scale fading. Otherwise motion of surrounding objects may be ignored, and only the speed of the mobile needs to be considered.
(4) The transmission bandwidth of the signal — If the transmitted radio signal bandwidth is greater than the coherence bandwidth of the multipath channel, the received signal will be distorted, but the received

signal strength will not fade much over a local area (i.e., the small-scale signal fading will be insignificant). If the transmitted signal has a narrow bandwidth as compared to the channel the amplitude of the signal will change rapidly, but the signal will not be distorted in time.

The **coherence bandwidth** is a statistical measure of the range of frequencies over which the channel can be considered "flat" (i.e., a channel which passes all spectral components with approximately equal gain and linear phase). In an indoor environment, multipath is almost always present and tends to be constantly varying. Severe fading due to multipath can result in a signal reduction of more than 30 dB, and may cause failure in communication. The rate of power decrease in a multipath environment is $1/R^n (n > 2)$. One method of overcoming this problem is to transmit more power since signal cancellation is never complete. The amount of extra RF power radiated to overcome signal fading is called **fade margin**. The exact amount of fade margin required depends on the desired reliability of the communication link. A good rule-of-thumb is 20 dB to 30 dB.

Another method of reducing the effects of multipath is antenna diversity (space diversity, frequency diversity, and polarization diversity). Since the cancellation is geometry-dependent, use of two or more antennas separated by at least half of a wavelength can drastically mitigate this problem, which is called **space diversity**. On acquisition of a signal, the receiver checks each antenna and simply selects the antenna with the best signal quality. This reduces the required link margin that would otherwise be needed for a system without employing diversity. The disadvantage is that this approach requires more antennas and a more complicated receiver design. It can be shown that the probability of two different frequencies to be in a fade at the same time is statistically unlikely therefore **frequency diversity** (also called frequency hopping) is often used. For example, the frequency changes in a GSM system occur 217 times a second. Due to the radio wave interacting with its surrounding, there may be some alternations in the polarization of the wave. To combat this, dual polar antennas can be used, which is called **polarization diversity**.

One can also use an adaptive channel equalizer to deal with the multipath problem. Generally speaking, equalization is a process of correcting irregularities in the parameters of a given link by means of preset and/or adjustable networks. In the ideal situation, an equalizer can be represented by a two-port network with a particular transfer function $H_{\text{eq}}(\omega)$, such that when cascaded with the transmission system with transfer function $G(\omega)$ (Figure 6.17), the overall response is distortionless. Thus, we

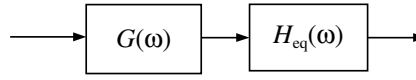


Figure 6.17 Equalizer.

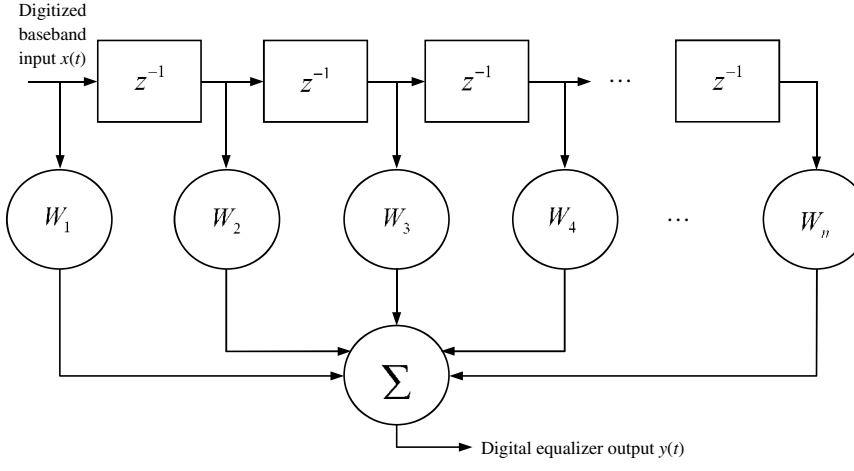


Figure 6.18 Digital equalizer.

must have

$$G(\omega) \cdot H_{\text{eq}}(\omega) = K \exp(-j\omega t_d),$$

or

$$H_{\text{eq}}(\omega) = \frac{K \exp(-j\omega t_d)}{G(\omega)},$$

where K is a constant. The practical realization of the necessary network is often very difficult, and normally some compromise solution is reached such that the overall characteristic is operationally acceptable. One versatile form of adjustable equalizer is the tapped delay line or transversal filter (Figure 6.18). After the signal is received and digitized, it is fed through a series of adaptive delay stages, which are summed together via feedback loops. This technique is particularly effective in slowly changing environment. The main drawback is the increase of system cost and complexity. The output of the equalizer is

$$y(t) = W_1 x(t) + W_2 x(t-\tau) + W_3 x(t-2\tau) + \cdots + W_n x[t-(n-1)\tau].$$

where W_i $(i = 1, 2, \ldots, n)$ are weighting coefficients. In frequency domain, this equation becomes

$$Y(\omega) = X(\omega)\left[W_1 + W_2 e^{-j\omega\tau} + W_3 e^{-j2\omega\tau} + \cdots + W_n e^{-j(n-1)\tau}\right],$$

where $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of $x(t)$ and $y(t)$, respectively. By suitable choice of the weighting coefficients an approximation to the required transfer function can be produced. The greater the number of taps, the more flexible the overall equalizer becomes.

Spread spectrum systems are fairly robust in the presence of multipath. The term spread spectrum simply means that the energy radiated by the transmitter is spread out over a wider range of the frequency spectrum than would otherwise be used. The Directive Sequence Spread Spectrum (DSSS), Code Division Multiple Access (CDMA), and Frequency Hopping Spread Systems (FHSS) are all considered to be spread spectrum systems.

6.4 Propagation Models for Deterministic MIMO System

Site-specific deterministic propagation models are often preferred for more accurate predictions of radio wave propagations than would be available from statistical models. In this section, we provide a method to predict the propagation model for a general multiple-input multiple-output (MIMO) system. An important performance index for characterizing a communication system is the spectral efficiency measured in bit/s/Hz. Shannon's channel capacity theorem (see Chapter 8) reveals that there is a maximum spectral efficiency, called channel capacity, at which any communication system can operate reliably (Shannon, 1948; 1949). The MIMO system has emerged as one of the most promising technologies to increase the capacity of the wireless link. In a MIMO wireless system, multiple antenna elements are deployed and the data stream from a single user is demultiplexed into n_t (the number of transmitting antennas) substreams. Each substream is then encoded into channel symbols, and the signals are received by n_r receiving antennas. Various coding schemes, such as layered space-time codes, space-time Trellis codes and space-time block codes, have been proposed to exploit the benefits of MIMO channels.

6.4.1 Channel Matrix

A general linear time-invariant MIMO system with n_t inputs and n_r outputs is shown in Figure 6.19, where T_i $(i = 1, 2, \ldots, n_t + n_r)$ is the ith antenna

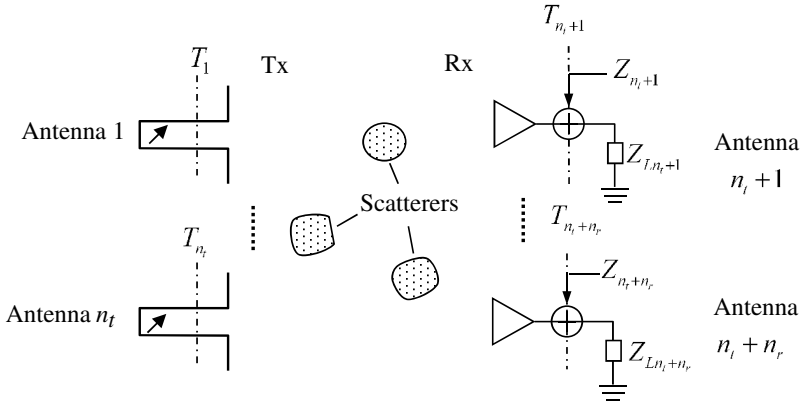


Figure 6.19 An arbitrary MIMO system.

input plane (i.e., the reference plane), and the n_r outputs are terminated by loads Z_{Li} at T_i $(i = n_t + 1, \ldots, n_t + n_r)$.

For a noisy MIMO network, the relationship between the normalized incident waves and reflected waves (all are considered as random processes, see Chapter 8) can be expressed as

$$\begin{bmatrix} b_1 \\ \vdots \\ b_{n_t} \\ b_{n_t+1} \\ \vdots \\ b_{n_t+n_r} \end{bmatrix} = \begin{bmatrix} S_{11} & \cdots & S_{1n_t} & S_{1(n_t+1)} & \cdots & S_{1(n_t+n_r)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ S_{(n_t)1} & \cdots & S_{n_tn_t} & S_{n_t(n_t+1)} & \cdots & S_{n_t(n_t+n_r)} \\ S_{(n_t+1)1} & \cdots & S_{(n_t+1)n_t} & S_{(n_t+1)(n_t+1)} & \cdots & S_{(n_t+1)(n_t+n_r)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ S_{(n_t+n_r)1} & \cdots & S_{(n_t+n_r)n_t} & S_{(n_t+n_r)(n_t+1)} & \cdots & S_{(n_t+n_r)(n_t+n_r)} \end{bmatrix} \times \begin{bmatrix} a_1 \\ \vdots \\ a_{n_t} \\ a_{n_t+1} \\ \vdots \\ a_{n_t+n_r} \end{bmatrix} + \begin{bmatrix} b_1^n \\ \vdots \\ b_{n_t}^n \\ b_{n_t+1}^n \\ \vdots \\ b_{n_t+n_r}^n \end{bmatrix} \tag{6.91}$$

where S_{ij} is the transmission coefficient from antenna j to antenna i, and b_i^n is the normalized noise wave which will be assumed to be a zero-mean white Gaussian noise. Let $\overline{\overline{f}}$ denote the ensemble average of f. For a stationary and ergodic random process f, the ensemble average equals the time average, i.e., $\overline{\overline{f}} = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} f(t)dt$. Taking the ensemble average of (6.91) and making use of the assumption that the channel is deterministic and b_i^n $(i = 1, 2, \ldots, n_t + n_r)$ have zero-mean lead to

$$\begin{bmatrix} \overline{\overline{b}}_1 \\ \vdots \\ \overline{\overline{b_{n_t}}} \\ \overline{\overline{b_{n_t+1}}} \\ \vdots \\ \overline{\overline{b_{n_t+n_r}}} \end{bmatrix} = \begin{bmatrix} S_{11} & \cdots & S_{1n_t} & S_{1(n_t+1)} & \cdots & S_{1(n_t+n_r)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ S_{n_t1} & \cdots & S_{n_tn_t} & S_{n_t(n_t+1)} & \cdots & S_{n_t(n_t+n_r)} \\ S_{(n_t+1)1} & \cdots & S_{(n_t+1)n_t} & S_{(n_t+1)(n_t+1)} & \cdots & S_{(n_t+1)(n_t+n_r)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ S_{(n_t+n_r)1} & \cdots & S_{(n_t+n_r)n_t} & S_{(n_t+n_r)(n_t+1)} & \cdots & S_{(n_t+n_r)(n_t+n_r)} \end{bmatrix} \times \begin{bmatrix} \overline{\overline{a_1}} \\ \vdots \\ \overline{\overline{a_{n_t}}} \\ \overline{\overline{a_{n_t+1}}} \\ \vdots \\ \overline{\overline{a_{n_t+n_r}}} \end{bmatrix}. \tag{6.92}$$

Hereafter all antennas will be assumed to be matched so that $\overline{\overline{b_1}} = \cdots = \overline{\overline{b_{n_t}}} = 0$ and $\overline{\overline{a_{n_t+1}}} = \cdots = \overline{\overline{a_{n_t+n_r}}} = 0$. Then (6.92) reduces to

$$\begin{bmatrix} \overline{\overline{b_{n_t+1}}} \\ \vdots \\ \overline{\overline{b_{n_t+n_r}}} \end{bmatrix} = \begin{bmatrix} S_{(n_t+1)1} & \cdots & S_{(n_t+1)n_t} \\ \vdots & \ddots & \vdots \\ S_{(n_t+n_r)1} & \cdots & S_{(n_t+n_r)n_t} \end{bmatrix} \begin{bmatrix} \overline{\overline{a_1}} \\ \vdots \\ \overline{\overline{a_{n_t}}} \end{bmatrix}. \tag{6.93}$$

We mention in passing that the matrix

$$\mathbf{H} = \begin{bmatrix} S_{(n_t+1)1} & \cdots & S_{(n_t+1)n_t} \\ \vdots & \ddots & \vdots \\ S_{(n_t+n_r)1} & \cdots & S_{(n_t+n_r)n_t} \end{bmatrix} \tag{6.94}$$

can be identified as the channel matrix of MIMO system (Geyi, 2007b).

6.4.2 Computation of Channel Matrix Elements

The channel matrix can be determined from electromagnetic theory. Consider a system consisting of n antennas contained in a region V_∞ bounded by S_∞. Let the fields (ensemble averages) generated by antenna i ($i = 1, 2, \ldots, n$) when antenna j ($j \neq i$) are receiving with all scatterers being in place be denoted by $\mathbf{E}_i$, $\mathbf{H}_i$, and V_i be the source region of antenna i, which is chosen in such a way that its boundary, denoted by S_i, is coincident with the metal surface of the antennas except for a portion of the reference plane T_i. This state of operation is illustrated in Figure 6.20, where the medium around the antenna is assumed to be isotropic and inhomogeneous. The frequency-domain reciprocity theorem for the ensemble averages of the complex envelopes of the fields may be written as

$$\int_S \left(\overline{\overline{\mathbf{E}_i}} \times \overline{\overline{\mathbf{H}_j}} - \overline{\overline{\mathbf{E}_j}} \times \overline{\overline{\mathbf{H}_i}} \right) \cdot \mathbf{u}_n \, dS = 0,$$

where it is assumed that the closed surface S does not contain any impressed sources. Similar to the two-antenna system discussed in Chapter 5, we have

$$S_{ij} = \left. \frac{\overline{\overline{b_i^{(j)}}}}{\overline{\overline{a_j^{(j)}}}} \right|_{a_l^{(j)}=0, l\neq j} = -\frac{1}{2\overline{\overline{a_i^{(i)} a_j^{(j)}}}} \int_{S_i'} \left(\overline{\overline{\mathbf{E}_i}} \times \overline{\overline{\mathbf{H}_j}} - \overline{\overline{\mathbf{E}_j}} \times \overline{\overline{\mathbf{H}_i}} \right) \cdot \mathbf{u}_n \, dS, \tag{6.95}$$

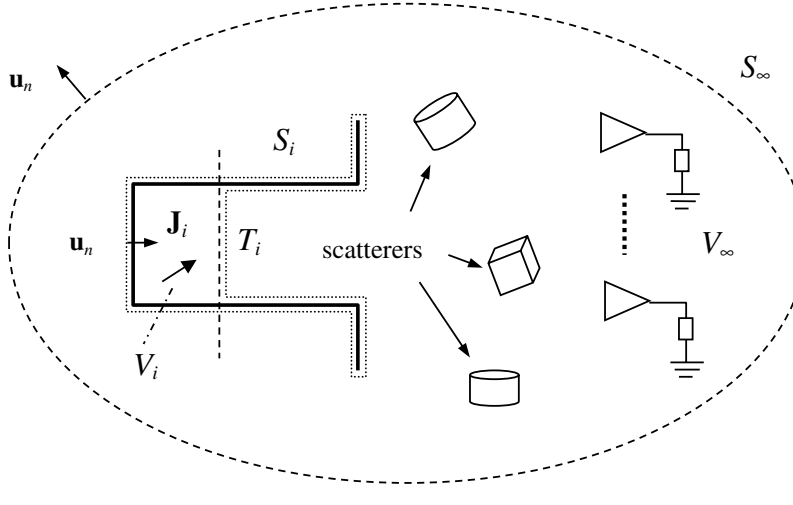


Figure 6.20 Derivation of scattering parameters.

where S_i' is a closed surface containing antenna i only. Note that the fields $\overline{\overline{\mathbf{E}}}_i$, $\overline{\overline{\mathbf{H}}}_i$ should be determined with all antenna elements and scatterers being in place.

The multipath fading due to the scatterers plays an important role in a MIMO system. The presence of significant scatterers promises that the waves from different paths will add differently at each receiving antenna element so that the n_r receiving signals are independent and can be used to unscramble the n_t transmitted signals. To predict how the MIMO channel matrix changes with environments, the general expression (6.95) can be used via numerical simulations with all scatterers being in place. If the presence of the scatterers does not change the field distributions significantly, a perturbation procedure may be adopted. Consider an arbitrary region V_p enclosed by a surface S_p in which the medium is assumed to be linear, isotropic and free of impressed source (Figure 6.21). The medium in V_p may be inhomogeneous with a permeability $\mu(\mathbf{r})$, permittivity $\varepsilon(\mathbf{r})$, and conductivity $\sigma(\mathbf{r})$. Thus, one may write

$$\begin{cases} \nabla \times \overline{\overline{\mathbf{H}}}(\mathbf{r}) = [\sigma(\mathbf{r}) + j\omega\varepsilon(\mathbf{r})]\overline{\overline{\mathbf{E}}}(\mathbf{r}) + \overline{\overline{\mathbf{J}}}(\mathbf{r}) \\ \nabla \times \overline{\overline{\mathbf{E}}}(\mathbf{r}) = -j\omega\mu(\mathbf{r})\overline{\overline{\mathbf{H}}}(\mathbf{r}) \end{cases}. \tag{6.96}$$

If the medium parameters $\mu(\mathbf{r})$, $\varepsilon(\mathbf{r})$, $\sigma(\mathbf{r})$are changed to $\mu'(\mathbf{r})$, $\varepsilon'(\mathbf{r})$, $\sigma'(\mathbf{r})$ in V_p, the perturbed fields in V_p will be governed by

$$\begin{cases} \nabla \times \overline{\overline{\mathbf{H}'}}(\mathbf{r}) = [\sigma'(\mathbf{r}) + j\omega\varepsilon'(\mathbf{r})]\overline{\overline{\mathbf{E}'}}(\mathbf{r}) + \overline{\overline{\mathbf{J}}}(\mathbf{r}) \\ \nabla \times \overline{\overline{\mathbf{E}'}}(\mathbf{r}) = -j\omega\mu'(\mathbf{r})\overline{\overline{\mathbf{H}'}}(\mathbf{r}) \end{cases}, \tag{6.97}$$

which may be rewritten as

$$\begin{cases} \nabla \times \overline{\overline{\mathbf{H}'}}(\mathbf{r}) = [\sigma(\mathbf{r}) + j\omega\varepsilon(\mathbf{r})]\overline{\overline{\mathbf{E}'}}(\mathbf{r}) + \overline{\overline{\mathbf{J}'}}(\mathbf{r}) + \overline{\overline{\mathbf{J}}}(\mathbf{r}) \\ \nabla \times \overline{\overline{\mathbf{E}'}}(\mathbf{r}) = -j\omega\mu(\mathbf{r})\overline{\overline{\mathbf{H}'}}(\mathbf{r}) - \overline{\overline{\mathbf{J}'_m}}(\mathbf{r}) \end{cases}, \tag{6.98}$$

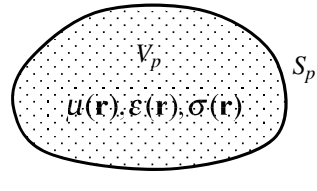


Figure 6.21 An arbitrary region where medium property changes.

where

$$\begin{cases} \overline{\overline{\mathbf{J}'}}(\mathbf{r}) = \{\sigma'(\mathbf{r}) - \sigma(\mathbf{r}) + j\omega[\varepsilon'(\mathbf{r}) - \varepsilon(\mathbf{r})]\}\overline{\overline{\mathbf{E}'}}(\mathbf{r}) \\ \overline{\overline{\mathbf{J}'_m}}(\mathbf{r}) = j\omega\overline{\overline{\mathbf{H}'}}(\mathbf{r})\,[\mu'(\mathbf{r}) - \mu(\mathbf{r})] \end{cases}. \tag{6.99}$$

Comparing (6.96) and (6.98), it may be found that the perturbed fields can be determined by introducing an equivalent electric current source $\overline{\overline{\mathbf{J}'}}$ and an equivalent magnetic current source $\overline{\overline{\mathbf{J}'_m}}$ in the region V_p, as if the medium parameters had not changed in V_p. This is what the compensation theorem implies (Geyi, 2010). The differences of the fields $\Delta\overline{\overline{\mathbf{E}}} = \overline{\overline{\mathbf{E}'}} - \overline{\overline{\mathbf{E}}}$, $\Delta\overline{\overline{\mathbf{H}}} = \overline{\overline{\mathbf{H}'}} - \overline{\overline{\mathbf{H}}}$ satisfy the following equations

$$\begin{cases} \nabla \times \Delta\overline{\overline{\mathbf{H}}}(\mathbf{r}) = [\sigma(\mathbf{r}) + j\omega\varepsilon(\mathbf{r})]\Delta\overline{\overline{\mathbf{E}}}(\mathbf{r}) + \overline{\overline{\mathbf{J}'}}(\mathbf{r}) \\ \nabla \times \Delta\overline{\overline{\mathbf{E}}}(\mathbf{r}) = -j\omega\mu(\mathbf{r})\Delta\overline{\overline{\mathbf{H}}}(\mathbf{r}) - \overline{\overline{\mathbf{J}'_m}}(\mathbf{r}) \end{cases}. \tag{6.100}$$

Therefore, the equivalent sources (6.99) generate the differential fields.

The influences of the change of the medium parameters on the scattering parameters can be studied by means of compensation theorem. Figure 6.22 shows any two antenna element i and j and a region V_p enclosed by S_p, where the changes of medium parameters take place. Two scenarios may be considered:

Scenario 1: The medium parameters are assumed to be μ, ε and σ. The antenna i produces the fields $\overline{\overline{\mathbf{E}_i}}$, $\overline{\overline{\mathbf{H}_i}}$ when all other antennas are receiving. The transmission coefficient between antenna i and antenna j ($j \neq i$) is denoted by S_{ij}.

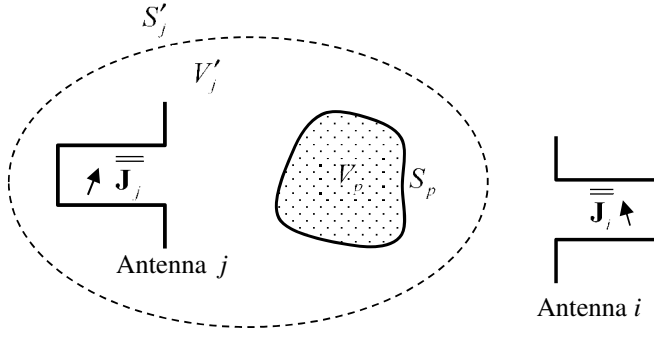


Figure 6.22 Coupling between two antenna elements in a scattering environment.

Scenario 2: The medium parameters μ, ε and σ in V_p are changed to μ', ε' and σ' respectively. The antenna i produces the field $\overline{\overline{\mathbf{E}'_i}}, \overline{\overline{\mathbf{H}'_i}}$ when all other antennas are receiving. The transmission coefficient between antenna i and antenna j $(j \neq i)$ is denoted by S'_{ij}.

From (6.95) and the reciprocity theorem in a region with impressed sources, the transmission coefficient for Scenario 1 may be expressed as

$$\begin{aligned} S_{ij} &= -\frac{1}{\overline{\overline{2a_i^{(i)} a_j^{(j)}}}} \int_{S'_i} \left(\overline{\overline{\mathbf{E}_i}} \times \overline{\overline{\mathbf{H}_j}} - \overline{\overline{\mathbf{E}_j}} \times \overline{\overline{\mathbf{H}_i}} \right) \cdot \mathbf{u}_n \, dS \\ &= -\frac{1}{\overline{\overline{2a_i^{(i)} a_j^{(j)}}}} \int_{V'_j} \overline{\overline{\mathbf{J}_j}} \cdot \overline{\overline{\mathbf{E}_i}} \, dV, \end{aligned} \tag{6.101}$$

where S'_i is the surface enclosing antenna i only (V_p is not contained in S'_i) and V'_j is the region enclosed by S'_j, which contains both antenna j and V_p and $\overline{\overline{\mathbf{J}_j}}$ is the current distribution of antenna j. Similarly, the perturbed transmission coefficient for Scenario 2 can be expressed as (assuming that the impressed $\overline{\overline{\mathbf{J}_j}}$ remains unchanged)

$$\begin{aligned} S'_{ij} &= -\frac{1}{\overline{\overline{2a_i^{(i)} a_j^{(j)}}}} \int_{S'_i} \left(\overline{\overline{\mathbf{E}'_i}} \times \overline{\overline{\mathbf{H}'_j}} - \overline{\overline{\mathbf{E}'_j}} \times \overline{\overline{\mathbf{H}'_i}} \right) \cdot \mathbf{u}_n \, dV \\ &= -\frac{1}{\overline{\overline{2a_i^{(i)} a_j^{(j)}}}} \int_{V'_j} \overline{\overline{\mathbf{J}_j}} \cdot \overline{\overline{\mathbf{E}'_i}} \, dV. \end{aligned} \tag{6.102}$$

Subtracting (6.101) from (6.102) gives

$$S'_{ij} - S_{ij} = -\frac{1}{\overline{\overline{2a_i^{(i)} a_j^{(j)}}}} \int_{V'_j} \overline{\overline{\mathbf{J}_j}} \cdot \left(\overline{\overline{\mathbf{E}'_i}} - \overline{\overline{\mathbf{E}_i}} \right) dV. \tag{6.103}$$

Considering that V'_j contains the region V_p and the sources producing the differential fields $\Delta\overline{\overline{\mathbf{E}}} = \overline{\overline{\mathbf{E}'}} - \overline{\overline{\mathbf{E}}}$ and $\Delta\overline{\overline{\mathbf{H}}} = \overline{\overline{\mathbf{H}'}} - \overline{\overline{\mathbf{H}}}$ are given by (6.99), (6.103) may be written as

$$\begin{aligned} S'_{ij} - S_{ij} &= -\frac{1}{\overline{\overline{2a_i^{(i)} a_j^{(j)}}}} \int_{V_p} \left(\overline{\overline{\mathbf{J}'}} \cdot \overline{\overline{\mathbf{E}_j}} - \overline{\overline{\mathbf{J}'_m}} \cdot \overline{\overline{\mathbf{H}_j}} \right) dV \\ &= \frac{1}{\overline{\overline{2a_i^{(i)} a_j^{(j)}}}} \int_{V_p} \Big\{ j\omega(\mu' - \mu) \overline{\overline{\mathbf{H}'_i}} \cdot \overline{\overline{\mathbf{H}_j}} \\ &\qquad - [\sigma' - \sigma + j\omega(\varepsilon' - \varepsilon)] \overline{\overline{\mathbf{E}'_i}} \cdot \overline{\overline{\mathbf{E}_j}} \Big\} dV \end{aligned} \tag{6.104}$$

from the reciprocity theorem. This formula is useful to study the effect of changes in permittivity and permeability of the medium in a finite three dimensional region. But it is not convenient to study the changes in highly conducting bodies where the fields are confined to a shallow surface layer. In this case, a surface integral will be more suitable. Making use of reciprocity theorem again, (6.104) may be expressed as

$$\begin{aligned} S'_{ij} - S_{ij} &= -\frac{1}{\overline{\overline{2a_i^{(i)} a_j^{(j)}}}} \int_{V_p} \overline{\overline{\mathbf{J}'}} \cdot \overline{\overline{\mathbf{E}_j}} - \overline{\overline{\mathbf{J}'_m}} \cdot \overline{\overline{\mathbf{H}_j}}\, dV \\ &= -\frac{1}{\overline{\overline{2a_i^{(i)} a_j^{(j)}}}} \int_{S_p} \left[\left(\overline{\overline{\mathbf{E}'_i}} - \overline{\overline{\mathbf{E}_i}} \right) \times \overline{\overline{\mathbf{H}_j}} - \overline{\overline{\mathbf{E}_j}} \times \left(\overline{\overline{\mathbf{H}'_i}} - \overline{\overline{\mathbf{H}_i}} \right) \right] \cdot \mathbf{u}_n\, dS \\ &= \frac{1}{\overline{\overline{2a_i^{(i)} a_j^{(j)}}}} \int_{S_p} \left(\overline{\overline{\mathbf{E}_j}} \times \overline{\overline{\mathbf{H}'_i}} - \overline{\overline{\mathbf{E}'_i}} \times \overline{\overline{\mathbf{H}_j}} \right) \cdot \mathbf{u}_n\, dS. \end{aligned} \tag{6.105}$$

Note that only the field components tangential to S_p contributes to (6.105). Let Z_s and Z'_s be the surface impedances before and after the change of the medium parameters respectively. Considering the relations $\overline{\overline{\mathbf{E}_{jt}}} = Z_s \mathbf{u}_n \times \overline{\overline{\mathbf{H}_{it}}}$, $\overline{\overline{\mathbf{E}'_{jt}}} = Z'_s \mathbf{u}_n \times \overline{\overline{\mathbf{H}'_i}}$, (6.105) may be rewritten as

$$S'_{ij} - S_{ij} = \frac{1}{\overline{\overline{2a_i^{(i)} a_j^{(j)}}}} \int_{S_p} (Z'_s - Z_s) \overline{\overline{\mathbf{H}_{jt}}} \cdot \overline{\overline{\mathbf{H}'_{it}}}\, dS, \tag{6.106}$$

where the subscript t is used to represent the tangential component.

If there exist m scatterers and each scatterer occupies a region V_p ($p = 1, 2, \ldots, m$), then the integrations in (6.104)–(6.106) become a summation of integrations over each scatterer. For instance, (6.104) may be written as

$$\begin{aligned} S'_{ij} = S_{ij} + \frac{1}{\overline{\overline{2a_i^{(i)} a_j^{(j)}}}} \sum_{p=1}^{m} \int_{V_p} \Big\{ & j\omega(\mu' - \mu) \overline{\overline{\mathbf{H}'_i}} \cdot \overline{\overline{\mathbf{H}_j}} \\ & - [\sigma' - \sigma + j\omega(\varepsilon' - \varepsilon)] \overline{\overline{\mathbf{E}'_i}} \cdot \overline{\overline{\mathbf{E}_j}} \Big\} dV. \end{aligned} \tag{6.107}$$

The first term of (6.107) corresponds to the contribution due to the direct path from antenna i to antenna j. The second term represents the m multipath components introduced by the m scatterers and usually improve the condition number of the channel matrix, which is important for a wireless MIMO system to be effective.

So far our discussion is exact. If the parameters ξ_1 and ξ_2 defined by

$$\begin{cases} \xi_1 = \sigma'(\mathbf{r}) - \sigma(\mathbf{r}) + j\omega[\varepsilon'(\mathbf{r}) - \varepsilon(\mathbf{r})] \\ \xi_2 = j\omega[\mu'(\mathbf{r}) - \mu(\mathbf{r})] \end{cases}$$

are small numbers, a perturbation method may be introduced to predict S'_{ij}. In this case, the fields $\overline{\overline{\mathbf{E}'_i}}$ and $\overline{\overline{\mathbf{H}'_i}}$ may be expanded in terms of ξ_1 and ξ_2 as follows

$$\begin{cases} \overline{\overline{\mathbf{E}'_i}} = \overline{\overline{\mathbf{E}_i}} + \xi_1 \overline{\overline{\mathbf{E}_{i1}}} + \xi_2 \overline{\overline{\mathbf{E}_{i2}}} + \cdots \\ \overline{\overline{\mathbf{H}'_i}} = \overline{\overline{\mathbf{H}_i}} + \xi_1 \overline{\overline{\mathbf{H}_{i1}}} + \xi_2 \overline{\overline{\mathbf{H}_{i2}}} + \cdots \end{cases}.$$

As a first order approximation, (6.104), (6.105) and (6.106) can then be approximated by

$$S'_{ij} - S_{ij} \approx \frac{1}{2a_i^{(i)} a_j^{(j)}} \int_{V_p} \Big\{ j\omega(\mu' - \mu) \overline{\overline{\mathbf{H}_i}} \cdot \overline{\overline{\mathbf{H}_j}} - \big[\sigma' - \sigma + j\omega_c(\varepsilon' - \varepsilon)\big] \overline{\overline{\mathbf{E}_i}} \cdot \overline{\overline{\mathbf{E}_j}} \Big\} dV, \tag{6.108}$$

$$S'_{ij} - S_{ij} \approx \frac{1}{2a_i^{(i)} a_j^{(j)}} \int_{S_p} \left(\overline{\overline{\mathbf{E}_j}} \times \overline{\overline{\mathbf{H}_i}} - \overline{\overline{\mathbf{E}_i}} \times \overline{\overline{\mathbf{H}_j}} \right) \cdot \mathbf{u}_n \, dS, \tag{6.109}$$

$$S'_{ij} - S_{ij} \approx \frac{1}{2a_i^{(i)} a_j^{(j)}} \int_{S_p} (Z'_s - Z_s) \overline{\overline{\mathbf{H}_{jt}}} \cdot \overline{\overline{\mathbf{H}_{it}}} \, dS. \tag{6.110}$$

It is far better to foresee even without certainty than not to foresee at all.

—Jules Henri Poincaré

This page intentionally left blank

Chapter 7

Electromagnetic Compatibility

> One thing I have learned in a long life: that all our science, measured against reality, is primitive and childlike — and yet is the most precious thing we have.
>
> —Albert Einstein

Electromagnetic compatibility (EMC) studies the unintentional generation, transmission and reception of electromagnetic energy, and deals with the electromagnetic interferences (EMIs) or disturbance that the unintentional electromagnetic energy (as an external source) may induce. Its aim is to ensure that the electronic devices or systems will not interfere with each other's normal operation. EMC is also referred to as EMI control so that the interference effects can be prevented.

Before the electronic devices are brought to the market, they must meet the EMC standards set by national and international organizations, such as the Federal Communications Commission (FCC) in United States. Compliance with national or international standards is usually required by laws passed by individual nations. Different nations can require compliance with different standards. An electronic system is said to be compatible to its environment if it satisfies the criteria that

(1) It does not cause interference with other systems.
(2) It is not susceptible to emissions from other systems.

Therefore, the objective of EMC design for an electronic system is to suppress its emissions and to reduce its susceptibility to other incoming electromagnetic energy. The interference source, the energy coupling mechanism and the receptor constitute three basic components of an EMI problem, as illustrated in Figure 7.1.

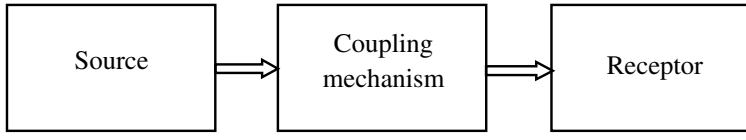


Figure 7.1 Three components of EMI problem.

The origin of the interference sources can be natural (e.g., lightening electromagnetic pulse, electrostatic discharge and solar activity) or man-made (e.g., spark gaps, power transmission lines, digital circuits, nuclear electromagnetic pulse, radio frequency transmissions), and the source may emit continuous wave at a narrow band of frequencies, or a transient wave which is usually broadband.

Basically, there are three coupling mechanisms: conductive coupling, near-field coupling and far-field coupling. Conductive coupling occurs when the coupling path between the source and the receptor is formed by direct contact with a conducting body, such as a transmission line, wire, cable, PCB trace or metal enclosure. The near field coupling refers to the situation where the interference source and the receptor are separated by a distance less than a wavelength, and can be further categorized into inductive coupling and capacitive coupling. If the stored electric (or magnetic) energy in the vicinity of emission source is higher than the stored magnetic (or electric) energy, the emission source produces more capacitive (or inductive) coupling. The far-field coupling refers to the situation where the source and the receptor are separated by a distance more than a wavelength. In this case, the source radiates an electromagnetic wave, which propagates across the open space in between and is received by the receptor. Various coupling mechanisms are illustrated in Figure 7.2.

7.1 Fields and Circuits

The circuit theories for various electromagnetic systems can be derived from the field theory. For each field quantity, there is a circuit quantity which is defined as the integral of the corresponding field quantity. Some typical circuit quantities are summarized in Table 7.1.

Note that all circuit quantities are algebraic quantities and they depend on the selection of reference direction $\mathbf{u}_l\, dl$ in the line integrals and the reference direction $\mathbf{u}_n\, dS$ in the surface integrals. The line-integral quantity (such as the voltage) is positive reference at the start of the path of integration [Figure 7.3(a)]. The surface-integral quantity (such as current)

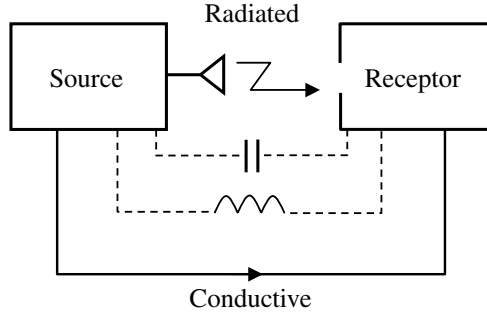


Figure 7.2 Various coupling mechanisms.

Table 7.1 Circuit quantities

Quantity	**Definition**
Voltage v **(volts)**	$v = \int_{a-b} \mathbf{E}(\mathbf{r}) \cdot \mathbf{u}_l \, dl(\mathbf{r})$
Current i **(amperes)**	$i = \int_S \mathbf{J}(\mathbf{r}) \cdot \mathbf{u}_n \, dS(\mathbf{r})$
Electric charge q **(coulombs)**	$q = \int_V \rho(\mathbf{r}) \, dV(\mathbf{r})$
Electric flux ψ_e **(coulombs)**	$\psi_e = \int_S \mathbf{D}(\mathbf{r}) \cdot \mathbf{u}_n \, dS(\mathbf{r})$
Magnetic flux ψ_m**(webers)**	$\psi_m = \int_S \mathbf{B}(\mathbf{r}) \cdot \mathbf{u}_n \, dS(\mathbf{r})$
Magnetomotive force u **(amperes)**	$u = \int_{a-b} \mathbf{H}(\mathbf{r}) \cdot \mathbf{u}_l \, dl(\mathbf{r})$

is positive reference in the direction of $\mathbf{u}_n dS$ [Figure 7.3(b)]. Charge is a net-amount quantity, which equals the amount of positive charge minus the amount of negative charge.

For a thin wire, we may choose $\mathbf{u}_l = \mathbf{u}_n$. This is called **passive sign convention**, which implies that whenever the reference direction or the current in an element is in the reference direction of voltage drop across the element [see Figure 7.3(c)], a positive sign is used in any expression that relates the voltage to the current. If we choose $\mathbf{u}_l = -\mathbf{u}_n$, we have an **active sign convention**.

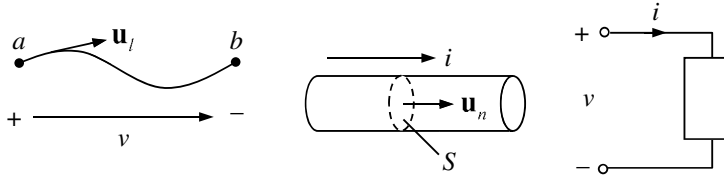


Figure 7.3 Reference convention.

7.1.1 Impressed Field and Scattered Field

The current density is related to the total field by the Ohm's law

$$\mathbf{J} = \sigma \mathbf{E}, \tag{7.1}$$

where $\mathbf{E}$ is the sum of the incident field $\mathbf{E}_{\text{in}}$ and the scattered field $\mathbf{E}_s$ produced by the charges and currents in the system

$$\mathbf{E} = \mathbf{E}_{\text{in}} + \mathbf{E}_s. \tag{7.2}$$

The scattered field $\mathbf{E}_s$ due to charges and currents in the system can then be expressed in terms of the scalar and vector potential

$$\mathbf{E}_s = -\nabla\phi - \frac{\partial \mathbf{A}}{\partial t}, \tag{7.3}$$

where

$$\begin{aligned} \phi(\mathbf{r}) &= \frac{1}{4\pi\varepsilon} \int_V \frac{\rho(\mathbf{r}')e^{-jk|r-r'|}}{|\mathbf{r}-\mathbf{r}'|} dV(\mathbf{r}'), \\ \mathbf{A}(\mathbf{r}) &= \frac{\mu}{4\pi} \int_V \frac{\mathbf{J}(\mathbf{r}')e^{jk|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|} dV(\mathbf{r}'). \end{aligned} \tag{7.4}$$

Substituting (7.1) and (7.3) into (7.2), we obtain

$$\mathbf{E}_{\text{in}} = \frac{\mathbf{J}}{\sigma} + \nabla\phi + \frac{\partial \mathbf{A}}{\partial t}. \tag{7.5}$$

This is called the **cause and effect relationship**. The incident field induces an ohmic term and terms due to the charges and currents of the system.

7.1.2 Kirchhoff's Laws

According to the continuity equation

$$\nabla \cdot \mathbf{J} = -\frac{\partial \rho}{\partial t},$$

we have

$$\int_{S_0} \mathbf{J} \cdot \mathbf{u}_n \, dS = -\frac{\partial}{\partial t} \int_{V_0} \rho \, dV, \tag{7.6}$$

where V_0 is the region bounded by a closed surface S_0 surrounding a junction of conducting wires as shown in Figure 7.4(a). The only conduction current flowing out of the surface is that in the wires, so the left side of (7.6) becomes the algebraic sum of currents flowing out in the wires. Thus (7.6) can be rewritten as

$$\sum_{n=1}^{N} i_n = -\frac{dQ}{dt}.$$

This is called **Kirchhoff's first law** or **Kirchhoff's current law**. If there is no accumulation of charge at the junction we have

$$\sum_{n=1}^{N} i_n = 0.$$

For an arbitrary circuit path shown in Figure 7.4(b), we may take the integration of (7.5) along the path

$$\int_{a-b} \mathbf{E}_{\text{in}} \cdot \mathbf{u}_l \, dl = \int_{a-b} \frac{\mathbf{J}}{\sigma} \cdot \mathbf{u}_l \, dl + \int_{a-b} \nabla\varphi \cdot \mathbf{u}_l \, dl + \int_{a-b} \frac{\partial \mathbf{A}}{\partial t} \cdot \mathbf{u}_l \, dl, \tag{7.7}$$

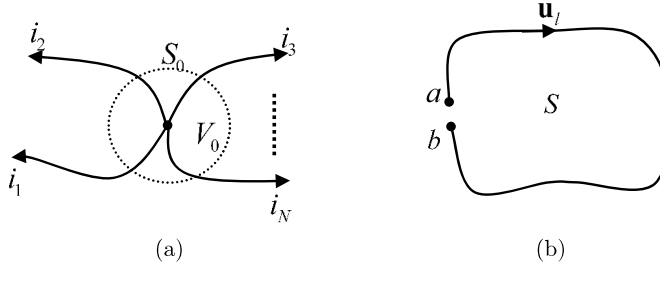


Figure 7.4 (a) Kirchhoff's current law. (b) Kirchhoff's voltage law.

where $\mathbf{u}_l$ is the current reference direction in the path, and each term on the right-hand side of (7.7) represents a voltage drop in the direction of the current flow

$$v_{\text{in}} = \int_{a-b} \mathbf{E}_{\text{in}} \cdot \mathbf{u}_l \, dl,$$

$$v_R = \int_{a-b} \frac{\mathbf{J}}{\sigma} \cdot \mathbf{u}_l \, dl,$$

$$v_C = \int_{a-b} \nabla\varphi \cdot \mathbf{u}_l \, dl,$$

$$v_L = \int_{a-b} \frac{\partial \mathbf{A}}{\partial t} \cdot \mathbf{u}_l \, dl.$$

They are respectively called applied voltage, internal impedance voltage drop, capacitive voltage drop and inductive voltage drop. Equation (7.7) is called **Kirchhoff's second law** or **Kirchhoff's voltage law**.

7.1.3 Low-frequency Approximations and Lumped Circuit Parameters

A low-frequency circuit refers to the case where the circuit size is small compared with wavelength or equivalently the following assumptions are made:

(1) Current is to be taken the same about the entire path.
(2) Retardation is to be neglected in computing the potentials $\mathbf{A}$ and ϕ.

7.1.3.1 *RLC Circuits*

Since the term $\mathbf{J}/\sigma$ gives the electric field $\mathbf{E}$, we have

$$v_R = \int_{a-b} \frac{\mathbf{J}}{\sigma} \cdot \mathbf{u}_l \, dl = i \int_{a-b} \frac{\mathbf{E}}{i} \cdot \mathbf{u}_l \, dl = i \int_{a-b} R' dl = iR,$$

where R' is the internal impedance per unit length and is defined by $R' = \mathbf{E} \cdot \mathbf{u}_l / i$, and i is the current flowing through the conductor defined by $i = \int_\Omega \mathbf{J} \cdot \mathbf{u}_l d\Omega$. Here Ω is the cross-section of the conducting path.

The vector potential in (7.4) may be written as

$$\mathbf{A}(\mathbf{r}) = \frac{\mu}{4\pi} \int_V \frac{\mathbf{J}(\mathbf{r}')dV(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}.$$

If the gap between the two ends a and b is small, we have

$$\int_{a-b} \mathbf{A} \cdot \mathbf{u}_l \, dl = \int_S \nabla \times \mathbf{A} \cdot \mathbf{u}_n \, dS = \int_S \mathbf{B} \cdot \mathbf{u}_n \, dS, \tag{7.8}$$

where S is the region bounded by the circuit path. We may define the **self-inductance** as follows

$$L = \frac{1}{i} \int_{a-b} \mathbf{A} \cdot \mathbf{u}_l \, dl = \frac{1}{i} \int_S \mathbf{B} \cdot \mathbf{u}_n \, dS. \tag{7.9}$$

If the circuit is stationary, the inductive voltage drop becomes

$$v_L = \int_{a-b} \frac{\partial \mathbf{A}}{\partial t} \cdot \mathbf{u}_l \, dl = \frac{d}{dt} \int_{a-b} \mathbf{A} \cdot \mathbf{u}_l \, dl = \frac{d}{dt}(Li) = L\frac{di}{dt}.$$

Even though the self-inductance is defined in terms of the magnetic flux and current, it is actually determined by the permeability μ of the medium and the size, length, and spacing of the conductors that form the current path. The self-inductance increases if the current loop area increases.

For a broken path, charge may accumulate on both ends a and b as shown in Figure 7.4(b). We assume that the break is rather small compared with other dimensions of the circuit. Then the capacitive voltage drop may be written as

$$v_C = \int_{a-b} \nabla\phi \cdot \mathbf{u}_l \, dl = \int_{a-b} \frac{\partial \phi}{\partial l} dl = \phi_b - \phi_a.$$

Neglecting the retardation as before, the scalar potential ϕ in (7.4) can be written as

$$\phi(\mathbf{r}) = \frac{1}{4\pi\varepsilon} \int_V \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} dV(\mathbf{r}').$$

If the stray capacitances are negligible so that all significant charge is concentrated at the discontinuity, q on one end and $-q$ on the other, the

value of ϕ will be proportional to q and so will the voltage drop $\varphi_b - \varphi_a$. The **capacitance** C is defined by

$$v_C = \frac{q}{C}.$$

The charge q may be calculated as $q = \int_{-\infty}^{t} i\,dt$ by continuity equation. As a result, (7.7) can be written as

$$v_{\text{in}} = iR + L\frac{di}{dt} + \frac{1}{C}\int_{-\infty}^{t} i\,dt, \tag{7.10}$$

and Figure 7.5 shows the equivalent circuit. The input power p_{in} into the circuit may be expressed as follows:

$$p_{\text{in}} = p_R + p_L + p_C, \tag{7.11}$$

where

$$p_{\text{in}} = v_{\text{in}} i, \quad p_R = Ri^2, \quad p_L = Li\frac{di}{dt}, \quad p_C = Cv_C\frac{dv_C}{dt}.$$

The first term on the right-hand side of (7.11) is the rate of the energy absorbed by the resistor R; the second term is the rate of magnetic energy stored in the inductor; the third term is the rate of electric energy stored in the capacitor. Equation (7.11) can be rewritten as

$$\overbrace{v_{\text{in}} i}^{p_{\text{in}}} = \overbrace{Ri^2}^{p_R} + \overbrace{\frac{dW_L}{dt}}^{p_L} + \overbrace{\frac{dW_C}{dt}}^{p_C}, \tag{7.12}$$

where

$$W_L = \frac{1}{2}Li^2, \quad W_C = \frac{1}{2}Cv_C^2 \tag{7.13}$$

are the stored electric energy and magnetic energy, respectively.

Example 7.1 (Sinusoidal excitation)**:** We assume $v_{\text{in}}(t) = V_{\text{in}}\sin(\omega t + \varphi_v)$, where V_{in} is the amplitude. The current in the RLC circuit satisfy the

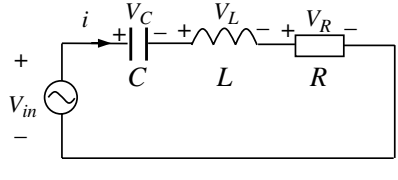


Figure 7.5 RLC circuit.

following equation and initial conditions

$$Ri(t) + L\frac{di(t)}{dt} + \frac{1}{C}q(t) = V_{\text{in}}\sin(\omega t + \varphi_v) \tag{7.14}$$

with initial conditions

$$i(0) = 0, \quad q(0) = q_0.$$

The solution of (7.14) is the sum of the steady state solution and a transient component:

$$i(t) = I(t) + e^{-\alpha t}\left\{\frac{V_d}{Lb}\frac{e^{bt} - e^{-bt}}{2} - \frac{V_{\text{in}}}{|Z(j\omega)|}\frac{e^{bt} + e^{-bt}}{2}\sin(\varphi_v - \theta)\right\}, \tag{7.15}$$

where

$$I(t) = \frac{V_{\text{in}}}{|Z(j\omega)|}\sin(\omega t + \varphi_v - \theta)$$

is the steady state component of current and

$$V_d = V_{\text{in}}\sin\phi_v - \frac{V_{\text{in}}\omega L}{|Z(j\omega)|}\cos(\varphi_v - \theta) - \frac{q_0}{C} - \frac{V_{\text{in}}R}{2\,|Z(j\omega)|}\sin(\varphi_v - \theta),$$

$$|Z(j\omega)| = \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}, \quad \theta = \tan^{-1}\left[\frac{1}{R}\left(\omega L - \frac{1}{\omega C}\right)\right],$$

$$\alpha = \frac{R}{2L}, \quad b = \sqrt{\frac{R^2}{4L^2} - \frac{1}{LC}}.$$

When the circuit is in the steady state, we have

$$P_{\text{in}} = \frac{V_{\text{in}}^2}{|Z(j\omega)|}\sin(\omega t + \varphi_v)\sin(\omega t + \varphi_v - \theta),$$

$$P_R = \frac{V_{\text{in}}^2 R}{|Z(j\omega)|^2}\sin^2(\omega t + \varphi_v - \theta),$$

$$P_L = \frac{V_{\text{in}}^2\omega L}{2\,|Z(j\omega)|^2}\sin 2(\omega t + \varphi_v - \theta),$$

$$P_C = -\frac{V_{\text{in}}^2}{2\,|Z(j\omega)|^2\,\omega C}\sin 2(\omega t + \varphi_v - \theta).$$

If the circuit is at resonance, we have

$$P_{\text{in}} = \frac{V_{\text{in}}^2}{R}\sin^2(\omega t + \varphi_v),$$

$$P_R = \frac{V_{\text{in}}^2}{R}\sin^2(\omega t + \varphi_v - \theta),$$

$$P_L = \frac{1}{2}\sqrt{\frac{L}{C}}\frac{V_{\text{in}}^2}{R^2}\sin 2(\omega t + \varphi_v - \theta),$$

$$P_C = -\frac{1}{2}\sqrt{\frac{L}{C}}\frac{V_{\text{in}}^2}{R^2}\sin 2(\omega t + \varphi_v - \theta).$$

Note that the resistor directly dissipates all the energy from the source while the inductor and capacitor exchange energy from each another when the RLC circuit is at resonance. □

7.1.3.2 *Lumped Circuit Elements*

The lumped circuit parameters R, L, and C can also be defined by the energy associated with them. The power dissipated in a region V with a current density distribution $\mathbf{J}$ is

$$P = \int_V \mathbf{J}\cdot\mathbf{E}\,dV,$$

and the resistance for the loop shown in Figure 7.4(b) is defined by

$$R = \frac{P}{i^2}. \tag{7.16}$$

Remark 7.1: In the case where the conductors have a uniform cross section with uniform flow of electric current, the **resistivity** ρ_r is defined by

$$\rho_r = R\frac{\Omega}{l}, \tag{7.17}$$

where R is the resistance of the conductor, Ω is the cross sectional area of the conductor and l is the length of the conductor. The resistivity defined this way makes resistivity a material property. The **conductivity** σ is defined as the inverse of the resistivity. □

The magnetic energy stored in the region V is

$$W_m = \int_V \frac{1}{2}\mathbf{H}\cdot\mathbf{B}\,dV, \tag{7.18}$$

and the inductance for the loop is defined by

$$L = \frac{2W_m}{i^2}. \tag{7.19}$$

The electric energy stored in the region V is

$$W_e = \int_V \frac{1}{2}\mathbf{E} \cdot \mathbf{D}\, dV, \tag{7.20}$$

and the capacitance for the loop is defined by

$$C = \frac{2W_e}{v_C^2}. \tag{7.21}$$

Note that the volume integrals in (7.18) and (7.20) must include the entire region where the fields are not zero. The inductance defined by (7.19) can be divided into **internal inductance** and **external inductance**

$$L = L_{\text{int}} + L_{\text{ext}},$$

where

$$L_{\text{int}} = \frac{2W_m^{\text{int}}}{i^2}, \quad L_{\text{ext}} = \frac{2W_m^{\text{ext}}}{i^2}, \tag{7.22}$$

with W_m^{int} and W_m^{ext} being respectively the magnetic energy stored inside and outside the conductor.

The equivalent circuit for a practical resistor is shown in Figure 7.6(a), where R is the designed value of resistance, L is the inductance including

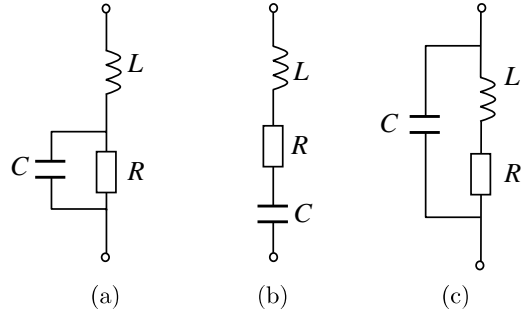


Figure 7.6 Equivalent circuits for practical elements. (a) A resistor. (b) A capacitor. (c) An inductor.

the lead inductance and the inductance from the resistor itself, and C is the shunt capacitance of the resistor. The terminal impedance of the equivalent circuit for the resistor can be written as

$$Z(\omega) = R\frac{1-(\omega/\omega_2)^2 + j(\omega\omega_1/\omega_2\omega_2)}{1+j\omega/\omega_1},$$

where

$$\omega_1 = \frac{1}{RC}, \quad \omega_2 = \frac{1}{\sqrt{LC}}. \tag{7.23}$$

Usually $\omega_1 \ll \omega_2$, and the impedance of the resistor has the following properties:

$$Z(\omega) \approx \begin{cases} R, & \text{for } \omega < \omega_1, \\ 1/j\omega C, & \text{for } \omega_1 < \omega < \omega_2, \\ j\omega L, & \text{for } \omega > \omega_2. \end{cases}$$

Thus a practical resistor behaves like a capacitor or inductor when the frequency becomes very high.

The equivalent circuit for a practical capacitor is shown in Figure 7.6(b), where C is the designed value of capacitance, L is the lead inductance, and R is the equivalent series resistance characterizing the losses in the capacitor. The terminal impedance of the equivalent circuit for the capacitor can be written as

$$Z(\omega) = \frac{1}{j\omega C}\left[1 - \left(\frac{\omega}{\omega_2}\right)^2 + j\frac{\omega R}{\omega_2^2 L}\right],$$

where ω_1 and ω_2 are given by (7.23). Usually $R/L \ll 1$, and the terminal impedance of the capacitor has the following properties:

$$Z(\omega) \approx \begin{cases} 1/j\omega C, & \text{for } \omega \ll \omega_2, \\ j\omega L, & \text{for } \omega \gg \omega_2. \end{cases}$$

The capacitor behaves like an inductor when the frequency becomes very high.

The equivalent circuit for a practical inductor is shown in Figure 7.6(c), where L is the designed value of inductance, C is the parasitic capacitance between the windings of the inductor, and R represents the losses in the wire and magnetic material. The terminal impedance of the equivalent circuit

for the inductor can be written as

$$Z(\omega) = j\omega L \frac{1 - j(\omega_1/\omega)}{1 - (\omega/\omega_2)^2 + j(\omega\omega_1/\omega_2^2)},$$

where

$$\omega_1 = \frac{R}{L}, \quad \omega_2 = \frac{1}{\sqrt{LC}}, \quad \omega_1 \leq \omega_2.$$

The impedance of the inductor has the following properties:

$$Z(\omega) \approx \begin{cases} R, & \text{for } \omega < \omega_1, \\ j\omega L, & \text{for } \omega_1 < \omega < \omega_2, \\ 1/j\omega C, & \text{for } \omega > \omega_2. \end{cases}$$

The inductor behaves like a capacitor when frequency becomes very high.

7.1.4 Mutual Coupling between Low-Frequency Circuits

The electromagnetic energy may be transferred from one circuit to another by mutual coupling such as the mutual induction in a transformer.

7.1.4.1 *Inductive Coupling*

Mutual inductive coupling occurs when the current in one branch of the circuit produces an induced field in another branch. Figure 7.7 shows two circuits coupled through inductive effects. The total electric field in circuit 1 is made up of two parts, one is generated by the current i_1 and the other part is induced by the current i_2. To compute the induced voltage in circuit 1 due to the changing current i_2 in circuit 2, we consider the inductive voltage drop

$$\int_{C_1} \frac{\partial \mathbf{A}_2}{\partial t} \cdot \mathbf{u}_{l_1} dl_1 = \frac{d}{dt} \int_{C_1} \mathbf{A}_2 \cdot \mathbf{u}_{l_1} dl_1, \tag{7.24}$$

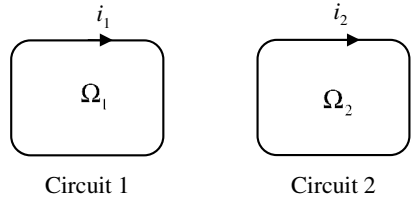


Figure 7.7 Inductive coupling.

where $\mathbf{A}_2$ is the vector potential associated with the current i_2, which, neglecting the retardation, may be written as

$$\mathbf{A}_2 = \frac{\mu i_2}{4\pi} \int_{C_2} \frac{\mathbf{u}_{l_2} dl_2}{R}. \tag{7.25}$$

The current i_2 has been taken as a constant along the path. We can define the mutual inductance by

$$M_{12} = \frac{1}{i_2} \int_{C_1} \mathbf{A}_2 \cdot \mathbf{u}_{l_1} \, dl_1 = \frac{1}{i_2} \int_{\Omega_1} \mathbf{B}_2 \cdot \mathbf{u}_n \, d\Omega. \tag{7.26}$$

Hence (7.24) can be written as

$$\int_{C_1} \frac{\partial \mathbf{A}_2}{\partial t} \cdot \mathbf{u}_l \, dl = \frac{d}{dt}(M_{12} i_2) = M_{12} \frac{di_2}{dt}.$$

Substituting (7.25) into (7.26), we obtain the Neumann form

$$M_{12} = \frac{\mu}{4\pi} \int_{C_1} \int_{C_2} \frac{dl_1 \mathbf{u}_{l_1} \cdot dl_2 \mathbf{u}_{l_2}}{R}. \tag{7.27}$$

Apparently, the following reciprocal relation holds

$$M_{12} = M_{21}$$

from (7.27).

7.1.4.2 *Capacitive Coupling*

By a mutual capacitive coupling we mean that the charge of one branch of the circuit produces an induced field in another branch. Figure 7.8 shows two circuits coupled through capacitive effects. The induced voltage drop

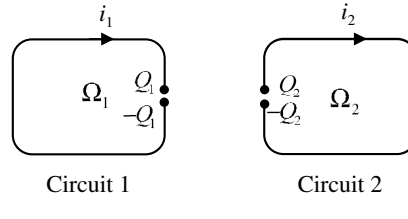


Figure 7.8 Capacitive coupling.

in circuit 1 due to charges in circuit 2 may be written as

$$\int_a^b \nabla\varphi_2 \cdot \mathbf{u}_{l_1} dl_1 = \varphi_{2b} - \varphi_{2a} = \frac{Q_2}{C_{12}}$$

where C_{12} is the mutual capacitance.

7.2 Electromagnetic Emissions and Susceptibility

Electromagnetic interference refers to the unwanted electromagnetic energy emitted by quickly changing signals in electrical circuits, which degrades or limit the effective performance of normal signals. Mobile phones, for example, may interfere with other instruments such as medical equipment and airplane controls. Integrated circuits are often a source of EMI, which usually couple their energy to larger objects such as heat sinks, circuit boards and cables to radiate significantly. The FCC requires that the radiated emissions by a digital device at distances of 3 and 10 meters from the device under test (DUT) are less than some specified limits.

Susceptibility is the sensitivity of a device's function to incoming EMI. In other words, susceptibility is the ability of the device to operate correctly in the presence of EMIs. A device which has high susceptibility has low immunity.

7.2.1 Rules for Emission Reductions

Contrary to the antenna design, EMC design endeavors to diminish the unwanted emissions and makes the unwanted emitting sources less efficient. Many circuit parts of a digital electronic device, such as long traces and various discontinuities, may radiate electromagnetic energy. It will thus be beneficial to have some rules of reducing the electromagnetic emissions in order for a device to meet the EMC standards.

The vector wave equations in time domain can be derived from Maxwell equations as follows

$$\nabla^2\mathbf{E}(\mathbf{r},t) - \mu\varepsilon\frac{\partial^2\mathbf{E}(\mathbf{r},t)}{\partial t^2} = \mathbf{S}_E(\mathbf{r},t), \tag{7.28}$$

$$\nabla^2\mathbf{H}(\mathbf{r},t) - \mu\varepsilon\frac{\partial^2\mathbf{H}(\mathbf{r},t)}{\partial t^2} = \mathbf{S}_H(\mathbf{r},t), \tag{7.29}$$

where

$$\mathbf{S}_E(\mathbf{r},t) = \mu\frac{\partial\mathbf{J}(\mathbf{r},t)}{\partial t} + \nabla\times\mathbf{J}_m(\mathbf{r},t) + \frac{1}{\varepsilon}\nabla\rho(\mathbf{r},t),$$

$$\mathbf{S}_H(\mathbf{r},t) = \varepsilon\frac{\partial\mathbf{J}_m(\mathbf{r},t)}{\partial t} - \nabla\times\mathbf{J}(\mathbf{r},t) + \frac{1}{\mu}\nabla\rho_m(\mathbf{r},t).$$

If all the sources of the fields are confined in a finite volume V, the solutions of (7.28) and (7.29) may be expressed by

$$\mathbf{E}(\mathbf{r},t) = -\int_V \frac{\mathbf{S}_E(\mathbf{r}', t-|\mathbf{r}-\mathbf{r}'|/v)}{4\pi R} dV(\mathbf{r}'), \tag{7.30}$$

$$\mathbf{H}(\mathbf{r},t) = -\int_V \frac{\mathbf{S}_H(\mathbf{r}', t-|\mathbf{r}-\mathbf{r}'|/v)}{4\pi R} dV(\mathbf{r}'), \tag{7.31}$$

where $v = 1/\sqrt{\mu\varepsilon}$. It can be seen that the contributions of the sources to the fields are not through the sources themselves but through their time and space variations. As a result, wires with concentrated or distributed loadings along their length radiate more efficiently than unloaded wires and the major contributions to the radiation fields may come from the ends of the wires. Physically, the loading and the discontinuities increase the gradient of charges along the wires.

First Rule for Emission Reduction: All unintentional current carriers should be properly designed to keep the time and space variations of the current and charge distributions on the carriers as small as possible. For example, avoiding long traces, sharp bends, sharp tips, and gaps are effective measures in reducing radiations.

The emission can also be reduced by increasing the rise-time of the pulse. To demonstrate this point, we may consider the current distribution

$$\mathbf{J}(\mathbf{r},t) = \mathbf{J}(\mathbf{r}) f(t)\delta(z), \quad \mathbf{r} \in \Omega,$$

and its radiated electric field on the z-axis (Geyi, 1996a):

$$\mathbf{E}(0,0,z,t) = -\frac{\mu}{4\pi z}\frac{df(t-z/v)}{dt}\int_\Omega \mathbf{J}(\mathbf{r}')d\Omega(\mathbf{r}').$$

The time-integrated Poynting vector of the fields is

$$\begin{aligned} S(0,0,z) &= \frac{1}{\eta}\int_{-\infty}^{\infty} |\mathbf{E}(0,0,z,t)|^2\, dt \\ &= \eta\left(\frac{1}{4\pi z v}\right)^2 \left|\int_\Omega \mathbf{J}(\mathbf{r}')d\Omega(\mathbf{r}')\right|^2 \int_{-\infty}^{\infty} \left|\frac{df(t)}{dt}\right|^2 dt. \end{aligned} \tag{7.32}$$

This indicates that the shorter the rise time of the exciting pulse, the stronger is the radiating energy. The property implies that the radiation intensity can be increased by decreasing the rise time of the pulse. Let us consider an interesting case where the exciting pulse is a modulated signal with a finite duration T and a carrier whose cycle is $T_0 = T/n$ $(n > 1)$:

$$f(t) = Ag(t)\sin\left(\frac{\pi t}{T_0}\right), \quad 0 < t < T. \tag{7.33}$$

If the energy of the pulse is normalized, i.e.,

$$\int_0^T g^2(t)dt = 1, \quad \int_0^T f^2(t)dt = 1,$$

we have $A \geq 1$. Substituting (7.33) into (7.32) gives

$$S(0,0,z) = \eta\left(\frac{1}{4\pi z v}\right)^2 \left|\int_\Omega \mathbf{J}\, d\Omega\right|^2 \cdot \left\{\left(\frac{n\pi}{T}\right)^2 + \frac{A^2}{2}\int_0^T \left[\frac{dg(t)}{dt}\right]^2 dt - \frac{A^2}{2}\int_0^T \left[\frac{dg(t)}{dt}\right]^2 \cos\left(\frac{2n\pi t}{T}\right) dt\right\}. \tag{7.34}$$

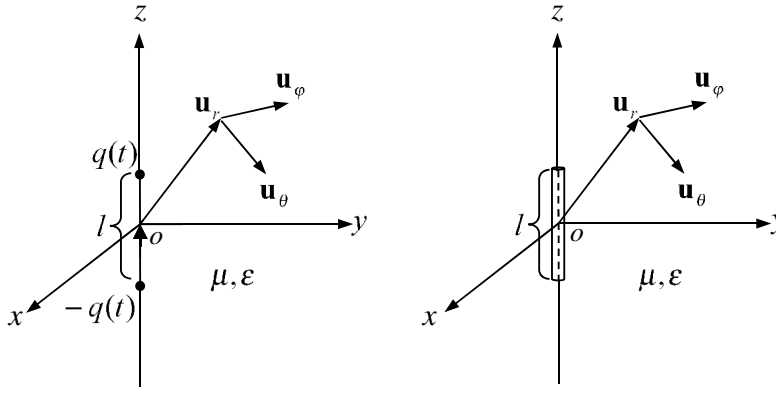
The last term in the curved brackets decreases rapidly as n increases. As a result, the time integrated energy on the z-axis increases as n increases. In other words, the energy density of the radiated electromagnetic pulse can be enhanced by increasing the frequency of the carrier. The above discussion leads to the second rule for Emission reduction.

Second Rule for Emission Reduction: The waveform of the high-speed signal should be properly designed so that the rise time of the high-speed signals is under control. A modulated signal with higher carrier frequency intends to emit more efficiently.

7.2.2 Fields of Electric Dipoles

An **electric dipole** is a system consisting of two point charges of equal magnitude $q(t)$ but opposite sign separated by a fixed distance l, as shown in Figure 7.9(a). According to the continuity equation, we have

$$\int_{\partial V_0} \mathbf{J}\cdot\mathbf{u}_n\, dS = -\frac{\partial}{\partial t}\int_{V_0} \rho\, dV,$$



(a) (b)

Figure 7.9 Electric dipole.

where V_0 is a volume enclosing one of the charges. Hence

$$i(t) = \frac{dq(t)}{dt}. \tag{7.35}$$

If we define the **current moment j** and electric **dipole moment p**

$$\mathbf{j}(t) = j(t)\mathbf{u}_z = i(t)l\mathbf{u}_z, \tag{7.36}$$

$$\mathbf{p}(t) = p(t)\mathbf{u}_z = q(t)l\mathbf{u}_z. \tag{7.37}$$

then (7.35) can be written as

$$\mathbf{j}(t) = \frac{d\mathbf{p}(t)}{dt}. \tag{7.38}$$

Equation (7.38) indicates that an electric dipole is equivalent to an electric current element $i(t)l$, as shown in Figure 7.9(b).

7.2.2.1 *Infinitesimal Electric Dipole*

An **infinitesimal dipole** or a **Hertzian dipole** is obtained when the length of the dipole approach to zero while the charges are increased to infinity so that the dipole moment remains finite. To fully understand how the dipole radiates, one should go to the time domain. In spherical coordinate system, the electromagnetic fields generated by an infinitesimal

dipole in time domain can be easily obtained as follows

$$\begin{aligned}
E_r(\mathbf{r},t) &= \frac{2\cos\theta}{4\pi\varepsilon}\left[\frac{1}{r^3}p(t_r)+\frac{1}{vr^2}\dot{p}(t_r)\right],\\
E_\theta(\mathbf{r},t) &= \frac{\sin\theta}{4\pi\varepsilon}\left[\frac{1}{r^3}p(t_r)+\frac{1}{vr^2}\dot{p}(t_r)+\frac{1}{v^2r}\ddot{p}(t_r)\right],\\
H_\varphi(\mathbf{r},t) &= \frac{\sin\theta}{4\pi}\left[\frac{1}{r^2}\dot{p}(t_r)+\frac{1}{vr}\ddot{p}(t_r)\right],\\
E_\varphi(\mathbf{r},t) &= H_r(\mathbf{r},t) = H_\theta(\mathbf{r},t) = 0,
\end{aligned} \tag{7.39}$$

where $t_r = t - r/v$. The electric field consists of three terms with the radial dependences $1/r^3$, $1/r^2$, and $1/r$, respectively. The first term $1/r^3$ is proportional to the dipole moment which predominates close to the dipole. This is the exact expression for a static dipole with the static moment replaced by time-varying moment. The second term $1/r^2$ is proportional to the time derivative of the dipole moment and the third term $1/r$ to the second derivative of the dipole moment. The radiated electromagnetic fields are

$$\begin{aligned}
E_\theta(\mathbf{r},t) &= \frac{\mu}{4\pi r}\ddot{p}(t_r)\sin\theta,\\
H_\varphi(\mathbf{r},t) &= \frac{1}{4\pi rv}\ddot{p}(t_r)\sin\theta.
\end{aligned} \tag{7.40}$$

For a small electric current element Il in frequency domain, we may write

$$\mathbf{p} = p\mathbf{u}_z = \frac{Il}{j\omega}\mathbf{u}_z, \tag{7.41}$$

and (7.39) become

$$\begin{aligned}
E_r(\mathbf{r}) &= \frac{2Il\cos\theta}{4\pi\omega\varepsilon}e^{-jkr}\left(\frac{k}{r^2}-\frac{j}{r^3}\right),\\
E_\theta(\mathbf{r}) &= \frac{Il\sin\theta}{4\pi\omega\varepsilon}e^{-jkr}\left(\frac{jk^2}{r}+\frac{k}{r^2}-\frac{j}{r^3}\right),\\
H_\varphi(\mathbf{r}) &= \frac{Il\sin\theta}{4\pi}e^{-jkr}\left(\frac{jk}{r}+\frac{1}{r^2}\right),\\
E_\varphi(\mathbf{r}) &= H_r(\mathbf{r}) = H_\theta(\mathbf{r}) = 0.
\end{aligned} \tag{7.42}$$

In the far-field region, the fields are

$$E_\theta(\mathbf{r}) = \frac{jk\eta Il\sin\theta}{4\pi r}e^{-jkr},$$
$$H_\varphi(\mathbf{r}) = \frac{jkIl\sin\theta}{4\pi r}e^{-jkr}. \tag{7.43}$$

where $\eta = \sqrt{\mu/\varepsilon}$.

Example 7.2: Two charges $+q$ and $-q$, located at the origin of the coordinate system, are suddenly separated by a distance l at $t = 0$. The dipole moment is then given by

$$\mathbf{p}(t) = U(t)ql\mathbf{u}_z,$$

where $U(t)$ is a unit step function. Substituting the above equation into (7.39) yields

$$E_r(\mathbf{r},t) = \frac{2ql\cos\theta}{4\pi\varepsilon}\left[\frac{1}{r^3}U(t_r) + \frac{1}{vr^2}\delta(t_r)\right],$$
$$E_\theta(\mathbf{r},t) = \frac{ql\sin\theta}{4\pi\varepsilon}\left[\frac{1}{r^3}U(t_r) + \frac{1}{vr^2}\delta(t_r) + \frac{1}{v^2r}\delta'(t_r)\right],$$
$$H_\varphi(\mathbf{r},t) = \frac{ql\sin\theta}{4\pi}\left[\frac{1}{r^2}\delta(t_r) + \frac{1}{vr}\delta'(t_r)\right]_{t_r=t-r/v}.$$

Note that the magnetic field only appears at $t = r/v$ and the fields are not continuous. □

To examine the energy flow around the small dipole, let us apply the time-domain Poynting theorem to a region V bounded by two spherical surfaces S_{R_1} and S_{R_2} with radius R_1 and R_2 respectively, where $R_1 < R_2$ and the smaller spherical surface encloses the dipole, as shown in Figure 7.10. Then

$$\int_{S_{R_2}} \mathbf{S}\cdot\mathbf{u}_r\,dS - \int_{S_{R_1}} \mathbf{S}\cdot\mathbf{u}_r\,dS = -\frac{\partial}{\partial t}\int_V (w_e + w_m)dV, \tag{7.44}$$

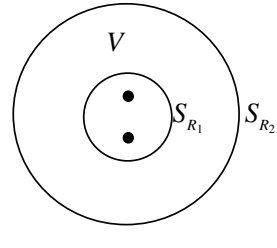


Figure 7.10 Energy flow around the small dipole.

where

$$\mathbf{S}\cdot\mathbf{u}_r = (\mathbf{E}\times\mathbf{H})\cdot\mathbf{u}_r = E_\theta H_\varphi, \quad w_e = \frac{1}{2}\varepsilon\mathbf{E}\cdot\mathbf{E}, \quad w_m = \frac{1}{2}\mu\mathbf{H}\cdot\mathbf{H}.$$

Taking the integration from t_1 and t_2, we obtain

$$\int_{t_1}^{t_2}\int_{S_{R_2}} \mathbf{S}\cdot\mathbf{u}_r\, dS\, dt - \int_{t_1}^{t_2}\int_{S_{R_1}} \mathbf{S}\cdot\mathbf{u}_r\, dS\, dt = \int_V (w_e + w_m)_{t_1} dV - \int_V (w_e + w_m)_{t_2} dV. \tag{7.45}$$

Assume that the generator emits a short pulse and then is turned off. If the pulse emitted by the dipole completely passes S_{R_1} after t_2, the second term on the left-hand side vanishes and the first term is just the radiated energy into space over the time interval $[t_1, t_2]$. The right-hand side denotes the decrease of the total stored energy over the time interval $[t_1, t_2]$ in the region V. This relationship indicates that the stored energy around the dipole is the source of the radiated energy when the generator is turned off.

7.2.2.2 *Electrically Short Dipole Antennas*

Electrically small dipole antennas can be related the electric dipole discussed above. The antenna is electrically small in the sense that the variation of the applied voltage source $v(t)$ is negligible during the time for the electromagnetic waves to travel the length of the antenna, say $2a$ as shown in Figure 7.11. Let the current distribution along the dipole antenna and the charge per unit length be noted by $J(z,t)$ and $\rho(z,t)$, respectively. At the antenna terminal, the current and charge will be noted by $J_0(t)$

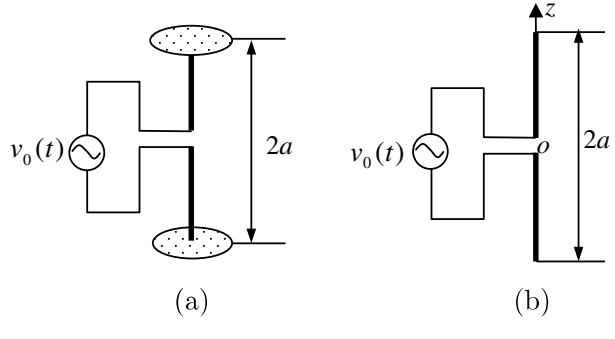


Figure 7.11 Small dipole antennas.

and $\rho_0(t)$, respectively. The total charge on one of the arm is assumed to be $Q(t)$. It is then appropriate to assume that the charges throughout the antenna change almost instantaneously in response to a change in the applied voltage. The antenna is then equivalent to a capacitor, i.e.,

$$Q(t) \approx C_a v_0(t).$$

For the top loaded antenna shown in Figure 7.11(a), we further assume that the current distribution is approximately uniform in the wires, i.e., $J(z,t) \approx J_0(t)$, which charges and discharges the plates. Based on this assumption, we have

$$\frac{\partial \rho}{\partial t} = -\frac{\partial J}{\partial z} \approx -\frac{\partial J_0}{\partial z} \approx 0,$$

which implies

$$\rho(z,t) \approx \int_{-\infty}^{t} 0\, dt = 0$$

and the total charge $Q(t)$ on the upper half of the antenna is on the upper plate. So the dipole moment for the top loaded antenna is

$$p(t) = 2aQ(t) = 2a \int_{-\infty}^{t} J_0(t')dt'. \tag{7.46}$$

For the dipole antenna shown in Figure 7.11(b), the current at two ends must be zero as there are no plates on which to deposit charge. We assume that the current distribution is a triangular

$$J(z,t) = J_0(t)\left(1 - \frac{|z|}{a}\right).$$

So we have

$$\frac{\partial \rho(z,t)}{\partial t} = -\frac{\partial J(z,t)}{\partial z} = \pm\frac{J_0(t)}{a}$$

and

$$\rho(z,t) = \pm\rho_0(t) = \frac{1}{a}\int_{-\infty}^{t} J_0(t')dt',$$

where the + sign applies for the upper arm ($z > 0$) and – sign for the lower arm ($z < 0$). The effective dipole moment is then

$$p(t) = \int_{-a}^{a} z\rho(z,t)dz = a^2\rho_0(t) = a\int_{-\infty}^{t} J_0(t')dt'. \tag{7.47}$$

Hence the radiated field of the uniform current distribution is twice that of the triangular current distribution when the current at the terminals is the same for both. When the dipoles are electrically small we can model them as an infinitesimal electric dipole and determine their fields by inserting (7.46) or (7.47) into (7.39). Note that

$$\ddot{p}(t) \propto \dot{J}_0(t) \propto \ddot{v}_0(t).$$

So the radiated field is proportional to the first derivative of current and second derivative of the applied voltage. Since the power radiated is proportional to $|\ddot{p}|^2$, the radiated power of uniform current distribution will be four times larger than the triangular distribution.

7.2.3 Fields of Magnetic Dipoles

The definition of a magnetic dipole is exactly the same as the electric dipole and we only need to replace $q(t)$ by $q_m(t)$, and $i(t)$ by $i_m(t)$. The magnetic dipole is fictitious entity due to the nonexistence of magnetic charge in nature. This theoretical conception has been proved to be very useful. Basically magnetic source is an idea from the equivalent principles. In fact, a magnetic dipole is equivalent to a small electric current loop in the sense that they both produce the same fields outside the source region. The **magnetic dipole moment** is defined by

$$\mathbf{m}(t) = m(t)\mathbf{u}_i = q_m(t)l\mathbf{u}_i. \tag{7.48}$$

It can be shown that the electromagnetic fields produced by (7.48) will be the same as that produced by a small electric current loop if we let

$$\mathbf{m}(t) = \mu i(t) S\mathbf{u}_n, \tag{7.49}$$

where S is the area of the loop, $i(t)$ is the current in the loop, and $\mathbf{u}_n$ is the unit normal whose direction is determined by the right hand rule as shown in Figure 7.12(a).

The multi-turn loop antenna is shown in Figure 7.12(b). When the loop is electrically small, the current is approximately uniform with the value $i_0(t)$. At any cross section of the multi-turn loop the total current is $ni_0(t)$, where n is the number of turns. So the magnetic dipole moment is

$$\mathbf{m}(t) = m(t)\mathbf{u}_n = \mu n i(t) S\mathbf{u}_n.$$

The electrically small loop antenna can be modeled as an infinitesimal magnetic dipole whose fields can be obtained by duality through (7.48)

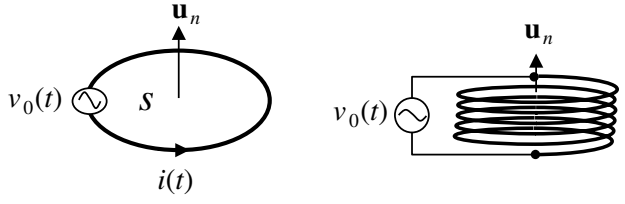


Figure 7.12 Loop antennas.

and (7.49),

$$
\begin{aligned}
E_\varphi(\mathbf{r},t) &= -\frac{\sin\theta}{4\pi}\left[\frac{1}{r^2}\dot{m}(t_r)+\frac{1}{vr}\ddot{m}(t_r)\right],\\
H_r(\mathbf{r},t) &= \frac{2\cos\theta}{4\pi\mu}\left[\frac{1}{r^3}m(t_r)+\frac{1}{vr^2}\dot{m}(t_r)\right],\\
H_\theta(\mathbf{r},t) &= \frac{\sin\theta}{4\pi\mu}\left[\frac{1}{r^3}m(t_r)+\frac{1}{vr^2}\dot{m}(t_r)+\frac{1}{v^2r}\ddot{m}(t_r)\right].
\end{aligned}
\tag{7.50}
$$

The antenna behaves essentially as an inductor L_a, thus

$$v_0(t) = L_a\frac{di(t)}{dt}.$$

The radiated field of the small loop is proportional to $\ddot{m}(t)$ given by

$$\ddot{m}(t) = \mu n\ddot{i}(t)S = \frac{\mu nS}{L_a}\dot{v}_0(t).$$

So the radiated field is proportional to the first derivative of the applied voltage or the second derivative of the current.

In frequency domain, (7.50) may be written as

$$
\begin{aligned}
E_\varphi(\mathbf{r}) &= -\frac{jm\omega\sin\theta}{4\pi}e^{-jkr}\left(\frac{jk}{r}+\frac{1}{r^2}\right),\\
H_r(\mathbf{r}) &= \frac{2m\cos\theta}{4\pi\mu}e^{-jkr}\left(\frac{jk}{r^2}+\frac{1}{r^3}\right),\\
H_\theta(\mathbf{r}) &= -\frac{jm\sin\theta}{4\pi\mu}e^{-jkr}\left(-\frac{jk^2}{r}-\frac{k}{r^2}+\frac{j}{r^3}\right).
\end{aligned}
\tag{7.51}
$$

7.2.4 Emissions from Common Mode Current and Differential Mode Current

Consider two parallel conducting wires as shown in Figure 7.13. The two wires carry currents I_1 and I_2 respectively. Then we may write

$$I_1 = I_c + I_d, \quad I_2 = I_c - I_d$$

where

$$I_c = \frac{1}{2}(I_1 + I_2), \quad I_d = \frac{1}{2}(I_1 - I_2)$$

are called **common mode current** and **differential mode current**, respectively. When the two wires are balanced we have $I_1 = -I_2$ and the common mode current does not occur. With a clamp-on current probe that encloses both wires, the differential mode current would read zero while the common mode current would give a non-zero reading. The desired signal currents exist only in the differential mode while the noise current may exist in either the differential or common mode forms. The common and differential mode emissions can be studied by the theory of array antennas. Consider two z-directed current elements $I_1\,dl$ and $I_2\,dl$ separated by a distance d as illustrated in Figure 7.14. Assuming $I_1 = I\angle 0°, I_2 = I\angle\alpha$, the array factor for the two-element array is then given by (see Section 5.11)

$$\text{AF} = 2Ie^{j\frac{\alpha}{2}} \cos\left(\frac{\alpha}{2} - \frac{kd}{2}\sin\theta\cos\varphi\right).$$

The far-field produced by the two current elements can be expressed as

$$E_\theta(\mathbf{r}) = \frac{j2\omega\mu I\,dl\sin\theta}{4\pi}\frac{e^{-jkr}}{r}e^{j\frac{\alpha}{2}}\cos\left(\frac{\alpha}{2} - \frac{kd}{2}\sin\theta\cos\varphi\right). \tag{7.52}$$

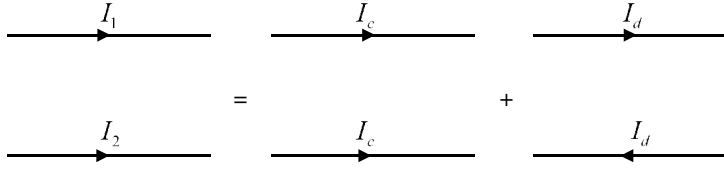


Figure 7.13 Common mode current and differential mode current.

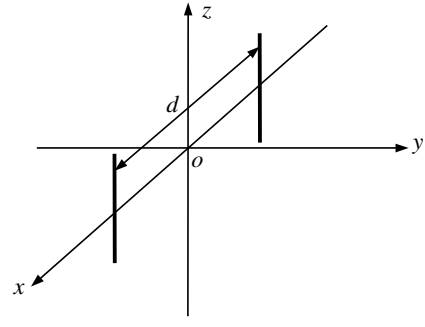


Figure 7.14 Two parallel current elements.

For two differential mode current elements, we may let $I = I_d, \alpha = \pi$ in (7.52) to get the far-field expression for the two-element system as follows:

$$E_{d\theta}(\mathbf{r}) = -\frac{2\omega\mu I_d\, dl}{4\pi}\frac{e^{-jkr}}{r}\sin\theta\sin\left(\frac{kd}{2}\sin\theta\cos\varphi\right). \tag{7.53}$$

For two common mode current elements, we may let $I = I_c, \alpha = 0$ in (7.52) to find the far-field expression for the two-element system as follows

$$E_{c\theta}(\mathbf{r}) = \frac{j2\omega\mu I_c\, dl}{4\pi}\frac{e^{-jkr}}{r}\sin\theta\cos\left(\frac{kd}{2}\sin\theta\cos\varphi\right). \tag{7.54}$$

Note that

$$\left|\frac{E_{d\theta}(\mathbf{r})}{E_{c\theta}(\mathbf{r})}\right| = \left|\frac{I_d}{I_c}\tan\left(\frac{kd}{2}\sin\theta\cos\varphi\right)\right|. \tag{7.55}$$

7.2.5 Multi-Conductor Transmission Line Models for Susceptibility

An electronic device must comply with the regulatory limits on radiated emissions and also must be insensitive to other interferences to ensure a reliable operation. In many situations, the internal circuits of electronic devices can be modeled as multi-conductor transmission lines. The voltage induced by an external incident field $\mathbf{E}_{\text{in}}$ at the terminations of the transmission line can be used to estimate the susceptibility. The external incident field $\mathbf{E}_{\text{in}}$ may be in the form of uniform plane waves generated by a distant radiator or non-uniform waves generated by a nearby radiator. The effects of the external sources can be incorporated into transmission line equation as distributed sources along the line (Paul, 2006).

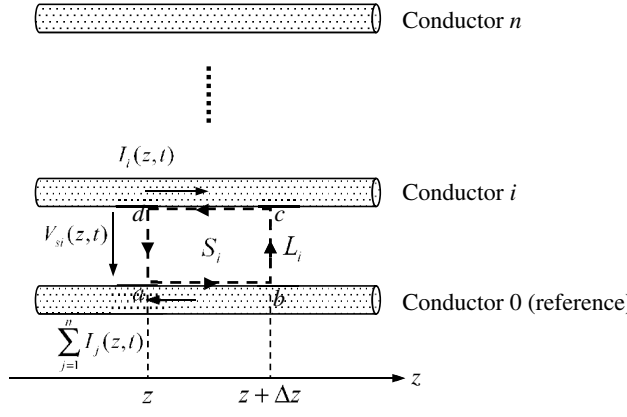


Figure 7.15 Multi-conductor transmission line.

Consider a multi-conductor transmission line consisting of $n+1$ uniform conductors parallel to the z-axis, as illustrated in Figure 7.15. We may draw a rectangular region S_i bounded by L_i between the reference conductor and conductor i. It follows from Maxwell equations and Stokes theorem that

$$\int_{S_i} \nabla \times \mathbf{E} \cdot \mathbf{u}_b \, dS = \int_{L_i} \mathbf{E} \cdot \mathbf{u}_l \, dl = -\frac{\partial}{\partial t} \int_{S_i} \mathbf{B} \cdot \mathbf{u}_b \, dS, \tag{7.56}$$

where $\mathbf{u}_l$ is the unit tangent vector along L_i in the anti-clockwise direction, and $\mathbf{u}_b$ is the unit vector pointing out of page. The above equation can be written as

$$\int_{ab} \mathbf{E} \cdot \mathbf{u}_l \, dl + \int_{bc} \mathbf{E} \cdot \mathbf{u}_l \, dl + \int_{cd} \mathbf{E} \cdot \mathbf{u}_l \, dl + \int_{da} \mathbf{E} \cdot \mathbf{u}_l \, dl$$
$$= -\frac{\partial}{\partial t} \int_{S_i} \mathbf{B} \cdot \mathbf{u}_b \, dS, \quad (i = 1, 2, \ldots, n). \tag{7.57}$$

The total field may be decomposed as the sum of the incident field and the scattered field

$$\mathbf{E} = \mathbf{E}_{\text{in}} + \mathbf{E}_s, \quad \mathbf{B} = \mathbf{B}_{\text{in}} + \mathbf{B}_s.$$

The scattered field $\mathbf{E}_s$ is generated by the induced currents and charges on the line conductors. If we assume that the currents on the line conductors are z-directed, the scattered magnetic field $\mathbf{B}_s$ will be transverse to the z-direction. As a result, a voltage corresponding to the scattered field may

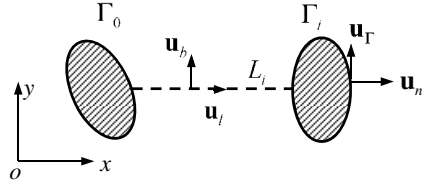


Figure 7.16 Cross section of multi-conductor transmission line.

be uniquely defined between the conductor i and the reference conductor

$$V_{si}(z,t) = -\int_{ad} \mathbf{E}_s \cdot \mathbf{u}_l \, dl, \quad V_{si}(z+\Delta z,t) = -\int_{bc} \mathbf{E}_s \cdot \mathbf{u}_l \, dl. \tag{7.58}$$

Referring to Figure 7.16, the current on the ith conductor is defined by the line integral of surface current $\mathbf{J}_s = \mathbf{u}_n \times \mathbf{H}$

$$I_i(z,t) = \int_{\Gamma_i} \mathbf{J}_s \cdot \mathbf{u}_z \, d\Gamma = \int_{\Gamma_i} \mathbf{H} \cdot \mathbf{u}_\Gamma \, d\Gamma, \tag{7.59}$$

where Γ_i is the boundary of ith conductor. Since the scattered magnetic field is transverse we may write

$$\lim_{\Delta z \to 0} \frac{1}{\Delta z} \int_{S_i} \mathbf{B}_s \cdot \mathbf{u}_b \, dS = -\sum_{j=1}^{n} l_{ij} I_j(z,t), \tag{7.60}$$

where l_{ij} $(j = 1, 2, \ldots, n)$ are inductances per unit length. The resistance per unit length r_i on the lines is defined by

$$\int_{ba} \mathbf{E} \cdot \mathbf{u}_l \, dl = r_0 \Delta z \sum_{j=1}^{n} I_j(z,t), \quad \int_{dc} \mathbf{E} \cdot \mathbf{u}_l \, dl = r_i \Delta z I_i(z,t). \tag{7.61}$$

Substituting (7.58), (7.60) and (7.61) into (7.57), we obtain

$$\sum_{j=1}^{n} r_0 I_j(z,t) + r_i I_i(z,t) + \frac{V_{si}(z+\Delta z,t) - V_{si}(z,t)}{\Delta z} + \sum_{j=1}^{n} l_{ij} \frac{\partial}{\partial t} I_j(z,t)$$
$$= \frac{1}{\Delta z} \frac{\partial}{\partial t} \int_{S_i} \mathbf{B}_{\text{in}} \cdot \mathbf{u}_b \, dS + \frac{1}{\Delta z} \int_{bc} \mathbf{E}_{\text{in}} \cdot \mathbf{u}_l \, dl + \frac{1}{\Delta z} \int_{da} \mathbf{E}_{\text{in}} \cdot \mathbf{u}_l \, dl,$$
$$i = 1, 2, \ldots, n. \tag{7.62}$$

As $\Delta z \to 0$, this becomes

$$\frac{\partial}{\partial z} V_{si}(z,t) + \sum_{j=1}^{n} r_0 I_j(z,t) + r_i I_i(z,t) + \frac{\partial}{\partial t} \sum_{j=1}^{n} l_{ij} I_j(z,t)$$

$$= \frac{\partial}{\partial t} \int_{ad} \mathbf{B}_{\text{in}} \cdot \mathbf{u}_b \, dl + \frac{\partial}{\partial z} \int_{ad} \mathbf{E}_{\text{in}} \cdot \mathbf{u}_l \, dl, \quad i = 1, 2, \ldots, n. \tag{7.63}$$

Note that (7.57) also holds for the incident fields:

$$\int_{ab} \mathbf{E}_{\text{in}} \cdot \mathbf{u}_l \, dl + \int_{bc} \mathbf{E}_{\text{in}} \cdot \mathbf{u}_l \, dl + \int_{cd} \mathbf{E}_{\text{in}} \cdot \mathbf{u}_l \, dl + \int_{da} \mathbf{E}_{\text{in}} \cdot \mathbf{u}_l \, dl$$

$$= -\frac{\partial}{\partial t} \int_{S_i} \mathbf{B}_{\text{in}} \cdot \mathbf{u}_b \, dS, \tag{7.64}$$

which yields

$$-\frac{\partial}{\partial t} \int_{S_i} \mathbf{B}_{\text{in}} \cdot \mathbf{u}_b \, dS - \int_{bc} \mathbf{E}_{\text{in}} \cdot \mathbf{u}_l \, dl - \int_{da} \mathbf{E}_{\text{in}} \cdot \mathbf{u}_l \, dl$$

$$= -\Delta z \, \mathbf{E}_{\text{in}}(z,t) \cdot \mathbf{u}_z|_{\text{conductor}\, i} + \Delta z \, \mathbf{E}_{\text{in}}(z,t) \cdot \mathbf{u}_z|_{\text{conductor}\, 0}. \tag{7.65}$$

As $\Delta z \to 0$, (7.65) becomes

$$\frac{\partial}{\partial t} \int_{ad} \mathbf{B}_{\text{in}} \cdot \mathbf{u}_b \, dl + \frac{\partial}{\partial z} \int_{ad} \mathbf{E}_{\text{in}} \cdot \mathbf{u}_l \, dl$$

$$= \mathbf{E}_{\text{in}}(z,t) \cdot \mathbf{u}_z|_{\text{conductor}\, i} - \mathbf{E}_{\text{in}}(z,t) \cdot \mathbf{u}_z|_{\text{conductor}\, 0}. \tag{7.66}$$

Thus (7.63) can be rewritten as

$$\frac{\partial}{\partial z} V_{si}(z,t) + \sum_{j=1}^{n} r_0 I_j(z,t) + r_i I_i(z,t) + \frac{\partial}{\partial t} \sum_{j=1}^{n} l_{ij} I_j(z,t)$$

$$= \mathbf{E}_{\text{in}}(z,t) \cdot \mathbf{u}_z|_{\text{conductor}\, i} - \mathbf{E}_{\text{in}}(z,t) \cdot \mathbf{u}_z|_{\text{conductor}\, 0}, \quad i = 1, 2, \ldots, n. \tag{7.67}$$

The above equation can be written in matrix form as

$$\frac{\partial}{\partial z} [V_s] + [R][I] + [L] \frac{\partial}{\partial t} [I] = [\Delta \mathbf{E}_{\text{in}} \cdot \mathbf{u}_z] \tag{7.68}$$

where

$$[V_s] = [V_{s1}, V_{s2}, \ldots, V_{sn}]^T, \quad [I] = [I_1, I_2, \ldots, I_n]^T,$$

$$[\Delta \mathbf{E}_{\text{in}} \cdot \mathbf{u}_z] = \begin{bmatrix} \mathbf{E}_{\text{in}}(z,t) \cdot \mathbf{u}_z|_{\text{conductor}\,1} - \mathbf{E}_{\text{in}}(z,t) \cdot \mathbf{u}_z|_{\text{conductor}\,0} \\ \mathbf{E}_{\text{in}}(z,t) \cdot \mathbf{u}_z|_{\text{conductor}\,2} - \mathbf{E}_{\text{in}}(z,t) \cdot \mathbf{u}_z|_{\text{conductor}\,0} \\ \vdots \\ \mathbf{E}_{\text{in}}(z,t) \cdot \mathbf{u}_z|_{\text{conductor}\,n} - \mathbf{E}_{\text{in}}(z,t) \cdot \mathbf{u}_z|_{\text{conductor}\,0} \end{bmatrix},$$

$$[R] = \begin{bmatrix} r_0 + r_1 & r_0 & \cdots & r_0 \\ r_0 & r_0 + r_2 & \cdots & r_0 \\ \vdots & \vdots & \ddots & \vdots \\ r_0 & r_0 & \cdots & r_0 + r_n \end{bmatrix}, \quad [L] = \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1n} \\ l_{21} & l_{22} & \cdots & l_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n1} & \cdots & l_{nn} \end{bmatrix}. \tag{7.69}$$

We now enclose the ith conductor with a cylinder of length Δz, as illustrated in Figure 7.17. The side surface of the ith cylinder is denoted by $S_{\rho i}$ and the two ends are denoted by S_{zi}. It follows from the continuity equation

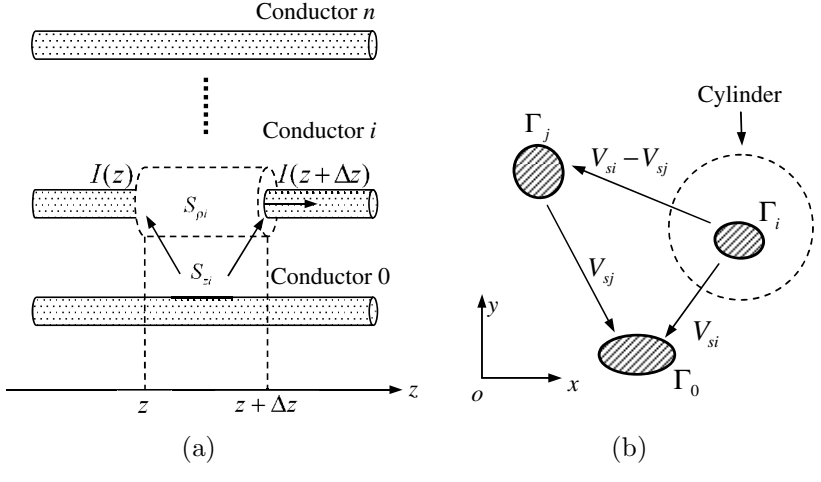


Figure 7.17 (a) Multi-conductor transmission line. (b) Cross-section of multi-conductor transmission line.

that

$$\int\limits_{S_{\rho i}+S_{zi}} \mathbf{J}\cdot\mathbf{u}_n\,dS = -\frac{\partial Q_i}{\partial t}, \quad i=1,2,\ldots,n, \tag{7.70}$$

where $\mathbf{u}_n$ is the unit outward normal of the cylinder, and Q_i is the net charge contained in the cylinder. Evidently, we have

$$\int\limits_{S_{zi}} \mathbf{J}\cdot\mathbf{u}_n\,dS = I_i(z+\Delta z,t) - I_i(z,t). \tag{7.71}$$

The transverse conduction current $I_{ti}(z,t)$ between the ith conductor and all other conductors is given by

$$\begin{aligned} I_{ti}(z,t) &= \lim_{\Delta z\to 0}\frac{1}{\Delta z}\int\limits_{S_{\rho i}} \mathbf{J}\cdot\mathbf{u}_n\,dS \\ &= g_{i1}(V_{si}-V_{s1}) + \cdots + g_{ii}V_{si} + \cdots + g_{in}(V_{si}-V_{sn}) \\ &= -g_{i1}V_{s1} - \cdots + V_{si}\sum_{j=1}^{n} g_{ij} - \cdots - g_{in}V_{sn}, \end{aligned} \tag{7.72}$$

where g_{ij} is the conductance per unit length between the ith conductor line and jth conductor line. The net charge per unit length can be expressed as

$$\begin{aligned} \lim_{\Delta z\to 0}\frac{Q_i}{\Delta z} &= c_{i1}(V_{si}-V_{s1}) + \cdots + c_{ii}V_{si} + \cdots + c_{in}(V_{si}-V_{sn}) \\ &= -c_{i1}V_{s1} - \cdots + V_{si}\sum_{j=1}^{n} c_{ij} - \cdots - c_{in}V_{sn}. \end{aligned} \tag{7.73}$$

Substituting (7.71)–(7.73) into (7.70) yields

$$\begin{aligned} &\frac{\partial}{\partial z}I_i(z,t) - g_{i1}V_{s1}(z,t) - \cdots + V_{si}(z,t)\sum_{j=1}^{n} g_{ij} - \cdots - g_{in}V_{sn}(z,t) \\ &\quad + \frac{\partial}{\partial t}\left[-c_{i1}V_{s1}(z,t) - \cdots + V_{si}(z,t)\sum_{j=1}^{n} c_{ij} - \cdots - c_{in}V_{sn}(z,t)\right], \\ &\qquad\qquad i=1,2,\ldots,n. \end{aligned} \tag{7.74}$$

This can be written in matrix form as

$$\frac{\partial}{\partial z}[I] + [G][V_s] + [C]\frac{\partial}{\partial t}[V_s] = 0, \tag{7.75}$$

where

$$[G] = \begin{bmatrix} \sum_{j=1}^{n} g_{1j} & -g_{12} & \cdots & -g_{1n} \\ -g_{21} & \sum_{j=1}^{n} g_{2j} & \cdots & -g_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -g_{n1} & -g_{n1} & \cdots & \sum_{j=1}^{n} g_{nj} \end{bmatrix},$$

$$[C] = \begin{bmatrix} \sum_{j=1}^{n} c_{1j} & -c_{12} & \cdots & -c_{1n} \\ -c_{21} & \sum_{j=1}^{n} c_{2j} & \cdots & -c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -c_{n1} & -c_{n1} & \cdots & \sum_{j=1}^{n} c_{nj} \end{bmatrix}. \tag{7.76}$$

The multi-conductor transmission line is characterized by (7.68) and (7.75), which are summarized below:

$$\begin{aligned} &\frac{\partial}{\partial z}[V_s] + [R][I] + [L]\frac{\partial}{\partial t}[I] = [\Delta \mathbf{E}_{\text{in}} \cdot \mathbf{u}_z], \\ &\frac{\partial}{\partial z}[I] + [G][V_s] + [C]\frac{\partial}{\partial t}[V_s] = 0. \end{aligned} \tag{7.77}$$

These are a set of $2n$ first-order partial differential equations, which are coupled by resistance matrix $[R]$, conductance matrix $[G]$, inductance matrix $[L]$, and the capacitance matrix $[C]$. All these matrices consist of parameters per unit length. The scattered voltages can be replaced by the total voltages through the relation

$$V_i(z,t) = V_{si} - \int_{ad} \mathbf{E}_{\text{in}} \cdot \mathbf{u}_l \, dl \Big|_{\text{conductor } i}. \tag{7.78}$$

Substituting this into (7.77) gives

$$\begin{aligned}
&\frac{\partial}{\partial z}[V] + [R][I] + [L]\frac{\partial}{\partial t}[I] = [\Delta \mathbf{E}_{\text{in}} \cdot \mathbf{u}_z] + \frac{\partial}{\partial z}[V_{\text{in}}], \\
&\frac{\partial}{\partial z}[I] + [G][V] + [C]\frac{\partial}{\partial t}[V] = [G][V_{\text{in}}] + [C]\frac{\partial}{\partial t}[V_{\text{in}}].
\end{aligned} \tag{7.79}$$

where

$$[V] = [V_1, V_2, \ldots, V_n]^T,$$

$$[V_{\text{in}}] = \left[-\int_{ad} \mathbf{E}_{\text{in}} \cdot \mathbf{u}_l \, dl \Big|_{\text{conductor 1}}, \quad -\int_{ad} \mathbf{E}_{\text{in}} \cdot \mathbf{u}_l \, dl \Big|_{\text{conductor 2}}, \ldots, -\int_{ad} \mathbf{E}_{\text{in}} \cdot \mathbf{u}_l \, dl \Big|_{\text{conductor } n} \right]^T.$$

Example 7.3: Consider a two-wire transmission line of length l illuminated by an incident field $(\mathbf{E}_{\text{in}}, \mathbf{H}_{\text{in}})$, as shown in Figure 7.18. The transmission line lies in the (x, z)-plane and is terminated with impedances Z_s and Z_L at both ends. For this arrangement, (7.79) reduces to

$$\begin{aligned}
&\frac{\partial}{\partial z}V(z,t) + L\frac{\partial}{\partial t}I(z,t) + RI(z,t) \\
&\quad = \mathbf{E}_{\text{in}}(z,t) \cdot \mathbf{u}_z|_{\text{conductor 1}} - \mathbf{E}_{\text{in}}(z,t) \cdot \mathbf{u}_z|_{\text{conductor 0}} - \frac{\partial}{\partial z}\int_0^d \mathbf{E}_{\text{in}} \cdot \mathbf{u}_z \, dx, \\
&\frac{\partial}{\partial z}I(z,t) + C\frac{\partial}{\partial t}V(z,t) + GV(z,t) \\
&\quad = -C\frac{\partial}{\partial t}\int_0^d \mathbf{E}_{\text{in}} \cdot \mathbf{u}_z \, dx - G\int_0^d \mathbf{E}_{\text{in}} \cdot \mathbf{u}_z \, dx,
\end{aligned} \tag{7.80}$$

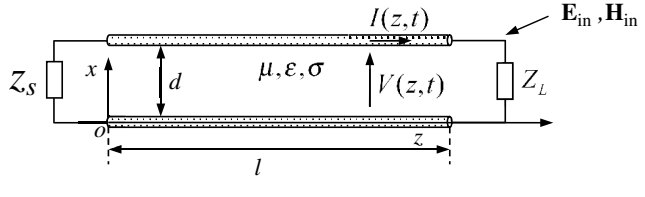


Figure 7.18 Two-wire transmission line illuminated by incident field ($l \gg d$).

where L and C are inductance and capacitance per unit length for the two-wire line. In terms of (7.66), we may rewrite the above equations as

$$\begin{aligned}
&\frac{\partial}{\partial z}V(z,t)+L\frac{\partial}{\partial t}I(z,t)+RI(z,t)=-\frac{\partial}{\partial t}\int_0^d \mathbf{B}_{\text{in}}\cdot\mathbf{u}_y\,dx,\\
&\frac{\partial}{\partial z}I(z,t)+C\frac{\partial}{\partial t}V(z,t)+GV(z,t)=-C\frac{\partial}{\partial t}\int_0^d \mathbf{E}_{\text{in}}\cdot\mathbf{u}_z\,dx-G\int_0^d \mathbf{E}_{\text{in}}\cdot\mathbf{u}_z\,dx
\end{aligned} \tag{7.81}$$

with the boundary conditions given by

$$V(0,t)=I(0,t)Z_s,\quad V(l,t)=I(l,t)Z_L. \tag{7.82}$$

It follows from (7.81) that

$$\begin{aligned}
&\frac{\partial^2}{\partial z^2}V(z,t)-LC\frac{\partial^2}{\partial t^2}V(z,t)-(LG+RC)\frac{\partial}{\partial t}V(z,t)-RGV(z,t)\\
&\quad=-\frac{\partial^2}{\partial z\partial t}\int_0^d \mathbf{B}_{\text{in}}\cdot\mathbf{u}_y\,dx+LC\frac{\partial^2}{\partial t^2}\int_0^d \mathbf{E}_{\text{in}}\cdot\mathbf{u}_z\,dx\\
&\qquad+(LG+RC)\frac{\partial}{\partial t}\int_0^d \mathbf{E}_{\text{in}}\cdot\mathbf{u}_z\,dx+RG\int_0^d \mathbf{E}_{\text{in}}\cdot\mathbf{u}_z\,dx.
\end{aligned} \tag{7.83}$$

This is the modified Klein–Gordon equation and its solution has been discussed in Chapter 2. □

7.3 Electromagnetic Coupling through Apertures

Electromagnetic coupling through apertures in conductors, intentional or unintentional, is widely encountered in microwave engineering and has been extensively investigated by many authors. The apertures are often intentionally introduced to couple the electromagnetic energy from one part to another in electronic devices, such as the coupling from waveguide to waveguide, waveguide to cavity, and cavity to cavity. Other desirable aperture coupling includes aperture antennas, slot antennas and microstrip patch antennas. Unintentional coupling often appears as leakage through apertures from electronic devices such as ventilation holes, visual access windows, and cracks around doors, and must be minimized or eliminated in the EMC designs. The rigorous analysis of field coupling through

apertures is very difficult in general. Many applications are based on the results from the coupling through an aperture in a planar conducting screen.

7.3.1 Coupling through Arbitrary Apertures

Consider a planar perfectly conducting screen S of infinite extent at $z = 0$, which is perforated with a finite aperture A, as shown in Figure 7.19(a) and 7.19(b). A current source $\mathbf{J}$ is assumed in Region 1 ($z < 0$), which produces fields $\mathbf{E}, \mathbf{H}$. When the aperture is closed (i.e., absent), the fields generated by the current $\mathbf{J}$ are denoted by $\mathbf{E}_{\text{in}}, \mathbf{H}_{\text{in}}$. According to equivalence principle, the fields in Region 1 may be determined by an equivalent magnetic current $\mathbf{J'}_{ms} = \mathbf{u}_z \times \mathbf{E}$ spread over the aperture with Region 2 filled with a perfect conductor and the original source $\mathbf{J}$,

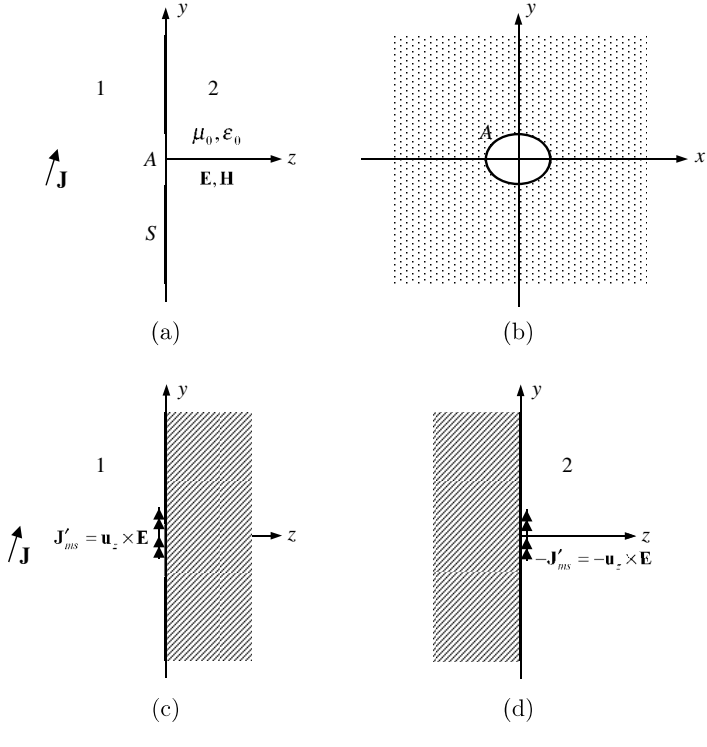


Figure 7.19 Aperture in planar conducting screen.

as illustrated in Figure 7.19(c). The fields in Region 1 can be written as

$$\begin{aligned}\mathbf{E}_1(\mathbf{r}) &= \mathbf{E}_{\text{in}}(\mathbf{r}) + \mathbf{E}_{s1}(\mathbf{r}),\\ \mathbf{H}_1(\mathbf{r}) &= \mathbf{H}_{\text{in}}(\mathbf{r}) + \mathbf{H}_{s1}(\mathbf{r}),\end{aligned} \tag{7.84}$$

where $\mathbf{E}_{s1}(\mathbf{r})$ and $\mathbf{H}_{s1}(\mathbf{r})$ are the fields generated by the equivalent magnetic current $\mathbf{J'}_{ms}$. By means of the image principle, we may write

$$\begin{aligned}\mathbf{E}_{s1}(\mathbf{r}) &= -\nabla \times \frac{1}{4\pi}\int_A \mathbf{J}_{ms}(\mathbf{r}')\frac{e^{-jkR}}{R}dS(\mathbf{r}')\\ &= -\frac{1}{4\pi}\int_A \mathbf{J}_{ms}(\mathbf{r}') \times \nabla'\frac{e^{-jkR}}{R}dS(\mathbf{r}'),\end{aligned} \tag{7.85}$$

$$\begin{aligned}\mathbf{H}_{s1}(\mathbf{r}) &= \frac{1}{j4\pi\omega\mu_0}\nabla \times \nabla \times \int_A \mathbf{J}_{ms}(\mathbf{r}')\frac{e^{-jkR}}{R}dS(\mathbf{r}')\\ &= -\frac{1}{4\pi\mu_0}\int_A \rho_{ms}\nabla\frac{e^{-jkR}}{R}dS(\mathbf{r}') - \frac{j\omega\varepsilon_0}{4\pi}\int_A \mathbf{J}_{ms}(\mathbf{r}')\frac{e^{-jkR}}{R}dS(\mathbf{r}'),\end{aligned} \tag{7.86}$$

where $\mathbf{J}_{ms} = 2\mathbf{u}_z \times \mathbf{E}$, $R = |\mathbf{r} - \mathbf{r}'|$. The fields in Region 2 can be determined by the source $-\mathbf{J'}_{ms} = -\mathbf{u}_z \times \mathbf{E}$ (the $-$ sign ensures that the tangential electrical field is continuous across the aperture). In a similar way, we have

$$\begin{aligned}\mathbf{E}_2(\mathbf{r}) &= \nabla \times \frac{1}{4\pi}\int_A \mathbf{J}_{ms}(\mathbf{r}')\frac{e^{-jkR}}{R}dS(\mathbf{r}')\\ &= \frac{1}{4\pi}\int_A \mathbf{J}_{ms}(\mathbf{r}') \times \nabla'\frac{e^{-jkR}}{R}dS(\mathbf{r}'),\end{aligned} \tag{7.87}$$

$$\begin{aligned}\mathbf{H}_2(\mathbf{r}) &= -\frac{1}{j4\pi\omega\mu_0}\nabla \times \nabla \times \int_A \mathbf{J}_{ms}(\mathbf{r}')\frac{e^{-jkR}}{R}dS(\mathbf{r}')\\ &= \frac{1}{4\pi\mu_0}\int_A \rho_{ms}\nabla\frac{e^{-jkR}}{R}dS(\mathbf{r}') + \frac{j\omega\varepsilon_0}{4\pi}\int_A \mathbf{J}_{ms}(\mathbf{r}')\frac{e^{-jkR}}{R}dS(\mathbf{r}').\end{aligned} \tag{7.88}$$

Now the tangential magnetic field must be continuous across the aperture. So we obtain the integro-differential equation for $\mathbf{J}_{ms}(\mathbf{r})$

$$\frac{1}{2\pi\mu_0}\nabla\int_A \rho_{ms}(\mathbf{r}')\frac{e^{-jkR}}{R}dS(\mathbf{r}') + \frac{j\omega\varepsilon_0}{2\pi}\int_A \mathbf{J}_{ms}(\mathbf{r}')\frac{e^{-jkR}}{R}dS(\mathbf{r}') = \mathbf{H}_{\text{in}}(\mathbf{r}). \tag{7.89}$$

The electric field must also be continuous across the aperture. This gives

$$\frac{1}{2\pi}\nabla\times\int_A \mathbf{J}_{ms}(\mathbf{r}')\frac{e^{-jkR}}{R}dS(\mathbf{r}') = \mathbf{E}_{\text{in}}(\mathbf{r}). \tag{7.90}$$

Equation (7.89) or (7.90) can be solved numerically. Once $\mathbf{J}_{ms}(\mathbf{r})$ is known, the fields can be determined from (7.85)–(7.88).

7.3.2 Coupling through Small Apertures

When the aperture is small compared with the wavelength an approximation solution to the electromagnetic coupling may be developed along the same line of thought as Bethe's early work (Bethe, 1944). For a small aperture with $kR \ll 1$, we may use the Taylor expansions

$$\frac{e^{-jkR}}{R} = \frac{1}{R} - jk - \frac{k^2R}{2} + \frac{jk^3R^2}{6} + \cdots,$$

$$\nabla\frac{e^{-jkR}}{R} = \nabla\frac{1}{R} - \frac{k^2}{2}\mathbf{u}_R + \frac{jk^3R}{3}\mathbf{u}_R + \cdots,$$

where $\mathbf{u}_R = (\mathbf{r}-\mathbf{r}')/R$. Thus we have

$$\int_A \mathbf{J}_{ms}(\mathbf{r}')\frac{e^{-jkR}}{R}dS(\mathbf{r}')$$

$$= \int_A \frac{\mathbf{J}_{ms}(\mathbf{r}')}{R}\,dS(\mathbf{r}') - jk\int_A \mathbf{J}_{ms}(\mathbf{r}')dS(\mathbf{r}') - \frac{k^2}{2}\int_A \mathbf{J}_{ms}(\mathbf{r}')R\,dS(\mathbf{r}'),$$

$$\int_A \rho_{ms}(\mathbf{r}')\nabla\frac{e^{-jkR}}{R}dS(\mathbf{r}')$$

$$= \int_A \rho_{ms}(\mathbf{r}')\nabla\frac{1}{R}dS(\mathbf{r}') - \frac{k^2}{2}\int_A \rho_{ms}(\mathbf{r}')\mathbf{u}_R\,dS(\mathbf{r}')$$

$$+ \frac{jk^3}{3}\int_A \rho_{ms}(\mathbf{r}')(\mathbf{r}-\mathbf{r}')dS(\mathbf{r}'),$$

$$\int_A \mathbf{J}_{ms}(\mathbf{r}') \times \nabla \frac{e^{-jkR}}{R} dS(\mathbf{r}')$$
$$= \int_A \mathbf{J}_{ms}(\mathbf{r}') \times \nabla \frac{1}{R} dS(\mathbf{r}') - \frac{k^2}{2} \int_A \mathbf{J}_{ms}(\mathbf{r}') \times \mathbf{u}_R \, dS(\mathbf{r}')$$
$$+ \frac{jk^3}{3} \int_A \mathbf{J}_{ms}(\mathbf{r}') \times (\mathbf{r} - \mathbf{r}') dS(\mathbf{r}').$$

Substituting these into (7.89) and (7.90) yields

$$\frac{1}{2\pi\mu_0} \int_A \rho_{ms}(\mathbf{r}') \nabla \frac{1}{R} dS(\mathbf{r}')$$
$$= \mathbf{H}_{\text{in}}(\mathbf{r}) + \frac{k^2}{4\pi\mu_0} \int_A \rho_{ms}(\mathbf{r}') \mathbf{u}_R \, dS(\mathbf{r}') - \frac{jk^3}{6\pi\mu_0} \int_A \rho_{ms}(\mathbf{r}')(\mathbf{r} - \mathbf{r}') dS(\mathbf{r}')$$
$$- \frac{jk}{2\pi\eta} \left[\int_A \frac{\mathbf{J}_{ms}(\mathbf{r}')}{R} dS(\mathbf{r}') - jk \int_A \mathbf{J}_{ms}(\mathbf{r}') dS(\mathbf{r}') \right.$$
$$\left. - \frac{k^2}{2} \int_A \mathbf{J}_{ms}(\mathbf{r}') R \, dS(\mathbf{r}') \right], \tag{7.91}$$

$$\frac{1}{2\pi} \int_A \mathbf{J}_{ms}(\mathbf{r}') \times \nabla \frac{1}{R} dS(\mathbf{r}') = -\mathbf{E}_{\text{in}}(\mathbf{r}) + \frac{k^2}{4\pi} \int_A \mathbf{J}_{ms}(\mathbf{r}') \times \mathbf{u}_R \, dS(\mathbf{r}')$$
$$- \frac{jk^3}{6\pi} \int_A \mathbf{J}_{ms}(\mathbf{r}') \times (\mathbf{r} - \mathbf{r}') dS(\mathbf{r}'). \tag{7.92}$$

For a small aperture, the right-hand sides of (7.91) and (7.92) may be assumed to be a constant and may be replaced by their values at the origin of the aperture. Thus

$$\frac{1}{2\pi} \int_A \rho_{ms}(\mathbf{r}') \nabla \frac{1}{R} dS(\mathbf{r}') = \mathbf{f}, \tag{7.93}$$

$$\frac{1}{2\pi} \int_A \mathbf{J}_{ms}(\mathbf{r}') \times \nabla \frac{1}{R} dS(\mathbf{r}') = g\mathbf{u}_z, \tag{7.94}$$

where

$$\mathbf{f} = \mu_0 \mathbf{H}_{\text{in}}(0) - \frac{k^2}{4\pi} \int_A \frac{\rho_{ms}(\mathbf{r}')\mathbf{r}'}{r'} dS(\mathbf{r}') + \frac{jk^3}{6\pi}\mu_0 \mathbf{m} - \frac{jk^3}{2\pi}\mu_0 \mathbf{m}$$

$$-\frac{jk\mu_0}{2\pi\eta}\left[\int_A \frac{\mathbf{J}_{ms}(\mathbf{r}')}{r'} dS(\mathbf{r}') - \frac{k^2}{2}\int_A \mathbf{J}_{ms}(\mathbf{r}')r' dS(\mathbf{r}')\right], \tag{7.95}$$

$$g\mathbf{u}_z = -\mathbf{E}_{\text{in}}(0) + \frac{k^2}{4\pi}\int_A \frac{\mathbf{r}' \times \mathbf{J}_{ms}(\mathbf{r}')}{r'} dS(\mathbf{r}') + \frac{jk^3}{3\pi}\frac{\mathbf{p}}{\varepsilon_0}, \tag{7.96}$$

with the magnetic dipole $\mathbf{m}$, the electric dipole $\mathbf{p}$ defined by

$$\mathbf{m} = \frac{1}{\mu_0}\int_A \rho_{ms}(\mathbf{r}')\mathbf{r}' dS(\mathbf{r}') = \frac{1}{j\omega\mu_0}\int_A \mathbf{J}_{ms}(\mathbf{r}') dS(\mathbf{r}'),$$

$$\mathbf{p} = \varepsilon_0 \int_A \frac{-\mathbf{r}' \times \mathbf{J}_{ms}(\mathbf{r}')}{2} dS(\mathbf{r}'). \tag{7.97}$$

For a circular aperture of radius a, (7.93) can be solved analytically and the solution in the cylindrical system (r, φ, z) is (e.g., Bladel, 1971)

$$\rho_{ms}(\mathbf{r}) = \frac{4r}{\pi\sqrt{a^2 - r^2}}\mathbf{u}_r \cdot \mathbf{f}. \tag{7.98}$$

For a small circular aperture, the electric field in the vicinity of the aperture can be derived from a scalar potential ϕ, i.e., $\mathbf{E} = -\nabla\phi$. Equation (7.94) can be written as

$$\frac{1}{\pi}\int_A \nabla\phi(\mathbf{r}') \cdot \nabla\frac{1}{R} dS(\mathbf{r}') = g. \tag{7.99}$$

The solution of above equation in the cylindrical system (r, φ, z) is

$$\phi(\mathbf{r}) = \frac{g}{\pi}\sqrt{a^2 - r^2}. \tag{7.100}$$

This gives

$$\mathbf{J}_{ms} = 2\mathbf{u}_z \times \mathbf{E} = \mathbf{u}_\varphi \frac{g}{\pi}\frac{2r}{\sqrt{a^2 - r^2}}. \tag{7.101}$$

Substituting (7.98) and (7.101) into (7.97), we obtain

$$\mu_0 \mathbf{m} = \frac{8}{3}a^3 \mathbf{f}, \quad \frac{\mathbf{p}}{\varepsilon_0} = \frac{4a^3}{3}\mathbf{u}_z g. \tag{7.102}$$

As a static-field approximation, only the first term on the right-hand side of (7.95) or (7.96) is important. So we have

$$\mathbf{m} = \frac{8}{3}a^3\mathbf{H}_{\text{in}}(0), \quad \frac{\mathbf{p}}{\varepsilon_0} = -\frac{4a^3}{3}\mathbf{E}_{\text{in}}(0). \tag{7.103}$$

These results have been widely used to study the small-hole coupling problems in microwave engineering (see Sections 2.4.5 and 4.4.4).

7.4 EMC Techniques

Any electronic circuit that carries electric signals will tend to radiate electromagnetic energy into space as a transmitter. At the same time, the circuit will tend to pick up radiated electromagnetic energy from other transmitters as a receiving antenna. A shield, a metallic enclosure, can be used to contain the radiated emission from the electronic circuit [Figure 7.20(a)], and can also be used to prevent unwanted radiated emission from the outside into a device [Figure 7.20(b)].

Although a circuit may be well-protected by a shield to prevent electromagnetic energy being radiated or being picked up by the circuit itself, unwanted (interfering) signals can enter or leave the circuit through its interconnections (wire lines). To reduce the levels of the unwanted signals which are usually out of the band that a useful signal occupies, EMC filters may be placed in the lines so that the useful signal is allowed to pass while the interfering signals are blocked, as illustrated in Figure 7.21.

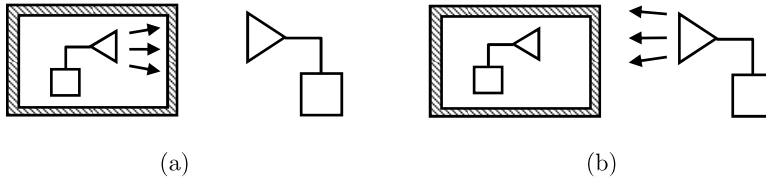


Figure 7.20 (a) A shield used to prevent the radiated emission from an electronic device into the outside. (b) A shield used to prevent unwanted radiated emission from the outside into a device.

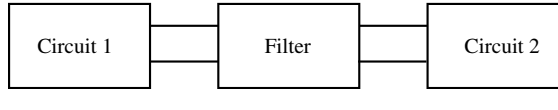


Figure 7.21 EMC filter between two circuits.

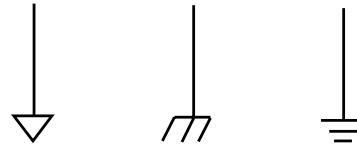


Figure 7.22 Ground symbols. (a) Signal ground. (b) Safety ground. (c) Earth ground.

The EMC filters mainly consist of two types. One type is used to absorb the unwanted energy. The other is used to reject the unwanted signal so that it is reflected back along the line. For EMC applications, the absorptive type is preferred.

In electrical engineering, ground or earth refers to the reference point in an electrical circuit from which voltages are measured, a common return path for electric current, or a direct physical connection to the Earth. The grounding scheme within equipment is very important. Poor grounding can lead to ground loops that can in turn lead to signals being radiated, or picked up within the equipment and hence poor EMC results.

Typical ground symbols are shown in Figure 7.22. Signal grounds serve as return paths for signals within equipment. Safety grounds, often referred to as chassis grounds, are required to provide protection against shock hazard, and also serve an important role in draining electrostatic discharge (ESD).

7.4.1 Shielding Method

Electromagnetic shielding is usually achieved by placing a conducting screen or a conductive enclosure between two regions so that it reduces or prevents transmission of electromagnetic fields from one side to the other. Perfect shielding of a device is impractical due to the input and output accesses by the device. To be cost effective, many devices do not use contiguous metallic enclosures but can still pass the EMC tests.

7.4.1.1 *Shielding Effectiveness: Far-Field Sources*

Consider a conductive screen of thickness d with medium parameters μ, ε, σ, as shown in Figure 7.23. An electromagnetic field $(\mathbf{E}_{\text{in}}, \mathbf{H}_{\text{in}})$ is assumed to be incident on the screen, which produces a reflected wave $(\mathbf{E}_r, \mathbf{H}_r)$ in the Region I $(z < 0)$ and a transmitted wave $(\mathbf{E}_t, \mathbf{H}_t)$ in the Region III

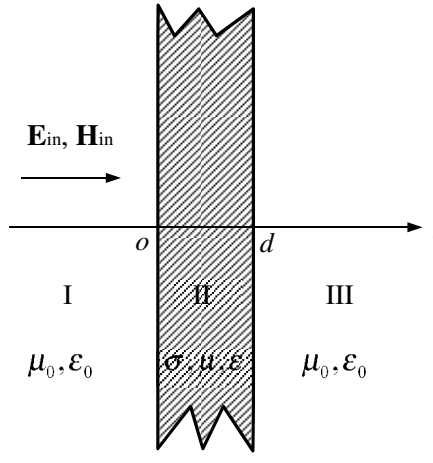


Figure 7.23 Conductive screen.

($z > d$). If we wish to shield the Region III from Region I, the **shielding effectiveness** of the screen is defined in decibels as

$$\mathrm{SE}_{dB} = 20 \log \frac{|\mathbf{E}_{\mathrm{in}}|}{|\mathbf{E}_t|}, \tag{7.104}$$

which is a positive number since the incident field is greater than the transmitted field in magnitude. Assume that the incident field is a plane wave:

$$\mathbf{E}_{\mathrm{in}} = \mathbf{u}_x E_{\mathrm{in}} e^{-jk_0 z}, \quad \mathbf{H}_{\mathrm{in}} = \mathbf{u}_y \frac{1}{\eta_0} E_{\mathrm{in}} e^{-jk_0 z},$$

where the amplitude E_{in} is assumed to be known. The fields in each region can be expressed as

$$\begin{aligned}
&\mathbf{E}_1 = \mathbf{u}_x (E_{\mathrm{in}} e^{-jk_0 z} + E_r e^{jk_0 z}), \quad \mathbf{H}_1 = \mathbf{u}_y \left(\frac{E_{\mathrm{in}}}{\eta_0} e^{-jk_0 z} - \frac{E_r}{\eta_0} e^{-jk_0 z} \right), \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad z < 0, \\
&\mathbf{E}_2 = \mathbf{u}_x (E_{2\mathrm{in}} e^{-\gamma z} + E_{2r} e^{\gamma z}), \quad \mathbf{H}_2 = \mathbf{u}_y \left(\frac{E_{2\mathrm{in}}}{\eta} e^{-\gamma z} - \frac{E_{2r}}{\eta} e^{\gamma z} \right), \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad 0 < z < d, \\
&\mathbf{E}_3 = \mathbf{u}_x E_t e^{-jk_0 z}, \quad \mathbf{H}_3 = \mathbf{u}_y \frac{E_t}{\eta_0} e^{-jk_0 z}, \quad z > d,
\end{aligned} \tag{7.105}$$

where

$$k_0 = \omega\sqrt{\mu_0\varepsilon_0}, \quad \eta_0 = \sqrt{\mu_0/\varepsilon_0},$$
$$\gamma = \sqrt{j\omega\mu(\sigma + j\omega\varepsilon)} = \alpha + j\beta,$$
$$\eta = \sqrt{\frac{j\omega\mu}{\sigma + j\omega\varepsilon}}.$$

At the boundaries $z = 0$ and $z = d$, the tangential fields must be continuous. Hence we have

$$\begin{aligned} &E_{\text{in}} + E_r = E_{2\text{in}} + E_{2r}, \\ &\frac{E_{\text{in}}}{\eta_0} - \frac{E_r}{\eta_0} = \frac{E_{2\text{in}}}{\eta} - \frac{E_{2r}}{\eta}, \\ &E_{2\text{in}}e^{-\gamma d} + E_{2r}e^{\gamma d} = E_t e^{-jk_0 d}, \\ &\frac{E_{2\text{in}}}{\eta}e^{-\gamma d} - \frac{E_{2r}}{\eta}e^{\gamma d} = \frac{E_t}{\eta_0}e^{-jk_0 d}. \end{aligned} \tag{7.106}$$

From these equations, we obtain

$$\frac{E_{\text{in}}}{E_t} = \frac{(\eta_0 + \eta)^2}{4\eta_0\eta}\left[1 - \left(\frac{\eta_0 - \eta}{\eta_0 + \eta}\right)^2 e^{-2\gamma d}\right] e^{\gamma d}e^{-jk_0 d}. \tag{7.107}$$

The shielding effectiveness can thus be written as

$$\text{SE}_{dB} = R_{dB} + A_{dB} + M_{dB}, \tag{7.108}$$

where

$$\begin{aligned} R_{dB} &= 20\log\left|\frac{(\eta_0 + \eta)^2}{4\eta_0\eta}\right|, \\ A_{dB} &= 20\log\left|e^{\gamma d}\right|, \\ M_{dB} &= 20\log\left|1 - \left(\frac{\eta_0 - \eta}{\eta_0 + \eta}\right)^2 e^{-2\gamma d}\right| \end{aligned} \tag{7.109}$$

respectively denote the reflection loss caused by the interfaces, absorption loss through the screen, and multiple reflection loss at interfaces. The reflection loss is the predominant shielding mechanism at the lower frequencies while the absorption loss is the predominant shielding mechanism at the higher frequencies.

The propagation constant γ in the conductive screen can be approximated by

$$\gamma = \alpha + j\beta \approx \frac{1+j}{\delta}, \tag{7.110}$$

where

$$\delta = \sqrt{\frac{2}{\omega\mu\sigma}} \tag{7.111}$$

is the skin depth for the conductive material. For a good conductor, we have $\eta << \eta_0$. Thus

$$\frac{\eta_0 - \eta}{\eta_0 + \eta} \approx 1.$$

Making use of the above approximations and taking the absolute value of (7.107) yield

$$\left|\frac{E_{\text{in}}}{E_t}\right| = \left|\frac{\eta_0}{4\eta}\right| \left|1 - e^{-\frac{2d}{\delta}} e^{-j\frac{2d}{\delta}}\right| e^{\frac{d}{\delta}}. \tag{7.112}$$

Therefore

$$\begin{aligned} R_{dB} &= 20\log\left|\frac{\eta_0}{4\eta}\right|, \\ A_{dB} &= 20\log e^{\frac{d}{\delta}}, \\ M_{dB} &= 20\log\left|1 - e^{-\frac{2d}{\delta}} e^{-j\frac{2d}{\delta}}\right|. \end{aligned} \tag{7.113}$$

7.4.1.2 *Shielding Effectiveness: Near-Field Sources*

Our previous study applies to the situation where the source of the incident fields is far from the shields. When the shields are located in the near-field region of the source, the discussion of the shielding effectiveness becomes very complicated. In such cases, the techniques for shielding will depend on the type of the sources.

To understand the behavior of the fields in the near-field and far-field regions, we may consider how the wave impedance changes with the distance from the source. In general, the **wave impedance** can be defined by

$$Z_w = \begin{cases} E_\theta / H_\varphi, & H_\varphi \neq 0 \\ -E_\varphi / H_\theta, & H_\theta \neq 0 \end{cases}. \tag{7.114}$$

For an infinitesimal electric dipole in free space, the wave impedance may be obtained from (7.42) as follows

$$Z_{we} = \frac{E_\theta}{H_\varphi} = \eta_0 \frac{j(k_0 r)^{-1} + (k_0 r)^{-2} - j(k_0 r)^{-3}}{j(k_0 r)^{-1} + (k_0 r)^{-2}}, \tag{7.115}$$

where $\eta_0 = \sqrt{\mu_0/\varepsilon_0}$ and $k_0 = \omega\sqrt{\mu_0\varepsilon_0}$. For an infinitesimal magnetic dipole in free space, the wave impedance can be obtained from (7.51) as follows

$$Z_{wm} = -\frac{E_\varphi}{H_\theta} = \eta_0 \frac{j(k_0 r)^{-1} + (k_0 r)^{-2}}{j(k_0 r)^{-1} + (k_0 r)^{-2} - j(k_0 r)^{-3}}. \tag{7.116}$$

The magnitude of the wave impedances for the electric dipole and magnetic dipole are plotted in Figure 7.24. It can be seen that the wave impedances for both the electric dipole and magnetic dipole approach to the intrinsic impedance η_0 of the medium as r increases. In the near-field region with $k_0 r < 0.707$, the electric (magnetic) dipole has a wave impedance greater (less) than the intrinsic impedance of the medium and hence is known as **high(low)-impedance source**.

As an approximation, when a shield is located in the near-field region of a dipole-like source, called **electric source**, or a loop-like source, called **magnetic source**, the shield effectiveness may be obtained from (7.108) and (7.109) by replacing η_0 with Z_{we} given by (7.115) or Z_{wm} given by (7.116). It follows from (7.113) that only reflection loss depends on the nature of the near fields. For the electric source, the reflection loss can be

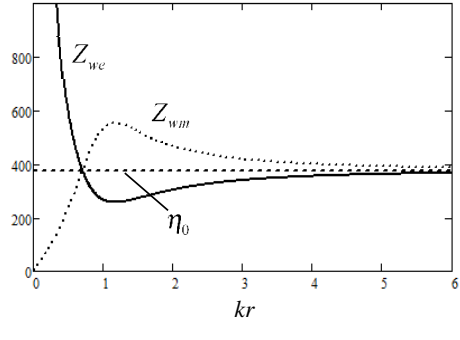


Figure 7.24 Wave impedances of infinitesimal electric dipole and magnetic dipole.

written as

$$R_{dB} = 20 \log \left| \frac{Z_{we}}{4\eta} \right|.$$

In the near-field region and for a good conductor, we may make the following approximation

$$|Z_{we}| \approx \frac{1}{\omega \varepsilon_0 r}, \quad |\eta| = \sqrt{\frac{\omega\mu}{\sigma}} = \sqrt{\frac{\omega \mu_r \mu_0}{\sigma_r \sigma_{cu}}},$$

where $\sigma = \sigma_r \sigma_{cu}$ and $\sigma_{cu} = 5.8 \times 10^7$ S/m is the conductivity of copper. The reflection loss is then given by

$$R_{dB} = 20 \log \frac{\sqrt{\sigma_{cu}/\varepsilon_0}}{8\sqrt[3]{2\pi}\sqrt{\mu_0 \varepsilon_0}} + 10 \log \frac{\sigma_r}{\mu_r f^3 r^2} = 322 + 10 \log \frac{\sigma_r}{\mu_r f^3 r^2}, \tag{7.117}$$

which indicates that the reflection loss increases as frequency or the distance between the electric source and the shield decreases.

In the near-field region of a magnetic source, the following approximation applies

$$|Z_{wm}| = \omega \mu_0 r.$$

The reflection loss for the magnetic source is

$$\begin{aligned} R_{dB} &= 20 \log \left| \frac{Z_{wm}}{4\eta} \right| = 20 \log \frac{\sqrt{2\pi \mu_0 \sigma_{cu}}}{4} + 10 \log \frac{f r^2 \sigma_r}{\mu_r} \\ &= 14.6 + 10 \log \frac{f r^2 \sigma_r}{\mu_r}. \end{aligned} \tag{7.118}$$

Therefore, the reflection loss is negligible at low frequencies, and absorption loss is dominant at all frequencies for a magnetic source. It is noted that both reflection and absorption losses are very small at low frequencies, more effective methods of shielding must be used. There are two basic methods for shielding against low frequency magnetic source. One is to use high μ_r material to provide a low-reluctance path to magnetic flux, and the other is to generate an opposing flux via Lenz's law (Paul, 2006).

7.4.1.3 *Electrostatic Shielding*

The proceeding shielding theory applies to the high frequency problems. Some electronic devices are very sensitive to electrostatic fields. The

damages caused by electrostatic fields may be caused either by direct contact of a charged objects with an electronic device or by the presence of electrostatic fields, which can be prevented by electrostatic shielding. The basic principles of electrostatic shielding are built on the theory of electrostatics. The field is electrostatic if the following conditions are satisfied

$$\frac{\partial}{\partial t}\mathbf{F} = 0, \quad \mathbf{J} = 0, \tag{7.119}$$

where $\mathbf{F}$ stands for the electric field or magnetic field. The fundamental equations for the static electric field generated by a volume charge distribution of density $\rho(\mathbf{r})$ are

$$\nabla \times \mathbf{E}(\mathbf{r}) = 0, \quad \nabla \cdot \mathbf{D}(\mathbf{r}) = \rho(\mathbf{r}), \tag{7.120}$$

with the boundary conditions on the interface of two regions given by

$$\mathbf{u}_n \times (\mathbf{E}_1 - \mathbf{E}_2) = 0, \quad \mathbf{u}_n \cdot (\mathbf{D}_1 - \mathbf{D}_2) = \rho_s. \tag{7.121}$$

where $\mathbf{u}_n$ is the unit normal of the boundary directed from medium 2 to medium 1 and ρ_s is the surface charge density. From (7.120), we may introduce the potential function ϕ such that $\mathbf{E} = -\nabla\phi$ to get the Poisson equation

$$\nabla^2\phi(\mathbf{r}) = -\frac{\rho(\mathbf{r})}{\varepsilon}. \tag{7.122}$$

It follows from (7.121) that

$$\phi_1(\mathbf{r}) = \phi_2(\mathbf{r}), \quad \varepsilon_1\frac{\partial\phi_1(\mathbf{r})}{\partial n} - \varepsilon_2\frac{\partial\phi_2(\mathbf{r})}{\partial n} = -\rho_s. \tag{7.123}$$

The field inside a conductor will cause electric current. Since $\mathbf{J} = \sigma\mathbf{E}$, the second equation of (7.119) implies that the static electric field must be zero inside the conductor. As a result, the conductor is equipotential. It follows from the second equation of (7.120) that the net charge inside the conductor must also be zero and the charges are distributed on the surface of the conductor. For a conductor, the boundary conditions (7.121) and (7.123) respectively reduce to

$$\mathbf{u}_n \times \mathbf{E} = 0, \quad \mathbf{u}_n \cdot \mathbf{D} = \rho_s, \tag{7.124}$$

and

$$\phi(\mathbf{r}) = \text{const}, \quad \frac{\partial\phi(\mathbf{r})}{\partial n} = -\frac{\rho_s}{\varepsilon}. \tag{7.125}$$

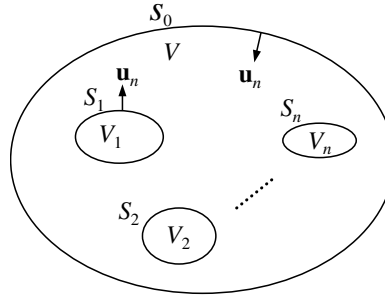


Figure 7.25 Multiply connected region.

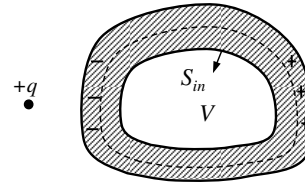


Figure 7.26 A metal cavity excited by an external charge.

Consider a multiply connected region V bounded by S_i $(i = 0, 1, 2, \ldots, n)$ as illustrated in Figure 7.25. The solution of (7.122) can then be expressed as

$$\phi(\mathbf{r}) = \int_V G(\mathbf{r}, \mathbf{r}') \frac{\rho(\mathbf{r}')}{\varepsilon} dV(\mathbf{r}') + \sum_{i=0}^{n} \int_{S_i} [\phi(\mathbf{r}') \mathbf{u}_n(\mathbf{r}') \cdot \nabla' G(\mathbf{r}, \mathbf{r}') - G(\mathbf{r}, \mathbf{r}') \mathbf{u}_n(\mathbf{r}') \cdot \nabla' \phi(\mathbf{r}')] dS(\mathbf{r}'), \tag{7.126}$$

where $G(\mathbf{r}, \mathbf{r}')$ is Green's function satisfying

$$\nabla^2 G(\mathbf{r}, \mathbf{r}') = -\delta(\mathbf{r} - \mathbf{r}'). \tag{7.127}$$

A conductive shell can be used as an electrostatic shield. Consider a closed metallic cavity excited by an external static charge, as shown in Figure 7.26. The cavity occupies the region V bounded by the inner surface S_{in} of the wall of the cavity. The wall of the cavity is assumed to be of finite thickness. It can be shown that the field inside the cavity must be zero. In fact the

potential inside the cavity can be expressed as

$$\phi(\mathbf{r}) = \int_{S_{\text{in}}} [\phi(\mathbf{r}')\mathbf{u}_n(\mathbf{r}') \cdot \nabla' G(\mathbf{r}, \mathbf{r}') - G(\mathbf{r}, \mathbf{r}')\mathbf{u}_n(\mathbf{r}') \cdot \nabla' \phi(\mathbf{r}')] \, dS(\mathbf{r}')$$

$$= \int_{S_{\text{in}}} \left[\phi(\mathbf{r}')\mathbf{u}_n(\mathbf{r}') \cdot \nabla' G(\mathbf{r}, \mathbf{r}') + G(\mathbf{r}, \mathbf{r}') \frac{\rho_s(\mathbf{r}')}{\varepsilon} \right] dS(\mathbf{r}'). \tag{7.128}$$

On the inner surface S_{in}, we have $\phi(\mathbf{r}) = \phi_0$ with ϕ_0 being a constant and $\rho_s = 0$ on S_{in}, the above equation becomes

$$\phi(\mathbf{r}) = \phi_0 \int_{S_{\text{in}}} \frac{\partial G(\mathbf{r}, \mathbf{r}')}{\partial n'} dS(\mathbf{r}'). \tag{7.129}$$

Considering the following identity

$$-1 = \int_V \nabla'^2 G(\mathbf{r}, \mathbf{r}') dV(\mathbf{r}') = -\int_{S_{\text{in}}} \frac{\partial G(\mathbf{r}, \mathbf{r}')}{\partial n'} dS(\mathbf{r}') \tag{7.130}$$

we obtain $\phi(\mathbf{r}) = \phi_0$ in V. As a result, the electric field vanishes inside the cavity.

When a positive charge is inside the originally neutral cavity, an equal negative charge is induced on the inner surface of the wall. An equal positive charge has to be present on the outer surface if the cavity is not grounded as indicated in Figure 7.27(a). In this case, the field outside the cavity is not zero. If the cavity is connected to an ideal reservoir of charges (i.e., the ground such as the Earth), the positive charges on the outer surface can be compensated by the negative charges coming from the ground and the

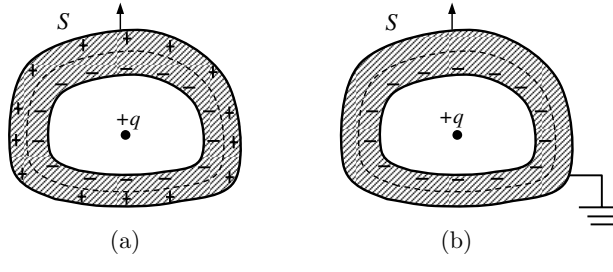


Figure 7.27 A metal cavity excited by a charge inside. (a) The cavity is not grounded. (b) The cavity is grounded.

grounded cavity is thus negatively charged, as illustrated in Figure 7.27(b). In this case, the field outside the cavity is zero. In fact, the potential outside the cavity can be expressed as

$$\phi(\mathbf{r}) = \int\limits_{S_{\text{out}}} \left[\phi(\mathbf{r}') \frac{\partial G(\mathbf{r}, \mathbf{r}')}{\partial n'} + G(\mathbf{r}, \mathbf{r}') \frac{\rho_s(\mathbf{r}')}{\varepsilon} \right] dS(\mathbf{r}'), \tag{7.131}$$

where S_{out} is the outer surface of the wall. Since the cavity is grounded, we have $\phi(\mathbf{r}) = 0$ and $\rho_s(\mathbf{r}) = 0$ on S_{out}. Thus the potential outside the cavity is zero. It is noted that a high dielectric constant medium can also be used as an electrostatic shield.

7.4.2 Filtering Method

Conducted emissions are the radio frequency noise current that flows in the physical wiring or traces of an electrical system, or alternatively, radio frequency voltage between traces. For the purposes of EMI analysis, conducted emissions are generally of interest over the frequency range from 150 kHz to 30 MHz, which is the frequency range over which most regulatory agencies specify conducted emissions limits.

EMC filters are very useful for the lines that only carry low frequency signals such as the AC power cord. The EMC filters can remove high frequency components that are coupled to the power cord from the internal subsystems or from the outside via a number of coupling paths. The EMC filters may be as simple as a resistor or a ferrite placed around a wire or cable. For more complicated applications, the EMC filters may consist of a number of components. In addition, the EMC filter design must match both the source and load impedances.

The traditional design theory of filters has been discussed in Chapter 4, which is, however, rarely used in the design of EMC filters. Instead of designing a filter stage for every new piece of electronic equipment, a ready-made module by the filter manufacturer, which is optimized to give the best reduction of conducted interference, can be deployed for most situations. These commercially available filters comply with the safety rules and are cost-effective.

7.4.2.1 *Line Impedance Stabilization Network*

When testing a device for compliance with the regulatory limits, a line impedance stabilization network (LISN) must be inserted between the AC power cord of the DUT and the commercial power outlet, as illustrated in

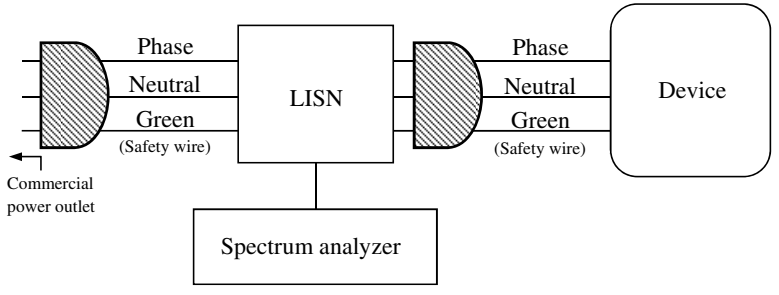


Figure 7.28 Conducted emission test.

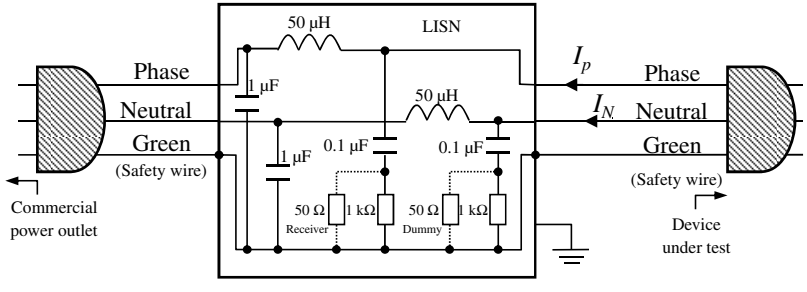


Figure 7.29 FCC-specified LISN for conducted emission measurement.

Figure 7.28. The AC power cord of the device is plugged into the input of the LISN. The output of the LISN is plugged into the commercial power outlet. One of the objectives of LISN is to stabilize the impedance seen by the device looking into the AC power cord, which varies considerably over the measurement frequency range from outlet to outlet and from building to building. The second objective of the LISN is to block external noise that exists on the power system net from entering the product's AC power cord since we are only interested in the conducted emissions that are due to the DUT.

The FCC-specified LISN for conducted emission measurement is shown in Figure 7.29. The purpose of 1 μF capacitors between phase and green wire and between neutral and green wire on the commercial power side is to divert external noise on the commercial power line and prevent that noise from flowing through the DUT and thereby contaminating the test data. Similarly, the purpose of 50 μH inductors is to block that noise. The purpose of 0.1 μF capacitors is to prevent DC from entering the test receiver input. The 1 kΩ resistors facilitate the discharge of 0.1 pF capacitors in case the

50 resistors are removed. One of the 50 Ω resistors is the input impedance of the spectrum analyzer or receiver while the other serves as a dummy load to make sure that the impedance between neutral and safety wire is 50 Ω at all times. Both the phase voltage V_P between the phase and safety wire and the neutral voltage V_N between the neutral and safety wire are measured and are required to be below the specified limit over the frequency range of the conducted emission measurement. In the measurement frequency range from 150 kHz to 30 MHz, all the capacitors of the LISN are essentially short circuits and all the inductors are open circuits. As a result, we have

$$V_P \approx 50 I_P, \quad V_N \approx 50 I_N. \tag{7.132}$$

7.4.2.2 *Common-Mode and Differential-Mode*

It follows from (7.132) that the LISN may be represented as 50 Ω resistors between phase wire and safety wire and between neutral wire and safety wire, seen by the DUT, as illustrated in Figure 7.30. The phase current and neutral current can be decomposed as follows

$$I_P = I_C + I_D, \quad I_N = I_C - I_D, \tag{7.133}$$

where I_C is the common-mode component that flows out through the phase conductor and the neutral conductor and returns on the safety wire; and I_D is the differential-mode component that flows out through the phase conductor and returns on the neutral conductor. It follows from (7.133) that

$$I_D = \frac{1}{2}(I_P - I_N), \quad I_C = \frac{1}{2}(I_P + I_N). \tag{7.134}$$

The measured voltages are then given by

$$V_P = 50(I_C + I_D), \quad V_N = 50(I_C - I_D). \tag{7.135}$$

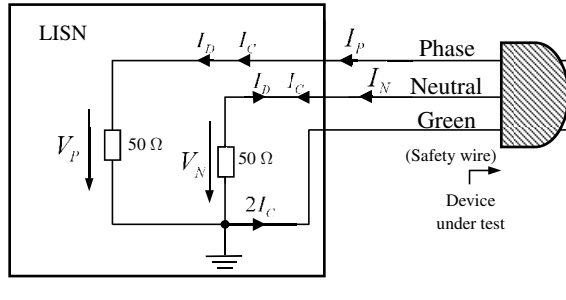


Figure 7.30 Equivalent circuit for the LISN.

7.4.2.3 *Power Supply Filters*

For all electronic devices with a power cord, a power supply filter must be used so that they can pass the conducted emission test. A generic power supply filter is shown in Figure 7.31. The differential- and common-mode currents at the input of the filter are denoted by I_D and I_C. At the output of the filter, the corresponding quantities are denoted by I'_D and I'_C. The function of the filter is to reduce the unprimed current levels to the primed current levels so that the measured voltages given by the primed quantities

$$V_P = 50(I'_C + I'_D), \quad V_N = 50(I'_C - I'_D) \tag{7.136}$$

are below the conducted emission limit.

The green-wire inductor L_{GW} is used to block common-mode current. The line-to-line capacitors C_{DR} and C_{DL} are introduced to divert differential-mode currents. The capacitors C_{CL} and C_{CR} are used to divert common-mode currents. The coupled inductors in the filter are the common-mode choke, where the self-inductances are denoted by L and the mutual inductance is denoted by M. The mutual inductance is approximately equal to the self-inductance $L \approx M$ so that the coupling coefficient is approximately unity:

$$k = \frac{M}{\sqrt{LL}} \approx 1.$$

The common-mode choke is used to block the common-mode currents. As shown in Figure 7.32(a), the voltage drop across one side of the choke with the common-mode currents is

$$V = j\omega(L + M)I_C.$$

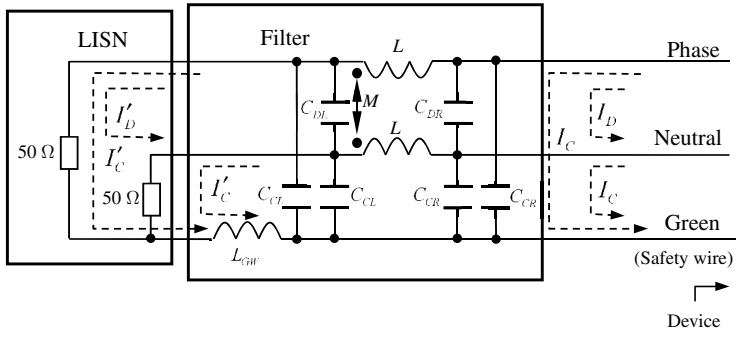


Figure 7.31 A typical power supply filter.

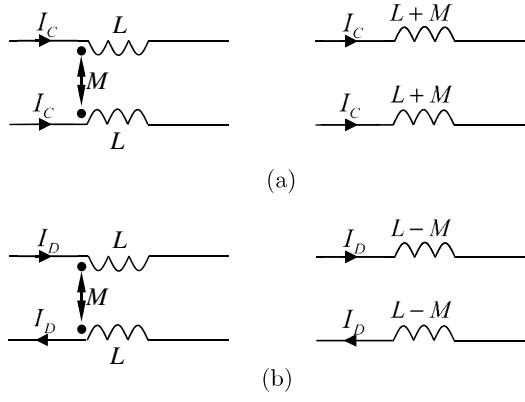


Figure 7.32 (a) Equivalent circuit for common-mode current. (b) Equivalent circuit for differential-mode current.

Therefore, the common-mode current is blocked by an inductance $L + M$. The common-mode choke should not affect the differential-mode current, and this can be observed by considering the voltage drop across one side of the choke with differential-mode current [Figure 7.32(b)]

$$V = j\omega(L - M)I_D \approx 0.$$

7.4.3 Grounding Method

The basic function of a ground is to provide a path to enable a current to return to its source, and therefore it should have low impedance. For this reason, the ground must be a good conductor of large surface area, and is close to the system. Typical ground structures include metal part of the building, the metal frame of a vehicle, the metal chassis of equipment, and the Earth. Electrical circuits may be connected to ground for several reasons:

(1) Safety ground (chassis ground): The purpose of safety grounding is to reduce the voltage difference between exposed conducting surfaces that might become energized. In power system, for example, exposed parts must be connected to ground to prevent user contact with dangerous voltage if electrical insulation fails. Connections to ground limit the build-up of static electricity and divert the ESD currents away from flammable products or electrostatic-sensitive devices.
(2) Signal reference: For measurement purposes at DC or low frequencies, the ground serves as a (reasonably) constant potential reference against

which other potentials can be measured. Different parts of a signal system, such as analog and digital circuits must operate at the same voltage reference. A voltage difference between reference points may cause common mode noise for the system. To reduce the voltage difference, the reference points may be connected together by a conductor, called signal grounding conductor.

(3) Signal return path: In portable electronic devices such as cell phones, the ground plane on a printed circuit board is introduced to serves as the common return path for current from many different components in the circuit. In some telegraph and power transmission circuits, the Earth itself can be used as the return conductor of the circuit to save the cost.

Therefore, there are basically two types of ground: safety ground and signal ground.

7.4.3.1 *Safety Ground*

As an example, let us consider an electronic device powered by an AC source in a metal chassis as shown in Figure 7.33. The exposed metal chassis might become energized due to the AC source, which will produce a voltage difference V between the chassis and the Earth and thus pose a potential shock hazard to anyone who might touch the chassis. In order to provide shock hazard protection, a safety wire (ground) must be connected to chassis and to the Earth to reduce the voltage difference.

7.4.3.2 *Signal Ground*

A signal ground may be used as the path for signal currents to return to their source. It should be noted that the currents will return to their source

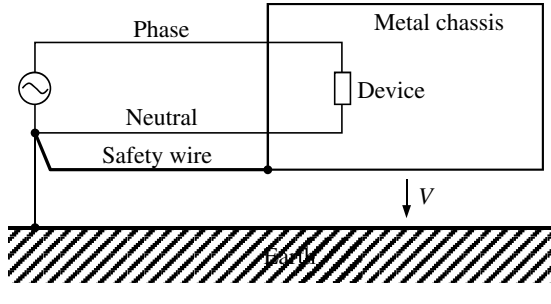


Figure 7.33 Safety ground.

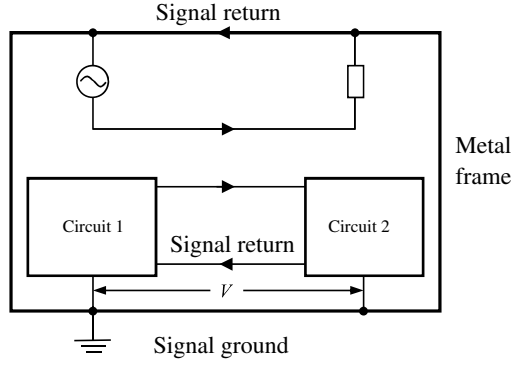


Figure 7.34 Signal return and signal ground.

along the path of least impedance. As a result, different components of the signal may take different paths to return their source, and some frequency components may follow a different path other than the one designated by the designer. Figure 7.34 shows two signals inside a metal frame used as a ground. One signal directly uses the ground as the signal return path, and the other uses a dedicated conductor as the return path between two circuits. In order that the Circuit 1 and Circuit 2 have the same voltage reference (i.e., $V = 0$), the two references are bonded together by the signal ground (part of the frame).

7.5 Lightning Protection

Lightning, or lightning discharge, is a massive ESD between the electrically charged regions within clouds or between a cloud and the Earth's surface to form a long electrical spark (lightning flash), which may extend from 5 to 100 km. Lightning may cause damages to a large variety of objects, such as electronic devices, buildings, power systems, and aircrafts.

7.5.1 Lightning Discharge and Lighting Terminology

In order for an ESD to occur, a high electric potential and a high-resistance medium must exist between two regions of space. All lightning discharges can be divided into two categories: those that bridge the gap between the cloud charge and the Earth, and those that do not. The latter group as a whole is referred to as "cloud discharges" and accounts for the majority of all lightning discharges. The cloud discharges can occur within a single cloud. This is called **intra-cloud lightning** and is the most common of all

the forms of lightning. The cloud discharges that occur between clouds are called **inter-cloud lightning**, and those that occur between one cloud and the surrounding air are called **cloud-to-air lightning**. The charged regions within the atmosphere temporarily equalize themselves through a lightning flash. A lightning discharge that involves an object on ground or in the atmosphere is called a **lightning strike**. The primary source of lightning is the thunderstorm or thundercloud. A simple model for the idealized gross charge structure in a thundercloud is shown in Figure 7.35, which consists of three vertically stacked point charges, positive at the top, negative in the middle and an additional smaller positive at the bottom. The top two charges, called **main charges**, form a dipole and are assumed to be equal in magnitude. The lower positive charge may not always be present.

The soft hail particles are heavy enough to fall in the thundercloud's updrafts and small ice crystals are light enough to be carried upward in those updrafts. The thundercloud charges are produced by the ice-hail interactions, which takes place at altitudes where the temperature is considerably cooler than freezing. After charge has been transferred between the colliding ice and hail particles, the positively charged ice crystals are carried further upward in updrafts to the top part of the thundercloud, to an altitude near 10 km above sea level in temperate summer storms while the negatively charged hail resides at an altitude of 6 to 8 km. In a typical thundercloud, a small positive charge is also formed below the main negative charge, at altitudes where the temperature is near or warmer than freezing.

Any self-propagating electrical discharge creating a channel of ionized air with electrical conductivity of the order $10^{-4}\,\mathrm{Sm}^{-1}$ is called a **leader**. When the conductivity is much lower, the channel of ionized air is called a

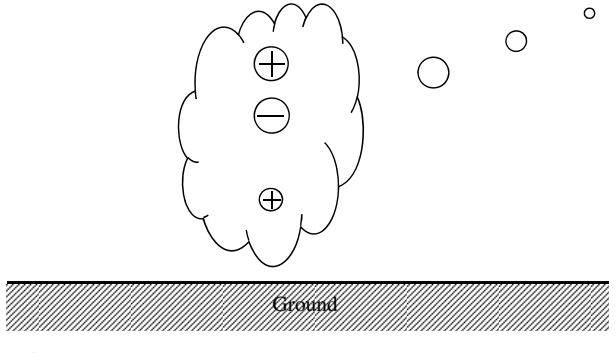


Figure 7.35 A simple model for the idealized gross charge structure in a thundercloud.

streamer. The air behind the streamer tip remains essentially an insulator. There are usually many separate paths of ionized air stemming from the cloud. These paths are typically referred to as **stepped leaders**. The stepped leader's movement from cloud-to-ground is not continuous. It moves downward in discrete luminous segments and each added length is called a **step**. Each leader step produces a pulse which contains a frequency spectrum extending from radio frequency to visible light and X-rays. The luminous diameter of the stepped leader is usually between 1 and 10 m. A typical stepped leader has about 5 coulombs of negative charge distributed over its length when it is near ground. To establish this charge on the leader channel an average current of about 100 to 200 amperes must flow during the whole leader process. The pulsed currents which flow in generating the leader steps have a peak current of the order of 1000 amperes.

For cloud-to-ground discharges, the negative charge normally collects in the cloud base, with a corresponding net positive charge in the ground under the cloud. Lightning strikes originating from this configuration are called **negative strikes**. In negative lightning, the free electrons over-run the lower positive charge region, neutralizing most of its small positive charge, and then continue their trip toward ground. Sometimes lightning originates from the upper part of the thundercloud, which is a region of the cloud that carries a big positive charge. In this case, the ground below has a net negative charge, and any lightning from this configuration is called a **positive strike**. Negative lightning usually strikes under the thunderstorm. Positive lightning often strikes near the edge of a thundercloud or even several miles from the cloud. Positive strikes usually have a stronger electric field than negative strikes. The energy in a positive strike may be 10 times higher than a negative strike, which makes positive lightning more lethal and damaging than negative lightning. There are four types of lightning corresponding to cloud-to-ground discharges as illustrated in Figure 7.36. They are (a) downward negative lightning, (b) downward positive lightning, (c) upward negative lightning, and (d) upward positive lightning.

The most common type of lightning is negative lightning. The downward negative lightning accounts for 90% or more of global cloud-to-ground lightning. When the stepped leader is near the ground, its relatively large negative charge induces (attracts) concentrated positive charge on the conducting Earth beneath it and especially on objects projecting above the Earth's surface. If the attraction between the opposite charges is strong enough, the positive charge on the Earth or Earth-bound objects will attempt to join and neutralize the negative charge above. This initiates an

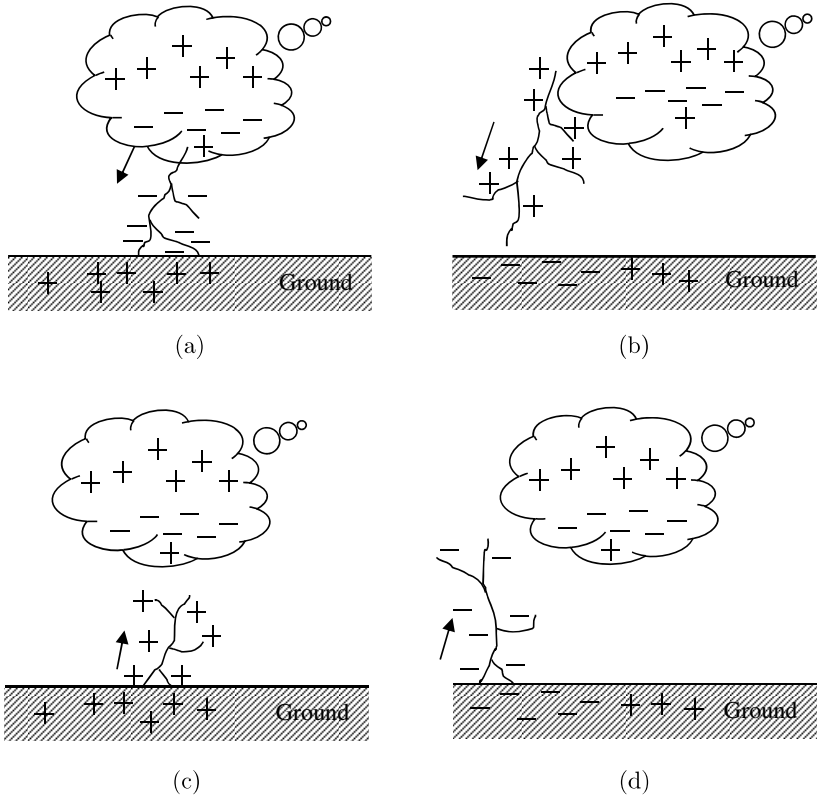


Figure 7.36 Types of cloud-to-ground lightning. (a) Downward negative lightning. (b) Downward positive lightning. (c) Upward negative lightning. (d) Upward positive lightning.

upward leader from the ground or from grounded objects. Once a downward leader connects to an upward leader, a low-resistance path is formed and discharge may occur. This process is referred to as **attachment**, which determines the lightning strike-point and the primary lightning current path (channel) between cloud and ground. Once a conductive channel bridges the ionized air between the negative charges in the cloud and the positive surface charges in the ground, a massive electrical discharge follows and enormous current of positive charges races up the ionic channel towards the thundercloud. This is called **return stroke** and is the most luminous and noticeable part of the lightning discharge. The massive flow of electrical current occurring during the return stroke combined with the rate at which it occurs rapidly superheats the completed leader channel, forming

a highly electrically-conductive plasma channel. The core temperature of the plasma during the return stroke may exceed 50,000 K, causing it to brilliantly radiate with a blue–white color. Once the electrical current stops flowing, the channel cools and dissipates over 10's or hundreds of milliseconds, often disappearing as fragmented patches of glowing gas. The nearly instantaneous heating during the return stroke causes the air to explosively expand, producing a powerful shock wave that is heard as thunder.

7.5.2 Lightning Protection

The electrical current within a typical negative cloud-to-ground lightning discharge rises very quickly to its peak value in 1–10 microseconds, then decays more slowly over 50–200 microseconds. The transient nature of the current within a lightning flash results in several phenomena that need to be addressed in the protection of ground-based structures. Rapidly changing currents tend to travel on the surface of a conductor due to skin effect. For this reason, conductors often used in the protection of facilities are multi-stranded small wires woven together to increase the surface area. The rapidly changing currents also create electromagnetic pulses that radiate outward from the ionic channel. When the pulses pass over conductive elements such as the electrical wires and transmission lines, they may induce a current which travels toward its termination. This is called **lighting surge** that often results in the destruction of delicate electronic devices. As a result, two factors must be considered in lightning protection:

(1) Diversion and shielding, which diverts the lightning current away from the protected structure and serves to reduce the lightning electric and magnetic fields within the structure.
(2) The limiting of currents and voltages on power and communication systems via surge protective devices (SPDs).

The protection system that is used to divert lightning current away from a protected structure and ultimately into the Earth consists of three electrically connected components:

(1) Air terminals, which may be vertical lightning rods connected together on the roof of the structure, or a mesh of horizontal wires on the roof, or overhead catenary wires above the roof, or a metal roof, with the purpose to intercept the descending lightning stepped leader by sending streamers upward.

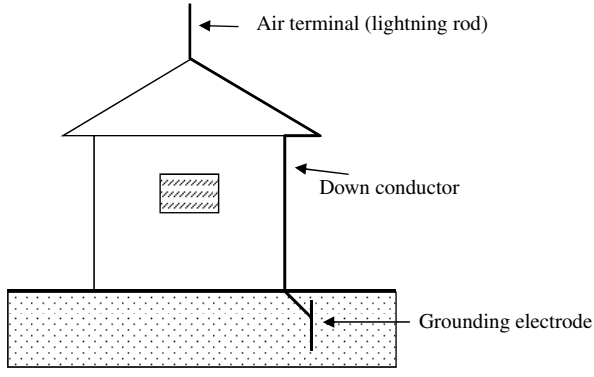


Figure 7.37 A simple lightning protection system.

(2) Down conductors to carry the lightning current to the grounding electrodes.
(3) Grounding electrodes to convey the current into the Earth.

All elements of the lightning protection system must be well-bonded electrically and all significant nearby conductors, including the ground wires on incoming utilities, must be bonded to the overall protection system to avoid voltage differences between the conductors that may lead to electrical breakdown between them. Figure 7.37 shows a simple lightning protection system.

The protection of electronic, power, or communication equipment within a structure should include the control of currents and voltages resulting both from direct strikes to the structure containing the equipment, and from lightning-induced current and voltage surges propagating into the structure on electric power, communication, or other metal wires and metal pipes entering the structure from outside. Four types of current- and voltage-limiting techniques are commonly used:

(1) Voltage crowbar devices, which limit the harmful voltages on the protected wires to small values compared with the operating voltage and attempt to short-circuit the associated current to ground. The older carbon block arresters and the modern gas-tube arresters used by telephone companies are good examples of crowbar devices. When the voltage across such a crowbar device reaches a value of many hundreds of volts, the arrester suffers an electrical breakdown in its gas component, reducing the voltage across the arrester terminals to

near zero. Silicon-controlled rectifiers and triacs are other examples of crowbar devices.

(2) Voltage clamps, which are solid-state devices such as metal oxide varistors, Zener and avalanche diodes, and p–n junction diodes that both reflect and absorb energy while clamping the applied voltage across their terminals to a more-or-less safe value, ideally 30 to 50% above the system operating voltage, rather than the very small voltages allowed by crowbar devices. Voltage clamps are nonlinear devices and can handle less energy than crowbar devices before failing. Both voltage clamps and voltage crowbar devices are referred to as SPDs.

(3) Circuit filters, which are linear electrical circuits that both reflect and absorb the frequencies that form the damaging lightning transient pulses while passing the operating waveforms. The simplest circuit filter is a series inductor whose impedance is much higher to the frequencies comprising the unwanted transient than to the operating frequency of the electronics being protected. Frequently, crowbar devices, clamps, and filters are used together in a coordinated way.

(4) Isolating devices, such as optical isolators and isolation transformers, which can suppress relatively large transients. Isolators are connected in series with the equipment to be protected and represent large series impedance to the unwanted transient signals.

The SPDs discussed above are generally connected at the input terminals of electrical devices or directly on circuit boards.

> I have long held an opinion, almost amounting to conviction, in common I believe with many other lovers of natural knowledge, that the various forms under which the forces of matter are made manifest have one common origin; or, in other words, are so directly related and mutually dependent, that they are convertible, as it were, one into another, and possess equivalents of power in their action.
>
> —Michael Faraday

Chapter 8

Information Theory and Systems

Information is the resolution of uncertainty.

—Claude Elwood Shannon (American mathematician and electrical engineer, 1916–2001)

In 1948, C. E. Shannon published his classic paper "A Mathematical Theory of Communication" in the *Bell System Technical Journal*, which founded the discipline of information theory for modern communication system. The foundations of communications technology also lay in the discovery of electromagnetics. A generic communication system is shown in Figure 8.1. The information source produces message to be communicated. The transmitter transforms the message into a signal suitable for going through a propagation channel during which the signal may be altered by noise and distortion. The channel assigns a probability distribution to the set of all possible outputs for each permissible input. The output of the channel is the received signal, which is then transformed into the original message by the receiver for delivery to the destination. Given a communication system, information theory attempts to build a mathematical model for each of the blocks of Figure 8.1 and studies the following problems:

(1) What is the minimum number of bits per symbol required to fully represent the source?
(2) What is the maximum rate at which reliable communication can take place over the channel?

The mathematical foundation of information theory is probability theory and statistics. The most important quantities of information are entropy and mutual information. The former stands for the information in a random variable and indicates how easily the message data can be compressed while the latter stands for the amount of information in common between two

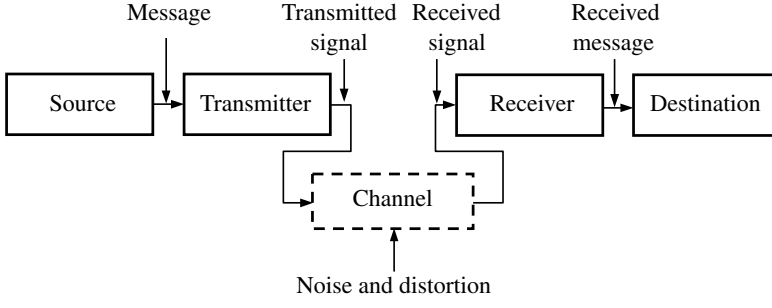


Figure 8.1 A generic communication system.

random variables and can be used to find the communication data rate across a channel. The fundamental theorem of information theory states that it is possible to transmit information through a noise channel at any rate less than channel capacity with an arbitrarily small probability of error.

8.1 Probability Theory and Random Process

The word "random" is used to describe various unpredictable (i.e., non-deterministic) phenomena. A most typical random system is quantum mechanics, which deals with small particle systems. A radio communication system is also random in nature due to the random interferences and noises. In these cases, we cannot predict the exact behavior of the systems. Instead we adopt a statistical description based on probability theory.

8.1.1 Probability Space

Let Ω be a set. A **σ-algebra** S_ω on Ω is a family of subsets of the set Ω with the following properties:

(1) The empty set $\emptyset$ belongs to S_ω: $\emptyset \in S_\omega$.
(2) If $E \in S_\omega$ then $\Omega - E \in S_\omega$.
(3) If $E_i \in S_\omega$ $(i = 1, 2, \ldots)$, then $\bigcup_{i=1}^{\infty} E_i \in S_\omega$.

The pair (Ω, S_ω) consisting of the set Ω and a σ-algebra S_ω is called a **measurable space**. The **probability measure** P on the measurable space (Ω, S_ω) is a function $P : S_\omega \to [0, 1]$ such that

(1) $P(\emptyset) = 0, P(\Omega) = 1$.
(2) If $E_i \in S_\omega$, $i = 1, 2, \ldots$ are disjoint $(E_i \cap E_j = \emptyset)$, then $P\left(\bigcup_i E_i\right) = \sum_i P(E_i)$.

A measurable space (Ω, S_ω) together with a probability measure P defined on the σ-algebra S_ω is called a **probability space**, denoted by a triple (Ω, S_ω, P). The set Ω is called **sample space**. An element in Ω is called a **sample** or an outcome. Any element E of S_ω is called an **event** and $P(E)$ is the probability that event E occurs. A major distinction between samples and events is that the samples are fixed and are not within our control while events can be chosen to suit our convenience. For example, the elements of Ω may be the occurrence or nonoccurrence of a signal pulse, S_ω may be a collection of possible sequences of a certain length of pulse and no pulse. Two events E_1 and E_2 are said to be **mutually exclusive** if $P(E_1 \cap E_2) = 0$. Two events E_1 and E_2 are said to be **statistically independent** if

$$P(E_1 \cap E_2) = P(E_1)P(E_2).$$

If $P(E_1) > 0$ the **conditional probability** $P(E|E_1)$ is defined as the probability that an event will occur when another event is known to occur or to have occurred

$$P(E|E_1) = \frac{P(E \cap E_1)}{P(E_1)}. \tag{8.1}$$

Let $\widetilde{\omega} = X(\omega)(\omega \in \Omega)$ be a function defined on the measurable space (Ω, S_ω), with values in a measurable space $(\widetilde{\Omega}, S_{\widetilde{\omega}})$, i.e., $\widetilde{\omega} \in \widetilde{\Omega}$. If for any set $\widetilde{E} \in S_{\widetilde{\omega}}$, the inverse image $X^{-1}(\widetilde{E})$ belongs to S_ω, X is called a **measurable function**. A measurable function $X(\omega)$ defined on a probability space (Ω, S_ω, P) and taking values in a measurable space $(\widetilde{\Omega}, S_{\widetilde{\omega}})$, is called a **random element**. If $\widetilde{\Omega}$ is a vector space, then $X(\omega)$ is called a **random vector**. If $\widetilde{\Omega} = R$, $X(\omega)$ is called a **random variable**. Random variables may be discrete, continuous, or mixed, depending on whether they take on a countable or uncountable number of values, or both.

Example 8.1: Let $\Omega = \{\omega_1, \omega_2, \ldots\}$ be finite or countable, S_ω be the collection of all subsets of Ω, and $\{p_1, p_2, \ldots\}$ be a sequence of non-negative numbers whose sum is unity. Then for any event $E = \{\omega_n | n \in J\}$ where J is an index set (a subset of natural numbers), we define $P(E) = \sum_{n \in J} p_n$. In this case, (Ω, S_ω, P) is called a **discrete probability space**. □

From now on, an upper case letter will be used to stand for a random vector and the corresponding lower case letters for its value. Sometimes the upper case letter and lower case letter will be used interchangeably if no confusion occurs.

8.1.2 Probability Distribution Function

The **probability distribution** $F_X(\cdot)$ of the random element X, defined on a probability space (Ω, S_ω, P) and taking values in a measurable space $(\widetilde{\Omega}, S_{\widetilde{\omega}})$, is defined by

$$F_X(\widetilde{E}) = P\{\omega|\omega \in X^{-1}(\widetilde{E})\}, \quad \widetilde{E} \in S_{\widetilde{\omega}}.$$

Especially, if X is a random variable, we may introduce $E_x = \{\omega|X(\omega) \leq x\} \in S_\omega$ and $F_X(x) = P(E_x)$, which is known as the probability distribution function of the random variable X. The first derivative of distribution function $p_X(x) = dF_X(x)/dx$ is called the **probability density function**. We have the following properties:

(1) $F_X(x)$ is monotone increasing.
(2) $0 \leq F_X(x) \leq 1$.
(3) $F_X(-\infty) = 0, F_X(\infty) = 1$.
(4) $F_X(x)$ is right continuous.
(5) The set of points on which $F_X(x)$ is discontinuous is at most countable.
(6) $P(\omega|x_1 < X(\omega) \leq x_2) = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} p_X(x)dx$.
(7) $\int_{-\infty}^{\infty} p_X(x)dx = 1$.

If $(\widetilde{\Omega}_1, S_{1\widetilde{\omega}})$ and $(\widetilde{\Omega}_2, S_{2\widetilde{\omega}})$ are two measurable spaces, their product $(\widetilde{\Omega}_1 \times \widetilde{\Omega}_2, S_{1\widetilde{\omega}} \times S_{2\widetilde{\omega}})$ consists of the space $\widetilde{\Omega}_1 \times \widetilde{\Omega}_2$ of all pairs (x, y) with $x \in \widetilde{\Omega}_1$, $y \in \widetilde{\Omega}_2$ and the σ-algebra $S_{1\widetilde{\omega}} \times S_{2\widetilde{\omega}}$ generated by all sets $\widetilde{E} \times \widetilde{F}$ with $\widetilde{E} \in S_{1\widetilde{\omega}}$, $\widetilde{F} \in S_{2\widetilde{\omega}}$. A pair of random elements X, Y defined on a fixed probability space (Ω, S_ω, P) with values in the spaces $(\widetilde{\Omega}_1, S_{1\widetilde{\omega}})$ and $(\widetilde{\Omega}_2, S_{2\widetilde{\omega}})$ respectively can be considered as a single random element (X, Y), called the **direct product** of X and Y, with values in the space $(\widetilde{\Omega}_1 \times \widetilde{\Omega}_2, S_{1\widetilde{\omega}} \times S_{2\widetilde{\omega}})$. The distribution $P_{(X,Y)}(\cdot) = P_{XY}(\cdot)$ of (X, Y) is called the **joint distribution** of X and Y. If X and Y are two random variables defined on a fixed probability space (Ω, S_ω, P), we introduce $E_x = \{\omega|X(\omega) \leq x\} \in S_\omega$, $E_y = \{\omega|Y(\omega) \leq y\} \in S_\omega$ and

$$F_{XY}(x, y) = P(E_x \cap E_y),$$

which is known as the **joint probability distribution function**. The **joint probability density function** is defined by $p_{XY}(x, y) = \partial^2 F_{XY}(x, y)/\partial x \partial y$. We have the following properties:

(1) $F_{XY}(x, y)$ is a monotone increasing function of both x and y.
(2) $p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x, \eta)d\eta$ (called marginal density).

The **conditional probability** of the event $Y \leq y$ given that $x < X \leq x + h$ is

$$P(\omega|Y(\omega) \leq y,\ x < X(\omega) \leq x + h) = \frac{\int_x^{x+h} \int_{-\infty}^{y} p_{XY}(\xi, \eta) d\xi d\eta}{\int_x^{x+h} p_X(\xi) d\xi}$$

from (8.1). Making $h \to 0$, we have

$$P(\omega|Y(\omega) \leq y,\ X(\omega) = x) = \frac{\int_{-\infty}^{y} p_{XY}(x, \eta) d\eta}{p_X(x)}.$$

The **conditional probability density function** of Y given that $X = x$ is defined by

$$p_Y(y|X = x) = \frac{p_{XY}(x, y)}{p_X(x)}.$$

Similarly, one can define $p_X(x|Y = y)$. Two random variables are said to be **statistically independent** when

$$p_{XY}(x, y) = p_X(x) p_Y(y).$$

In the case of statistical independence

$$p_X(x|Y = y) = p_X(x), \quad p_Y(y|X = x) = p_Y(y).$$

Therefore, the conditioning has no effects when there is statistical independence. Consider a one–one mapping from (x, y)-plane to the (ξ, η)-plane. The probability density functions in (x, y)-plane and (ξ, η)-plane are denoted by $p_{XY}(x, y)$ and $q(\xi, \eta)$ respectively. Suppose that an arbitrary domain D in (x, y)-plane is mapped to D' in (ξ, η)-plane. Then

$$P[(x, y) \in D] = P[(\xi, \eta) \in D'],$$

which implies

$$\iint_D p_{XY}(x, y) dx\, dy = \iint_{D'} q(\xi, \eta) d\xi\, d\eta = \iint_D q(\xi, \eta) \frac{\partial(\xi, \eta)}{\partial(x, y)} dx\, dy,$$

where $\partial(\xi, \eta)/\partial(x, y)$ is the Jacobian of the mapping

$$\frac{\partial(\xi, \eta)}{\partial(x, y)} = \begin{vmatrix} \partial\xi/\partial x & \partial\eta/\partial x \\ \partial\xi/\partial y & \partial\eta/\partial y \end{vmatrix}.$$

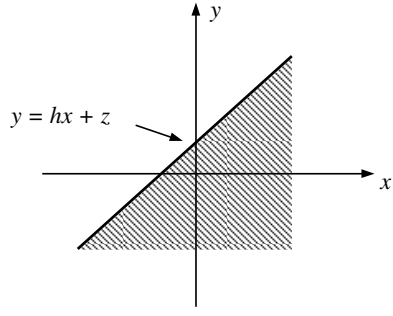


Figure 8.2 Area interpretation of probability $P(Y \leq y)$.

Since D is arbitrary, we have

$$p_{XY}(x, y) = q(\xi, \eta)\frac{\partial(\xi, \eta)}{\partial(x, y)}.$$

Example 8.2: Let X, Y and Z be random variables and $y = hx + z$, where h is a constant. Referring to Figure 8.2, the probability $P(Y \leq y)$ can be interpreted as the shaded area. Thus

$$P(Y \leq y) = P(hx + z \leq y) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{y-hx} p_{XZ}(x, z)dz$$

and

$$p_Y(y) = \frac{d}{dy}\int_{-\infty}^{\infty} dx \int_{-\infty}^{y-hx} p_{XZ}(x, z)dz = \int_{-\infty}^{\infty} p_{XZ}(x, y - hx)dx.$$

The above equation reduces to

$$p_Y(y) = \int_{-\infty}^{\infty} p_X(x)p_Z(y - hx)dx$$

if X and Z are independent. Thus

$$p_{XY}(x, y) = p_X(x)p_Z(y - hx).$$

The above discussions can be generalized to random vectors. Let $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ be random vectors and $\mathbf{y} = [H]\mathbf{x} + \mathbf{z}$, where $[H]$ is a matrix. Then we have

$$p_{\mathbf{Y}}(\mathbf{y}) = \int_{-\infty}^{\infty} p_{\mathbf{XZ}}(\mathbf{x}, \mathbf{y} - [\mathbf{H}]\mathbf{x})d\mathbf{x}.$$

If $\mathbf{X}$ and $\mathbf{Z}$ are independent, this reduces to

$$p_{\mathbf{Y}}(\mathbf{y}) = \int_{-\infty}^{\infty} p_{\mathbf{X}}(\mathbf{x}) p_{\mathbf{Z}}(\mathbf{y} - [H]\mathbf{x}) d\mathbf{x}.$$

Thus

$$p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = p_{\mathbf{X}}(\mathbf{x}) p_{\mathbf{Z}}(\mathbf{y} - [H]\mathbf{x}).$$

This is a useful relation and will be used later. □

8.1.3 Mathematical Expectations and Moments

Let X be a random variable and g be a function of X. The **mathematical expectation** (or **mean**) of g is defined by

$$\langle g(X) \rangle = \int_{-\infty}^{\infty} g(x) p_X(x) dx,$$

where $\langle \cdot \rangle$ is called the **expectation operator**. Especially, the **mean** of a random variable X is given by

$$m_X = \langle X \rangle = \int_{-\infty}^{\infty} x p_X(x) dx.$$

The **mean-square value** of X is defined by

$$\langle X^2 \rangle = \int_{-\infty}^{\infty} x^2 p_X(x) dx.$$

The nth **moment** of X is defined by

$$\langle X^n \rangle = \int_{-\infty}^{\infty} x^n p_X(x) dx.$$

The nth **central moment** of X is defined by

$$\langle (X - m_X)^n \rangle = \int_{-\infty}^{\infty} (x - m_X)^n p_X(x) dx.$$

The second central moment is called the **variance** of X

$$\mathrm{Var}(X) = \sigma_X^2 = \langle (X - m_X)^2 \rangle = \int_{-\infty}^{\infty} (x - m_X)^2 p_X(x) dx.$$

The square root of the variance σ_X is called the **standard deviation** of the random variable X.

For two random variables X and Y, the **joint moments** are defined by

$$\langle X^m Y^n \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^m y^n p_{XY}(x, y) dx\, dy.$$

Especially, the **correlation** of two random variables X and Y is $\langle XY \rangle$. The **covariance** of X and Y is

$$\mathrm{Cov}(XY) = \langle (X - m_X)(Y - m_Y) \rangle = \langle XY \rangle - m_X m_Y.$$

The covariance of X and Y normalized with respect to $\sigma_X \sigma_Y$ is called the **correlation coefficient** of X and Y

$$\rho = \frac{\mathrm{Cov}(XY)}{\sigma_X \sigma_Y}.$$

Two random variables X and Y are said to be **uncorrelated** iff $\mathrm{Cov}(XY) = 0$. If $\langle XY \rangle = 0$, we say that X and Y are **orthogonal**.

8.1.4 Stochastic Process

Let T be a subset of the real numbers. For every $t \in T$, let $X(\omega, t)$ be a random variable defined on a probability space (Ω, S_ω, P). Then $\{X(\omega, t) | t \in T\}$ is called a **real stochastic** (or **random**) **process** and T is called a **linear index set** or a **parameter set**. Let $Y(\omega, t)$ be another real stochastic process defined on the same probability space. Then $\{Z(\omega, t) = X(\omega, t) + jY(\omega, t) | t \in T\}$ is called a **complex stochastic process**. For both random variable and process, it is customary to drop the ω dependence. For $t_i \in T$ $(i = 1, 2, \ldots, n)$, we have n random variables $X(t_i)$ $(i = 1, 2, \ldots, n)$. The joint distribution of these random variables is

$$\begin{aligned} &F_{X(t_1)X(t_2)\ldots X(t_n)}(x_1, x_2, \ldots, x_n) \\ &\quad = P(X(t_1) \le x_1, X(t_2) \le x_2, \ldots, X(t_n) \le x_n) \end{aligned}$$

for a real random process. In the above $x_i = x(t_i)$, $i = 1, 2, \ldots, n$. The joint probability density function is given by

$$\begin{aligned} &p_{X(t_1)X(t_2)\ldots X(t_n)}(x_1, x_2, \ldots, x_n) \\ &\quad = \frac{\partial^n}{\partial x_1 \partial x_2 \ldots \partial x_n} F_{X(t_1)X(t_2)\ldots X(t_n)}(x_1, x_2, \ldots, x_n), \end{aligned}$$

which is called nth order probability density function. The random process $X(t)$ is said to be **stationary** of order n if

$$p_{X(t_1)X(t_2)\ldots X(t_n)}[x(t_1), x(t_2), \ldots, x(t_n)]$$
$$= p_{X(t_1+t_0)X(t_2+t_0)\ldots X(t_n+t_0)}[x(t_1+t_0), x(t_2+t_0), \ldots, x(t_n+t_0)]$$

for an arbitrary t_0. The random process $X(t)$ is said to be **strictly stationary** or **stationary in the strict sense** if it is stationary of any order. Many important properties of the stationary process can be described by first and second moments.

Remark 8.1: If the random process $Z(t)$ is complex, the joint n-dimensional distribution of random variables $Z(t_i)$ $(i = 1, 2, \ldots, n)$ will mean the joint $2n$-dimensional distribution of the real and imaginary components of $Z(t_i)$ $(i = 1, 2, \ldots, n)$. □

8.1.4.1 *Time-Average and Ensemble-Average*

Given a sample function $x(t)$, we may introduce the following time-average quantities

$$\overline{\overline{x(t)}} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)dt,$$

$$\overline{\overline{x^2(t)}} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x^2(t)dt,$$

$$\overline{\overline{[x(t) - \overline{\overline{x(t)}}]^2}} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} [x(t) - \overline{\overline{x(t)}}]^2 dt,$$

$$\overline{\overline{x(t+\tau)\overline{\overline{x(t)}}}} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t+\tau)\bar{x}(t)dt,$$

which are called **mean**, **mean square**, **variance**, and **autocorrelation**, respectively. Given a stochastic process $\{X(t)|t \in T\}$, we may introduce the ensemble-average quantities:

$$m_X(t) = \langle X(t)\rangle,$$
$$R_{XX}(t,s) = \langle X(t)\bar{X}(s)\rangle,$$
$$K_{XX}(t,s) = \langle [X(t) - m_X(t)][\bar{X}(s) - \bar{m}_X(s)]\rangle.$$

They are called **mean**, **autocorrelation**, and **autocovariance**, respectively. For a strictly stationary process, the mean of the random process is a constant and the autocorrelation and autocovariance functions depend on the time difference $t - s$ only

$$\begin{aligned} m_X(t) &= m_X = \text{constant}, \\ R_{XX}(t,s) &= R_{XX}(t-s), \\ K_{XX}(t,s) &= K_{XX}(t-s). \end{aligned} \tag{8.2}$$

The random process $X(t)$ is said to be **stationary in the wide sense** or **weakly stationary**, or **stationary to the second order**, if (8.2) holds. For convenience, the autocorrelation function of a stationary process is denoted as

$$R_{XX}(\tau) = \langle X(t+\tau)\bar{X}(t)\rangle.$$

A stationary process $X(t)$ is said to be **ergodic** if

$$m_X = \overline{\overline{x(t)}}, \quad R_{XX}(\tau) = \overline{\overline{x(t+\tau)\bar{x}(t)}},$$

where $x(t)$ is a sample function. In this case, the ensemble-average is equal to the time-average.

8.1.4.2 *Power Spectral Density*

The **power spectral density** (PSD) of the random process $X(t)$ is defined as the Fourier transform of the autocorrelation function

$$S_{XX}(\omega) = \int_{-\infty}^{\infty} R_{XX}(\tau)e^{-j\omega\tau}d\tau. \tag{8.3}$$

Thus

$$R_{XX}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{XX}(\omega)e^{j\omega\tau}d\omega. \tag{8.4}$$

Equations (8.3) and (8.4) are the Wiener–Khintchine relations. We have the following properties:

(1) $R_{XX}(\tau) = R_{XX}(-\tau)$.
(2) $|R_{XX}(\tau)| \leq R_{XX}(0)$.
(3) $S_{XX}(\omega) \geq 0$.
(4) $S_{XX}(\omega) = S_{XX}(-\omega)$.

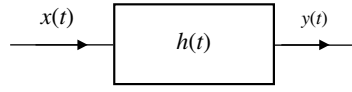


Figure 8.3 LTI system.

$X(t)$ is called **a white noise process** if the PSD is a constant

$$S_{XX}(\omega) = \frac{N_0}{2}.$$

The autocorrelation function of the white noise process may then be written as

$$R_{XX}(\tau) = \frac{N_0}{2}\delta(\tau).$$

Consider a linear time-invariant (LTI) system with impulse response $h(t)$ and transfer function $H(\omega)$ (Figure 8.3), where $h(t)$ and $H(\omega)$ constitute a Fourier transform pair. Suppose that $x(t)$ is a sample function of a stationary stochastic process $X(t)$ and is the input of the system. Then the output $y(t)$ is a sample function of a stochastic process $Y(t)$. We have the following properties:

(1) $m_Y = \langle Y(t)\rangle = \langle X(t)\rangle \cdot H(0) = m_X \cdot H(0)$,
(2) $R_{YY}(\tau) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\alpha)h(\beta)R_{XX}(\tau+\alpha-\beta)d\,\alpha d\,\beta$,
(3) $S_{YY}(\omega) = |H(\omega)|^2 S_{XX}(\omega)$.

We may adopt a time-average approach to investigate the relationship between correlation function and PSD. For two arbitrary functions $x_1(t)$ and $x_2(t)$, the **Parseval identity** is given by

$$\int_{-\infty}^{\infty} x_1(t)\bar{x}_2(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{x}_1(\omega)\bar{\tilde{x}}_2(\omega)d\omega, \tag{8.5}$$

where $\tilde{x}_1(\omega)$ and $\tilde{x}_2(\omega)$ are the Fourier transforms of $x_1(t)$, and $x_2(t)$, respectively. For a sample function $x(t)$, we introduce the truncated function

$$x_T(t) = \begin{cases} x(t), & |t| \le T/2 \\ 0, & |t| > T/2 \end{cases}$$

and its Fourier transform is denoted by $\tilde{x}_T(\omega)$. From (8.5), the autocorrelation function of $x_T(t)$ can be expressed as

$$\frac{1}{T}\int_{-T/2}^{T/2} x_T(t+\tau)\bar{x}_T(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{1}{T}\left|\tilde{x}_T(\omega)\right|^2 e^{j\omega\tau}d\omega.$$

Taking the ensemble-average gives

$$\frac{1}{T}\int_{-T/2}^{T/2}\langle x_T(t+\tau)\bar{x}_T(t)\rangle dt = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{1}{T}\langle|\tilde{x}_T(\omega)|^2\rangle e^{j\omega\tau}d\omega.$$

For a stationary process, we have

$$\lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2}\langle x_T(t+\tau)\bar{x}_T(t)\rangle dt = R_{XX}(\tau).$$

Thus

$$R_{XX}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\lim_{T\to\infty}\frac{1}{T}\langle|\tilde{x}_T(\omega)|^2\rangle e^{j\omega\tau}d\omega.$$

The PSD can be identified as

$$S_{XX}(\omega) = \lim_{T\to\infty}\frac{1}{T}\langle|\tilde{x}_T(\omega)|^2\rangle.$$

8.1.5 Gaussian Process

A real random variable X is said to have a **Gaussian distribution** (**normally distributed**) if its probability density function has the form

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X}e^{-(x-m_X)^2/2\sigma_X^2} \tag{8.6}$$

with $m_X = \langle X\rangle$ and $\sigma_X^2 = \langle(X-m_X)^2\rangle$. In general, a real random vector $\mathbf{X} = [X_1, X_2, \ldots, X_n]^T$ is said to be **normally distributed** if its probability density function has the form (n-dimensional Gaussian density function)

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}(\det[\Sigma])^{1/2}}e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m_X})^T[\Sigma]^{-1}(\mathbf{x}-\mathbf{m_X})} \tag{8.7}$$

with $\mathbf{m_X} = \langle\mathbf{X}\rangle$, $[\Sigma] = \langle[(\mathbf{X}-\mathbf{m_X})(\mathbf{X}-\mathbf{m_X})^T]\rangle$. A random process $X(t)$ is said to be a **Gaussian process** if the set of random variables

$X(t_1), X(t_2), \ldots, X(t_n)$, obtained by sampling $x(t)$ at times $t_1, t_2, \ldots, t_n$ are jointly Gaussian for any n. We have the following properties:

(1) If the input to a stable linear filter is a Gaussian process the output is also Gaussian.
(2) If a Gaussian process is wide-sense stationary, the process is also stationary in the strict-sense.
(3) If the set of random variables $X(t_1), X(t_2), \ldots, X(t_n)$, obtained by sampling a Gaussian process at times $t_1, t_2, \ldots, t_n$ are uncorrelated, i.e.,

$$\langle [X(t_k) - m_{X(t_k)}][X(t_i) - m_{X(t_i)}] \rangle = 0, \quad k \neq i,$$

then this set of random variables are statistically independent.

Let $Z(t) = X(t) + jY(t)$ be a complex random process. If the joint distribution of $X(t_1), Y(t_1), \ldots, X(t_n), Y(t_n)$ is $2n$-dimensional Gaussian for any choice of sample points $t_1, t_2, \ldots, t_n$ and for any n, then $Z(t)$ is said to be a **complex Gaussian process**.

8.1.6 Complex Gaussian Density Function

Let $[Q] \in C^{n \times n}$ be a positive definite Hermitian matrix. The **complex Gaussian density function** for an n-dimensional complex random variable $\mathbf{Z}$ is defined by

$$p_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{\pi^n \det[Q]} e^{-(\mathbf{z}-\boldsymbol{\mu})^\dagger [Q]^{-1} (\mathbf{z}-\boldsymbol{\mu})}, \quad \mathbf{z},\ \boldsymbol{\mu} \in C^n. \tag{8.8}$$

Then (Miller, 1974)

$$\begin{aligned}(\mathbf{z}-\boldsymbol{\mu})^\dagger [Q]^{-1} (\mathbf{z}-\boldsymbol{\mu}) &= \frac{1}{2}(\mathbf{z}_c - \boldsymbol{\mu}_c)^T [Q_a]^{-1} (\mathbf{z}_c - \boldsymbol{\mu}_c) \\ &= \frac{1}{2}(\mathbf{z}_c - \boldsymbol{\mu}_c)^\dagger [Q_d]^{-1} (\mathbf{z}_c - \boldsymbol{\mu}_c) \\ &= \frac{1}{2}(\widehat{\mathbf{z}} - \widehat{\boldsymbol{\mu}})^\dagger [\widehat{Q}]^{-1} (\widehat{\mathbf{z}} - \widehat{\boldsymbol{\mu}}),\end{aligned}$$

where

$$\widehat{\mathbf{z}} = \begin{bmatrix}\mathbf{x} \\ \mathbf{y}\end{bmatrix} = \begin{bmatrix}\mathrm{Re}(\mathbf{z}) \\ \mathrm{Im}(\mathbf{z})\end{bmatrix}, \quad \widehat{\boldsymbol{\mu}} = \begin{bmatrix}\mathrm{Re}(\boldsymbol{\mu}) \\ \mathrm{Im}(\boldsymbol{\mu})\end{bmatrix}, \quad [\widehat{Q}] = \frac{1}{2}\begin{bmatrix}\mathrm{Re}[Q] & -\mathrm{Im}[Q] \\ \mathrm{Im}[Q] & \mathrm{Re}[Q]\end{bmatrix},$$

$$\mathbf{z}_c = \begin{bmatrix}\mathbf{z} \\ \bar{\mathbf{z}}\end{bmatrix}, \quad \boldsymbol{\mu}_c = \begin{bmatrix}\boldsymbol{\mu} \\ \widehat{\boldsymbol{\mu}}\end{bmatrix}, \quad [Q_d] = \begin{bmatrix}[Q] & 0 \\ 0 & [\bar{Q}]\end{bmatrix}, \quad [Q_a] = \begin{bmatrix}0 & [Q] \\ [\bar{Q}] & 0\end{bmatrix}.$$

Note that

$$\det[\widehat{Q}] = 2^{-2n}(\det[Q])^2, \quad \det[Q_a] = (-1)^n(\det[Q])^2, \quad \det[Q_d] = (\det[Q])^2.$$

Thus (8.8) may be written as

$$\begin{aligned} p_{\mathbf{Z}}(\mathbf{z}) &= \frac{1}{\pi^n \det([Q_d])^{1/2}} e^{-\frac{1}{2}(\mathbf{z}_c-\boldsymbol{\mu}_c)^\dagger [Q_d]^{-1}(\mathbf{z}_c-\boldsymbol{\mu}_c)} \\ &= \frac{1}{(2\pi)^n \det[\widehat{Q}]^{1/2}} e^{-\frac{1}{2}(\widehat{\mathbf{z}}-\widehat{\boldsymbol{\mu}})^\dagger [\widehat{Q}]^{-1}(\widehat{\mathbf{z}}-\widehat{\boldsymbol{\mu}})}. \end{aligned}$$

It follows that

$$\langle \mathbf{Z} \rangle = \boldsymbol{\mu}, \quad \langle [(\mathbf{Z}-\boldsymbol{\mu})(\mathbf{Z}-\boldsymbol{\mu})^\dagger \rangle = [Q], \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(\mathbf{z}) d\mathbf{x}\, d\mathbf{y} = 1.$$

8.1.7 Analytic Representation

Let us consider a real random signal $s(t)$ with its Fourier transform denoted by $\tilde{s}(\omega)$. Then

$$\begin{aligned} s(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{s}(\omega) e^{j\omega t} d\omega \\ &= \frac{1}{2\pi}\left[\int_0^{\infty} \tilde{s}(\omega) e^{j\omega t} d\omega + \int_0^{\infty} \bar{\tilde{s}}(\omega) e^{-j\omega t} d\omega\right] = \operatorname{Re} s_a(t), \end{aligned}$$

where $s_a(t) = \frac{1}{\pi}\int_0^{\infty} \tilde{s}(\omega) e^{j\omega t} d\omega$ is referred to as the **analytic representation** of $s(t)$. The Fourier transform of the analytic signal is given by

$$\begin{aligned} \tilde{s}_a(\omega) &= \int_{-\infty}^{\infty} s_a(t) e^{-j\omega t} dt = \frac{1}{\pi}\int_{-\infty}^{\infty} dt \int_0^{\infty} \tilde{s}(\omega') e^{j(\omega'-\omega)t} d\omega' \\ &= \frac{1}{\pi}\int_0^{\infty} \tilde{s}(\omega') 2\pi\delta(\omega'-\omega) d\omega' = \begin{cases} 2\tilde{s}(\omega), & \omega > 0 \\ 0, & \omega < 0 \end{cases}. \end{aligned}$$

Therefore, the analytic representation has no negative frequency component, which indicates that the negative frequency components of a real valued signal are superfluous due to the spectral symmetry. The introduction of analytic representation allows us to use complex variable analysis to study the random process. The autocorrelation function of the

analytic representation is called the **coherence function**, denoted by

$$\Gamma(\tau) = \langle s_a(t+\tau)\bar{s}_a(t)\rangle.$$

If $\mathbf{E}_a(\mathbf{r},t)$ is the analytic representation of a random vector field. The **coherence tensor** of the electric field is defined as the ensemble-average of the dyad $\mathbf{E}_a(\mathbf{r}_1,t_1)\bar{\mathbf{E}}_a(\mathbf{r}_2,t_2)$

$$\overleftrightarrow{\mathbf{\Gamma}}(\mathbf{r}_1,\mathbf{r}_2,t_1,t_2) = \langle \mathbf{E}_a(\mathbf{r}_1,t_1)\bar{\mathbf{E}}_a(\mathbf{r}_2,t_2)\rangle,$$

which can be written as

$$\overleftrightarrow{\mathbf{\Gamma}}(\mathbf{r}_1,\mathbf{r}_2,t_1,t_2) = \begin{bmatrix} \Gamma_{11}(\mathbf{r}_1,\mathbf{r}_2,t_1,t_2) & \Gamma_{12}(\mathbf{r}_1,\mathbf{r}_2,t_1,t_2) & \Gamma_{13}(\mathbf{r}_1,\mathbf{r}_2,t_1,t_2) \\ \Gamma_{21}(\mathbf{r}_1,\mathbf{r}_2,t_1,t_2) & \Gamma_{22}(\mathbf{r}_1,\mathbf{r}_2,t_1,t_2) & \Gamma_{23}(\mathbf{r}_1,\mathbf{r}_2,t_1,t_2) \\ \Gamma_{31}(\mathbf{r}_1,\mathbf{r}_2,t_1,t_2) & \Gamma_{32}(\mathbf{r}_1,\mathbf{r}_2,t_1,t_2) & \Gamma_{33}(\mathbf{r}_1,\mathbf{r}_2,t_1,t_2) \end{bmatrix}$$

with $\Gamma_{ij}(\mathbf{r}_1,\mathbf{r}_2,t_1,t_2) = \langle E_i(\mathbf{r}_1,t_1)\bar{E}_j(\mathbf{r}_2,t_2)\rangle$, $i,j = 1,2,3$. If the field is stationary, we have $\overleftrightarrow{\mathbf{\Gamma}}(\mathbf{r}_1,\mathbf{r}_2,t_1,t_2) = \overleftrightarrow{\mathbf{\Gamma}}(\mathbf{r}_1,\mathbf{r}_2,\tau)$. If the field is ergodic, then

$$\overleftrightarrow{\mathbf{\Gamma}}(\mathbf{r}_1,\mathbf{r}_2,\tau) = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2}\mathbf{E}(\mathbf{r}_1,t+\tau)\bar{\mathbf{E}}(\mathbf{r}_2,t)dt.$$

Note that $\overleftrightarrow{\mathbf{\Gamma}}(\mathbf{r}_1,\mathbf{r}_2,\tau) = \overleftrightarrow{\mathbf{\Gamma}}^{\dagger}(\mathbf{r}_2,\mathbf{r}_1,-\tau)$, where "†" denotes the Hermitian. The **power spectral density tensor** is the Fourier transform of the coherence tensor

$$\overleftrightarrow{\mathbf{P}}(\mathbf{r}_1,\mathbf{r}_2,\omega) = \int_{-\infty}^{\infty}\overleftrightarrow{\mathbf{\Gamma}}(\mathbf{r}_1,\mathbf{r}_2,\tau)e^{-j\omega\tau}d\tau$$

with

$$\overleftrightarrow{\mathbf{P}}(\mathbf{r}_1,\mathbf{r}_2,\omega) = \begin{bmatrix} P_{11}(\mathbf{r}_1,\mathbf{r}_2,\omega) & P_{12}(\mathbf{r}_1,\mathbf{r}_2,\omega) & P_{13}(\mathbf{r}_1,\mathbf{r}_2,\omega) \\ P_{21}(\mathbf{r}_1,\mathbf{r}_2,\omega) & P_{22}(\mathbf{r}_1,\mathbf{r}_2,\omega) & P_{23}(\mathbf{r}_1,\mathbf{r}_2,\omega) \\ P_{31}(\mathbf{r}_1,\mathbf{r}_2,\omega) & P_{32}(\mathbf{r}_1,\mathbf{r}_2,\omega) & P_{33}(\mathbf{r}_1,\mathbf{r}_2,\omega) \end{bmatrix},$$

$$P_{ij}(\mathbf{r}_1,\mathbf{r}_2,\omega) = \int_{-\infty}^{\infty}\Gamma_{ij}(\mathbf{r}_1,\mathbf{r}_2,\tau)e^{-j\omega\tau}d\tau.$$

Similarly, we have $\overleftrightarrow{\mathbf{P}}(\mathbf{r}_1,\mathbf{r}_2,\omega) = \overleftrightarrow{\mathbf{P}}^{\dagger}(\mathbf{r}_2,\mathbf{r}_1,-\omega)$. It is easy to show that in free space, Γ_{ij} and P_{ij} satisfy the following equations

$$\nabla_1^2\Gamma_{ij}(\mathbf{r}_1,\mathbf{r}_2,\tau) - \frac{1}{c^2}\frac{\partial^2}{\partial\tau^2}\Gamma_{ij}(\mathbf{r}_1,\mathbf{r}_2,\tau) = 0,$$

$$\nabla_1^2 P_{ij}(\mathbf{r}_1,\mathbf{r}_2,\omega) + k^2 P_{ij}(\mathbf{r}_1,\mathbf{r}_2,\omega) = 0,$$

where $k = \omega/c$, and the subscript "1" stands for the differential operation with respect to $\mathbf{r}_1$.

8.1.8 Narrow-Band Stationary Stochastic Process

Digital modulation is a process by which digital symbols are transformed into waveforms that are compatible with the characteristics of the channel. In the case of baseband modulation, these waveforms are pulses, but in the case of bandpass modulation the desired information signal modulates a sinusoid called a **carrier**. Modulation methods determine the system bandwidth, power efficiency, sensitivity, and complexity. Digital modulation offers many advantages over analog modulation and may be broadly classified as **linear** and **nonlinear**. In linear modulation techniques, the amplitude of the transmitted signal varies linearly with modulating digital signal. Linear modulation techniques are bandwidth efficient and hence are very useful in wireless communication systems. As a linear modulation technique, an easy way to translate the spectrum of low-pass or baseband signal $a(t)$ to a higher frequency is to multiply or **heterodyne** the baseband signal with a carrier wave. The resulting waveform (assumed to be a sample function of a wide-sense stationary stochastic process) is called a **double-sided bandpass** (DSB) **signal** and can be represented by

$$s(t) = \begin{cases} a(t)\cos[\omega_c t + \varphi(t)], \\ x(t)\cos\omega_c t - y(t)\sin\omega_c t, \\ \operatorname{Re} s_{\text{en}}(t)e^{j\omega_c t}, \end{cases} \tag{8.9}$$

where $\omega_c = 2\pi f_c$, $a(t)$ and $\varphi(t)$ are the **carrier frequency**, **envelope**, and **phase** of the modulated signal respectively, and

$$\begin{aligned} s_{\text{en}}(t) &= x(t) + jy(t), \\ x(t) &= a(t)\cos\varphi(t), \\ y(t) &= a(t)\sin\varphi(t), \end{aligned}$$

where $s_{\text{en}}(t)$, $x(t)$, and $y(t)$ are called **complex envelope**, **in-phase component**, and **quadrature component** of the modulated signal. It is clear from (8.9) that the amplitude of the carrier varies linearly with the modulating signal. Linear modulation schemes in general do not have constant envelope. The spectrum of the DSB signal is given by

$$\tilde{s}(f) = \frac{1}{2}\left[\tilde{a}(f - f_c) + \tilde{a}(f + f_c)\right],$$

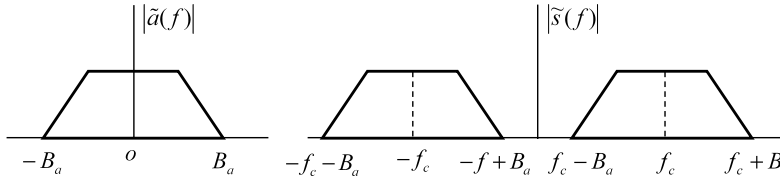


Figure 8.4 Spectrum of DSB signal.

where $\tilde{s}(f)$ is the Fourier transform of $s(t)$. If the baseband signal $a(t)$ has a bandwidth B_a, the bandwidth of the DSB will be $B_s = 2B_a$, as illustrated in Figure 8.4. That is, we need twice as much transmission bandwidth to transmit a DSB version of the signal than we do to transmit its baseband counterpart. The autocorrelation function $R_{SS}(\tau)$ of $s(t)$ is

$$\begin{aligned} R_{SS}(\tau) &= \langle S(t+\tau)S(t)\rangle \\ &= \frac{1}{2}\left[R_{XX}(\tau)+R_{YY}(\tau)\right]\cos\omega_c\tau \\ &\quad + \frac{1}{2}\left[R_{XX}(\tau)-R_{YY}(\tau)\right]\cos\omega_c(2t+\tau) \\ &\quad - \frac{1}{2}\left[R_{YX}(\tau)-R_{XY}(\tau)\right]\sin\omega_c\tau \\ &\quad - \frac{1}{2}\left[R_{YX}(\tau)+R_{XY}(\tau)\right]\sin\omega_c(2t+\tau). \end{aligned}$$

Since $s(t)$ is stationary, the right-hand side must be independent of t, which implies

$$R_{XX}(\tau) = R_{YY}(\tau), \quad R_{XY}(\tau) = -R_{YX}(\tau).$$

Note that the cross-correlation function satisfies

$$R_{XY}(-\tau) = R_{YX}(\tau), \quad R_{XY}(\tau) = -R_{XY}(-\tau)$$

and $R_{XY}(0) = 0$. Thus

$$R_{SS}(\tau) = R_{XX}(\tau)\cos\omega_c\tau - R_{YX}(\tau)\sin\omega_c\tau. \tag{8.10}$$

The autocorrelation function of the complex envelope is given by

$$R_{S_{\text{en}}S_{\text{en}}}(\tau) = \langle S_{\text{en}}(t+\tau)\overline{S}_{\text{en}}(t)\rangle = 2[R_{XX}(\tau)+jR_{YX}(\tau)]. \tag{8.11}$$

It follows from (8.10) and (8.11) that

$$R_{SS}(\tau) = \frac{1}{2}\mathrm{Re}\big[R_{S_{\mathrm{en}}S_{\mathrm{en}}}(\tau)e^{j\omega_c\tau}\big].$$

The PSD of $s(t)$ is the Fourier transform of $R_{SS}(\tau)$ and can be expressed as

$$S_{SS}(\omega) = \frac{1}{4}[S_{S_{\mathrm{en}}S_{\mathrm{en}}}(\omega-\omega_c) + S_{S_{\mathrm{en}}S_{\mathrm{en}}}(-\omega-\omega_c)].$$

The definition of signal bandwidth varies with context, and there is no single definition that suits all applications. All definitions are however based on some measure on the PSD of the signal. The stochastic process $s(t)$ is said to be a **narrowband bandpass process** if the width of the spectral density is much smaller than the carrier frequency f_c.

Similar to (8.9), a narrowband bandpass stochastic vector field $\mathbf{F}$ in the time domain can be expressed as

$$\mathbf{F}(\mathbf{r},t) = \begin{cases} \mathbf{a}(\mathbf{r},t)\cos[\omega_c t + \varphi(\mathbf{r},t)], \\ \mathbf{x}(\mathbf{r},t)\cos\omega_c t - \mathbf{y}(\mathbf{r},t)\sin\omega_c t, \\ \mathrm{Re}\,\mathbf{F}_{\mathrm{en}}(\mathbf{r},t)e^{j\omega_c t}, \end{cases}$$

where

$$\begin{aligned} \mathbf{F}_{\mathrm{en}}(\mathbf{r},t) &= \mathbf{x}(\mathbf{r},t) + j\mathbf{y}(\mathbf{r},t), \\ \mathbf{x}(\mathbf{r},t) &= \mathbf{a}(\mathbf{r},t)\cos\varphi(\mathbf{r},t), \\ \mathbf{y}(\mathbf{r},t) &= \mathbf{a}(\mathbf{r},t)\sin\varphi(\mathbf{r},t). \end{aligned}$$

If the complex envelopes of electromagnetic fields are slowly varying functions of time compared to $e^{j\omega_c t}$, we have

$$\begin{cases} \nabla\times\mathbf{H}_{\mathrm{en}}(\mathbf{r},t) = j\omega_c\varepsilon\mathbf{E}_{\mathrm{en}}(\mathbf{r},t) + \mathbf{J}_{\mathrm{en}}(\mathbf{r},t), \\ \nabla\times\mathbf{E}_{\mathrm{en}}(\mathbf{r},t) = -j\omega\mu\mathbf{H}_{\mathrm{en}}(\mathbf{r},t). \end{cases} \tag{8.12}$$

Therefore, the complex envelopes of electromagnetic fields satisfy the time-harmonic Maxwell equations, and most of the theoretical results about the time-harmonic fields can be applied to the complex envelopes. Let $\overline{\overline{\mathbf{F}}}$ denote the time-average of $\mathbf{F}$: $\overline{\overline{\mathbf{F}}} = \lim\limits_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2}\mathbf{F}(t)dt$. For a stationary and ergodic electromagnetic field, we may take the time-average of (8.12) to obtain

$$\begin{cases} \nabla\times\overline{\overline{\mathbf{H}_{\mathrm{en}}(\mathbf{r})}} = j\omega_c\varepsilon\overline{\overline{\mathbf{E}_{\mathrm{en}}(\mathbf{r})}} + \overline{\overline{\mathbf{J}_{\mathrm{en}}(\mathbf{r})}}, \\ \nabla\times\overline{\overline{\mathbf{E}_{\mathrm{en}}(\mathbf{r})}} = -j\omega_c\mu\overline{\overline{\mathbf{H}_{\mathrm{en}}(\mathbf{r})}}. \end{cases} \tag{8.13}$$

Hence, most of the theoretical results about the time-harmonic fields can also be applied to the time-averages of the complex envelopes of the fields.

The autocorrelation function $R_{\mathbf{FF}}(\tau)$ of vector field $\mathbf{F}$ is defined by

$$R_{\mathbf{FF}}(\tau) = \langle \mathbf{F}(t+\tau) \cdot \bar{\mathbf{F}}(t) \rangle = R_{\mathbf{xx}}(\tau) \cos \omega_c \tau - R_{\mathbf{yx}}(\tau) \sin \omega_c \tau,$$

where we have assumed that $\mathbf{F}$ is stationary. The autocorrelation function of the complex envelope is then given by

$$R_{\mathbf{F}_{\text{en}}\mathbf{F}_{\text{en}}}(\tau) = \langle \mathbf{F}_{\text{en}}(t+\tau) \cdot \bar{\mathbf{F}}_{\text{en}}(t) \rangle = 2[R_{\mathbf{xx}}(\tau) + jR_{\mathbf{yx}}(\tau)]$$

and we also have

$$R_{\mathbf{FF}}(\tau) = \frac{1}{2} \text{Re}[R_{\mathbf{F}_{\text{en}}\mathbf{F}_{\text{en}}}(\tau) e^{j\omega_c \tau}].$$

The PSD of $\mathbf{F}(\mathbf{r}, t)$ is the Fourier transform of $R_{\mathbf{FF}}(\tau)$, which can be written as

$$S_{\mathbf{FF}}(\omega) = \frac{1}{4}[S_{\mathbf{F}_{\text{en}}\mathbf{F}_{\text{en}}}(\omega - \omega_c) + S_{\mathbf{F}_{\text{en}}\mathbf{F}_{\text{en}}}(-\omega - \omega_c)].$$

8.2 Information Theory

Information theory is essentially a branch of applied mathematics and has found applications in many areas. The most important quantities of information are entropy, the information in a random variable, and mutual information, the amount of information in common between two random variables.

8.2.1 System with One Random Variable

Let X be a discrete random variable with $P(X = x_j) = p_j$, $0 \leq p_j \leq 1$ and $\sum_{j=1}^{n} p_j = 1$. The **entropy** of X, denoted by $H(X)$, is defined by

$$H(X) = -\sum_{j=1}^{n} p_j \log p_j. \tag{8.14}$$

The base of logarithm is not specified in the definition. The unit of entropy is bits if base 2 is used or nats if base e is used. If $p_i = 0$, the term $p_i \log p_i$ in (8.14) is defined to be zero. If $n = \infty$, the sum (8.14) may not converge. In this case, we define $H(X) = +\infty$. The **differential entropy**

of a continuous random vector $\mathbf{X} = [X_1, X_2, \ldots, X_n]^T$ is defined by

$$H(\mathbf{X}) = -\int_{-\infty}^{\infty} p_{\mathbf{X}}(\mathbf{x}) \log p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x},$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ and $d\mathbf{x} = dx_1 dx_2 \ldots dx_n$.

Remark 8.2: The entropy is a measure of uncertainty of a random variable or is a measure of unpredictability of information content. The entropy of a continuous random variable needs not to exist. When it does exist, it can be negative. □

Example 8.3: The entropy of the Gaussian distribution given by (8.6) is $\log[(2\pi e)^{1/2}\sigma_X]$. Thus the entropy increases as σ_X increases. Making use of $\mathbf{x}^\dagger\mathbf{x} = \mathrm{Tr}(\mathbf{x}\mathbf{x}^\dagger)$ and (8.8), the entropy of an n-dimensional complex Gaussian random vector $\mathbf{Z}$ with covariance $[Q]$ can be written as

$$\begin{aligned} H(\mathbf{Z}) &= \langle -\log p_{\mathbf{Z}}(\mathbf{z}) \rangle \\ &= \log\det(\pi[Q]) + (\log e)\langle [(\mathbf{z}-\boldsymbol{\mu})^\dagger [Q]^{-1} (\mathbf{z}-\boldsymbol{\mu})] \rangle \\ &= \log\det(\pi[Q]) + (\log e)\mathrm{Tr}\{\langle [(\mathbf{z}-\boldsymbol{\mu})(\mathbf{z}-\boldsymbol{\mu})^\dagger] \rangle [Q]^{-1}\} \\ &= \log\det(\pi[Q]) + n\log e = \log\det(\pi e[Q]). \end{aligned} \tag{8.15}$$

Similarly, it follows from (8.7) that the entropy of an n-dimensional real Gaussian random vector $\mathbf{X}$ with covariance $[\Sigma]$ is

$$\begin{aligned} H(\mathbf{X}) &= \langle -\log p_{\mathbf{X}}(\mathbf{x}) \rangle \\ &= \frac{1}{2}\log\det(2\pi[\Sigma]) + \frac{1}{2}(\log e)\langle [(\mathbf{x}-\mathbf{m}_{\mathbf{X}})^T [\Sigma]^{-1} (\mathbf{x}-\mathbf{m}_{\mathbf{X}})] \rangle \\ &= \frac{1}{2}\log\det(2\pi[\Sigma]) + \frac{1}{2}(\log e)\mathrm{Tr}\{\langle [(\mathbf{x}-\mathbf{m}_{\mathbf{X}})(\mathbf{x}-\mathbf{m}_{\mathbf{X}})^T] \rangle [\Sigma]^{-1}\} \\ &= \frac{1}{2}\log\det(2\pi[\Sigma]) + \frac{1}{2}n\log e = \frac{1}{2}\log\det(2\pi e\boldsymbol{\Sigma}). \end{aligned} \tag{8.16}$$

Note the difference between (8.15) and (8.16). □

We note that the probability density which gives the greatest differential entropy subject to the restriction

$$\int_{-\infty}^{\infty} x^2 p_X(x) dx = \sigma_X^2$$

is the Gaussian distribution with zero mean and variance σ_X^2 (e.g., Jones, 1979b).

8.2.2 System with Two Random Variables

Let X be a discrete random variable with $P(X = x_j) = p_j$, $0 \leq p_j \leq 1$ and $\sum_{j=1}^{n} p_j = 1$. Let Y be another discrete random variable with $P(Y = y_k) = q_k$, $0 \leq q_k \leq 1$ and $\sum_{k=1}^{m} q_k = 1$. The connection between X and Y is obtained by specifying

$$P(X = x_j, Y = y_k) = p_{jk}$$

subject to $p_{jk} \geq 0$, and $\sum_{j=1}^{n} \sum_{k=1}^{m} p_{jk} = 1$. Evidently, we have

$$p_j = \sum_{k=1}^{m} p_{jk}, \quad q_k = \sum_{j=1}^{n} p_{jk},$$

$$P(X = x_j | Y = y_k) = \frac{P(X = x_j, Y = y_k)}{P(Y = y_k)} = \frac{p_{jk}}{q_k}.$$

The **conditional entropy** $H(X|Y)$ is defined by

$$H(X|Y) = -\sum_{j=1}^{n} \sum_{k=1}^{m} p_{jk} \log P(X = x_j | Y = y_k) = -\sum_{j=1}^{n} \sum_{k=1}^{m} p_{jk} \log \frac{p_{jk}}{q_k}.$$

The **joint entropy** $H(X \cap Y)$ of X and Y is defined by

$$H(X \cap Y) = -\sum_{j=1}^{n} \sum_{k=1}^{m} p_{jk} \log p_{jk}.$$

It is easy to show that

(1) $H(X \cap Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$ (**chain rule**).
(2) $H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X \cap Y)$.
(3) $H(X|Y) \leq H(X)$ with equality only if X and Y are statistically independent.

The **mutual information** $I(X, Y)$ between X and Y is defined by

$$I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = I(Y, X).$$

The mutual information $I(X, Y)$ is the reduction in the uncertainty of X due to the knowledge of Y. Note that $I(X, Y) > 0$.

Let us consider two continuous random vectors $\mathbf{X} = [X_1, X_2, \ldots, X_n]^T$ and $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_n]^T$ with probability density $p_{\mathbf{X}}(\mathbf{x})$ and $p_{\mathbf{Y}}(\mathbf{y})$ respectively. The joint distribution of the two random vectors is $p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})$. The joint entropy, conditional entropy and mutual information are respectively defined by

$$H(\mathbf{X} \cap \mathbf{Y}) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) \log p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) d\mathbf{x}\, d\mathbf{y},$$

$$H(\mathbf{X}|\,\mathbf{Y}) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) \log \frac{p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})}{p_{\mathbf{Y}}(\mathbf{y})} d\mathbf{x}\, d\mathbf{y},$$

$$I(\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}) - H(\mathbf{X}|\mathbf{Y}). \tag{8.17}$$

We have the following properties:

(1) $H(\mathbf{X} \cap \mathbf{Y}) = H(\mathbf{Y}) + H(\mathbf{X}|\mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}|\mathbf{X})$.
(2) $I(\mathbf{X}, \mathbf{Y}) = I(\mathbf{Y}, \mathbf{X}) = H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{X} \cap \mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X})$.
(3) $I(\mathbf{X}, \mathbf{Y}) \geq 0$, $H(\mathbf{X}) \geq H(\mathbf{X}|\mathbf{Y})$ with equality iff $\mathbf{X}$ and $\mathbf{Y}$ are statistically independent.

Note that $I(\mathbf{X}, \mathbf{X}) = H(\mathbf{X})$. Thus entropy is actually the self-information of a random variable.

8.2.3 System with More Than Two Random Variables

Let X be a discrete random variable with $P(X = x_j) = p_j$, $0 \leq p_k \leq 1$ and $\sum_{j=1}^{n} p_j = 1$. Let Y be the second discrete random variable with $P(Y = y_k) = q_k$, $0 \leq q_k \leq 1$ and $\sum_{k=1}^{m} q_k = 1$. Let Z be the third discrete random variable with $P(Z = z_l) = s_l$, $0 \leq s_l \leq 1$ and $\sum_{l=1}^{r} s_l = 1$. The connection between the three random variables is obtained by specifying

$$P(X = x_j, Y = y_k, Z = z_l) = p_{jkl}$$

subject to $p_{jkl} \geq 0$ and $\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{r} p_{jkl} = 1$. It is easy to show that

$$p_{jk} = P(X = x_j, Y = y_k) = \sum_{l=1}^{r} p_{jkl},$$

$$p_{jl} = P(X = x_j, Z = z_l) = \sum_{k=1}^{r} p_{jkl},$$

$$p_j = P(X = x_j) = \sum_{k=1}^{m}\sum_{l=1}^{r} p_{jkl}.$$

X and Y are said to be **statistically independent when conditioned on** Z if

$$P(X = x_j, Y = y_k|Z = z_l) = P(X = x_j|Z = z_l)P(Y = y_k|Z = z_l)$$

for all j, k, and l such that $P(Z = z_l) \neq 0$. We introduce the following entropies

$$H(X|Y \cap Z) = -\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{r} p_{jkl} \log P(X = x_j|Y = y_k, Z = z_l),$$

$$H(X \cap Y|Z) = -\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{r} p_{jkl} \log P(X = x_j, Y = y_k|Z = z_l).$$

The **mutual information** $I(X \cap Y, Z)$ are defined by

$$I(X \cap Y, Z) = H(X \cap Y) - H(X \cap Y|Z).$$

The **mutual information between** X **and** Y **conditioned on** Z is defined by

$$I(X, Y|Z) = H(X|Z) - H(X|Y \cap Z).$$

The following properties hold

(1) $I(X, Y|Z) \geq 0$ with equality only if X and Y are said to be statistically independent when conditioned on Z.
(2) $I(X \cap Y, Z) = I(X, Z) + I(Y, Z|X) = I(Y, Z) + I(X, Z|Y)$.
(3) **Data processing theorem**: If X and Z are statistically independent when conditioned on Y, then $I(X, Z) \leq I(Y, Z), I(X, Z) \leq I(X, Y)$.
(4) $I(X, Y) \geq 0$ with equality only if X and Y are statistically independent.

Let us consider three random vectors $\mathbf{X} = [X_1, X_2, \ldots, X_n]^T$, $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_n]^T$, and $\mathbf{Z} = [Z_1, Z_2, \ldots, Z_n]^T$ with probability density $p_{\mathbf{X}}(\mathbf{x})$, $p_{\mathbf{Y}}(\mathbf{y})$, and $p_{\mathbf{Z}}(\mathbf{z})$ respectively. The joint distribution of the three random vectors is $p_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$. We introduce the following conditional entropies

$$H(\mathbf{X}|\mathbf{Y} \cap \mathbf{Z}) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z}) \log \frac{p_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z})}{p_{\mathbf{YZ}}(\mathbf{y}, \mathbf{z})} d\mathbf{x}\, d\mathbf{y}\, d\mathbf{z},$$

$$H(\mathbf{X} \cap \mathbf{Y}|\mathbf{Z}) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z}) \log \frac{p_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z})}{p_{\mathbf{Z}}(\mathbf{z})} d\mathbf{x}\, d\mathbf{y}\, d\mathbf{z},$$

and define the **conditional mutual information**

$$I(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) = H(\mathbf{X}|\mathbf{Z}) - H(\mathbf{X}|\mathbf{Y} \cap \mathbf{Z}).$$

Then

(1) $H(\mathbf{X} \cap \mathbf{Y}|\mathbf{Z}) = H(\mathbf{X}|\mathbf{Z}) + H(\mathbf{Y}|\mathbf{X} \cap \mathbf{Z})$.
(2) $I(\mathbf{X} \cap \mathbf{Y}, \mathbf{Z}) = I(\mathbf{X}, \mathbf{Z}) + I(\mathbf{Y}, \mathbf{Z}|\mathbf{X})$ (chain rule for the mutual information).

8.2.4 Channel Capacity of Deterministic MIMO System

The well-known **Shannon's continuous channel theorem** (Shannon, 1948) gives the fundamental limit on the rate of error-free transmission for a power-limited, band-limited Gaussian channel. This theorem indicates that the maximum information rate C depends on three key system parameters: channel bandwidth B, average transmitted power P and noise PSD N_0 at the channel output:

$$C = B \log\left(1 + \frac{P}{N_0 B}\right). \tag{8.18}$$

To approach this limit, the statistical properties of the transmitted signal must be like a white Gaussian noise. To stretch Shannon's above limit, one may use multiple antenna system. The block diagram of a multiple antenna system is shown in Figure 8.5. For an $n_t \times n_r$ MIMO system, we have n_t transmitting antennas and n_r receiving antennas. The input signal to the transmit antenna system is denoted by $\mathbf{X}(t) = [x_1(t), x_2(t), \ldots, x_{n_t}(t)]^T \in C^{n_t}$, which is a $n_t \times 1$ complex random matrix,

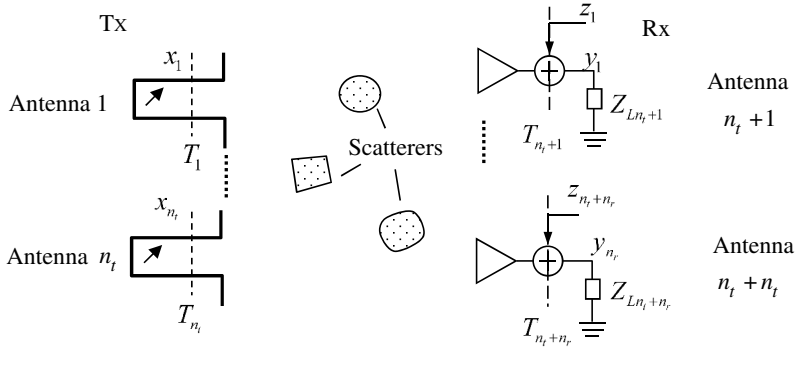


Figure 8.5 Generic MIMO system.

and the output of the receive antenna system is denoted by $\mathbf{Y}(t) = [y_1(t), y_2(t), \ldots, y_{n_t}(t)]^T \in C^{n_r}$, which is an $n_r \times 1$ complex random matrix given by

$$\mathbf{y}(t) = \int_{-\infty}^{\infty} [h(t,\tau)]\mathbf{x}(t-\tau)d\tau + \mathbf{z}(t), \tag{8.19}$$

where $\mathbf{Z}$ is $n_r \times 1$ additive white Gaussian noise (AWGN) matrix with zero mean at a given instant of time and $[h(t,\tau)]$ is the MIMO **channel response matrix**

$$[h(t,\tau)] = \begin{bmatrix} h_{11}(t,\tau) & h_{12}(t,\tau) & \cdots & h_{1n_t}(t,\tau) \\ h_{21}(t,\tau) & h_{22}(t,\tau) & \cdots & h_{2n_t}(t,\tau) \\ \vdots & \vdots & \ddots & \vdots \\ h_{n_r1}(t,\tau) & h_{n_r2}(t,\tau) & \cdots & h_{n_rn_t}(t,\tau) \end{bmatrix}$$

and $h_{ij}(t,\tau)$ is the time-varying impulse response between the jth $(j = 1, 2, \ldots, n_t)$ transmitting antenna and the ith receive antenna $(i = 1, 2, \ldots, n_r)$ and is the response at time t to an impulse transmitted at $t-\tau$. The vector $[h_{1j}(t,\tau), h_{2j}(t,\tau), \ldots, h_{n_rj}(t,\tau)]^T$ is referred to as **spatio-temporal signature** induced by the jth transmitting antenna across the receiving antenna array. In most applications, communications are carried out in a passband around a center frequency ω_c. In this case, we may write

$$\mathbf{x}(t) = \mathrm{Re}[\mathbf{x}_{\mathrm{en}}(t)e^{j\omega_c t}], \quad \mathbf{y}(t) = \mathrm{Re}[\mathbf{y}_{\mathrm{en}}(t)e^{j\omega_c t}], \quad \mathbf{z}(t) = \mathrm{Re}[\mathbf{z}_{\mathrm{en}}(t)e^{j\omega_c t}].$$

Let $[\tilde{h}(t,\omega)]$ be the Fourier transform of $[h(t,\tau)]$. If the bandwidth of the input signal vector, denoted by B, is narrow enough, $\tilde{h}(t,\omega)$ can be treated as constant over the band of interest. Introducing these into (8.19), we obtain

$$\begin{aligned} \mathbf{y}_{\mathrm{en}}(t) &= \int_{-\infty}^{\infty} [h(t,\tau)]e^{-j\omega_c\tau}\mathbf{x}_{\mathrm{en}}(t-\tau)d\tau + \mathbf{z}_{\mathrm{en}}(t) \\ &\approx \int_{-\infty}^{\infty} \left[\frac{1}{2\pi}[\tilde{h}(t,\omega_c)] \int_{\omega_c - B/2}^{\omega_c + B/2} e^{j\omega\tau}d\omega\right] e^{-j\omega_c\tau}\mathbf{x}_{\mathrm{en}}(t-\tau)d\tau + \mathbf{z}_{\mathrm{en}}(t) \\ &= \frac{1}{2\pi}B[\tilde{h}(t,\omega_c)] \int_{-\infty}^{\infty} \left[\frac{\sin(B\tau/2)}{B\tau/2}\right] \mathbf{x}_{\mathrm{en}}(t-\tau)d\tau + \mathbf{z}_{\mathrm{en}}(t) \end{aligned}$$

$$\approx \frac{1}{2\pi} B[\tilde{h}(t,\omega_c)]\mathbf{x}_{\text{en}}(t) + \mathbf{z}_{\text{en}}(t) = [h(t,0)]\mathbf{x}_{\text{en}}(t) + \mathbf{z}_{\text{en}}(t)$$
$$= [H]\mathbf{x}_{\text{en}}(t) + \mathbf{z}_{\text{en}}(t), \tag{8.20}$$

where $[H] = [h(t,0)]$ is the **narrowband MIMO channel matrix**. Thus, the convolution in (8.19) can be replaced by a product as in (8.20) for narrowband application. Equation (8.20) gives the relationship between the input and output symbols. Assuming that $\mathbf{X}$ and $\mathbf{Z}$ are statistically independent, we have

$$\begin{aligned}[R_{\mathbf{YY}}] &= \langle[\mathbf{Y}\bar{\mathbf{Y}}^T]\rangle = \langle[H]\mathbf{X}\bar{\mathbf{X}}^T[\bar{H}]^T\rangle + \langle[\mathbf{Z}\bar{\mathbf{Z}}^T]\rangle \\ &= [H][R_{\mathbf{XX}}][\bar{H}]^T + [R_{\mathbf{ZZ}}], \\ [R_{\mathbf{XY}}] &= \langle[\mathbf{X}\bar{\mathbf{Y}}^T]\rangle = \langle\mathbf{X}\bar{\mathbf{X}}^T[\bar{H}]^T\rangle + \langle[\mathbf{X}\bar{\mathbf{Z}}^T]\rangle = [R_{\mathbf{XX}}][\bar{H}]^T.\end{aligned} \tag{8.21}$$

The total transmitted power at the input is then given by $\text{Tr}[R_{\mathbf{XX}}]$.

Note that in order to decode n_t separate transmitted signal, a necessary condition is that the number of receiving antennas must be at least as many as the number of transmitting antennas. When there are sufficient scatterers in the environment, one may expect that the n_r receiving signals are linearly independent combinations of the transmitted signals. In this case, it is possible to deduce the value of $\mathbf{x}$ from $\mathbf{y}$ through (8.20) by inverting the matrix $[H]$ or performing a pseudoinverse if $[H]$ is not invertible. It should be noted that the linear independence depends on the environment and antennas. For example, if two antennas receive the same electromagnetic field, one of them becomes redundant and the linear independence fails. Another extreme situation is when no scatterers exist, i.e., the line-of-sight case, where the n_r receiving antennas receive essentially the same combination of the n_t different transmitted signals up to a global phase shift.

The MIMO channel capacity is defined as the maximum mutual information over all possible transmitted vector signals. Let $\mathbf{X}$ and $\mathbf{Y}$ be two complex random vectors, which represent the input and output of a memoryless wireless channel. The mutual information between $\mathbf{X}$ and $\mathbf{Y}$ is denoted by $I(\mathbf{X},\mathbf{Y})$. The capacity (in bits/s/Hz) of a deterministic MIMO channel is then given by

$$C = \sup_{f_{\mathbf{X}}(\mathbf{x})|\, Tr[R_{\mathbf{XX}}]=P} I(\mathbf{X},\mathbf{Y}), \tag{8.22}$$

where the supremum is taken overall input probability distribution with the total input power limited to P, regardless of the number of transmit antennas. Since $\mathbf{X}$ and $\mathbf{Z}$ are independent, the mutual information can be expressed as

$$I(\mathbf{X}, \mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) = H(\mathbf{Y}) - H(\mathbf{Z}).$$

The value of $H(\mathbf{Z})$ is set by the noise alone. Thus maximizing $I(\mathbf{X}, \mathbf{Y})$ is equivalent to maximizing $H(\mathbf{Y})$. It can be shown that $H(\mathbf{Y})$ is maximized if $\mathbf{Y}$ is Gaussian with zero-mean. From (8.15), the maximum mutual information is given by

$$\begin{aligned}
&\sup_{\mathrm{Tr}[R_{\mathbf{XX}}]=P} I(\mathbf{X}, \mathbf{Y}) \\
&\quad = \sup_{\mathrm{Tr}[R_{\mathbf{XX}}]=P} \log\det\big(e\pi\big[[H][R_{\mathbf{XX}}][\bar{H}]^T + [R_{\mathbf{ZZ}}]\big]\big) \\
&\qquad - \log\det\big(e\pi[R_{\mathbf{ZZ}}]\big) \\
&\quad = \sup_{\mathrm{Tr}[R_{\mathbf{XX}}]=P} \log\det\big([I_{n_r}] + [H][R_{\mathbf{XX}}][\bar{H}]^T[R_{\mathbf{ZZ}}]^{-1}\big)
\end{aligned} \tag{8.23}$$

and the channel capacity is

$$C = \sup_{\mathrm{Tr}[R_{\mathbf{XX}}]=P} \log\det\big([I_{n_r}] + [H][R_{\mathbf{XX}}][\bar{H}]^T[R_{\mathbf{ZZ}}]^{-1}\big). \tag{8.24}$$

Remark 8.3: Equation (8.23) has an intuitive expression which relates the channel capacity to the covariance matrix of the linear minimum mean squared error estimate of the input $\mathbf{X}$ (Stocia *et al.*, 2005). □

It is common to assume that the noises in the receiver branches are uncorrelated so that one can write $[R_{\mathbf{ZZ}}] = \sigma^2[I_{n_r}]$. Thus (8.23) may be written as

$$\begin{aligned}
&\sup_{\mathrm{Tr}[R_{\mathbf{XX}}]=P} I(\mathbf{X}, \mathbf{Y}) \\
&\quad = \sup_{\mathrm{Tr}[R_{\mathbf{XX}}]=P} \log\det\left([I_{n_r}] + \frac{1}{\sigma^2}[H][R_{\mathbf{XX}}][\bar{H}]^T\right).
\end{aligned} \tag{8.25}$$

Making use of the determinant identity,

$$\det([I] + [AB]) = \det([I] + [BA]), \tag{8.26}$$

(8.25) can also be expressed as

$$\sup_{\mathrm{Tr}[R_{\mathbf{XX}}]=P} I(\mathbf{X},\mathbf{Y}) = \sup_{\mathrm{Tr}[R_{\mathbf{XX}}]=P} \log\det\left([I_{n_t}] + \frac{1}{\sigma^2}[R_{\mathbf{XX}}][\bar{H}]^T[H]\right) \tag{8.27}$$

and channel capacity (8.24) may be written as

$$C = \sup_{\mathrm{Tr}[R_{\mathbf{XX}}]=P} \log\det\left([I_{n_r}] + \frac{1}{\sigma^2}[H][R_{\mathbf{XX}}][\bar{H}]^T\right) = \sup_{\mathrm{Tr}[R_{\mathbf{XX}}]=P} \log\det\left([I_{n_t}] + \frac{1}{\sigma^2}[R_{\mathbf{XX}}][\bar{H}]^T[H]\right). \tag{8.28}$$

It now remains to choose $[R_{\mathbf{XX}}]$ to maximize (8.25) or (8.27) subject to the constraint $\mathrm{Tr}[R_{\mathbf{XX}}] = P$.

Case 1: In practice the transmitter has no channel knowledge. In this case, it is reasonable to choose $\mathbf{X}$ to be spatially white (i.e., signals transmitted by each antenna are independent) and use a uniform power distribution (i.e., each antenna is equi-powered). Thus, the covariance matrix of $\mathbf{X}$ is given by $[R_{\mathbf{XX}}] = (P/n_t)[I_{n_t}]$. In this case, the capacity (in bit/s/Hz) for a complex AWGN MIMO channel can be expressed as

$$C = \log\det\left([I_{n_r}] + \frac{P}{\sigma^2 n_t}[H][\bar{H}]^T\right) = \log\det\left([I_{n_t}] + \frac{P}{\sigma^2 n_t}[\bar{H}]^T[H]\right). \tag{8.29}$$

Using the singular value decomposition, $[H]$ can be written as

$$[H] = [U][\Sigma][\bar{V}]^T = \sum_{i=1}^{q} \delta_i \mathbf{u}_i \bar{\mathbf{v}}_i^T, \tag{8.30}$$

where $[U]$ and $[V]$ are $n_r \times n_r$ and $n_t \times n_t$ unitary matrices respectively; $[\Sigma]$ is a $n_r \times n_t$ matrix whose only nonzero entries are $\Sigma(i,i) = \delta_i$, $i = 1, 2, \ldots, q$; $\mathbf{u}_i(\mathbf{v}_i)$ are column vectors of $[U]([V])$ respectively and they are orthonormal so that $[\bar{U}]^T[U] = I_{n_r}$ and $[\bar{V}]^T[V] = I_{n_t}$; q is the number of nonzero singular values and is the rank of the channel matrix $[H]$

$$q = \mathrm{rank}[H] \leq \min(n_t, n_r).$$

It follows from (8.30) that

$$\begin{aligned}[\bar{H}]^T[H] &= [V][\bar{\Sigma}]^T[\bar{U}]^T[U][\Sigma][\bar{V}]^T = \sum_{i=1}^{q} \delta_i^2 \mathbf{v}_i \bar{\mathbf{v}}_i^T, \\ [H][\bar{H}]^T &= [U][\Sigma][\bar{V}]^T[V][\overline{\Sigma}]^T[\bar{U}]^T = \sum_{i=1}^{q} \delta_i^2 \mathbf{u}_i \bar{\mathbf{u}}_i^T.\end{aligned} \tag{8.31}$$

Thus (8.29) can be written as

$$\begin{aligned}C &= \begin{cases} \log\det\left([I_{n_r}] + \dfrac{P}{\sigma^2 n_t}[U][\Sigma][\bar{\Sigma}]^T[\bar{U}]^T\right) \\ \log\det\left([I_{n_t}] + \dfrac{P}{\sigma^2 n_t}[V][\bar{\Sigma}]^T[\Sigma][\bar{V}]^T\right) \end{cases} \\ &= \begin{cases} \log\det\left([I_{n_r}] + \dfrac{P}{\sigma^2 n_t}[\Sigma][\bar{\Sigma}]^T\right) \\ \log\det\left([I_{n_t}] + \dfrac{P}{\sigma^2 n_t}[\bar{\Sigma}]^T[\Sigma]\right) \end{cases} = \sum_{i=1}^{q} \log\left(1 + \frac{P}{\sigma^2 n_t}\delta_i^2\right).\end{aligned} \tag{8.32}$$

Case 2: When the channel is known at the transmitter, the capacity will be higher and the maximum capacity can be achieved by using the water-filling principle, where power is unevenly distributed among the transmitting antennas. Since $[\bar{H}]^T[H]$ is Hermitian it can be diagonalized with

$$[\bar{H}]^T[H] = [\bar{U}]^T[\Lambda][U], \tag{8.33}$$

where $[U]$ is the eigenvector matrix with orthonormal columns and

$$[\Lambda] = \mathrm{diag}(\lambda_1, \lambda_2, \ldots \lambda_{n_t}). \tag{8.34}$$

Thus

$$\begin{aligned}&\det\left([I_{n_r}] + \frac{1}{\sigma^2}[H][R_{\mathbf{XX}}][\bar{H}]^T\right) \\ &\quad = \det\left([I_{n_t}] + \frac{1}{\sigma^2}[R_{\mathbf{XX}}][\bar{H}]^T[H]\right) \\ &\quad = \det\left([I_{n_t}] + \frac{1}{\sigma^2}[R_{\mathbf{XX}}][\bar{U}]^T[\Lambda]^{1/2}[\Lambda]^{1/2}[U]\right) \\ &\quad = \det\left([I_{n_t}] + \frac{1}{\sigma^2}[\Lambda]^{1/2}[U][R_{\mathbf{XX}}][\bar{U}]^T[\Lambda]^{1/2}\right) \\ &\quad = \det\left([I_{n_t}] + \frac{1}{\sigma^2}[\Lambda]^{1/2}[\tilde{R}_{\mathbf{XX}}][\Lambda]^{1/2}\right),\end{aligned} \tag{8.35}$$

where $[\tilde{R}_{\mathbf{XX}}] = [U][R_{\mathbf{XX}}][\bar{U}]^T$. Note that $[\tilde{R}_{\mathbf{XX}}]$ is positive definite if and only if $[R_{\mathbf{XX}}]$ is and that $\mathrm{Tr}[\tilde{R}_{\mathbf{XX}}] = \mathrm{Tr}[R_{\mathbf{XX}}]$. As a result, the maximization with respect to $[R_{\mathbf{XX}}]$ is equivalent to the maximization with respect to $[\tilde{R}_{\mathbf{XX}}]$. Also note that for a positive definite matrix $[A]$ we have $\det[A] \leq \prod_i A(i,i)$ from Hadamard's inequality (Horn and Johnson, 1985), where $A(i,i)$ denote the diagonal matrix elements. Thus

$$\det\left([I_{n_t}] + \frac{1}{\sigma^2}[\Lambda]^{1/2}[\tilde{R}_{\mathbf{XX}}][\Lambda]^{1/2}\right) \leq \prod_{i=1}^{n_t}\left(1 + \frac{1}{\sigma^2}\lambda_i \tilde{R}_{\mathbf{XX}}(i,i)\right). \tag{8.36}$$

The equality holds only when $[\tilde{R}_{\mathbf{XX}}]$ is diagonal. Thus (8.27) may be written as

$$\begin{aligned}\sup_{\mathrm{Tr}[R_{\mathbf{XX}}]=P} I(\mathbf{X},\mathbf{Y}) &= \sup_{\mathrm{Tr}[\tilde{R}_{\mathbf{XX}}]=P} \log \prod_{i=1}^{n_t}\left(1 + \frac{1}{\sigma^2}\lambda_i \tilde{R}_{\mathbf{XX}}(i,i)\right) \\ &= \sup_{\mathrm{Tr}[\tilde{R}_{\mathbf{XX}}]=P} \sum_{i=1}^{n_t} \log\left(1 + \frac{1}{\sigma^2}\lambda_i \tilde{R}_{\mathbf{XX}}(i,i)\right).\end{aligned}$$

Let $P_i = \tilde{R}_{\mathbf{XX}}(i,i)$ and consider the Lagrangian

$$L(P_1, P_2, \ldots, P_{n_t}) = \sum_{i=1}^{n_t} \log\left(1 + \frac{\lambda_i P_i}{\sigma^2}\right) - \lambda\left(\sum_{i=1}^{n_t} P_i - P\right).$$

The Kuhn–Tucker condition for the optimality of a power allocation is

$$\frac{\partial L}{\partial P_i}\begin{cases} = 0, & P_i > 0 \\ \leq 0, & P_i = 0 \end{cases},$$

from which the optimal diagonal entries can be determined to be

$$P_i = \left(\mu - \frac{\sigma^2}{\lambda_i}\right)^+, \quad i = 1, 2, \ldots, n_t, \tag{8.37}$$

where a^+ denotes $\max(0, a)$, $\mu = \log e/\lambda$, and μ is chosen such that

$$\mathrm{Tr}[\tilde{R}_{\mathbf{XX}}] = \sum_{i=1}^{n_t} P_i = \sum_{i=1}^{n_t}\left(\mu - \frac{\sigma^2}{\lambda_i}\right)^+ = P. \tag{8.38}$$

From (8.27), (8.35) and (8.36), the maximum mutual information is given by

$$\sup_{\mathrm{Tr}[R_{\mathbf{XX}}]=P} I(\mathbf{X},\mathbf{Y}) = \log\prod_{i=1}^{n_t}\left[1+\frac{\lambda_i}{\sigma^2}\left(\mu-\frac{\sigma^2}{\lambda_i}\right)^+\right] = \sum_{i=1}^{n_t}\left[\log\left(\frac{\lambda_i}{\sigma^2}\mu\right)\right]^+. \tag{8.39}$$

In this case, the ergodic capacity (in bit/s/Hz) for a complex AWGN–MIMO channel can be expressed as

$$C = \sum_{i=1}^{n_t}\left[\log\left(\frac{\lambda_i}{\sigma^2}\mu\right)\right]^+. \tag{8.40}$$

The optimal power allocation strategy (8.37) is called **water-filling**.

Let us consider some special situations.

(1) SISO channel capacity ($n_t = n_r = 1$)
In this case, $[H] = [h]$ and (8.28) becomes

$$C = \log\left(1+\frac{P}{\sigma^2}|h|^2\right) = \log\left(1+\rho|g|^2\right), \tag{8.41}$$

where $g = \sqrt{P/P_{\mathrm{rec}}}h$, $\rho = P_{\mathrm{rec}}/\sigma^2$, and P_{rec} is the received power at the output of the receiving antenna.

(2) SIMO channel capacity ($n_t = 1$)
In this case, $[H] = [h_1, h_2, \ldots, h_{n_r}]^T$ and (8.28) becomes

$$C = \log\left(1+\frac{P}{\sigma^2}\sum_{i=1}^{n_r}|h_i|^2\right). \tag{8.42}$$

(3) MIMO channel capacity ($n_t = n_r = q$, $\delta_1 = \delta_2 = \cdots = \delta$)
From (8.32), we obtain

$$C = n_t \log\left(1+\frac{P}{\sigma^2 n_t}\delta^2\right). \tag{8.43}$$

Comparing to $n_t = n_r = 1$ (SISO), the capacity of a multiple antenna system is much higher.

8.3 Digital Communication Systems

The primary advantage of the digital communication system is the ease with which digital signals, compared to analog signals, are regenerated. A typical

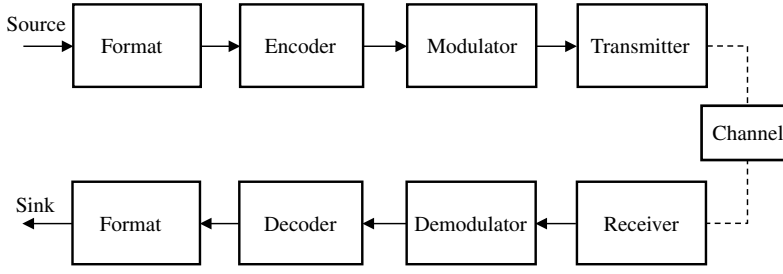


Figure 8.6 Digital communication system.

block diagram of the digital communication system is shown in Figure 8.6. The upper blocks indicate the signal transformation from the source to the transmitter. The lower blocks indicate the signal transformation from the receiver to the sink, which reverses the signal processing steps performed by upper blocks. The functions of each block are summarized as follows:

Format: The format processes performed in the transmitter typically consists of a sampler and a quantizer, and a coder, which transforms the source information (analog information) into digital information (digital symbols) and makes the information compatible with the signal processing within digital communication system. The format processes performed in the receiver usually consists of a decoder and a low-pass filter, which transforms digital information into analog information.

Encoder: The encoder mainly involves a source encoder and a channel encoder. The source-encoding process removes redundant or unneeded information and produces code-words in binary form. The goal of source coding is either to improve the signal-to-noise ratio (SNR) for a given bit rate or to reduce the bit rate for a given SNR. For the source encoder to be efficient, the knowledge of statics of the source is required. The average code word length must be greater than the entropy of the source (Shannon's first theorem). The channel encoder introduces redundancy in a prescribed manner to increase the resistance of a digital communication system to channel noise (error control encoding). Channel coding can reduce the probability of error (P_E) for a given data rate, or reduce SNR requirement, at the expense of bandwidth or decoder complexity. The Shannon's second theorem (or channel coding theorem) states that the channel capacity C is the fundamental limit on the rate at which the transmission of error-free message can take place over a memoryless channel. For communications security, encryption must be introduced in the encoder to prevent the

unauthorized users from understanding messages and from injecting false messages into the system.

Modulator and demodulator: After possible source and error control encoding, a sequence of message symbols to be transmitted on the channel is obtained. Each symbol belongs to a finite set of alphabet $\{1, 2, \ldots, M\}$. The modulator performs the function to convert the digital symbols $\{1, 2, \ldots, M\}$ to digitally encoded waveforms $\{s_1(t), s_2(t), \ldots, s_M(t)\}$ that are compatible with the transmission channel. The demodulator is a device that performs the inverse operation of modulation.

Transmitter and receiver: The transmitter usually consists of a frequency up-conversion stage, a high-power amplifier, and an antenna. The receiver portion usually consists of an antenna, a low-noise amplifier (LNA), and a down-converter stage, typically to an intermediate frequency (IF).

8.3.1 Digital Modulation Techniques

A sinusoidal wave has three properties: amplitude, frequency, and phase. Digital modulation refers the process of varying one or more properties of a sine waveform, called the **carrier**, with a digital bit stream (baseband signal) which contains information to be transmitted.

8.3.1.1 *Baseband Transmission*

An original analog waveform can be sampled to obtain a natural-sampled data or pulse amplitude modulation (PAM), which is the output of the sampling process and is then transformed into a quantized PAM signal (called **quantization**), as illustrated in Figure 8.7. Each quantized sample

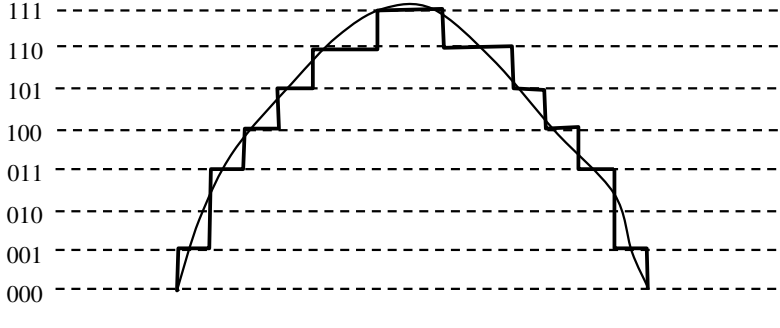


Figure 8.7 An example of 3-bit quantization.

is then encoded into a **digital word** (**code word**) to obtain a **PCM** (pulse code modulation) **sequence**. The PCM word size l depends on allowable quantization distortion, which satisfies

$$l \geq \log_2 \frac{1}{2p} \text{bits},$$

where p is a number chosen such that the quantization distortion error does not exceed a fraction p of the peak-to-peak analog voltage. The sampling process can be implemented in several ways and the most popular one is the sample-and-hold operation. According to the uniform sampling theorem, the sampling rate f_s (the number of samples per second) should satisfy

$$f_s \geq 2f_m,$$

where f_m is the absolute bandwidth of a bandlimited signal (i.e., the signal has no spectral components above f_m). The sampling rate $f_s = 2f_m$ is called **Nyquist rate**.

Remark 8.4 (Sampling Theorem)**:** If $g(t)$ has finite bandwidth f_m, i.e., its Fourier transform is zero for $|f| > f_m$, then

$$g(t) = \sum_{i=-\infty}^{\infty} g(iT)\text{sinc}\left(\frac{t-iT}{T}\right),$$

where $T = 1/2f_m$, $\text{sinc}(t)$ is the normalized sinc function defined by

$$\text{sinc}(t) = \frac{\sin \pi t}{\pi t}.$$

It is easy to show that

$$\int_{-\infty}^{\infty} \text{sinc}\left(\frac{t-iT}{T}\right)\text{sinc}\left(\frac{t-jT}{T}\right)dt = \begin{cases} T, & i = j \\ 0, & i \neq j \end{cases}.$$

The sampling theorem indicates that a signal of bandwidth f_m is totally determined by its sampling values at the times iT ($T = 1/2f_m$ is called **Nyquist interval**). If $g(t)$ is small outside the time interval $(0, T_0)$, then only sampling points within $(0, T_0)$ need to be considered. The number of sampling points inside $(0, T_0)$ is given by $2f_mT_0$. Thus, we have the following expansion

$$g(t) = \sum_{i=1}^{2f_mT_0} g(iT)\,\text{sinc}\left(\frac{t-iT}{T}\right).$$

The $2f_mT_0$ samples need not be the equally spaced. Any set of independent $2f_mT_0$ numbers associated with the function can be used to characterize it. □

These PCM binary digits are then represented by electrical pulses in order to transmit them through a baseband channel. To obtain particular spectral characteristics of a pulse train, digital baseband signals (PCM sequence) often use line codes. The most common codes for mobile communication are return-to-zero (**RZ**), non-return-to-zero (**NRZ**), and **Manchester codes**, as illustrated in Figure 8.8. All of these may either be unipolar or bipolar. RZ implies that the pulse returns to zero within every bit period. This leads to spectral widening, but improves time synchronization. NRZ codes do not return to zero during a bit period, i.e., the signal stays at constant levels throughout a bit period. NRZ codes are more spectrally efficient than RZ codes, but offer poorer synchronization capabilities. Since NRZ has a large DC component, it is used for data that

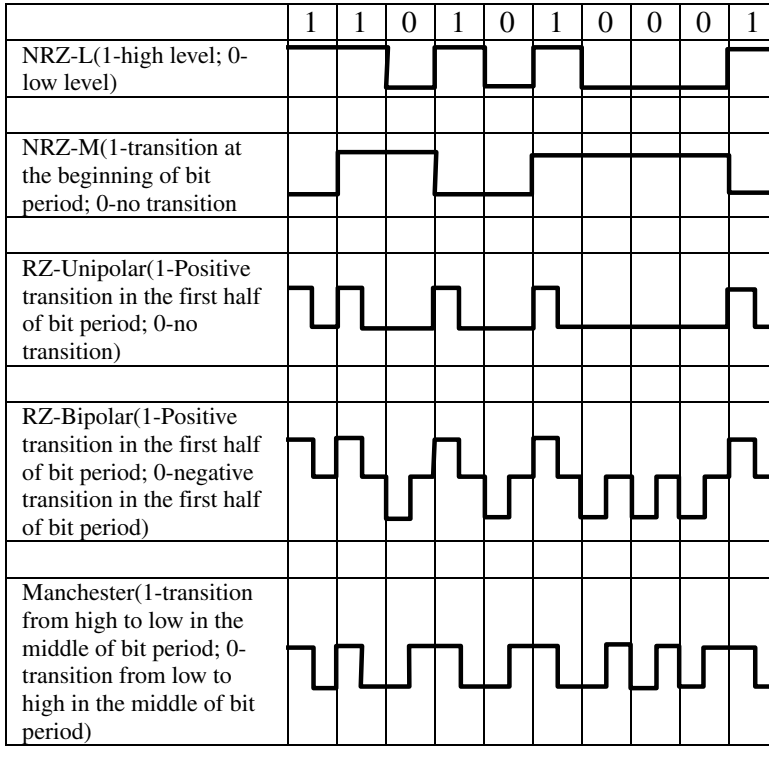


Figure 8.8 Digital signal encoding formats.

does not have to be passed through DC blocking circuits such as audio amplifiers or phone lines. The Manchester code is special type of NRZ line code that is ideally suited for signaling that must pass through phone lines and other DC blocking circuits as it has no DC component and offers simple synchronization. Manchester code use two pulses to represent each binary symbol, and thereby provide easy clock recovery since zero crossings are guaranteed in very bit period.

The system bandwidth required for binary PCM signaling may be very large. In recent years, more and more telecommunications services are needed out of a limited amount of spectrum. To relax the bandwidth requirement one can use multilevel digital signals to modulate the carrier to offer greater bandwidth efficiency. This is called **M-ary signaling**. The original binary data stream can be subdivided into groups of k bits and each group, which is called a **symbol**, is converted to one of $M = 2^k$ possible levels by means of a D/A converter. Thus, the multilevel signaling can be used to reduce the number of symbols transmitted per second (equal to R_b/k), or thus to reduce the bandwidth requirement of the channel. The resulting waveform has fewer transitions per unit time but requires higher amplitude resolution in the detector and a greater amount of energy for equivalent detection performance as compared to the two-level signaling, as illustrated in Figure 8.9. Since one of M symbols is transmitted during each symbol duration, denoted T_s, the data rate R_b can be expressed as

$$R_b = k/T_s = kR_s.$$

The **bandwidth efficiency** of a digital system that transmits $k = \log_2 M$ bits in T_s seconds using a bandwidth B is defined by

$$\eta_B = \frac{R_b}{B} = \frac{\log_2 M}{BT_s} = \frac{1}{BT_b},$$

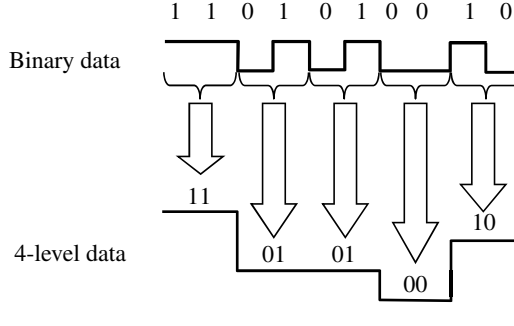


Figure 8.9 Four-level digital representation of a binary data stream.

which characterizes the capability for a modulation scheme to accommodate data within a specified bandwidth. From the above equation the smaller the BT_s product, the more bandwidth efficient will be a digital communication system. Thus signals with small BT_s are often used with bandwidth-limited systems. For example, GSM uses Gaussian minimum-shift keying (GMSK) modulation having a BT_s product equal to 0.3 Hz/(bit/s), where B is the bandwidth of a Gaussian filter.

One must not confuse the idea of the number of bits per PCM word, denoted l, with the M-level transmission concept of k data bits per symbol. Here is an example about quantization levels and multilevel signaling to clarify the distinction. The information in an analog waveform, with maximum frequency $f_m = 3$ kHz, is to be transmitted over an M-level PCM system, where the number of pulse levels is $M = 16$. The quantization distortion is specified not to exceed $\pm 1\%$ of the peak-to-peak analog signal. Then

(1) The PCM word size $l \geq \log_2 \frac{1}{0.02} = 5.6$, so $l = 6$.
(2) Using Nyquist sampling criterion, the minimum sampling rate $f_s = 2f_m = 6000$ (samples/s). Since each sample will give rise to a PCM word of 6 bits, the bit transmission rate $R_b = lf_s = 36{,}000$ bits/s.
(3) Since the multilevel pulses are to be used with $M = 2^k = 16$ levels, $k = \log_2 16 = 4$ bits/symbol. Therefore, the bit stream will be partitioned into groups of 4 bits to form new 16-level PCM digits, and the resulting symbol rate $R_s = R_b/k = 9000$ symbols/s.

When rectangular pulses pass through a bandlimited channel (most mobile communication systems operate with minimal bandwidth), the pulses will spread in time, and the pulses for each symbol will smear into the time intervals of succeeding symbols. This causes **intersymbol interference** (ISI) and leads to an increased probability of the receiver making an error in detecting a symbol. To reduce the intersymbol effects and the spectral width of a modulated signal, pulse shaping techniques are often used, which is done through baseband or IF processing as it is much easier to manipulate the transmitter spectrum at lower frequency. Nyquist showed that the theoretical minimum system bandwidth needed to detect R_s symbols/s, without ISI, is $R_s/2$. A **Nyquist filter** results in the minimum required transmission bandwidth that yields zero ISI. Nyquist observed that the effect of ISI could be completely nullified if the overall response of the communication system (including transmitter, channel, and receiver) is designed so that at every sampling instant at the receiver, the response due to all symbols except the current one is equal to zero. If $h_{\mathit{eff}}(t)$ is

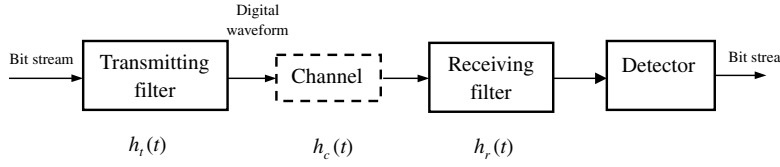


Figure 8.10 Effective transfer function.

the impulse response of the overall communication system, this condition, called **Nyquist ISI criterion**, can be mathematically stated as

$$h_{eff}(nT_s) = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases},$$

where T_s is the symbol period, n is the integer. The effective transfer function of the system can be represented as

$$h_{eff}(t) = h_t(t) * h_c(t) * h_r(t),$$

where $h_t(t)$ is the transmitter impulse response, $h_c(t)$ is the channel impulse response, and $h_r(t)$ is the receiver impulse response, as shown in Figure 8.10. The receiving filter, called **equalizing filter**, should be configured to compensate for the distortion caused by the transmitter and the channel.

To reduce ISI, the most popular pulse-shaping filter satisfying the Nyquist criterion is the **raised cosine filter**,

$$H(f) = \begin{cases} T_s, & 0 \leq |f| \leq \dfrac{1-\alpha}{2T_s} \\ \dfrac{T_s}{2}\left[1 + \cos\dfrac{\pi T_s}{\alpha}\left(|f| - \dfrac{1-\alpha}{2T_s}\right)\right], & \dfrac{1-\alpha}{2T_s} \leq |f| \leq \dfrac{1+\alpha}{2T_s} \\ 0, & |f| > \dfrac{1+\alpha}{2T_s} \end{cases},$$

where T_s is the symbol period. The impulse response of such a filter is given by

$$h(t) = \text{sinc}\left(\frac{t}{T_s}\right)\frac{\cos\left(\frac{\pi\alpha}{T_s}t\right)}{1 - \frac{4\alpha^2}{T_s^2}t^2},$$

where α is called roll-off factor. Let B be the absolute filter bandwidth $B = (1+\alpha)/2T_s$. The symbol rate that can be passed through a baseband raised cosine filter is

$$R_s = 1/T_s = 2B/(1+\alpha),$$

or

$$B_{\text{PCM}} = (1 + \alpha)R_s/2,$$

where B_{PCM} stands for the minimum system PCM bandwidth requirement for a symbol rate R_s. Bandpass-modulated signals (baseband signals that have been shifted in frequency), such as ASK and PSK require twice the transmission bandwidth of the equivalent baseband signals (DSB signals). Therefore, for the RF systems, the RF passband bandwidth doubles and

$$R_s = 1/T_s = B/(1 + \alpha),$$

or

$$B_{\text{DSB}} = (1 + \alpha)R_s,$$

where B_{DSB} represents the minimum required DSB bandwidth for transmitting the modulated PCM sequence.

As an example, we consider the bandwidth requirement for the baseband transmission of a four level PCM pulse sequence having a data rate of $R_b = 2400$ bits/s if the system transfer characteristic consists of a raised cosine spectrum with 100% excess bandwidth ($\alpha = 1$). Since $M = 4$, then $k = 2$. The symbol rate is $R_s = R_b/k = 1200$ symbols/s. The minimum bandwidth is $B_{\text{PCM}} = (1 + \alpha)R_s/2 = 1200\,\text{Hz}$. If the same PCM sequence is modulated onto a carrier wave, so that the baseband spectrum is shifted and centered at frequency f_c. The DSB bandwidth will be $B_{\text{DSB}} = (1 + \alpha)R_s = 2400\,\text{Hz}$.

In practical systems, the frequency response of the channel is not known with sufficient precision to allow for a receiver design that will compensate for the ISI for all time. In practice, the filter for handling ISI at the receiver contains various parameters that are adjusted on the basis of measurements of the channel characteristics. The process of correcting the channel-induced distortion is called **equalization**.

8.3.1.2 *Modulation and Demodulation*

The analog modulation methods include amplitude modulation (AM), frequency modulation (FM), and phase modulation (PM). The digital counterparts are amplitude shift keying (**ASK**), frequency shift keying (**FSK**), and phase shift keying (**PSK**), respectively. FSK and PSK are less sensitive to the amplitude noise and are widely used in RF systems.

The digital modulator performs the function to map a digital symbol into a digital waveform:

$$i \in \{1, 2, \ldots, M\} \rightarrow s_i(t) \in \{s_1(t), s_2(t), \ldots, s_M(t)\}. \tag{8.44}$$

If the symbol to waveform mapping (8.44) is fixed from one interval to the next, the modulation is called memoryless. If the mapping in the ith symbol interval depends on previously transmitted symbols, the modulation is said to have memory.

We may define an N-dimensional orthogonal space spanned by $N(N < M)$ linearly independent functions, $\{\psi_j(t)\}$, called **basis functions**. Then the digital waveforms after modulation (i.e., the transmitted signal waveforms) can be expanded as follows

$$\begin{cases} s_1(t) = a_{11}\psi_1(t) + a_{12}\psi_2(t) + \cdots + a_{1N}\psi_N(t) \\ s_2(t) = a_{21}\psi_1(t) + a_{22}\psi_2(t) + \cdots + a_{2N}\psi_N(t) \\ \cdots\cdots \\ s_M(t) = a_{M1}\psi_1(t) + a_{M2}\psi_2(t) + \cdots + a_{MN}\psi_N(t) \end{cases}.$$

Once a set of N orthogonal functions has been chosen, each of the digital waveforms $s_i(t)$ is completely determined by the vector of its coefficients

$$\mathbf{s}_i = (a_{i1}, a_{i2}, \ldots, a_{iN})^T, \quad i = 1, 2, \ldots, M,$$

which is called **signal vector**. Table 8.1 shows the general analytic expressions of the digital waveforms of various digital modulators for general M-ary signaling and the corresponding basis functions. In the table, T_s is the symbol time duration and E_s is the symbol energy. Note that M-ary QAM does not have constant energy per symbol, nor does it have constant distance between possible symbol states. For M-ary QAM, $E_{s\,\min}$ is the symbol energy of the signal with the lowest amplitude, and

$$\{a_i, b_i\} = \begin{bmatrix} (-\sqrt{M}+1,\ \sqrt{M}-1) & (-\sqrt{M}+3,\ \sqrt{M}-1) & \cdots \\ (-\sqrt{M}+1,\ \sqrt{M}-3) & (-\sqrt{M}+3,\ \sqrt{M}-3) & \cdots \\ \cdots & \cdots & \cdots \\ (-\sqrt{M}+1,\ -\sqrt{M}+1) & (-\sqrt{M}+3,\ -\sqrt{M}L+1) & \cdots \end{bmatrix}$$

$$\begin{bmatrix} (\sqrt{M}-1,\ \sqrt{M}-1) \\ (\sqrt{M}-1,\ \sqrt{M}-3) \\ \cdots \\ (\sqrt{M}-1,\ -\sqrt{M}+1) \end{bmatrix}$$

Table 8.1 Digital waveforms

Digital modulation schemes and basis functions	**Digital waveforms**	**Power spectral density (PSD)/Bandwidth B**
Phase shift keying (PSK)		
$\psi_1(t) = \sqrt{\frac{2}{T_s}} \cos \omega_c t$ $\psi_2(t) = \sqrt{\frac{2}{T_s}} \sin \omega_c t$	$s_i(t) = \sqrt{\frac{2E_s}{T_s}} \cos[\omega_c t + \varphi_i(t)]$ $= \sqrt{E_s} \cos \varphi_i \psi_1(t) + \sqrt{E_s} \sin \varphi_i \psi_2(t)$ $\varphi_i = \frac{2\pi i}{M}, \quad 0 \le t \le T_s, \quad i = 1, \ldots, M$	$PSD = \frac{E_s}{2} \left[\frac{\sin \pi(f - f_c)T_s}{\pi(f - f_c)T_s} \right]^2 + \frac{E_s}{2} \left[\frac{\sin \pi(f - f_c)T_s}{\pi(f - f_c)T_s} \right]^2,$ $B = 2R_b / \log_2 M$
Frequency shift keying (FSK)		
$\psi_j(t) = \sqrt{\frac{2}{T_s}} \cos \omega_j t$ $j = 1, 2, \ldots, N$	$s_i(t) = \sqrt{\frac{2E_s}{T_s}} \cos(\omega_i t) = \sqrt{E_s} \psi_i(t)$ $0 \le t \le T_s, \quad i = 1, \ldots, M$	$B = \frac{R_b(M+3)}{2 \log_2 M}$ (coherent) $B = \frac{R_b M}{2 \log_2 M}$ (non-coherent)
Quadrature amplitude modulation (QAM)		
$\psi_1(t) = \sqrt{\frac{2}{T_s}} \cos \omega_c t$ $\psi_2(t) = \sqrt{\frac{2}{T_s}} \sin \omega_c t$	$s_i(t) = \sqrt{\frac{2E_{s\min}}{T_s}} a_i \cos \omega_c t + \sqrt{\frac{2E_{s\min}}{T_s}} b_i \sin \omega_c t$ $= \sqrt{E_{s\min}} a_i \psi_1(t) + \sqrt{E_{s\min}} b_i \psi_2(t)$ $0 \le t \le T_s, \quad i = 1, \ldots, M$	$B = \frac{2R_b}{\log_2 M}$
Amplitude shift keying (ASK)	$s_i(t) = \sqrt{\frac{2E_{si}}{T_s}} \cos(\omega_c t + \varphi)$ $0 \le t \le T_s, \ i = 1, \ldots, M$	
Amplitude phase keying (APK)	$s_i(t) = \sqrt{\frac{2E_{si}}{T_s}} \cos[\omega_c t + \varphi_i(t)]$ $0 \le t \le T_s, \ i = 1, \ldots, M$	

Unlike the MPSK or MQAM, MFSK signals are bandwidth inefficient. However, the power efficiency has been improved as M increases because the M signals are orthogonal and they are less crowding in the signal space as compared to MPSK. Actually for MFSK, the distance between any two reference signal vectors is

$$\|\mathbf{s}_i - \mathbf{s}_j\| = \sqrt{2E_s}.$$

The geometrical representation of signal vectors in signal space is called **constellation diagram,** which provides a graphical representation of the complex envelope of each possible symbol state. Some of the properties of a modulation scheme can be inferred from its constellation diagram. For example, if a modulation scheme has a constellation diagram that is densely (sparsely) packed, it is more bandwidth (power) efficient than a modulation scheme with sparsely (densely) packed constellation. However, it should be noted that the bandwidth occupied by a modulated signal increases with the dimension N of the constellation. For example, consider the set of BPSK signals

$$s_1(t) = \sqrt{\frac{2E_b}{T_b}}\cos(2\pi f_c t), \quad s_2(t) = -\sqrt{\frac{2E_b}{T_b}}\cos(2\pi f_c t), \quad 0 \le t \le T_b.$$

The basis function for this signal set consists of a single waveform

$$\psi_1(t) = \sqrt{\frac{2}{T_b}}\cos(2\pi f_c t), \quad 0 \le t \le T_b,$$

and the BPSK signal can be represented as

$$s_1(t) = \sqrt{E_b}\psi_1(t), \quad s_2(t) = -\sqrt{E_b}\psi_1(t), \quad 0 \le t \le T_b.$$

The constellation diagram is shown in Figure 8.11(a). Similarly, a QPSK signal can be expressed as

$$s_i(t) = \sqrt{E_s}\cos\left[(i-1)\frac{\pi}{2}\right]\psi_1(t) - \sqrt{E_s}\sin\left[(i-1)\frac{\pi}{2}\right]\psi_2(t), \quad i = 1, 2, 3, 4$$

and the constellation is shown in Figure 8.11(b). The x-axis of the constellation diagram represents the in-phase component of the complex envelope, and y-axis represents the quadrature component of the complex component.

The noise can be partitioned into two components $n(t) = \hat{n}(t) + \tilde{n}(t)$, where $\hat{n}(t) = \sum_{j=1}^{N} n_j\psi_j(t)$ is the noise component in the signal space spanned by $\{\psi_j(t)\}$, and $\tilde{n}(t)$ is the noise component outside the signal

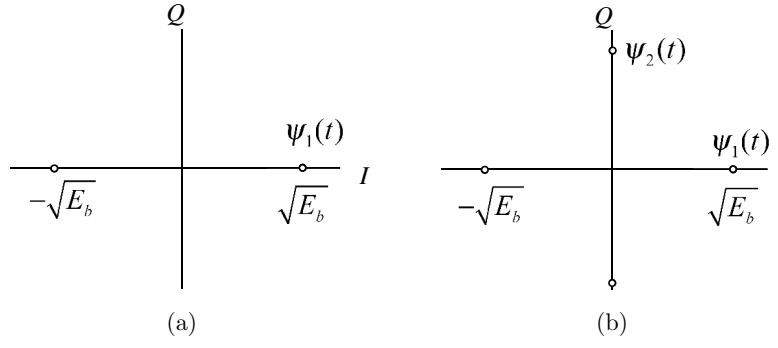


Figure 8.11 Constellation diagram.

space, which satisfy

$$\int_0^{T_s} \tilde{n}(t)\psi_j(t)dt = 0, \quad j = 1, 2, \ldots, N.$$

Physically, $\hat{n}(t)$ is the noise component that will interfere with the detection process, and $\tilde{n}(t)$ is the noise component that will be tuned out by the detector. The interfering portion of the noise $\hat{n}(t)$ will henceforth be referred to simply as $n(t)$. In a similar manner, we may introduce the noise vector $\mathbf{n} = (n_1, n_2, \ldots, n_N)^T$. In digital communications, the terms demodulation and detection are used somewhat interchangeably, although the demodulation emphasizes removal of the carrier, and detection includes the process of symbol decision. A typical detection problem can be conveniently viewed in terms of signal vectors, as is geometrically shown in Figure 8.12. Vectors $\mathbf{s}_j$ and $\mathbf{s}_k$ are the prototype or reference signals belong to the set of M waveforms $\{s_i(t)\}$. The receiver knows the location in the signal space of each reference signals belonging to the M-ary set as *a priori*. During the transmission the signal is perturbed by noise so that the resultant vector received is a perturbed version of the original one (e.g., $\mathbf{s}_j + \mathbf{n}$). Since the noise is additive and has a Gaussian distribution, the resulting distribution of possible received signals is a cluster or cloud of points around $\mathbf{s}_j$. The cluster is dense in the center and becomes sparse with increasing distance from the reference signal. The task of the detector is to decide which of the reference signals within the signal space is closest in distance to the received signal during some symbol interval, denoted $\mathbf{r} = \mathbf{s} + \mathbf{n}$ in Figure 8.12.

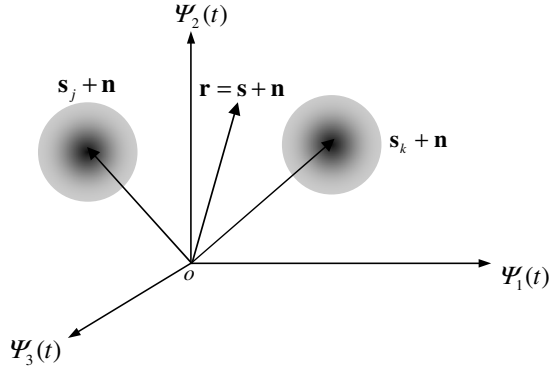


Figure 8.12 Signal space.

The selection of modulation method depends on the several factors such as signal quality (e.g., SNR), bandwidth efficiency (bits/s/Hz), power efficiency (Joule/bit), implementation cost etc. When transmitted power and channel attenuation (path loss) are given, the signal quality of the output of the detector depends on the type of the modem (**mod**ulator-**dem**odulator). If the modems can achieve higher tolerance of noise, either the transmitted power can be reduced or a higher path loss can be accommodated. Bandwidth efficiency describes the ability of a modulation scheme to accommodate data within a limited bandwidth. In general, increasing the data rate implies decreasing the pulse width of a digital symbol, which increases the bandwidth of the signal. Therefore, there is a tradeoff between the data rate and bandwidth occupancy. Some modulation scheme performs better than the others in making this tradeoff. The typical bandwidths and other information for various digital modulation methods are listed in Table 8.2.

Remark 8.5: The system capacity of a digital mobile communication system is directly related to bandwidth efficiency of the modulation scheme, since modulation with a greater value of bandwidth efficiency will transmit more data in a given spectrum allocation. According to Shannon's channel capacity theorem there is an upper bound on achievable bandwidth efficiency, i.e.,

$$\eta_B = \frac{C}{B} = \log_2\left(1 + \frac{P}{N_0 B}\right).$$

So the maximum bandwidth efficiency is limited by the noise in the channel. □

Table 8.2 Digital modulation methods

Modulation M-ary signaling	Number of states (M logic levels)	Theoretical bandwidth efficiency (bits/s/Hz)	Bits sent each symbol	Required bandwidth (symbol rate) Hz
2-PSK	2	1	1	bit rate
4-PSK	4	2	2	1/2 bit rate
8-PSK	8	3	3	1/3 bit rate
16-PSK	16	4	4	1/4 bit rate
64-QAM	64	6	6	1/6 bit rate
256-QAM	256	8	8	1/8 bit rate

Remark 8.6: Power efficiency describes the ability of a modulation scheme to preserve the fidelity of the digital message (i.e., an acceptable bit error probability) at low power levels. In a digital communication system, in order to increase noise immunity it is necessary to increase the signal power. However, the amount by which the signal power should be increased to obtain a certain level of fidelity depends on the particular type of modulation employed. The **power efficiency** η_P of a digital modulation scheme is a measure of how favorably this tradeoff between fidelity and signal power is made, and is often expressed as the ratio of the signal energy per bit, denoted by E_b, to noise PSD, denoted by N_0, required at the receiver input for a certain probability of error (say 10^{-5})

$$\eta_P = E_b/N_0. \qquad \square$$

Remark 8.7: The signal quality of a digital modem can be expressed in terms of the **bit error rate** (**BER**), or **bit error probability**, defined as the average number of erroneous bits observed at the output of the detector divided by the total number of bits received in a unit time. $\square$

The manner in which the baseband signal is extracted from the modulated waveform has great impact on the overall system performance, in particular the signal quality in the output of the detector. The detectors can be categorized as either coherent or non-coherent. A coherent detector requires the knowledge of the phase of the carrier wave to demodulate the signal while a non-coherent detector does not. Given a received signal

$$r(t) = s_i(t) + n(t), \quad 0 \le t \le T_s, \quad i = 1, 2, \ldots, M,$$

where $s_i(t)$ are known signal and $n(t)$ is AWGN. The received signal is used as an input to a linear, time-invariant filter with transfer function

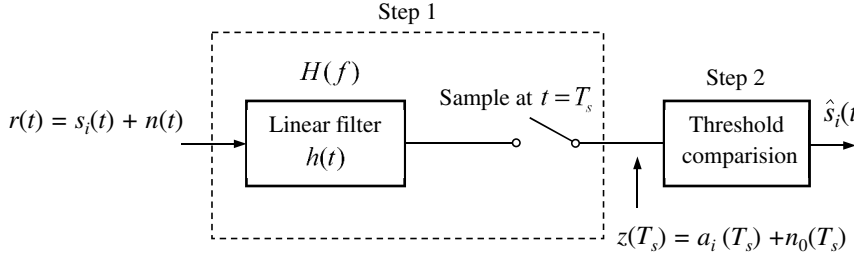


Figure 8.13 Detection.

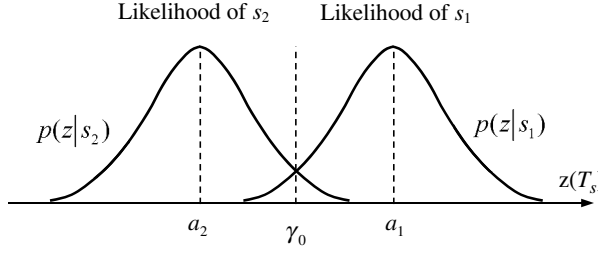


Figure 8.14 Conditional probability density functions.

$H(f)$ followed by a sampler, as shown in Figure 8.13. The output of the sampler is

$$z(T_s) = a_i(T_s) + n_0(T_s), \quad i = 1, 2, \ldots, M, \tag{8.45}$$

where $a_i(T_s)$ is the desired signal component and $n_0(T_s)$ is the noise component. The probability density function of the Gaussian random noise n_0 may be expressed as

$$p(n_0) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{n_0}{\sigma_0}\right)^2\right] \tag{8.46}$$

where σ_0^2 is the noise variance. For $M = 2$ (binary signaling), the conditional probability density functions $f(z|s_i)(i = 1, 2)$ can be obtained from (8.45) and (8.46)

$$p(z|s_i) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{z - a_i}{\sigma_0}\right)^2\right], \quad i = 1, 2. \tag{8.47}$$

These conditional probability density functions are depicted in Figure 8.14 The conditional probability density function $f(z|s_i)$ represents the

probability density function of the random variable $z(T_s)$ given that the symbol s_i was transmitted, and is called the likelihood of s_i.

The detection shown in Figure 8.13 can be carried out by choosing the hypothesis from the threshold measurement

$$z(T_s) \underset{H_2}{\overset{H_1}{\gtrless}} \gamma. \tag{8.48}$$

The inequality implies that the hypothesis H_1 (the signal s_1 was sent) is chosen if $z(T_s) > \gamma$ and the hypothesis H_2 (the signal s_2 was sent) is chosen if $z(T_s) < \gamma$. An optimum threshold $\gamma = \gamma_0$ can be obtained by minimizing the probability of error. One may start with the likelihood ratio test

$$\frac{p(z|s_1)}{p(z|s_2)} \underset{H_2}{\overset{H_1}{\gtrless}} \frac{P(s_2)}{P(s_2)}, \tag{8.49}$$

where $P(s_i)(i = 1, 2)$ denotes the *a priori* probability of the signal s_i being present. This decision criterion corresponds to (8.48) and is called the maximum *a posteriori* (MAP) criterion or minimum error criterion. Assume that the signals s_1 and s_2 are equally likely. The substitution of (8.47) into (8.49) gives

$$z(T_s) \underset{H_2}{\overset{H_1}{\gtrless}} \gamma_0 = \frac{1}{2}(a_1 + a_2). \tag{8.50}$$

The two-sided PSD of the input noise is $N_0/2$. The variance of the output noise (average noise power) is denoted by σ_0^2, so that the ratio of the instantaneous signal power to average noise power out of the receiver at time $t = T_s$ is

$$\left.\frac{S}{N}\right|_{t=T_s} = \left.\frac{a_i^2}{\sigma_0^2}\right|_{t=T_s}, \tag{8.51}$$

where a_i is the signal component after sampler. We wish to find the filter transfer function that maximizes the above equation. Substituting

$$a_i(t) = \int_{-\infty}^{\infty} H(f)\tilde{s}_i(f)e^{-j2\pi ft}df$$

and

$$\sigma_0^2 = \frac{N_0}{2}\int_{-\infty}^{\infty} |H(f)|^2 df$$

into (8.51), we obtain

$$\left.\frac{S}{N}\right|_{t=T_s} = \frac{\left|\int_{-\infty}^{\infty} H(f)\tilde{s}_i(f)e^{j2\pi fT_s}df\right|^2}{\frac{N_0}{2}\int_{-\infty}^{\infty}|H(f)|^2df} \le \frac{2}{N_0}\int_{-\infty}^{\infty}|\tilde{s}_i(f)|df = \frac{2E_s}{N_0},$$

where we have used Schwartz inequality. Thus the maximum output S/N at $t = T_s$ depends on the input signal energy and the PSD, not on the particular shape of the waveform. In digital communication system, E_s usually stands for the symbol energy. The equality in the above equation holds if the filter transfer function satisfies

$$H(f) = c\tilde{\bar{s}}_i(f)e^{-j2\pi fT_s}$$

or the impulse response of the filter takes the form

$$h(t) = \begin{cases} cs(T_s - t) & 0 \le t \le T_s \\ 0 & \text{elsewhere} \end{cases}.$$

The output of the filter is

$$z(T_s) = \int_0^{T_s} r(\tau)h(T_s - \tau)d\tau = \int_0^{T_s} r(\tau)s(\tau)d\tau,$$

which is called the **correlation** of $r(t)$ with $s(t)$.

According to the above discussion, the detection process consists of two basic steps. In the first step, the received waveform $r(t)$ is reduced to a single random variable $z(T_s)$, or a set of random variables, $z_i(T_s)$, $i = 1, 2, \ldots, M$ at time $t = T_s$, where T_s is the symbol duration. In the second step, a symbol decision is made by comparing $z_i(T_s)$ to a threshold or choosing the maximum $z_i(T_s)$. A detector can be optimized in the sense of minimizing the error probability by using matched filters or correlators in Step 1 and optimizing the decision criterion in Step 2. A correlation receiver is shown in Figure 8.15.

8.3.2 Probability of Error

An error occurs if hypothesis H_i is chosen when the signal s_j $(j \neq i)$ was actually transmitted. The probability of error is the sum of all the probabilities that an error can occur. For the binary decision-making shown

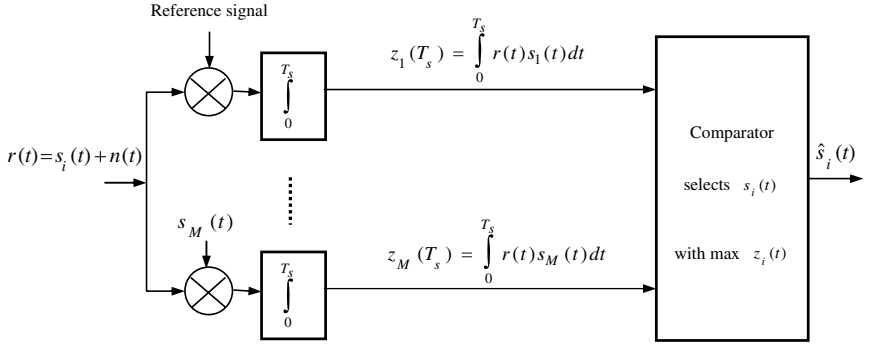


Figure 8.15 Correlation receiver.

in Figure 8.14, the probability of bit error can be expressed as

$$P_b = P(H_2 \cap s_1) + P(H_1 \cap s_2) = P(H_2|s_1)P(s_1) + P(H_1|s_2)P(s_2).$$

If the *a priori* probabilities are equal, i.e., $P(s_1) = P(s_2) = 1/2$, we have

$$P_b = \frac{1}{2}[P(H_2|s_1) + P(H_1|s_2)] = P(H_2|s_1) = P(H_1|s_2),$$

where we have used the symmetry of the probability density functions. Therefore, we may write

$$P_b = \int_{-\infty}^{\gamma_0} p(z|s_1)dz = \int_{\gamma_0}^{\infty} p(z|s_2)dz.$$

Substituting (8.47) into the above equation yields

$$P_b = Q\left(\frac{a_1 - a_2}{2\sigma_0}\right), \tag{8.52}$$

where Q stands for the complementary error function

$$Q(x) = \int_x^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du.$$

In order to minimize the probability of bit error, we need to choose the optimum linear filter and the optimum decision threshold shown in Figure 8.13. For the binary system, the optimum decision threshold has

already been given in (8.48). Thus we only need to determine the linear filter that maximizes $(a_1 - a_2)/2\sigma_0$, or equivalently, $(a_1 - a_2)^2/\sigma_0^2$. Assume that the filter is matched to the input difference signal $s_1(t) - s_2(t)$. The maximum output SNR at time $t = T_s$ is then given by

$$\left.\frac{S}{N}\right|_{T_s} = \frac{(a_1 - a_2)^2}{\sigma_0^2} = \frac{2E_d}{N_0}, \tag{8.53}$$

where $N_0/2$ is the two-sided PSD of the noise at the filter input and

$$E_d = \int_0^{T_s} [s_1(t) - s_2(t)]^2 dt$$

is the energy of the difference signal at the filter input. Note that

$$\begin{aligned} E_d &= \int_0^{T_s} s_1^2(t)dt + \int_0^{T_s} s_2^2(t)dt - 2\int_0^{T_s} s_1(t)s_2(t)dt \\ &= 2E_b(1-\rho), \end{aligned} \tag{8.54}$$

where

$$\rho = \frac{1}{E_b}\int_0^{T_s} s_1(t)s_2(t)dt$$

is called time cross-correlation coefficient. Considering (8.53) and (8.54), (8.52) can be written as

$$P_b = Q\left(\sqrt{\frac{E_b(1-\rho)}{N_0}}\right). \tag{8.55}$$

If the two signals s_1 and s_2 are orthogonal, (8.55) reduces to

$$P_b = Q\left(\sqrt{\frac{E_b}{N_0}}\right). \tag{8.56}$$

If the two signals s_1 and s_2 are antipodal (i.e., the angle between the signal vectors is 180°), we have $\rho = -1$. In this case, (8.55) can be written as

$$P_b = Q\left(\sqrt{\frac{2E_b}{N_0}}\right). \tag{8.57}$$

Table 8.3 Probability of bit error

Modulation	P_b (probability of bit error)
Equally likely coherent BPSK	$Q\left(\sqrt{\frac{2E_b}{N_0}}\right)$
Differentially non-coherent DPSK	$\frac{1}{2}\exp\left(-\frac{E_b}{N_0}\right)$
Equally likely coherent FSK	$Q\left(\sqrt{\frac{E_b}{N_0}}\right)$
Equally likely non-coherent FSK	$\frac{1}{2}\exp\left(-\frac{1}{2}\frac{E_b}{N_0}\right)$

Table 8.4 Probability of symbol error

Modulation	P_s (probability of symbol error)
Equally likely coherent MPSK	$2Q\left(\sqrt{\frac{2E_s}{N_0}}\sin\frac{\pi}{M}\right)$
Differentially coherent MDPSK	$2Q\left(\sqrt{\frac{2E_s}{N_0}}\sin\frac{\pi}{\sqrt{2}M}\right),\ \frac{E_s}{N_0} >> 1$
Equally likely coherent MFSK	$\leq(M-1)Q\left(\sqrt{\frac{E_s}{N_0}}\right)$
Equally likely non-coherent MFSK	$\frac{1}{M}\exp\left(-\frac{E_s}{N_0}\right)\sum_{j=2}^{M}(-1)^j\frac{M!}{j!(M-j)!}\exp\left(\frac{E_s}{jN_0}\right)$
Coherent MQAM	$4\left(1-\frac{1}{\sqrt{M}}\right)Q\left(\sqrt{\frac{2E_{s\,\min}}{N_0}}\right)$

The typical probability of bit error P_b vs. E_b/N_0 curve has a waterfall-like shape. The error performances for binary systems are summarized in Table 8.3. The error performances for M-ary systems are summarized in Table 8.4, where $E_s = E_b \log_2 M$ is the energy per symbol, and $M = 2^k$ is the size of the symbol set, and P_s is the probability of symbol error.

The ratio E_b/N_0 is a natural figure-of-merit for digital communication systems, which allow us to compare one system with another. At the bit level, it indicates how much energy per bit is required for a given bit error probability.

Remark 8.8: The digital waveform may contain 1 bit (binary), 2 bits (4-ary), 3 bits (8-ary), ..., a description of the digital waveforms in terms of S/N is thus virtually useless. □

Remark 8.9: It can be shown that for an M-ary orthogonal signal set, the relationship between probability of bit error P_b and probability of symbol error P_s is

$$\frac{P_b}{P_s} = \frac{2^{k-1}}{2^k - 1} = \frac{M/2}{M-1} \to \frac{1}{2}$$

for large k. □

The design of any digital communication system begins with a description of the channel (received power, available bandwidth, noise statistics, and other impairments such as fading), and a definition of the system requirements (data rate and error performance). Two primary communication resources are the received power and the available transmission bandwidth. In many communication systems, one of these resources may be more precious than the other, and hence most systems can be classified as either bandwidth-limited or power-limited. In bandwidth-limited systems, spectrally efficient modulation techniques can be used to save bandwidth at the expenses of power. In power-limited systems, power-efficient modulation techniques can be used to save power at the expense of bandwidth. In both bandwidth- and power-limited systems, error correction coding (often called channel coding) can be used to save power or to improve error performance at the expense of bandwidth. Given the available bandwidth, the available received SNR (determined by transmit power, antenna gains, path loss, etc.), the required data rate, and the required probability of bit error, a modulation scheme can be chosen to meet the performance requirements (Sklar, 1993).

Example 8.4: Consider a bandwidth-limited system with an AWGN radio channel of available bandwidth $B = 4$ kHz. The received SNR is assumed to be $S/N_0 = 53\,\text{dB}$. The required data rate is $R_b = 9600\,\text{bit/s}$, and the required error performance is $P_b \leq 10^{-5}$. Since the required data rate is much higher than the available bandwidth, the channel is bandwidth limited. Therefore, we may select MPSK as the modulation scheme. From the relationship

$$R_s = \frac{R_b}{\log_2 M},$$

we may find the smallest possible value of M that satisfies $R_s < B$ is $M = 8$. The power efficiency is given by

$$\eta_p = \frac{E_b}{N_0}(\text{dB}) = \frac{S}{N_0}(\text{dB}) - R_b(\text{dB}) = 13.2\,\text{dB}.$$

It can be shown that the above power efficiency value meets the requirement of bit-error performance for 8-PSK modulation scheme. □

8.3.3 Link Budget Analysis

The link budget is a balance sheet of gains and losses. By examining the link budget we can learn many things about overall system design and performance. The link budget outlines the distribution of transmission and reception resources, noise sources, signal attenuators, and effects of processes throughout the link. As radio waves propagate in free space, power falls off as the square of range. This effect is due to the spreading of the radio waves as they propagate. For an isotropic point source, the power density on a sphere at distance R is $\frac{P_{\text{rad}}}{4\pi R^2}(W/m^2)$, where P_{rad} is the radiated power of the point source. The power accepted by receiving antenna can then be written

$$P_{\text{rec}}(\mathbf{u}_r) = \frac{P_{\text{rad}}}{4\pi R^2} A_e(\mathbf{u}_r),$$

where $A_e(\mathbf{u}_r)$ is the effective area of the receiving antenna in the direction of $\mathbf{u}_r$ (a unit vector directed from receiving antenna to the point source). If the transmitting antenna is not an isotropic source, the received power in general will be of the form

$$\begin{aligned} P_{\text{rec}}(\mathbf{u}_t, \mathbf{u}_r) &= \frac{P_a G_t(\mathbf{u}_t)}{4\pi R^2} A_e(\mathbf{u}_r) \\ &= \text{EIRP}(\mathbf{u}_t)\frac{A_e(\mathbf{u}_r)}{4\pi R^2} = \text{EIRP}(\mathbf{u}_t)\frac{G_r(\mathbf{u}_r)}{L_s}, \end{aligned} \tag{8.58}$$

where P_a is the input power to the antenna terminal, $\mathbf{u}_t$ is the unit vector directed from the transmitting antenna to receiving antenna,

$$L_s = \left(\frac{4\pi R}{\lambda}\right)^2$$

is the **free space path loss**, G_t and G_r are the transmitting and receiving antenna gain respectively, and EIRP is the **effective isotropic radiated**

power, which is defined by

$$\text{EIRP} = P_a G_t(\mathbf{u}_t).$$

The **received isotropic power** is defined as

$$p_{\text{rec}} = \frac{\text{EIRP}}{L_s}, \tag{8.59}$$

which is the power received by an isotropic antenna ($G_r = 1$).

8.3.3.1 *Link Margin, Noise Figure and Noise Temperature*

The SNR after the receiving antenna as shown in Figure 8.16 (before the LNA) is

$$\text{SNR} = \frac{P_{\text{rec}}}{N} = \text{EIRP}\frac{G_r}{L_s L_o N},$$

where L_o represents other possible losses.

For optimum (matched filter) detection signal bandwidth is equal to the noise bandwidth. Therefore noise is usually normalized, i.e., the noise spectral density $N_0 = kT_i$ is used instead of the total noise power N:

$$\text{SNR} = \frac{P_{\text{rec}}}{N_0} = \text{EIRP}\frac{G_r}{L_s L_o N_0}.$$

Figure 8.17 shows a typical waterfall-like curve of error probability vs. E_b/N_0 for the digital communication system, where E_b is the energy per

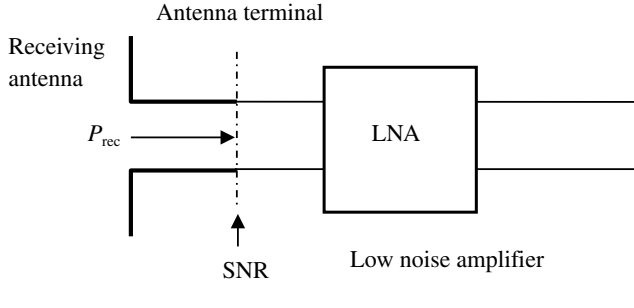


Figure 8.16 Signal-to-noise ratio.

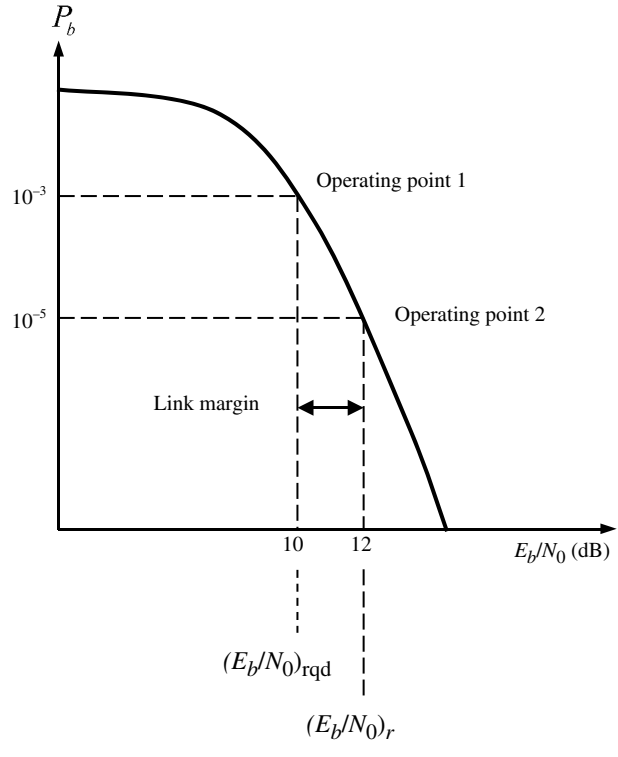


Figure 8.17 Error probability.

bit. Once a modulation scheme has been chosen, the requirement to meet a particular error probability dictates a particular point on the curve (operating point 1). In other words, the required error performance dictates the value of E_b/N_0 that must be made available at the receiver input to meet that performance. The task of a link analysis is to determine an actual system operating point (operating point 2) and to establish that the error probability associated with that point is less than or equal to the system requirement. The **link margin** (or **safety factor**) is defined by

$$M = \frac{(E_b/N_0)_r}{(E_b/N_0)_{\text{rqd}}}.$$

where $(E_b/N_0)_{\text{rqd}}$ is the value required to yield a specified bit error ratio, $(E_b/N_0)_r$ is the value actually received. Since $P_{\text{rec}} = E_b R_b$, where R_b is the

bit rate, we have

$$M = \frac{(E_b/N_0)_r}{(E_b/N_0)_{\text{rqd}}} = \frac{P_{\text{rec}}/N_0}{(E_b/N_0)_{\text{rqd}} R_b} = \text{EIRP}\frac{G_r}{L_s L_o k T_i (E_b/N_0)_{\text{rqd}} R_b}.$$

The noise figure of an amplifier is given by

$$F = \frac{S_{\text{in}}/N_{\text{in}}}{S_{\text{out}}/N_{\text{out}}} = \frac{S_{\text{in}}/N_{\text{in}}}{G_p S_{in}/G_p(N_{\text{in}} + N_{\text{internal}})} = 1 + \frac{N_{\text{internal}}}{N_i}, \qquad (8.60)$$

where G_p is the amplifier gain, N_{in} is the noise power into amplifier and N_{internal} is the amplifier noise referred to input. The noise figure is a measure of the amount of noise added by the amplifier itself. If we introduce the **amplifier equivalent noise temperature** T_r, which is defined by

$$N_{\text{internal}} = kT_r B,$$

then from (8.60) and $N_{\text{in}} = kT_i B$, we obtain amplifier equivalent noise temperature

$$T_r = (F - 1)T_i.$$

The line loss can be treated like noise figure, which is defined by

$$F_l = \frac{S_{\text{in}}/N_{\text{in}}}{S_{\text{out}}/N_{\text{out}}} = \frac{S_{\text{in}}/N_{\text{in}}}{AS_{\text{in}}/A(N_{\text{in}} + N_l)} = 1 + \frac{N_l}{N_{\text{in}}},$$

where A is the attenuation of the line, N_{in} is the noise power into the lossy line and N_l is the equivalent line noise referred to input. If we introduce the line equivalent noise temperature T_l, i.e.,

$$N_l = kT_l B,$$

we have

$$T_l = (F_l - 1)T_i.$$

The overall noise figure for a cascade of stages can be obtained in terms of the noise figure and gain of each stage. Consider the system shown in Figure 8.18, the composite noise figure is of the form

$$F_{\text{comp}} = F_1 + \frac{F_2 - 1}{G_1}.$$

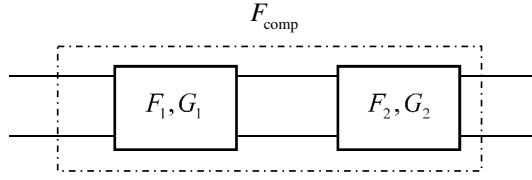


Figure 8.18 Noise figure of composite system.

8.3.3.2 *Link Budget Analysis for Mobile Systems*

The link budget analysis for a wireless system is based on the following relation

$$P_{\text{rec}} = \text{EIRP}\frac{G_r}{L_s L_o}, \tag{8.61}$$

where L_o stands for all possible losses other than free space path loss. We can also start from (8.60) to obtain

$$P_{\text{rec}}(\text{dBm}) = F(\text{dB}) + E_b/N_0(\text{dB}) + R_b(\text{dB}) + kT \text{ (dBm/Hz)}. \tag{8.62}$$

The **sensitivity** of an RF receiver is defined as the minimum signal level that the system can detect with acceptable SNR, which can be obtained from the above equation by requiring the BER not exceed a specified value (say 2%, thus the required E_b/N_0 is known).

When we do the link budget analysis, two major margins should be added to the loss L_o, which are model margin and fading margin. Model margin is considered because of the lack of confidence of the propagation model accuracy. Sometimes this is not necessary if we can make direct measurement or have more accurate modeling. The fading margin is necessary unless we can eliminate fading by other means. In the following, we give an example from GSM system design to explain the link budget analysis using (8.62).

1. The Minimum Signal Level at Base Station

The first step in performing the link budget is to determine the required signal strength at the receiver input, i.e., receiver sensitivity. The receiver input sensitivity at base station can be determined from

$$P_{\text{rec},b}(\text{dBm}) = F(\text{dB}) + E_b/N_0(\text{dB}) + R_b(\text{dB}) + kT(\text{dBm/Hz}),$$

where $kT = -174\,\text{dBm/Hz@290 K}$. For 10^{-2} BER, we have $E_b/N_0 = 8$ dB. Assume that

$$R_b = 10\log 270.833\,\text{kbit/s} = 54.32\,\text{dB}, \quad F = 8\,\text{dB}.$$

Thus, the receiver sensitivity is $P_{\text{rec},b} = -104\,\text{dBm}$. The received isotropic power is then given by

$$p_{\text{rec},b} = \frac{\text{EIRP}}{L_s} = \frac{P_{\text{rec},b}}{G_r}.$$

Hence, the minimum isotropic power at base station required from the mobile station is

$$\begin{aligned}\min p_{\text{rec},b} = {} & \text{Receiver input sensitivity } (-104\,\text{dBm}) \\ & - \text{Receiving antenna gain } (12\,\text{dBi}) \\ & + \text{Interference margin } (3\,\text{dB}) \\ & + \text{Cable loss } (4\,\text{dB}) \\ & + \text{log normal fading margin } (5\,\text{dB}) \\ & + \text{Rayleigh fading margin } (10\,\text{dB}) \\ = {} & -94\,\text{dBm}. \end{aligned} \tag{8.63}$$

2. The Minimum Mobile Transmit Power

The minimum required power from the mobile station to maintain a 10^{-2} BER can be found by (8.63) after the characterization of the path loss. Here we use Hata propagation model. If $f = 900\,\text{MHz}$, $h_b = 62\,\text{m}$, $h_m = 2\,\text{m}$ and the city is large we have

$$L_s = 121 + 33\log R(\text{dB}).$$

According to (8.59), the minimum power needed from the mobile station to maintain a 10^{-2} BER is

$$\begin{aligned}\min \text{EIRP} &= \min p_{\text{rec},b}(\text{dBm}) + L_s(\text{dB}) \\ &= \begin{cases} -94 + 121 = 27\,\text{dBm@1 km } (500\,\text{mW}) \\ -94 + 121 + 33\log 1.5 = 32.8\,\text{dBm@1.5 km } (1.9\,\text{W}) \\ -94 + 121 + 33\log 2 = 36.9\,\text{dBm@2 km } (4.9\,\text{W}) \end{cases}. \end{aligned}$$

8.3.4 Mobile Antennas and Environments

In a mobile environment, there is a strong correlation between antenna pattern and the statistics of the received signal strength.

8.3.4.1 *Incident Signal*

Consider two reference systems S and $\tilde{S}$, where $\tilde{S}$ is moving in the positive x-direction with a velocity $\mathbf{v} = v\mathbf{u}_x$ as viewed from S (Figure 8.19). The two reference systems are related by the Lorentz transformation (e.g., Geyi, 2010)

$$\begin{aligned} \mathbf{r} &= \overleftrightarrow{\boldsymbol{\alpha}} \cdot \tilde{\mathbf{r}} + \gamma \boldsymbol{\beta} c\tilde{t}, \\ ct &= \gamma(c\tilde{t} + \boldsymbol{\beta} \cdot \tilde{\mathbf{r}}), \end{aligned} \tag{8.64}$$

where $\boldsymbol{\beta} = \mathbf{v}/c$, $\gamma = 1/\sqrt{1-\beta^2}$, $\beta = v/c$, and $\overleftrightarrow{\boldsymbol{\alpha}}$ is a dyadic defined by

$$\overleftrightarrow{\boldsymbol{\alpha}} = \overleftrightarrow{\mathbf{I}} + (\gamma - 1)\frac{\boldsymbol{\beta}\boldsymbol{\beta}}{\beta^2},$$

and $\overleftrightarrow{\mathbf{I}}$ is the identity dyad.

An incident plane wave (modulated signal) in the S system can be represented by

$$\mathbf{E}_{\text{in}} = \mathbf{E}_0 \cos(\omega_c t + \mathbf{k} \cdot \mathbf{r}),$$

where $\mathbf{k} = k\mathbf{u}_k$ is the wave vector with $k = \omega_c\sqrt{\mu\varepsilon}$ being the wavenumber, and $\mathbf{E}_0$ is a constant vector. The plane wave in the $\tilde{S}$ system can be expressed as

$$\tilde{\mathbf{E}}_{\text{in}} = \tilde{\mathbf{E}}_0 \cos(\tilde{\omega}_c\tilde{t} + \tilde{\mathbf{k}} \cdot \tilde{\mathbf{r}}).$$

Under the Lorentz transformation (8.64), the field components that parallel to the velocity $\mathbf{v}$ is invariant. So we have

$$E_{0x} \cos(\omega_c t + \mathbf{k} \cdot \mathbf{r}) = \tilde{E}_{0x} \cos(\tilde{\omega}_c\tilde{t} + \tilde{\mathbf{k}} \cdot \tilde{\mathbf{r}}),$$

where the subscript x denotes the x component. The Lorentz transformation assumes that the two origins of the systems S and $\tilde{S}$ coincide at

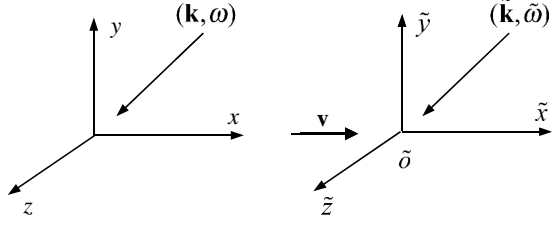


Figure 8.19 Transformation of plane wave.

$t = \tilde{t} = 0$. As a result, we have $E_{0x} = \widetilde{E}_{0x}$. This implies that the phase $\varphi = \omega_c t + \mathbf{k} \cdot \mathbf{r}$ is an invariant

$$\omega_c t + \mathbf{k} \cdot \mathbf{r} = \widetilde{\omega}_c \tilde{t} + \tilde{\mathbf{k}} \cdot \tilde{\mathbf{r}}. \tag{8.65}$$

Introducing the Lorentz transformation into the above equation, we obtain

$$\begin{aligned} \tilde{\mathbf{k}} &= \vec{\overleftrightarrow{\alpha}} \cdot \mathbf{k} + \frac{\omega_c \gamma}{c} \boldsymbol{\beta}, \\ \widetilde{\omega}_c &= \gamma(\omega_c + \mathbf{v} \cdot \mathbf{k}). \end{aligned} \tag{8.66}$$

Equation (8.66) indicates that the relative motion between two observers introduces a Doppler shift. Let us consider a mobile terminal traveling in the x direction with speed v, illuminated by N-plane waves of the same carrier frequency ω_c. The incident electric field at the mobile terminal can be written as

$$\begin{aligned} \mathbf{E}_{\text{in}} &= \sum_{n=1}^{N} \tilde{\mathbf{E}}_{0n} \cos(\widetilde{\omega}_c t + \tilde{\mathbf{k}}_n \cdot \tilde{\mathbf{r}}) \approx \sum_{n=1}^{N} \mathbf{E}_{0n} \cos[(\omega_c + \mathbf{v} \cdot \mathbf{k}_n)t + \mathbf{k}_n \cdot \mathbf{r}] \\ &= \sum_{n=1}^{N} \mathbf{E}_{0n} \cos(\omega_c t + \theta_n) \end{aligned}$$

where $\theta_n = \omega_n t + \varphi_n$ with $\omega_n = \mathbf{v} \cdot \mathbf{k}_n$, $\varphi_n = \mathbf{k}_n \cdot \mathbf{r}$. According to the central limit theorem (e.g., Proakis, 1995), the electric field $\mathbf{E}_{\text{in}}$ may be considered as a Gaussian random process as $N \to \infty$. The above equation can be written as

$$\mathbf{E}_{\text{in}} = \mathbf{x}(t) \cos \omega_c t - \mathbf{y}(t) \sin \omega_c t = \text{Re}\, \mathbf{E}_{\text{in,en}}(t) e^{j\omega_c t}, \tag{8.67}$$

where $\mathbf{E}_{\text{in,en}}(t) = \mathbf{x}(t) + j\mathbf{y}(t)$ is the complex envelope of the modulated signal $\mathbf{E}_{\text{in}}$, and

$$\begin{aligned} \mathbf{x}(t) &= \sum_{n=1}^{N} \mathbf{E}_{0n} \cos(\omega_n t + \varphi_n), \\ \mathbf{y}(t) &= \sum_{n=1}^{N} \mathbf{E}_{0n} \sin(\omega_n t + \varphi_n) \end{aligned} \tag{8.68}$$

are the in-phase and quadrature components of $\mathbf{E}_{\text{in}}$ respectively. They are also Gaussian random processes.

Remark 8.10 (Rayleigh distribution and Rician distribution)**:** Rayleigh distributions and Rician distributions are used to model scattered signals that reach a receiver by multiple paths. Rayleigh distributions are used to characterize dense scatterers, while Rician distributions are used for the

scenario with a line-of-sight path between the transmitter and the receiver. Consider a field component seen at the mobile

$$E = \sum_{n=1}^{N} E_n \cos(\omega_c t + \theta_n) = x \cos \omega_c t - y \sin \omega_c t = \mathrm{Re} E_{\mathrm{en}} e^{j\omega_c t}, \quad (8.69)$$

where

$$x = \sum_{n=1}^{N} E_n \cos \theta_n, \quad y = \sum_{n=1}^{N} E_n \sin \theta_n, \quad E_{\mathrm{in}} = x + jy.$$

If x and y are Gaussian random processes, they have probability density functions of the form

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2},$$

where σ is the variance. It can be shown that the envelope

$$r = \sqrt{x^2 + y^2}$$

obeys Raleigh distribution

$$p(r) = \begin{cases} \dfrac{r}{\sigma^2} e^{-r^2/2\sigma^2}, & r \geq 0 \\ 0, & r < 0 \end{cases}.$$

If the incident field consists of a strong sinusoidal wave and a random noise E given by (8.69)

$$\begin{aligned} F &= q \cos \omega_c t + E \\ &= (q + x) \cos \omega_c t - y \sin \omega_c t \\ &= r \cos \theta \cos \omega_c t - r \sin \theta \sin \omega_c t \\ &= r \cos(\omega_c t + \theta), \end{aligned} \quad (8.70)$$

where the envelope r and the phase angle θ are defined by

$$r \cos \theta = q + x, \quad r \sin \theta = y.$$

It can be shown that the envelope r obeys Rician distribution

$$p(r) = \frac{r}{\sigma^2} I_0\Big(\frac{rq}{\sigma^2}\Big) e^{-(r^2+q^2)/2\sigma^2}, \quad (8.71)$$

where I_0 is the zero-order modified Bessel function of the first kind. □

8.3.4.2 *Received Signal by Mobile Antenna*

In terms of the reciprocity theorem, the open circuit voltage at the antenna terminal induced by the incident fields $\mathbf{E}_{\text{in}}$, $\mathbf{H}_{\text{in}}$ can be written as (see Chapter 5)

$$\begin{aligned} V_{\text{oc}} &= -\frac{1}{I}\int_{V_0} \mathbf{E}_{\text{in}}(\mathbf{r}')\cdot\mathbf{J}(\mathbf{r}')dV(\mathbf{r}') \\ &= -\frac{1}{I}\int_{S} [\mathbf{E}(\mathbf{r}')\times\mathbf{H}_{\text{in}}(\mathbf{r}') - \mathbf{E}_{\text{in}}(\mathbf{r}')\times\mathbf{H}(\mathbf{r}')]\cdot\mathbf{u}_n dS(\mathbf{r}'), \end{aligned} \tag{8.72}$$

where $\mathbf{J}$ is the current distribution confined in the source region V_0 and it generates the fields $\mathbf{E}, \mathbf{H}$ when the antenna is used as a transmitting antenna; I is the terminal current; and S is a closed surface containing the source region V_0. Let S be a sphere located in the far-field region of the antenna. Then we have

$$\eta\mathbf{H} = \mathbf{u}_n \times \mathbf{E}.$$

Substituting the above equation into (8.72) gives

$$\begin{aligned} V_{\text{oc}} &= -\frac{1}{I}\int_{S} [\mathbf{E}\times\mathbf{H}_{\text{in}} - \eta^{-1}\mathbf{E}_{\text{in}}\times(\mathbf{u}_n\times\mathbf{E})]\cdot\mathbf{u}_n dS \\ &= -\frac{1}{I}\int_{S} [-\mathbf{E}\cdot(\mathbf{u}_n\times\mathbf{H}_{\text{in}}) - \eta^{-1}(\mathbf{E}_{\text{in}}\cdot\mathbf{E}) + \eta^{-1}(\mathbf{u}_n\cdot\mathbf{E})(\mathbf{u}_n\cdot\mathbf{E}_{\text{in}})]dS \\ &= \frac{1}{I\eta}\int_{S} \mathbf{E}\cdot(\eta\mathbf{u}_n\times\mathbf{H}_{\text{in}} + \mathbf{E}_{\text{in}})dS. \end{aligned}$$

Let us consider a mobile receiving antenna. If the incident field (8.67) consists of plane waves such that the approximation $\eta\mathbf{u}_n \times \mathbf{H}_{\text{in}} \approx \mathbf{E}_{\text{in}}$ is valid, we have

$$V_{\text{oc}}(t) = \frac{2}{I\eta}\int_{S} \mathbf{E}(\mathbf{r})\cdot\mathbf{E}_{\text{in}}(\mathbf{r},t)dS(\mathbf{r}). \tag{8.73}$$

When the antenna is conjugately matched to a load R_L, the received power is given by

$$\begin{aligned} P_{\text{rec}}(t) = \frac{|V_{\text{oc}}(t)|^2}{8R_L} = \frac{1}{2|I|^2R_L\eta^2}\int_{S}\int_{S} &\left[\mathbf{E}(\mathbf{r})\cdot\mathbf{E}_{\text{in}}(\mathbf{r},t)\right] \\ &\times\left[\bar{\mathbf{E}}(\mathbf{r}')\cdot\bar{\mathbf{E}}_{\text{in}}(\mathbf{r}',t)\right]dS(\mathbf{r})dS(\mathbf{r}'). \end{aligned} \tag{8.74}$$

In spherical coordinate system, we may write

$$\mathbf{E} = E_\theta \mathbf{u}_\theta + E_\varphi \mathbf{u}_\varphi, \quad \mathbf{E}_{\text{in}} = E_{\text{in},\theta}\mathbf{u}_\theta + E_{\text{in},\varphi}\mathbf{u}_\varphi.$$

Substituting these into (8.74) yields

$$\begin{aligned} P_{\text{rec}}(t) = \frac{1}{2|I|^2 R_L \eta^2} \Bigg[& \int_S \int_S E_\theta(\mathbf{r})\bar{E}_\theta(\mathbf{r}')E_{\text{in},\theta}(\mathbf{r})\bar{E}_{\text{in},\theta}(\mathbf{r}')dS(\mathbf{r})dS(\mathbf{r}') \\ & + \int_S \int_S E_\varphi(\mathbf{r})\bar{E}_\varphi(\mathbf{r}')E_{\text{in},\varphi}(\mathbf{r})\bar{E}_{\text{in},\varphi}(\mathbf{r}')dS(\mathbf{r})dS(\mathbf{r}') \\ & + \int_S \int_S E_\theta(\mathbf{r})\bar{E}_\varphi(\mathbf{r}')E_{\text{in},\theta}(\mathbf{r})\bar{E}_{\text{in},\varphi}(\mathbf{r}')dS(\mathbf{r})dS(\mathbf{r}') \\ & + \int_S \int_S E_\varphi(\mathbf{r})\bar{E}_\theta(\mathbf{r}')E_{\text{in},\varphi}(\mathbf{r})\bar{E}_{\text{in},\theta}(\mathbf{r}')dS(\mathbf{r})dS(\mathbf{r}') \Bigg]. \end{aligned} \tag{8.75}$$

We now assume that the field components (either co-polarized or cross-polarized) of plane waves arriving from different directions are independent:

$$\begin{aligned} \langle E_{\text{in},\theta}(\mathbf{r})\bar{E}_{\text{in},\theta}(\mathbf{r}')\rangle &= \langle E_{\text{in},\theta}(\mathbf{r})\bar{E}_{\text{in},\theta}(\mathbf{r})\rangle\delta(\mathbf{r}-\mathbf{r}'), \\ \langle E_{\text{in},\theta}(\mathbf{r})\bar{E}_{\text{in},\varphi}(\mathbf{r}')\rangle &= \langle E_{\text{in},\varphi}(\mathbf{r})\bar{E}_{\text{in},\theta}(\mathbf{r}')\rangle = 0. \end{aligned} \tag{8.76}$$

The received power averaged over a random route is then given by

$$\begin{aligned} \langle P_{\text{rec}}(t)\rangle = \frac{1}{|I|^2 R_L \eta} \int_\Omega \frac{1}{2\eta} r^2 \big[& |E_\theta(\mathbf{r})|^2 \langle |E_{\text{in},\theta}(\mathbf{r},t)|^2\rangle \\ & + |E_\varphi(\mathbf{r})|^2 \langle |E_{\text{in},\varphi}(\mathbf{r},t)|^2\rangle \big] d\Omega(\mathbf{r}), \end{aligned} \tag{8.77}$$

where Ω is the unit sphere. Let

$$\langle |E_{\text{in},\theta}(\mathbf{r},t)|^2\rangle = C_1 p_\theta(\mathbf{u}_r), \quad \langle |E_{\text{in},\varphi}(\mathbf{r},t)|^2\rangle = C_2 p_\varphi(\mathbf{u}_r),$$

where C_1 and C_2 are two constants; $p_\theta(\mathbf{u}_r)$ and $p_\varphi(\mathbf{u}_r)$ are angular power density function satisfying

$$\int_\Omega p_\theta(\mathbf{u}_r)d\Omega(\mathbf{r}) = 1, \quad \int_\Omega p_\varphi(\mathbf{u}_r)d\Omega(\mathbf{r}) = 1.$$

Equation (8.77) can be rewritten as

$$\langle P_{\text{rec}}(t)\rangle = \int_\Omega [P_1 G_\theta(\mathbf{u}_r)p_\theta(\mathbf{u}_r) + P_2 G_\varphi(\mathbf{u}_r)p_\varphi(\mathbf{u}_r)]d\Omega(\mathbf{r}), \tag{8.78}$$

where P_1 and P_2 are two constants and can be interpreted as the received power (averaged over the same route) by an isotropic antenna polarized in $\mathbf{u}_\theta$ and $\mathbf{u}_\varphi$ respectively.

8.3.4.3 *Mean Effective Gain*

In a mobile environment, the evaluation of the antenna performance is not a trivial task due to the multipath propagations. The **mean effective gain** (MEG) is a statistical measure of the antenna gain in a mobile environment. It is defined as the ratio between the mean received power of the antenna and the total mean incident power when moving the antenna over a random route. In terms of (8.78), the MEG can be expressed as

$$\begin{aligned}\text{MEG} &= \frac{\langle P_{\text{rec}}(t)\rangle}{P_1+P_2} \\ &= \int_\Omega \left[\frac{\text{XPR}}{1+\text{XPR}} G_\theta(\mathbf{u}_r)p_\theta(\mathbf{u}_r) + \frac{1}{1+\text{XPR}} G_\varphi(\mathbf{u}_r)p_\varphi(\mathbf{u}_r)\right] d\Omega(\mathbf{r}),\end{aligned} \tag{8.79}$$

where XPR is the cross-polarization power ratio defined by $\text{XPR} = P_1/P_2$.

8.4 Radar Systems

Radar is an object detection system which transmits radio waves toward various targets including aircraft, spacecraft, missiles, vehicles, ships, weather formations, and terrain, etc. The targets reflect a tiny part of the wave's energy to the receiver of the radar. By comparing the transmitting and receiving waves, the properties of the targets such as the distance, altitude, direction, or speed can be determined. The term RADAR was coined by the United States Navy as an acronym for Radio Detection And Ranging in 1940. The applications of radar systems are highly diverse, including various surveillance systems, air and marine navigation, remote sensing, meteorological precipitation monitoring, and geological observations etc.

8.4.1 Radar Signals

There are basically two different radar systems, monostatic system and bistatic system. A **monostatic system** uses the same antenna for both transmit and receive, while the **bistatic system** deploys two separate antennas. Figures 8.20(a) and 8.20(b) respectively show the basic block diagram of a pulsed radar system and the pulse radar signals. The range

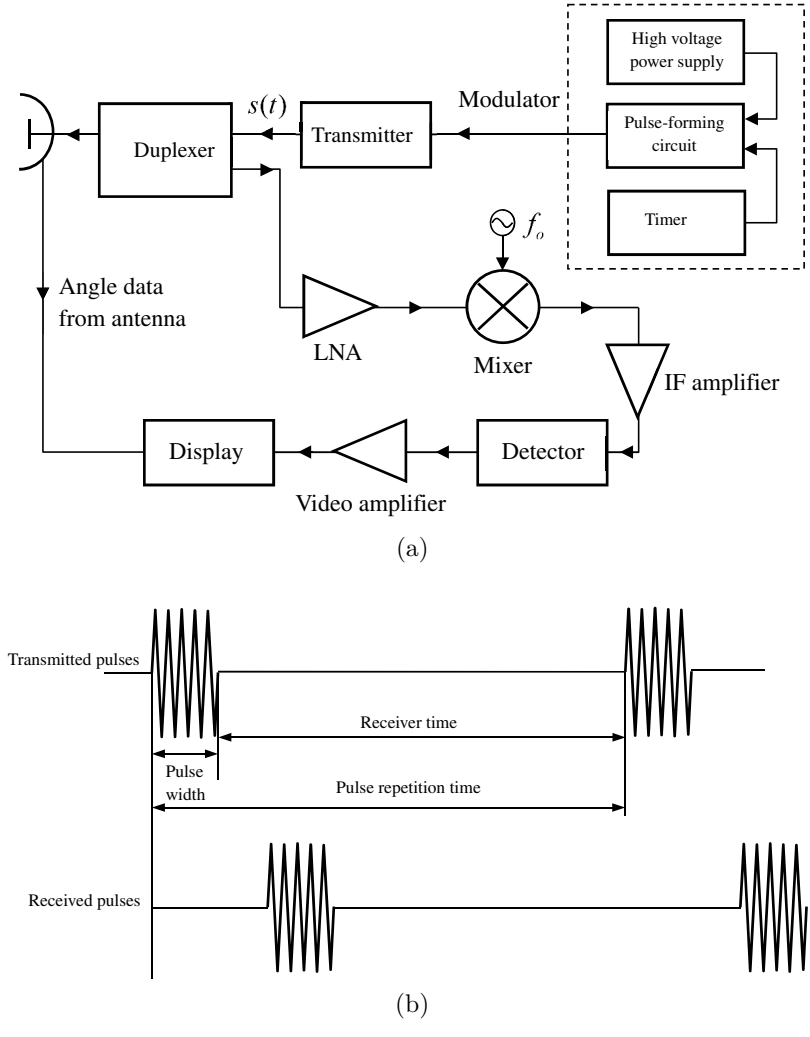


Figure 8.20 (a) Basic block diagram of a pulsed radar. (b) Transmitted and received pulses.

of the target is determined by measuring the round-trip time of pulsed microwave signal. If a narrow beam antenna is used the target's direction may be determined from the beam direction of the antenna. The duration of the pulse is called **pulse width** during which the transmitter is radiating energy. The number of pulses transmitted per second is called **pulse repetition frequency** (PRF). The time from the beginning of one pulse

to the beginning of next pulse is called **pulse repetition time** (PRT), denoted T_r. The time between pulses is called **receiver time** during which the transmitter is turned off. The modulator consists of a pulse-forming circuit, a high-voltage power supply and a timer. The timer controls the PRF, and triggers the pulse-forming circuit and generates high-voltage pulses of rectangular shape. These pulses are used as supply voltage to turn the transmitter on and off, which generates RF carrier. The duplexer is a device that allows bi-directional communication over a single path so that a common antenna can be used for both transmitting and receiving. It also isolates the receiver from the transmitter. The receiver is of common superheterodyne type.

The output of the transmitter can be modeled as a Thevenin's equivalent circuit consisting of a voltage source $s(t)$ in series with an output impedance $Z_{\text{out}} = R + jX$. The voltage source may be expressed by

$$s(t) = a(t)\cos[\omega_c t + \varphi(t) + \varphi_0]$$

where $a(t)$ denotes the amplitude; ω_c is the carrier frequency; $\varphi(t)$ is the phase and φ_0 is a random phase angle. We assume that both $a(t)$ and $\varphi(t)$ are slowly varying function and can thus be considered as constants during one cycle of the carrier. When the antenna is conjugately matched to the transmitter, the maximum instantaneous power that can be delivered to the antenna is (called **available instantaneous power**)

$$P_m = \frac{1}{4R}s^2(t) = \frac{1}{4R}\frac{a^2(t)}{2}\{1 + \cos[2\omega_c t + 2\varphi(t) + 2\varphi_0]\}.$$

The (**average**) **peak transmitted power**, denoted P_t, is defined as the available instantaneous power at the output of the transmitter averaged over one cycle of the carrier when $s(t)$ has maximum amplitude. Thus we have

$$P_t = \frac{1}{4R}\frac{\max a^2(t)}{2}. \tag{8.80}$$

The **average transmitted power**, denoted by P_{av}, is defined as the available instantaneous power at the output of the transmitter averaged over the time interval T_r

$$P_{\text{av}} = \frac{1}{4R}\frac{1}{T_r}\int_0^{T_r} s^2(t)dt. \tag{8.81}$$

Usually, we use normalized power so that the factor $1/4R$ does not appear in (8.80) and (8.81). Let the pulse width be denoted by $\tau(\tau < T_r)$. The average transmitted power is related to peak power by

$$P_{\text{av}} \approx D_r P_t, \tag{8.82}$$

where $D_r = \tau/T_r$ is called the **duty cycle** of radar.

8.4.2 Radar Cross Section

The basic function of a radar system is to find targets. In order to characterize how a target interacts with the electromagnetic waves, a quantity called **radar cross section** (RCS) is often introduced. The RCS of a target is defined as a hypothetical area required to intercept the transmitted power density at the target such that if the total intercepted power were re-radiated isotropically, the power density actually observed at the receiver is produced (Skolnick, 1980). Quantitatively, RCS is calculated in three-dimensions as

$$\sigma(\mathbf{u}_t, \mathbf{u}_s) = \lim_{R_2 \to \infty} 4\pi R_2^2 \frac{p_s}{p_{\text{in}}} = \lim_{R_2 \to \infty} 4\pi R_2^2 \frac{|\mathbf{E}_s|^2}{|\mathbf{E}_{\text{in}}|^2}, \tag{8.83}$$

where p_{in} and p_s are the incident and scattered power density in $\mathbf{u}_t$ and $\mathbf{u}_s$ direction respectively; $\mathbf{E}_{\text{in}}$ and $\mathbf{E}_s$ are the corresponding electric field intensities; R_2 is the observation distance from the target, as illustrated in Figure 8.21.

If the target is not very big, the incident wavefronts may be considered plane. Thus the calculation of RCS becomes the scattering problem of

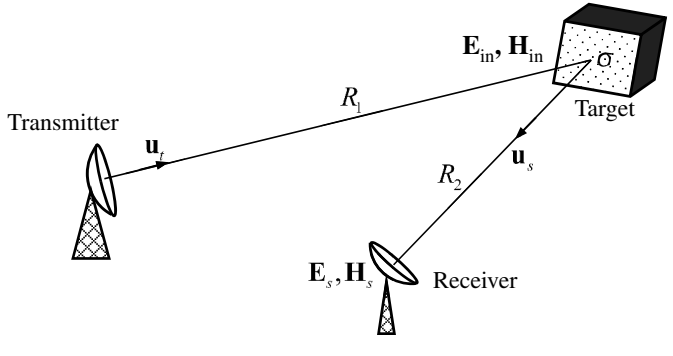


Figure 8.21 Radar cross section.

a plane wave in free space. The problem can be solved exactly only for some regular shapes. In many situations, we have to resort to numerical solutions.

8.4.2.1 *Scattering by Conducting Targets*

Let us consider the scattering of a perfectly conducting target bounded by S, which is illuminated by a plane wave coming from z-direction, as illustrated in Figure 8.22. Let $\mathbf{u}_n$ be the unit normal pointing out of S. The locus of points such that $\mathbf{u}_n \cdot \mathbf{u}_z = 0$ is the shadow boundary and will be denoted by Γ. The shadow boundary Γ divides the surface S into two parts $S = S_1 + S_2$, where S_1 is the shadowed side and S_2 the illuminated side. Let the coordinate system be oriented so that the z-axis lies along the backscattering direction, and passes through the point on the surfaces S, where the normal to the surface $\mathbf{u}_n$ also points in the direction of backscattering. This point is called **specular point** whose position is denoted by $(0, 0, a)$. We further assume that the positive directions of x and y-axes will be along the two principal directions of curvature at the specular point. The fields on the surface will be assumed to drop abruptly to zero as one moves from the illuminated side to the shadowed side. This assumption implies that the fields suffer a discontinuity at the shadow boundary. In this case, the scattered fields can be represented by (e.g., Geyi, 2010)

$$\mathbf{E}_s(\mathbf{r}) = -\int_S [jk\eta G(\mathbf{r}, \mathbf{r}')\mathbf{u}_n(\mathbf{r}') \times \mathbf{H}(\mathbf{r}') + \mathbf{u}_n(\mathbf{r}') \cdot \mathbf{E}(\mathbf{r}')\nabla' G(\mathbf{r}, \mathbf{r}')]dS(\mathbf{r}')$$
$$+ \frac{j}{\omega\varepsilon}\int_\Gamma \mathbf{u}_t(\mathbf{r}') \cdot \mathbf{H}(\mathbf{r}')\nabla' G(\mathbf{r}, \mathbf{r}')d\Gamma(\mathbf{r}'), \tag{8.84}$$

where $\mathbf{E}$ and $\mathbf{H}$ are the total fields on S_2, $G(\mathbf{r}, \mathbf{r}') = e^{-jk|\mathbf{r}-\mathbf{r}'|}/4\pi|\mathbf{r} - \mathbf{r}'|$, and $\mathbf{u}_t$ is the unit tangent vector of Γ in a positive sense that an observer

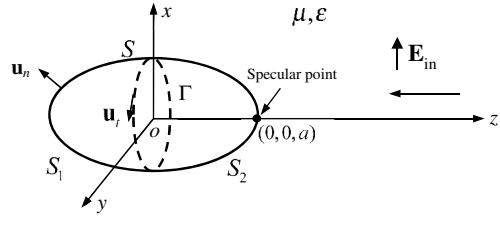


Figure 8.22 Backscattering of conducting target.

moving in the direction $\mathbf{u}_t$ will have the illuminated side of S on his left. Making use of the following identity

$$\begin{aligned}
&\int_{\Gamma} [\mathbf{u}_t(\mathbf{r}') \cdot \mathbf{H}(\mathbf{r}')] \nabla' G(\mathbf{r}, \mathbf{r}') d\Gamma(\mathbf{r}') \\
&\quad = \int_{S_2} \{[\mathbf{u}_n(\mathbf{r}') \times \nabla'] \cdot \mathbf{H}(\mathbf{r}')\} \nabla' G(\mathbf{r}, \mathbf{r}') dS(\mathbf{r}') \\
&\quad = \int_{S_2} \nabla' G(\mathbf{r}, \mathbf{r}') [\mathbf{u}_n(\mathbf{r}') \cdot \nabla' \times \mathbf{H}(\mathbf{r}')] dS(\mathbf{r}') \\
&\qquad - \int_{S_2} \{[\mathbf{u}_n(\mathbf{r}') \times \mathbf{H}(\mathbf{r}')] \cdot \nabla'\} \nabla' G(\mathbf{r}, \mathbf{r}') dS(\mathbf{r}') \\
&\quad = \int_{S_2} j\omega\varepsilon [\mathbf{u}_n(\mathbf{r}') \cdot \mathbf{E}(\mathbf{r}')] \nabla' G(\mathbf{r}, \mathbf{r}') dS(\mathbf{r}') \\
&\qquad - \int_{S_2} \{[\mathbf{u}_n(\mathbf{r}') \times \mathbf{H}(\mathbf{r}')] \cdot \nabla'\} \nabla' G(\mathbf{r}, \mathbf{r}') dS(\mathbf{r}'),
\end{aligned}$$

we obtain

$$\begin{aligned}
\mathbf{E}_s(\mathbf{r}) = &\frac{1}{j\omega\varepsilon} \int_{S_2} k^2 G(\mathbf{r}, \mathbf{r}') \mathbf{u}_n(\mathbf{r}') \times \mathbf{H}(\mathbf{r}') dS(\mathbf{r}') \\
&+ \frac{1}{j\omega\varepsilon} \int_{S_2} \{[\mathbf{u}_n(\mathbf{r}') \times \mathbf{H}(\mathbf{r}')] \cdot \nabla'\} \nabla' G(\mathbf{r}, \mathbf{r}') dS(\mathbf{r}'). \qquad (8.85)
\end{aligned}$$

In order to convert the above integral equations into definite integrals we must relate the scattered fields at the surface to the incident fields. This is difficult in general except when the surface of the reflecting target is such that its radii of curvature at all points are large enough that one can approximate the surface at and near any point as an infinite plane tangent to the surface at that point. Then the relationship between incident and reflected fields at the surface is easily obtained from elementary considerations of plane wave reflection from an infinite plane. These relations are given below:

$$\mathbf{u}_n(\mathbf{r}') \times \mathbf{H}(\mathbf{r}') = 2\mathbf{u}_n(\mathbf{r}') \times \mathbf{H}_{\text{in}}(\mathbf{r}'). \qquad (8.86)$$

Introducing this into (8.85), we have

$$\mathbf{E}_s(\mathbf{r}) = \frac{2}{j\omega\varepsilon}\int\limits_{S_2} k^2 G(\mathbf{r},\mathbf{r}')\mathbf{u}_n(\mathbf{r}')\times\mathbf{H}_{\text{in}}(\mathbf{r}')dS(\mathbf{r}') + \frac{2}{j\omega\varepsilon}\int\limits_{S_2}\left\{[\mathbf{u}_n(\mathbf{r}')\times\mathbf{H}_{\text{in}}(\mathbf{r}')]\cdot\nabla'\right\}\nabla' G(\mathbf{r},\mathbf{r}')dS(\mathbf{r}'). \quad (8.87)$$

Let $\mathbf{r} = (0,0,z)$. For large z, we have the following approximation

$$[(\mathbf{u}_n\times\mathbf{H}_{\text{in}})\cdot\nabla']\nabla' G \approx -k^2[(\mathbf{u}_n\times\mathbf{H}_{\text{in}})\cdot\mathbf{u}_z]G\mathbf{u}_z.$$

Then, we can rewrite (8.87) as

$$\mathbf{E}_s(\mathbf{r}) = -\frac{2k^2}{j\omega\varepsilon}\mathbf{u}_z\times\int\limits_{S_2} G(\mathbf{r},\mathbf{r}')\mathbf{u}_z\times[\mathbf{u}_n(\mathbf{r}')\times\mathbf{H}_{\text{in}}(\mathbf{r}')]\,dS(\mathbf{r}'). \quad (8.88)$$

We can further use the approximation $|\mathbf{r}-\mathbf{r}'|\approx z$ in the denominator of G and

$$|\mathbf{r}-\mathbf{r}'|\approx r-\mathbf{u}_r\cdot\mathbf{r}' = z-\mathbf{u}_z\cdot\mathbf{r}' \quad (8.89)$$

inside the exponential of G. Evidently, $(x',y') = (0,0)$ is a stationary point of (8.89), and in the vicinity of this point the phase of the integrand in (8.88) varies very slowly. So the dominant contribution of the integration (2 1) comes from the immediate neighborhood of $(x',y') = (0,0)$. In this neighborhood, the surface S_2 can be approximated by its osculating quadric, which, in the coordinate system chosen, is given by

$$z' = a - \left[\frac{(x')^2}{2R_1}+\frac{(y')^2}{2R_2}\right]. \quad (8.90)$$

Substituting (8.89) and (8.90) into (8.88) and neglecting the higher-order terms than quadratic one in the exponential, we have

$$\mathbf{E}_s(\mathbf{r}) = -\frac{k\eta e^{-jkz}}{j2\pi z}\mathbf{u}_z\times\int\limits_{S_2}\mathbf{u}_z\times[\mathbf{u}_n(\mathbf{r}')\times\mathbf{H}_{\text{in}}(\mathbf{r}')]\,e^{jk\mathbf{u}_z\cdot\mathbf{r}'}dS(\mathbf{r}'). \quad (8.91)$$

Example 8.5: Let $\mathbf{E}_{\text{in}}(\mathbf{r}) = \mathbf{u}_x E_{\text{in}} e^{jkz}$. Then we have $\mathbf{H}_{\text{in}}(\mathbf{r}) = -\mathbf{u}_y\frac{E_{\text{in}}}{\eta}e^{jkz}$. For a conducting plate as shown in Figure 8.23, we have $R_1 = R_2 = \infty$.

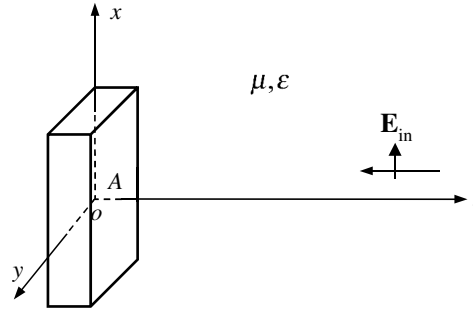


Figure 8.23 Backscattering by a conducting plate.

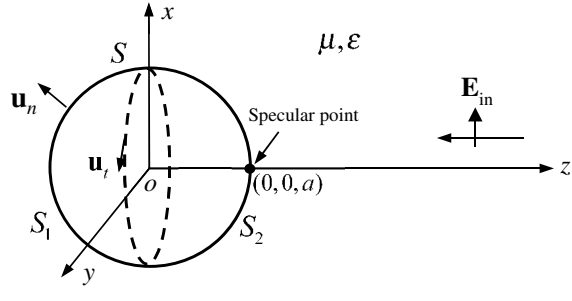


Figure 8.24 Backscattering of conducting sphere.

For large z, (8.91) reduces to

$$\mathbf{E}_s(\mathbf{r}) = \mathbf{u}_x E_{\text{in}} \frac{ke^{-jkz}}{j2\pi z} A, \tag{8.92}$$

where A is surface area of the plate. From (8.83), we may obtain the radar backscatter cross section for a conducting plate as follows

$$\sigma = \frac{4\pi A^2}{\lambda^2}. \tag{8.93}$$

This is an important result and is valid for a conducting plate of any shape. □

Example 8.6: Let $\mathbf{E}_{\text{in}}(\mathbf{r}) = \mathbf{u}_x E_{\text{in}} e^{jkz}$. For a conducting sphere of radius a as shown in Figure 8.24, we have $R_1 = R_2 = a$. For large z, (8.91)

reduces to

$$\begin{aligned}\mathbf{E}_s(\mathbf{r}) &= \frac{k\eta e^{-jkz}}{j2\pi z}\frac{E_{\text{in}}}{\eta}\mathbf{u}_z \times \int_{S_2} \mathbf{u}_z \times \left[\mathbf{u}_r \times \mathbf{u}_y e^{jkz'}\right] e^{jk\mathbf{u}_z\cdot\mathbf{r}'} dS(\mathbf{r}') \\ &= \mathbf{u}_x \frac{kE_{\text{in}}e^{-jkz}}{j2\pi z}\int_0^{\pi/2}\int_0^{2\pi} a^2 \sin\theta' \cos\theta' e^{j2ka\cos\theta'} d\theta' d\varphi' \\ &= -\mathbf{u}_x \frac{E_{\text{in}}ae^{-jkz}e^{j2ka}}{2z}\left(1 + \frac{j}{2ka} - \frac{je^{-j2ka}}{2ka}\right). \end{aligned} \tag{8.94}$$

In terms of (8.83), the radar backscatter cross-section for a conducting sphere may be written as

$$\sigma = \pi a^2 \left|1 + \frac{j}{2ka} - \frac{je^{-j2ka}}{2ka}\right|^2. \tag{8.95}$$

Evidently, the RCS σ approaches the geometrical cross section πa^2 as the sphere becomes electrically large (i.e., $ka \to \infty$). □

8.4.2.2 *Scattering by Rain*

The rain droplet can be approximated by a dielectric sphere with a complex dielectric constant $\tilde{\varepsilon} = \tilde{\varepsilon}_r\varepsilon_0$ with $\tilde{\varepsilon}_r = \tilde{\varepsilon}_r' - j\tilde{\varepsilon}_r''$, as shown in Figure 8.25. An incident plane wave $\mathbf{E}_{\text{in}} = \mathbf{u}_z E_0 e^{jk_0 x}$ with $k_0 = \omega\sqrt{\mu_0\varepsilon_0}$ is assumed to be incident upon the dielectric sphere from positive x-direction. The scattered field in the far-field region is then given by (see (6.79))

$$\mathbf{E}_s(\mathbf{r}) = -\omega k_0 \eta_0 P_0 \sin\theta \frac{e^{-jk_0 r}}{4\pi r}\mathbf{u}_\theta.$$

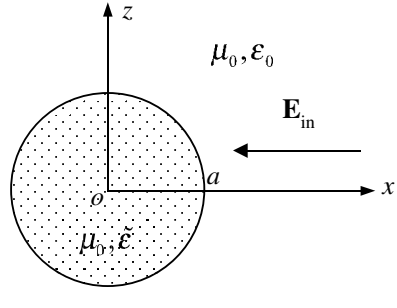


Figure 8.25 Scattering by rain droplet.

The backscattered power density is

$$p_s = \frac{1}{2\eta_0} \left|\mathbf{E}_s(\mathbf{r})\right|^2\big|_{\theta=\pi/2} = \frac{\omega^2 k_0^2 \eta_0 |P_0|^2}{32\pi^2 r^2}.$$

The backscatter cross section is then given by

$$\sigma_{\rm bs} = \lim_{r\to\infty} 4\pi r^2 \frac{\left|\mathbf{E}_s\right|^2}{\left|\mathbf{E}_{\rm in}\right|^2} = 4\pi a^2 (k_0 a)^4 \left|\frac{\tilde{\varepsilon}_r - 1}{\tilde{\varepsilon}_r + 2}\right|^2. \tag{8.96}$$

Note that the scattering cross section is very weak at the longer wavelengths since the cross section depends on $(k_0 a)^4$. Now we assume that the dielectric sphere is located at (r, θ, φ), where r is the distance from the radar and θ and φ are the polar and azimuthal angles measured relative to the bore-sight direction of the radar antenna, as illustrated in Figure 8.26. The power density at the location of the dielectric sphere is

$$p = \frac{P_t G_t(\theta, \varphi)}{4\pi r^2}, \tag{8.97}$$

where P_t is the input power to the antenna terminal, $G_t(\mathbf{u}_t)$ is the gain of the transmitting antenna in the direction (θ, φ). The backscattered power density at the radar location is

$$p_s = p \frac{\sigma_{\rm bs}}{4\pi r^2}. \tag{8.98}$$

The received power by the radar antenna is obtained by multiplying the antenna equivalent area by the backscattered power density p_s. If there is no polarization and impedance mismatch at the antenna, the received power by the radar antenna from a single rain drop is given by

$$dP_{\rm rec} = A_e(\theta, \varphi) p_s = \frac{\lambda^2}{(4\pi)^3 r^4} P_t G_t^2(\theta, \varphi) \sigma_{\rm bs}, \tag{8.99}$$

where λ is the wavelength in free space. For an extended volume of rain, we let $N(a)da$ denote the number of droplets with radii in the interval

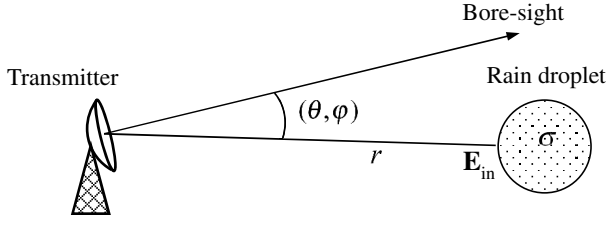


Figure 8.26 Rain droplet illuminated by radar.

$[a, a + da]$ per unit volume. The average backscattering cross section per unit volume, denoted by $\langle \sigma_{\rm bs} \rangle$, is thus given by

$$\langle \sigma_{\rm bs} \rangle = \int_0^\infty \sigma_{\rm bs} N(a) da = \int_0^\infty 4\pi a^2 (k_0 a)^4 \left| \frac{\tilde{\varepsilon}_r - 1}{\tilde{\varepsilon}_r + 2} \right|^2 N(a) da. \tag{8.100}$$

The average received power by the radar antenna from a unit volume of rain drops can be written as

$$\langle dP_{\rm rec} \rangle = \frac{\lambda^2}{(4\pi)^3 r^4} P_t G_t^2(\theta, \varphi) \langle \sigma_{\rm bs} \rangle. \tag{8.101}$$

For non-uniform rain rate, the drop size distribution $N(a)$ depends on the position (r, θ, φ). Therefore, the average cross-section $\langle \sigma_{\rm bs} \rangle$ is a function of (r, θ, φ). The total backscattered power is thus given by

$$P_{\rm rec} = \int_V \langle dP_r \rangle dV = P_t \frac{\lambda^2}{(4\pi)^3} \int_V G_t^2(\theta, \varphi) \frac{\langle \sigma_{\rm bs} \rangle}{r^2} \sin\theta dr\, d\theta\, d\varphi, \tag{8.102}$$

where V stands for the volume of rain illuminated by the radar. For a pulsed radar with pulse width τ, the integration along r is only needed to be carried out in an interval of length $c\tau/2$, where $c = 1/\sqrt{\mu_0 \varepsilon_0}$ This can be understood as follows. We assume that the leading edge of the pulse signal leaves at time $t = 0$, which returns a signal to the receiver from drops at range r_0 (the range of the rain cell being explored and chosen by radar operator) at time $2r_0/c$. The signal that leaves the transmitter at time $t (0 < t \leq \tau)$ will be returned to the receiver at the same time $2r_0/c$ as the leading edge by the drops located at a range $r_0 - ct/2$. Thus, the range interval that returns signals at the same instant of time is $c\tau/2$ long. The average received power from a single transmission of pulse is then given by the integration over the illuminated region by pulse radar (Figure 8.27)

$$P_{\rm rec} = \int_V \langle dP_r \rangle dV = P_t \frac{\lambda^2}{(4\pi)^3 r_0^2} \frac{c\tau}{2} \int_\Omega G_t^2(\theta, \varphi) \langle \sigma_{\rm bs} \rangle \sin\theta\, d\theta\, d\varphi, \tag{8.103}$$

where Ω is the solid angle of the antenna beam.

8.4.2.3 *Effect of Polarization*

All antennas transmit radio waves in a particular polarization and most of them are linearly polarized. The reflection and scattering will typically

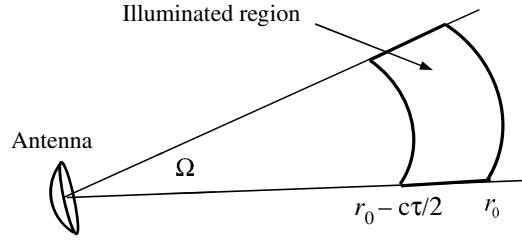


Figure 8.27 Region illuminated by pulsed radar.

introduce a change in polarization. For instance, reflections from airplane or ionosphere can change the wave's polarization. As a result, the received power at the radar will be lowered due to the mismatch of polarization. When the rain drops are spherical in shape and multiple scattering can be neglected, a circularly polarized incident wave is returned as a circularly polarized wave of the opposite sense and is not received by the radar antenna. Therefore, the back scattering of radar pulses by rain drops can be avoided by using circular polarization and this phenomenon may be used to reduce the clutter interference produced by rain.

The far-field of the radar antenna generated by a current distribution $\mathbf{J}$ in a homogeneous and isotropic medium can be expressed as

$$\mathbf{E}(\mathbf{r}) = -\frac{jk_0\eta_0 I}{4\pi r} e^{-jk_0 r} \mathbf{L}(\mathbf{u}_r). \tag{8.104}$$

Here I is the exciting current at the feeding plane, and $\mathbf{L}$ is the antenna vector effective length. The induced dipole moment in a single rain drop is given by

$$\mathbf{P} = 3\frac{\tilde{\varepsilon}_r - 1}{\tilde{\varepsilon}_r + 2}\varepsilon_0 \mathbf{E}. \tag{8.105}$$

The equivalent current is $j\omega\mathbf{P}$, which produces a field $\mathbf{E}_s$. The open-circuit voltage at the radar antenna-feeding plane induced by the incident field $\mathbf{E}_s$ generated by a single rain drop is (see Chapter 5)

$$dV_{\text{oc}}(\mathbf{u}_r) = -\frac{1}{I}\int_{V_0} \mathbf{E}_s(\mathbf{r}') \cdot \mathbf{J}(\mathbf{r}') dV(\mathbf{r}') = -\frac{1}{I}\int_{V_r} \mathbf{E}(\mathbf{r}') \cdot [j\omega\mathbf{P}(\mathbf{r}')]\, dV(\mathbf{r}'),$$

where V_r denotes the volume occupied by the rain drop, and $\mathbf{u}_r$ is the unit vector in the direction of the rain drop. In the above, the reciprocity

theorem has been used. For a small rain drop, this may be approximated by

$$dV_{\rm oc} = -\frac{1}{I}\mathbf{E}\cdot j\omega\mathbf{P}\frac{4}{3}\pi a^3 = -\frac{jk_0 4\pi a^3}{I\eta_0}\frac{\tilde{\varepsilon}_r-1}{\tilde{\varepsilon}_r+2}\mathbf{E}\cdot\mathbf{E}$$
$$= \frac{jI\eta_0 k_0^3}{4\pi r^2}\frac{\tilde{\varepsilon}_r-1}{\tilde{\varepsilon}_r+2}a^3 e^{-2jk_0 r}\mathbf{L}\cdot\mathbf{L}. \tag{8.106}$$

In spherical coordinates, we may write

$$\mathbf{L} = L_\theta\mathbf{u}_\theta + L_\varphi\mathbf{u}_\varphi,$$

where L_θ and L_φ are the components along θ and φ direction. For a circularly polarized antenna, we have $L_\varphi = \pm jL_\theta$. This leads to

$$\mathbf{L}\cdot\mathbf{L} = L_\theta^2 + L_\varphi^2 = 0.$$

This implies a complete mismatch in polarization between the antenna and the scattered field by the rain drop, and the received voltage is zero. The received open-circuit voltage from a volume of drops can be expressed as the sum of (8.106) overall drops

$$V_{\rm oc} = \frac{jI\eta_0 k_0^3}{4\pi}\frac{\tilde{\varepsilon}_r-1}{\tilde{\varepsilon}_r+2}\sum_i \frac{a_i^3}{r_i^2}e^{-j2k_0 r_i}(\mathbf{L}\cdot\mathbf{L})_i, \tag{8.107}$$

where the subscript i denotes the value of the corresponding parameter for the ith drop located at $(r_i, \theta_i, \varphi_i)$ at time t. When the antenna is conjugately matched to a load R_L, the received power at time t is

$$\frac{|V_{\rm oc}|^2}{8R_L} = \frac{|I|^2\eta_0^2 k_0^6}{128\pi^2 R_L}\left|\frac{\tilde{\varepsilon}_r-1}{\tilde{\varepsilon}_r+2}\right|^2\sum_i\sum_j \frac{a_i^3 a_j^3}{r_i^2 r_j^2}e^{-j2k_0(r_i-r_j)}(\mathbf{L}\cdot\mathbf{L})_i(\bar{\mathbf{L}}\cdot\bar{\mathbf{L}})_j. \tag{8.108}$$

The relative phase angle $\Delta_{ij} = 2k_0(r_i - r_j)$ changes with time t and may be considered uniformly distributed over the range $[0, \pi]$ with a probability density $1/2\pi$. The ensemble-average of $e^{-j\Delta_{ij}}$ is

$$\langle e^{-j\Delta_{ij}}\rangle = \int_0^{2\pi}\frac{e^{-j\Delta_{ij}}}{2\pi}d\Delta_{ij} = \begin{cases}0, & i\neq j\\ 1, & i=j\end{cases}.$$

Taking the ensemble-average of (8.108) gives the average received power as follows

$$P_{\rm rec} = \frac{\langle|V_{\rm oc}|^2\rangle}{8R_L} = \frac{|I|^2\eta_0^2 k_0^6}{128\pi^2 R_L}\left|\frac{\tilde{\varepsilon}_r-1}{\tilde{\varepsilon}_r+2}\right|^2\sum_i\frac{a_i^6}{r_i^4}|(\mathbf{L}\cdot\mathbf{L})_i|^2. \tag{8.109}$$

For a large number of rain drops with different radii, the above sum becomes an integral

$$P_{\text{rec}} = \frac{\langle |V_{\text{oc}}|^2 \rangle}{8R_L}$$
$$= \frac{|I|^2 \eta_0^2 k_0^6}{128\pi^2 R_L} \left| \frac{\tilde{\varepsilon}_r - 1}{\tilde{\varepsilon}_r + 2} \right|^2 \int_0^\infty da \int_V \frac{|\mathbf{L} \cdot \mathbf{L}|^2}{r^2} N(a) a^6 \sin\theta \, dr \, d\theta \, d\varphi, \quad (8.110)$$

where $N(a)da$ is the number of drops per unit volume, and V is the volume occupied by the rain drops.

8.4.3 Radar Range Equation

The power density at a distance R_1 (target position) from the radiating antenna of the radar is given by (Figure 8.21)

$$p(\mathbf{u}_t) = \frac{P_t G_t(\mathbf{u}_t)}{4\pi R_1^2}, \quad (8.111)$$

where P_t is the input power to the antenna terminal, $G_t(\mathbf{u}_t)$ is the gain of the transmitting antenna in the direction $\mathbf{u}_t$. When the target is illuminated by the incident wave from the transmitting antenna, it will reflect the incident wave in various directions. Power captured by the target that will be re-radiated and the scattered power density in the direction $\mathbf{u}_s$ may be expressed as the product of incident power density times the RCS divided by $4\pi R_2^2$

$$p_s(-\mathbf{u}_s) = \frac{P_t G_t(\mathbf{u}_t)}{4\pi R_1^2} \frac{\sigma(\mathbf{u}_t, \mathbf{u}_s)}{4\pi R_2^2}. \quad (8.112)$$

The power received by the antenna is

$$P_{\text{rec}} = p_s(-\mathbf{u}_s) A_e(-\mathbf{u}_s) = A_e(-\mathbf{u}_s) \frac{P_t G_t(\mathbf{u}_t)}{4\pi R_1^2} \frac{\sigma(\mathbf{u}_t, \mathbf{u}_s)}{4\pi R_2^2}. \quad (8.113)$$

Here A_e is the equivalent area of the receiving antenna, which can be expressed as (Chapter 5)

$$A_e(-\mathbf{u}_s) = \left| \frac{\mathbf{E}_s(\mathbf{u}_s) \cdot \mathbf{L}(-\mathbf{u}_s)}{Z + Z_L} \right|^2 \frac{\eta \text{Re} Z_L}{|\mathbf{E}_s(\mathbf{u}_s)|^2} = \frac{|\mathbf{L}(-\mathbf{u}_s)|^2 \eta \text{Re} Z_L}{|Z + Z_L|^2} |\mathbf{u}_{E_s} \cdot \mathbf{u}_L|^2,$$

where $\mathbf{u}_{E_s}$ and $\mathbf{u}_L$ are the polarization unit vector of the scattered field and the receiving antenna respectively. If the receiving antenna is conjugately

matched, the above equation reduces to

$$A_e(-\mathbf{u}_s) = \frac{\lambda^2}{4\pi}\frac{\pi\eta}{(R_{\text{rad}}+R_{\text{loss}})}\frac{|\mathbf{L}(-\mathbf{u}_s)|^2}{\lambda^2}\left|\mathbf{u}_{E_s}\cdot\mathbf{u}_L\right|^2$$
$$= \frac{\lambda^2}{4\pi}G_r(-\mathbf{u}_s)\left|\mathbf{u}_{E_s}\cdot\mathbf{u}_L\right|^2, \tag{8.114}$$

where $G_r(-\mathbf{u}_s)$ stands for the gain of the receiving antenna in $-\mathbf{u}_s$ direction. Introducing this into (8.113), we obtain

$$\frac{P_{\text{rec}}}{P_t} = \frac{1}{(4\pi)^3}\frac{\lambda^2\sigma(\mathbf{u}_t,\mathbf{u}_s)}{R_1^2R_2^2}G_t(\mathbf{u}_t)G_r(-\mathbf{u}_s)\left|\mathbf{u}_{E_s}\cdot\mathbf{u}_L\right|^2. \tag{8.115}$$

This is called **radar range equation**, a relationship between radar range, transmitted power, received power, antenna gain, and the target's RCS. For polarization-matched receiving antenna, (8.115) reduces to

$$\frac{P_{\text{rec}}}{P_t} = \frac{1}{(4\pi)^3}\frac{\lambda^2\sigma(\mathbf{u}_t,\mathbf{u}_s)}{R_1^2R_2^2}G_t(\mathbf{u}_t)G_r(-\mathbf{u}_s). \tag{8.116}$$

For a monostatic system, (8.116) may be simplified as below

$$\frac{P_{\text{rec}}}{P_t} = \frac{G_t^2}{(4\pi)^3}\frac{\lambda^2\sigma}{R^4}. \tag{8.117}$$

This relation can be used to measure the RCS.

> If we consider what science already has enabled men to know-the immensity of space, the fantastic philosophy of the stars, the infinite smallness of the composition of atoms, the macrocosm whereby we succeed only in creating outlines and translating a measure into numbers without our minds being able to form any concrete idea of it-we remain astounded by the enormous machinery of the universe.
>
> —Guglielmo Marconi

Bibliography

Abraham, R., J. E. Marsden and T. Ratiu, *Manifolds, Tensor Analysis, and Applications*, Springer-Verlag, 1988.

Adler, R. B., L. J. Chu and R. M. Fano, *Electromagnetic Energy Transmission and Radiation*, John Wiley & Sons Inc., 1960.

Aharoni, J., *Antennas*, Oxford, Clarendon Press, 1946.

Ahner, J. F. and R. E. Kleimann, "The exterior Neumann problem for the Helmholtz equation", *Arch. Rational Mech. Anal.*, Vol. 52, 26–43, 1973.

Albert, G. E. and J. L. Synge, "The general problem of antenna radiation and the fundamental integral equation with application to an antenna of revolution-Part 1", *Quart. Appl. Math.*, Vol. 6, 117–131, April 1948.

Albertsen, N. C., J. E. Hansen and N. E. Jensen, "Computation of radiation from wire antennas on conducting bodies", *IEEE Trans. Antennas and Propagat.* Vol. AP-22, 200–206, No. 2, Mar. 1974.

Alexandrov, O. and G. Ciraolo, "Wave propagation in a 3-D optical waveguide", *Mathematical Models and Methods in Applied Sciences*, Vol. 14, No. 6, 819–852, 2004.

Ancona, C., "On small antenna impedance in weakly dissipative media", *IEEE Trans. Antennas and Propagat.*, Vol. AP-26, 341–343, Mar. 1978.

Angell, T. S. and A. Kirsh, *Optimization Method in Electromagnetic Radiation*, Springer, 2004.

Ash, R. B., *Information Theory*, John Wiley & Sons, 1965.

Aydin, K. and A. Hizal, "On the completeness of the spherical vector wave functions", *J. Math. Anal. & Appl.*, Vol. 117, 428–440, 1986.

Bahl, I., *Lumped Elements for RF and Microwave Circuits*, Artech House, 2003.

Bahl, I. and P. Bhartia, *Microwave Solid State Circuit Design*, 2nd Edition, John Wiley & Sons, 2003.

Bahl, I. and D. K. Trivedi, "A designer's guide to micorstrip line", *Microwaves*, 174–182, May 1977.

Balanis, C. A., *Antenna Theory: Analysis and Design*, 2nd Edition, John Wiley & Sons, 2005.

Bamberger, A. and A. S. Bonnet, "Mathematical analysis of the guided modes of an optical fiber", *SIAM J. Math. Anal.*, Vol. 21, No. 6, 1487–1510, Nov. 1990.

Barut, A. O., *Electromagnetics and Classical Theory of Fields and Particles*, Macmillan, New York, 1964.

Baum, C. E., "Emerging technology for transient and broad-band analysis and synthesis of antennas and scatters", *Proc. IEEE*, Vol. 64, 1598–1616, 1976.

Baum, C. E., E. J. Rothwell, K. M. Chen and D. P. Nyquist, "The singularity expansion method and its application to target identification", *Proc. IEEE*, Vol. 79, No. 10, 1481–1492, Oct. 1991.

Benford, J., J. A. Swegle and E. Schamiloglu, *High Power Microwaves*, 2nd Edition, Taylor & Francis, 2007.

Bertoni, H. L., *Radio Propagation for Modern Wireless Systems*, Prentice Hall, 1999.

Best, A. C., "Empirical formulae for the terminal velocity of water drops fall in through the atmosphere", *Quart. J. Met. Soc.*, Vol. 76, 302–311, 1950.

Bethe, H. A., "Theory of Diffraction by Small Holes", *Physical Review*, Vol. 66, Nos. 7–8, 163–182, 1944.

Bhat, B. and S. K. Koul, *Stripline-like Transmission Lines for Microwave Integrated Circuits*, New Age International, 1989.

Bladel, J. V., "Small holes in a waveguide", *Proc. IEE*, Vol. 118, No. 1, 43–50, Jan. 1971.

Bladel, J. V., *Electromagnetic Fields*, IEEE Press, 2007.

Bluck, M. J., M. D. Pocock and S. P. Walker, "An accurate method for the calculation of singular integrals arising in time-domain integral equation analysis of electromagnetic scattering", *IEEE Trans. Antennas and Propagat.* Vol. AP-45, 1793–1798, No. 12, Dec. 1997.

Bogosanovic, M. and A. G. Williamson, "Antenna array with beam focused in near-field zone," *Electron. Lett.*, Vol. 39, pp. 704–705, May 2003.

Bogosanovic, M. and A. G. Williamson, "Microstrip antenna array with a beam focused in the near-field zone for application in noncontact microwave industrial inspection," *IEEE Trans. Instrument. Meas.*, Vol. 56, 2186–2195, Dec. 2007.

Booker, H. G., "Slot aerials and their relation to complementary wire aerials", *J. IEE* (London) Pt IIIA, Vol. 93, No. 4, 620–626, 1946.

Bondeson, A., T. Rylander and P. Ingelström, *Computational Electromagnetics*, Springer, 2005.

Borgiotti, G. V., "Maximum power transfer between two planar apertures in the Fresnel zone", *IEEE Trans. Antennas Propagat.*, Vol. AP-14, 158–163, Mar. 1966.

Borgiotti, G. V., "On the reactive energy of an antenna", *IEEE Trans. Antennas and Propagat.* Vol. AP-15, 565–566, 1967.

Born, M. and E. Wolf, *Principles of Optics*, 6th Edition, Pergamon Press, 1980.

Boxleitner, N., *Electrostatic Discharge and Electronic Equipment: A Practical Guide for Designing to Prevent ESD Problems*, IEEE Press, New York, 1989.

Boyd, G. D. and H. Kogelnik, "Generalized confocal resonator theory", *Bell Sys. Tech. J.*, Vol. 41, 1347–1369, 1962.

Boyd, G. D. and J. P. Gordon, "Confocal multimode resonator for millimeter through optical wavelength masers", *Bell Sys. Tech. J.*, Vol. 40, 489–508, 1961.

Brau, C. A., *Modern Problems in Classical Electrodynamics*, Oxford University Press, 2004.

Brekhovskikh, L. M., *Waves in Layered Media*, Academic Press, 1960.

Brezis, H. and F. Browder, "Partial differential equations in the 20th century", *Advances in Mathematics*, Vol. 135, 76–144, 1998.

Brillouin, L., *Wave Propagation and Group Velocity*, Academic Press, New York, 1960.

Brinkman, W. F. and D. V. Lang, "Physics and the communication industry", *Reviews of Modern Physics*, Vol. 71, No. 2, 480–488, 1999.

Brown, W. C., "The history of power transmission by radio waves", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-32, 1230–1242, Sept. 1984.

Bucci, O. and G. Di Massa, "Open resonator powered by rectangular waveguide", *IEE Proceedings-H*, Vol. 139, 323–329, 1992.

Buffi, A., P. Nepa and G. Manara, "Design criteria for near-field-focused planar arrays", *IEEE Antennas Propag. Mag.*, Vol. 54, 40–50, Feb. 2012.

Burton, A. J. and G. F. Miller, "The application of integral equation methods to the numerical solution of some exterior boundary value problems", *Proc. Roy. Soc. Lond. A*, Vol. 323, 201–210, 1971.

Butler, C. M., Y. Rahmat-Samii and R. Mittra, "Electromagnetic penetration through apertures in conducting surfaces", *IEEE Trans. Electromagn. Comp.*, Vol. EMC-20, 82–93, Feb. 1978.

Byron, F. W. and R. W. Fuller, *Mathematics of Classical and Quantum Physics*, Addison-Wesley, 1969.

Calderon, A. P., "The multiple expansion of radiation fields", *J. Rational Mech. Anal.*, Vol. 3, 523–537, 1954.

Carin, L. and L. B. Felsen (Eds.), *Ultra-Wideband Short-Pulse Electromagnetics*, New York: Plenum, 1995.

Carson, J. R., "A generalization of reciprocity theorem", *Bell Syst. Tech. J.*, Vol. 3, 393, 1924.

Carter, P. S., "Circuit relations in radiating systems and applications to antenna problems", *Proc. IRE*, Vol. 20, No. 6, 1004–1041, June 1932.

Celozzi, S., R. Araneo and G. Lovat, *Electromagnetic Shielding*, John Wiley & Sons, 2008.

Chen, C. A. and D. K. Cheng, "Optimum element lengths for Yagi-Uda arrays", *IEEE Trans. Antennas Propag.*, Vol. AP-23, 8–15, Jan. 1975.

Chang, K. (Ed.), *Encyclopedia of RF and Microwave Engineering*, John Wiley & Sons, 2005.

Cheng, D. K. and C. A. Chen, "Optimum element spacings for Yagi-Uda arrays," *IEEE Trans. Antennas Propag.*, Vol. AP-21, 615–623, Sept. 1973.

Chew, W. C., *Waves and Fields in Inhomogeneous Media*, Van Nostrand Reinhold, 1990.

Chew, W. C., *Fast and Efficient Algorithms in Computational Electromagentics*, Artech House, 2001.

Christopoulos, C., *Principles and Techniques of Electromagnetic Compatibility*, CRC Press, Ann Arbor, 1995.

Chu, L. J., "Physical limitations of omni-directional antennas", *J. Appl. Phys.*, Vol. 19, 1163–1175, 1948.

Clark, P. O., "A self consistent field analysis of spherical-mirror Fabry Perot resonators", *Proc. IEEE*, Vol. 53, No. 1, 36–41, 1964.

Cochran, J. A., *The Analysis of Linear Integral Equations*, McGraw-Hill, New York, 1972.

Cohn, S. B., "Determination of aperture parameters by electrolytic tank measurements", *Proc. IRE*, Vol. 39, 1416–1421, Nov. 1951.

Cohn, S. B., "Parallel-coupled transmission-line resonator filters", *IRE Transactions: Microwave Theory and Techniques*, Vol. MTT-6, 223–231, April 1958.

Cohn, S. B., "Microwave bandpass filters containing high-Q dielectric resonators", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-16, 218–227, April 1968.

Coleman, C., *An Introduction to Radio Frequency Engineering*, Cambridge University Press, 2004.

Collardey, S., A. Sharaiha and K. Mahdjoubi, "Evaluation of antenna radiation Q using FDTD method", *Electronics Letters*, Vol. 41, No. 12, 675–677, 9th June 2005.

Collardey, S., A. Sharaiha and K. Mahdjoubi, "Calculation of small antennas quality factor using FDTD method", *IEEE Antennas and Wireless Propagation Letters*, Vol. 5, No. 1, 191–194, 2006.

Collin, R. E., "Stored energy Q and frequency sensitivity of planar aperture antennas", *IEEE Trans. Antennas and Propagat.*, Vol. AP-15, 567–568, 1967.

Collin, R. E. and F. J. Zucker, *Antenna Theory*, McGraw-Hill, New York, 1969.

Collin, R. E., *Antennas and Radio Wave Propagation*, McGraw-Hill, New York, 1985.

Collin, R. E., *Field Theory of Guided Waves*, IEEE Press, 1991.

Collin, R. E., "Minimum Q of small antennas", *Journal of Electromagnetic Waves and Applications*, Vol. 12, 1369–1393, 1998.

Collin, R. E., *Foundations for Microwave Engineering*, 2nd Edition, IEEE Press, 2001.

Collin, R. E. and S. Rothschild, "Evaluation of antenna Q", *IEEE Trans. Antennas and Propagat.*, Vol. AP-12, 23–27, Jan. 1964.

Collin, R. E., "Limitations of the Thévenin and Norton equivalent circuits for a receiving antenna", *IEEE Antennas and Propagat. Magazine*, Vol. 45, No. 2, 119–124, 2003.

Collin, R. E., "Generalization of a fundamental theorem for the scattering from a receiving antenna", *AP-S/URSI Symposium*, 559, 2003.

Colton, D. and R. E. Kleimann, "The direct and inverse scattering problems for an arbitrary cylinder: Dirichlet boundary conditions", *Proc. Royal Soc., Edinburgh*, 86A, 29–42, 1980.

Colton, D. and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley & Sons, 1983.

Colton, D. and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag, 1998.

Correia, L. M., "A comparison of integral equations with unique solution in the resonance region for scattering by conduction bodies", *IEEE Trans. Antennas and Propagat.*, Vols. AP-41, 52–58, Jan. 1993.

Costabel, M., "A coercive bilinear form for Maxwell equations", *J. Math. Anal. & Appl.*, 157, 527–541, 1991.

Counter, V. A., "Miniature cavity antennas," Rep. No. 2, Contract No. W28-099-ac-382, Microwave Lab., Stanford University, June 30, 1948.

Courant, R. and D. Hilbert, *Methods of Mathematical Physics*, Vols. 1–2, John Wiley & Sons, 1953.

Cover, T. M. and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.

Cristal, E. G. and S. Frankel, "Hairpin line/half-wave parallel-coupled-line filters", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-20, 719–728, November 1972.

Deschamps, G. A., "Impedance of an antenna in a conducting medium", *IRE Trans. Antennas Propagat.*, Vol. Ap-10, 648–649, Sept. 1962.

Dhia, S. B., M. Ramdani and E. Sicard (Eds.), *Electromagnetic Compatibility of Integrated Circuits: Techniques for low emission and susceptibility*, Springer, 2006.

Dicke, R. H., "Molecular amplification and generation systems and methods", U.S. Patent 2,851,652, Sept. 9, 1958.

Dolph, C. L. and S. K. Cho, "On the relationship between the singularity expansion method and the mathematical theory of scattering", *IEEE Trans. Antennas and Propagat.*, Vol. AP-28, No. 6, 888–897, Nov. 1980.

Dudley, D. G., *Mathematical Foundations for Electromagnetic Theory*, IEEE Press, 1994.

DuHamel, R. H. and D. E. Isbell, Broadband logarithmically periodic antenna structures, *IRE National Convention Record*, Pt. 1, 119–128, 1957.

Dydyk, M., "Master the T-Junction and Sharpen Your MIC Designs", *Microwaves*, 184–186, May 1977.

Edwards, M. L. and J. H. Sinksy, "A new criterion for linear 2-port stability using a single geometrically derived parameter", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-40, 2803–2811, Dec. 1992.

Eisenhart, L. P., "Separable systems of Stäckel", *Ann. Math.*, Vol. 35, No. 2, 284–305, 1934.

Elliott, R. S., *Electromagnetics-History, Theory and Applications*, IEEE Press, 1993.

Elliott, R. S., *An Introduction to Guided Waves and Microwave Circuits*, Prentice-Hall, New York, 1993.

Elliott, R. S., *Antenna Theory and Design*, Prentice-Hall, New York, 1981.

Erdelyi, A. (Ed.), *Tables of Integral Transform*, Bateman Manuscript Project, Vol. 1, McGraw-Hill, 1954.

Fabry, C. and A. Perot, "Theorie et Applications d'une Nouvélle Method de Spectroscopie Interférentielle", *Ann Chim. Phys.*, Vol. 16, 115–146, 1899.

Fano, R. M., L. J. Chu and R. B. Adler, *Electromagnetic Fields, Energy, and Forces*, John Wiley & Sons, New York and MIT Press, Cambridge, MA, 1960.

Fante, R. L., "Quality factor of general idea antennas", *IEEE Trans. Antennas and Propagat.*, Vol. AP-17, 151–155, 1969.

Fante, R. L., "Maximum possible gain for an arbitrary ideal antenna with specified quality factor", *IEEE Trans. Antennas and Propagat.*, Vol. AP-40, 1586–1588, Dec. 1992.

Farago, P. S., *An Introduction to Linear Network Analysis*, English Universities Press, 1961.

Felsen, L. B. (Ed.), *Transient Electromagnetic Fields*, Springer-Verlag, 1976.

Felsen, L. B. and N. Marcuwitz, *Radiation and Scattering of Electromagnetic Waves*, Prentice Hall, Englewood Cliffs, New Jersey, 1973.

Fooks, E. H. and R. A. Zakarevicius, *Microwave Engineering using Microstrip Circuits*, Prentice Hall, 1990.

Ford, P. J. and G. A. Saunders, *The Rise of the Superconductors*, CRC Press, 2005.

Foschini, G. J. and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas", *Wireless Personal Communications*, Vol. 40, No. 6, 311–335, 1998.

Fox, A. G. and T. Li, "Resonant modes in a maser interferometer", *Bell Syst. Tech. J.*, Vol. 40, No. 2, 453–488, Mar. 1961.

Fox, A. G. and T. Li, "Modes in a maser interferometer with curved and tilted mirrors", *Proc. IEEE*, Vol. 51, 80–89, January 1963.

Franceschetti, G. and C. H. Papas, "Pulsed antennas", *IEEE Trans. Antennas and Propagat.*, Vol. AP-22, 651–661, Sept. 1974.

Frankl, D. R., *Electromagnetic Theory*, Prentice Hall, New York, 1986.

Frey, J., *Microwave Integrated Circuits*, Artech House, 1975.

Friedman, B., *Principles and Techniques of Applied Mathematics*, John Wiley & Sons, Inc., 1956.

Friis, H. T., "A note on a simple transmission formula", *Proc. IRE*, Vol. 34, 254–256, 1946.

Fujimoto, K. and J. R. James, *Mobile Antenna Systems Handbook*, Artech House, 2001.

Fujimoto, K. and H. Morishita, *Modern Small Antennas*, Cambridge University Press, 2014.

Fusco, V. F., *Microwave Circuits*, Prentice Hall, 1987.

Gallager, R. G., *Information Theory and Reliable Communication*, John Wiley & Sons, 1968.

Gandhi, O. P., *Microwave Engineering and Applications*, Pergamon Press, 1981.

Garg, R. and I. J. Bahl, "Microstrip Discontinuities," *Int. J. Electronics*, Vol. 45, No. 1, 81–87, 1978.

Geyi, W. and W. Hongshi, "Solution of the resonant frequencies of cavity resonator by boundary element method", *IEE Proc., Microwaves, Antennas and Propagation*, Vol. 135, Pt.H, No. 6, 361–365, 1988a.

Geyi, W. and W. Hongshi, "Solution of the resonant frequencies of a microwave dielectric resonator using boundary element method", *IEE Proc., Microwaves, Antennas and Propagation*, Vol. 135, Pt.H, No. 5, 333–338, 1988b.

Geyi, W., L. Xueguan and W. Wanchun, "Solution of the characteristic impedance of an arbitrary shaped TEM transmission line using complex variable boundary element method", *IEE Proc., Microwaves, Antennas and Propagation*, Vol. 136, Pt. H, No. 1, 73–75, 1989.

Geyi, W., "On the spurious solutions in boundary integral formulation for waveguide eigenvalue problems", *Proc. of European Microwave Conference*, Vol. 2, 1311–1316, 1990a.

Geyi, W., "Numerical solution of the transmission line problems by a network model decomposition method based on polygon discretization", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-38, No. 8, 1086–1091, 1990b.

Geyi, W., "Numerical analysis of waveguide discontinuity problems by using network model decomposition method", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-39, No. 10, 1766–1770, 1991.

Geyi, W., Y. Chengli and L. Weigan, "Unified theory of the backscattering of electromagnetic missiles by a perfectly conducting target", *J. Appl. Phys.*, Vol. 71, 3103–3106, Apr. 1992.

Geyi, W., "Neumann series solutions for low frequency electromagnetic scattering problems", *Chinese Journal of Electronics* (English version), Vol. 4, No. 3, 89–92, 1995.

Geyi, W., "Further research on the behavior of energy density of electromagnetic pulse", *Microwave and Optical Technology Letters*, Vol. 9, 331–335, Aug. 20, 1996a.

Geyi, W., "Enhancement of backscattering by electromagnetic focusing," *Journal of UEST of China*, Vol. 25, 177–184, Aug. 1996b.

Geyi, W., "Theoretical study of microwave power transmission," *Journal of Electronics*, Vol. 20, 538–545, July 1998 (in Chinese).

Geyi, W., *Advances in Electromagnetic Theory*, National Defense Publishing House of China, 1999 (in Chinese).

Geyi, W., P. Jarmuszewski and Y. Qi, "Foster reactance theorems for antennas and radiation Q", *IEEE Trans. Antennas and Propagat.*, Vol. AP-48, 401–408, Mar. 2000.

Geyi, W., "Physical Limitations of Antenna", *IEEE Trans., Antennas and Propagat.*, Vol. AP-51, 2116–2123, Aug. 2003a.

Geyi, W., "A Method for the Evaluation of Small Antenna Q", *IEEE Trans. Antennas and Propagat.*, Vol. AP-51, 2124–2129, Aug. 2003b.

Geyi, W., "Derivation of equivalent circuits for receiving antenna", *IEEE Trans. Antennas and Propagat.*, Vol. AP-52, 1620–1624, June 2004.

Geyi, W., "A time-domain theory of waveguide", *Progress in Electromagnetics Research*, PIER 59, 267–297, 2006a.

Geyi, W., "New magnetic field integral equation for antenna system", *Progress in Electromagnetics Research*, PIER 63, 153–176, 2006b.

Geyi, W., "Reply to comments on 'The Foster reactance theorem for antennas and radiation Q'", *IEEE Trans. Antennas and Propagat.*, Vol. AP-55, 1014–1016, 2007a.

Geyi, W., "Multi-antenna information theory", *Progress in Electromagnetics Research*, PIER 75, 11–50, 2007b.

Geyi, W., "Time-domain theory of metal cavity resonator", *Progress in Electromagnetics Research*, PIER 78, 219–253, 2008.

Geyi, W., Q. Rao, S. Ali and D. Wang, Handset antenna design: Practice and theory, *Progress in Electromagnetics Research*, PIER 80, 123–160, 2008b.

Geyi, W., *Foundations of Applied Electrodynamics*, New York: Wiley, 2010.

Geyi, W., "A New Derivation of the upper bounds for the ratio of gain to Q", *IEEE Trans. Antennas and Propagat.*, Vol. AP-60, No. 7, 1916–1922, July 2012.

Geyi, W., "Optimization of the ratio of gain to Q", *IEEE Trans. Antennas and Propagat.*, Vol. AP-61, No. 4, 3488–3490, April 2013.

Geyi, W., "Optimal design of antenna arrays (Invited)," *International Workshop on Antenna Technology*, Sydney, March 2014.

Ginzburg, V. L., *The Propagation of Electromagnetic Waves in Plasmas*, Oxford: Pergamon Press, 1964.

Glisson, A. W. and D. R. Wilton, "Simple and efficient numerical methods for problems of electromagnetic radiation and scattering from surfaces", *IEEE Trans. Antennas and Propagat.*, Vol. AP-28, No. 5, 593–603, Sept. 1980.

Golio, J. M., *The RF and Microwave Handbook*, CRC Press, 2001.

Gonschorek, K.-H. and R. Vick, *Electromagnetic Compatibility for Device Design and System Integration*, Springer, 2009.

Good, R. H., "Particle aspect of the electromagnetic field equations", *Phys. Rev.*, Vol. 105, No. 6, 1914–1919, 1957.

Gosling, W., *Radio Antennas and Propagation*, Newnes, 1998.

Goubao, G. and F. Schwering, "On the guided propagation of electromagnetic wave beams", *IRE Trans. Antennas and Propagat.*, Vol. AP-9, 248–256, May 1961.

Goubau, G. (Ed.), *Electromagnetic Waveguides and Cavities*, London: Pergmon, 1961.

Gradsheyn, L. S. and I. M. Ryzhik, *Tables of Integrals, Series, and Products*, Academic Press, 1994.

Graglia, R. D., "On the numerical integration of the linear shape functions times the 3-D Green's function or its gradient on a plane triangle", *IEEE Trans. Antennas and Propagat.* Vol. AP-41, 1448–1455, Oct. 1993.

Graham, W. J., "Analysis and synthesis of axial field patterns of focused apertures", *IEEE Trans. Antennas Propag.*, Vol. 31, 665–668, July 1962.

Green, R. B., "The general theory of antenna scattering," Antenna Lab., Ohio State Research Foundation, Rept., 1223–17, Nov. 30, 1963.

Grieg, D. D. and H. F. Engelmann, "Microstrip-A new transmission technique for the kilomegacycle range", *Proceedings of the IRE*, Vol. 40(12), 1644–1650, Dec. 1952.

Griffiths, D. J., *Introduction to Electrodynamics*, Prentice Hall, 1999.

Gupta, K. C., R. Garg and I. Bahl, *Microstrip Lines and Slotlines*, Artech House, 1979.

Gupta, K. C., R. Garg and R. Chadha, *Computer-aided Design of Microwave Circuits*, Artech House, 1981.

Gustafsson, M. and S. Nordebo, "Optimal antenna currents for Q, superdirectivity, and radiation patterns using convex optimization," *IEEE Trans. Antennas Propagat.*, Vol. AP-61, 1109–1118, Mar. 2013.

Hallén, E., "Theoretical investigations into transmitting and receiving qualities of antennae", *Nova Acta Regial Soc. Sci. Upsaliensis*, Ser. IV, Vol. 2, No. 4, 1–44, Nov. 1938.

Hammond, P., *Energy Methods in Electromagnetism*, Clarendon Press Oxford, 1981.

Hammerstad, E. and O. Jensen, "Accurate models for microstrip computer-aided design," *IEEE MTT-S Digest*, Vol. 80, 407–409, May 1980.

Hammerstad, E., "Computer-aided design of microstrip couplers with accurate discontinuity models", *IEEE MTT-S Digest*, Vol. 81, 54–56, June 1981.

Hanson, G. W. and A. B. Yakovlev, *Operator Theory for Electromagnetics*: *An Introduction*, Springer, 2002.

Hansen, R. C., "Fundamental limitations in antennas", *Proc. IEEE*, Vol. 69, 170–182, Feb. 1981.

Hansen, R.C., *Electrically Small, Superdirective, and Superconducting Antennas*, Wiley & Sons, 2006.

Hansen, W. W., "A new type of expansion in radiation problems", *Phys. Rev.*, Vol. 47, 139–143, 1935.

Harrington, R. F., "On the gain and beamwidth of directional antennas", *IRE Trans. on Antennas and Propagat.* Vol. 6, 219–225, 1958.

Harrington, R. F., "Effect of antenna size on gain, bandwidth, and efficiency", *Journal of Research of the National Bureau of Standards-D. Radio Propagation*, Vol. 64D, No. 1, Jan.–Feb. 1960.

Harrington, R. F., *Time-Harmonic Electromagnetic Fields*, McGraw-Hill Book Company, Inc, 1961.

Harrington, R. F., *Field Computation by Moment Methods*, MacMillan, 1968.

Harrington, R. F. and J. R. Mautz, "A generalized formulation for aperture problems", *IEEE Trans. Antennas and Propagat.*, Vol. AP-24, 870–873, Nov. 1976.

Harrington, R. F. and A. T. Villeneuve, "Reciprocal relationships for gyrotropic media", *IRE Trans. Microwave Theory and Techniques*, Vol. MTT-6, 308–310, July 1958.

Hartemann, F. V., *High Field Electrodynamics*, CRC Press, 2002.

Hata, M., "Empirical formula for propagation loss in land mobile radio service", *IEEE Trans. on Veh. Tech.*, VT-29, Vol. 3, 317–325, Aug. 1980.

Hazard, C. and M. Lenoir, "On the solution of time-harmonic scattering problems for Maxwell equations", *SIAM J. Math. Anal.*, Vol. 27, No. 6, 1597–1630, 1996.

Heras, J. A., "How the potentials in different gauges yield the same retarded electric and magnetic field", *Am. J. Phys.*, Vol. 75, No. 2, 176–183, 2007.

Heurtley, J. C., "Maximum power transfer between two finite antennas", *IEEE Trans. Antennas Propagat.*, Vol. Ap-15, 298–300, Mar. 1967.

Hoffmann, R. K., *Handbook of Microwave Integrated Circuits*, Artech House, 1987.

Holzman, E., *Essentials of RF and Microwave Grounding*, Artech House, 2006.

Hondros, D., "Ueber elektromagnetische Drahtwelle," *Annalen der Physik*, Vol. 30, 905–949, 1909.

Hong, J.-S. and M. J. Lancaster, *Microstrip Filters for RF/Microwave Applications*, John Wiley & Sons, 2001.

Hooft, G., "A confrontation with infinity", *Reviews of Modern Physics*, Vol. 72, No. 2, 333–339, April 2000.

Hoop, T. A., *Handbook of Radiation and Scattering of Waves: Acoustic Waves in Fluids, Elastic Waves in Solids, Electromagnetic Waves*, Academic Press, 1995.

Horn, R. A. and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

Howard J. and W. C. Lin, "Simple Rules Guide Design of Wideband Stripline Couplers", *Microwaves and RF*, Vol. 27, No. 5, 201–211, May 1988.

Hsiao, G. C. and R. E. Kleinmann, "Mathematical foundations for error estimation in numerical solutions of integral equations in electromagnetics", *IEEE Trans. Antennas and Propagat.* Vol. AP-45, 316–328, Mar. 1997.

Hu, M. K., "Near zone power transmission formula", *IRE Nat'l Conv. Rec.*, Part 8, 128–135, 1958.

Huang, K., "On the interaction between the radiation field and ionic crystals", *Proc. Roy. Soc.* (*London*) *A*, Vol. 208, 352–365, Sept. 1951.

Huurdeman, A. A., *The Worldwide History of Telecommunications*, Wiley-IEEE, 2003.

Huygens, C., *Treatise on Light*, Dover Publications INC, New York, 1962; first published in 1690.

Idemen, M., "The Maxwell equations in the sense of distribution", *IEEE Trans. Antennas and Propagat.*, Vol. AP-21, 736–738, Jul. 1973.

Iizsuka, K., R. King and C. Harrison, Jr., "Self and mutual admittances of two identical circular loop antennas", *IEEE Trans. Antennas and Propagat.*, Vol. AP-14, No. 4, 440–450, July 1966.

Ito, M., "Dispersion of very short microwave pulses in waveguide", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-13, 357–364, May 1965.

Itoh, T. and R. Mittra, "Spectral-domain approach for calculating the dispersion characteristics of microstrip lines", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-21, No. 7, 496–499, July 1973.

Jackson, J. D., *Classical Electrodynamics*, 3rd Edition, John Wiley & Sons, New York, 1999.

Jakes, W. C., *Microwave Mobile Communications*, IEEE Press, 1994.

James J. R. and P. S. Hall, *Handbook of Microstrip Antennas*, INSPEC, Inc., 1988.

Jarry, P. and J. Beneat, *Design and Realizations of Miniaturized Fractal Microwave and RF Filters*, John Wiley and Sons, 2009.

Järvenpää, S., M. Taskinen and P. Ylä-Oijala, "Singularity extraction technique for integral equation methods with higher order basis functions on plane triangles and tetrahedral", *Int. J. Numer. Meth. Engng.*, Vol. 58, 1149–1165, 2003.

Jin, J. M., *The Finite Element Method in Electromagnetics*, Wiley, 2002.

Johnson, J., "Thermal agitation of electricity in conductors", *Phys. Rev.*, Vol. 32, 97–109, 1928.

Jones, D. S., *The Theory of Electromagnetism*, Pergamon Press, 1964.

Jones, D. S., *Methods in Electromagnetic Wave Propagation*, Clarendon Press, Oxford, 1979a.

Jones, D. S., *Elementary Information Theory*, Clarendon Press, 1979b.

Joss, J., J. C. Thams and A. Waldvogel, "The variation of raindrop size distribution at Locarno", *Proc. of the International Conference on Cloud Physics*, Toronto, 1968.

Kahn, W. K. and H. Kurss, "Minimum-scattering antennas", *IEEE Trans. Antennas and Propagat.*, Vol. AP-13, 671–675, Sep. 1965.

Kajfez, D. and P. Guillon, *Dielectric Resonators*, Artech House, 1986.

Kajfez, D., *Q Factor Measurements Using MATLAB*, Artech House, 2011.

Kalafus, R. M., "On the evaluation of antenna quality factors", *IEEE Trans. Antennas and Propagat.*, Vol. AP-17, 729–732, 1969.

Kang, E. W., *Radar System Analysis, Design, and Simulation*, Artech House, 2008.

Karimkashi, S. and A. A. Kishk, "Focused microstrip array antenna using a Dolph–Chebyshev near-field design", *IEEE Trans. Antennas Propagat.*, Vol. AP-57, 3813–3820, Dec. 2009.

Kartchevski, E. M. *et al.*, "Mathematical analysis of the generalized natural modes of an inhomogeneous optical fiber", *SIAM J. Appl. Math.*, Vol. 65, 2033–2048, 2005.

Kay, A. F., "Near field gain of aperture antennas", *IRE Trans. Antennas and Propagat.*, Vol. AP-8, 586-593, Nov. 1960.

Kellogg, O. D., *Foundations of Potential Theory*, New York: Dover, 1953.

Kerr, D. E. (Ed.), *Propagation of Short Radio Waves*, McGraw-Hill, 1951.

Kerr, D. E. and P. J. Rubenstein, "Introduction to Microwave Propagation", *RL Report No. 406*, Sept. 16, 1943.

Kinayman, N. and M. I. Aksun, *Modern Microwave Circuits*, Artech House, 2005.

King, R. W. P., *The Theory of Linear Antennas*, Cambridge, MA: Harvard University Press, 1956.

King, R. W. P. and C. W. Harrison Jr., "The receiving antenna," *Proc. IRE*, 32, 18–34, Jan. 1944.

King, R. W. P., J. C. W. Harrison and D. H. Denton, Transmission line missile antennas, *IRE Trans. on Antennas and Propagat.*, Vol. 8, No. 1, 88–90, 1960.

Kleimann, R. E., "Iterative solutions of boundary value problems", in *Function Theoretic Methods for Partial Differential Equations*, Springer, New York, 1976.

Kleimann, R. E., "Low frequency electromagnetic scattering", in *Electromagnetic Scattering*, (Ed. P. L.E. Uslenghi), 1978.

Kleimann, R. E. and W. Wendland, "On the Neumann's method for the exterior Neumann problem for the Helmholtz equation", *J. Math. Anal. & Appl.*, Vol. 57, 170–202, 1977.

Klein, C. and R. Mittra, "Stability of matrix equations arising in electromagnetics", *IEEE Trans. Antennas and Propagat.*, Vol. AP-21, 902–905, No. 6, Nov. 1973.

Kline, M. and I. W. Kay, *Electromagnetic Theory and Geometrical Optics*, Interscience, 1965.

Knepp, D. L. and J. Goldhirsh, "Numerical analysis of electromagnetic radiation properties of smooth conducting bodies of arbitrary shape", *IEEE Trans. Antennas and Propagat.*, Vol. AP-20, 383–388, No. 3, May 1972.

Kneppo, I., *Microwave Integrated Circuits*, Springer, 1994.

Kodali V. P., *Engineering Electromagnetic Compatibility*, IEEE Press, New York, 1991.

Kogelnik, H. and T. Li, "Laser beam and resonator", *Proc. IEEE*, Vol. 54, 1312–1329, 1966.

Kolner, B. H., "Space-time duality and the theory of temporal imaging", *IEEE J. Quantum Electron.*, Vol. 30, 1951–1963, Aug. 1994.

Kong, J. A., *Electromagnetic Wave Theory*, New York: Wiley-Interscience, 1990.

Konishi, Y., *Microwave Integrated Circuits*, CRC Press, 1991.

Kovetz, A., *Electromagnetic Theory*, Oxford University Press, 2000.

Kraus, J. D., *Antennas*, 2nd Edition, McGraw-Hill, New York, 1988.

Kraus, J. D., *Electromagnetics*, McGraw-Hill, New York, 1984.

Kraus, J. D. and D. A. Fleisch, *Electromagnetics with Applications*, McGraw-Hill, New York, 1999.

Kreyszig, E., *Introductory Functional Analysis with Applications*, John Wiley & Sons, 1978.

Kristensson, G., "Transient electromagnetic wave propagation in waveguides", *J. Electromagnetic Waves and Applications*, Vol. 9, 645–671, 1995.

Kron, G., "Equivalent circuits to represent the electromagnetic field equations", *Phys. Rev.*, Vol. 64, 126–128, 1943.

Kron, G., "Equivalent circuit of the field equations of Maxwell-I", *Proc. IRE*, 289–299, May 1944.

Kron, G., "Numerical solution of ordinary and partial differential equations by means of equivalent circuits", *J. Appl. Phys.*, Vol. 16, 172–186, 1945.

Kumar, G. and K. P. Ray, *Broadband Microstrip Antennas*, Artech House, 2003.

Kurokawa, K., *An Introduction to Microwave Circuits*, New York, Academic Press, 1969.

Lamensdorf, D. and L. Susman, "Baseband-pulse-antenna techniques", *IEEE Antennas and Propagation Magazine*, Vol. 36, No. 1, 20–30, 1994.

Landau, L. D., E. M. Lifshitz and L. P. Pitaevskii, *Electrodynamics of Continuous Media*, Pergamon, Oxford 1984.

Lattarulo, F. (Ed.), *Electromagnetic Compatibility in Power Systems*, Elsevier, 2007.

Lécuyer, C., *Making Silicon Valley: Innovation and the Growth of High Tech.*, The MIT Press, Cambridge, MA, 2005.

Lee, K. F., *Principles of Antenna Theory*, John Wiley & Sons, New York, 1984.

Lee, T. H., *Planar Microwave Engineering: A Practical Guide to Theory, Measurement, and Circuits*, Cambridge University Press, 2004.

Leenaerts, D., J. Tang and C. Vaucher, *Circuit Design for RF Transceivers*, Kluwer Academic Publishers, 2001.

Leis, R., *Initial Boundary Value Problems in Mathematical Physics*, John Wiley & Sons, 1986.

Levine, H. and J. Schwinger, "On the theory of electromagnetic wave diffraction by an aperture in an infinite plane conducting screen", *Comm. Pure Appl. Math.* III **4**, 355–391, 1950.

Levy, R., "Theory of direct coupled-cavity filters", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-15, 340–348, June 1967.

Levy, R., "Improved single and multiaperture waveguide coupling theory, including explanation of mutual interactions", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-28, 331–338, April 1980.

Levy, R. and S. B. Cohn, "A History of microwave filter research, design, and development", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-32, No. 9, 1055–1067, Sept. 1984.

Lewin, L., *Advanced Theory of Waveguides*, lliffe and Sons, London, 1951.

Lo, Y. T. and S. W. Lee, *Antenna Handbook-Theory, Applications, and Design*, VNR, 1988.

Loane, J. T. and S. Lee, "Gain optimization of a near-field focusing array for hyperthermia application," *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-37, 1629–1635, Oct. 1989.

Lovelock, D. and H. Rund, *Tensors, Differential forms, and Variational Principles*, John Wiley & Sons, New York, 1975.

Luk, K. M., K. W. Leung and J. R. James, *Dielectric Resonator Antennas*, Research Studies Press, 2002.

MacMillan, W. D., *The Theory of the Potential*, New York: Dover, 1958.

Makimoto, M. and S. Yamashita, *Microwave Resonators and Filters for Wireless Communication: Theory, Design, and Application*, Springer, 2001.

Marcuse, D., *Light Transmission Optics*, 2nd Edition, Van Nostrand, 1982.

Marcuvitz, N., *Waveguide Handbook*, McGraw-Hill Book Company Inc., 1951.

Marion, J. B. and M. A. Heald, *Classical Electromagnetic Radiation*, 2nd Edition, Academic Press, 1980.

Marin, L., "Natural-mode representation of transient scattered fields", *IEEE Trans. Antennas and Propagat.*, Vol. AP-21, No. 6, 809–818, Nov. 1973.

Marks, R. B., "Application of the singular function expansion to an integral equation for scattering", *IEEE Trans. Antennas and Propagat.*, Vol. AP-34, 725–728, May 1986.

Marks, R. B., "The singular function expansion in time-dependent scattering", *IEEE Trans. Antennas Propagat.*, Vol. AP-37, 1559–1565, Dec. 1989.

Marshall, J. S. and W. M. Palmer, "The distribution of raindrops with size", *Journal Meteorol.*, Vol. 5, 165–166, 1948.

Mason, W. P. and R. A. Sykes, "The use of coaxial and balanced transmission lines in filters and wide band transformers for high radio frequencies", *Bell Syst. Tech. J.*, Vol. 16, 275–302, 1937.

Massa, G. Di, D. Cuomo, A. Cutolo and G. Delle Cave, "Open resonator for microwave application", *IEE Proc., Microwaves, Antennas and Propagation, Pt. H*, Vol. 136, 159–164, 1989.

Matthaei, G. L., L. Young and E. M. T. Jones, *Microwave Filters, Impedance-Matching Networks, and Coupling Structures*, McGraw-Hill, 1964.

Mautz, J. R. and R. F. Harrington, "Radiation and scattering from bodies of revolution", *Appl. Sce. Res.*, Vol. 20, 405–435, June 1969.

Mautz, J. R. and R. F. Harrington, "H-field, E-field, and combined-field solutions for conducting bodies of revolution", *Arch. Elektron, Übertragungstech., Elctron. Commun.*, Vol. 32, 19–164, 1978.

Maxwell, J. C., *A Treatise on Electricity and Magnetism*, 3rd Edition, Vol. 1, Dover Publications, Inc., New York, 1954; first published in 1891.

McLean, J. S., "A re-examination of the fundamental limits on the radiation Q of electrically small antennas", *IEEE Trans. Antennas and Propagat.*, Vol. AP-44, 672–676, 1996.

McIntosh, R. E. and J. E. Sarna, "Bounds on the optimum performance of planar antennas for pulse radiation", *IEEE Trans. Antennas and Propagat.*, Vol. AP-30, 381–389, July 1983.

McSpaddan, J. O. and J. C. Mankins, "Space solar power programs and microwave wireless power transmission technology," *IEEE Microw. Mag.*, Vol. 3, 46–57, Dec. 2002.

Meier, P. J., "Two integrated-circuit media with special advantages at millimeter wavelengths", *IEEE MTT-S Int. Microwave Symp. Digest*, 221–223, 1972.

Meikle, H., Modern Radar Systems, Artech House, 2008.

Menzel, W. and I. Wolff, "A Method for Calculating the Frequency-Dependent Properties of Microstrip Discontinuities," *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-25, 107–112, Feb. 1977.

Mikhlin, S. G., *Variational Methods in Mathematical Physics*, Oxford, Pergamon Press, 1964.

Mikhlin, S. G., *Linear Integral Equations*, Delhi: Hindustan, 1960.

Miller, K. S., *Complex Stochastic Processes*, Addison-Wesley Publishing Company, 1974.

Miller, R. F., "On the completeness of sets of solutions to the Helmholtz equation", *IMA J. Appl. Math.*, Vol. 30, 27–37, 1983.

Mills, J. P., *Electromagnetic Interference Reduction in Electronic Systems*, PTR Prentice-Hall, Englewood Cliffs, NJ, 1993.

Misra, D. K., *Radio Frequency and Microwave Communication Circuits*, John Wiley & Sons, 2001.

Mittra, R., *Computer Techniques for Electromagnetics*, Pergamon Press, 1973.

Mittra, R. and W. W. Lee, *Analytical Techniques in the Theory of Guided Waves*, Macmillan, New York, 1971.

Molan, J., *The Physics of Lightning*, English Universities Press, London, 1963.

Monteath, G. D., *Applications of the Electromagnetic Reciprocity Principle*, Pergamon Press, 1973.

Montgomery, C. G., R. H. Dicke and E. M. Purcell, *Principles of Microwave Circuits*, McGraw-Hill, 1948.

Montrose, M. I., *Printed Circuits Board Design Techniques for EMC Compliance*, 2nd Edition, IEEE Press, New York, 2000.

Montrose, M. I. and E. D. Nakauchi, *Testing for EMC Compliance*: *Approaches and Techniques*, IEEE Press, 2004.

Moon, P. and D. E. Spencer, *Field Theory Handbook*, Springer, 1988.

Morgan, D. A., *A Handbook for EMC Testing and Measurement Series 8*, Peter Peregrinus, London, 1994.

Morita, N., "Surface integral representations for electromagnetic scattering from dielectric cylinders", *IEEE Trans. Antennas and Propagat.*, Vol. AP-26, 261–266, No. 2, March 1978.

Morita, N., "Another method of extending the boundary condition for the problem of scattering by dielectric cylinders", *IEEE Trans. Antennas and Propagat.*, Vol. AP-27, 97–99, No. 1, Jan. 1979.

Morita, N., "Resonant solutions involved in the integral equation approach to scattering from conducting and dielectric cylinders", *IEEE Trans. Antennas and Propagat.*, Vol. AP-27, 869–871, No. 6, Nov. 1979.

Morita, N., N. Kumagai and J. R. Mautz, *Integral Equation Methods for Electrmagnetics*, Artech House, 1990.

Morse, P. M. and H. Feshbach, *Methods of Theoretical Physics*, McGraw-Hill, 1953.

Moses, H. E., "Solutions of Maxwell equations in terms of a spinor notation: The direct and inverse problem", *Phys. Rev.*, Vol. 113, No. 6, 1670–1679, 1959.

Moses, H. E. and R. T. Prosser, "Initial conditions sources, and currents for prescribed time-dependent acoustic and electromagnetic fields in three dimensions, Part 1: The inverse initial value problem, Acoustic and electromagnetic bullets, expanding waves, and imploding waves", *IEEE Trans. Antennas and Propagat.*, Vol. AP-34, 188–196, Feb. 1986.

Müller, C., "Electromagnetic radiation patterns and sources", *IRE Trans. Antennas and Propagat.*, Vol. AP-4, 224–232, July 1956.

Müller, C., *Foundations of the Mathematical Theory of Electromagnetic Waves*, Springer, 1969.

Mushiake, Y., *Self-Complementary Antennas*, Springer-Verlag, 1996.

Namiki, M. and K. Horiuchi, "On the transient phenomena in the wave guide", *J. Phys. Soc. Japan*, Vol. 7, 190–193, 1952.

Niehenke, E. C., R. A. Pucel and I. J. Bahl, "Microwave and millimeter-wave integrated circuits", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-50, No. 3, 846–857, March 2002.

Nguyen, C., *Analysis Methods for RF, Microwave and Millimeter-Wave Planar Transmission Line Structures*, John Wiley & Sons, 2001.

Norton, K. A., "The propagation of radio waves over the surface of the Earth an in the upper atmosphere", *Proc. IRR*, Vol. 24, 1367–1387, 1936 and Vol. 25, 1203–1236, 1937.

Nyquist, H., "Thermal agitation of electric charge in conductors", *Phys. Rev.*, Vol. 32, 110–113, 1928.

Oguchi, T., "Electromagnetic wave propagation and scattering in rain and other hydrometeors", *Proc. IEEE*, Vol. 71, 1029–1078, Sept. 1983.

Okumura, Y. *et al.*, "Field strength and its variability in VHF and UHF land mobile services", *Review Electrical Engineering Lab*, Vol. 16, Nos. 9–10, 825–873, Sept.–Oct. 1968.

Oliner, A., "Historical perspectives on microwave field theory", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-32, 1022–1045, Sept. 1984.

Oseen, C. W., Uber die Electromagnetische Spektrum einen Dunnen Ringes, *Arkiv. Mat. Astron. Fysik*, Vol. 9, No. 28, 1–34, 1913.

Ott, H. W., *Noise Reduction Techniques in Electronic Systems*, 2nd Edition, Wiley, New York, 1988.

Owens, R. P., "Accurate analytical determination of quasi-static microstrip line parameters", *The Radio and Electronic Engineers*, Vol. 46, No. 7, 360–364, July 1976.

Packard, K. S., "The origin of waveguides: A case of multiple rediscovery", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-32, 961–969, Sept. 1984.

Panofsky, W. K. H. and M. Phillips, *Classical Electricity and Magnetism*, 2nd Edition, Addison-Wesley, Reading, MA, 1962.

Page, L., "A derivation of the fundamental relations of electrodynamics from those of electrostatics", *Am. J. Sci.*, Vol. 44, 57–68, 1912.

Papas, C. H., *Theory of Electromagnetic Wave Propagation*, McGraw-Hill, New York, 1965.

Parsons, J. D., *The Mobile Radio Propagation Channel*, 2nd Edition, John Wiley & Sons, 2000.

Paul, C. R., *Analyses of Multiconductor Transmission Lines*, Wiley, New York, 1994.

Paul, C. R., *Introduction to Electromagnetic Compatibility*, Wiley, 2006.

Paulraj, A., R. Nabar and D. Gore, *Introduction to Space-Time Wireless Communications*, Cambridge University Press, 2003.

Peebles, P. Z., *Radar Principles*, John Wiley & Sons, 1998.

Pelzer, H., "Energy density of monochromatic radiation in a dispersive medium", *Proc. Roy. Soc.* (*London*) *A*, Vol. 208, 365–366, Sept. 1951.

Peterson, A. F., "The interior resonance problem associated with surface integral equations of electromagnetics: Numerical consequences and a survey of remedies", *Electromagnetics*, Vol. 10, 293–312, July–Sept. 1990.

Peterson, A. F., S. L. Ray and R. Mittra, *Computational Methods for Electromagnetics*, Oxford University Press, 1998.

Pincherle, L., Electromagnetic waves in metal tubes filled longitudinally with two dielectrics, *Phys. Rev.*, Vol. 66, No. 5, 118–130, 1944.

Pinsker, M. S., *Information and Information Stability of Random Process*, Holden Bay, San Francisco, 1964.

Pocock, M., M. J. Bluck and S. P. Walker, "Electromagnetic scattering from 3-D curved dielectric bodies using time-domain integral equations", *IEEE Trans. Antennas and Propagat.*, Vol. AP-46, 1212–1219, No. 8, Aug. 1998.

Poggio, A. J. and E. K. Miller, "Integral equation solution of three dimensional scattering problems", in *Computer Techniques for Electromagnetics*, New York: Pergamon Press, 1973.

Popvić, B. D., "Electromagnetic field theorems", *IEE Proc.*, Pt. A, Vol. 128, 47–63, Jan. 1981.

Popvić, B. D., M. B. Dragovic and A. R. Djordjevic, *Analysis and Synthesis of Wire Antennas*, Research Studies Press, John Wiley & Sons, New York, 1982.

Pospieszalski, M. W., "Cylindrical dielectric resonators and their applications in TEM line microwave circuits," *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-27, 233–238, March 1979.

Pozar, D. M., *Microwave Engineering*, John Wiley & Sons, 1998.

Proakis, J. G., *Digital Communications*, McGraw-Hill, 3rd Edition, 1995.

Prokhorov, A. M., "Molecular amplifier and generator for submillimeter waves", *Sov. Phys. JETP*, Vol. 7, 1140–1141, Dec. 1958.

Ragan, G. L. (Ed.), *Microwave transmission circuits*, Massachusetts Institute of Technology Radiation Laboratory, Dover Publications, 1965.

Ramm, A. G., "Non-selfadjoint operators in diffraction and scattering", *Math. Methods in Applied Science*, Vol. 2, 327–346, 1980.

Ramm, A. G., "Theoretical and practical aspects of the singularity and eigenmode expansion methods", *IEEE Trans. Antennas and Propagat.*, Vol. AP-28, 897–901, Nov. 1980.

Ramm, A. G., "Mathematical foundations of the singularity and eigenmode expansion methods (SEM and EEM)", *J. Math. Anal. Appl.*, Vol. 86, 562–591, 1982.

Ramo, S. and J. R. Whinnery, *Fields and Waves in Modern Radio*, John Wiley & Sons, 1953.

Rappaport, T. S., *Wireless communications*: *Principles and practice*, Prentice Hall, 1996.

Rao, S. M., D. R. Wilton and A. W. Glisson, "Electromagnetic scattering by surfaces of arbitrary shape", *IEEE Trans. Antennas and Propagat.*, Vol. AP-30, 409–417, No. 3, May 1982.

Rayleigh, L, "On the Passage of waves through tubes, or the vibration of dielectric cylinders", *Philosophical Magazine*, Vol. 43, 125–132, February 1897.

Razavi, B., *RF microelectronics*, Prentic Hall, 1998.

Read, F. H., *Electromagnetic Radiation*, John Wiley & Sons, New York, 1980.

Reddy, J. N., *Applied Functional Analysis and Variational Methods in Engineering*, McGraw-Hill, 1986.

Reitz J. R., F. J. Milford and R. W. Christy, *Foundations of Electromagnetic Theory*, Addison-Wesley, 1979.

Rhodes, D. R., "On the stored energy of planar apertures", *IEEE Trans. Antennas and Propagat.*, Vol. AP-14, 676–683, 1966.

Rhodes, D. R., Author's reply, *IEEE Trans. Antennas and Propagat.*, Vol. AP-15, 568–569, 1967.

Rhodes, D. R., "Observable stored energies of electromagnetic systems", *Journal of Franklin Institute*, Vol. 302, No. 3, 225–237, 1976.

Rhodes, D. R., "A reactance theorem", *Proceedings of the Royal Society of London*, Series A (Mathematical and Physical Science), Vol. 353, No. 1672, 1–10, 1977.

Richards, M. A., J. A. Scheer and W. A. Holm, *Principles of Modern Radar*, SciTech Publishing, 2010.

Richards, P. I., "Resistor-transmission-line circuits", *Proceedings of the IRE*, Vol. 36, 217–220, Feb. 1948.

Richmond, J. H., "A reaction theorem and its application to antenna impedance calculations", *IRE Trans. Antennas Propagat.*, Vol. AP-9, 515–520, Nov. 1961.

Richtmyer, R. D., "Dielectric resonators", *J. Appl. Phys.*, Vol. 10, 391–398, 1939.

Rizzi, P. A., *Microwave Engineering*, Prentice Hall, 1988.

Ruck, G. T. *et al.* (Ed.), *Radar Cross Section Handbook*, Vol. 1–2, Plenum Press, 1970.

Rumsey, V. H., "A new way of solving Maxwell equations", *IRE Trans. Antennas Propagat.*, Vol. AP-9, 461–463, Sept. 1961.

Rumsey, V. H., "A short way of solving advanced problems in electromagnetic fields and other linear systems", *IEEE Trans. Antennas and Propagat.*, Vol. AP-11, 73–86, Jan. 1963.

Rumsey, V. H., "Some new forms of Huygens' principle", *IRE Trans. Antennas and Propagat.*, Vol. AP-7, 103–116, Dec. 1959.

Rumsey, V. H., "Reaction concept in electromagnetic theory", *Phys. Rev.*, Vol. 17, 952–956, 1954.

Rumsey, V. H., "Frequency independent antennas", IRE National Convention Record, pt. 1, 114–118, 1957.

Rumsey, V. H., *Frequency Independent Antennas*, Academic Press, 1966.

Ruppin, R., "Electromagnetic energy density in a dispersive and absorptive material," *Physics Letters A*, Vol. 299, Nos. 2–3, 309–312, July 1, 2002.

Samaddar, S. N. and E. L. Mokole, "Biconical antennas with unequal cone angles", *IEEE Trans. Antennas and Propagat.*, Vol. AP-46, 181–193, 1998.

Sarkar, T. K., *History of Wireless*, John Wiley and Sons, 2006.

Schawlow, A. L. and C. H. Townes, "Infrared and optical masers", *Phys. Rev.*, Vol. 29, 1940–1949, Dec. 1958.

Schantz, H. G., "Electromagnetic energy around Hertzian dipoles", *IEEE Antennas and Propagation Magazine*, Vol. 43, No. 2, 50–62, Apr. 2001.

Schelkunoff, S. A., *Antennas*: *Theory and Practice*, John Wiley & Sons, 1952.

Schulz-DuBois, E. O., "Sommerfeld Pre- and Postcursors in the context of waveguide transients", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-18, 455–460, Aug. 1970.

Schwinger, J. and D. S. Saxon, *Discontinuities in Waveguides*, Gordon and Breach, New York, 1968.

Schwinger, J., L. L. DeRaad Jr., K. A. Milton and W. Tsai, *Classical Electrodynamics*, Perseus Books, 1998.

Sengupta, D. L. and V. V. Liepa, *Applied Electromagnetics and Electromagnetic Compatibility*, John Wiley & Sons, 2006.

Shannon, C. E., "A mathematical theory of communication", *Bell Syst. Tech. J.*, Vol. 27, 379–423, July 1948; Vol. 27, 623–656, Oct. 1948.

Shannon, C. E., "Communication in the presence of noise", *Proc. IRE*, Vol. 37, 10–21, Jan. 1949.

Sherman, J. W., "Properties of focused apertures in the Fresnel region", *IRE Trans. Antennas Propagat.*, Vol. AP-10, 399–408, July 1962.

Shlivinski, A. *et al.*, "Antenna characterization in time domain", *IEEE Trans. Antennas and Propagat.*, Vol. AP-45, 1140–1149, July 1997.

Shlivinski, A. and E. Heyman, "Time domain near field analysis of short pulse antennas — Part 1 and Part 2", *IEEE Trans. Antennas and Propagat.*, Vol. AP-47, 271–286, Feb. 1999.

Silva, E., *High Frequency and Microwave Engineering*, Butterworth-Heinemann, 2001.

Silver, S., *Microwave Antenna Theory and Design*, Dover publications, New York, 1949.

Silvester, P. and P. Benedek, "Equivalent capacitance of microstrip open circuits," *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-20, 511–516, Aug. 1972.

Silvester, P. and P. Benedek, "Equivalent capacitance of microstrip gaps and steps", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-20, 729–733, Nov. 1972.

Silvester, P. and Peter Benedek, "Microstrip discontinuity capacitances for right-angle bends, T-junctions, and crossings," *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-21, 341–346, May 1973 (Corrections: May 1975).

Singh, H. M. and J. Howard, "Fundamentals of wideband stripline directional coupler design", *Microwave Journal*, Vol. 32, No. 11, 99–106, Nov. 1989.

Siragusa, R., P. Lemaitre-Auger and S. Tedjini, "Near field focusing circular microstrip antenna array for RFID applications," *in IEEE Int. Symp. Antennas Propag. Dig.*, 1–4, June 2009.

Sizun, H., *Radio Wave Propagation for Telecommunication Applications*, Springer, 2005.

Sklar, B., *Digital Communications-Fundamentals and Applications*, Prentice Hall, 1988.

Sklar, B., "Defining, designing, and evaluating digital communication systems", *IEEE Communications Magazine*, Vol. 31, No. 11, 91–101, Nov. 1993.

Skolnik, M. I., *Introduction to Radar Systems*, McGraw-Hill, New York, 1980.

Skolnik, M. I., *Radar Handbook*, 3rd Edition, McGraw-Hill, New York, 2008.

Slater, J. C., *Microwave Electronics*, Van Nostrand, Princeton, New Jersey, 1950.

Slepian, D. and H. O. Pollak, "Prolate spherical wave functions, Fourier analysis and uncertainty-I, II", *Bell Sys. Tech. J.*, Vol. 40, 43–84, Jan. 1961.

Smith, G. S. and T. W. Hertel, "On the transient radiation of energy from simple current distributions and linear antennas", *IEEE Antennas and Propagation Magazine*, Vol. 43, No. 3, 49–63, Jun. 2001.

Sobol, H., "Microwave communications-A historical perspective", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-32, 1170–1181, Sept. 1984,

Soejima, T., "Fresnel gain of aperture aerials", *Proc. IEE* (London), Vol. 110, 1021–1027, 1963.

Sommerfeld, A., *Electrodynamics*, Academic Press, New York, 1949.

Spohn, H., *Dynamics of Chareged Particles and Their Radiation Field*, Cambridge University Press, 2004.

Stavroulakis, P., *Interference Analysis and Reduction for Wireless Systems*, Artech House, 2003.

Sten, J. C.-E. *et al.*, "Quality factor of an electrically small antenna radiating close to a conducting plane", *IEEE Trans. Antennas and Propagat.*, Vol. AP-49, 829–837, May 2001.

Stephan, K. D., J. B. Mead, D. M. Pozar, L. Wang and J. A. Pearce, "A near field focused microstrip array for a radiometric temperature sensor", *IEEE Trans. Antennas Propag.*, Vol. 55, 1199–1203, April 2007.

Stevenson, A. F., "Relations between the transmitting and receiving properties of antennas", *Quart. Appl. Math.*, Vol. 5, 140–148, Jan. 1948.

Stevenson, A. F., "Solution of electromagnetic scattering problems as power series in the ratio (dimension of scatter/wavelength)", *J. Appl. Phys.*, Vol. 24, 1134–1142, 1953.

Stinson, D. C., *Intermediate Mathematics of Electromagnetics*, Prentice-Hall, 1976.

Stocia, P., Y. Jiang and J. Li, "On MIMO channel capacity: An intuitive discussion", *IEEE Signal Processing Magazine*, Vol. 22, No. 3, 83–84, May 2005.

Storer, J. E., "Impedance of thin-wire loop antennas", *Trans. AIEEE* (Communication and Electronics), Vol. 75, 606–619, Nov. 1956.

Strassner, B. and K. Chang, "5.8-GHz circularly polarized rectifying antenna for wireless microwave power transmission", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-50, 1870–1876, Aug. 2002.

Stratton, J. A., *Electromagnetic Theory*, McGraw-Hill, New York, 1941.

Stutzman, W. L. and G. A. Thiele, *Antenna Theory and Design*, John Wiley & Sons, New York, 1981.

Taga, T. and K. Tsunekawa, Performance analysis of a built-in inverted-F antenna for 800 MHz band portable radio units, *IEEE Journal on Selected Areas in Comm.*, Vol. 5, No. 5, 921–929, June, 1987.

Taga, T., Analysis for mean effective gain of mobile antennas in land mobile radio environments, *IEEE Trans. Veh. Technol.*, Vol. 39, 117–131, 1990.

Tai, C.-T., *Dyadic Green Functions in Electromagnetic Theory*, IEEE Press, 1994.

Tai, C.-T., "On the transposed radiating systems in an anisotropic medium", *IRE Trans. Antennas and Propagat.*, Vol. AP-9, 502–503, Sept. 1961.

Takeshita, S., "Power transfer efficiency between focused circular antennas with Gaussian illumination in Fresnel region", *IEEE Trans. Antennas and Propagat.*, Vol. AP-16, 305–309, May 1968.

Telatar, I. E., "Capacity of multi-antenna Gaussian channels", *Europ. Trans. Telecomm.*, Vol. 10, 585–595, Nov. 1999.

Tesche, F. M., M. V. Ianuz and T. Karlsson, *EMC Analysis Methods and Computational Models*, Wiley, New York, 1997.

Tonning, A., "Energy density in continuous electromagnetic media", *IEEE Trans. Antennas and Propagat.*, Vol. AP-8, 428–434, July 1960.

Tsang, L., J. A. Kong and K. Ding, *Scattering of Electromagnetic Waves*, John Wiley & Sons, 2000.

Tse, D. and P. Viswanath, *Fundamentals of Wireless Communications*, Cambridge University Press, 2005.

Uda, S., "Wireless Beam of Short Electric Waves", *J. IEEE* (Japan), 273–282, 1926; 1209–1219, 1927.

Umashankar, K. and A. Taflove, *Computational Electromagnetics*, Artech House, 1993.

Vandenbosch, G. A. E., "Reactive energies, impedance, and Q factor of radiating structures," *IEEE Trans. Antennas Propagat.*, Vol. 58, No. 4, 1112–1127, 2010.

Vartanian, P. H., W. P. Ayres and A. L. Helgesson, "Propagation in dielectric slab loaded rectangular waveguide", *IRE Trans. Microwave Theory and Tech.*, Vol. 6, No. 2, 215–222, April 1958.

Vekua, I. N., "About the completeness of the system of metaharmonic functions", *Dokl. Akad. Nauk.*, USSR 90, 715–718, 1953.

Vendelin, G. D., A. M. Pavio, and U. L. Rhode, *Microwave Circuit Design Using Linear and Nonlinear Techniques*, 2nd Edition, Wiley-Interscience, 2005.

Volakis, J. L., *Antenna Engineering Handbook*, 4th Edition, McGraw-Hill, 2007.

Wadell, B. C., *Transmission Line Design Handbook*, Artech House, 1991.

Wait, J. R., *Wave Propagation Theory*, Pergamon Press, 1981.

Wallace, J. W. and M. A. Jensen, "Mutual coupling in MIMO wireless systems: A rigorous network theory analysis", *IEEE Trans. Wireless Comm.*, Vol. 3, No. 4, 1317–1325, July 2004.

Weng, C. H., C. F. Yang, Y. S. Lin, F. S. Chen, Y. C. Huang and C. W. Hsu, "Design of RFID near-field focusing circular patch array antenna at 2.4 GHz with applications," *Internet of Things* (*IOT*), 1–4, Dec. 2010.

Wasylkiwskyj, W. and W. K. Kahn, "Theory of mutual coupling among minimum-scattering antennas", *IEEE Trans. Antennas and Propagat.*, Vol. AP-18, 204–216, March 1970.

Waterman, P. C., "Matrix formulation of electromagnetic scattering", *Proc. IEEE*, Vol. 53, 806–812, Aug. 1965.

Waterman, P. C., "Symmetry, unitarity, and geometry in electromagnetic scattering", *Phys. Rev.*, Vol. D3, 825–839, 1971.

Weinberger, H. F., *Variational Methods for Eigenvalue Approximation*, Philadelphia: Society for Industrial and Applied Mathematics, 1974.

Wheeler, H. A., "Transmission-line properties of parallel strips separated by a dielectric sheet", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-13, 172–185, No. 2, 1965.

Wheeler, H. A., "Small antennas", *IEEE Trans. Antennas and Propagat.*, Vol. AP-23, 462–469, No. 4, July 1975.

White, J. F., *High Frequency Techniques*: *A Introduction to RF and Microwave Engineering*, John Wiley & Sons, 2004.

Widrow, B., P. E. Mantey, L. J. Griffiths and B. B. Goode, "Adaptive antenna systems", *Proc. IEEE*, Vol. 55, No. 12, 2143–2159, Aug. 1967.

Wiener, N., *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, John Wiley and Sons, New York, 1949.

Wilcox, C. H., "An expansion theorem for electromagnetic fields", *Comm. Pure Appl. Math.*, Vol. 9, 115–134, 1956.

Wilcox, C. H., "Debye potentials", *J. Math. Mech.*, Vol. 6, 167–202, 1957.

Wilkinson, E. J., "An N-way Power Divider", *IRE Trans. Microwave Theory and Tech.*, Vol. 8, 116–118, Jan. 1960.

Wilton, D. R., S. M. Rao, A. W. Glisson and D. H. Schaubert, "Potential integrals for uniform and linear source distributions on polygonal polyhedral domains", *IEEE Trans. Antennas and Propagat.*, Vol. AP-32, 276–281, No. 3, Mar. 1984.

Wu, T. T. and R. W. P. King, "Transient response of linear antennas driven from a coaxial line", *IEEE Trans. Antenna and Propagat.*, Vol. AP-11, 17–23, Jan. 1963.

Xie, F., G. M. Yang and W. Geyi, "Optimal design of antenna array for energy harvesting", *IEEE Antennas Wireless Propag. Letts.*, Vol. 12, 155–158, Jan. 2013.

Yacoub, M. D., *Foundations of Mobile Radio Engineering*, CRC Press, Boca Raton, Feb. 1993.

Yagi, H., "Beam transmission of ultra short waves", *IRE Proceedings.*, Vol. 16, 715–741, 1928.

Yee, K. S., "Numerical solution of initial boundary value problems involving Maxwell's equations", *IEEE Trans. Antenna and Propagat.*, Vol. AP-14, 302–307, May 1966.

Young, L., "Direct-coupled cavity filters for wide and narrow bandwidths", *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-11, 162–178, May 1963.

Yu, K. and B. Ottersten, "Models for MIMO propagation channels: A review", *Wirel. Commun. Mob. Comput.*, Vol. 2, No. 7, 653–666, Nov. 2002.

Zahn, H., "Ueber den Nachweis elektromagnetischer Wellen an dielektrischen Draehten", *Annalen der Physik*, Vol. 37, 907–933, 1916.

Ziemer, R. E. and R. L. Peterson, *Introduction to Digital Communication*, Prentice Hall, 2001.

Zhang, W. X., *Engineering Electromagnetism*: *Functional Methods*, Ellis Horwood, 1993.

Index