20

The Skin Effect

20.1 Introduction

We are now in a position to formulate any electromagnetic problem in terms of Maxwell's equations. This chapter deals with the skin effect, the first practical electromagnetic problem we will solve as an example of this kind. A related effect, the proximity effect, is then considered briefly.

We know that a time-invariant current in a homogeneous cylindrical conductor is distributed uniformly over the conductor cross section. If the conductor is not cylindrical, the time-invariant current in it is not distributed uniformly, but it exists in the *entire* conductor. We shall see in this chapter that a time-varying current has a tendency to concentrate near the surfaces of conductors. If the frequency is very high, the current is restricted to a very thin layer near the conductor surfaces, practically on the surfaces themselves. Because of this extreme case, the entire phenomenon of nonuniform distribution of time-varying currents in conductors is known as the *skin effect*.

The cause of the skin effect is electromagnetic induction. A time-varying magnetic field is accompanied by a time-varying induced electric field, which in turn creates secondary time-varying currents (induced currents) and a secondary magnetic field. We know from Lentz's law that the induced currents produce the magnetic flux, which is opposite to the external flux (which "produced" the induced currents), so that the total flux is reduced. The larger the conductivity, the larger the induced currents are, and the larger the permeability, the more pronounced is the flux reduction.

Consequently, both the total time-varying magnetic field and induced currents inside conductors are reduced when compared with the dc case.

The skin effect is of considerable practical importance. For example, at very high frequencies a very thin layer of conductor carries most of the current, so we can coat any conductor with silver (the best available conductor) and have practically the entire current flow through this thin silver coating. (Unfortunately silver oxidizes easily, so gold is often used instead because it is inert.) Even at low, power-line frequencies (60 Hz in the United States and Canada, and 50 Hz in Europe), in the case of high currents the use of thick, solid conductors is not efficient; bundled conductors are used instead.

The skin effect exists in all conductors, but as mentioned, the tendency of current and magnetic flux to be restricted to a thin layer on the conductor surface is much more pronounced for a ferromagnetic conductor than for a nonferromagnetic conductor of the same conductivity. For example, for iron at 60 Hz the thickness of this layer is on the order of only 0.5 mm. Consequently, solid ferromagnetic cores for alternating current electric motors, generators, transformers, etc., would have very high eddy-current losses. Therefore laminated cores made of thin, mutually insulated sheets are used instead. At very high frequencies, ferrites (ferrimagnetic materials) are used because they have very low conductivity when compared to metallic ferromagnetic materials.

Questions and problems: Q20.1 to Q20.10

20.2 Skin Effect

Consider an idealized case of a sinusoidal current in a homogeneous conducting half-space, as sketched in Fig. 20.1. Let the angular frequency of the current be ω and let the medium have a conductivity σ and permeability μ. Finally, assume that the

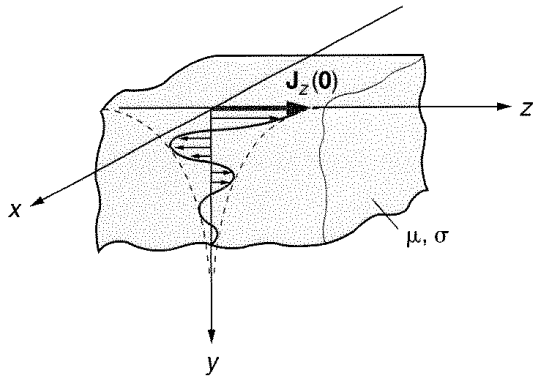

Figure 20.1 Homogeneous conducting half-space with sinusoidal current. The amplitude of the current density vector at an instant of time versus the distance y from the boundary surface is as indicated.

current density vector is parallel to the boundary surface, and that it has a single component, for example, $\mathbf{J} = J_z \mathbf{u}_z$, depending on the coordinate y (the distance from the interface) only. We wish to determine the distribution of current in the conducting half-space.

At first glance, one might be tempted to think this problem is purely academic. We will see, however, that it has important practical implications. After solving Maxwell's equations, we will find that the intensity of the current density vector and of all the field vectors decreases exponentially with the distance from the boundary surface. This decrease is more rapid at higher frequencies and for higher conductivities and permeabilities. For conductors used in everyday practice (copper, for instance), and frequencies higher than about 1 MHz, the thickness of the current layer becomes less than a fraction of a millimeter. If we consider *any* conductor whose radius of curvature is much larger than the current layer thickness, the results we will obtain can be applied with high accuracy. Therefore this section has considerable practical importance and applicability.

We start the analysis from the differential form of Maxwell's equations in complex form. Because we assume the medium to be a good conductor, the displacement current density in the second equation can be neglected. We start from

$$\nabla \times \mathbf{E} = -\mathrm{j}\omega \mathbf{B} \qquad \nabla \times \mathbf{H} = \mathbf{J}. \tag{20.1}$$

For simplicity we do not underline the complex vectors $\mathbf{E}$, $\mathbf{B}$, $\mathbf{H}$, and $\mathbf{J}$. Since $\mathbf{E} = \mathbf{J}/\sigma$ and $\mathbf{H} = \mathbf{B}/\mu$, Eqs. (20.1) become

$$\nabla \times \mathbf{J} = -\mathrm{j}\omega\sigma \mathbf{B} \qquad \nabla \times \mathbf{B} = \mu \mathbf{J}. \tag{20.2}$$

We assumed that the current density vector has only a z component, which depends only on y. From the Biot-Savart law and symmetry it therefore follows that there is only an x component of the vector $\mathbf{B}$. According to the expression for the curl in a rectangular coordinate system, Eqs. (20.2) become

$$\frac{\mathrm{d}J_z}{\mathrm{d}y} = -\mathrm{j}\omega\sigma B_x \qquad -\frac{\mathrm{d}B_x}{\mathrm{d}y} = \mu J_z. \tag{20.3}$$

We use ordinary derivatives (not partial derivatives) because J_z and B_x depend only on y.

From Eqs. (20.3) we can eliminate B_x to obtain an equation in J_z:

$$\frac{\mathrm{d}^2 J_z}{\mathrm{d}y^2} = \mathrm{j}\omega\mu\sigma J_z. \tag{20.4}$$

This equation has a simple solution,

$$J_z(y) = J_1 \mathrm{e}^{Ky} + J_2 \mathrm{e}^{-Ky}, \tag{20.5}$$

where

$$K = \sqrt{\mathrm{j}\omega\mu\sigma} = (1+\mathrm{j})\sqrt{\frac{\omega\mu\sigma}{2}} = (1+\mathrm{j})k \qquad k = \sqrt{\frac{\omega\mu\sigma}{2}}. \tag{20.6}$$

Assume that for $y = 0$ the current density is $J_z(0)$. For $y \to \infty$, the current density cannot increase indefinitely, so $J_1 = 0$. Thus we finally have

$$J_z(y) = J_z(0)e^{-ky}e^{-jky}. \tag{20.7}$$

The intensity of the current density vector decreases exponentially with increasing y. At a distance

$$\delta = \frac{1}{k} = \sqrt{\frac{2}{\omega\mu\sigma}} \quad \text{(m)}, \tag{20.8}$$

(Definition of skin depth)

the amplitude of the current density vector decreases to 1/e of its value $J_z(0)$ at the boundary surface. This distance is known as the *skin depth*.

As mentioned, although derived for a special case of currents in a half-space, the preceding analysis is valid for a current distribution in any conductor whose radius of curvature is much larger than the skin depth.

Example 20.1—Skin depth for some common materials. As an illustration, let us determine the skin depth for copper ($\sigma = 57 \cdot 10^6$ S/m, $\mu = \mu_0$), iron ($\sigma = 10^7$ S/m, $\mu_r = 1000$), seawater ($\sigma = 4$ S/m, $\mu = \mu_0$), and wet soil ($\sigma = 0.01$ S/m, $\mu = \mu_0$) at 60 Hz (power-line frequency), 10^3 Hz, 10^6 Hz, and 10^9 Hz. The results are summarized in Table 20.1. Note that for iron the skin depth is very small (significantly less than a millimeter) even at the low power-line frequency. For seawater, the power-frequency skin depth is also relatively small (about 35 m), and for a radio frequency of 1 MHz it is less than 25 cm. For copper, at 1 MHz the skin depth is less than one-tenth of a millimeter.

Example 20.2—Why not use cheap iron instead of expensive copper for distributing electric power? The skin depth for iron at 60 Hz in Table 20.1 answers an important question. If iron has a conductivity that is only about six times less than that of copper, and copper is much more expensive than iron, why do we not use iron wires to carry electric power to our homes? With the millions of kilometers of such wires, that would mean very large savings.

Unfortunately, due to its large relative permeability, iron has a very small skin depth at powerline frequency, so the losses in iron wire are large, outweighing the savings. Thus we have to use copper or aluminum.

TABLE 20.1 Skin depth (δ) for some common materials

Material	$f = 60$ Hz	$f = 10^3$ Hz	$f = 10^6$ Hz	$f = 10^9$ Hz
Copper	8.61 mm	2.1 mm	0.067 mm	2.11 μm
Iron	0.65 mm	0.16 mm	5.03 μm	0.016 μm
Seawater	32.5 m	7.96 m	0.25 m	7.96 mm
Wet soil	650 m	159 m	5.03 m	0.16 m

Example 20.3—Mutual inductance between cables laid on the bottom of the sea. Assume we have three single-phase 60-Hz cables laid at the bottom of the sea (for example, to supply electric power to an island). The cables are spaced by a few hundred meters and are parallel. (Three distant single-phase instead of one three-phase cable are often used for safety reasons: if a ship accidentally pulls and breaks one cable with an anchor, two are left. In addition, usually a spare single-phase cable is laid to enable quick replacement of a damaged one.) If the length of the cables is long (in practice, it can be many kilometers), we might expect very large mutual inductance between these cables, due to the huge loops they form. The skin depth of seawater at 60 Hz (Table 20.1), however, tells us that there will be practically *no* mutual inductance between the cables.

According to Eq. (20.7), with increasing y the current density changes not only in amplitude *but also in phase*. Thus, at a distance $y = \pi/k$ from the boundary surface the vector **J** has at all times *the opposite actual direction* to that near the boundary surface. The distribution of current density as a function of y at an instant is sketched in Fig. 20.1.

An important problem that we are now ready to solve is Joule's losses in the conductor per unit area of the boundary surface. Because we know the current density vector, one possibility is to integrate $[|J_z^2(y)|/\sigma]\,\mathrm{d}y$ from $y = 0$ to infinity, which is not too difficult to do. There is an easier way, however, that does not require integration but uses the concept of the Poynting vector. This derivation is given as problem P20.7 at the end of the chapter. Here we quote and discuss the final result of this derivation. It is found that the power of Joule's losses and the internal reactive power inside the conductor, per area S, are given by

$$P_J = \int_S R_s |H_0|^2 \mathrm{d}S = (P_{\text{reactive}})_{\text{internal}} \quad \text{(W)}, \tag{20.9}$$

(Evaluation of Joule's losses and reactive power in conductors at high frequencies)

where H_0 is the complex rms value of the tangential component of the vector **H** *on the conductor surface*. (By assumption, the normal component of **H** does not exist.) R_s is the *surface resistance* of the conductor, given by

$$R_s = \sqrt{\frac{\omega\mu}{2\sigma}} \quad (\Omega). \tag{20.10}$$

(Definition of surface resistance)

This formula for the surface resistance can be obtained if we consider the following rough approximation, illustrated on the square metal slab in Fig. 20.2. We assume that the entire high-frequency current is flowing uniformly over the cross

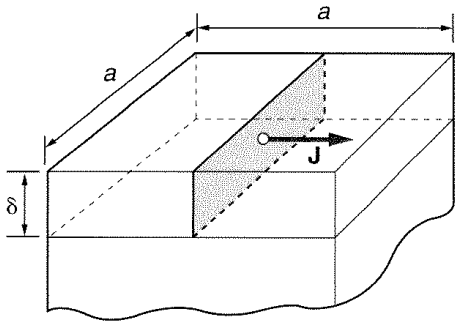

Figure 20.2 The surface resistance of a conductive slab with a uniform current flowing through a cross section δ deep and a wide

section defined by the skin depth and the width a of the conductor slab. Then the resistance of the slab is given by

$$R = \frac{1}{\sigma}\frac{a}{a\delta} = \frac{1}{\sigma}\sqrt{\frac{\omega\mu\sigma}{2}} = \sqrt{\frac{\omega\mu}{2\sigma}},$$

which can be obtained in an exact manner using the complex Poynting's vector.

Equation (20.9) is used to determine the attenuation in all metallic systems for guiding electromagnetic energy, like two-wire lines, coaxial lines, and rectangular waveguides. We illustrate this with two examples.

Example 20.4—Resistance and internal inductance per unit length of a cylindrical wire at high frequencies. Consider a straight round wire of radius a, conductivity σ, and permeability μ, carrying a sinusoidal current of angular frequency ω and with an rms value I. The magnetic field intensity on the wire surface is $H(0) = I/(2\pi a)$, so the Joule's losses per unit length of the wire, according to Eq. (20.9), are

$$P_J' = R_s \frac{I^2}{(2\pi a)^2} 2\pi a = R_s \frac{I^2}{2\pi a}.$$

Because the resistance per unit length is defined by the relation $P_J' = R'I^2$, we obtain that, at high frequencies,

$$R' = \frac{R_s}{2\pi a} \quad (\Omega/\text{m}). \tag{20.11}$$

(Resistance per unit length of round conductor at high frequencies)

According to Eq. (20.9), the reactive power at high frequencies inside the conductor per unit area is the same as the power of Joule's losses. We know from circuit theory that the internal reactive power per unit length of the wire can be expressed as $X_{\text{int}}' I^2$. The power in Eq. (20.9) refers to the field *inside the conductor* (i.e., the wire) *only*. Since it is positive, the internal reactance is inductive, that is, $X_{\text{int}}' = R' = \omega L_{\text{int}}'$. Therefore the *internal inductance* of the

wire at high frequencies per unit length is given by

$$L'_{\text{int}} = \frac{R'}{\omega} = \frac{R_s}{2\pi a\omega} \quad \text{(H/m)}. \tag{20.12}$$

(Internal inductance per unit length of round conductor at high frequencies)

This formula for L'_{int} was given in Table 18.1.

Example 20.5—Resistance and internal inductance per unit length of a thin two-wire line at high frequencies. Using the results from Example 20.4, it is a simple matter to calculate the resistance and internal inductance per unit length of a thin two-wire line. Let the line have conductors of radius a, and let the distance between the wires be much larger than a. Then the influence of the current in one wire on the current distribution inside the other can be neglected. That means that the current distribution in each wire is practically axially symmetric, as for a single wire in Example 20.4. Therefore, the resistance and internal inductance per unit length are just twice those calculated in the preceding example (because we have two wires). This is the formula given in Table 18.1.

Questions and problems: Q20.11 and Q20.12, P20.1 to P20.12

20.3 Proximity Effect

The term *proximity effect* refers to the influence of alternating current in one conductor on the current distribution in another, nearby conductor. Qualitatively, it can also be explained by Lentz's law.

Consider a coaxial cable of finite length. Assume for the moment that there is an alternating current only in the inner conductor (for example, that it is connected to a generator), and that the outer conductor is not connected to anything. If the outer conductor is much thicker than the skin depth, there is practically no magnetic field inside the outer conductor. If we apply Ampère's law to a coaxial circular contour contained in that conductor, it follows that the induced current on the *inside* surface of the outer conductor is exactly equal and opposite to the current in the inner conductor. This is an example of the proximity effect.

The current from the inner surface of the outer conductor must close into itself over the *outside* surface of the outer conductor, so that on that surface the same current exists as in the inner conductor.

Let us now, in addition, have normal cable current in the outer conductor. It is the same, but opposite, to the current on the conductor outer surface, so the two cancel out. We are left with a current over the inner conductor, and a current over the inside surface of the outer conductor. This is a combined skin effect and proximity effect. Normally, this *combined* effect is what is actually encountered, but it is usually called just the proximity effect.

If the skin effect is not pronounced, the situation is similar except that there is an appreciable current density at all points of the inner and outer cable conductors, as sketched in Fig. 20.3.

The analysis of the proximity effect (i.e., of the combined proximity effect and skin effect) is rather complicated. We shall not, therefore, illustrate the proximity ef-

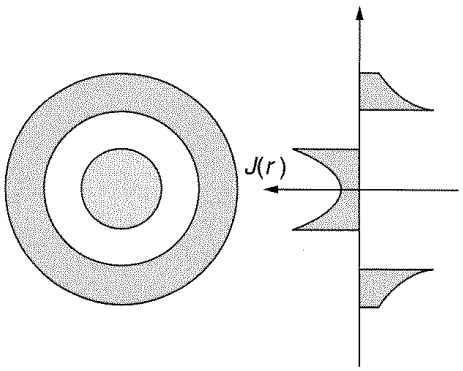

Figure 20.3 Combined proximity and skin effects in a coaxial cable

fect for the general case. If the skin effect is very pronounced, however, in some cases it becomes quite simple, as the next example shows.

Example 20.6—Resistance and internal inductance per unit length of a coaxial cable at high frequencies. Consider again a coaxial cable, with an inner conductor of radius a and an outer conductor of inner radius b. Assume that a sinusoidal current of rms value I flows through the cable at a frequency for which the skin effect is very pronounced. In this case, we have two thin current layers: one over the inner conductor, and one over the inside surface of the outer conductor, as explained previously. According to Eq. (20.9), Joule's losses per unit length of the cable (in both conductors) are given by the sum of losses in the cylinders of radius a and of radius b. Therefore, the resistance per unit length is the sum of that in Eq. (20.11) and of the same expression with a substituted by b:

$$R' = \frac{R_s}{2\pi a} + \frac{R_s}{2\pi b} = \frac{R_s}{2\pi}\left(\frac{1}{a} + \frac{1}{b}\right). \tag{20.13}$$

We know that the internal reactance per unit length has the same value as R'. The internal inductance of the cable at high frequencies is therefore $L'_{\text{int}} = R'/\omega$. These are the formulas given earlier in Table 18.1 without explanation.

20.4 Chapter Summary

1. Sinusoidal currents in good conductors are not distributed uniformly over their cross section. Rather, as frequency increases, the current tends to concentrate near the conductor surfaces, a phenomenon known as the *skin effect*.
2. At very high frequencies, the skin effect is so pronounced that current exists only over a very thin layer of any good (metallic) conductor.
3. The penetration of current in a good conductor is characterized by the *skin depth*. At this depth, the current density is $1/e \simeq 0.37$ of that at the conductor surface. At 60 Hz, it is on the order of 1 cm for copper and 1 mm for iron.
4. The skin depth is inversely proportional to the square root of frequency, permeability, and conductivity.

5. A time-varying current in one conductor influences the current distribution in nearby conductors, a phenomenon known as the proximity effect.
6. Both skin effect and proximity effect are consequences of electromagnetic induction.
7. In conducting magnetic materials, the time-varying magnetic field also exhibits the skin effect. For this reason, ferromagnetic cores of alternating-current machinery are made of thin, mutually insulated sheets. At very high frequencies, transformer and inductor cores are made of ferrites, which have a relatively high permeability, but are also relatively good insulators, so that the skin effect for the magnetic field almost does not exist.

QUESTIONS

Q20.1. Three long parallel wires a distance d apart are in one plane. At their ends they are connected together. These common ends are then connected by a large loop to a generator of sinusoidal emf. Are the currents in the three wires the same? Explain. [*Hint:* Have in mind Eq. (14.3), where $\mathbf{J}\,dv$ is substituted by $i\,\mathrm{d}\mathbf{l}$.]

Q20.2. N long parallel thin wires are arranged uniformly around a circular cylinder. At their ends the wires are connected by a large loop to a generator of sinusoidal emf. Are the currents in the N wires the same? Explain.

Q20.3. Another wire is added in question Q20.2 along the axis of the cylinder. Is the current in the added wire the same as in the rest? Is it smaller or greater? Explain, having in mind Eq. (14.3).

Q20.4. A thin metallic strip of width d carries a sinusoidal current of a high frequency. What do you expect the distribution of current in the strip to be like?

Q20.5. The two conductors of a coaxial line are connected in parallel to a generator of sinusoidal emf. Is the current intensity in the two conductors the same? If it is not, does the difference depend on frequency? Explain.

Q20.6. A metal coin is situated in a time-harmonic uniform magnetic field, with faces normal to the field lines. What are the lines of eddy currents in the coin like? What are the lines of the induced electric field of these currents?

Q20.7. So-called induction furnaces are used for melting iron by producing large eddy currents in iron pieces. Assume that the iron in the furnace is first in the form of small ferromagnetic objects (nails, screws, etc.). What do you expect to happen if they are exposed to a very strong time-harmonic magnetic field? What happens once they melt?

Q20.8. Two parallel, coplanar thin strips carry equal time-harmonic currents. What do you think the current distribution in the strips is like if the currents in the strips are (1) in the same direction, and (2) in opposite directions?

Q20.9. A thick copper conductor of square cross section carries a large time-harmonic current. Where do you expect the most intense Joule's heating of the conductor? Explain.

Q20.10. A ferromagnetic core of a solenoid is made of thin sheets. If the current in the solenoid is time-harmonic, where do you expect the strongest heating of the core due to eddy currents?

Q20.11. Describe the procedure of determining the resistance and internal inductance per unit length of a stripline at high frequencies. Neglect edge effect.

Q20.12. When compared with current density on the surface, what is the magnitude of current density in a thick conducting sheet one skin depth below the surface, and what is it at two skin depths below the surface?

PROBLEMS

P20.1. Check all skin depth values given in Table 20.1.

P20.2. Starting from Eq. (20.7), prove that the total current in the half-space in Fig. 20.1 is the same as if a current of constant density $J_z(0)/(1+\mathrm{j})$ exists in a slab $0 \leq y \leq \delta$.

P20.3. Determine the total Joule's losses per unit area of the half-space in Fig. 20.1 by integrating the density of Joule's losses. Compare the result with Eq. (20.9).

P20.4. Using Poynting's theorem in complex form, prove that for any conductor with two close terminals, at very high frequencies the conductor resistance and internal reactance are equal. Find the (integral) expression for these quantities.

P20.5. A stripline of strip width $a = 2\,\mathrm{cm}$, distance between them $d = 2\,\mathrm{mm}$, and the thickness of the strips $b = 1\,\mathrm{mm}$ carries a time-harmonic current of rms value $I = 0.5\,\mathrm{A}$ and frequency $f = 1\,\mathrm{GHz}$. The strips are made of copper. Neglecting fringing effect, determine the line resistance and total inductance per unit length.

P20.6. Starting from Eqs. (20.3), determine the distribution of current in a flat conducting sheet of thickness d. The sheet conductivity is σ, permeability μ, and angular frequency of the current is ω. Set the origin of the y coordinate at the sheet center, and assume that the rms value of the current density at the center is $J_z(0)$. Plot the resulting current distribution.

***P20.7.** Find $H_x(y)$ from Eqs. (20.3) and (20.7), and $E_z(y)$ from Eq. (20.7). Use these expressions and Poynting's theorem to prove Eq. (20.9).

P20.8. Starting from Eq. (20.7), derive the expression for the instantaneous value of the current density, $J_z(y, t)$.

P20.9. Calculate the resistance per unit length of a round copper wire of radius $a = 1\,\mathrm{mm}$, from the frequency for which the skin depth is one-tenth of the wire radius, to the frequency $f = 10\,\mathrm{GHz}$. Plot this resistance as a function of frequency.

P20.10. Assume that in a ferromagnetic round wire of radius a, conductivity σ, and permeability μ, there is an axial magnetic field of angular frequency ω and of rms flux density B practically constant over the wire cross section. Find the expressions for eddy currents in the wire and eddy current losses in the wire per unit length.

P20.11. A bunch of N insulated round wires of radius a, conductivity σ, and permeability μ is exposed to an axial time-harmonic magnetic field of angular frequency ω. The frequency is sufficiently low that the field can be considered uniform over the cross section of the wires. If the rms value of the magnetic flux density is B_0, determine the time-average eddy current power losses in the bunch, per unit volume *of the wires*. Use the result of the preceding problem. Specifically, calculate the losses per unit volume assuming $B_0 = 0.1\,\mathrm{T}$, $a = 0.5\,\mathrm{mm}$, $\sigma = 10^7\,\mathrm{S/m}$, $\mu = 1000\mu_0$, and $f = 60\,\mathrm{Hz}$.

***P20.12.** Consider a straight wire of radius a, conductivity σ, and permeability μ. Let the wire axis be the z axis of a cylindrical coordinate system. Assume there is a current in the wire of rms value I and angular frequency ω. Starting from Maxwell's equations in cylindrical coordinates, derive the differential equation for the only existing, J_z component of the current density vector in the wire. Note that, by symmetry, the only component of **H** is H_ϕ. Do *not* attempt to solve the equation you obtain. (If your equation is correct, it is known as a Bessel differential equation, and its solutions are known as Bessel functions.)

21

Uniform Plane Waves

21.1 Introduction

Maxwell's theory predicts the existence of specific electromagnetic fields, known as *electromagnetic waves*. These fields, once created by time-varying currents and charges, continue to move with a finite velocity independent of the sources that produced them.

Much of the rest of this book is devoted to the analysis of various types of electromagnetic waves. In this chapter the simplest type of wave, known as a *uniform plane wave*, is considered. Although they are the simplest, uniform plane waves are of extreme practical importance: actual waves radiating from sources are spherical, but at large distances from sources they become practically plane waves; and in addition, more complicated wave types can be represented as a superposition of plane waves.

21.2 The Wave Equation

The wave equation is a second-order partial differential equation that is satisfied by all electromagnetic fields in *homogeneous linear media*.

Assume that an electromagnetic field exists in a homogeneous linear medium with parameters ϵ, μ, and σ. Suppose that there are neither free charges nor field sources (impressed electric fields) in the medium considered. Maxwell's equations in

that case have the form

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t}, \qquad \nabla \cdot \mathbf{E} = 0, \tag{21.1}$$

$$\nabla \times \mathbf{H} = \sigma \mathbf{E} + \epsilon \frac{\partial \mathbf{E}}{\partial t}, \qquad \nabla \cdot \mathbf{H} = 0. \tag{21.2}$$

To eliminate $\mathbf{H}$ from the first equation pair, let us apply the curl operator to the first equation. Since the curl implies differentiation with respect to space coordinates, it is independent of the differentiation with respect to time. Therefore, $\nabla \times (\partial \mathbf{H}/\partial t)$ can be written as $\partial(\nabla \times \mathbf{H})/\partial t$. With this in mind, making use of the first of Eqs. (21.2), the first of Eqs. (21.1) becomes

$$\nabla \times (\nabla \times \mathbf{E}) = -\mu\sigma \frac{\partial \mathbf{E}}{\partial t} - \epsilon\mu \frac{\partial^2 \mathbf{E}}{\partial t^2}. \tag{21.3}$$

In a similar manner we eliminate $\mathbf{E}$ from the first equation of the second pair, to obtain

$$\nabla \times (\nabla \times \mathbf{H}) = -\mu\sigma \frac{\partial \mathbf{H}}{\partial t} - \epsilon\mu \frac{\partial^2 \mathbf{H}}{\partial t^2}. \tag{21.4}$$

We know from vector analysis that for any vector function $\mathbf{F}$ that can be differentiated twice, $\nabla \times (\nabla \times \mathbf{F}) = \nabla(\nabla \cdot \mathbf{F}) - \nabla^2 \mathbf{F}$ [see Appendix 2, No. 28]. According to the second parts of Eqs. (21.1) and (21.2), $\nabla \cdot \mathbf{E} = 0$ and $\nabla \cdot \mathbf{H} = 0$, so Eqs. (21.3) and (21.4) become

$$\nabla^2 \mathbf{E} - \epsilon\mu \frac{\partial^2 \mathbf{E}}{\partial t^2} - \mu\sigma \frac{\partial \mathbf{E}}{\partial t} = 0, \tag{21.5}$$

(Wave equation for vector $\mathbf{E}$*)*

and

$$\nabla^2 \mathbf{H} - \epsilon\mu \frac{\partial^2 \mathbf{H}}{\partial t^2} - \mu\sigma \frac{\partial \mathbf{H}}{\partial t} = 0. \tag{21.6}$$

(Wave equation for vector $\mathbf{H}$*)*

These are the *wave equations.*

If the field is time-harmonic and we use complex notation, we obtain

$$\nabla^2 \underline{\mathbf{E}} + (\omega^2 \epsilon\mu - \mathrm{j}\omega\mu\sigma)\underline{\mathbf{E}} = 0, \tag{21.7}$$

[Helmholtz equation (complex wave equation) for vector $\mathbf{E}$*]*

and

$$\nabla^2 \underline{\mathbf{H}} + (\omega^2 \epsilon\mu - \mathrm{j}\omega\mu\sigma)\underline{\mathbf{H}} = 0. \tag{21.8}$$

[Helmholtz equation (complex wave equation) for vector $\mathbf{H}$*]*

Although these are just wave equations in complex form, sometimes they are referred to as the *Helmholtz equations.*

Note that the wave equations were derived by using the conditions $\nabla \cdot \mathbf{E} = 0$ and $\nabla \cdot \mathbf{H} = 0$, but these conditions are *not* implicit in the wave equations. Because they must be satisfied, a solution must be sought using the following procedure:

1. We find a solution to the wave equation, for example, Eq. (21.5), that satisfies the condition $\nabla \cdot \mathbf{E} = 0$.
2. We determine vector $\mathbf{H}$ from the first of Eqs. (21.1).

The second step guarantees that also $\nabla \cdot \mathbf{H} = 0$ (recall that the divergence of the curl is zero). The first of Eqs. (21.2) is thereby also satisfied: the time derivative of that equation, with $\partial \mathbf{H}/\partial t$ from the first of Eqs. (21.1), becomes the wave equation for $\mathbf{E}$. Consequently, the solution found in this way satisfies all the necessary equations and is a legitimate solution of Maxwell's equations.

Example 21.1—Wave equation in a rectangular coordinate system. The wave equations (21.5) to (21.8) are valid in any coordinate system. What do they become in a rectangular coordinate system?

Note first that these equations are *vector equations*, that is, they represent *three scalar equations*. Consider, for example, the wave equation for vector $\mathbf{E}$, Eq. (21.5). Recalling the expression for $\nabla^2\mathbf{E}$ in a rectangular coordinate system, we obtain for the x component of the wave equation

$$\frac{\partial^2 E_x}{\partial x^2} + \frac{\partial^2 E_x}{\partial y^2} + \frac{\partial^2 E_x}{\partial z^2} - \epsilon\mu\frac{\partial^2 E_x}{\partial t^2} - \mu\sigma\frac{\partial E_x}{\partial t} = 0. \tag{21.9}$$

The y and z components of the equation are of exactly the same form, with E_x replaced by E_y and E_z, respectively.

Questions and problems: Q21.1 and Q21.2

21.3 Uniform Plane Electromagnetic Waves in Perfect Dielectrics

We now use the wave equation to analyze a specific electromagnetic field. Let the field satisfy the following two conditions:

1. Both vectors $\mathbf{E}$ and $\mathbf{H}$ depend only on coordinate z and time.
2. The field exists in a homogeneous, lossless medium, with parameters ϵ, μ, and $\sigma = 0$.

We first stipulate that $\nabla \cdot \mathbf{E} = 0$. Because in a rectangular coordinate system

$$\nabla \cdot \mathbf{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z},$$

and because we assumed $\mathbf{E}$ to depend only on z and time, the first two terms on the right-hand side are zero. So the condition $\nabla \cdot \mathbf{E} = 0$ reduces to $\partial E_z/\partial z = 0$. This is satisfied if $E_z = 0$, or possibly $E_z =$ constant. The constant solution (with respect to z) is not of interest because we search for a field varying along the z axis, so $E_z = 0$.

No generality is lost if we assume that only one of the two components normal to z, E_x or E_y, is nonzero. Let E_x be nonzero, and $E_y = 0$. Then the wave equation

for the vector **E** has only an x component, so that Eq. (21.9) is the only one existing. Because the derivatives with respect to x and y are zero (field components depend on the coordinate z only), and $\sigma = 0$ for a perfect dielectric, Eq. (21.9) becomes

$$\frac{\partial^2 E_x}{\partial z^2} - \epsilon\mu \frac{\partial^2 E_x}{\partial t^2} = 0. \tag{21.10}$$

This equation has the same form as Eqs. (18.5) that we derived for the voltage and current along a transmission line. Notice that instead of $L'C'$ in Eqs. (18.5), we now have $\epsilon\mu$, and we know these products are equal for lossless lines. As we have seen in Chapter 19, the solution to this equation is of the form

$$E_x(z, t) = E_1 f_1\left(t - \frac{z}{c}\right) + E_2 f_2\left(t + \frac{z}{c}\right). \tag{21.11}$$

(E field consisting of incident and reflected plane waves)

In this equation,

$$c = \frac{1}{\sqrt{\epsilon\mu}} \tag{21.12}$$

(Velocity of propagation of plane waves)

is the velocity of propagation of plane waves, E_1 and E_2 are constants, and f_1 and f_2 are *any* functions of the arguments $(t - z/c)$ and $(t + z/c)$, respectively. That the expression in Eq. (21.11) is the solution of Eq. (21.10) can be proved by substitution, as we did in Chapter 18. In fact, there are an infinite number of solutions to the wave equation (infinite number of different fields), since $f_1(t - z/c)$ and $f_2(t + z/c)$ are *arbitrary* functions.

Example 21.2—Some specific wave functions. Let us construct a few specific wave functions. For this, we need to consider *any* function of a single variable, and to replace this variable by $(t \pm z/c)$. For example, consider the function $\sin \omega t$. The corresponding wave function is $\sin \omega(t \pm z/c)$. A wave function $e^{\pm j\omega(t \pm z/c)}$ corresponds to the function $e^{\pm j\omega t}$. As a final example, consider the following function:

$$f(x) = 1 \quad \text{for} \quad a < x < b \quad \text{else} \quad f(x) = 0.$$

The corresponding wave function, for example of the form $f(t - z/c)$, is

$$f(t - z/c) = 1 \quad \text{for} \quad a < (t - z/c) < b \quad \text{else} \quad f(t - z/c) = 0.$$

We know from Chapter 18 what the physical meaning of the functions $f_1(t - z/c)$ and $f_2(t + z/c)$ is: $f_1(t - z/c)$ is an incident (forward) traveling (electric field) wave, and $f_2(t + z/c)$ is a reflected (backward) traveling wave (with respect to the z axis). This important conclusion for $f_1(t - z/c)$ is illustrated again in Fig. 21.1. Such moving fields, as already mentioned, are known as *electromagnetic waves*. Note again that the

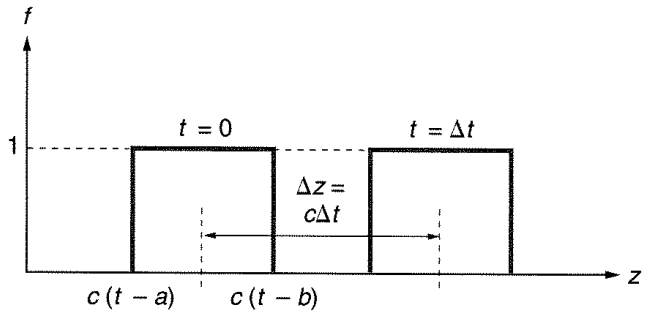

Figure 21.1 A pulse at $t = 0$ and at a later instant Δt, as described in Example 21.2

E-field component satisfies the same equation and that this equation has the same solutions as that for the voltage along a transmission line.

We have thus determined the electric field of a wave. Let us now determine the magnetic field of the wave, to obtain a complete solution to Maxwell's equations. This is analogous to determining the current along a transmission line from the transmission-line equations, and then finding the characteristic impedance as a ratio of voltage and current at a point. Again, the analogy is not surprising, as the magnetic field is produced by the current.

To simplify the derivation, consider only the wave with the argument $(t - z/c)$. We proceed as outlined in the preceding section and determine $\mathbf{H}$ from the first of Eqs. (21.1). Having in mind the expression for the curl in a rectangular coordinate system, and that $\mathbf{E}(z, t) = E_x(z, t)\mathbf{u}_x$, this *vector* equation results in the following three scalar equations:

$$0 = -\mu\frac{\partial H_x}{\partial t} \qquad \frac{\partial E_x}{\partial z} = E_1\frac{\partial f_1}{\partial z} = -\mu\frac{\partial H_y}{\partial t} \qquad 0 = -\mu\frac{\partial H_z}{\partial t}. \tag{21.13}$$

We are not interested in time-constant field components (the components having zero time derivative), so $H_x = H_z = 0$. The only existing time-varying component of the magnetic field intensity, H_y, is obtained by integrating the second of Eqs. (21.13). Having in mind Eq. (18.9), we obtain

$$H_y(z, t) = \frac{1}{\mu c}E_1 f_1\left(t - \frac{z}{c}\right) = \frac{1}{\mu c}E_x(z, t), \tag{21.14}$$

or, since $c = 1/\sqrt{\epsilon\mu}$,

$$H_y(z, t) = \sqrt{\frac{\epsilon}{\mu}}E_x(z, t). \tag{21.15}$$

(H_y component associated with the E_x component of incident plane wave)

An interesting conclusion follows from this relation: for a plane wave, the energy density of the magnetic field, $\mu H^2/2$, at all points and at all instants, equals the energy

density of the electric field, $\epsilon E^2/2$. Of course, this is true only for a single incident plane wave, not for a possible superposition of waves.

The electric and magnetic field vectors of the electromagnetic wave we considered are in planes perpendicular to the direction of propagation of the wave (the z direction). This is why the wave is known as a *plane wave*. In addition, these vectors are constant at any of these planes at a given instant. For this reason it is said that this particular plane wave is *uniform*. (Some plane waves are not uniform—for example, the electric and magnetic field waves along transmission lines are plane, but not uniform. Why?) Finally, since the vectors **E** and **H** are transverse to the direction of propagation, this kind of wave is known as a *transverse electromagnetic wave*, or *TEM wave*.

It is a simple matter to show that the unit for $\sqrt{\mu/\epsilon}$ is the ohm, that is, this quantity has the dimension of impedance. For this reason it is known as the *intrinsic impedance* (or sometimes just *impedance*) of the medium in which the wave exists. Note that Eq. (21.15) is analogous to Eq. (18.19), where the characteristic impedance of a transmission line was given by $\sqrt{L'/C'}$. The intrinsic impedance of a plane wave is usually denoted by η (or sometimes Z),

$$\eta = \sqrt{\frac{\mu}{\epsilon}} \quad (\Omega). \tag{21.16}$$

(Intrinsic impedance of the medium)

The most important (and frequent) waves in practice are those propagating in a vacuum (or air). In that case, the intrinsic impedance of the medium and the velocity of propagation become

$$\eta_0 = \sqrt{\frac{\mu_0}{\epsilon_0}} \simeq 120\pi\ \Omega \simeq 377\ \Omega \tag{21.17}$$

(Intrinsic impedance of a vacuum)

$$c_0 = \frac{1}{\sqrt{\epsilon_0 \mu_0}} \simeq 3 \cdot 10^8\ \text{m/s}. \tag{21.18}$$

(Velocity of propagation of plane waves in a vacuum)

Thus, *the velocity of propagation of plane waves in a vacuum is equal to the velocity of light in a vacuum*. This fact was the basis on which Maxwell argued that light is nothing but a rapidly oscillating electromagnetic wave. This mathematically obtained number was the basis for Maxwell's electromagnetic theory of light.

Note that the cross product of the vectors **E** and **H**, that is, the Poynting vector, is in the direction of propagation of the wave:

$$\mathbf{E} \times \mathbf{H} = E_x(z, t)H_y(z, t)\mathbf{u}_z = \mathcal{P}(z, t)\mathbf{u}_z. \tag{21.19}$$

This means that the wave *transports electromagnetic energy*.

If the E field has a y component (instead of an x component), the Poynting theorem tells us that the H field will have a $-x$ component, since the cross product $\mathbf{E} \times \mathbf{H}$ must be in the $+z$ direction (the direction of propagation of the wave). The relation between the two is

$$H_x(z, t) = -\sqrt{\frac{\epsilon}{\mu}}E_y(z, t). \tag{21.20}$$

This conclusion is important for understanding reflections of plane waves.

Example 21.3—Plane waves propagating in the $-z$ direction. Consider now the other solution of the wave equation, $E_x(z, t) = E_2 f_2(t + z/c)$. To determine the corresponding vector **H**, we just replace c by $-c$ in the preceding derivations. So for the "reflected" wave

$$H_y(z, t) = -\sqrt{\frac{\epsilon}{\mu}}E_x(z, t),$$

and

$$H_x(z, t) = \sqrt{\frac{\epsilon}{\mu}}E_y(z, t).$$

It is left as an exercise for the reader to prove that the Poynting vector in both of these cases is directed in the $-z$ direction.

Let us summarize the properties of a uniform plane electromagnetic wave:

1. The vectors **E** and **H** are mutually perpendicular, and perpendicular to the direction of wave propagation.
2. The direction of the Poynting vector is in the direction of wave propagation.
3. At any instant, the magnitudes of vectors **E** and **H** are the same in any individual plane normal to the direction of propagation.
4. For a single plane wave (but not for a wave obtained as a superposition of several waves propagating in arbitrary directions), the ratio of the magnitudes of vectors **E** and **H** is the same at all points and at all instants. It is equal to the intrinsic impedance of the medium in which the wave exists, $\eta = \sqrt{\mu/\epsilon}$. In air and vacuum, $\eta_0 \simeq 120\pi\ \Omega \simeq 377\ \Omega$.
5. The velocity of propagation of plane waves is $c = 1/\sqrt{\epsilon\mu}$; in a vacuum, this is equal to the velocity of light in a vacuum, $c_0 \simeq 3 \cdot 10^8$ m/s.

Questions and problems: Q21.3 and Q21.4, P21.1 and P21.2

21.4 Time-Harmonic Uniform Plane Waves and Their Complex Form

Electromagnetic waves in electrical engineering are most often harmonic in time, or at least harmonic during a certain time interval. Suppose a wave at a fixed coordinate z varies in time as $\cos \omega t$. In that case, for a wave propagating in the $+z$ direction we have

$$E_x(z,t) = E\sqrt{2}\cos\omega(t - z/c) \qquad H_y(z,t) = \sqrt{\frac{\epsilon_0}{\mu_0}}E_x(z,t), \tag{21.21}$$

(Electric and magnetic fields of a sinusoidal plane wave)

where E is the rms value of the electric field.

We know that $\cos(\alpha + n \cdot 2\pi) = \cos\alpha$ for all integer values of n. This means that the values of the electric and magnetic fields periodically repeat themselves in *time and space*. Namely, for a given $z = z_0$, the fields are changing according to $\cos\omega(t - z_0/c)$. On the other hand, for a given $t = t_0$, the fields change along the z axis according to

$$E_x(z, t_0) = E\sqrt{2}\cos(\omega t_0 - \omega z/c). \tag{21.22}$$

$E_x(z, t_0)$ has the same value at all points for which

$$|\omega t_0 - \omega z_n/c| = n \cdot 2\pi \qquad n = 0, 1, 2, \ldots. \tag{21.23}$$

The distance along the z axis between two such points z_{n+1} and z_n is

$$\lambda = \frac{2\pi c}{\omega}. \tag{21.24}$$

We know from Chapter 18 that this distance is called the *wavelength* of the time-harmonic (sinusoidal) plane wave. Since $\omega = 2\pi f$, where f is the frequency of the wave, the wavelength can also be expressed as

$$\lambda = \frac{c}{f} \quad \text{(m)}, \tag{21.25}$$

(Wavelength of plane waves)

which we already know from Chapter 18.

Figure 21.2 shows the way an electromagnetic field changes in space (along the z axis) when frozen in time. In time, the whole picture moves in the direction of the z axis with a velocity c.

Time-harmonic electromagnetic waves used in engineering and physics have frequencies in a very wide range—from about 1 Hz to about 10^{22} Hz. This corre-

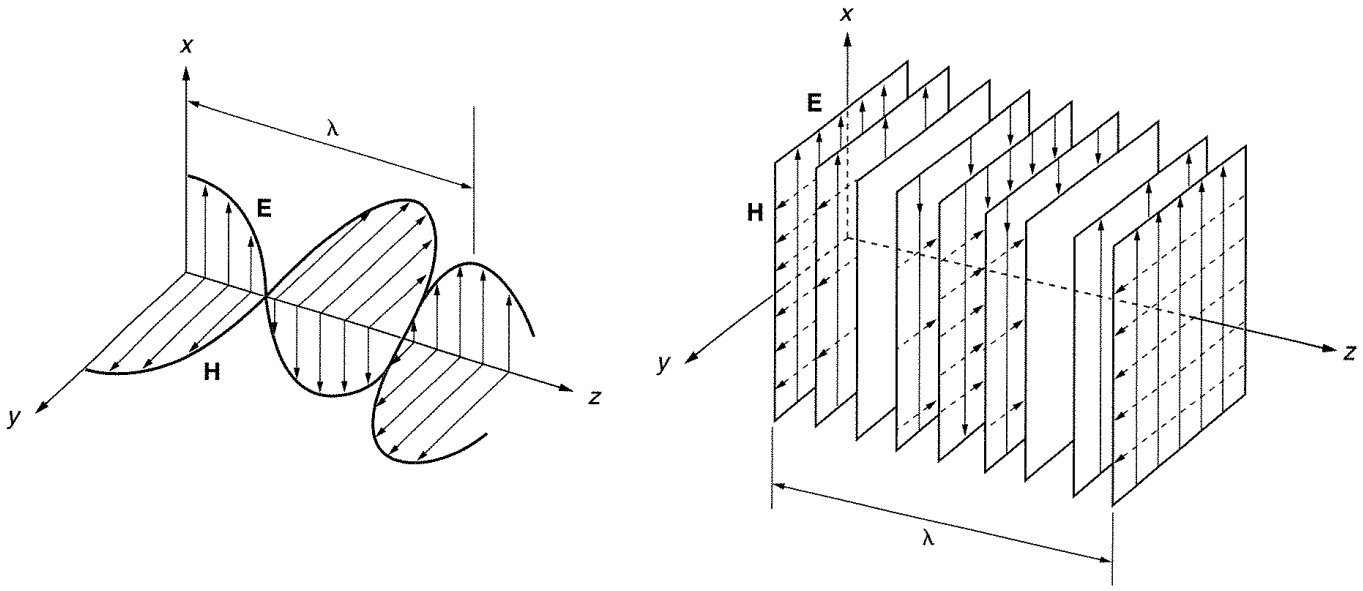

Figure 21.2 Two schematic ways of representing a time-harmonic incident plane wave frozen in time. As time passes, the picture moves in the $+z$ direction with a velocity c.

sponds to wavelengths from about 10^8 m to about 10^{-14} m. This range of frequencies is known as the *electromagnetic spectrum*. It is sketched in Fig. 21.3.

We can always determine the wavelength of a plane wave of a given frequency (or the frequency of a wave of a given wavelength) by means of the formula in Eq. (21.25). Electronics engineers usually use frequency, optical engineers use wavelength, and both are used in the microwave frequency range (from about 1 to 300 GHz). For the radio frequency (rf) and microwave region and for waves in a vacuum, the following simple rule is convenient to use: the wavelength in centimeters is equal to 30 divided by the frequency in GHz. For example, the wavelength at 10 GHz is 3 cm, and the wavelength at 900 MHz = 0.9 GHz is $33\frac{1}{3}$ cm.

Time-harmonic electromagnetic waves can be represented in complex form, as in the case of time-harmonic currents and voltages. We know that in the case of cur-

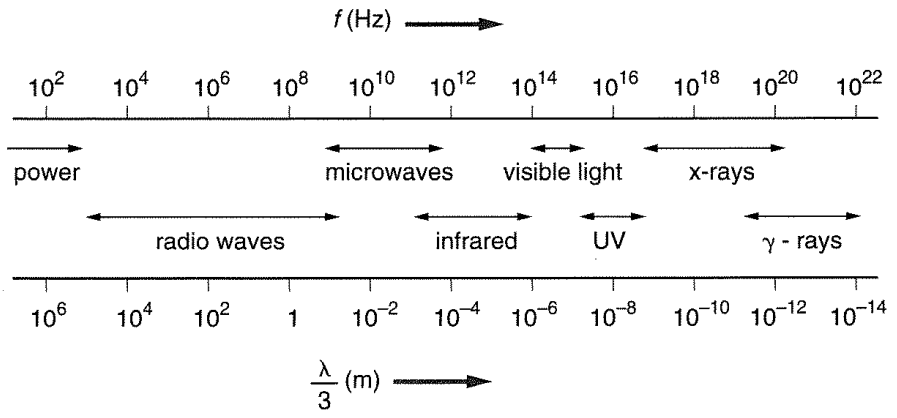

Figure 21.3 The electromagnetic spectrum

rents and voltages their form in time domain is obtained by multiplying their complex form by $\sqrt{2}e^{j\omega t}$, and taking the real part of the expression obtained in this way.

Let us write the expression in Eq. (21.21) for the E field of a plane wave as

$$E_x(z, t) = E\sqrt{2}\cos(\omega t - \beta z + \theta), \tag{21.26}$$

where

$$\beta = \frac{\omega}{c} = \frac{2\pi}{\lambda} \quad \text{(radian/m)}, \tag{21.27}$$

(Phase coefficient for plane waves)

as earlier, is termed the *phase coefficient*, and θ is the initial phase of the field [which in Eq. (21.21) was assumed to be zero]. (Because for various media in which plane waves may propagate the phase coefficient may *not* be constant, the term "phase constant" for waves is not quite appropriate.) The complex form of the expression in Eq. (21.26) should be such that we obtain Eq. (21.26) from it in the same way we obtain time-domain forms of currents and voltages from their complex forms:

$$E_x(z, t) = \sqrt{2}\,\mathrm{Re}\left\{\underline{E}_x(z)e^{j\omega t}\right\}. \tag{21.28}$$

By comparing the last expression with that in Eq. (21.26), we identify immediately the complex form of the E field (and H field) of a plane wave propagating along the z axis:

$$\underline{E}_x(z) = \underline{E}e^{-j\beta z} \qquad \underline{H}_y(z) = \frac{1}{\eta}\underline{E}e^{-j\beta z} \qquad \text{where} \qquad \underline{E} = Ee^{j\theta}. \tag{21.29}$$

Thus, as we already know from transmission lines, a factor $e^{-j\beta z}$ in the complex form of a quantity indicates propagation of that quantity in the $+z$ direction with a velocity $c = \omega/\beta$. Evidently, a factor of the form $e^{+j\beta z}$ indicates propagation in the $-z$ direction.

Questions and problems: Q21.5 to Q21.11, P21.3 to P21.10

21.5 Polarization of Plane Waves

Any number of plane waves propagating in the same direction add up to a complex plane wave. The time variation of the component waves may be arbitrary. Fortunately, such a general case is of quite minor engineering interest. We are mostly interested in time-harmonic plane waves. For these simple uniform plane waves, the vector **E** is at all times parallel to an axis (the x axis in our case). It does vary in time, but the tip of the vector traces a line parallel to the x axis. We say that the polarization of this wave is *linear*.

What happens if we have two otherwise arbitrary plane waves, whose frequencies are the same, propagating in the same direction? We shall now show that the resulting wave is a plane wave with the tip of vector **E** tracing an ellipse at every point in space. We say that such a resulting wave is *elliptically polarized*. In the special case when the major and minor semi-axes of the ellipse are of equal lengths, the tip of vector **E** at a fixed point in space traces a circle. We say that the polarization of this wave is *circular*.

To demonstrate elliptic polarization of plane waves, consider two plane waves of the same frequency and propagating in the same direction, but with vectors **E** of the two waves perpendicular to each other and of different amplitudes. Let the directions of the two E vectors be the x and y direction, and let one of them vary as the cosine function, and the other as the sine function:

$$E_x(z, t) = E_1 \cos(\omega t - \beta z), \tag{21.30}$$

$$E_y(z, t) = E_2 \sin(\omega t - \beta z). \tag{21.31}$$

In this simple case

$$\left[\frac{E_x(z, t)}{E_1}\right]^2 + \left[\frac{E_y(z, t)}{E_2}\right]^2 = 1, \tag{21.32}$$

since $\cos^2 \alpha + \sin^2 \alpha = 1$.

For a fixed z, this is the equation of an *ellipse* with semi-axes E_1 and E_2. This means that indeed, the tip of the total electric field intensity vector, $\mathbf{E}_{\text{tot}}(z, t) = E_x(z, t)\mathbf{u}_x + E_y(z, t)\mathbf{u}_y$, for any fixed z, in the course of time describes this ellipse.

If $E_1 = E_2$, the tip of $\mathbf{E}_{\text{tot}}(z, t)$, for any fixed z, describes a circle of radius E_1, so the total field is circularly polarized.

Figure 21.4 illustrates elliptic, circular, and linear polarizations. A representation of an elliptically polarized incident plane wave frozen in time is shown in Fig. 21.5. In the figure, the density of the E lines (solid) and H lines (dashed) is proportional to the local intensity of these vectors. (Note the different density of lines along the z axis.) The whole picture moves with a velocity c_0 in the $+z$ direction.

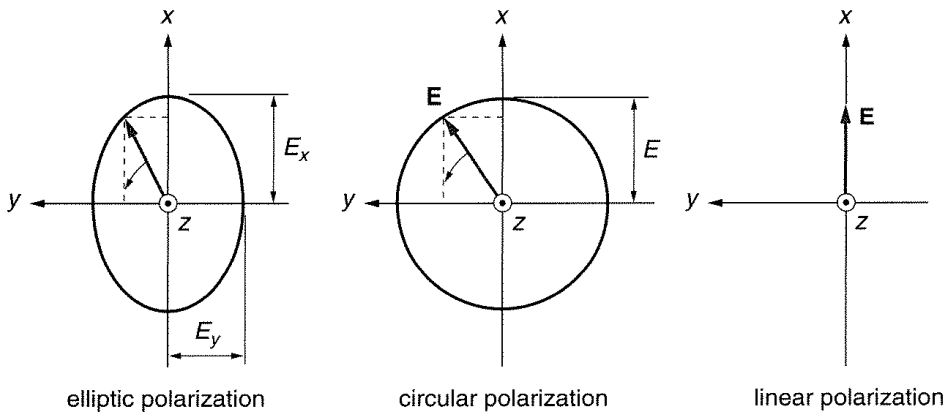

Figure 21.4 Illustration of elliptic, circular, and linear polarizations

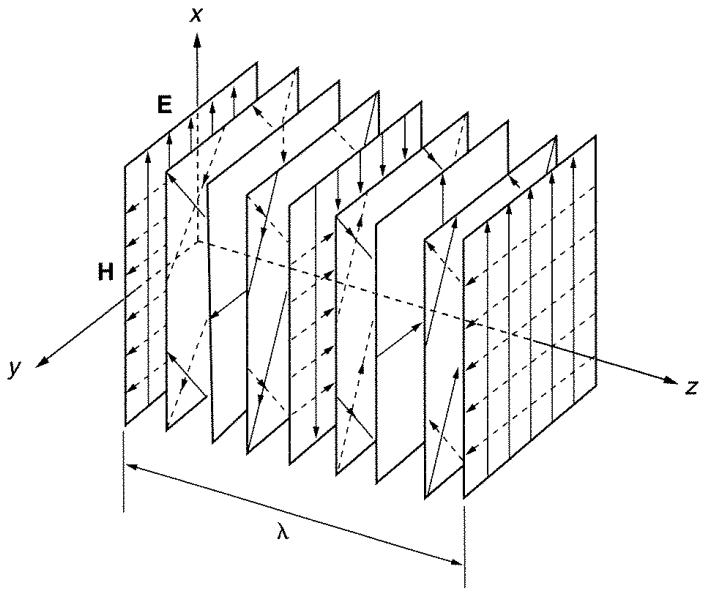

Figure 21.5 A representation of an elliptically polarized plane wave

So, *for a fixed z*, the vector **E** (and, of course, also the vector **H**) rotates and changes the magnitude, its tip describing an ellipse. This rotation can be in one or the other direction. If the direction of rotation is given by the right-hand rule with respect to the direction of wave propagation, it is said that the polarization of the wave is *right-handed*. If the field vectors rotate in the opposite direction, the polarization is said to be *left-handed*.

Questions and problems: Q21.12 to Q21.17, P21.11 to P21.15

21.6 Phase Velocity and Group Velocity: Dispersion

We know that the velocity of propagation of uniform plane electromagnetic waves is $c = 1/\sqrt{\epsilon\mu}$. In a vacuum, this velocity equals the velocity of light, and is the same for any frequency of the wave. In other media, the permittivity and permeability depend on frequency at least to some extent. Therefore, except in a vacuum, the phase coefficient, and therefore also the velocity of propagation of uniform plane waves, depend on the wave frequency. This is known as *dispersion*.

In many media, dispersion can be ignored. In quite a number of important cases, however, it must be taken into account, as it results in signal distortion. Namely, any signal is composed of time-harmonic components contained in a certain frequency band. Its shape is determined by relative amplitudes *and phases* of the time-harmonic components in this band. If the velocities of the time-harmonic components are not the same, their relative positions, which means their relative phases, change as the signal propagates, which means that the signal shape changes

as well. For this reason, the frequency band of a signal is usually made small enough so that distortion can be ignored.

The velocity c of plane waves in Eq. (21.21) is the velocity with which the *phase* of the wave propagates. To be more specific, it determines the progression of the z coordinates in the argument of the cosine function, which ensures that as time passes, the argument (i.e., the phase) of the cosine function remains unchanged. For this reason, the velocity c is termed the *phase velocity*. According to Eq. (21.27), the phase velocity, $v_{\text{ph}} = c$, can be expressed as

$$v_{\text{ph}}(\omega) = \frac{\omega}{\beta(\omega)} \quad \text{(m/s)}. \tag{21.33}$$

(Definition of phase velocity)

There is another important concept connected to dispersion, known as the *group velocity*. It represents the velocity of the signal in a dispersive medium and can be defined only for cases where dispersion is small.

To determine the group velocity, consider a simple signal in a weakly dispersive medium. Let the signal be obtained as a superposition of two plane waves propagating in the same direction and with slightly different angular frequencies, ω_1 and ω_2, and slightly different phase coefficients, β_1 and β_2 (due to dispersion). Without loss of generality, we can assume the amplitudes of both waves are the same, for example equal to 1, and consider a signal $f(z, t)$ of the form

$$f(z, t) = \cos(\omega_1 t - \beta_1 z) + \cos(\omega_2 t - \beta_2 z). \tag{21.34}$$

Now, $\cos a_1 + \cos a_2 = 2\cos[(a_2 - a_1)/2]\cos[(a_2 + a_1)/2]$, so that

$$f(z, t) = 2\cos(\Delta\omega\, t - \Delta\beta_1 z)\cos(\omega t - \beta z), \tag{21.35}$$

where

$$\Delta\beta = \frac{\beta_2 - \beta_1}{2}, \qquad \Delta\omega = \frac{\omega_2 - \omega_1}{2}, \qquad \omega = \frac{\omega_1 + \omega_2}{2}, \qquad \beta = \frac{\beta_1 + \beta_2}{2}. \tag{21.36}$$

Since $\Delta\omega \ll \omega$, the shape of this signal frozen in time is as sketched in Fig. 21.6. This is a rapidly varying wave (of frequency ω) modulated by a slowly varying wave (of frequency $\Delta\omega$). The velocity of propagation of the rapidly varying modulated wave (the solid line in Fig. 21.6) is simply ω/β. The velocity of the modulating wave (signal), however, is different—it is equal to the velocity of the *envelope* of the rapidly varying wave, indicated by the dashed lines. This means that the group velocity, v_g, is given by

$$v_g = \frac{\Delta\omega}{\Delta\beta} = \frac{1}{\Delta\beta/\Delta\omega}. \tag{21.37}$$

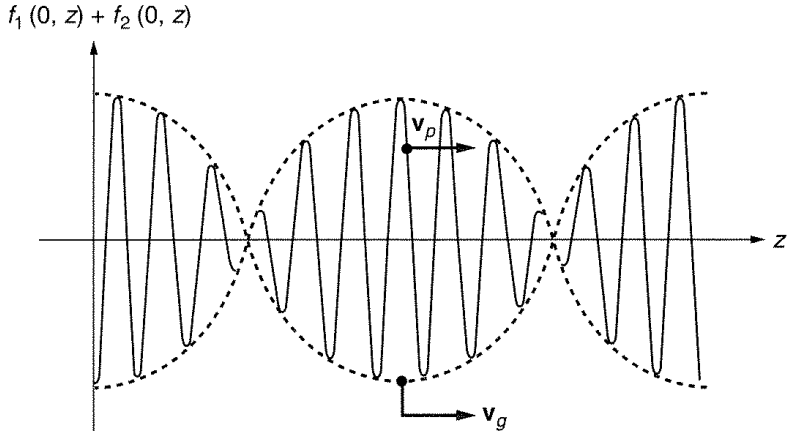

Figure 21.6 Sum of two time-harmonic functions of equal amplitudes and slightly different frequencies

What is the meaning of this expression? We know that β is a function of ω. For example, for a homogeneous dispersive dielectric of parameters $\epsilon(\omega)$ and $\mu(\omega)$ we know that $\beta(\omega) = \omega\sqrt{\epsilon(\omega)\mu(\omega)}$. Since $\Delta\beta$ is a small difference in β corresponding to a small difference in ω, the expression in the denominator in the last expression is in fact the derivative $\mathrm{d}\beta(\omega)/\mathrm{d}\omega$. We finally have

$$v_g(\omega) = \frac{1}{\mathrm{d}\beta(\omega)/\mathrm{d}\omega} \qquad \textit{(group velocity)}. \tag{21.38}$$

Because the envelope in Fig. 21.6 represents a signal (this is called *amplitude modulation*, or AM), the *group velocity* is the velocity of propagation of information transmitted by electromagnetic waves.

Example 21.4—Group velocity in nondispersive media. Assume that the medium is not dispersive. In that case, $\beta(\omega) = \omega/c$. The formula in Eq. (21.38) for the group velocity thus yields $v_g = c$. Of course, this was expected because in nondispersive media all time-harmonic components propagate with the same velocity, and therefore the group velocity is the same as the phase velocity.

Example 21.5—Phase velocity and group velocity in ionized gases and hollow metal waveguides. We will show in Chapters 23 and 25 that the phase coefficient for both a hollow metal waveguide and a homogeneous ionized gas is of the form

$$\beta(\omega) = \frac{\omega}{c_0}\sqrt{1 - \frac{{\omega_c}^2}{\omega^2}},$$

where ω_c is a constant that depends on the medium through which the wave propagates. The phase velocity is given by Eq. (21.33):

$$v_{\mathrm{ph}}(\omega) = \frac{\omega}{\beta(\omega)} = \frac{c_0}{\sqrt{1 - {\omega_c}^2/\omega^2}} \quad (\mathrm{m/s}). \tag{21.39}$$

The group velocity is obtained from Eq. (21.38). The final result is

$$v_g(\omega) = \frac{1}{\mathrm{d}\beta(\omega)/\mathrm{d}\omega} = c_0\sqrt{1 - \frac{{\omega_c}^2}{\omega^2}} \quad \text{(m/s)}. \tag{21.40}$$

So we see that the group velocity is less than c_0 (the velocity of light in a vacuum), but the phase velocity is larger than c_0! How this can be, when we know that c_0 is the largest possible velocity? Note that the phase velocity is a purely *geometrical* velocity, not the velocity of a particle or of a wave, so it can have any value, even larger than c_0. However, a wave does not transport power or information at the phase velocity, but rather at the group velocity, which is always smaller than c_0.

21.7 Chapter Summary

1. Maxwell's equations predict the existence of a specific type of electromagnetic field that, once created by time-varying currents and charges, continues to exist with no connection whatsoever with its sources. Such fields are known as *electromagnetic waves.*
2. The simplest electromagnetic waves are uniform plane waves. Their electric and magnetic field vectors are normal to each other and to the direction of propagation, and constant in planes normal to that direction. They are therefore known as *uniform transverse electromagnetic (TEM) waves.*
3. The speed of plane waves equals $1/\sqrt{\epsilon\mu}$, which in a vacuum equals the speed of light.
4. The ratio of the electric and magnetic fields at any point of a plane wave, and at all instants, is a constant, equal to the intrinsic impedance of the medium, $\eta = \sqrt{\mu/\epsilon}$.
5. If the properties of a medium depend on frequency, it is called a *dispersive medium.* For a signal composed of a narrow frequency band, and if dispersion is small, the signal propagates with a velocity known as the *group velocity.* The group velocity is different from the phase velocity, which is a geometrical velocity with which a fixed phase of the wave propagates.
6. The tip of the electric field vector of a time-harmonic field at a fixed point in space may trace in the course of time a straight-line segment (linear polarization), a circle (circular polarization) or an ellipse (elliptic polarization). No other traces are possible for a time-harmonic field of a single frequency.
7. The circular or elliptic polarizations are said to be right-handed if the rotation is clockwise, looking in the direction of the wave propagation. For waves rotating in the opposite direction, the polarization is said to be left-handed.

QUESTIONS

Q21.1. What would Eqs. (21.1) and (21.2) be like for an inhomogeneous perfect dielectric? Would it be possible to obtain the wave equation in that case?

Q21.2. Derive the Helmholtz equations, (21.7) and (21.8), from the wave equations, (21.5) and (21.6).

Q21.3. Write the expressions for at least three functions representing forward and backward traveling waves.

Q21.4. Is a plane electromagnetic wave with a component of the electric or magnetic field in the direction of propagation possible? Explain.

Q21.5. A perfect dielectric medium is not homogeneous, but ϵ is a smooth function of position, $\epsilon = \epsilon(x, y, z)$. Is a uniform plane wave possible in such a medium? Explain.

Q21.6. Write Eq. (21.10) in complex form and find its solutions.

Q21.7. What is the ratio of the wavelengths of a sinusoidal plane wave of frequency f if it propagates in perfect dielectrics of permittivities ϵ_1 and ϵ_2, and permeability μ_0?

Q21.8. What is the wavelength in a vacuum corresponding to the following frequencies of a plane wave: (1) 60 Hz, (2) 10 kHz, (3) 1 MHz, (4) 100 MHz, (5) 1 GHz, (6) 10 GHz, (7) 100 GHz, (8) 300 THz?

Q21.9. Does the expression $\mathbf{E} = E_1 \cos(\omega t - \beta z)\mathbf{u}_z$ represent a possible electric field of a plane wave? Explain.

Q21.10. Does the expression $\mathbf{E} = E_1 e^{j\beta x}\mathbf{u}_z$ represent a possible phasor expression for the electric field of a plane wave? Explain.

Q21.11. A circular loop of radius a is situated in the field of a plane electromagnetic wave of wavelength $\lambda = a$. Is it possible in principle to evaluate the emf induced in the loop? If you think it is, can it be used for the evaluation of current intensity in the loop by means of circuit theory? Explain.

Q21.12. Does the concept of linear wave polarization make sense if the wave is not time-harmonic? What about the concept of circular and elliptical polarization? Explain.

Q21.13. A time-harmonic, linearly polarized plane wave propagates along the z axis. Located along the z axis is a row of small free charges. How do the charges move in time? Sketch their approximate position over a few time intervals.

Q21.14. Repeat question Q21.13 assuming that the polarization of the wave is circular.

Q21.15. What is the complex representation of the electric field of an elliptically polarized plane wave defined by Eqs. (21.30) and (21.31)?

Q21.16. Assuming that the wave propagates into the paper, is the polarization of the wave represented in the two sketches in Fig. Q21.16 right-handed or left-handed? Explain.

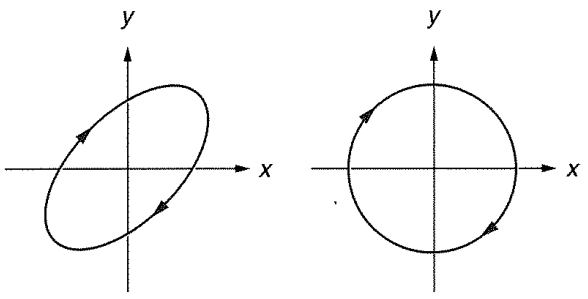

Figure Q21.16 Elliptic and circular polarization

Q21.17. Is the polarization of the wave represented in Fig. Q21.17 right-handed or left-handed? Explain.

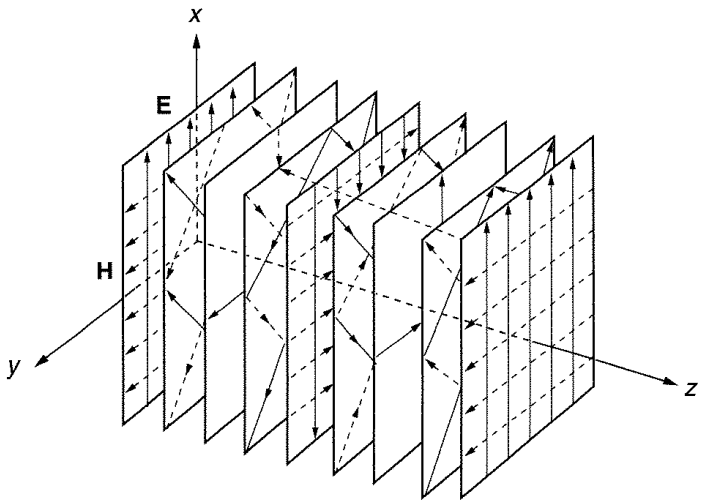

Figure Q21.17 Elliptically polarized wave

PROBLEMS

P21.1. Assuming an E_y component of the electric field, *derive* the corresponding component of the magnetic field in Eq. (21.20), following the same procedure as when E_x was assumed.

P21.2. Prove that for a plane wave with both E_x and E_y components the vector **E** is normal to vector **H**. Evaluate the Poynting vector in that case.

P21.3. Repeat the entire derivation of the plane waves for time-harmonic plane waves and starting from complex forms of all the equations.

P21.4. A time-harmonic plane wave with an rms value of the electric field vector $E = 10\,\text{mV/m}$ propagates in a vacuum, and is normally incident on a screen that totally absorbs the energy of the wave. Find the absorbed energy per square meter of the screen in one hour.

P21.5. By measurements it was found that the time-average power of the sun's radiation on the surface of the earth is about $1.35\,\text{kW/m}^2$, for normal incidence of the plane waves from the sun. This radiation is composed of a very wide band of frequencies, and the components of different frequencies are generally polarized elliptically. Assuming, for simplicity, that the entire radiation is a linearly polarized wave of a single frequency, determine the rms value of its electric and magnetic field.

P21.6. The radius of the earth is about 6350 km. Assuming the entire energy of the sun's radiation reaching the earth is absorbed by the earth, calculate the time-average power of the absorbed energy, and the energy absorbed by the earth in one day. Compare this with the total man-produced energy, assuming that the time-average power of this energy during the day is about 12,500 GW.

P21.7. Due to various effects, human exposure to electromagnetic radiation is considered to be harmful above a certain time-average value of the Poynting vector. This estimated value depends on frequency, and differs greatly among different countries in the world. Assuming that above 10 GHz this value is on the order of $10\,\mathrm{mW/cm^2}$, compute the corresponding rms value of the electric and magnetic field of the plane wave with this time-average value of the Poynting vector. Compare this value of the electric field with the rms value of TV and broadcasting stations, which is on the order of mV/m.

P21.8. A circular wire loop of radius a is situated in a vacuum in the electromagnetic field of a plane wave, of wavelength λ ($\lambda \gg a$), and the rms value of the electric field strength E. How should the loop be positioned in order that the emf induced in it be maximal? Determine the rms value of the emf in that case.

P21.9. A rectangular wire loop with sides a and b is situated in a vacuum in the electromagnetic field of a time-harmonic plane wave. The amplitude of the electric field strength of the wave is E, and its wavelength is λ ($\lambda \gg a, b$). The loop is oriented so that the maximal emf is induced in it. In case (1) the sides a are parallel to the electric field of the wave, and in case (2) the sides b are parallel to the electric field of the wave. Evaluate in both cases the emf (a) starting from Faraday's law of electromagnetic induction in its usual form ($e = -\mathrm{d}\Phi/\mathrm{d}t$), and (b) as an integral of the electric field strength of the wave around the contour.

P21.10. A sinusoidal plane wave, of frequency f and time-average value of the Poynting vector $\mathcal{P}$, propagates through distilled water ($\mu = \mu_0, \epsilon = \epsilon_r\epsilon_0$). Find the rms value of the emf induced in a small circular loop of radius a, oriented so that the emf is maximal. What condition must be met in order that the loop can be considered as a quasi-static system?

P21.11. Prove that an elliptically polarized wave can be represented as a sum of two circularly polarized waves.

P21.12. Determine the time-average Poynting vector of a circularly polarized plane wave (1) starting from the expressions for the plane wave in time domain, and (2) starting from the phasor expressions for the plane wave.

P21.13. The electric field of a plane wave in complex (phasor) form is given by $\mathbf{E}(z) = (E_x\mathbf{u}_x + E_y\mathbf{u}_y)\mathrm{e}^{-\mathrm{j}\beta z}$. The components E_x and E_y are arbitrary complex numbers. Assuming that the wave propagates in the $+z$ direction, discuss the polarization of the wave, stating whether it is right-handed or left-handed for circular and elliptic polarization, if (1) $E_x = 1$, $E_y = 0$; (2) $E_x = 0$, $E_y = 5$; (3) $E_x = \mathrm{j}$, $E_y = -\mathrm{j}$; (4) $E_x = \mathrm{j}$, $E_y = 2$; (5) $E_x = (1 + \mathrm{j})$, $E_y = 0$; (6) $E_x = 1$, $E_y = \mathrm{j}$; or (7) $E_x = (1 + 2\mathrm{j})$, $E_y = (1 - \mathrm{j})$.

P21.14. Two plane waves of equal frequencies and phases propagate in the same z direction. Both are circularly polarized, but in opposite directions (one is right-handed, the other left-handed). The amplitudes of the electric field strength of the two waves are E_1 and E_2. Find the polarization of the resultant wave in terms of E_1 and E_2, starting from the expressions of the waves in time domain.

P21.15. Two linearly polarized sinusoidal waves of the same frequency propagate in the z direction. The electric field vectors of the two waves, $\mathbf{E}_1$ and $\mathbf{E}_2$, are along the x and y axes, respectively. Plot the trace that the tip of the resulting vector, $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$, traces at $z = 0$ in time, as a function of the ratio of amplitudes E_1 and E_2, and their relative phase, ϕ.

22

Reflection and Refraction of Plane Waves

22.1 Introduction

In reality, plane electromagnetic waves frequently encounter obstacles along their propagation paths: hills, buildings, metallic antennas aimed at receiving the messages the waves carry, objects from which they are supposed to partly reflect (as when the wave is a radar beam), and so on. In such cases, the wave induces conduction currents in the object (if the object is metallic), or polarization currents (if the object is made of an insulator). These currents are, of course, sources of a secondary electromagnetic field. This field is known as the *scattered field*, and the process that creates it is known as *scattering of electromagnetic waves*. The objects, or obstacles, are called *scatterers*.

The determination of scattered fields is a difficult problem even in the case of simple scatterers, and can rarely be solved analytically. Numerical analysis offers various solutions. There is one class of problems, however, for which the determination of the scattered field is remarkably simple. When a *plane* electromagnetic wave is incident on a *planar* boundary between two *homogeneous* media, the scattered waves are also plane waves. One of these waves is radiated back into the half-space of the incident wave: this wave is known as the *reflected wave*. There is also a wave in the other half-space (except in the case of a perfect conductor), propagating generally

in a different direction from the incident wave; it is therefore called the *refracted or transmitted wave.*

Naturally, the described geometry is an idealized one. Nevertheless, the results we will arrive at have great practical importance because many real problems can be solved in this manner with sufficient accuracy.

22.2 Plane Waves Normally Incident on a Perfectly Conducting Plane

The simplest case of wave reflection is when a uniform plane wave is incident normally on the planar interface between a perfect dielectric, of parameters ϵ and μ, and a perfect conductor. Let the interface be at $z = 0$, and let the wave of angular frequency ω have E_x and H_y components, as indicated in Fig. 22.1. We wish to determine the resulting wave for $z \leq 0$. (We know that inside the perfect conductor there is no field.)

The physics of wave scattering in this case is fairly obvious. The incident wave induces currents and charges only on the surface of the perfect conductor. (For a perfect conductor, the skin depth is infinitely small.) Since inside the conductor there is no field, we can consider this layer of currents and charges to exist in a homogeneous dielectric of parameters ϵ and μ. The distribution of these sources must be such that their field exactly cancels the incident field inside the conductor (that is, for $z > 0$). So we know that the scattered field for $z > 0$ is exactly the same in amplitude as the incident field, but is π out of phase. The current sheet obviously produces a symmetrical field in the half-space $z < 0$, that is, a plane wave propagating back in the $-z$ direction. This reradiated wave is the reflected wave. From this reasoning, the reflected wave has the same amplitude as the incident wave. At $z = 0$, its E-field vector is the same as that of the incident field, but in the opposite direction.

To put these conclusions into equations, let the incident wave (in phasor form) be represented by

$$\mathbf{E}_i(z) = E\mathrm{e}^{-\mathrm{j}\beta z}\mathbf{u}_x \qquad \mathbf{H}_i(z) = H\mathrm{e}^{-\mathrm{j}\beta z}\mathbf{u}_y, \tag{22.1}$$

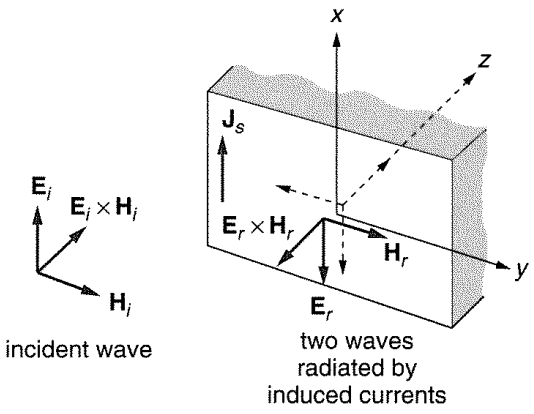

Figure 22.1 Uniform plane wave normally incident on the planar interface between a perfect dielectric and perfect conductor

where $E/H = \eta$ ($\eta = \sqrt{\mu/\epsilon}$, the intrinsic impedance of the medium). The reflected wave is then of the form

$$\mathbf{E}_r(z) = -Ee^{+j\beta z}\mathbf{u}_x \qquad \mathbf{H}_r(z) = He^{+j\beta z}\mathbf{u}_y. \tag{22.2}$$

The Poynting vector (which represents power flow) for the reflected wave is $-z$ oriented, which determines the sign of $\mathbf{H}_r(z)$.

The total field for $z < 0$ is obtained as a superposition of the waves in Eqs. (22.1) and (22.2):

$$\mathbf{E}_{\text{tot}}(z) = \mathbf{E}_i(z) + \mathbf{E}_r(z) = E\left(e^{-j\beta z} - e^{+j\beta z}\right)\mathbf{u}_x = -2jE\sin\beta z\,\mathbf{u}_x, \tag{22.3}$$

$$\mathbf{H}_{\text{tot}}(z) = \mathbf{H}_i(z) + \mathbf{H}_r(z) = H\left(e^{-j\beta z} + e^{+j\beta z}\right)\mathbf{u}_y = 2H\cos\beta z\,\mathbf{u}_y. \tag{22.4}$$

The instantaneous values of the two vectors are therefore

$$\mathbf{E}_{\text{tot}}(z,t) = 2E\sqrt{2}\sin\beta z\cos(\omega t - \pi/2)\,\mathbf{u}_x = 2E\sqrt{2}\sin\beta z\sin\omega t\,\mathbf{u}_x, \tag{22.5}$$

$$\mathbf{H}_{\text{tot}}(z,t) = 2H\sqrt{2}\cos\beta z\cos\omega t\,\mathbf{u}_y. \tag{22.6}$$

The total wave does *not* contain the factor $e^{\pm j\beta z}$. We already had such a case for voltage and current along open and shorted transmission lines. We know that it is not a progressive, traveling wave in either direction, but a *standing wave*. As along such transmission lines, there are planes in which $\mathbf{E}_{\text{tot}}(z,t)$ is zero at all times. These planes are defined by $\beta z = -n\pi$, $n = 0, 1, 2, \ldots$. Similarly, the magnetic field is zero at all times in planes defined by $\beta z = -(2n+1)\pi/2$, $n = 0, 1, 2, \ldots$. Thus, the total wave actually stays where it is, only pulsating in time according to the sine law (the E field) or the cosine law (the H field). The expressions describing a wave in which the time and space coordinates are as in Eqs. (22.3) and (22.4) in complex notation, or as in Eqs. (22.5) and (22.6) in the time domain, always represent standing waves. A sketch of instantaneous values of the total E and H field in front of the interface, for $\omega t = 0, \pi/4, 2\pi/4$, and $3\pi/4$, is shown in Fig. 22.2.

Example 22.1—The Fabry-Perot resonator. Consider again the case of a plane wave reflecting off a perfectly conducting plane, which behaves like a mirror. Note that according to Eq. (22.5), $\mathbf{E}_{\text{tot}}(-n\lambda/2, t) = 0$ *at all times* in the planes $z = -n\lambda/2$, $n = 1, 2, \ldots$. The electric field vector is tangential to these planes. Therefore, if we insert a perfectly conducting sheet (also a mirror) in any of these planes (i.e., for any n), nothing will change, since the boundary condition on the plane for vector $\mathbf{E}$ is satisfied automatically. In this manner, we obtained a semi-infinite region to the left of the sheet with the standing wave, and a region between the original mirror and the sheet in which *the electric and magnetic fields oscillate* as in Fig. 22.2. When the electric field is maximum, the magnetic field is zero, and conversely. This means that in the region between the two mirrors, the electric energy is being converted into magnetic energy, and vice versa. This is typical behavior for resonant electric circuits. We can conclude that from the energy point of view, just like in an LC circuit, this is a resonator, but a *spatial resonator*. This particular type of spatial resonator is known as the *Fabry-Perot resonator*, and is used extensively in optics and at millimeter-wave and infrared frequencies.

The Fabry-Perot resonator has a very useful property. Note that losses exist only in the original plane and in the sheet, due to the skin effect and finite conductivity of the metal. The

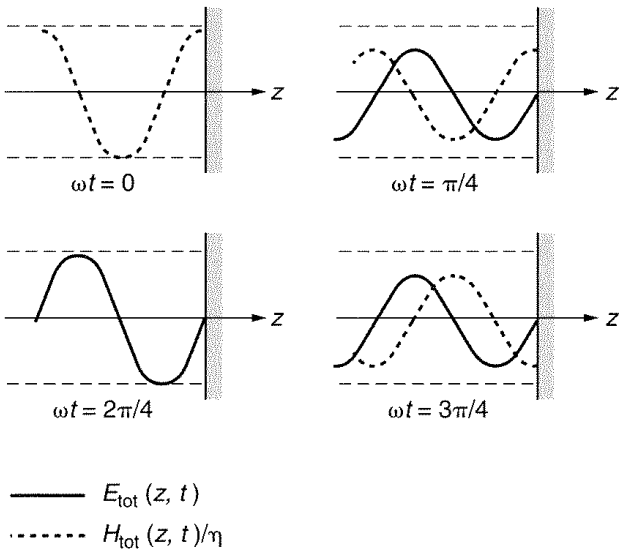

Figure 22.2 Sketch of instantaneous values of $E_{tot}(z, t)$ and $H_{tot}(z, t)$ in Eqs. (22.5) and (22.6) for a standing wave

electromagnetic energy located in the resonator, however, increases with the number n, that is, with the number of half wavelengths of the wave contained in the region between the two mirrors. At high frequencies (e.g., in the microwave region or in optics), n can be made very large with reasonable dimensions of the resonator. (For example, at 30 GHz, the wavelength is 1 cm, and a 10-cm-long resonator has $n = 20$.) Therefore, the Fabry-Perot resonator can have an arbitrarily large ratio between the energy stored in the resonator and losses in one cycle. We know that this ratio is proportional to the quality factor of the resonator (the Q factor). Therefore, the Q factor of a Fabry-Perot resonator can be extremely large (on the order of tens of thousands) when compared with that of a resonant circuit (which has a maximum Q factor of about one hundred). Of course, due to the finite size of the plates in reality, there will always be some leakage of electromagnetic energy from the resonator, which we do not take into account in this simplified analysis. Also, usually energy is purposely taken out of the resonator: for example, one of the mirrors may be partially transparent so that not all of the energy is reflected back into the cavity. This also is not taken into account in our simplified analysis.

Questions and problems: Q22.1 to Q22.4, P22.1 to P22.6

22.3 Reflection and Transmission of Plane Waves Normally Incident on a Planar Boundary Surface Between Two Dielectric Media

Let us consider two lossless dielectric media, 1 and 2, of parameters ϵ_1 and μ_1, and ϵ_2 and μ_2, respectively, separated by a planar interface, as in Fig. 22.3. Let the incident wave, with an electric field E_{1i} and of angular frequency ω, propagate in medium 1 toward the interface, normal to it, with the vector **E** parallel to the x axis (Fig. 22.3).

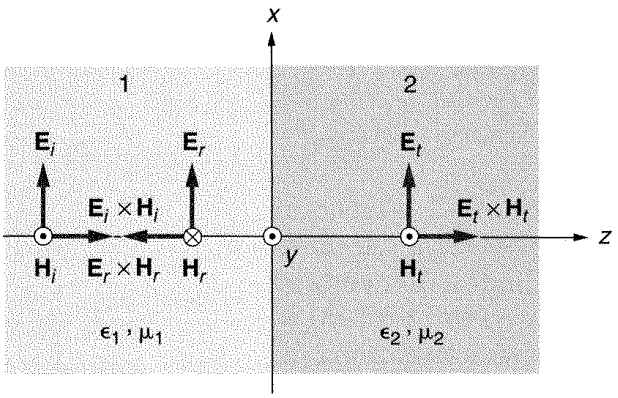

Figure 22.3 Two lossless dielectric media divided by a plane interface. The incident wave from medium 1 is partly reflected, and partly transmitted to medium 2.

A part of the incident electromagnetic energy will be reflected from the interface, and a part will be transmitted into medium 2. Assume the reference directions of the E field for the reflected and transmitted waves as indicated. The reference directions of the H field for the three waves are then as shown in the figure.

We wish to determine the relative intensities E_{1r} and E_2 at $z = 0$ of the E field for the reflected and transmitted waves. For this, we first need the expressions for these fields. With the adopted reference directions of the vectors in Fig. 22.3, they have the forms

$$\mathbf{E}_i(z) = E_{1i}e^{-j\beta_1 z}\mathbf{u}_x, \quad \mathbf{H}_i(z) = \frac{E_{1i}}{\eta_1}e^{-j\beta_1 z}\mathbf{u}_y, \tag{22.7}$$

$$\mathbf{E}_r(z) = E_{1r}e^{+j\beta_1 z}\mathbf{u}_x, \quad \mathbf{H}_r(z) = -\frac{E_{1r}}{\eta_1}e^{+j\beta_1 z}\mathbf{u}_y, \tag{22.8}$$

$$\mathbf{E}_t(z) = E_2e^{-j\beta_2 z}\mathbf{u}_x, \quad \mathbf{H}_t(z) = \frac{E_2}{\eta_2}e^{-j\beta_2 z}\mathbf{u}_y. \tag{22.9}$$

We can now write the boundary conditions, i.e., the requirements that the tangential components of the total vectors $\mathbf{E}$ and $\mathbf{H}$ on two sides of the interface be the same:

$$E_{1i} + E_{1r} = E_2, \qquad \frac{E_{1i}}{\eta_1} - \frac{E_{1r}}{\eta_1} = \frac{E_2}{\eta_2}. \tag{22.10}$$

By solving these two equations for E_{1r} and E_2, we obtain

$$E_{1r} = \frac{\eta_2 - \eta_1}{\eta_1 + \eta_2}E_{1i}, \qquad E_2 = \frac{2\eta_2}{\eta_1 + \eta_2}E_{1i}. \tag{22.11}$$

Note that these are the same expressions we found in Chapter 18 for the incident (forward) and reflected (backward) voltages along a line of characteristic impedance Z_1 terminated in an infinite line of characteristic impedance Z_2 (Example 18.8). As

in the case of transmission lines, the ratio E_{1r}/E_{1i} is known as the *reflection coefficient*, and the ratio E_2/E_{1i} the *transmission coefficient*:

$$\rho = \frac{\eta_2 - \eta_1}{\eta_1 + \eta_2} \qquad \text{(the reflection coefficient, dimensionless)} \tag{22.12}$$

$$\tau = \frac{2\eta_2}{\eta_1 + \eta_2} \qquad \text{(the transmission coefficient, dimensionless).} \tag{22.13}$$

Note also that ρ and τ as just derived are defined with respect to the same reference directions of all three components of the *electric field* of the three waves.

In medium 2 there is only the progressive transmitted wave. In medium 1, however, we have the incident wave and the reflected wave, the total field being the sum of the two:

$$\mathbf{E}_1(z) = \mathbf{E}_i(z) + \mathbf{E}_r(z) = E_{1i}e^{-j\beta_1 z}\left(1 + \rho e^{+j2\beta_1 z}\right)\mathbf{u}_x. \tag{22.14}$$

This is the same expression as for the voltage along a transmission line terminated in a load, Eq. (18.22*a*). The electric field in medium 1 is therefore of the same form as the voltage in the analogous transmission-line case. The following analysis, which parallels that from Chapter 18, shows this clearly.

If $\rho > 0$ (that is, if $\eta_2 > \eta_1$), the expression in parentheses is the largest, equal to $(1 + \rho)$, in planes defined by the following equation (note that medium 1 occupies the half-space $z < 0$):

$$2\beta_1 z_{\max} = -2n\pi, \qquad \text{or} \qquad z_{\max} = -\frac{n\pi}{\beta_1} = -\frac{n\lambda_1}{2} \qquad n = 0, 1, \ldots. \tag{22.15}$$

This expression is minimal, equal to $(1 - \rho)$, in planes

$$z_{\min} = -(2n + 1)\frac{\lambda_1}{4} \qquad n = 0, 1, \ldots. \tag{22.16}$$

If $\rho < 0$ (that is, if $\eta_2 < \eta_1$), $z_{\max}$ and $z_{\min}$ simply exchange places.

The resultant wave in medium 1 can be visualized as a sum of a progressive wave of rms value $(1 - |\rho|)E_{1i}$ and a standing wave of rms value (in the maximum of the standing wave) $2|\rho|E_{1i}$. This becomes evident if Eq. (22.14) is rewritten in the form

$$\mathbf{E}_1(z) = (1 - \rho)E_{1i}e^{-j\beta_1 z}\mathbf{u}_x + 2\rho E_{1i}\cos\beta_1 z\,\mathbf{u}_x \qquad (\rho > 0), \tag{22.17}$$

namely,

$$\mathbf{E}_1(z) = (1 + \rho)E_{1i}e^{-j\beta_1 z}\mathbf{u}_x + 2j\rho E_{1i}\sin\beta_1 z\,\mathbf{u}_x \qquad (\rho < 0). \tag{22.18}$$

The ratio analogous to the voltage standing-wave ratio, VSWR, from transmission lines,

$$\text{SWR} = \frac{|E_1(z)|_{\max}}{|E_1(z)|_{\min}} = \frac{1 + |\rho|}{1 - |\rho|} \qquad \text{(standing-wave ratio, dimensionless)} \tag{22.19}$$

is known as the *standing-wave ratio*. Since $|\rho| < 1$, it increases with the increase of $|\rho|$.

Questions and problems: Q22.5, P22.7 to P22.11

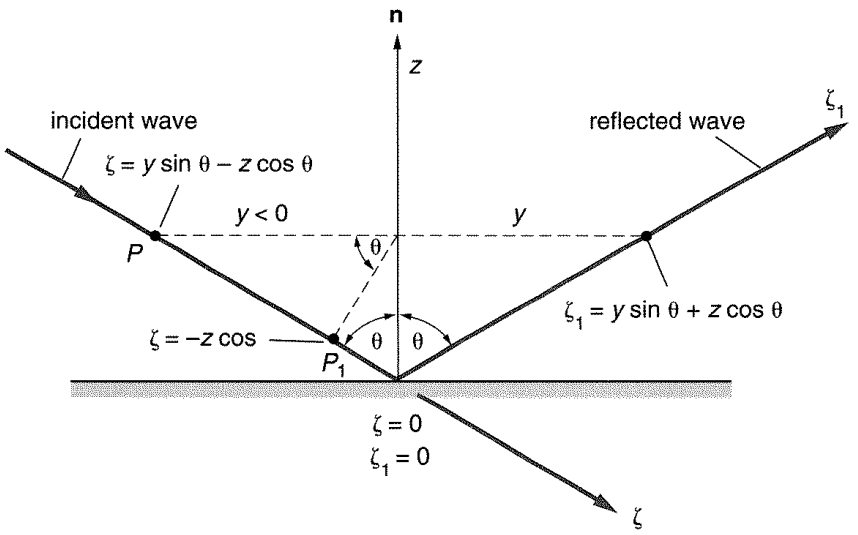

Figure 22.4 The direction of propagation of a wave reflected off a perfectly conducting plane is at the same angle with respect to the normal to the plane as the incident wave direction.

22.4 Plane Waves Obliquely Incident on a Perfectly Conducting Plane

Assume that a uniform plane wave propagating in a perfect dielectric of parameters ϵ and μ is obliquely incident on a perfectly conducting plane. As in the case of normal incidence, the scattered field due to the currents and charges induced on the plane must be such that it cancels the incident field in the perfect conductor. Thus, these currents and charges will produce a wave inside the conductor. This wave will be exactly the same as the incident wave, propagating in the same direction, but of opposite phase. The same field will be produced on the other side of the plane, resulting in a "reflected wave." Therefore, the direction of propagation of the reflected wave will be at the same angle θ (with respect to the normal to the plane) as the incident wave (Fig. 22.4).

The plane containing the vector **n** and the directions of propagation of the incident and reflected waves is known as the *plane of incidence*. Any incident plane wave can be represented as a superposition of two plane waves, one with the vector **E** normal to the plane of incidence, and the other with the vector **E** parallel to it. These two cases are simpler to analyze than any other. Therefore, we consider these two special cases only, knowing that any other case can be obtained by superposition.

22.4.1 VECTOR E NORMAL TO THE PLANE OF INCIDENCE

It is customary to say that this wave has *normal* or *horizontal polarization*. (The term "normal" refers to the plane of incidence, and "horizontal" refers to the fact that frequently the reflection plane is the earth's surface, in which case the vector **E** of this wave is horizontal.) The case of a horizontally polarized incident wave is sketched in Fig. 22.5, with the adopted reference directions for vectors **E** and **H**.

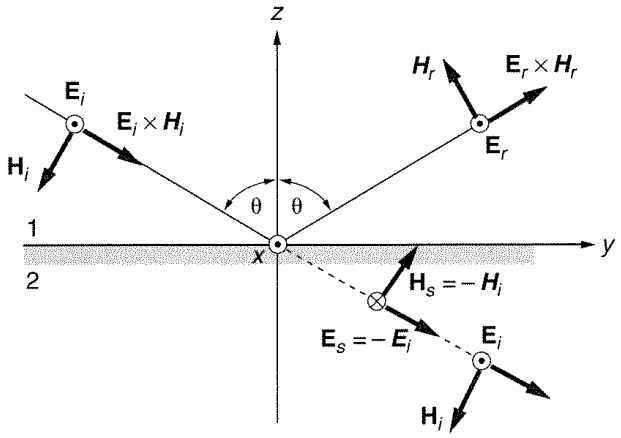

Figure 22.5 A horizontally polarized plane wave reflected from a perfectly conducting plane

The wave propagates along an axis (ζ axis, Fig. 22.4) not coinciding with a coordinate axis x, y, or z. To write the expression for the wave in terms of the rectangular coordinates, we need to determine the distance of a point P (Fig. 22.4) from the origin of the ζ axis ($\zeta = 0$) in terms of x, y, and z. From Fig. 22.4 it is seen that for the incident wave this distance equals $\zeta = y \sin\theta - z\cos\theta$. For P note that $y < 0$ and $z > 0$, and that P is in the negative part of the ζ axis. So the factor $e^{-j\beta\zeta}$ (the factor for the wave propagating in the direction of the ζ axis) becomes $e^{-j\beta(y\sin\theta - z\cos\theta)}$. The expression for the complex electric field of the incident wave is thus

$$\mathbf{E}_i(y, z) = Ee^{-j\beta(y\sin\theta - z\cos\theta)}\,\mathbf{u}_x. \tag{22.20}$$

In the same wave we conclude that the E field of the reflected wave is given by

$$\mathbf{E}_r(y, z) = -Ee^{-j\beta(y\sin\theta + z\cos\theta)}\,\mathbf{u}_x. \tag{22.21}$$

The minus sign comes from the requirement that the total tangential E field on the plane $z = 0$ must be zero for any y.

The total electric field has only an x component, given by

$$E_{\text{tot}}(y, z) = E_i(y, z) + E_r(y, z) = Ee^{-j\beta y\sin\theta}\left(e^{j\beta z\cos\theta} - e^{-j\beta z\cos\theta}\right),$$

from which

$$E_{\text{tot}}(y, z) = 2jE\sin(\beta z\cos\theta)\,e^{-j\beta y\sin\theta}. \tag{22.22}$$

We see that the total electric field is a standing wave in the z direction, and a traveling wave in the y direction. The wavelength in the z direction is given by

$$\lambda_z = \frac{2\pi}{\beta\cos\theta} = \frac{\lambda}{\cos\theta}, \tag{22.23}$$

where λ is the wavelength of the incident (and reflected) wave. The vector $\mathbf{E}$ is zero in the planes in which $\beta z\cos\theta = n\pi$, $n = 0, 1, 2, \ldots$, or

$$z_{E=0} = \frac{n\lambda_z}{2} = \frac{n\lambda}{2\cos\theta}, \quad n = 0, 1, 2, \ldots. \tag{22.24}$$

In the direction of the y axis, the total field behaves as a traveling wave, with a phase velocity along the y axis

$$v_{\text{ph}} = \frac{\omega}{\beta_y} = \frac{\omega}{\beta \sin\theta} = \frac{c}{\sin\theta}, \qquad c = \frac{1}{\sqrt{\epsilon\mu}} \tag{22.25}$$

(note that the phase coefficient with respect to the y axis is the entire factor of jy in the exponent, that is, $\beta_y = \beta \sin\theta$), and with a wavelength along the y axis

$$\lambda_y = \frac{2\pi}{\beta_y} = \frac{\lambda}{\sin\theta}. \tag{22.26}$$

Example 22.2—The rectangular waveguide. Because in the planes $z_{E=0}$ the magnitude of the E field is zero at all times, we can insert into any one of these planes a perfectly conducting sheet. Assume that in some way we switch the field above the sheet off. What remains is a system of two perfectly conducting planes guiding a specific wave propagating in the y direction.

We can go a step further. The vector **E** is normal to the planes defined by $x =$ constant. Introducing a perfectly conducting sheet in one or more such planes will not change the field—it will only induce surface charges of opposite signs on the two faces of the sheet. However, if we imagine two such sheets together with the first sheet parallel to the plane $z = 0$, we obtain a rectangular tube through which an electromagnetic wave propagates just like water flows through a pipe. Such a rectangular metallic tube is known as the *rectangular waveguide.* It is used extensively for guiding electromagnetic energy at microwave and millimeter-wave frequencies.

We will learn in the next chapter that this type of electromagnetic wave is only one of an infinite number of wave types that can propagate through such rectangular metallic tubes.

Example 22.3—Determination of the total H field. With reference to Fig. 22.5, the total H field has two components, H_y and H_z. Let us determine them as an exercise. The two components of the total H field are obtained as the sum of these components for the incident and the reflected waves:

$$\begin{aligned} H_{\text{tot }y}(y,z) &= H_{iy}(y,z) + H_{ry}(y,z) \\ &= -\frac{E}{\eta} e^{-j\beta(y\sin\theta - z\cos\theta)} \cos\theta - \frac{E}{\eta} e^{-j\beta(y\sin\theta + z\cos\theta)} \cos\theta. \end{aligned}$$

After simple rearrangements similar to those in deriving Eq. (22.22), we obtain

$$H_{\text{tot }y}(y,z) = -2\frac{E}{\eta} \cos\theta \cos(\beta z \cos\theta) e^{-j\beta y \sin\theta}.$$

The H_z component is obtained in a similar way, which is left as an exercise for the reader. The result is

$$H_{\text{tot }z}(y,z) = 2j\frac{E}{\eta} \sin\theta \sin(\beta z \cos\theta) e^{-j\beta y \sin\theta}.$$

Note that H_z is zero on the perfectly conducting plane, as it should be (the magnetic field can have no normal component on a perfect conductor—see Example 20.2).

22.4.2 VECTOR E PARALLEL TO THE PLANE OF INCIDENCE

Assume now that vector **E** is parallel to the plane of incidence, as sketched in Fig. 22.6. Because the tangential component (y component) at the plane must be zero, again the reflected wave has the same amplitude. The directions of the E and H vectors indicated in the figure represent their reference directions.

We now have two E-field components of the incident and the reflected waves, the y and the z components. Both must be of the form in Eqs. (22.20) and (22.21):

$$E_{iy}(y, z) = E\cos\theta e^{-j\beta(y\sin\theta - z\cos\theta)}, \tag{22.27}$$

$$E_{iz}(y, z) = E\sin\theta e^{-j\beta(y\sin\theta - z\cos\theta)}, \tag{22.28}$$

and

$$E_{ry}(y, z) = -E\cos\theta e^{-j\beta(y\sin\theta + z\cos\theta)}, \tag{22.29}$$

$$E_{rz}(y, z) = E\sin\theta e^{-j\beta(y\sin\theta + z\cos\theta)}. \tag{22.30}$$

The total components are the sum of these:

$$E_{\text{tot }y}(y, z) = E_{iy}(y, z) + E_{ry}(y, z) = 2jE\cos\theta\sin(\beta z\cos\theta)e^{-j\beta y\sin\theta}, \tag{22.31}$$

$$E_{\text{tot }z}(y, z) = E_{iz}(y, z) + E_{rz}(y, z) = 2E\sin\theta\cos(\beta z\cos\theta)e^{-j\beta y\sin\theta}. \tag{22.32}$$

Example 22.4—Determination of the total H field. With reference to Fig. 22.6, the total H field in this case has the x component only. Because we know that $H = E/\eta$, the expressions

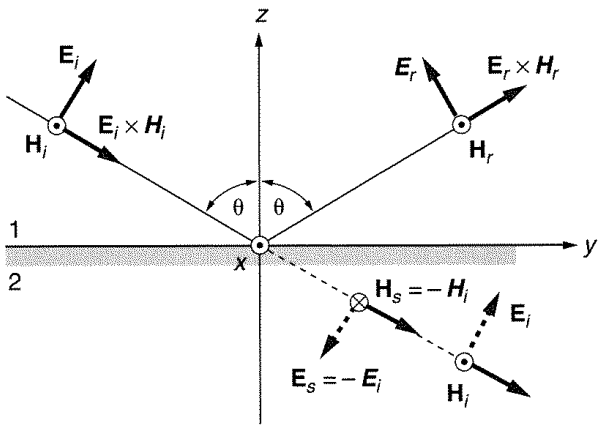

Figure 22.6 Reflection of a vertically polarized plane wave from a perfectly conducting plane

for the H field of the incident and reflected waves are

$$H_{ix}(y, z) = \frac{E}{\eta} e^{-j\beta(y \sin\theta - z\cos\theta)} \quad \text{and} \quad H_{rx}(y, z) = \frac{E}{\eta} e^{-j\beta(y\sin\theta + z\cos\theta)}.$$

The total H field is hence

$$H_{\text{tot } x}(y, z) = 2\frac{E}{\eta} \cos(\beta z \cos\theta) e^{-j\beta y \sin\theta}.$$

Example 22.5—Maximal emf induced in a small loop above a perfectly conducting plane. Assume that we wish to receive a signal contained in the incident wave. One way, which is quite easy to understand, is to use a loop of wire (e.g., a circular one of radius a) much smaller than the wavelength of the wave. The emf induced in the loop is then obtained according to Faraday's law. So to obtain a maximal emf, the principal thing we have to determine is where the magnetic field is maximal, and what its direction is at that point.

The magnetic field has a maximum at $z = 0$, with a value equal to

$$H_{\text{tot } x}(y, 0) = 2\frac{E}{\eta} e^{-j\beta y \sin\theta}.$$

The last factor in this expression determines just the initial phase of the field along the y axis. To simplify, let $y = 0$. The maximal possible complex emf [note that the complex counterpart of the expression $e(t) = -d\Phi(t)/dt$ is $\underline{e} = -j\omega\underline{\Phi}$] induced in the loop is thus

$$\mathcal{E}_{\max} = -2j\frac{E}{\eta}\omega\mu_0 a^2 \pi.$$

In this case, we have used the small loop as a receiving antenna. This type of antenna, which develops a voltage between its terminals due to a time-variable flux of the magnetic field through its contour, is called a *loop antenna.* We could also have used two short straight wires connected to a voltmeter that measures the emf. In this case, there is an induced emf due to the integral of the induced electric field along the wires. This type of antenna is called a *short dipole antenna.*

Questions and problems: Q22.6 and Q22.7, P22.12 and P22.13

22.5 Reflection and Transmission of Plane Waves Obliquely Incident on a Planar Boundary Surface Between Two Dielectric Media

When a plane wave is obliquely (at an angle) incident on a plane interface between two media, the formulation of boundary conditions becomes more complex than when incidence is normal. Of course, again a part of the incident energy is reflected back into medium 1 (of parameters ϵ_1 and μ_1), and a part is transmitted into medium 2 (of parameters ϵ_2 and μ_2). We shall see that the direction of propagation of the reflected wave makes the same angle with the normal to the interface as the incident wave, as before. However, the transmitted wave is deflected with respect to this normal. The transmitted wave in this case is therefore frequently termed the *refracted wave.*

The amplitudes of the reflected and refracted waves depend, among other things, on the polarization of the wave (i.e., on the electric field vector being parallel or normal to the plane of incidence). However, the angles that the direction of propagation of the two secondary waves make with the normal to the interface are the same for any polarization.

Figure 22.7 shows equiphase planes (planes of constant phase) and the directions of propagation of the incident, reflected, and refracted waves. These planes in medium 1 are moving with a velocity $c_1 = 1/\sqrt{\epsilon_1 \mu_1}$, and in medium 2 with a velocity $c_2 = 1/\sqrt{\epsilon_2 \mu_2}$. Indicated in the figure are a few equiphase planes of the three waves. Let the boundary conditions be satisfied at the instant for which Fig. 22.7 is valid. In order that they remain satisfied at all times, it is necessary that the relative amplitudes and phases of the three waves at the interface remain unchanged. This is possible only if the intersections of the equiphase planes with the interface move *along the interface* at the same speed.

From Fig. 22.7, this velocity for the incident and reflected wave is $c_1/\sin\theta_i$ and $c_1/\sin\theta_r$, and for the refracted wave $c_2/\sin\theta_2$. To satisfy this condition we conclude that, first, $\theta_r = \theta_i$. This angle we shall therefore denote as θ_1. Second, the condition $c_1/\sin\theta_1 = c_2/\sin\theta_2$ must also hold, so

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{c_1}{c_2} = \sqrt{\frac{\epsilon_2\mu_2}{\epsilon_1\mu_1}}. \qquad (22.33)$$

(Snell's law)

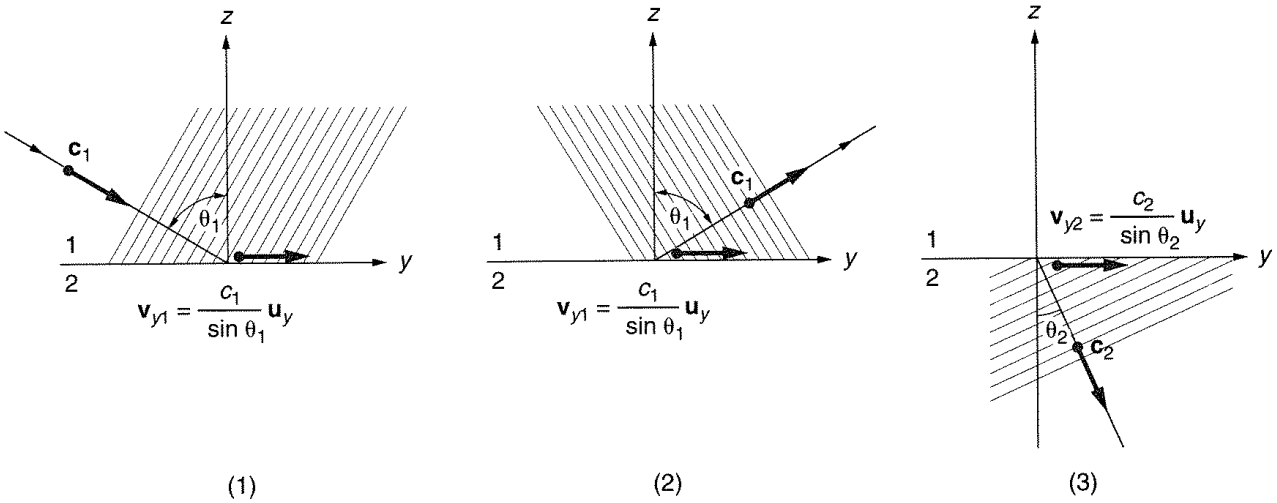

Figure 22.7 Equiphase planes and the directions of propagation of the incident, reflected, and refracted waves

This relation is known as *Snell's law*. The ratio c_1/c_2 is termed the *index of refraction* for media 1 and 2, and is often denoted as n_{12}, especially in optics. If $\mu_1 = \mu_2 = \mu_0$ (which is most often the case),

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{c_1}{c_2} = \sqrt{\frac{\epsilon_2}{\epsilon_1}} \qquad (\mu_1 = \mu_2). \tag{22.34}$$

We know that if $0 \leq \beta < \alpha \leq \pi/2$, then $\sin\alpha > \sin\beta$. Snell's law and Eq. (22.34) therefore tell us that for $\epsilon_1 < \epsilon_2$, $\theta_1 > \theta_2$. This means that the wave is refracted toward the normal. The refracted wave exists for any θ_1.

If $\epsilon_1 > \epsilon_2$, however, $\theta_1 < \theta_2$. This means that the direction of propagation of the refracted wave makes a greater angle with the normal than that of the incident wave. So for a certain angle θ_1 the angle θ_2 will become $\pi/2$, and cannot increase further. From Eq. (22.34), this limiting angle $\theta_1 = \theta_t$ is defined by

$$\frac{\sin\theta_t}{\sin(\pi/2)} = \sin\theta_t = \sqrt{\frac{\epsilon_2}{\epsilon_1}} \qquad (\epsilon_1 > \epsilon_2, \mu_1 = \mu_2). \tag{22.35}$$

[Sine of critical angle (angle of total reflection)]

This particular angle $\theta_1 = \theta_t$ is known as the *critical angle*, or the *angle of total reflection*.

For $\theta_1 > \theta_t$, the sine of θ_2 must be *greater than one* in order for the boundary conditions to be satisfied. At first glance it might seem as if we made a mistake. The sine of a *real* angle cannot be greater than one. However, the sine of a complex angle can be larger than unity. This is easily understood if we set $\theta_2 = \pi/2 - \mathrm{j}x$ and recall the expression for the sine in terms of the exponential function:

$$\sin(\pi/2 - \mathrm{j}x) = \frac{1}{2\mathrm{j}}\left[\mathrm{e}^{\mathrm{j}(\pi/2-\mathrm{j}x)} - \mathrm{e}^{-\mathrm{j}(\pi/2-\mathrm{j}x)}\right] = \frac{1}{2\mathrm{j}}\left(\mathrm{j}e^{x} + \mathrm{j}e^{-x}\right) = \cosh x,$$

since $\mathrm{e}^{\pm\mathrm{j}\pi/2} = \pm\mathrm{j}$. The *hyperbolic consine* function of x, $\cosh x = (e^x + e^{-x})/2$ can have any positive value between one and infinity.

What happens then if $\theta_1 > \theta_t$? Obviously, there can be no refracted wave in medium 2, so all of the energy of the incident wave is reflected back into medium 1. Example 22.9 will show that indeed, the magnitude of the reflection coefficient is then equal to one. This is known as *total reflection*. It has many applications and is encountered on many occasions.

Example 22.6—Apparent shape of an oar observed from a rowboat. If you are in a rowboat on clear, calm water, and observe the oar immersed in the water, the oar looks as if it is broken at the water level: the immersed part of the oar appears higher than expected. This is easy to explain using Snell's law. You see the immersed part of the oar because light rays, i.e., electromagnetic waves, are reflected from the oar toward your eyes. They pass the water-air interface and are refracted in the air away from the normal because the permittivity of water is greater than that of air. Therefore, the oar looks broken.

If you observe the oar from a distant point, you will not be able to see the submerged part of the oar. This is because the rays from the submerged part of the oar in that case are incident on the water-air interface at an angle greater than the critical angle, and there are no transmitted rays in the air in your direction anymore.

Questions and problems: Q22.8, P22.14

22.6 Fresnel Coefficients

Snell's law and the phenomenon of total reflection are valid for any polarization of the incident wave. The reflection and transmission coefficients, which are defined in the same way as for normal incidence, are different for normal and parallel polarization. In this section we derive the so-called *Fresnel coefficients*, which are reflection and transmission coefficients written in terms of the angle of incidence and the material properties (wave impedances) of the two media.

For waves obliquely incident on the interface between two dielectrics, we need to consider the two polarizations separately, similarly to the conductor case.

22.6.1 VECTOR E NORMAL TO THE PLANE OF INCIDENCE (TRANSVERSE ELECTRIC CASE)

Let the reference directions of the field vectors of the incident, reflected, and refracted waves be as in Fig. 22.8. Let E_{1i}, E_{1r}, and E_2 and H_{1i}, H_{1r}, and H_2 be the complex rms values of the vectors of the three waves at the interface ($z = 0$). The boundary conditions require that the tangential components of the total E field and the total

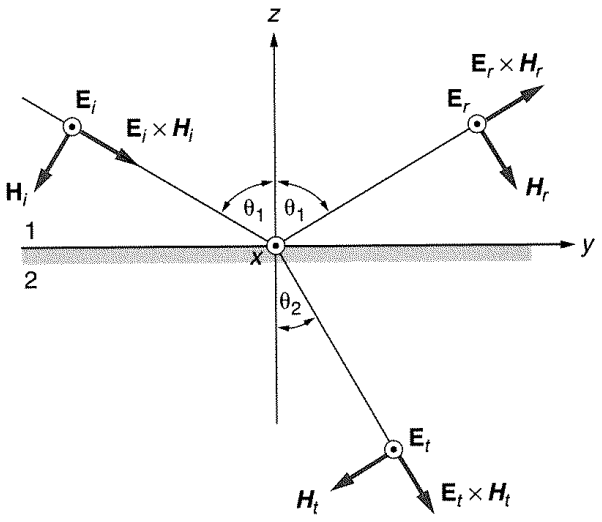

Figure 22.8 Reference directions of the field vectors of the incident, reflected, and refracted waves for a normally polarized incident wave

H field on the two sides of the interface be equal. This results in the following two equations:

$$E_{1i} + E_{1r} = E_2 \qquad (H_{1i} - H_{1r}) \cos\theta_1 = H_2 \cos\theta_2. \tag{22.36}$$

Since $H_{1i} = E_{1i}/\eta_1$, $H_{1r} = E_{1r}/\eta_1$, and $H_2 = E_2/\eta_2$, we have two linear equations in two unknowns, E_{1r} and E_2. Solving these equations we obtain

$$\rho_n = \left(\frac{E_{1r}}{E_{1i}}\right)_n = \frac{\eta_2 \cos\theta_1 - \eta_1 \cos\theta_2}{\eta_2 \cos\theta_1 + \eta_1 \cos\theta_2}, \tag{22.37}$$

$$\tau_n = \left(\frac{E_2}{E_{1i}}\right)_n = \frac{2\eta_2 \cos\theta_1}{\eta_2 \cos\theta_1 + \eta_1 \cos\theta_2}. \tag{22.38}$$

[Fresnel's coefficients for normal (TE) polarization]

The reflection and transmission coefficients, ρ_n and τ_n, are known as the *Fresnel coefficients* for normal polarization. They are also sometimes called the *transverse electric (TE)* Fresnel coefficients. In these expressions, according to Snell's law in Eq. (22.33), $\cos\theta_2$ must be calculated as

$$\cos\theta_2 = \sqrt{1 - \sin^2\theta_2} = \frac{c_2}{c_1}\sqrt{\left(\frac{c_1}{c_2}\right)^2 - \sin^2\theta_1}. \tag{22.39}$$

The expressions for the ρ and τ coefficients are general. For perfect dielectrics, having real intrinsic impedances, they are real. As a consequence, the reflected wave on the interface is either in phase (if $\rho > 0$) or in counterphase (if $\rho < 0$) with respect to the incident wave. If either of the two media is not a perfect dielectric, the intrinsic impedance of that medium is complex, so that both ρ and τ are complex as well, and the phase difference between the field vectors on the interface is different from π or zero.

22.6.2 VECTOR E PARALLEL TO THE PLANE OF INCIDENCE (TRANSVERSE MAGNETIC CASE)

Assume that the reference directions of the field vectors of the incident, reflected, and refracted waves in this case is as in Fig. 22.9. Again let E_{1i}, E_{1r}, and E_2 and H_{1i}, H_{1r}, and H_2 be the complex rms values of the field vectors of the three waves at $z = 0$. The boundary conditions in this case are

$$(E_{1i} - E_{1r}) \cos\theta_1 = E_2 \cos\theta_2 \qquad H_{1i} + H_{1r} = H_2. \tag{22.40}$$

Expressing the magnetic field intensities as E/η with appropriate subscripts, we again obtain two linear equations in unknowns E_{1r} and E_2. The solution of these equations is straightforward. The reflection and transmission coefficients are found to be

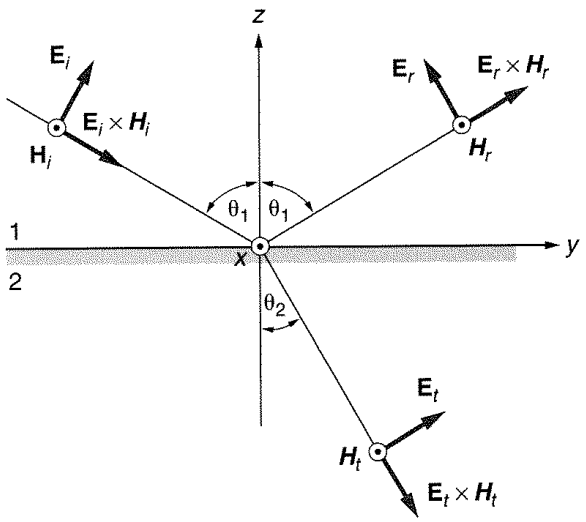

Figure 22.9 Reference directions of the field vectors of the incident, reflected, and refracted waves for the parallel polarization of the incident wave

$$\rho_p = \left(\frac{E_{1r}}{E_{1i}}\right)_p = \frac{\eta_1 \cos\theta_1 - \eta_2 \cos\theta_2}{\eta_1 \cos\theta_1 + \eta_2 \cos\theta_2}, \tag{22.41}$$

$$\tau_p = \left(\frac{E_2}{E_{1i}}\right)_p = \frac{2\eta_2 \cos\theta_1}{\eta_1 \cos\theta_1 + \eta_2 \cos\theta_2}. \tag{22.42}$$

[Fresnel's coefficients for parallel (TM) polarization]

Of course, in these two expressions $\cos\theta_2$ must also be calculated as in Eq. (22.39). The coefficients ρ_p and τ_p in Eqs. (22.41) and (22.42) are the parallel polarization Fresnel coefficients, sometimes also called the *transverse magnetic (TM)* Fresnel coefficients.

Example 22.7—Transmission-line models for oblique incidence of plane waves. We mentioned earlier that transmission-line theory can be used for plane waves incident normally to any interface. It turns out that with a slight modification, we can also use transmission-line theory for oblique incidence. This can be seen if we rewrite the Fresnel coefficients, Eqs. (22.37) and (22.42), as

$$\rho_n = \frac{\dfrac{\eta_2}{\cos\theta_2} - \dfrac{\eta_1}{\cos\theta_1}}{\dfrac{\eta_2}{\cos\theta_2} + \dfrac{\eta_1}{\cos\theta_1}} = \frac{\eta_{2n} - \eta_{1n}}{\eta_{2n} + \eta_{1n}}, \tag{22.43}$$

$$\rho_p = -\frac{\eta_2 \cos\theta_2 - \eta_1 \cos\theta_1}{\eta_2 \cos\theta_2 + \eta_1 \cos\theta_1} = -\frac{\eta_{2p} - \eta_{1p}}{\eta_{2p} + \eta_{1p}}, \tag{22.44}$$

where we have now defined the normal and parallel wave impedances as $\eta_{\text{in}} = \eta_i/\cos\theta_i$ and $\eta_{\text{ip}} = \eta_i \cos\theta_i$. (The minus sign in ρ_p results from the adopted reference directions, and is of no importance.) The transmission coefficients can, obviously, be written in the same way. As an exercise, it is suggested that the reader determine ρ for a normally polarized wave incident at a 45-degree angle from air on a dielectric with $\epsilon_r = 4$ and $\mu = \mu_0$.

Example 22.8—Fresnel coefficients for perfect dielectrics with equal permeabilities. In practice the most common case is actually the special case of the two media being perfect dielectrics of equal permeabilities. Then $\eta_1/\eta_2 = \sqrt{\epsilon_2/\epsilon_1}$, and the reflection and transmission coefficients for the normal polarization in Eqs. (22.37) and (22.38) become

$$\rho_n = \frac{\cos\theta_1 - \sqrt{\epsilon_2/\epsilon_1}\cos\theta_2}{\cos\theta_1 + \sqrt{\epsilon_2/\epsilon_1}\cos\theta_2}, \qquad \tau_n = \frac{2\cos\theta_1}{\cos\theta_1 + \sqrt{\epsilon_2/\epsilon_1}\cos\theta_2} \qquad (\mu_1 = \mu_2). \tag{22.45}$$

For the parallel polarization, the reflection and transmission coefficients in this case become

$$\rho_p = \frac{\sqrt{\epsilon_2/\epsilon_1}\cos\theta_1 - \cos\theta_2}{\sqrt{\epsilon_2/\epsilon_1}\cos\theta_1 + \cos\theta_2}, \qquad \tau_p = \frac{2\cos\theta_1}{\sqrt{\epsilon_2/\epsilon_1}\cos\theta_1 + \cos\theta_2} \qquad (\mu_1 = \mu_2). \tag{22.46}$$

Example 22.9—The Brewster angle. From Example 22.8, a few simple conclusions can be drawn:

1. It is not difficult to understand that ρ_n can never be zero. This would require that, simultaneously, $\sin\theta_1/\sin\theta_2 = \sqrt{\epsilon_2/\epsilon_1}$ (Snell's law) and $\cos\theta_1/\cos\theta_2 = \sqrt{\epsilon_2/\epsilon_1}$ (the equation resulting from $\rho_n = 0$), which is not possible.
2. If θ_1 is greater than the critical angle for total reflection, we know that $\sin^2\theta_1 > \epsilon_2/\epsilon_1$, so that $\cos\theta_2$ is purely imaginary. We see from Eq. (22.45) that ρ_n is then in the form $(a - jb)/(a + jb)$. This means that the magnitude of ρ_n is equal to one, that is, that the entire energy of the incident wave is reflected back into medium 1. The same conclusion can be reached for ρ_p.
3. The reflection coefficient in the parallel polarization case *can* be zero. For that to happen, it is necessary that $\cos\theta_1/\cos\theta_2 = \sqrt{\epsilon_1/\epsilon_2}$. This is now not in contradiction with Snell's law, but both equations must simultaneously be satisfied. If we multiply the two equations, we obtain that the reflected wave does not exist if

$$\sin\theta_1\cos\theta_1 = \sin\theta_2\cos\theta_2, \qquad \text{or} \qquad \sin 2\theta_1 = \sin 2\theta_2. \tag{22.47}$$

This equation is satisfied if $2\theta_1 = (\pi - 2\theta_2)$, that is, if $(\theta_1 + \theta_2) = \pi/2$. But for two angles adding to $\pi/2$ the sine of one equals the cosine of the other, so that from Snell's law,

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{\sin\theta_1}{\cos\theta_1} = \tan\theta_1 = n_{12}. \tag{22.48}$$

(The Brewster angle, parallel polarization only)

This particular angle of incidence of a wave with parallel polarization, for which the reflected wave disappears, is known as the *Brewster angle* or the *polarization angle.*

Example 22.10—Polarization of reflected waves incident at the Brewster angle. We know that an arbitrarily polarized wave can always be represented as a superposition of two

linearly polarized waves. Therefore, if we have, for example, an elliptically polarized wave incident at the Brewster angle, the reflected wave will not contain the component with parallel polarization, i.e., it will have only a normally polarized electric field. In other words, any wave incident on a plane interface of two dielectric media at the Brewster angle will be reflected as a *linearly polarized wave.*

Example 22.11—Elimination of the reflected wave in the case of an arbitrarily polarized incident wave. Suppose that we introduce in the path of the reflected wave from the preceding example a dielectric slab oriented so that the wave is incident on it at the Brewster angle, and that the polarization of the wave (recall that it is defined with respect to the plane of incidence) is parallel. The reflected wave is then going to disappear completely. This is exactly how this phenomenon was discovered experimentally by Brewster, using electromagnetic waves in the visible light frequency region.

Questions and problems: Q22.9 and Q22.10, P22.15 to P22.19

22.7 Chapter Summary

1. If a plane wave is incident on a plane boundary surface between two media, boundary conditions can be satisfied by assuming that the wave resulting from the discontinuity (the scattered wave) consists of a reflected plane wave, and (if the other medium is not perfectly conducting) of a transmitted, or refracted, plane wave. This enables relatively simple analysis of electromagnetic scattering of plane waves, similar to transmission-line analysis.
2. If a plane wave is normally incident on a perfectly conducting plane, a standing wave in front of the plane results. If incidence is not normal, there is a standing wave in the direction normal to the plane, and a traveling wave parallel to it.
3. If a plane wave is incident on a plane boundary surface between two dielectric media, a plane wave is reflected from the interface, and a plane wave is transmitted into the other medium.
4. For an arbitrary angle of the incident wave, the plane of incidence is defined as the plane normal to the boundary and containing the direction of propagation of the incident wave.
5. The incident wave is said to have normal polarization if **E** is normal to the plane of incidence, and to have parallel polarization if **E** is parallel to that plane.
6. The ratios of the amplitudes of the reflected and incident waves, and of the transmitted and incident waves, are known as the Fresnel coefficients, with one set for normal polarization and one for parallel polarization.

QUESTIONS

Q22.1. For what orientation and position of a small wire loop, Fig. 22.1, is the emf induced in it maximal?

Q22.2. Prove that the time-average value of the Poynting vector at any point in Fig. 22.1 is zero.

Q22.3. If waves are represented in phasor form, how can you distinguish a standing wave from a traveling wave?

Q22.4. If waves are represented in the time domain, how can you distinguish a standing wave from a traveling wave?

Q22.5. Does the emf induced in a small loop of area S placed in Fig. 22.3 at a coordinate $z > 0$ depend on z? Does it depend on z if $z < 0$? Explain.

Q22.6. Can a small wire loop be placed in Fig. 22.5 so that the emf induced in it is practically zero irrespective of the orientation of the loop?

Q22.7. Repeat question Q22.6 for Fig. 22.6.

Q22.8. Is total reflection possible if a wave is incident from air onto a dielectric surface? Explain.

Q22.9. Why is there no counterpart of the Brewster angle for a wave with vector **E** normal to the plane of incidence?

Q22.10. A linearly polarized plane wave is incident from air onto a dielectric half-space, with the vector **E** at an angle $\alpha (0 < \alpha < \pi/2)$ with respect to the plane of incidence. Is the polarization of the transmitted and reflected wave linear? If not, what is the polarization of the two waves? Does it depend, for a given α, on the properties of the dielectric medium?

PROBLEMS

P22.1. A linearly polarized plane wave of rms electric field strength E and angular frequency ω is normally incident from a vacuum on a large, perfectly conducting flat sheet. Determine the induced surface charges and currents on the sheet.

P22.2. Note that the induced surface currents in problem P22.1 are situated in the magnetic field of the incident wave. Determine the time-average force per unit area (the pressure) on the sheet. (This is known as *radiation pressure.*)

P22.3. Repeat problems P22.1 and P22.2 assuming the wave is polarized circularly.

P22.4. If the conductivity σ of the sheet in problem P22.1 is large, but finite, its permeability is μ, and the frequency of the wave is ω, find the time-average power losses in the sheet per unit area. Specifically, find these losses if $f = 1\,\text{MHz}$, $E = 1\,\text{V/m}$, $\sigma = 56 \cdot 10^6\,\text{S/m}$ (copper), and $\mu = \mu_0$.

P22.5. A plane wave, of wavelength λ, is normally incident from a vacuum on a large, perfectly conducting sheet. A circular loop of radius $a (a \ll \lambda)$ should be at a location at which the induced emf is maximal, as near as possible to the sheet. If the electric field of the incident wave is E, calculate this maximal emf.

P22.6. Assume that a time-harmonic surface current of density $J_{sx} = J_{s0} \cos \omega t$ exists over an infinitely large plane sheet. Write the integral expression for the electric field strength vector at a distance z from the sheet. Do not evaluate the integral, but reconsider problem P22.1 to see if you know what the result must be.

P22.7. A linearly polarized plane wave, of frequency $f = 1\,\text{MHz}$, is normally incident from a vacuum on the planar surface of distilled water ($\mu = \mu_0$, $\epsilon = 81\epsilon_0$, $\sigma \simeq 0$). The rms value of the electric field strength of the incident wave is $E = 100\,\text{mV/m}$. A loop of area $S = 100\,\text{cm}^2$ wound with $N = 5$ turns is situated in water so that the emf induced in it is maximal. Determine the rms value of the emf.

P22.8. A plane wave propagating in dielectric 1, of permittivity ϵ_1 and permeability μ_1, impinges normally on a dielectric slab 2, of permittivity ϵ_2, permeability μ_2, and thickness d. To the right of the slab there is a semi-infinite medium of permittivity ϵ_3 and permeability μ_3. Determine the reflection coefficient at the interface between media 1 and 2. Plot the reflection coefficient as a function of the slab thickness, d, for given permittivities. Consider cases when (1) $\epsilon_1 > \epsilon_2 > \epsilon_3$, (2) $\epsilon_3 > \epsilon_2 > \epsilon_1$, (3) $\epsilon_2 > \epsilon_1 > \epsilon_3$, and (4) $\epsilon_2 > \epsilon_3 > \epsilon_1$.

P22.9. Assume that in the preceding problem the thickness of the slab is (1) half a wavelength, and (2) a quarter of a wavelength in the slab. Determine the relationship between the intrinsic impedances of the three media for which in the two cases there will be no reflected wave into medium 1. (The first of these conditions is used in antenna covers, called radomes. The second is used in optics, for so-called anti-reflection, or AR, coatings. The thickness and relative permittivity of a thin transparent layer over lenses can be designed in this way so that the reflection of light from the lens is minimized.)

P22.10. Find the reflection and transmission coefficients for the interface between air and fresh water ($\epsilon = 81\epsilon_0, \sigma \simeq 0$), in the case of perpendicular incidence.

P22.11. A plane wave is normally incident on the interface between air and a dielectric having a permeability $\mu = \mu_0$, and an unknown permittivity ϵ. The measured standing-wave ratio in air is 1.8. Determine ϵ.

P22.12. What is the position of a small loop of area S in Fig. 22.6 in order that the emf induced in it be maximal? If the electric field of the wave is E and its frequency f, calculate this maximal emf.

P22.13. Repeat problem P22.12 for a small loop placed in the wave in Fig. 22.5.

P22.14. Determine the minimal relative permittivity of a dielectric medium for which the critical angle of total reflection from the dielectric into air is less than 45 degrees. Is it possible to make from such a dielectric a right-angled isosceles triangular prism that returns the light wave as in Fig. P22.14? Is there reflection of the light wave when it enters the prism?

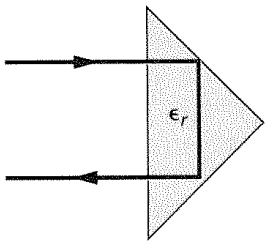

Figure P22.14 Reflection of a light wave by a prism

P22.15. A plane wave with parallel polarization is incident at an angle of $\pi/4$ from air on a perfect dielectric with $\epsilon_r = 4$ and $\mu = \mu_0$. Find the Fresnel coefficients. What fraction of the incident power is reflected, and what is transmitted into the dielectric? More generally, plot the Fresnel coefficients and the reflected and transmitted power as a function of ϵ_r, assuming its value is between 1 and 80.

P22.16. Repeat problem P22.15 for a normally polarized wave.

P22.17. A plane wave with normal polarization is incident at an angle of 60° from air onto deep fresh water with $\epsilon_r = 81$ ($\sigma = 0$). The rms value of the incident electric field is 1 V/m. Find the rms value of the reflected and transmitted electric field.

P22.18. Repeat problem P22.17 for parallel polarization.

P22.19. Is there an incident angle in problems P22.17 and P22.18 for which the reflected wave is eliminated? If so, calculate this angle for the two polarizations.

23

Waveguides and Resonators

23.1 Introduction

Waveguides are structures that direct electromagnetic energy along a desired path, transmission lines being just one example. We know that transmission lines consist of two conductors, but some may have more than two, as in three-phase power lines. Maxwell's equations predict that electromagnetic waves can also be guided through hollow metallic tubes, like water is "guided" through pipes. There are a variety of such hollow metallic waveguides, differing in the shape of their cross section; the most common shape is rectangular.

Maxwell's equations also predict that electromagnetic waves can be guided by dielectric slabs or rods, known as *dielectric waveguides*. For example, an optical fiber is a specific type of dielectric waveguide used for guiding electromagnetic waves at optical frequencies.

Transmission lines can support waves with vectors **E** and **H** in planes *transversal* (normal) to the direction of wave propagation. We know that such waves are termed *transverse electromagnetic waves*, or TEM waves. We already know that plane waves are also TEM waves, but for a plane wave the vectors **E** and **H** in transversal planes are *constant*, whereas in transmission lines they are not.

Waveguides in the form of metallic tubes and dielectric plates or rods cannot support TEM waves. Instead, waves along such waveguides may have *either* the **E** vector *or* the **H** vector in the transversal plane alone, but the other vector must have

a component in the direction of propagation. These two wave types are called *transverse electric*, or TE, waves, and *transverse magnetic*, or TM, waves.

We know that lossless transmission lines guide TEM waves of *any* frequency, and *with the same velocity*. We will see that TE and TM waves can propagate only above a certain critical frequency, and that their velocity depends on frequency. So structures supporting TE and TM waves behave as high-pass filters.

We have seen that among other purposes, transmission lines are used as circuit elements (to obtain an element with desired reactance, to act as a transformer, etc.). Waveguides are also used for such purposes, but only in the microwave range because they would be very large and impractical at low frequencies. They are used as building blocks of various microwave components: attenuators, phase shifters, transformers, and so on. We will consider only one such component, the so-called resonant cavity, which is an analogue to an *LC* resonant circuit with a lumped inductor and a lumped capacitor. A resonant cavity, however, is a *spatial* resonator, in the form of a box in which electromagnetic energy oscillates, similar to the way acoustic energy oscillates in a hallway.

The theory of waveguides is significantly more complex than any theory we have considered so far, and a complete presentatoin is beyond the scope of this introductory text. Since the waveguides are of great practical importance at higher frequencies, every electrical engineer should know at least the basic concepts of these electromagnetic structures. A compromise is therefore made in what follows, and most of the basic waveguide equations are given without proof. The interested reader can find these proofs in Appendix 8.

23.2 Wave Types (Modes)

Consider a hollow, perfectly conducting waveguide pipe, filled with a perfect dielectric of parameters ϵ and μ, as in Fig. 23.1. Let the complex vectors $\mathbf{E}$ and $\mathbf{H}$ in the waveguide be of the form $\mathbf{E}_{\text{tot}} = \mathbf{E}(x, y)e^{-\gamma z}$, and $\mathbf{H}_{\text{tot}} = \mathbf{H}(x, y)e^{-\gamma z}$. Here γ is the propagation coefficient in the z direction. First we allow γ to be complex, and later we will discuss what that physically means. After performing vector differentiation

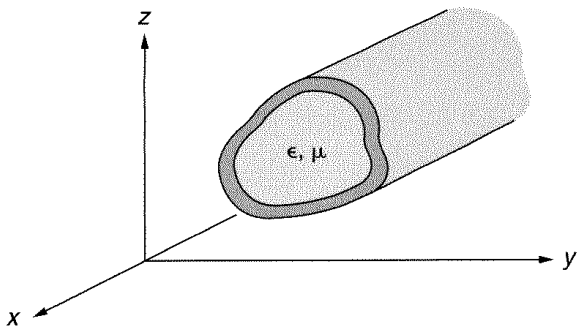

Figure 23.1 Cross section of a general waveguide. It is assumed that the waveguide is lossless, and that the dielectric is homogeneous.

on Maxwell's equations in complex form, as shown in detail in Appendix 8, the following expressions for the electric and magnetic field components are obtained:

$$E_x = -\frac{1}{K^2}\left(\gamma\frac{\partial E_z}{\partial x} + \mathrm{j}\omega\mu\frac{\partial H_z}{\partial y}\right), \tag{23.1}$$

$$E_y = -\frac{1}{K^2}\left(\gamma\frac{\partial E_z}{\partial y} - \mathrm{j}\omega\mu\frac{\partial H_z}{\partial x}\right), \tag{23.2}$$

$$H_x = -\frac{1}{K^2}\left(-\mathrm{j}\omega\epsilon\frac{\partial E_z}{\partial y} + \gamma\frac{\partial H_z}{\partial x}\right), \tag{23.3}$$

$$H_y = -\frac{1}{K^2}\left(\mathrm{j}\omega\epsilon\frac{\partial E_z}{\partial x} + \gamma\frac{\partial H_z}{\partial y}\right), \tag{23.4}$$

where

$$K^2 = \gamma^2 + \beta^2, \qquad \beta^2 = \omega^2\epsilon\mu. \tag{23.5}$$

That Eqs. (23.1) to (23.5) are solutions to Maxwell's equations can be checked by substitution. Note that the propagation coefficient γ (and therefore also the coefficient K) *is not known*. So there are seven scalar unknowns (the six field components and γ).

We can reach an interesting conclusion by looking carefully at the preceding equations: if we can find $E_z(x, y)$ and $H_z(x, y)$, we know the complete electromagnetic field for a given waveguide shape and size. [Thus the functions $E_z(x, y)$ and $H_z(x, y)$ play a role analogous to a potential function because all the other components are obtained from them by differentiation.]

These equations have three *classes* of solution:

1. Both $E_z = 0$ and $H_z = 0$, that is, only transversal components of the wave exist. [The possibility of such a solution is not evident from Eqs. (23.1) to (23.5), but will be demonstrated in the next section.] This solution corresponds to a TEM wave.
2. $E_z = 0$, but $H_z \neq 0$. This solution corresponds to a transverse electric (TE) wave.
3. $E_z \neq 0$, and $H_z = 0$. This solution corresponds to a transverse magnetic (TM) wave.

We now examine these three classes of solutions, often called *modes*, in turn.

23.2.1 TRANSVERSE ELECTROMAGNETIC (TEM) WAVES

The salient properties of TEM wave types, or TEM modes, are the TEM propagation coefficient γ, the wave impedance Z_{TEM}, and the unique quasi-static nature of TEM wave types.

Propagation Coefficient

If both $E_z = 0$ and $H_z = 0$, the expressions in parentheses in Eqs. (23.1) to (23.4) are zero. One might be tempted to conclude that all the other components are also zero.

Note, however, that *K is not known*, so it can also be zero. We know that the expression of the form 0/0 need not be undefined (for example, $\sin x/x \to 1$ if $x \to 0$). So the solution could exist only if $K^2 = \gamma^2 + \beta^2 = 0$, or

$$\gamma = \pm j\omega\sqrt{\epsilon\mu}. \tag{23.6}$$

(Propagation coefficient of TEM waves)

We recognize in γ the propagation coefficient of plane waves, and also the propagation coefficient along lossless transmission lines. We will see that indeed, waves propagating along lossless transmission lines *are* of the TEM type.

Wave Impedance

In Eqs. (23.1) and (23.4) let $\gamma = \pm j\omega\sqrt{\epsilon\mu}$. After simple manipulations we find that in such a case, the ratio of the transverse electric and magnetic field components is

$$\frac{E_x}{H_y} = \pm Z_{\text{TEM}}, \quad \frac{E_y}{H_x} = \mp Z_{\text{TEM}}, \quad \text{where} \quad Z_{\text{TEM}} = \sqrt{\frac{\mu}{\epsilon}} \tag{23.7}$$

(Wave impedance of TEM waves)

for any E_z and H_z (which cancel out). Z_{TEM} is known as the *wave impedance of TEM waves*.

From this we can draw three conclusions. First, vector **H** is normal to vector **E** (both are, of course, in a transverse plane). Second, the ratio of the electric and magnetic fields for a forward wave (the upper sign) is the intrinsic impedance of the medium, and for the backward wave it is the negative of that. Third, for the forward and for the backward waves **E** and **H** are such that their cross product results in the Poynting vector (power flow) in the respective direction. How do these properties compare to those of a plane wave in free space?

Quasi-Static Nature of TEM Waves

There is an interesting general conclusion about TEM wave types. It turns out (see Appendix 8, section A8.2) that the electric field is derivable from a potential function which at $z = 0$ satisfies Laplace's equation in x and y:

$$\frac{\partial^2 V(x, y)}{\partial x^2} + \frac{\partial^2 V(x, y)}{\partial y^2} = 0. \tag{23.8}$$

Because boundary conditions require that the tangential **E** on conductor surfaces be zero, we reach the following conclusion: for TEM waves, the electric field in planes where z is constant is the same as the electrostatic field corresponding to the potentials of waveguide conductors at that cross section.

Example 23.1—A TEM wave cannot propagate through a hollow metal tube. Consider a waveguide in the form of a hollow metal tube. For a TEM wave to exist inside the tube, the field must be the same as in the electrostatic case. But we know that inside a hollow conductor

with no charges there can be no electrostatic field. Consequently, TEM waves cannot propagate through hollow metallic waveguides.

Note that a coaxial cable does have another conductor in the tube, and that an electrostatic field can exist in the cable if the two cable conductors are at different potentials. Therefore, a TEM wave can propagate inside a coaxial cable (which we already know is true).

Example 23.2—Transmission lines must have at least two conductors. For the electrostatic field to exist in a cylindrical system, we must have at least two conductors. Indeed, a single charged conductor is a fiction—it implies infinite electrical energy per unit length, since the potential of the conductor with respect to a reference point at infinity is infinite. So TEM waves cannot propagate along a single wire.

However, any electrostatic system of two or more conductors *with a zero total charge per unit length* is feasible, because it has a finite electrical energy per unit length. Consequently, TEM waves can propagate along such waveguides. Note that this also implies a zero total current at any cross section of a transmission line, a proof of which is left as an exercise for the reader. Thus, equations of TEM waves propagating along waveguides are actually equations of wave propagation along lossless transmission lines.

23.2.2 TRANSVERSE ELECTRIC (TE) WAVES

Now let us briefly examine the general properties of TE wave types (for details, see Appendix 8, section A8.3).

Propagation Coefficient

Using the condition $E_z(x, y) = 0$, the wave equation for the H_z component in this case is given by

$$\frac{\partial^2 H_z}{\partial x^2} + \frac{\partial^2 H_z}{\partial y^2} + \gamma^2 H_z + \omega^2 \epsilon \mu H_z = \frac{\partial^2 H_z}{\partial x^2} + \frac{\partial^2 H_z}{\partial y^2} + K^2 H_z = 0. \tag{23.9}$$

To solve this equation we need to know the geometry of the waveguide. We will see that for specific boundary conditions this equation can be satisfied only for certain distinct values of the parameter K. These values of K, for which both the wave (Helmholtz) equation and boundary conditions are satisfied, are known as its *eigenvalues*, or *characteristic values*. We will see that, for example, eigenvalues of K for a rectangular waveguide are given by a double infinite set of pairs of numbers dependent on the waveguide dimensions and frequency. An eigenvalue of K determines the propagation coefficient γ according to Eq. (23.5).

Wave Impedance

From Eqs. (23.1) to (23.4) it follows that if $E_z = 0$, the *transverse electric and magnetic field vectors in a TE wave are normal to each other*, and that

$$Z_{\text{TE}} = \frac{E_x}{H_y} = -\frac{E_y}{H_x} = \frac{\mathrm{j}\omega\mu}{\gamma} \tag{23.10}$$

(Wave impedance of TE waves)

is a constant, the same at all points of the field in a waveguide. This is known as the *wave impedance of TE modes*. We will see that it is *not* the same as that for TEM waves, because γ for TE waves is different from $j\omega\sqrt{\epsilon\mu}$.

23.2.3 TRANSVERSE MAGNETIC (TM) WAVES

Finally, let us look at the general properties of TM wave types.

Propagation Coefficient

Using the condition $H_z(x, y) = 0$, we can obtain E_z from the Helmholtz equation, which now reads

$$\frac{\partial^2 E_z}{\partial x^2} + \frac{\partial^2 E_z}{\partial y^2} + K^2 E_z = 0. \tag{23.11}$$

To solve this equation we need to know the geometry of the waveguide, and a solution exists only for specific values of K (its eigenvalues), as in the case of TE modes.

Wave Impedance

As in the case of TE modes, from Eqs. (23.1) to (23.4) it follows that for $H_z = 0$, the transverse electric and magnetic field vectors in a TM wave are normal to each other, and that

$$\frac{E_x}{H_y} = -\frac{E_y}{H_x} = Z_{\text{TM}} = \frac{\gamma}{j\omega\epsilon} \tag{23.12}$$

(Wave impedance of TM waves)

is the same at all points. This is known as the *wave impedance of TM wave types*.

Note that for a forward wave and the same value of the propagation coefficient γ,

$$Z_{\text{TE}} Z_{\text{TM}} = Z_{\text{TEM}}^2 = \frac{\mu}{\epsilon}. \tag{23.13}$$

Example 23.3—Power transmitted along waveguides. Let us now derive a general expression for the power transmitted along a waveguide. Let only a forward wave exist in a waveguide. The power transmitted along the waveguide can be determined by integrating the complex Poynting vector over the structure cross section at $z = 0$:

$$P = \int_{S_{\text{transv}}} \text{Re}\{(\mathbf{E}_{\text{transv}} \times \mathbf{H}^*_{\text{transv}}) \cdot \mathbf{u}_z\}\, dS_{\text{transv}}, \tag{23.14}$$

where the subscript "transv" relates to components normal to the direction of propagation.

The transverse components of vectors **E** and **H** for all three wavetypes (TEM, TE, and TM) are normal to each other. In addition, their ratio equals the wave impedance of the wave,

and the cross product $\mathbf{E}_{\text{transv}} \times \mathbf{H}^*_{\text{transv}}$ is in the direction of vector $\mathbf{u}_z$. So

$$(\mathbf{E}_{\text{transv}} \times \mathbf{H}^*_{\text{transv}}) \cdot \mathbf{u}_z = E_{\text{transv}} H^*_{\text{transv}} = Z_{\text{wave type}} |H_{\text{transv}}|^2 = \frac{1}{Z_{\text{wave type}}} |E_{\text{transv}}|^2,$$

where $Z_{\text{wave type}}$ is the wave impedance of the wave propagating along the waveguide. Thus for the power transmitted along the waveguide we obtain

$$\begin{aligned} P_{\text{wave type}} &= Z_{\text{wave type}} \int_{S_{\text{transv}}} |H_{\text{transv}}|^2 \mathrm{d}S_{\text{transv}} \\ &= \frac{1}{Z_{\text{wave type}}} \int_{S_{\text{transv}}} |E_{\text{transv}}|^2 \mathrm{d}S_{\text{transv}} \qquad \text{(valid for forward wave),} \end{aligned} \tag{23.15}$$

where "wave type" stands for TEM, TE, or TM.

Questions and problems: Q23.1 to Q23.7, P23.1 to P23.4

23.3 Rectangular Metallic Waveguides

At frequencies above about 1 GHz, losses in transmission-line conductors due to skin effect become pronounced, so lower-loss hollow waveguides are used for guiding waves up to frequencies of several hundred gigahertz. Most often such waveguides are of rectangular cross section, but circular and some other cross sections are also used. We restrict our attention to rectangular waveguides.

A sketch of a rectangular waveguide is shown in Fig. 23.2. We assume the waveguide to be lossless and straight. From Example 23.1 we know that TEM waves cannot propagate along hollow waveguides. So we consider TE modes in more detail, and also TM wave types briefly. Let us start with the TE wave types.

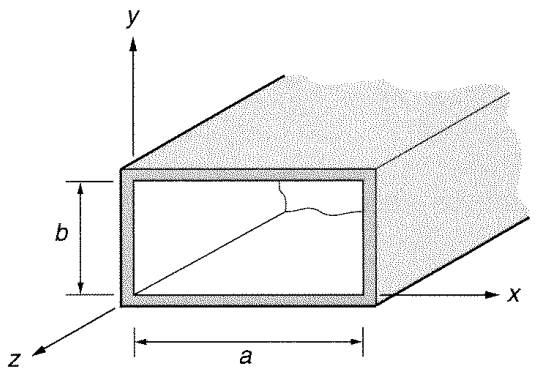

Figure 23.2 Sketch of a waveguide of rectangular cross section

23.3.1 TE WAVES IN RECTANGULAR WAVEGUIDES

The complete derivation for TE modes in a rectangular metallic waveguide is given in Appendix 8, section A8.3. At the introductory level, it suffices to quote the most important expressions and discuss their practical meaning.

Complete Expression for TE_{mn} Wave Types

After applying the boundary conditions to the perfectly conducting waveguide walls at $x = 0$, $x = a$, $y = 0$, and $y = b$, the H_z component in the cross section $z = 0$ of the waveguide is found to be

$$H_z(x, y) = H_0 \cos\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right) \qquad (\text{at } z = 0), \tag{23.16}$$

where H_0 is a constant depending on the level of excitation of the wave in the waveguide. The other components at $z = 0$ are (note that $E_z = 0$)

$$E_x(x, y) = \frac{j\omega\mu}{K^2}\frac{n\pi}{b}H_0 \cos\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right) \qquad (\text{at } z = 0), \tag{23.17}$$

$$E_y(x, y) = -\frac{j\omega\mu}{K^2}\frac{m\pi}{a}H_0 \sin\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right) \qquad (\text{at } z = 0), \tag{23.18}$$

$$H_x(x, y) = \frac{\gamma}{K^2}\frac{m\pi}{a}H_0 \sin\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right) \qquad (\text{at } z = 0), \tag{23.19}$$

$$H_y(x, y) = \frac{\gamma}{K^2}\frac{n\pi}{b}H_0 \cos\left(\frac{m\pi}{a}x\right)\sin\left(\frac{n\pi}{b}y\right) \qquad (\text{at } z = 0). \tag{23.20}$$

Since the cosine and sine functions are periodic, we see that there is a double infinite number of TE wave types, corresponding to any possible pair of m and n. Note that m represents the number of half-waves along the x axis, and n the number of half-waves along the y axis. The wave determined by m and n is known as a *TE_{mn} mode*. From Eqs. (23.16) to (23.20) we see that for $m = n = 0$ all components are zero. Thus, a TE_{00} mode does not exist. Waves for any other combinations of numbers m and n may exist, for example, TE_{10}, TE_{01}, TE_{11}, TE_{21}. The values of the wave components for any z are obtained by simply multiplying the preceding expressions by $e^{-\gamma z} = e^{-j\beta z}$, where γ and β are as given in the next section.

Propagation Coefficient of TE Waves

Noting that $\omega = 2\pi f$, the expression for the propagation coefficient of a TE wave along a rectangular waveguide is

$$\gamma = j\beta \qquad \beta = \omega\sqrt{\epsilon\mu}\sqrt{1 - \frac{f_c^2}{f^2}}, \tag{23.21}$$

(Propagation and phase coefficients of rectangular waveguides)

where f_c is known as the *cutoff frequency*, and it depends on the mode numbers m and n, and on the dimensions of the waveguide, a and b.

Cutoff Frequency of TE Wave Types

Noting that $1/\sqrt{\epsilon\mu} = c$,

$$f_c = \frac{c}{2}\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2}, \qquad c = \frac{1}{\sqrt{\epsilon\mu}}. \tag{23.22}$$

(Cutoff frequency of rectangular waveguides)

Why f_c is known as the cutoff frequency is explained next.

Phase and Group Velocity of TE_{mn} Waves

Let us now investigate more closely the properties of TE_{mn} modes for different pairs of values of m and n. First, note that the *phase velocity* of the TE_{mn} mode is given by

$$v_{\text{ph}} = \frac{\omega}{\beta} = \frac{c}{\sqrt{1 - f_c^2/f^2}}. \tag{23.23}$$

(Phase velocity of waves propagating along rectangular waveguides)

You may recall Example 21.5, in which we showed that for this dependence of the phase velocity on frequency, the group velocity is given by

$$v_g = c\sqrt{1 - f_c^2/f^2}. \tag{23.24}$$

(Group velocity of waves propagating along rectangular waveguides)

Since the phase (and group) velocity depend on frequency, rectangular waveguides are *dispersive structures*. The dependence of the phase and group velocity on normalized frequency, f/f_c, is sketched in Fig. 23.3.

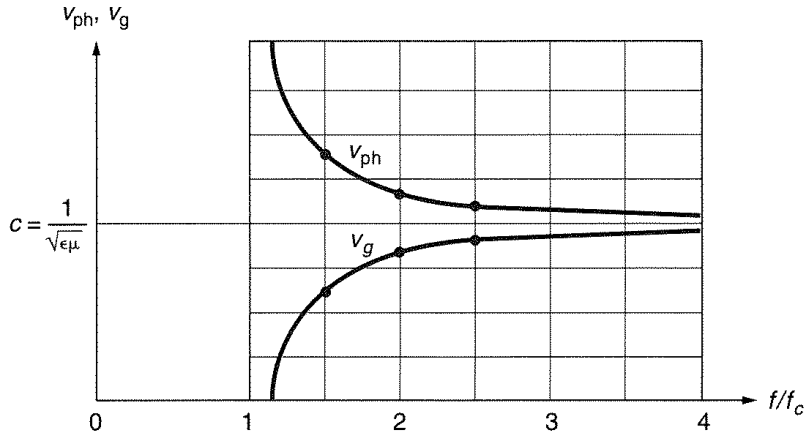

Figure 23.3 Dependence of phase velocity and group velocity on normalized frequency, f/f_c

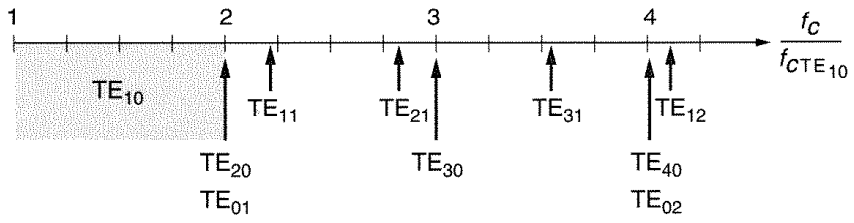

Figure 23.4 Cutoff frequencies of the first few higher-order TE modes normalized to that of the dominant TE_{10} mode, for $a/b = 2$

Assume that for a given waveguide $f > f_c$. Then the phase coefficient is real, as are the phase velocity and the group velocity. This means that the wave of this frequency propagates along the waveguide.

If, however, $f < f_c$, the expression under the square root in Eqs. (23.21), (23.23), and (23.24) is *negative*. The phase coefficient becomes imaginary, $-j|\beta|$ (negative value of the root is taken to avoid exponentially increasing wave amplitudes with increasing z), so that the propagation factor $e^{-j\beta z}$ becomes $e^{-|\beta|z}$. *This means that waves of frequencies lower than f_c cannot propagate along rectangular waveguides.* This is why f_c is termed the "cutoff frequency." Because the attenuation of the wave is exponential, the wave of a frequency $f < f_c$ is attenuated very rapidly with z. Thus, as mentioned in the introduction to this chapter, rectangular waveguides behave as high-pass filters.

Modes that propagate through a given waveguide are the *propagating modes*, and those that do not propagate are the *evanescent modes*. To transmit energy, we use propagating modes. Evanescent modes are used, for example, when making an attenuator out of a section of a waveguide.

Rectangular waveguides are always made such that $a > b$. Let $a = 2b$, which is fairly standard. The first few cutoff frequencies of the TE_{mn} modes, Eq. (23.22), relative to the cutoff frequency of the TE_{10} mode are shown in Fig. 23.4. Note that between the cutoff frequency of the TE_{10} mode and the next one, that of the TE_{01} mode, only the TE_{10} mode can propagate. This is a remarkable property of the TE_{10} mode. A discontinuity in the waveguide, like a bend or a slot in the guide wall, will always produce a multitude of modes. If we use a frequency of a wave to be within these limits (shown shaded in Fig. 23.4), out of all of these modes only the TE_{10} mode will propagate further—all other modes will be evanescent.

23.3.2 TM WAVES IN RECTANGULAR WAVEGUIDES

As in the case of TE modes, there are an infinite number of TM_{mn} modes, corresponding to all possible pairs of numbers m and n. The expressions for the cutoff frequency, propagation coefficient, phase velocity, etc., are very similar to the TE case. There is an important difference, however: the lowest TM mode is TM_{11}, that is, there is no TM mode for which either $m = 0$ or $n = 0$. As an illustration, Fig. 23.5b shows the comparison of the lowest order TE mode fields and the TM_{11} mode fields.

23.4 TE_{10} Mode in Rectangular Waveguides

We have seen that the cutoff frequencies of all TE modes higher than the 10 mode, as well as the cutoff frequencies of all TM modes, are higher than that of the TE_{10} mode. For this reason the TE_{10} mode is known as the *dominant mode* in rectangular waveguides. It is by far the most commonly used wave type in hollow metallic waveguides, so we consider it in more detail.

The field components of the TE_{10} mode are given by Eqs. (23.16) to (23.20) for $m = 1$ and $n = 0$:

$$H_z(x, y) = H_0 \cos\left(\frac{\pi}{a}x\right) \quad (\text{TE}_{10}\text{ mode}), \tag{23.25}$$

$$E_x(x, y) = E_z(x, y) = H_y(x, y) = 0 \quad (\text{TE}_{10}\text{ mode}), \tag{23.26}$$

$$E_y(x, y) = -\mathrm{j}\omega\mu\frac{a}{\pi}H_0 \sin\left(\frac{\pi}{a}x\right) \quad (\text{TE}_{10}\text{ mode}), \tag{23.27}$$

$$H_x(x, y) = \mathrm{j}\beta\frac{a}{\pi}H_0 \sin\left(\frac{\pi}{a}x\right) \quad (\text{TE}_{10}\text{ mode}). \tag{23.28}$$

(Field components of TE_{10} *mode in a rectangular waveguide)*

The cutoff frequency of the TE_{10} mode is

$$f_{c\,\text{TE10}} = \frac{c}{2a}, \tag{23.29}$$

(Cutoff frequency of TE_{10} *mode in rectangular waveguide)*

and the phase velocity, wave impedance, and so on for the TE_{10} mode are obtained from the general expressions with this cutoff frequency.

The wave impedance of the TE_{10} mode is obtained from Eqs. (23.10), (23.21), and (23.29):

$$Z_{\text{TE10}} = \frac{\sqrt{\mu/\epsilon}}{\sqrt{1 - f_c^2/f^2}} > \sqrt{\mu/\epsilon}, \tag{23.30}$$

where $f_c = c/2a$. This expression tells us that the field inside the waveguide is different from that in free space (a plane wave). Consequently, if we cut a waveguide, only part of the energy will be radiated from its open end, and the rest will be reflected back.

Example 23.4—Wavelength along waveguide. The wavelength inside the waveguide, along the z axis, is determined simply as $\lambda_z = 2\pi/\beta$, where β is the phase coefficient of the mode of interest. So the wavelength along the waveguide is

$$\lambda_z = \frac{2\pi}{\beta} = \frac{c}{f\sqrt{1 - f_c^2/f^2}} = \frac{\lambda_0}{\sqrt{1 - f_c^2/f^2}}, \tag{23.31}$$

(Wavelength along a rectangular waveguide)

where λ_0 is the wavelength of a plane wave of the same frequency and in the same medium.

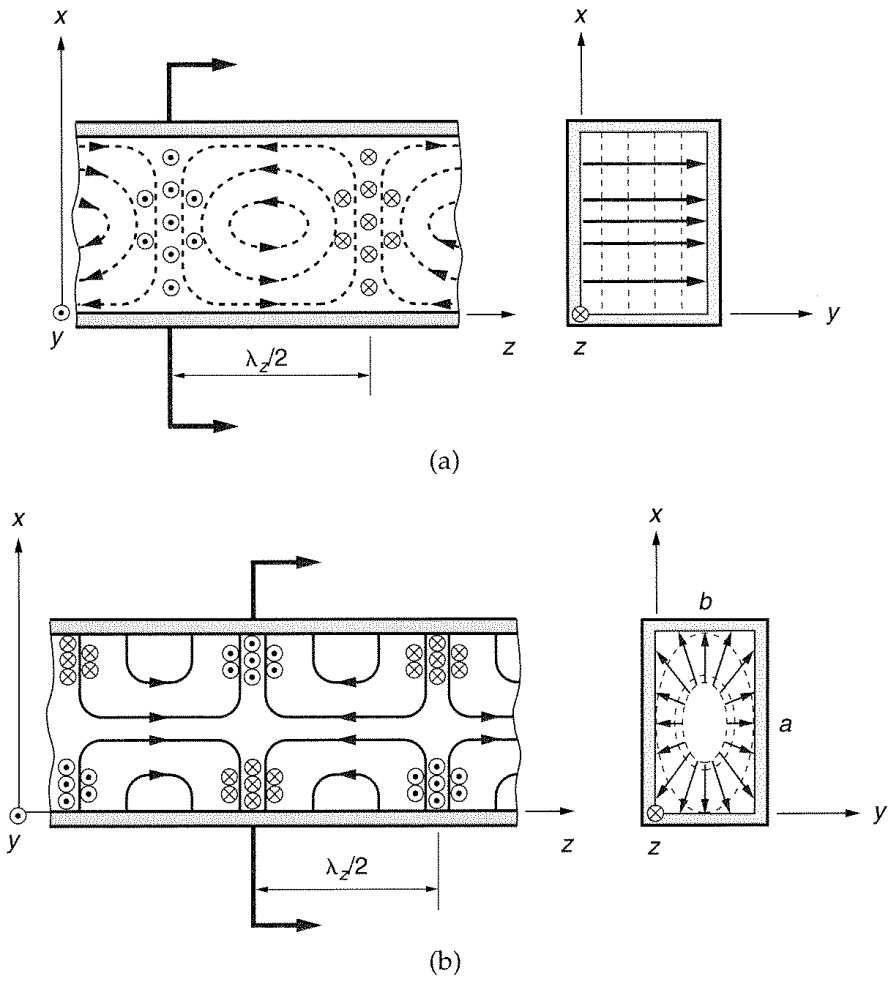

Figure 23.5 (a) Sketch of the E-field and H-field lines of the TE_{10} mode in a rectangular waveguide, frozen in time. (b) Sketch of the E and H lines of the TM_{11} mode (the mode with the lowest cutoff frequency of all TM modes). Solid lines show the E lines, and dashed lines and $\odot$ and $\otimes$ symbols show the H lines. The entire picture moves at the phase velocity in the $+z$ direction.

Example 23.5—Sketch of the field of the TE_{10} mode. When we multiply Eqs. (23.25) to (23.28) by $e^{-j\beta z}$, we obtain the phasor wave components at any point (x, y, z). To obtain a picture of the fields in the waveguide at an instant, we need to obtain the time-domain expressions, fix an instant in time (e.g., $t = 0$), and then plot the field lines. Although time-domain expressions are easy to obtain (this is left as an exercise for the reader), plotting the field is not simple. A sketch of the fields of the TE_{10} mode is shown in Fig. 23.5a. In time, the entire picture moves in the $+z$ direction with the phase velocity of the TE_{10} mode. For comparison, the E and H lines of the TM_{11} mode (the mode with the lowest cutoff frequency of all TM modes) are sketched in Fig. 23.5b.

Example 23.6—Surface current distribution on waveguide walls for the TE_{10} mode. The surface currents are obtained from the time-domain expressions of the magnetic field on waveguide walls and the boundary condition in Eq. (19.12) for a perfect conductor. We need to

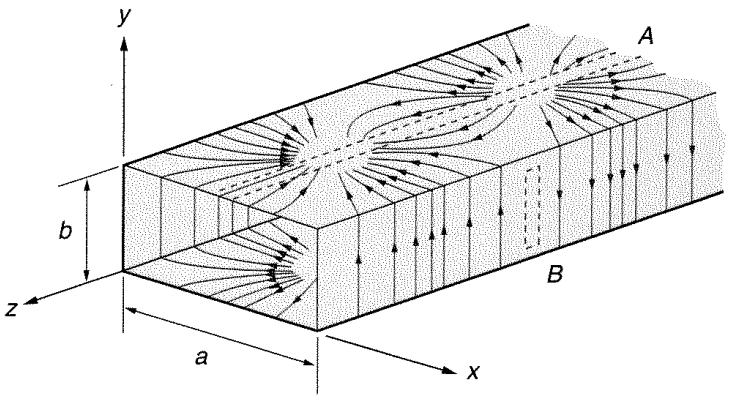

Figure 23.6 Sketch of surface current distribution for the TE_{10} mode in a rectangular waveguide, frozen in time. The entire picture moves at the phase velocity in the $+z$ direction.

fix an instant of time (e.g., $t = 0$), and then plot the lines of the surface-current density vector. This, again, is not an easy task. A sketch of the lines of the current density vector is shown in Fig. 23.6. In time, the surface current density distribution moves with the phase velocity of the TE_{10} mode in the $+z$ direction.

Note that if we cut a slot in the waveguide wall, in such a way that the slot is always tangential to the lines of the surface current, only a small disturbance of the wave propagation in the waveguide will result. Therefore we can cut narrow slots of types *A* and *B* indicated in the figure without changing the fields in the waveguide. The slot of type *A* is made to obtain a slotted waveguide used for measurements similar to those done by a slotted coaxial line (Example 18.10).

Example 23.7—Excitation of TE_{10} mode. How can we produce a TE_{10} mode? First, we need to close one end of the waveguide, to prevent propagation in both directions. We do this with a metal plate perpendicular to the waveguide, as in Fig. 23.7. In waveguide terminology, this is known as a *shorted waveguide.*

To excite the TE_{10} mode, we can excite either the *E* field or the *H* field. The *E* field can be excited by a small coaxial probe. A short extension of the inner conductor of a coaxial line is inserted into the waveguide, and the outer conductor of the line is connected to the waveguide wall, as in Fig. 23.7. Roughly, the position of the probe should be in the middle of the wave-

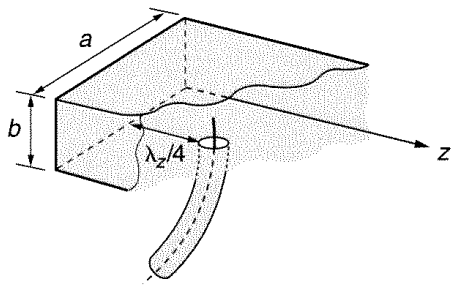

Figure 23.7 Sketch of a probe for exciting a TE_{10} mode in a rectangular waveguide

guide (where the E field is the strongest, Fig. 23.5), and about a quarter of a guided wavelength λ_z from the short circuit. If the latter condition is fulfilled, the wave from the probe needs one quarter of a period to reach the short circuit, is reflected there and changes phase (which is equivalent to losing another two quarters of a period), and then loses another quarter of a period to go back to the probe. So the wave reflected from the short circuit will be in phase with the wave radiated in the $+z$ direction. This simple reasoning, however, is only a rough estimate, and the actual position of the probe needs to be determined either experimentally or using accurate numerical methods.

Another possibility is to excite the TE_{10} mode by a small loop intended to excite the H field at the place where it is the strongest, e.g., in the middle of the waveguide short circuit. It is left as an exercise for the reader to sketch this type of waveguide excitation.

Example 23.8—TE_{10} mode in X-band waveguide. The microwave frequency range is divided into so-called bands, and a waveguide of certain dimensions can support waves at frequencies covering the entire band. A commonly used range is the X band (about 8.2 to 12.4 GHz), for which a standard waveguide has $a = 23$ mm and $b = 10$ mm. Using the formulas for TE_{10} modes with these waveguide dimensions, we find that the cutoff frequency is $f_c = 6.52$ GHz. The guided wavelength and impedance are different at different frequencies inside the band. At the center of the band, $f = 10$ GHz, the guided wavelength $\lambda_g = 3.96$ cm, and the impedance $Z_{TE10} = 497\,\Omega$. So, the characteristic impedance is much larger than that of a coaxial cable, which is usually $50\,\Omega$. That means that the probe described in Example 23.7 needs to match the coaxial impedance to that of the waveguide dominant mode.

Example 23.9—Power transmitted by the TE_{10} mode. The power transmitted by a forward wave through any waveguide is given in Eq. (23.15). We know the transverse components and the wave impedance of the TE_{10} mode, so we need only to substitute these expressions into Eq. (23.15) and to integrate over the cross-sectional area of the waveguide,

$$P_{TE10} = \frac{\beta}{\omega\mu}\int_{y=0}^{b}\int_{x=0}^{a}\omega^2\mu^2\frac{a^2}{\pi^2}|H_0|^2\sin^2\left(\frac{\pi}{a}x\right)\mathrm{d}x\,\mathrm{d}y = \frac{\omega\mu\beta a^2|H_0|^2}{\pi^2}\,b\,\frac{a}{2} = \frac{\omega\mu\beta a^3 b|H_0|^2}{2\pi^2}.$$

H_0 is the rms value of the magnetic field at the magnetic field maximum. If β is replaced by its expression in Eq. (23.21), this becomes

$$P_{TE10} = \frac{ab}{2}\sqrt{\frac{\mu}{\epsilon}}\frac{f^2}{f_c^2}\sqrt{1-\frac{f_c^2}{f^2}}\,|H_0|^2. \tag{23.32}$$

It is very instructive to evaluate this power for a specific case. Let the frequency be $f = 10$ GHz, and let $a = 2$ cm and $b = 1$ cm. The cutoff frequency for this waveguide for the TE_{10} mode is $f_c = c/2a = 3\cdot 10^8/0.04 = 7.5$ GHz. Let us calculate the maximal possible power that can be transmitted through this waveguide if it is filled with air of dielectric strength $E_{max} = 30$ kV/cm. The maximal electric field is at the coordinate $x = a/2$. At that point, the electric field *amplitude* has to be less than E_{max}. This enables us to calculate the maximal H_0 from Eq. (23.27):

$$H_{0max} = \frac{\pi E_{max}/\sqrt{2}}{\omega\mu a}.$$

Substituting this value of H_0 into Eq. (23.32), we obtain that the maximal power that can be transmitted is about 800 kW. This is much more than the power that could be transmitted through a coaxial cable or printed line (why?), so waveguides are the guiding medium of choice for high-power applications, such as some radars.

Questions and problems: Q23.8 to Q23.22, P23.5 to P23.10

23.5 The Microstrip Line (Hybrid Modes)

We have discussed in detail only one of the TE and briefly one of the TM modes in a rectangular waveguide. There are an infinite number of other TE and TM modes in such guiding structures. However, other structures can support wave types that are a combination of TEM, TE, and TM modes. These wave types are called *hybrid modes*.

As an illustration, we consider a commonly used waveguide, called a *microstrip line*, sketched in Fig. 23.8. A microstrip line is made on a dielectric slab, called the *substrate*. One side of the substrate is coated with metal and acts as the ground electrode, similar to the outer conductor of a coax. A metal strip on the other side of the substrate enables a wave to propagate mostly in the dielectric. The role of the strip is similar to that of the center conductor of a coax. Unlike the coax, however, this structure has an inhomogeneous dielectric. The fields are not contained completely in the dielectric but are partly in air, as sketched in Fig. 23.8. This electric field is usually referred to as a fringing field.

Because there are two conductors in this guide, according to Example 23.2 we may suspect that a microstrip can support a TEM wave. Let us check if this is possible. The boundary condition for the (fringing) tangential electric field tells us that $E_{x,\text{diel}} = E_{x,\text{air}}$. We replace E_x with spatial derivatives of $\mathbf{H}$ from Maxwell's second equation in differential form, $\mathbf{E} = -\mu\partial(\nabla \times \mathbf{H})/\partial t$. Taking into account the boundary condition for the magnetic flux density vector (normal components equal on the two sides of

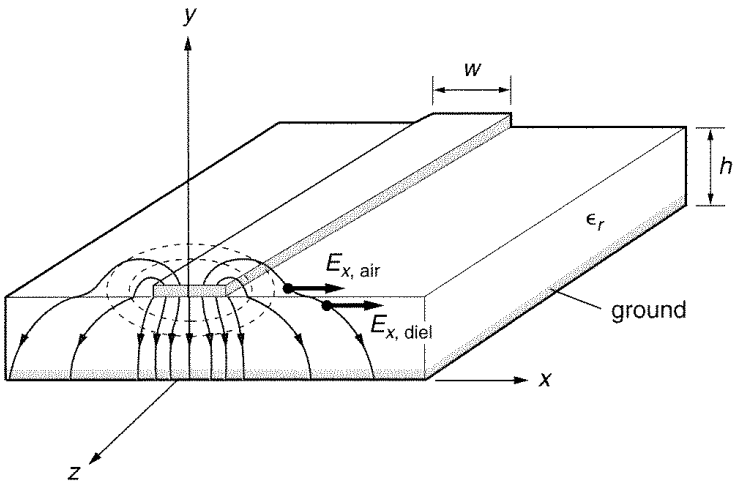

Figure 23.8 Sketch of a microstrip line. The electric field lines are sketched in solid line, and the magnetic field lines in dashed line.

the boundary), the boundary condition for the electric field can be written as (see problem P23.11)

$$\epsilon_r \frac{\partial H_{z,\text{air}}}{\partial y} - \frac{\partial H_{z,\text{diel}}}{\partial y} = (\epsilon_r - 1)\frac{\partial H_y}{\partial z}. \tag{23.33}$$

Let us see what this boundary condition tells us. The right-hand side is not zero, because $\epsilon_r > 1$, and H_y is not zero. That means that the left-hand side is not zero, which means that H_z is not zero. It can be shown in a similar manner that E_z cannot be zero. So if Maxwell's equations are satisfied for this structure, the wave type that propagates has to have nonzero H_z and E_z components, which means it contains TE and TM modes, and is a hybrid mode.

The components of the electric and magnetic field vectors along the direction of propagation are small compared to the other components, and this structure supports a wave type referred to as a "quasi-TEM" mode, similar to a TEM mode. This means that we define a characteristic impedance and a propagation constant, and then use TEM mode, or transmission-line equations. These line parameters are expressed in terms of an *effective dielectric constant*, which depends on the relative permittivity and thickness of the substrate (see problems P23.12 and P23.13).

Microstrip lines are used extensively at microwave frequencies because of their small size and ease of manufacturing (using printed-circuit board technology). Their loss is higher than that in waveguides, so they are not used for high power levels.

Questions and problems: Q23.23 and Q23.24, P23.11 to P23.13

23.6 Electromagnetic Resonators

Classical resonant circuits with lumped elements cannot be used above about 100 MHz. On one hand, losses due to skin effect and dielectric losses become very pronounced. On the other hand, the circuit needs to be sufficiently small not to radiate energy. Therefore at high frequencies, instead of resonant circuits, closed (usually air-filled) metallic structures are used, *inside* which the electromagnetic field is excited to oscillate. Between about 500 MHz and 3 GHz, resonators are usually in the form of shorted segments of shielded transmission lines (e.g., coaxial line, shielded two-wire line). From about 3 GHz to a few tens of GHz, metallic boxes of various shapes are often used instead (most often in the form of a parallelepiped or circular cylinder). Such boxes are known as *cavity resonators*. We have seen in Example 22.1 that at still higher frequencies we use so-called Fabry-Perot resonators, consisting of two parallel, highly polished metal plates.

The basic parameters of an electromagnetic resonator are its *resonant frequency*, f_r, the type of wave inside it, and its *quality factor*, Q.

The resonant frequency and the type of wave depend on the resonator shape, size, and excitation, while the quality factor can be defined in general terms. It is defined as the ratio of the electromagnetic energy contained in the resonator, W_{em}, and the total energy lost in one cycle $W_{\text{lost/cycle}}$ in the resonator containing this energy, multiplied by 2π. Since the cycle duration at resonance is $T_r = 1/f_r$, where f_r is the

resonant frequency of the resonator, $W_{\text{lost/cycle}} = P_{\text{losses}} \cdot T_r = P_{\text{losses}}/f_r$. So the Q factor can be written in the following two equivalent forms:

$$Q = 2\pi \frac{W_{\text{em}}}{W_{\text{lost/cycle}}} = \omega_{\text{r}} \frac{W_{\text{em}}}{P_{\text{losses}}} \qquad \text{(dimensionless)} \tag{23.34}$$

(General definition of Q factor of electromagnetic resonators)

Example 23.10—Q factor of an *LC* circuit. The general definition of the Q factor is valid for resonant circuits also. Consider a parallel connection of a capacitor of capacitance C and a coil of inductance L. Let the (small) series resistance of the circuit be R. Provided that losses are small, we know that the angular resonant frequency of the circuit is $\omega_r = 1/\sqrt{LC}$. We also know that energy oscillates between that in the capacitor and that in the coil. When the energy is completely in the coil, the current in the circuit is maximal, for example, I_m. The magnetic energy contained in the coil at that instant is the energy of the circuit, and is simply

$$W_{\text{em}} = \frac{1}{2} L I_m^2.$$

Time-average Joule's losses in the circuit corresponding to this current *amplitude* are

$$P_{\text{losses}} = \frac{1}{2} R I_m^2.$$

Thus the circuit Q factor is $Q = \omega_r L/R$, as defined in circuit theory. It is almost impossible to obtain a resonant circuit with a Q factor greater than about 100. (Why do you think this is so? See question Q23.25.)

23.6.1 TRANSMISSION-LINE SEGMENTS AS ELECTROMAGNETIC RESONATORS

Consider a two-conductor lossless transmission line segment of length ζ shorted at its end. We assume the ζ axis to be directed from the shorted end toward the generator, as in Chapter 18. The input impedance of the segment was derived in Example 18.6:

$$Z(\zeta) = \mathrm{j}Z_0 \tan(\beta\zeta) = \mathrm{j}Z_0 \tan\left(\frac{2\pi}{\lambda}\zeta\right), \qquad \lambda = \frac{c}{f}, \qquad c = \frac{1}{\sqrt{L'C'}}.$$

We see that $Z(\zeta) = 0$ for $\zeta = n\lambda/2$, $n = 1, 2, \ldots$. This means that a shorted transmission line connected to a generator can be short-circuited at any such point (cross section) without affecting the voltage and current along the line. We can even cut off such a section (shorted at both ends) of the *excited* line, and the current and voltage along it will not be affected. Thus, we obtained an electromagnetic resonator of length $x = n\lambda/2$.

Assume that the rms value of the voltage of the forward wave is V_+. The rms value of the forward current wave is $I_+ = V_+/Z_0$. We know from Example 18.3 that the voltage reflection coefficient in this case is -1. Therefore, the voltage and current distribution along the line segment, given in Eqs. (18.22), in this case become

$$V(\zeta) = \mathrm{j}2V_+ \sin\beta\zeta, \qquad I(\zeta) = 2I_+ \cos\beta\zeta, \tag{23.35}$$

which represent standing voltage and current waves. (Note that in these equations we needed to replace z by $-\zeta$.) We see that the phase difference between the two is $\pi/2$, which means that the voltage is zero everywhere when the current is maximum, and vice versa. Thus we have the same situation as in resonant circuits: when, for example, the current in the resonator is maximal, the entire energy is in the magnetic field. Such resonators can be made with any transmission line, e.g., twin lead, coaxial cable, or microstrip line. The next example discusses a coaxial cable resonator.

Example 23.11—Q factor of a coaxial resonator. The energy in a segment $\lambda/2$ of a coaxial transmission line is

$$W_{\text{em}} = \int_0^{\lambda/2} \frac{1}{2} L' |I(\zeta)\sqrt{2}|^2 \, \mathrm{d}\zeta = \frac{1}{2} L' \frac{8 I_+^2 \lambda}{4} = L' I_+^2 \lambda. \tag{23.36}$$

If the conductors have a small resistance R' per unit length, the time-average power losses in the resonator are

$$P_{\text{losses}} = \int_0^{\lambda/2} R' |I(\zeta)|^2 \, \mathrm{d}\zeta = R' \int_0^{\lambda/2} 4 I_+^2 \cos^2(\beta\zeta) \, \mathrm{d}\zeta = R' I_+^2 \lambda. \tag{23.37}$$

According to the definition of the Q factor, we obtain that for such a resonator

$$Q = \frac{\omega_r L'}{R'}, \tag{23.38}$$

which is of the same form as for a resonant circuit.

This simple theory is valid for *all* transmission lines. At higher frequencies, however, in open structures like two-wire lines, there may be significant losses due to radiation. Therefore resonators of this type are most often made of a coaxial-line segment, as in Fig. 23.9. The resonator may be excited (and the energy from it extracted) by either a small probe *a* or a small loop *b*. The probe and the loop are positioned at the voltage maximum (i.e., the electric-field maximum), namely the current maximum (i.e., the magnetic-field maximum).

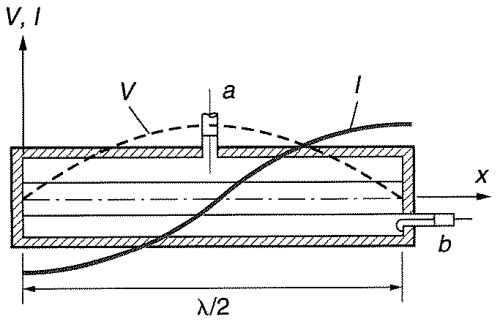

Figure 23.9 Sketch of a resonant ($\lambda/2$) section of a coaxial line shorted at both ends. The resonator may be excited by a small probe *a*, or a small loop *b*. Note the positions of the two excitation elements.

As an example, consider a coaxial resonator with the inner conductor of radius $a = 0.5\,\text{cm}$, the inner radius of the outer conductor $b = 1.5\,\text{cm}$, made of copper ($\sigma = 56 \cdot 10^6\,\text{S/m}$, $\mu = \mu_0$), at a frequency $f = 1\,\text{GHz}$. Using the data from Table 18.1 and Eq. (23.38), we obtain that $Q = 4614$. This is a very large value compared with those for resonant circuits (as mentioned, at the most about 100).

23.6.2 RESONANT CAVITIES

Resonant cavities may be of diverse shapes, but we will analyze the simplest, in the form of a parallelepiped (rectangular box). It can be obtained by introducing appropriate short circuits (transverse metallic walls) into a rectangular waveguide.

Consider a shorted rectangular waveguide, as in Fig. 23.10. Let a distant generator at left (not shown) excite in the waveguide the dominant, TE_{10} mode. The wave is reflected at the short circuit, giving rise to a backward wave. The backward wave is of the same form as the forward wave, except that the phase coefficient is now $-\beta$. At the short circuit the E_y component of the reflected wave must have the opposite phase with respect to the incident wave. From Eqs. (23.25) to (23.28) we thus obtain the following expressions for the resulting field in the shorted waveguide:

$$E_{y\,\text{res}}(x, y, z) = \text{j}\omega\mu\frac{a}{\pi}H_0 \sin\left(\frac{\pi}{a}x\right)\left(\text{e}^{\text{j}\beta z} - \text{e}^{-\text{j}\beta z}\right)$$

$$= -2\omega\mu\frac{a}{\pi}H_0 \sin\left(\frac{\pi}{a}x\right)\sin\beta z, \tag{23.39}$$

$$H_{x\,\text{res}}(x, y, z) = 2\text{j}\beta\frac{a}{\pi}H_0 \sin\left(\frac{\pi}{a}x\right)\cos\beta z, \tag{23.40}$$

$$H_{z\,\text{res}}(x, y, z) = -2\text{j}H_0 \cos\left(\frac{\pi}{a}x\right)\sin\beta z. \tag{23.41}$$

We see that the factor $\text{e}^{\pm\text{j}\beta z}$ is not present, so we have a standing wave in the waveguide. The other components are zero.

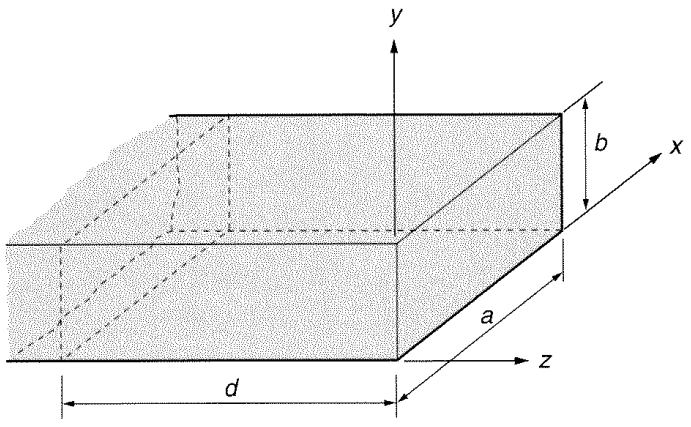

Figure 23.10 Shorted rectangular waveguide. Indicated in dashed lines is another "short circuit" of the waveguide, resulting in a resonant cavity in the form of a rectangular box.

According to Eqs. (23.39) and (23.41), the total y component of the electric field, and the total z component of the magnetic field, are zero not only at $z = 0$ (at the shorted end), but also in planes $z = -\pi p/\beta = -p\lambda_z/2$, $p = 1, 2, \ldots$. Consequently, if we place a thin metal foil in any of these planes and thus short the waveguide at one more place, the field will not change, since the boundary conditions are automatically satisfied in these planes. So we obtain a standing wave in a rectangular box of sides a, b, and $p\lambda_z/2$. This type of standing wave is known as the TE_{10p} mode in the cavity.

The TE_{101} Mode

Let us consider the simplest mode, the TE_{101} mode. Let $\lambda_z/2 = d$ in Fig. 23.10. We then have $\beta = 2\pi/\lambda_z = \pi/d$, so that Eqs. (23.39) to (23.41) become

$$E_{y\,\mathrm{res}}(x, y, z) = -2\omega\mu\frac{a}{\pi}H_0 \sin\left(\frac{\pi}{a}x\right)\sin\left(\frac{\pi}{d}z\right), \tag{23.42}$$

$$H_{x\,\mathrm{res}}(x, y, z) = 2\mathrm{j}\frac{a}{d}H_0 \sin\left(\frac{\pi}{a}x\right)\cos\left(\frac{\pi}{d}z\right), \tag{23.43}$$

$$H_{z\,\mathrm{res}}(x, y, z) = -2\mathrm{j}H_0 \cos\left(\frac{\pi}{a}x\right)\sin\left(\frac{\pi}{d}z\right). \tag{23.44}$$

From these equations we can deduce how electromagnetic oscillations in the cavity are maintained. There are induced electric charges on the upper and lower cavity walls because the normal component $E_{y\,\mathrm{res}}$ of the electric field vector exists there. Surface currents in the y direction appear only on the side walls, where $H_{x\,\mathrm{res}}$ and $H_{z\,\mathrm{res}}$ are nonzero. So the oscillations of the electromagnetic field inside the cavity are accompanied by charges and currents on its walls.

According to Eqs. (23.42) to (23.44), the electric and magnetic fields are shifted in phase by $\pi/2$ (the factor j). Therefore, at some points in time there are no currents flowing in the walls, and at others there are no charges. These instants are separated in time by $T/4$, where $T = 1/f$ is the period of the oscillations. Figure 23.11 shows the distribution of charges and currents during one period, starting at the instant when the upper face carries the maximum positive charge.

To determine the quality factor of the cavity, we need to calculate the energy contained in the cavity and the time-average power of losses in the cavity walls corresponding to this energy. At this introductory level, we will just give a numerical example: for a cubic cavity ($a = b = d$) filled with air, designed to operate at 3 GHz in the TE_{101} mode, we find that the side length of the cube is $a = c/(f\sqrt{2}) = 7.07\,\mathrm{cm}$. (Of course, in that case the TE_{101} mode will be the same as the TE_{110} or TE_{011} modes.) Let the cavity be made of copper ($\sigma = 56 \cdot 10^6\,\mathrm{S/m}$, $\mu = \mu_0$). Using the losses in the surface resistance of the cavity walls, one can calculate that $Q \simeq 19{,}200$. To obtain this value, the surface resistance is calculated assuming a perfectly polished wall, i.e., with unevenness much less than the skin depth. So as frequency increases, it becomes more and more difficult to avoid increases in loss.

Questions and problems: Q23.25 to Q23.32, and P23.14 to P23.16

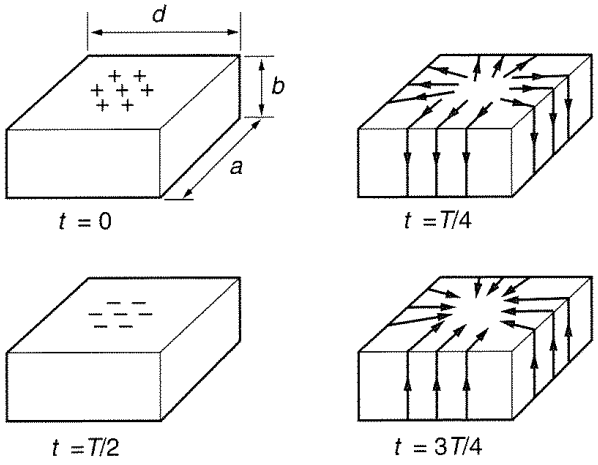

Figure 23.11 Sketch of charge distribution (plus and minus signs) and surface currents over the cavity walls of a rectangular cavity at four different moments in time

23.7 Chapter Summary

1. Electromagnetic waves can be guided along a desired route not only by transmission lines but also by hollow pipes, dielectric-coated surfaces, or dielectric rods.
2. It is typical of all these hollow waveguides that (1) they cannot support the TEM wave type, but support the TE and TM wave types, and (2) they cannot transmit energy below a certain frequency, known as the cutoff frequency.
3. Waves of frequencies above the cutoff frequency propagate without attenuation (except due to losses in the materials).
4. Waves of frequencies lower than the cutoff frequency, known as evanescent modes, are exponentially attenuated, and do not propagate at all.
5. Among hollow metallic waveguides, the most important are those of rectangular cross section. There are an infinite number of both TE_{mn} and TM_{mn} modes that can propagate along such waveguides.
6. The higher-order modes (with larger m and n values) for the same waveguide must be of increasingly higher frequencies. So there is a frequency range in which only one mode, the TE_{10} mode, can propagate. This mode is therefore termed the *dominant mode.*
7. Commonly used printed microstrip lines support a hybrid mode, consisting of both TE and TM wave types. However, because the longitudinal components of the electric and magnetic field vectors are small, we can approximately treat the wave as a "quasi-TEM" wave, similar to that in a transmission line.
8. Electromagnetic resonators support oscillating electromagnetic fields. At high frequencies, two types of such resonators are mostly used, the coaxial-line (or

other similar line) resonator and the cavity-type resonator. The latter is obtained as a special case of a waveguide, short-circuited at two ends.

9. The quality factor of waveguide resonators may be by two orders of magnitude greater than that of lumped-element resonant circuits.

QUESTIONS

Q23.1. Write the instantaneous value of $\mathbf{E}_{\text{tot}}(x, y, z) = \mathbf{E}(x, y)e^{-\gamma z}$, where $\gamma = \alpha + j\beta$.

Q23.2. Complete the derivation of Eq. (23.7).

Q23.3. Define in your own words the TEM, TE and TM waves. What does "mode" mean?

Q23.4. Can the complex propagation coefficient γ in Eq. (23.5) be real? Can it have a real part?

Q23.5. What are eigenvalues (characteristic values) of a parameter in a boundary-value problem? What do they depend on?

Q23.6. The wave impedance of a TEM wave is always real. Are the wave impedances of TE and TM waves also always real? Explain.

Q23.7. Under which conditions is the relation (23.13) valid?

Q23.8. What is the physical meaning of the coefficients m and n in the field components inside a rectangular waveguide in Eqs. (23.16) to (23.20)?

Q23.9. What is the phase and group velocity in a rectangular waveguide in these three cases? (1) $f < f_c$, (2) $f = f_c$, and (3) $f > f_c$

Q23.10. What is the attenuation constant in a rectangular waveguide in these three cases? (1) $f < f_c$, (2) $f = f_c$, and (3) $f > f_c$

Q23.11. What are the parameters that determine the cutoff frequency in a waveguide?

Q23.12. A signal consisting of frequencies in the vicinity of a frequency f_1, and a signal consisting of frequencies in the vicinity of a frequency f_2, propagate unattenuated along a rectangular waveguide in the TE_{10} mode. If $f_1 < f_2$, which is faster?

Q23.13. What will eventually happen with the signals from the preceding question if the waveguide is long?

Q23.14. A signal consisting of frequencies in the vicinity of a frequency f_1 propagates along a rectangular waveguide as a TE_{10} mode. What happens if the bandwidth of the signal is relatively large?

Q23.15. What are *propagating modes* and *evanescent modes* in a waveguide?

Q23.16. You would like to have openings for airing a shielded room (a Faraday's cage) without enabling electromagnetic energy to enter or leave the cavity. You are aware that a field of a certain microwave frequency is particularly pronounced around the room, but you do not know its polarization. Can you make the openings in the form of waveguide sections? What profile of the waveguide would you use?

Q23.17. You are using a square waveguide that is bent and twisted along its way. The waveguide is excited with the TE_{10} mode (the E field parallel to the y axis). Can you be certain about the polarization of the wave at the receiving point? Explain.

Q23.18. What is the physical meaning of the *dominant mode* in a waveguide?

Q23.19. A rectangular waveguide along which waves of many frequencies and modes propagate is terminated in a large metal box. Can you extract from the box a signal of a specific frequency and a desired mode by connecting a section of the same waveguide at another point of the box?

Q23.20. How would you construct a high-pass filter (i.e., a filter transmitting only frequencies above a certain frequency), using sections of rectangular waveguides?

Q23.21. Propose a method for exciting the TE_{11} mode in a rectangular waveguide.

Q23.22. You would like for a rectangular waveguide with a TE_{10} wave to *radiate* (leak) from a series of narrow slots you made in its walls. For this, you need slots that would force the internal waveguide currents to appear on its outer surface. How do the slots need to be oriented to accomplish this?

Q23.23. Sketch the electric and magnetic field lines for two microstrip lines, one with a substrate twice the thickness of the other, but with the same permittivity. In which case is the quasi-TEM approximation more accurate? Explain.

Q23.24. Sketch the electric and magnetic field lines for two microstrip lines on substrates of equal thicknesses, but where one has a permittivity two times higher than the other. In which case is the quasi-TEM approximation more accurate? Explain.

Q23.25. In the resonant circuit of Example 23.10, explain why it is hard to achieve a large Q factor. Why do losses go up as the frequency increases?

Q23.26. You would like to have a coaxial-line resonator with as large a Q factor as possible for a given outer resonator size. What would you do?

Q23.27. Propose two methods for the excitation and energy extraction from a coaxial resonator.

Q23.28. Find the energy contained in coaxial resonators of lengths λ, $\frac{3}{2}\lambda$, and 2λ, using Eq. (23.36). What is the Q factor of these resonators?

Q23.29. Sketch the current and voltage along an open-ended microstrip line resonator that is half of a guided wavelength long. What is the impedance at the center of the resonator, and what at the two ends?

Q23.30. What loss mechanisms can you think of in an open-ended microstrip line resonator?

Q23.31. A rectangular waveguide with a TE_{10} mode is terminated in a large rectangular cavity (e.g., of a microwave oven). Describe qualitatively what happens.

Q23.32. Propose two methods for the excitation and energy extraction from a cavity resonator with a TE_{101} wave type in it.

PROBLEMS

P23.1. Prove that for any TEM wave the electric and magnetic field vectors are normal to each other at all points.

P23.2. Prove that at any cross section of a two-conductor transmission line with a forward traveling wave, the ratio of the voltage between the conductors and the current in them equals Z_{TEM} in Eq. (23.7). Show that for two-conductor transmission lines, $C'L' = \epsilon\mu$.

P23.3. Prove that since at any cross section of a multiconductor transmission line $\sum Q'(z) = 0$, it follows that $\sum I(z) = 0$, where the sum refers to all the conductors of the line.

P23.4. Prove that Eqs. (23.1) to (23.4) imply that the electric and magnetic field vectors of a TE wave are normal to each other at all points.

P23.5. Write the instantaneous values of all the components of the TE_{10} wave in a rectangular waveguide. From these equations, sketch the distribution of the **E**-field and the **H**-field in the waveguide at $t = 0$.

P23.6. Determine the cutoff frequencies of an air-filled waveguide with $a = 2.5\,\text{cm}$ and $b = 1.25\,\text{cm}$, for the following wave types: (1) TE_{01}, (2) TE_{10}, (3) TE_{11}, (4) TE_{21}, (5) TE_{12}, and (6) TE_{22}.

P23.7. Plot the mode impedances between 8 and 12 GHz for an air-filled rectangular waveguide with $a = 2.5\,\text{cm}$ and $b = 1.25\,\text{cm}$, for the following wave types: (1) TE_{01}, (2) TE_{10}, (3) TE_{11}, (4) TE_{21}, (5) TE_{12}, and (6) TE_{22}.

P23.8. Plot the wavelength λ_z along a rectangular waveguide with $a = 2\,\text{cm}$, $b = 1\,\text{cm}$, and air as the dielectric, if the wave is of the TE_{10} type, for frequencies between 8 and 10 GHz. Is the wavelength shorter or longer than in an air-filled coaxial line?

P23.9. Plot the phase and group velocities in problem P23.8.

P23.10. In a rectangular waveguide from problem P23.8, two signals are launched at the same instant. The frequency range of the first is in the vicinity of $f_1 = 10\,\text{GHz}$, and of the second in the vicinity of $f_2 = 12\,\text{GHz}$. Both signals propagate as TE_{10} waves. Find the time intervals the two signals need to cover a distance $L = 10\,\text{m}$, and the difference between the two intervals. Which signal is faster?

P23.11. Derive Eq. (23.33) starting from the tangential electric field boundary condition.

P23.12. The *effective dielectric constant* of a microstrip line depends on its dimensions approximately as

$$\epsilon_e = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2} \frac{1}{\sqrt{1 + 12h/w}},$$

where the parameters are explained in Fig. P23.12. Plot the effective dielectric constant for h/w ratios between 0.1 and 10 (this is the approximate range for practical use), and for substrates that have relative permittivities of 2.2 (Teflon-based Duroid), 4.6 (FR4 laminate), 9 (aluminum nitride), 12 (high-resistivity silicon), and 13 (gallium arsenide).

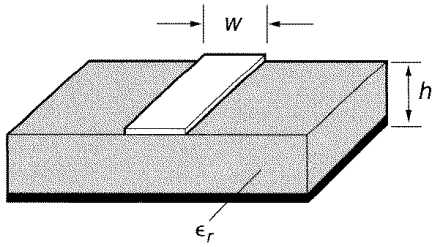

Figure P23.12 A microstrip line

P23.13. The approximate formulas for microstrip line impedance and propagation constant based on the quasi-TEM approximation are given by

$$\beta = \omega\sqrt{\epsilon_0\mu_0}\sqrt{\epsilon_e},$$

$$Z_0 = \begin{cases} \dfrac{60}{\sqrt{\epsilon_e}} \ln\left(\dfrac{8h}{w} + \dfrac{w}{4h}\right), & \dfrac{w}{h} \leq 1 \\ \dfrac{120\pi}{\sqrt{\epsilon_e}\left\{(w/h) + 1.393 + 0.667\ln[(w/h) + 1.444]\right\}}, & \dfrac{w}{h} > 1 \end{cases}$$

Plot the characteristic impedance as a function of the ratio w/h (between 0.1 and 10), and for the relative permittivities from problem P23.12. What can you conclude about the impedance as the line gets narrower?

P23.14. Plot the current, voltage, and impedance along a half-wavelength coaxial resonator short-circuited at both ends. If you want to feed the resonator with another piece of the same kind of cable, at which place along the resonator would you do it and why?

P23.15. Plot the current, voltage, and impedance along a half-wavelength coaxial line resonator open-circuited at both ends. You want to feed the resonator with a 50-Ω coaxial line. Propose (sketch) a way to do it, and explain.

P23.16. Determine the maximum possible energy stored in a cubical resonant air-filled cavity with $a = b = d = 10\,\text{cm}$, at a resonant frequency corresponding to the TE_{101} wave. The electric strength of air is $30\,\text{kV/cm}$.

24

Fundamentals of Electromagnetic Wave Radiation and Antennas

24.1 Introduction

We know that plane electromagnetic waves propagate through space and carry energy, but we do not still know how such waves can be produced. For the creation of electromagnetic waves we need specific structures with time-varying charges and currents. The process of producing electromagnetic waves, which then propagate with no connection to the sources, is known as *electromagnetic radiation.*

Theoretically, any system containing time-varying charges and currents radiates a certain amount of energy. In many cases in actual practice the radiation can either be ignored or is intentionally suppressed. For example, radiation from 60-Hz (or 50-Hz) power transmission lines exists in theory, but the power radiated is so small that it can practically not be detected. At high frequencies, coaxial cables or hollow waveguides are used to transmit energy precisely because they do not radiate at any frequency.

Specific structures aimed at efficient radiation of electromagnetic waves are referred to as *transmitting antennas.* Typically, transmitting antennas do not radiate

equally in all directions, i.e., they have certain *directional radiation properties.* An entire complicated science of designing and analyzing antennas has been developed in the last hundred years. Although a great number of antennas are available today, the analysis and clever design of antennas for ever-increasing numbers of new applications is an engineering challenge. The present-day powerful numerical methods enable efficient design of many classes of antennas, but a profound knowledge of electromagnetic field theory is needed to make optimal use of such methods.

The electromagnetic energy radiated by an antenna invariably carries a certain signal (information) to be transmitted to one or many receivers. How is energy, and the signal, extracted from an electromagnetic wave? Basically, structures the same as transmitting antennas are used for that purpose, in which case they are called *receiving antennas.* We will see that the most important properties of a receiving antenna can be evaluated if they are known for the same antenna when it is transmitting. In fact, frequently one antenna is used for both transmitting and receiving (e.g., antennas in mobile phones).

Transmitting and receiving antennas are the vocal cords and ears of all radio wave communication systems. Although quite different in shape and size, rabbit-ear antennas and rod antennas for our portable radios, various antennas for TV reception, antennas for the reception of satellite TV programs, antennas used for communication with satellites surveying distant planets, and antennas for astrophysical research that can be the size of an entire valley all operate on the same principles. This chapter is devoted to explaining basic principles and concepts related to radiation and propagation of electromagnetic waves.

24.2 Transmitting and Receiving Antennas

Antennas can be used to transmit or receive signals in the form of electromagnetic waves. In either case, the antenna is connected through its feed to some circuit. For example, in reception the antenna is usually followed by a low-noise amplifier because the signals are often very small and the amplifier serves to overcome the noise that the signal is buried in. In transmission, a power amplifier is often connected before the antenna feed. An engineer needs to know how the antenna behaves as part of the rest of the circuit. In the following sections, we discuss transmitting and receiving antennas as circuit elements.

24.2.1 NOTES ON TRANSMITTING ANTENNAS

A transmitting antenna takes energy from a source (e.g., an oscillator) and radiates a part of this energy in the form of a free electromagnetic wave. Frequently, the source is connected to the antenna via a transmission line, called the *antenna feed.* Looking from the source (or the feed end terminals), a transmitting antenna is just a receiver of energy. Most frequently, the source feeds the antenna with a time-harmonic voltage, so that in the complex (phasor) domain, the source sees the transmitting antenna as a complex impedance Z_A. This impedance is known as the (transmitting) *antenna impedance* and it is, in general, a function of frequency.

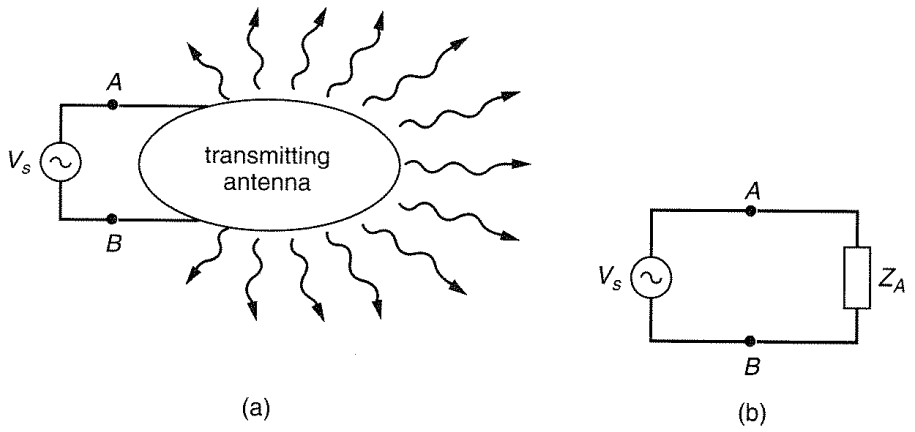

Figure 24.1 (a) A symbolic picture of a transmitting antenna, and (b) its equivalent circuit as seen by the generator. Wavy arrows of unequal lengths symbolize that the radiation differs in different directions in space.

Neglecting possible losses in the antenna, the time-average power delivered to the antenna is the radiated power. Therefore, the time-average radiated power is given by $P_{\text{rad}} = R_{\text{rad}} I_0^2$, where R_{rad} is the real part of the antenna impedance, and I_0 is the rms value of the current at the antenna terminals.

Most frequently, the transmitting antenna is designed to radiate electromagnetic waves in specific directions, depending on the application. For example, a broadcasting antenna should radiate in all horizontal directions, whereas a satellite antenna should radiate in a very narrow beam toward the corresponding satellite. A symbolic picture of a transmitting antenna and its equivalent circuit are shown in Fig. 24.1.

24.2.2 NOTES ON RECEIVING ANTENNAS

A receiving antenna transforms a part of the energy carried by an electromagnetic wave into voltage between two antenna terminals, which are connected to a receiver. This voltage is next amplified and processed as needed. So, from a viewpoint at the receiver terminals, a receiving antenna acts as a voltage generator. In the frequency domain and in complex notation, it behaves as a generator of an emf and an internal impedance, so it can be described by a Thévenin equivalent.

We know that the emf of a Thévenin generator is found as the open-circuit voltage across its terminals (in this case, across the antenna terminals). The internal impedance of the Thévenin generator is that seen by a source connected to the antenna terminals, if the emf (i.e., the incident wave) is not present. This is precisely the *transmitting antenna impedance* mentioned earlier. We reach a very important conclusion:

The internal impedance of the Thévenin generator of a receiving antenna is the same as the impedance of the antenna when transmitting. (24.1)

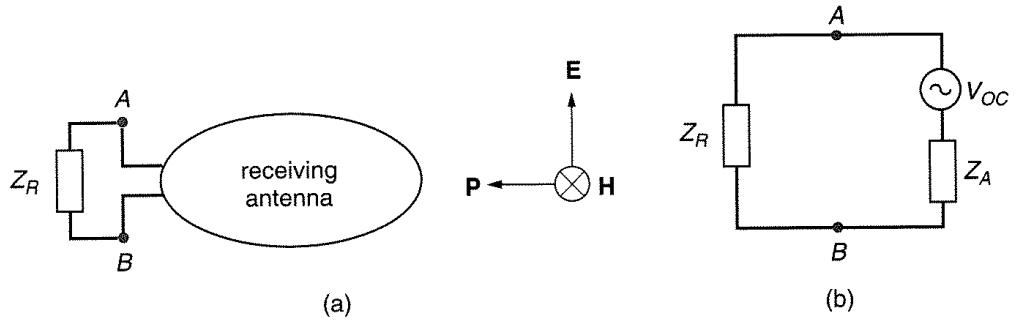

Figure 24.2 (a) A symbolic picture of a receiving antenna, and (b) its equivalent circuit with respect to the receiver. The emf of the equivalent generator depends on the incident wave direction and polarization.

A symbolic picture of a receiving antenna and its Thévenin equivalent with respect to the receiver are sketched in Fig. 24.2.

The Thévenin emf of a receiving antenna depends on the antenna shape and the direction of the incident electromagnetic wave exciting it. For example, assume the receiving antenna to be in the form of two straight wire segments connected to the two receiver terminals (this is known as a *wire dipole antenna*). If the electric field of the wave is parallel to the wire, it is natural to expect the largest emf, and if it is perpendicular, we would expect no emf at all. Thus, although the impedance of the Thévenin generator equivalent to a receiving antenna depends only on the antenna itself, its emf depends greatly on the direction and polarization of the incident electromagnetic wave. So a receiving antenna *has directional properties*, as does a transmitting antenna. We will show that the directional properties of any receiving antenna are known if they are known for the same antenna in transmitting mode.

24.2.3 PRINCIPAL ISSUES IN ANTENNA ANALYSIS AND DESIGN

From the preceding simple reasoning, we see that antenna engineers need to know the circuit and radiation properties of antennas in the *transmitting* mode only. In addition to choosing an antenna that radiates properly, they will usually need to match the antenna to the transmitter or receiver, just as we match any passive or active element in circuit theory. Finally, to design a radio communication link, engineers need to know basic properties of electromagnetic-wave propagation in realistic circumstances.

24.2.4 ANTENNAS ABOVE CONDUCTING SURFACES

Many antennas are close to approximately flat conducting surfaces, such as antennas above ground, or small antennas on aircraft fuselage or wings, or on cars. If the conducting surface is approximated by a perfectly conducting plane, the antenna can be analyzed by image theory. Figure 24.3 shows a few typical cases. The images at corresponding points should have opposite charges, and currents in opposite directions,

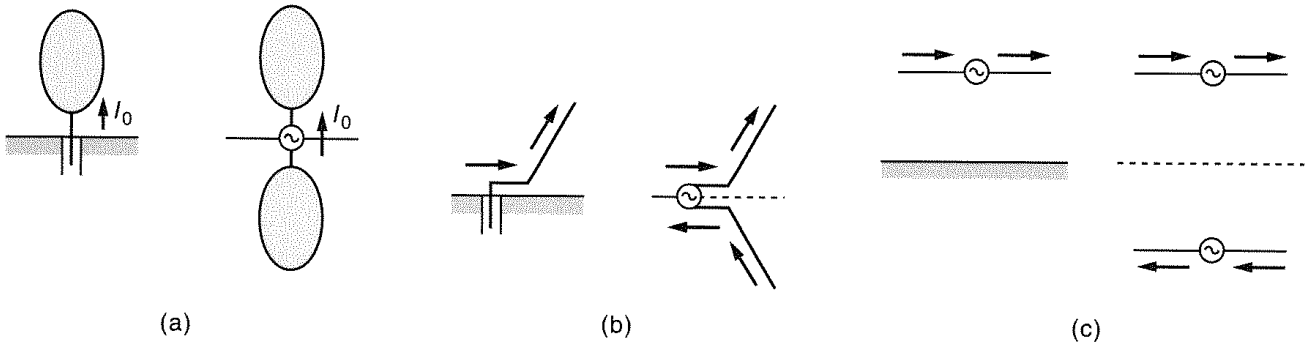

Figure 24.3 Three examples of antennas close to a perfectly conducting plane and their equivalents with respect to the upper half-space, showing images with the necessary charges and currents: (a, b) monopole antennas fed at the ground plane; (c) a dipole antenna above a ground plane

with respect to the conducting plane. Such charges and currents guarantee a zero tangential component of the electric field vector on the perfectly conducting plane, i.e., the original boundary conditions remain satisfied.

Antennas of the type (c) in Fig. 24.3 are known as *dipole antennas* (antennas with two poles, i.e., two arms). Antennas of the types (a) and (b), which are excited at the surface as indicated, have only one arm, and are called *monopole antennas*. Note that in reality, the radiation field exists *only in the upper half-space*. This means that for cases (a) and (b) in Fig. 24.3, the voltages driving the antennas with their images have to be twice those of the original monopole antennas. Since the current is the same, this means that the impedance of a monopole antenna above a perfect ground equals half the impedance of the corresponding dipole antenna. Note that this conclusion does not apply to the dipole antenna (c), although its impedance will also be different from that of the isolated antenna, due to the presence of the conducting plane.

Questions and problems: Q24.1 to Q24.3, P24.1

24.3 Electric Dipole Antenna (Hertzian Dipole)

The electric dipole antenna, or the Hertzian dipole, is the simplest of all radiating systems. It consists of a straight, thin wire conductor of length l with two small conducting spheres or disks at the ends, as sketched in Fig. 24.4. The spheres serve as capacitor electrodes, and make the current $i(t)$ along the dipole wire constant along its length. A sinusoidal generator of angular frequency ω is connected somewhere along the wire. The derivation of the electric and magnetic field vector components is given in most higher-level electromagnetics textbooks (see, e.g., S. Ramo et al., *Fields and waves in communication electronics*, 3d ed., section 12.3, John Wiley & Sons, 1993). For our introduction to Hertzian dipoles, it is sufficient to quote the expressions so that we can make some important conclusions about the radiated fields. Note that the dipole radiates in a sphere around it, and therefore it is natural to use the spherical coordinate system to describe the radiated fields. At distances far from the dipole in terms of the free-space wavelength (more than about 10 wavelengths), and assuming

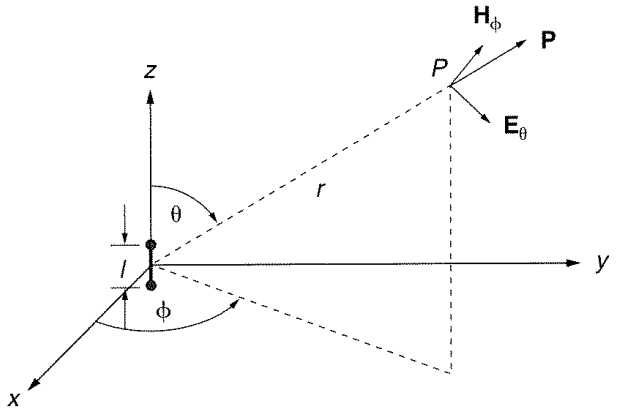

Figure 24.4 The electric dipole antenna (the Hertzian dipole)

that the dipole is situated in a homogeneous dielectric of parameters ϵ and μ, the complex expressions for the electric and magnetic fields are given by

$$E_\theta(r,\theta) = \frac{\mathrm{j}\beta Il\sin\theta}{4\pi r}\sqrt{\frac{\mu}{\epsilon}}\mathrm{e}^{-\mathrm{j}\beta r}, \tag{24.2}$$

$$H_\phi(r,\theta) = \frac{\mathrm{j}\beta Il\sin\theta}{4\pi r}\mathrm{e}^{-\mathrm{j}\beta r}. \tag{24.3}$$

(Electric and magnetic fields far from a Hertzian dipole)

Thus, the only component of the electric field is in the direction of the unit vector $\mathbf{u}_\theta$, and that of the magnetic field in the direction of the unit vector $\mathbf{u}_\phi$. (In Fig. 24.4, at point P the latter is directed into the paper.) All the other components are zero. Hence the ratio of the amplitudes of the two vectors is the same as for a plane wave,

$$\frac{E_\theta(r,\theta)}{H_\phi(r,\theta)} = \eta = \sqrt{\frac{\mu}{\epsilon}}. \tag{24.4}$$

(Relation between E and H fields of the wave radiated by a Hertzian dipole)

This was to be expected because at large distances from the dipole the spherical wave is locally a plane wave. In addition, we note that the vectors E_θ and H_ϕ are normal to each other, and that the Poynting vector is directed away from the dipole, as expected, because the radiated wave carries power away from the antenna.

Note that both field components depend on the distance r from the dipole as $1/r$. No static field has this dependence on r. This type of field is thus different from any electromagnetic field we considered so far, and is termed the *radiation field*, or the *far field*, of the Hertzian dipole. We are interested here only in this field (the field

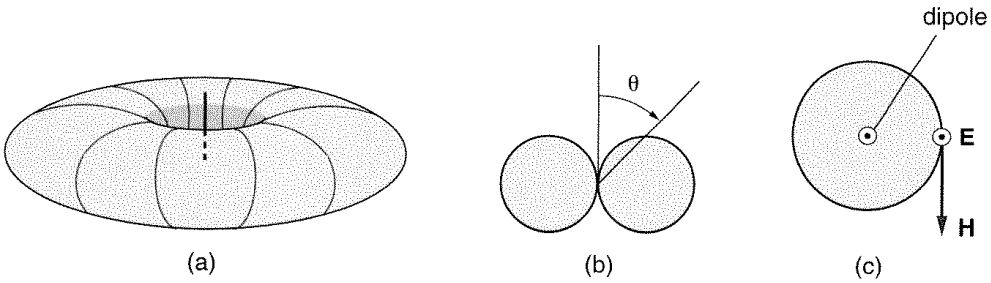

Figure 24.5 Radiation patterns of a Hertzian dipole versus angle θ, normalized to the maximal radiation. (a) The three-dimensional pattern, (b) the E-plane pattern, and (c) the H-plane pattern are plotted.

closer to the dipole has other components in addition). Because all antennas can be considered as large assemblies of Hertzian dipoles, *the far fields of all antennas also depend on r as* $1/r$.

Example 24.1—Spatial distribution of radiation from a Hertzian dipole. Note that the simplest radiating element, the Hertzian dipole, does not radiate equally in all directions. This is evident from the factor $\sin\theta$, which tells us that the radiation is the strongest in the plane of symmetry of the dipole (for $\theta = \pi/2$), and zero along the direction defined by the dipole wire ($\theta = 0$ and $\theta = \pi$). Consequently, no real antenna can radiate equally in all directions. To characterize the distribution of radiated field in different directions, it is customary to normalize the field with respect to its maximal value, and to plot a graph of this function. Such a graph is known as the *antenna radiation pattern*. Antenna radiation patterns may be plotted in three dimensions, as in Fig. 24.5a, but it is usually more convenient to plot cuts of the three-dimensional pattern. Usual cuts are those containing the **E** vector (known as the *E-plane pattern*, Fig. 24.5b) or the **H** vector (the *H-plane pattern*, Fig. 24.5c). We will see that with appropriate shapes of antennas, we can obtain a great variety of radiation patterns.

Any radiating system can always be subdivided into a large number of Hertzian dipoles, provided that the current distribution of the system is known. It is important to understand that if this simplest radiating system produces a radiation field with the electric and magnetic field as in a plane wave propagating along the (local) r axis, by superposition (and because of linearity) the same will be true for *any* radiating structure.

Example 24.2—The half-wave dipole antenna. A frequently used antenna is in the form of a straight wire dipole of total length equal to about half a wavelength. For such a dipole, it turns out that the current distribution can be roughly approximated by a sine function (Fig. 24.6a). At an introductory level, it is important to remember that the impedance of a half-wave dipole is, very roughly, 73 Ω. The radiation pattern of a half-wave dipole is very similar to that of the Hertzian dipole. The only difference is that the E-plane pattern is not in the form of a "number eight" consisting of two circles, but instead of an "eight" consisting approximately of two ellipses, as in Fig. 24.6b.

Questions and problems: Q24.4 to Q24.8, P24.2 to P24.5

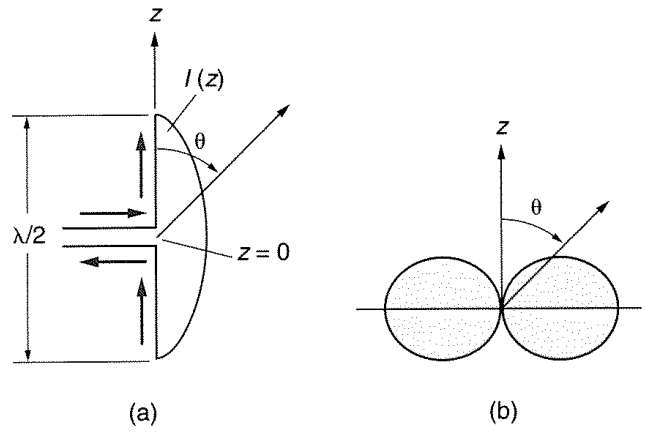

Figure 24.6 (a) The half-wave dipole antenna; (b) E-plane radiation pattern of the dipole

24.4 Antenna Directivity

The radiation pattern is only one of several parameters that describe the spacial distribution of antenna radiation. We now describe another quantity, which is used more frequently than radiation patterns.

Assume that an antenna radiates equally in all directions. Such a hypothetical antenna is known as an *isotropic*, or *omnidirectional, antenna*. For an isotropic antenna, the power density is the same for all points of a sphere of radius r centered at the antenna. This power density is denoted by S_i, and is equal to

$$S_i = \frac{P_{\text{rad}}}{4\pi r^2},$$

in watts per m^2. We can now define the antenna *directivity* as the ratio of the power density radiated in a given direction (θ, ϕ) to the isotropic power density:

$$D(\theta, \phi) = \frac{S(\theta, \phi)}{S_i} = \frac{4\pi r^2 S(\theta, \phi)}{P_{\text{rad}}}, \qquad P_{\text{rad}} = R_{\text{rad}} |I_0|^2. \tag{24.5a}$$

(Definition of directivity of an antenna)

From this expression, we see that the directivity of an isotropic antenna is unity.

We can relate the directivity to the electric and magnetic fields through the Poynting vector. From the Poynting theorem, Eq. (19.42), the surface integral of the Poynting vector is the radiated power. The power density is therefore equal to the magnitude of the Poynting vector, which we remember is $\mathbf{E} \times \mathbf{H}$. We know that the electric and magnetic far fields at a point are proportional to $1/r$, where r is the distance of the observation point from the antenna. Consequently, the time-average Poynting vector, $\mathcal{P}(r, \theta, \phi)$, at a point in the far field is proportional to $1/r^2$, and it is also proportional to the power radiated by the antenna, P_{rad}. So we get an alternate

expression for the antenna directivity *relative to an isotropic antenna*:

$$D(\theta,\phi)=\frac{4\pi r^2|\mathcal{P}(r,\theta,\phi)|}{P_{\text{rad}}}=\sqrt{\frac{\epsilon}{\mu}}\frac{4\pi r^2|\mathbf{E}(r,\theta,\phi)|^2}{P_{\text{rad}}},\qquad P_{\text{rad}}=R_{\text{rad}}|I_0|^2. \quad (24.5b)$$

(Definition of directivity of an antenna)

We see that the directivity *depends only on spherical angles* θ *and* ϕ and can be used as a measure of the antenna directional properties. Often, the directivity is given in decibels,

$$[D(\theta,\phi)]_{\text{dB}}=10\log\{D(\theta,\phi)\}\quad(\text{dB}). \quad (24.6)$$

(Directivity in decibels)

The multiplier of the logarithm is 10, not 20, because the directivity is defined through power, not field intensity.

The plot of the directivity in space is known as the *antenna power pattern*. As in the case of field patterns, more frequently the antenna E-plane or H-plane power pattern or both are plotted, although the patterns in other planes may also be of interest. From the definition in Eqs. (24.5), we see that the power pattern is proportional to the *square* of the field pattern.

If, in referring to the directivity, the direction (defined by angles θ and ϕ or in some other way) is not specified, by convention this means that *the maximum value of the directivity is implied*, i.e.,

$$D=[D(\theta,\phi)]_{\max}. \quad (24.7)$$

(Definition of directivity with no reference to direction)

In this text, we will use the term "maximal directivity" for clarity. It is of significant practical interest because in many applications antennas are used to radiate in a specific direction (for example, in satellite links).

Manufacturers often specify *antenna gain* and maximum antenna gain, $G(\theta,\phi)$ and $G=[G(\theta,\phi)]_{\max}$, instead of directivity. The difference between these two parameters is that the antenna gain includes any power losses in the antenna (such as losses due to the finite conductivity of the metal that the antenna is made of). Therefore, the gain of an antenna is a number that is at most as large as the directivity of the same antenna.

Example 24.3—Directivity of the Hertzian dipole. From the far electric field of the Hertzian dipole, the power radiated by the dipole, and hence its directivity, can be calculated using Poynting's theorem. The derivation is given in most higher-level electromagnetic textbooks, and here we give only the final expression for the directivity of the Hertzian dipole:

$$[D(\theta,\phi)]_{\text{Hertzian dipole}}=1.5\ \sin^2\theta. \quad (24.8)$$

(Directivity of Hertzian dipole)

This is proportional to the square of the electric field pattern, which is proportional to $\sin\theta$. The maximal directivity of the Hertzian dipole is thus

$$D_{\text{Hertzian dipole}} = 1.5. \tag{24.9}$$

(Maximal directivity of Hertzian dipole)

Note that this means that the power density (magnitude of the Poynting vector) of the field radiated by the Hertzian dipole in the plane $\theta = \pi/2$ is simply 1.5 times that of an isotropic antenna radiating the same power and located at the position of the dipole. The maximal radiated electric field strength of the Hertzian dipole is thus $\sqrt{1.5} \simeq 1.22$ times that of an isotropic antenna.

Example 24.4—Directivity of the half-wave dipole. To determine the directivity of the half-wave dipole, we need to know the radiation field of the dipole. Because we know the (approximate) current distribution along it, this is not too complicated, but at this introductory level we will skip the derivation. The final result for the directivity is

$$D(\theta) = \sqrt{\frac{\mu}{\epsilon}} \frac{1}{\pi R_{\text{rad}}} \frac{\cos^2(\pi/2\cos\theta)}{\sin^2\theta}, \tag{24.10}$$

where θ is the angle between the dipole axis (z axis in Fig. 24.6a), and the direction toward the point in the radiation field. R_{rad} is the antenna resistance (we know that for a half-wave dipole, $R_{\text{rad}} \simeq 73\,\Omega$). Therefore the maximal directivity of the half-wave dipole is

$$D = D(\pi/2) = \frac{120}{73} \simeq 1.64. \tag{24.11}$$

(Maximal directivity of half-wave dipole)

Questions and problems: Q24.9 and Q24.10

24.5 The Receiving Antenna

A receiving antenna is always very far from the transmitting antenna (practical reasons for this fact are given in Chapter 25). Therefore, the current induced in it is very small compared to that in the transmitting antenna. The field produced at the transmitting antenna by the current induced in the receiving antenna is evidently negligible. Therefore, the equivalent circuit for the transmitting-receiving antenna system is as in Fig. 24.7.

The transfer impedance (or mutual impedance), Z_{12}, has exactly the same meaning as in circuits coupled by the induced electric field ("magnetic coupling"). In this case also the induced electric field is the one that induces the emf and current in the receiving antenna, the only difference being that the finite velocity of propagation of electromagnetic waves must now be taken into account. Note that in circuits the effect of finite wave velocity can normally be neglected because all the parts of a circuit are close when compared with the wavelength.

Although this is not necessary for the following derivations, assume that both antennas are matched to the feed lines, as indicated in Fig. 24.7, which is most often

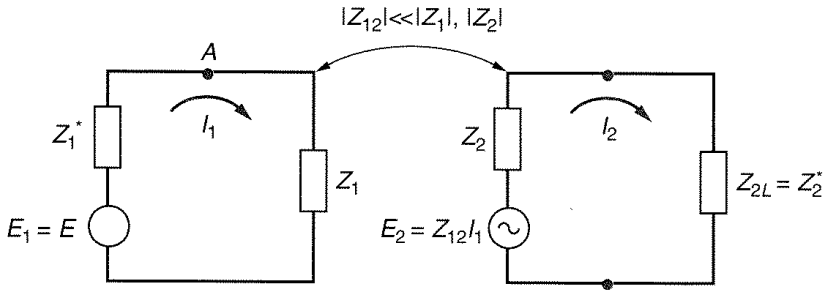

Figure 24.7 Equivalent circuit for the transmitting-receiving antenna system, when antenna 1 is transmitting

the case. If antenna 1 is transmitting and is excited by a generator of voltage V, the current in the receiving antenna, antenna 2, is given by (see Fig. 24.7)

$$I_{2\text{rec}} = \frac{Z_{12}(\theta_1, \phi_1, \theta_2, \phi_2) I_1}{2R_2} = Z_{12}(\theta_1, \phi_1, \theta_2, \phi_2) \frac{V}{4R_2 R_1}. \tag{24.12a}$$

The angles θ_1 and ϕ_1 are local spherical angles of antenna 1, defining the direction from antenna 1 toward antenna 2. Similarly, θ_2 and ϕ_2 are local spherical angles of antenna 2, defining the direction from antenna 2 toward antenna 1.

If the generator is moved to antenna 2, and antenna 1 is receiving, the current in antenna 1 is similarly

$$I_{1\text{rec}} = Z_{21}(\theta_1, \phi_1, \theta_2, \phi_2) \frac{V}{4R_1 R_2}. \tag{24.12b}$$

Note that we first determined the current in branch 2 of the (coupled) circuits due to the generator in branch 1. We next moved the generator to branch 2, and determined the current in branch 1. This is a typical example of circuit reciprocity, which dictates that the currents in Eqs. (24.12*a*) and (24.12*b*) must be the same. Consequently, $Z_{21}(\theta_1, \phi_1, \theta_2, \phi_2) = Z_{12}(\theta_1, \phi_1, \theta_2, \phi_2)$.

The transfer impedances implicitly contain the direction of radiation of the transmitting antenna, as well as the direction from which the incident wave is arriving to the receiving antenna. Therefore the transfer impedance contains the directional properties of the two antennas in both the transmitting and the receiving mode. For $Z_{21} = Z_{12}$ to hold in all cases, the transmitting and receiving patterns of an antenna must be the same. We reach an extremely important conclusion:

The receiving pattern of a receiving antenna is the same as the radiation pattern of the same antenna in transmitting mode. (24.13)

The antenna's *effective area* is a quantity used frequently for describing a receiving antenna. Assume a receiving antenna 2 *matched to its load.* Assume that the incident wave arrives from the direction defined by the angles θ_2 and ϕ_2 with respect to

the receiving antenna spherical coordinate system. The power density (time-average of the Poynting vector) at the antenna terminals is $S(\theta, \phi) = \mathcal{P}_1(\theta, \phi) = E_1^2/\eta$, where E_1 is the rms value of the electric field vector due to the transmitting antenna 1 in the direction (θ, ϕ). Assume, also, that vector $\mathbf{E}_1$ is oriented so that the emf induced in the receiving antenna is the largest possible. The effective area of the antenna, $A_{\text{eff}}(\theta_2, \phi_2)$, is then defined by the equation

$$A_{\text{eff2}}(\theta, \phi) = \frac{P_{2 \text{ matched load, optimal reception}}}{S(\theta, \phi)} = \frac{P_{2 \text{ matched load, optimal reception}}}{\mathcal{P}_1(\theta, \phi)}. \tag{24.14}$$

(Definition of the effective area of a receiving antenna)

Questions and problems: Q24.11

24.6 The Friis Transmission Formula

Let us now examine a line-of-sight link (i.e., a radio link in which two antennas can "see" each other) between two antennas, as in Fig. 24.8. A transmitting antenna, of directivity $D_1(\theta_1, \phi_1)$ in the direction of the receiving antenna, radiates a power $P_{1\text{rad}}$, and a receiving antenna, with an effective area $A_2(\theta_2, \phi_2)$ in the direction of the transmitter, is matched to the receiver. The power delivered to the load connected to the receiving antenna terminals, from Eq. (24.14), is

$$P_{2 \text{ matched load, optimal reception}} = A_{\text{eff2}}(\theta_2, \phi_2)\mathcal{P}_1,$$

or, using Eq. (24.5*b*),

$$P_{2 \text{ matched load, optimal reception}} = A_{\text{eff2}}(\theta_2, \phi_2)\frac{D_1(\theta_1, \phi_1)P_{1\text{rad}}}{4\pi r^2},$$

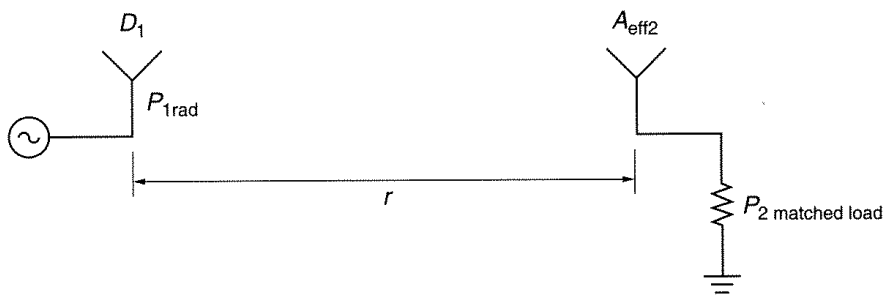

Figure 24.8 A line-of-sight link between two antennas

so that

$$P_{2\text{ matched load, optimal reception}} = P_{1\text{rad}} \frac{D_1(\theta_1, \phi_1) A_{\text{eff2}}(\theta_2, \phi_2)}{4\pi r^2}. \tag{24.15}$$

(Friis transmission formula)

This is known as the *Friis transmission formula* and it describes the power transmission in a line-of-sight antenna link. Note that we expressed it in terms of the directivity of the transmitting antenna, and the effective area of the receiving antenna.

The Relation Between Directivity and Effective Area

We can use the Friis transmission formula to show that the effective area of a receiving antenna can be expressed in terms of the antenna directivity, i.e., *in terms of the properties of the antenna in transmitting mode.* In the link shown in Fig. 24.8, we first assume antenna 1 is transmitting, and antenna 2 is receiving. The Friis formula gives for this case (labeled with a prime):

$$P'_{2\text{rec}} = P'_{1\text{rad}} \frac{D_1 A_{\text{eff2}}}{4\pi r^2}, \tag{24.16}$$

where the (θ, ϕ) dependence was omitted for brevity, i.e., the expression is valid for any direction. Now let us look at the second case (labeled with a double prime), when antenna 2 transmits, and antenna 1 receives. The Friis formula is now

$$P''_{1\text{rec}} = P''_{2\text{rad}} \frac{D_2 A_{\text{eff1}}}{4\pi r^2}. \tag{24.17}$$

Now we can apply reciprocity, as discussed in section 24.5. If instead of currents and voltages, we apply the reciprocity to transmitted and received powers, we obtain

$$\frac{P'_{2\text{rec}}}{P''_{1\text{rec}}} = \frac{P'_{1\text{rad}}}{P_{2\text{rad}}''}.$$

[This is obtained, with a little manipulation, from Eqs. (24.12*a*) and (24.12*b*); see problem P24.7.] Now Eqs. (24.16) and (24.17) are substituted in the last equation, and after canceling the powers and the term $4\pi r^2$, we obtain an interesting result:

$$\frac{A_{\text{eff1}}(\theta, \phi)}{D_1(\theta, \phi)} = \frac{A_{\text{eff2}}(\theta, \phi)}{D_2(\theta, \phi)}.$$

Here we have inserted the angular dependence again so as not to forget that D and A_{eff} change when the angle between the two antennas changes. What does this equation tell us? We did not assume anything about the two antennas, and we conclude that the ratio of the effective area and the directivity is always the same constant! If we knew what this constant was, we could always relate the directivity to the effective area, and therefore, the transmitting properties of an antenna to its receiving properties.

It is relatively easy to show that the integral of the directivity over all angles (sphere of any radius) is 4π. It can also be shown (but this is not at all easy and is

well beyond the scope of this book) that the integral of the effective area over all angles is equal to λ^2, where λ is the free-space wavelength. Then the ratios in the last equation are the same as the ratios of their integrals, or $\lambda^2/(4\pi)$, giving

$$A_{\text{eff}}(\theta, \phi) = \frac{\lambda^2}{4\pi} D(\theta, \phi). \tag{24.18}$$

(Relationship between antenna effective area and directivity)

This equation tells us that the directivity of an antenna is proportional to the effective area (and therefore size) of the antenna measured in free-space wavelengths.

Example 24.5—Effective area of a half-wave dipole with sinusoidal current distribution. From Eq. (24.10) in Example 24.4, we know the directivity of the half-wave dipole. From Eq. (24.18), the effective area of the half-wave dipole is thus

$$[A_{\text{eff}}(\theta)]_{\text{half-wave dipole}} = \sqrt{\frac{\mu}{\epsilon}} \frac{\lambda^2}{4\pi^2 R_{\text{rad}}} \frac{\cos^2(\pi/2 \cos\theta)}{\sin^2\theta}.$$

The maximal possible power delivered to the antenna is obtained if the wave is incident from the direction $\theta = \pi/2$ (and if, of course, the electric field vector is parallel to the dipole). In that case, with $R_{\text{rad}} = 73\,\Omega$, the preceding expression yields

$$[A_{\text{eff}}(\pi/2)]_{\text{max half-wave dipole}} \simeq 0.13\,\lambda^2.$$

This means that a matched half-wave dipole extracts from the wave the power contained in approximately $\lambda^2/8$ of the wavefront, as sketched in Fig. 24.9.

Example 24.6—Directivity of a 3-meter dish antenna. As another example, consider a 3-meter diameter dish for satellite TV at 12 GHz. Assuming the effective area of the dish (for $\theta = 0°$) is the same as its geometric area (which is approximately true for reflector antennas),

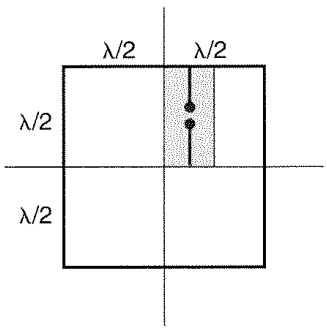

Figure 24.9 A matched half-wave dipole extracts from the wave the power contained in approximately $\lambda^2/8$ of the wavefront (shaded area with the sketch of the dipole)

from Eq. (24.17) we obtain $D = (4\pi \cdot \pi \cdot 1.5^2)/(0.025^2) = 141{,}975$. From this example, we see that antennas that are large when measured in wavelengths have high directivities, i.e., narrow beams.

Example 24.7—A radio communication link with half-wave dipole antennas. Assume that both the transmitting and receiving antennas in a line-of-sight link are half-wave dipoles a distance r apart. Let the dipoles be parallel to each other, and normal to the line connecting them, which ensures the maximal possible transmission of power under matched conditions. Specifically, let the frequency be $f = 900\,\text{MHz}$ (a cellular phone frequency), and the distance between the antennas $r = 100\,\text{m}$. The wavelength corresponding to $f = 900\,\text{MHz}$ is $\lambda = c/f = 0.333\,\text{m}$. We can express the Friis transmission formulas only in terms of the directivities, using Eq. (24.18):

$$P_{2\text{ matched load, optimal reception}} = P_{1\text{rad}} \frac{D_1 D_2}{r^2 \lambda^2}. \tag{24.19}$$

The ratio of the power received by the receiver matched load and the power radiated by the transmitting antenna is in this case

$$\frac{P_{2\text{ matched load, optimal reception}}}{P_{1\text{rad}}} = \left(\frac{D}{r\lambda}\right)^2 = \left(\frac{0.333}{4\pi\ 100}\right)^2 1.64^2 \simeq 1.9 \cdot 10^{-7}.$$

Thus, for a transmitted power of 10 W, the received power will be only about $1.9\,\mu\text{W}$.

Example 24.8—Other forms of the Friis formula. The Friis formula can also be expressed in terms of effective areas. From Eqs. (24.15) and (24.18) we obtain

$$P_{2\text{ matched load, optimal reception}} = \frac{A_{\text{eff1}}(\theta_1, \phi_1) A_{\text{eff2}}(\theta_2, \phi_2)}{(\lambda r)^2} P_{1\text{rad}}. \tag{24.20}$$

Usually, the quantity given for an antenna by the manufacturer is the maximal directivity in decibels, $D_{\text{dB}}(\theta, \phi) = 10 \log D(\theta, \phi)$. Also, commonly the radiated and received power are expressed with respect to a certain reference power level, the most frequent being 1 mW or 1 W. As an example, let the reference power level be 1 mW. We divide Eq. (24.19) by 1 mW, take the decimal logarithm of both sides, and multiply by 10 (to obtain decibels as calculated for power). The Friis formula expressed in decibels thus becomes

$$P_{2\text{ matched load, optimal reception, dBm}} = P_{1\text{rad, dBm}} + D_{1\text{dB}}(\theta_1, \phi_1) + D_{2\text{dB}}(\theta_2, \phi_2) + 20 \log \frac{\lambda}{r} - 21.984, \tag{24.21}$$

where the subscript "dBm" stands for "decibels over one milliwatt," and $-21.984 = 20 \log (1/4\pi)$. As an example of directivity expressed in decibels, the directivity of the satellite dish from Example 24.6 is $10 \log 141{,}975 = 51.5\,\text{dB}$.

Questions and problems: Q24.12, P24.6 to P24.9

24.7 Brief Overview of Other Antenna Types and Additional Concepts

There is a very wide variety of antennas used for different frequency ranges and different purposes. We conclude this brief chapter with a description of some frequently used antenna types. In addition, we will define some of the important antenna parameters that were not mentioned so far.

An antenna is said to be *narrowband* if it can efficiently emit and receive signals of frequencies in a relatively narrow frequency range, not exceeding more than a few percent of the central frequency. Antennas that can emit and receive efficiently in a broader frequency range are known as *broadband* antennas.

Antennas may have radiation patterns with several maxima. The largest is known as the *main lobe*, and the others as the *side lobes*. The *sidelobe level*, usually in decibels, is the level of the sidelobes with respect to the main lobe. The control of sidelobe levels frequently is not very strict, but in some applications it may be quite strict, requiring sidelobe levels of less than, for example, -40 dB.

An important property of the main antenna beam is the *beamwidth*. By definition, this is the angle between the directions for which the field intensity is $1/\sqrt{2} = 0.707$ that at the beam maximum. This corresponds to half the power density at the beam maximum.

Antennas are given additional descriptions relating principally to the frequency range in which they are used. We encounter low frequency (LF) antennas (30 kHz to 300 kHz), medium frequency (MF) antennas (300 kHz to 3 MHz), high frequency (HF) antennas (3 MHz to 30 MHz), very high frequency (VHF) antennas (30 MHz to 300 MHz), and ultra high frequency (UHF) antennas (300 MHz to 3000 MHz). The term "microwave antennas" usually implies frequencies above about 3000 MHz. Although some antennas can be used in several frequency ranges, these names also point to certain antenna types.

Sketched in Fig. 24.10a and 24.10b are two basic antenna types we have already met: the wire dipole and the wire monopole antenna. They are used in various forms at virtually all frequencies.

The antenna in Fig. 24.10c is known as the *loop antenna*. It does not radiate (and therefore does not receive a signal) normal to its plane, and the received signal is maximal from the directions in that plane. Therefore two or three such antennas at different points can be used for locating a transmitting antenna ("direction finding"). It is used at frequencies below about 1 GHz.

A great variety of relatively simple directional antennas are used for TV reception (6-MHz-wide channels in several frequency bands in the range 54 MHz to 890 MHz). One such narrowband antenna is the Yagi-Uda array sketched in Fig. 24.10d. It consists of a driven dipole backed by a *passive* (not directly excited) wire, known as the *reflector*, and several passive wires in front of the dipole (toward the transmitter), known as the *directors*. The electromagnetic field induces currents in the passive elements to influence the antenna radiation pattern. If properly designed, a Yagi antenna radiates a relatively narrow beam, but it is a frequency-sensitive structure. Until recently, the design of Yagi arrays was mostly experimental, but nowadays computer codes exist that enable the analysis and design of Yagi antennas with great precision.

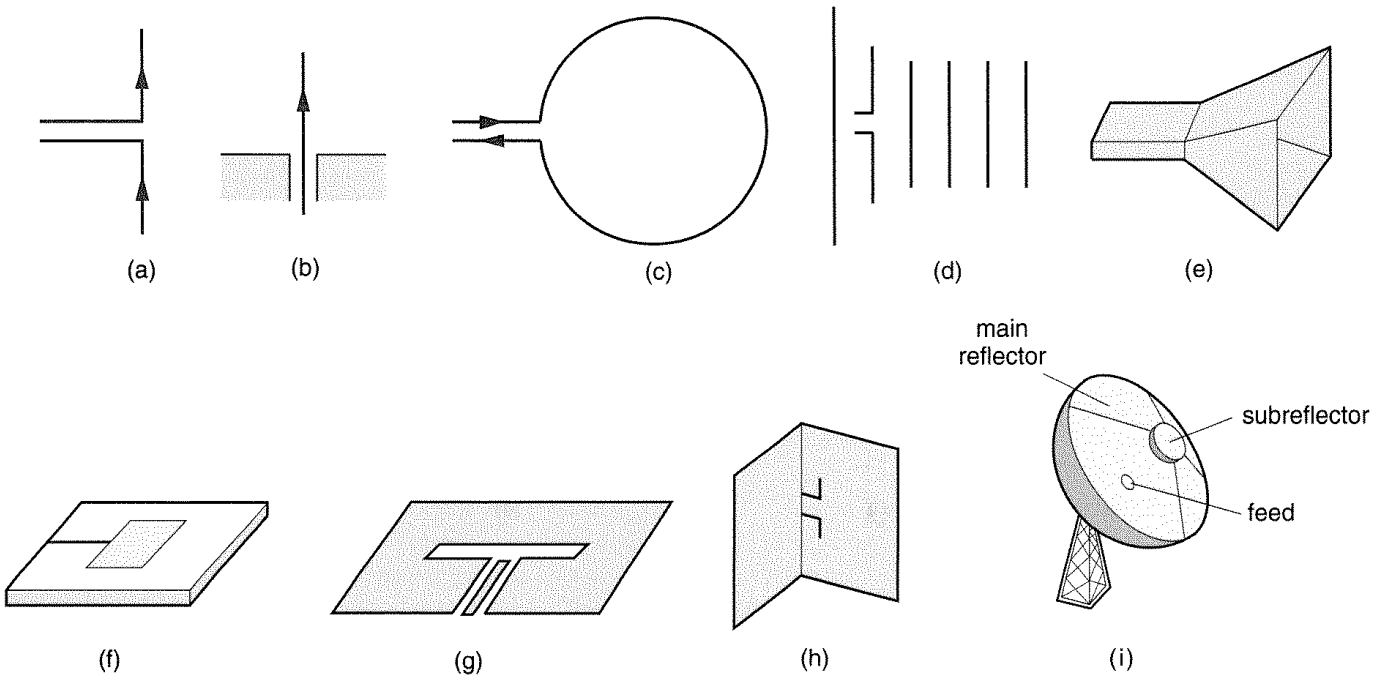

Figure 24.10 Sketches of some basic antennas: (a) a wire dipole; (b) a wire monopole; (c) a loop; (d) a Yagi-Uda array; (e) a horn; (f) a microstrip patch; (g) a slot; (h) a corner reflector; and (i) a parabolic reflector (dish)

If we terminate a rectangular waveguide in a rectangular open horn, as in Fig. 24.10e, the horn enhances the radiation from the open waveguide end. Such antennas are known as *horn antennas*, and are used exclusively as microwave antennas. The radiation of such antennas can be analyzed by assuming a field distribution over the horn aperture, from which, using certain advanced electromagnetic theory methods, the antenna far field can be calculated. Therefore, such antennas are also called *aperture antennas*.

Figure 24.10f is a sketch of the so-called *microstrip patch antenna*. On a dielectric substrate with ground metallization on the other side, a metallic patch is made and fed with a narrow metallic strip. The other terminal of the generator is connected to the ground metallization. (The transmission line obtained in this way is known as a *microstrip line*.) Microstrip patch antennas are narrowband, but due to their flat shape they have many applications where flush mounting is desirable (sometimes these flush-mounted antennas are called "conformal").

An entirely different family of antennas are *slot antennas*, obtained by cutting slots of various shapes in metal screens, as in Fig. 24.10g. There is a theory that enables the analysis of radiation and circuit properties of slot antennas in terms of those of "complementary" antennas, i.e., metallic antennas in the form of the slot, with the screen removed. Slot antennas are used as microwave antennas.

To enhance radiation in a specific direction, metallic reflectors are used as in Fig. 24.10h, where a dipole antenna is backed by a *corner reflector*, and Fig. 24.10i, where a parabolic reflector is used to concentrate the antenna radiation into a very narrow beam, sometimes termed a "pencil beam." Whereas antennas of the form in

Fig. 24.10h are used above about 100 MHz, antennas with a parabolic reflector are meaningless for frequencies below about 1 GHz, because the reflector must be many wavelengths in diameter in order to concentrate the radiation efficiently.

Example 24.9—High-directivity antennas. Since $D = (4\pi/\lambda^2)\,A_{\text{eff}}$, antennas with high directivity are large when measured in wavelengths. For example, an antenna that has a maximal directivity of 30 dB should be at least $1000/4\pi \simeq 80$ wavelengths square, or if it were a square, about 9 wavelengths on the side. At 300 MHz, this would be about 9 meters on the side, at 1 GHz about 2.7 m on the side, and at 30 GHz about 9 cm on the side. So, if an airplane or a satellite needs to carry a highly directional antenna, choosing a higher frequency of operation allows for a smaller and lighter antenna, which translates to less fuel, more room, and ultimately lower cost.

When high directivity is required, antenna arrays are frequently used instead of a single very large antenna. In antenna arrays, a large number of smaller, low-directivity elements are fed with a common feed, which makes the array look like one antenna. The elements are usually about a half wavelength apart. This antenna has a geometric area that can be many wavelengths across, and can therefore have a very high directivity. Further, the radiation pattern can be tailored to satisfy different requirements at different angles. For example, an array in one dimension (say, a 10 by 1 array) of antennas will have a high directivity in the plane of the 10 elements, and a low directivity in the orthogonal plane, and its effective area will be very roughly $5\lambda^2$.

24.8 Chapter Summary

1. Antennas are metallic and/or dielectric structures used for radiation and reception of electromagnetic waves. As a rule, they have two closely spaced terminals in the circuit-theory sense.
2. A transmitting antenna excited by a sinusoidal generator behaves as a load of a certain frequency-dependent impedance, known as the *antenna impedance*.
3. The simplest antenna is the short electric dipole antenna, or the Hertzian dipole. The current distribution along the Hertzian dipole is assumed to be constant. The radiation properties of the Hertizan dipole can be obtained in a relatively simple manner.
4. Antennas do not radiate equally in all directions. The directional radiation properties of antennas are described by two basic quantities, the antenna radiation pattern and its directivity. The maximal directivity is usually of particular interest.
5. The gain of an antenna is at most as high as its directivity, since it includes any losses in the antenna.
6. If an antenna is used for receiving, it transfers a part of the energy of the incident electromagnetic wave to its load (which is usually the input to an amplifier). A receiving antenna behaves with respect to the load as an equivalent Thévenin generator and has the same internal impedance as when it is transmitting.
7. Directional properties of a receiving antenna are the same as those when it is used for transmitting.

QUESTIONS

Q24.1. You have a black box with two terminals. You connect a generator to these terminals and find out that the black box behaves as an impedance. Can you check by observing the measured impedance whether the terminals belong to a transmitting antenna inside the box? Explain.

Q24.2. You have a black box with two terminals. You connect a load to these terminals and find out that the black box behaves as a generator. Can you check by observing the measured current in the load whether the terminals belong to a receiving antenna inside the box? Explain.

Q24.3. Why are the images of antennas in Fig. 24.3 as indicated?

Q24.4. On many short antennas there are small conducting balls at each end. What are these balls for?

Q24.5. Assuming that the dipole shown in Fig. 24.4 has no spheres at the ends, will there be a current in the two short wire segments? If the answer is yes, what do you expect this current distribution to be like?

Q24.6. What is the relationship between the phasor current I in the dipole in Fig. 24.4, and the charges Q and $-Q$ on the dipole end spheres?

Q24.7. Take a pencil and assume it is a Hertzian dipole. What is its radiation pattern in space like?

Q24.8. In the preceding question, define an E plane and an H plane of the radiation pattern.

Q24.9. What is an isotropic antenna? Can it be made? If you think it cannot, explain why.

Q24.10. Why is the directivity of an isotropic antenna equal to unity, or zero dB?

Q24.11. What are the conditions implicit in the definition of the effective antenna area?

Q24.12. Can the Friis transmission formula be used for the analysis of a radio communication channel if the transmitting antenna is not matched? Or if the receiving antenna is not matched? How would you modify the formula?

PROBLEMS

P24.1. Prove that the impedance of any antenna above a perfectly conducting ground, with the generator driving the antenna connected between the ground and the antenna terminal, is one half that of the symmetrical antenna obtained with the image of the antenna.

P24.2. A thin two-wire transmission line with conductors of radius $a = 1\,\text{mm}$ and distance between them $d = 5\,\text{cm}$ is driven at one end by a generator with a rms value of the emf $\mathcal{E} = 10\,\text{V}$ and frequency $f = 100\,\text{MHz}$. The line length is $b = 50\,\text{cm}$, and the other end of the line is open-circuited. Assuming that the line conductors do not radiate, but that the short segment with the generator does, determine approximately the electric field strength at a distance $r = 1\,\text{km}$ from the antenna. (*Hint:* consider the short segment with the generator as a Hertzian dipole.)

P24.3. A Hertzian dipole of length $l = 1\,\text{m}$ is fed with a current of rms value $I = 1\,\text{A}$ and of frequency $f = 1\,\text{MHz}$. Find the rms values of E_θ and H_ϕ in the equatorial plane (plane $\theta = \pi/2$) of the dipole at a distance of $r = 10\,\text{km}$.

P24.4. Using a system of two half-wave dipoles, construct an antenna system that radiates a circularly polarized wave in one direction. State clearly how you would make the feed.

P24.5. A short vertical transmitting antenna of height h has a conducting plate at the top, so that the current along the antenna is practically uniform, of rms value I. At the receiving point, at a distance d from the antenna, only the wave reflected by the ionosphere arrives, as shown in Fig. P24.5. The ionosphere can be approximated by a perfectly conducting plane at a height H above the surface of the ground. Assuming that the ground at both the transmitting and receiving point is perfectly conducting, and neglecting the curvature of the earth, determine the rms value of the electric field intensity at the receiving point. The wavelength of the radiated wave is λ.

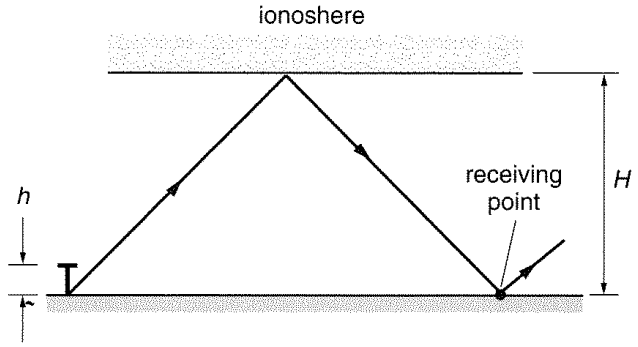

Figure P24.5 Vertical antenna above ground

P24.6. In Example 24.7, replace one of the antennas (for example, in a cellular phone, this would be the base-station antenna) by an antenna with a higher directivity and calculate the received power for a directivity of (1) 6 dB, (2) 10 dB, and (3) 20 dB.

P24.7. Assume that in a communications link two matched lossless antennas, A and B, are r apart in each other's far fields. The antenna directivities and effective areas in the line-of-sight direction are D_A, A_A, and D_B, A_B, respectively. First antenna A transmits a power P_{A1}, while antenna B receives a power P_{B1}. Then antenna B transmits a power P_{B2}, while antenna A receives a power P_{A2}. Using the reciprocity condition, which says that $P_{A1}/P_{B1} = P_{B2}/P_{A2}$ (think about what this means), show that the ratio of the directivity to the effective area is a constant for any antenna.

P24.8. Derive the Friis formula in terms of effective area only. In a microwave relay system for TV each antenna is a reflector with an effective area of 1 m^2, independent of frequency. The antennas are 10 km apart. If the required received power is $P_r = 1$ nW, what is the minimum transmitted power P_t required for transmission at 1 GHz, 3 GHz, and 10 GHz?

P24.9. Derive the Friis formula in terms of directivities only.

25

Some Practical Aspects of Electromagnetic Waves

25.1 Introduction

Applications of electromagnetic waves are numerous, ranging from cooking food to controlling a faraway spacecraft and receiving information from it. Electromagnetic waves cover a very broad frequency (wavelength) spectrum, and it is impossible in this text to even attempt to cover applications in all regions of the spectrum. Therefore, we confine ourselves mainly to waves used in the versatile area of communications, as the readers of this text are likely to spend a part of their professional lives dealing with this subject. The word "communications" refers broadly to sending and receiving a signal that contains some useful information. This signal might be sent along a coaxial cable or through a waveguide, or radiated or received by an antenna, or propagated along an optical fiber, for example. We will describe some issues related to these different ways of communicating, but we will not deal with the information itself. At the end of the chapter, some other common applications of electromagnetic waves, such as cooking, are described briefly.

25.2 Power Attenuation of Electromagnetic Waves

When one sends or receives information using electromagnetic (radio) waves, the maximum distance at which this can be done is limited by the amount of power available at the sending end, and the loss of the wave energy by the time it gets to the receiving end, assuming a certain receiver sensitivity. The path loss varies with the medium through which the wave is propagating, as well as the frequency (wavelength) of the wave. So let us calculate the loss per unit distance for a few different cases, and then do a performance comparison. In the following examples, we calculate the loss in a coaxial cable, a rectangular waveguide, an optical fiber, and a line-of-sight radio link (Fig. 25.1), referring to knowledge we gained in previous chapters. In each case, we consider a link where the power at the transmitting (sending) end is P_T and the received power $P(r)$ is a function of the distance r between the transmitter and receiver. So, at a point r away from the transmitter, the received power is $P(r) = P_T f(r) < P_T$. What is this function $f(r)$ in different cases?

Example 25.1—Attenuation of an electromagnetic wave in a coaxial cable. In a coaxial cable, if the losses are not large, the power along the line as a function of distance r from the generator can be expressed as

$$P(r) = P_T e^{-2\alpha r} \tag{25.1}$$

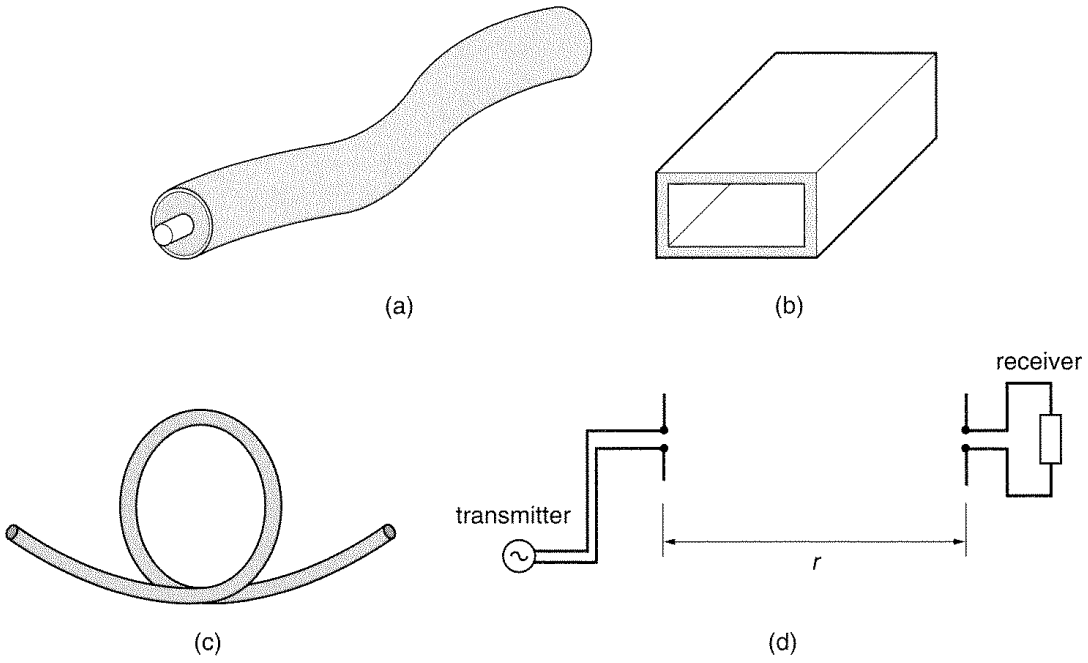

Figure 25.1 (a) A coaxial cable, (b) a rectangular waveguide, (c) an optical fiber, and (d) antennas in a line-of-sight radio link are used in communications as ways of transferring information contained in an electromagnetic wave. The coaxial cable was discussed in detail in Chapter 18, the rectangular waveguide in Chapter 23, and the line-of-sight link in Chapter 24.

(see section 18.4, and note that power is the product of the voltage and current, giving the factor of 2 in the exponent). Therefore,

$$-\frac{dP(r)}{dr} = 2\alpha P_T e^{-2\alpha r} = 2\alpha P(r) = \frac{dP_{\text{losses}}(r)}{dr}. \tag{25.2}$$

The attenuation coefficient α comes from losses in the conductor, described by the resistance per unit length, R', and losses in the dielectric, described by the conductance per unit length, G', as described in Fig. 18.10. Usually the conductive losses are dominant, and

$$\frac{dP_{\text{losses}}(r)}{dr} = R'|I(r)|^2.$$

The power transmitted to a point r away from the line's beginning can be expressed in terms of the current $I(r)$ and the characteristic impedance (assuming no losses) as $P(r) = Z_0|I(r)|^2$, where $Z_0 \simeq \sqrt{L'/C'}$. So

$$\alpha = \frac{R'}{2Z_0} = \frac{R'}{2}\sqrt{\frac{C'}{L'}}. \tag{25.3}$$

As one example, let us calculate the loss at 1 MHz in a coaxial cable made of copper, filled with a dielectric of permittivity $\epsilon_r = 3$, and of dimensions such that the inner conductor radius $a = 0.45\,\text{mm}$ and inner radius of the outer conductor $b = ae$. The skin depth is $\delta = 0.067\,\text{mm}$ (Example 20.1), which means that the current is not distributed through the entire cross section. We obtain in this case that $R' \simeq \rho/(2\pi a\delta) = 0.093\,\Omega/\text{m}$, and $Z_0 = 50\,\Omega$. This gives $\alpha = 0.00093\,\text{Np/m} = 0.016\,\text{dB/m}$, so that $f(r) = e^{-0.0019r}$, where r is in meters. This just corresponds to the loss in the inner conductor. In the outer conductor, the losses are lower (see problem P25.5).

What happens at higher frequencies with the loss in coaxial cables? As another example, let us look at losses at 0.1, 1, and 10 GHz in a high-frequency 50-Ω RG-58 cable, made of copper with polyethylene dielectric, with $a = 0.45\,\text{mm}$ and $b = 1.47\,\text{mm}$. The loss can usually be found in manufacturer's data sheets, and for this cable at 0.1 GHz, the loss is about 0.2 dB/m, and at 1 and 10 GHz it increases to 0.66 and 2.6 dB/m, respectively. This means that at 10 GHz almost half of the power (which would be 3 dB) is lost after only 1 m of propagation. In this case, the function $f(r) = e^{-0.6r}$, where r is in meters. Why is the loss this high?

Example 25.2—History: the transatlantic telegraphy cable. Reduction of losses along lines by increasing inductance per unit length. When the first transatlantic cable was laid in the Atlantic Ocean, the engineers did not understand that the loss over several thousand kilometers would make the cable impractical (see problem P25.1—calculate the loss for $r = 5000$ km at 10 kHz for practice). The famous British physicist Oliver Heaviside had warned the engineers about losses, but they did not listen. Later, Mihailo Pupin, a Columbia University professor, noticed that in practice, the first term of α in parentheses in Eq. (18.36), repeated here for convenience,

$$\alpha \simeq \frac{1}{2}\left(R'\sqrt{\frac{C'}{L'}} + G'\sqrt{\frac{L'}{C'}}\right), \tag{18.36}$$

is much greater (several orders of magnitude) than the second, due to the relatively large value of R'. He next realized that it is possible to reduce this term considerably by increasing L',

without making the second term prohibitively large. He then proposed to reduce α by placing series inductive coils along the cable at regular distances. These are today called Pupin coils, and they enabled transmission of signals along transmission lines of great lengths, including the transatlantic cable.

Let us consider a typical coaxial line and estimate the first and second terms in parentheses of Eq. (18.36). Let the line dielectric have a relative permittivity $\epsilon_r = 3$, permeability μ_0, and conductivity 10^{-12} S/m. Let the line conductors be copper, of conductivity 56×10^6 S/m, and let the ratio of conductor radii (see Table 18.1) be $b/a = \mathrm{e} = 2.71828$.

Using the expressions for C', G', L', and R' in Table 18.1, we find that $C' \simeq 167\,\mathrm{pF/m}$, $G' \simeq 0.3\,\mathrm{pS/m}$, $L' \simeq 0.2\,\mu\mathrm{H/m}$, and $R' \simeq 0.0055\,\Omega/\mathrm{m}$. With these values, the first term in the expression for α (including 1/2) is about 0.8×10^{-4} Np/m, and the second term is about 5×10^{-12} Np/m. The first term, due to imperfect cable conductors, is indeed much greater than the second, due to imperfect cable dielectric.

By increasing L' artificially, as Pupin did by connecting lumped series coils in the cable, the first term can be substantially reduced, and therefore α made smaller. Note that the attenuation of the wave is proportional to $\mathrm{e}^{-\alpha r}$, so a 10-fold increase in inductance per unit length results in about a 24-fold decrease in signal attenuation. This means that if a cable with no Pupin coils has an attainable range of 100 km, with a 10-fold artificial increase of the line series inductance per unit length the range is increased to about 2400 km.

Example 25.3—Attenuation of electromagnetic waves in a rectangular waveguide for the dominant mode. Losses in hollow metal waveguides depend on the mode propagating in the waveguide, the type of metal and dielectric used to make the waveguide, the geometry of the waveguide, and frequency. It can be shown (see, e.g., S. Ramo et al., *Fields and waves in communication electronics*, 3d ed., J. Wiley & Sons, 1993, p. 423) that the attenuation coefficient for the dominant TE_{10} mode in a rectangular waveguide of sides a and b is given by

$$\alpha = R_s \sqrt{\frac{\epsilon}{\mu}} \frac{a/b + 2f_c^2/f^2}{a\sqrt{1 - f_c^2/f^2}} \qquad (\mathrm{TE}_{10}), \tag{25.4}$$

where R_s is the surface resistance of the metal walls, Eq. (20.10). Consequently, the losses depend on the metal conductivity and frequency. For a typical X-band (8.2 to 12.4 GHz) waveguide ($a = 25.4$ mm, $b = 12.7$ mm), made of brass plated with silver and a rhodium anticorrosion coating, the conductivity is $6.17 \cdot 10^7$ S/m, and the skin depth at 10 GHz is δ =0.64 μm, yielding $R_s = 2.53\,\Omega$ and $\alpha = 0.0883\,\mathrm{Np/m} = 0.767\,\mathrm{dB/m}$.

Note that at very high frequencies, when the skin depth is on the order of the conductor surface imperfections (the surface cannot be made absolutely flat), the surface resistance becomes substantially larger than its theoretical value for a perfectly flat surface.

It should be noted that some other kinds of metallic waveguides have field profiles that result in lower losses than in the previous case. An example is a TM_{11} mode in a waveguide with circular cross section, with a typical $\alpha = 0.01\,\mathrm{dB/m}$. Initially there was a large development effort, mostly in Bell Labs, to use this kind of waveguide for the entire phone network across the United States, but before the network was built, optical fibers were shown to have lower loss and cost, so the waveguide technology was never implemented. In Socorro, New Mexico, however, there is a large radio telescope [Very Large Array (VLA), nicely shown in the 1997 movie *Contact*] in which the signals received from 27 large dish antennas (each 25 m in diameter) formerly were propagated at 44 GHz through a circular waveguide to the operations center that is about 60 km away from the telescope. Recently, the circular waveguide was replaced by optical fibers, but the waveguide is still used as the pipe for the fibers.

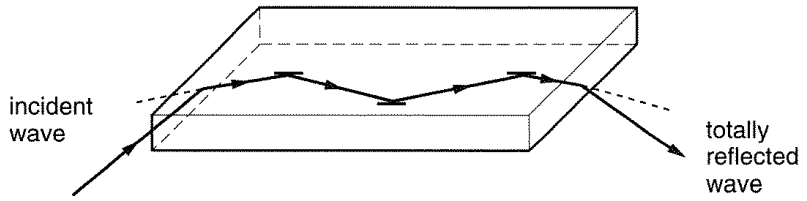

Figure 25.2 Sketch of wave propagation along a dielectric slab

Example 25.4—Dielectric waveguides and optical fibers. Optical fibers are used as waveguides for electromagnetic waves in the visible and infrared part of the frequency spectrum, with wavelengths between roughly 300 nm and 10 μm (optical engineers usually think in terms of wavelength, whereas radio engineers think in terms of frequency). Fibers are so-called *dielectric waveguides*. The simplest dielectric waveguide is a flat dielectric slab. Because the permittivity of the slab is always greater than ϵ_0, a possibility exists that the wave propagating in the slab is totally reflected at the interface.

If we excite in the slab a plane wave incident on one slab face at an angle greater than the critical angle, it will be reflected totally at the same angle toward the other face, then reflected totally from that other face, etc. So the wave will bounce between the two slab faces, and the slab will serve as a guiding medium of the wave, as sketched in Fig. 25.2.

The principle of optical fibers is essentially the same, although the wave types propagating along them are more complicated. Thanks to total reflection, however, these waves also are restricted to the domain of the fiber. The fiber is made of an inhomogeneous dielectric (quartz), and roughly speaking it has a core and an outer cladding layer, sketched in Fig. 25.3. The permittivity of the core is typically a fraction of a percent higher than that of the outer part (it is germanium doped). The cladding has a permittivity of about $\epsilon_r = 2.1$ (an optical index of $n = \sqrt{\epsilon_r} = 1.46$). The typical attenuation of a so-called single-moded fiber at a wavelength of 1.55 μm is 1 dB/km = 0.001 dB/m (Corning specification sheets, 1998), and for very specialized low-loss fibers it can be as low as 0.1 dB/km.

Example 25.5—Attenuation of electromagnetic waves in a line-of-sight radio link through a vacuum. In a line-of-sight radio link (which means that the radio wave travels between two antennas directly, with no reflections), the power loss function $f(r)$ is given by the Friis transmission formula:

$$P(r) = P_T \frac{G_T A_R}{4\pi r^2} = P_T \frac{A_T A_R}{\lambda^2 r^2}, \tag{25.5}$$

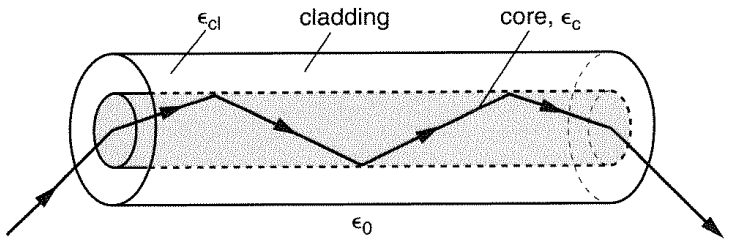

Figure 25.3 Sketch of wave propagation along an optical fiber, based on total internal reflection

where G_T is the gain of the transmitting antenna, and A_R and A_T are effective areas of the receiving and transmitting antennas used in the link. We can measure these effective areas in terms of the operating wavelength λ used in the link. If the two antennas are equal, $n\lambda^2$ large, and we assume they are well designed so that the effective areas are roughly equal to their geometric areas, we get

$$\frac{P(r)}{P_T} = f(r) = \frac{n^2\lambda^2}{r^2}. \tag{25.6}$$

This means that the larger the antennas are (measured in wavelengths), the lower the loss of power between the transmitter and receiver. Large antennas, of course, correspond to high directivity (gain). As an example, a standard X-band horn antenna measures about 2 by 2.6 wavelengths at 10 GHz, yielding $f(r) = 0.025/r^2$. If instead of this horn, 3-m round dishes are used, $f(r) = 55{,}000/r^2$. In the second case, a much larger distance can be spanned with the same receiver sensitivity.

In the preceding examples we have seen that in coaxial cable, waveguides, and optical fiber, the power decays exponentially away from the transmitter, with very different decay constants (on the order of 1 dB/m in a coaxial cable, 0.1 dB/m in a waveguide, and 0.01 dB/m in a fiber). If antennas are used, however, the power drops as $1/r^2$, which is a function that decays more slowly than an exponential for large distances. A plot that shows the attenuation as a function of distance is shown in Fig. 25.4, for the attenuation coefficient values calculated in Examples 25.1 to 25.5. The lower attenuation for long distances is one of the reasons that antennas are used

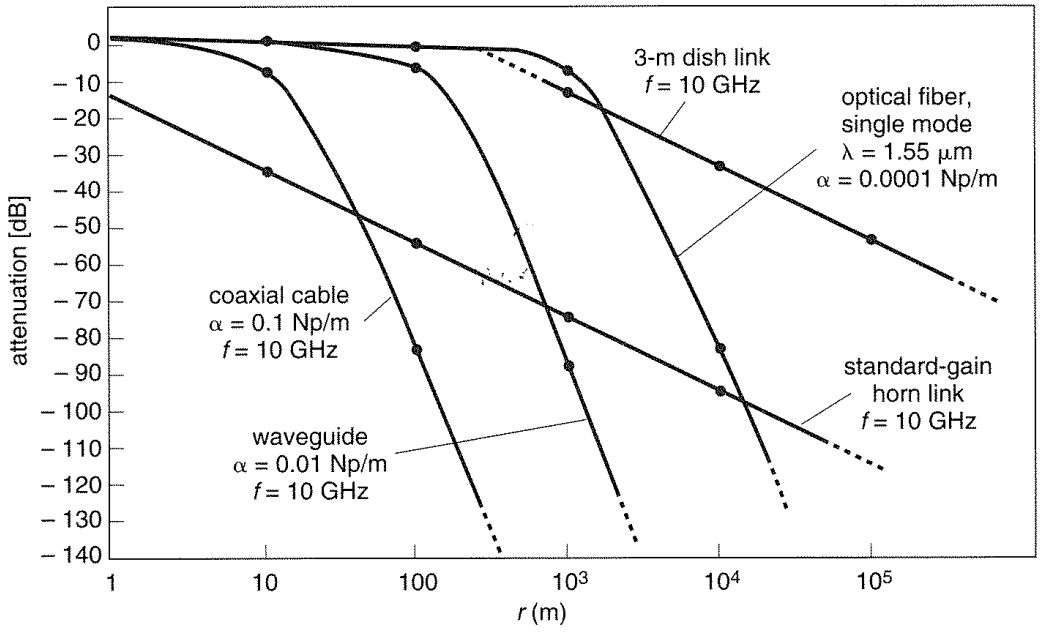

Figure 25.4 The attenuation function $\log f(r)$ for coaxial cable, rectangular waveguide, and a 3-m diameter dish antenna line-of-sight link at 10 GHz. Attenuation in an optical fiber at 1.55-μm wavelength is also shown. Note the logarithmic scale.

for communications. Another reason is that in many cases, for example in aircraft guidance, satellite communications, portable phones, and pagers, it would not be practical to use cables.

Example 25.6—Curvature of the earth and effective earth radius in line-of-sight links. AM broadcasting systems rely on surface wave transmission between two points on the earth's surface. Shortwave radio systems bounce waves off the ionosphere and the earth's surface. The UHF and VHF radio waves used for communications by airplanes, as well as microwaves in radio relay links, propagate along a direct path. As mentioned, this is called *line-of-sight* propagation, illustrated in Fig. 25.5. The figure shows how the range is limited by the curvature of the earth. That is why almost all radio relay stations are put on high peaks, even though the weather conditions at these places often complicate the design (and very few people want to work there). From the figure,

$$(R+h)^2 = R^2 + r^2, \tag{25.7}$$

where R is the radius of the earth, h is the height of the antenna above ground, and r is the range of the communication link. If we solve for r,

$$r = \sqrt{h^2 + 2Rh} \simeq \sqrt{2Rh} \tag{25.8}$$

since $R \gg h$.

Equation (25.8) predicts shorter ranges than the ones achievable in reality. The reason is the change in the refractive index of the atmosphere, so that the waves really follow a curved path, not a straight one. This is sketched in Fig. 25.6. The waves bend toward the denser (lower) layers, and this gives longer ranges. This effect varies quantitatively depending on the location on the earth and the hour of the day. A reasonable approximation is to use an effective radius for the earth, R_{eff}, somewhat larger than the real radius. It turns out that if this radius is taken to be about 4/3 of the actual radius, or about 8500 km, the wave refraction is fairly accurately taken into account. If both the receiving and transmitting antennas are above ground, the line-of-sight approximate range formula becomes

$$r = \sqrt{2R_{\text{eff}}h_1} + \sqrt{2R_{\text{eff}}h_2}. \tag{25.9}$$

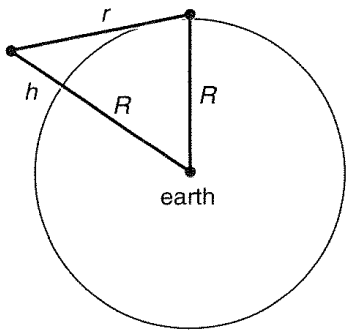

Figure 25.5 Line-of-sight path limit on the curved earth

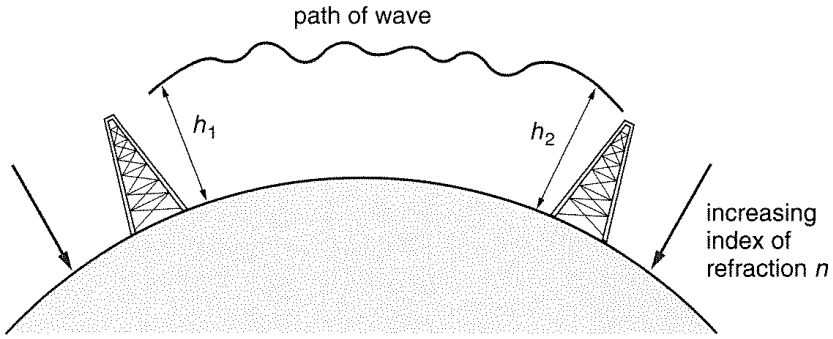

Figure 25.6 The variation of the refractive index of the atmosphere makes the paths of the waves longer.

When deriving the antenna link path loss, we assumed no extra attenuation due to the atmosphere. This is a good assumption in clear weather close to the earth's surface, but only in a certain range of frequencies. Rain and snow degrade the path loss significantly, but even clear atmosphere has a frequency-dependent attenuation curve that has strong peaks due to specific properties of oxygen, the hydroxide (OH) radical, water vapor, and other constituents of the atmosphere. This dependence dictates the frequencies used for specific purposes, as described later in this chapter. In satellite communications, the waves pass through the ionosphere, a layer of the atmosphere that has unique properties and that also significantly affects wave propagation. We will be able to analyze the effects of the ionosphere in more detail after we develop an understanding of plane wave propagation through ionized gases, described in the next section.

Questions and problems: Q25.1 to Q25.5, P25.1 to P25.8

25.3 Effects of the Ionosphere on Wave Propagation

The upper layer of the atmosphere, between about 50 and 500 km above the earth's surface, is a highly rarefied ionized gas. This ionized layer of the atmosphere is known as the *ionosphere*. It has a pronounced influence on the propagation of electromagnetic waves in a wide frequency range. It is therefore important to understand this influence when dealing with any kind of radio communications in which the waves travel through the ionosphere. The influence of the ionosphere on radio-wave propagation is of great interest starting from very low frequencies (10 to 100 kHz), to short waves (up to 30 MHz), but also for higher frequencies than these.

25.3.1 PLANE WAVE PROPAGATION THROUGH THE IONOSPHERE (AN IONIZED GAS)

This section is aimed at presenting the basic theory of propagation of uniform plane electromagnetic waves in ionized gases. To simplify the analysis, the collisions of moving charged particles with neutral gas molecules, resulting in wave attenuation, will be ignored.

Consider a uniform plane electromagnetic wave of angular frequency ω propagating in an ionized gas. Let there be N ions per unit volume of charge Q and mass m. Assume that at a fixed point inside the gas the electric field strength of the wave varies in time as $\mathbf{E}(t) = \mathbf{E}_m \cos \omega t$. The equation of motion of a single ion under the influence of the electric and magnetic field of the wave has the form

$$m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = Q\mathbf{E}_m \cos \omega t + Q\mathbf{v} \times (\mu_0 \mathbf{H}_m) \cos \omega t. \tag{25.10}$$

We know that for a uniform plane wave $H_m = \sqrt{\epsilon_0/\mu_0}E_m$, and that $\sqrt{\epsilon_0\mu_0} = 1/c_0$. Therefore, the second term on the right side of this equation is approximately proportional to the first term multiplied by the ratio v/c_0. Because the velocities that ions can acquire in a time-harmonic electric field are much smaller than the velocity of light in a vacuum, the second term can be ignored. If we multiply the equation thus obtained by $\mathrm{d}t$ and then integrate, we obtain

$$\mathbf{v} = \frac{Q}{\omega m}\mathbf{E}_m \sin \omega t. \tag{25.11}$$

The integration constant, a time-constant velocity, is zero because a time-harmonic electric field cannot produce a steady drift of ions.

Knowing the velocity of ions, we know also the current density that they produce,

$$\mathbf{J} = NQ\mathbf{v} = \frac{NQ^2}{\omega m}\mathbf{E}_m \sin \omega t. \tag{25.12}$$

The second Maxwell's equation now becomes

$$\nabla \times \mathbf{H} = \mathbf{J} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} = \left(\frac{NQ^2}{\omega m} - \epsilon_0 \omega\right) \mathbf{E}_m \sin \omega t, \tag{25.13}$$

or

$$\nabla \times \mathbf{H} = -\omega \left(\epsilon_0 - \frac{NQ^2}{\omega^2 m}\right) \mathbf{E}_m \sin \omega t. \tag{25.14}$$

For $N = 0$ (a vacuum), the second term in the parentheses does not exist. Therefore, the presence of the ions can be represented by an equivalent *reduction* in permittivity. Because this reduction is proportional to Q^2/m, we conclude that the sign of the ions is unimportant, and that those ions having the largest ratio Q^2/m have the most pronounced influence. Because this ratio is the largest for free electrons, they are the dominant factor for plane wave propagation in an ionized gas.

Let us define the equivalent (or effective) permittivity of an ionized gas by

$$\epsilon' = \epsilon_0 \left(1 - \frac{NQ^2}{\omega^2 \epsilon_0 m}\right) = \epsilon_0 \left(1 - \frac{\omega_c^2}{\omega^2}\right), \tag{25.15}$$

where

$$\omega_c = \sqrt{\frac{NQ^2}{\epsilon_0 m}} \tag{25.16}$$

is known as the *critical angular frequency*, and $f_c = \omega_c/2\pi$ as the *critical frequency*, of the ionized gas.

The propagation coefficient can now be written as

$$\beta = \omega\sqrt{\epsilon'\mu_0} = \frac{\omega}{c_0}\sqrt{1 - \frac{\omega_c^2}{\omega^2}}, \tag{25.17}$$

and the phase velocity of the wave is given by

$$v_{\text{ph}} = \frac{\omega}{\beta} = \frac{c_0}{\sqrt{1 - \omega_c^2/\omega^2}}. \tag{25.18}$$

If $\omega > \omega_c$, the expression under the square root is positive, so that $v_{\text{ph}} > c_0$. We know that this is only a geometrical velocity, and that the velocity of propagation of a signal, or of energy, is less than c_0 (see Example 21.6).

If $\omega < \omega_c$, however, the expression under the square root is negative and β becomes imaginary, which means that waves of angular frequencies less than ω_c cannot propagate in this ionized gas. This is why ω_0 is called "the critical angular frequency," and f_c "the critical frequency."

Example 25.7—Penetration of plane waves through the ionosphere. The critical frequency of the ionosphere varies greatly with the distance from the earth's surface, as well as with the hour of the day and month of the year, and with the sun's activity. Roughly, for a plane wave propagating vertically, this frequency ranges from about 3 to 8 MHz. Therefore, no wave of a frequency below about 3 MHz can escape the earth, nor can such a wave reach us from outer space. Therefore, for communications via satellites we must use higher frequencies than these.

It is interesting that at very low frequencies (less than about 100 Hz), the wavelength is much larger than the ionosphere thickness, and such waves can penetrate the ionosphere.

25.3.2 REFLECTION AND REFRACTION OF PLANE WAVES IN THE IONOSPHERE

We shall now see what happens if a plane wave is emitted from the earth's surface toward the ionosphere at an arbitrary angle. The ionosphere is an ionized layer of the atmosphere, and we have shown earlier that free electrons have the most pronounced influence on wave propagation. The concentration of electrons changes with the height and depends greatly on the hour of the day, and significantly on the season of the year, the latitude, and the activity of the sun.

During the day, the variation of the concentration of electrons with height above the earth's surface shows certain regularities, resembling four blurred layers. Starting from the surface of the earth, the layers are designated as D, E, F_1, and F_2. The corresponding heights are 50 to 70 km (the D layer), 100 to 150 km (the E layer),

about 200 km (the F_1 layer), and between 250 and 300 km (the F_2 layer). During the night, the D and E layers practically disappear, and the layers F_1 and F_2 merge into a single layer, F, between about 250 and 400 km above earth. The critical frequency of layers E, F_1, F_2, and F are about 3 to 4 MHz, 4 to 5 MHz, 6 to 8 MHz, and 3 to 5 MHz, respectively. The lowest layer, D, in which collisions of electrons with neutral atoms and molecules are most pronounced, dominates the attenuation of waves propagating through the ionosphere. This is why attenuation of waves reflected from the ionosphere is the largest during the day, when this layer is present.

Assume that the ionosphere critical frequency versus height h above the earth's surface is as in Fig. 25.7, with a maximum critical frequency $f_{c\,\max}$ at a certain height. Assume that an antenna radiates a plane wave of frequency f so that it is incident at an angle θ_0 on the lower boundary of the ionosphere (which we assume to be plane), as in Fig. 25.7. We know that in the case of a homogeneous ionized gas, it can be considered as a medium of equivalent apparent permittivity from Eq. (25.15):

$$\epsilon' = \epsilon_0 \left(1 - \frac{\omega_c^2}{\omega^2}\right) = \epsilon_0 \left(1 - \frac{f_c^2}{f^2}\right). \tag{25.19}$$

We can imagine the ionosphere consisting of many thin homogeneous layers of slightly different critical frequencies, as indicated in Fig. 25.7. The incident wave then

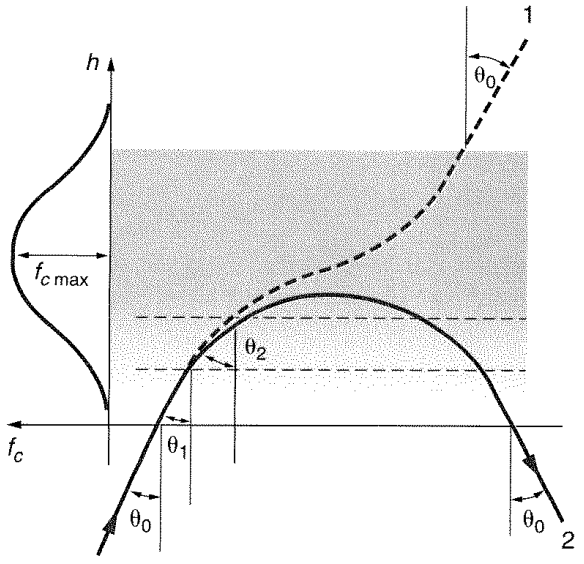

Figure 25.7 A model of the ionosphere critical frequency versus the height, h, above earth's surface (diagram on the left), and a sketch of propagation of two waves incident on the ionosphere. The solid line represents a wave of frequency less than the maximum critical frequency. The dashed line represents a wave of frequency greater than the maximum critical frequency.

refracts on the first layer, with no reflected wave because the apparent permittivity of that layer is almost the same as ϵ_0. It is next incident at a new angle, θ_1, on the next layer, and so on. By applying Snell's law, we obtain the following sequence of equations:

$$\frac{\sin\theta_0}{\sin\theta_1} = \sqrt{\frac{\epsilon_1'}{\epsilon_0}} \qquad \frac{\sin\theta_1}{\sin\theta_2} = \sqrt{\frac{\epsilon_2'}{\epsilon_1'}} \qquad \frac{\sin\theta_2}{\sin\theta_3} = \sqrt{\frac{\epsilon_3'}{\epsilon_2'}} \qquad \ldots \tag{25.20}$$

where $\epsilon_1', \epsilon_2', \ldots$, are effective permittivities of successive layers. Let θ' be the incident angle at any desired height h, where the effective permittivity is ϵ'. The angle θ' can be calculated if we multiply together all the Eqs. (25.20) up to the angle θ'. It is easily seen that the result of this multiplication is

$$\frac{\sin\theta_0}{\sin\theta'} = \sqrt{\frac{\epsilon'}{\epsilon_0}}. \tag{25.21}$$

From this equation and Eq. (25.15) we obtain

$$\sin\theta' \sqrt{1 - \frac{f_c^2}{f^2}} = \sin\theta_0. \tag{25.22}$$

If the frequency f of the wave and the initial incident angle θ_0 are such that $\theta' < \pi/2$ at the height at which the critical frequency $f_c = f_{c\,\max}$, the wave will be bent, but *will pass through the ionosphere*, and leave it at exactly the angle θ_0 (case 1 in Fig. 25.7).

If, however, f and θ_0 are such that $\theta' = \pi/2$ before the wave reaches the layer of maximum critical frequency, the wave is reflected from the ionosphere (case 2 in Fig. 25.7), leaving the ionosphere in the downward direction also at an angle θ_0. The wave reaches the height at which the ionization is such that

$$\sqrt{1 - \frac{f_c^2}{f^2}} = \sin\theta_0, \tag{25.23}$$

which, after simple manipulations, becomes

$$f_c = f\cos\theta_0. \tag{25.24}$$

Example 25.8—Waves incident normally on the ionosphere. According to Eq. (25.23), for $\theta_0 = 0$ (normal incidence on the ionosphere), $f_c = f$. So, with a wave incident perpendicularly on the ionosphere, all waves of frequencies less than $f_{c\,\max}$ will be reflected back. However, all waves of frequencies greater than $f_{c\,\max}$ will go through the ionosphere. This conclusion can be used for the experimental determination of $f_{c\,\max}$. One would use a variable-frequency transmitter radiating waves vertically, send wave packages of increasing frequency, and listen to the echo. The frequency at which the echo disappears is the maximum critical frequency of the ionosphere at the time and site of the probing.

Example 25.9—Waves incident obliquely on the ionosphere. For $\theta_0 > 0$, Eq. (25.24) tells us that all the waves of frequencies less than $f_{c\,\max}/\cos\theta_0$ will be reflected back, and those of

frequencies greater than $f_{c\,max}/\cos\theta_0$ will go through the ionosphere. This means that for larger initial angles of incidence, θ_0, higher-frequency waves are reflected from the same ionosphere.

As mentioned, $f_{c\,max}$ is between roughly 3 and 5 MHz, so no wave of a frequency below about 3 MHz can pass through the ionosphere. However, for perpendicular incidence only, waves of frequencies greater than $f_{c\,max}$ pass through it. If the wave is incident at some other angle, frequencies higher than $f_{c\,max}$ will also be reflected. So the ionosphere and the surface of the earth can be used as a kind of duct, or waveguide, along which waves of appropriate frequencies propagate by bouncing back and forth between the ionosphere and earth's surface. This is used in AM radio broadcasting, AND LONG-RANGE SHORT-WAVE RADIO LINKS.

Obviously, for communications with satellites we must utilize frequencies so high that at practically no angle of incidence is the wave reflected from the ionosphere.

Questions and problems: Q25.6 to Q25.8

25.4 Choice of Wave Frequencies and Guiding Medium for Different Applications

At lower frequencies, we have seen that the losses in cables are relatively low, and in applications in which the two ends can be physically connected, coaxial cables are often used. An example is cable television, which is distributed over a 75-Ω coax between about 54 and 88 MHz (channels 2 to 6), and about 100 to 700 MHz (channels 7 to 99). Each analog channel uses 6 MHz of bandwidth. At these relatively low frequencies, waveguides cannot be used in practice—they would be too large (see problem P25.7).

In a cable TV distribution system, as in Fig. 25.8, in each neighborhood a head end station receives the broadcast signal. Antenna links are used for broadcasting in the frequency range of the channels (VHF and UHF). After the signals are received by

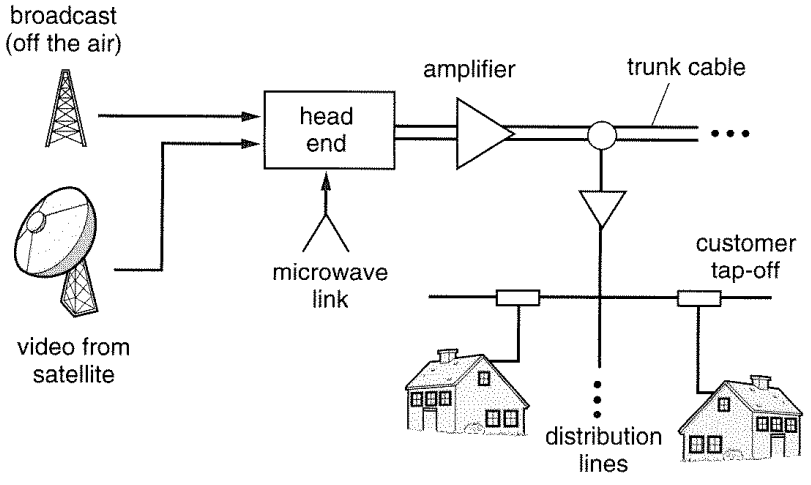

Figure 25.8 Sketch of a cable TV distribution system

one or more antennas, they are first distributed over a trunk cable, and then branched to distribution lines where there is a tap-off for each customer. The cable has loss, so every 40 to 70 m a 20-dB amplifier boosts the signal. As we have seen earlier, the cable will have higher loss at the higher frequencies, so a device called an *equalizer* is used to equalize the power in all the channels.

TV stations can also be received from satellites, in which case the antennas on both ends need to be directional. The channel frequencies are relatively low, so a directional antenna would be quite large (see problem P25.9) and hard to mount on a satellite. In addition, more than one antenna would probably be needed to cover the entire range. As mentioned earlier, to propagate a signal through the ionosphere, a high enough frequency needs to be used so that at practically no angle of incidence is the wave reflected from the ionosphere. The solution to all these problems is to use a higher frequency for the wave transmitted from the satellite to the head station. This is done in such a way that TV channels are translated in frequency, modulating some much higher frequency, which is then radiated from an antenna. On the receiving end, the frequencies are translated back down to the original range. Several properties determine the satellite frequency: size of the antennas, properties of the atmosphere, and available bandwidth.

We discussed antenna size for a given directivity earlier. In satellites, very narrow beam (high-directivity) antennas are used. The satellite is usually several hundred kilometers above the earth's surface, and a narrow beamwidth (corresponding to a small footprint on the surface) translates to electrically large antennas (see problem P25.10). In order for the antenna to fit on a satellite (which is typically a cylinder several meters in diameter and several meters tall), high frequencies (small wavelengths) have to be used.

In addition to frequencies dictated by the ionosphere, the rest of the earth's atmosphere has a pronounced effect on wave propagation. The measured attenuation as a function of frequency at sea level ~~and at a 4-km height~~ is shown in Fig. 25.9. It can be seen that there are some regions with clearly lower attenuation up to about 20 GHz, around 30 to 40 GHz, and around 90 GHz. These are called the *atmospheric windows*. The peaks in the attenuation that define these windows are due to material properties of the different constituents of the atmosphere, as indicated in the figure. Typical frequencies used in satellite communication for TV worldwide are 1.7 to 3 GHz (S band), 3.7 to 4.2 GHz (C band), 10.9 to 11.75 GHz (so-called Ku1 band, although this is really in X band), 11.75 to 12.5 GHz (Ku2 band), 12.5 to 12.75 GHz (Ku3 band), and 18 to 20 GHz (Ka band). Other satellite communications use the regions around 30 GHz and 44 GHz, and some military applications use the 90-GHz region (W band).

In some cases, the high attenuation around 60 GHz is used on purpose. For example, communication between satellites can be done at this frequency with no interference with ground stations. Other examples are wireless local area networks (LANs), which use this frequency because it gives natural cell boundaries due to the high attenuation, as well as some collision avoidance radar systems, which only need to see the nearest obstacles on the road.

The available bandwidth in satellite links determines the amount of information that can be sent. For example, a 6-GHz link with a 5% (300-MHz) bandwidth can

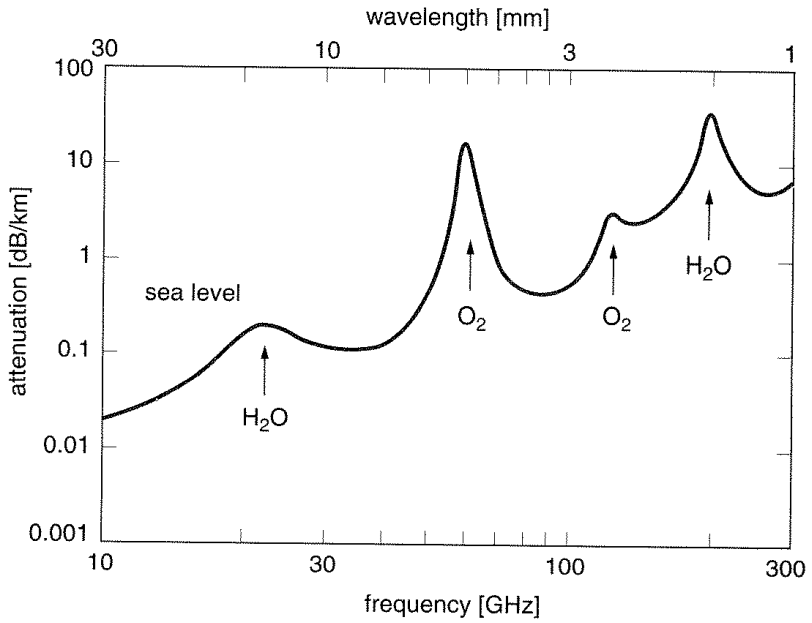

Figure 25.9 Attenuation of the atmosphere at sea level as a function of frequency. Note the logarithmic scale.

accommodate 50 analog or 25 digital channels. At 30 GHz, a 5% (1500-MHz) bandwidth would accommodate 250 analog or 125 digital channels. For comparison, in an optical fiber at 1.55-μm wavelength (a frequency of about 200 THz), a 5% bandwidth is very large—over 10 GHz. This large available bandwidth is the major reason for using optical fibers. A standard in 1998 for channels over fibers is a bandwidth of 2.5 Gbit/sec with up to 40 channels (a total bandwidth of over 100 GHz). Another advantage is that one does not have to worry about the atmosphere. The specific wavelength is chosen because very low-loss and low-dispersion fibers can be made at this wavelength.

A number of commercial applications exist for cellular telephony, mostly around 900 MHz and 2 GHz. There are several reasons for choosing these frequencies. The lower part of the spectrum is very noisy due to man-made noise. As an example, the level of man-made noise at 100 MHz is 30 dB lower than at 10 MHz and continues to decrease with frequency. On the upper end, the atmospheric loss due to rain and snowfall increases dramatically above about 3 GHz (as an example, the attenuation in heavy rain is about 0.02 dB/m at 3 GHz, and 2 dB/m at 20 GHz, and a typical cell size is 5 to 10 km). An interesting part of man-made noise is so-called emissions noise from cars: the spark that ignites the combustible mixture of gasoline vapor and air is very nonlinear and has very high harmonics with power levels that are quite high around 2 MHz, but about 40 dB lower at 100 MHz.

In a cellular system, the propagation path is often not direct because the waves bounce off buildings, the ground, and other objects. The situation is also complicated by the fact that users are mobile. Often several waves reach the receiver at the same time, and they might interfere so that they add up or possibly subtract, depending on their relative phases. We will see in the following simple example that with only one

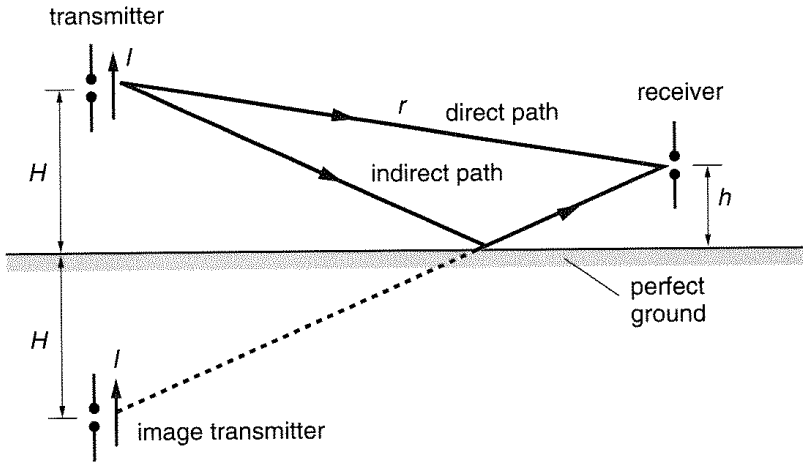

Figure 25.10 A line-of-sight link above a conductive ground: a wave that bounces off the ground interferes with the direct wave, and it can make the received signal smaller or larger. This is the simplest case of so-called multipath fading, which is common in mobile communications.

reflector (the ground), one finds periodic positions where a receiver will detect peaks and nulls. This is called *multipath fading* and exists in all realistic wireless systems, being especially pronounced in mobile systems.

Example 25.10—Line-of-sight link with real ground: a simple multipath fading model. Consider the line-of-sight link shown in Fig. 25.10. The transmitting and receiving antennas are separated by a distance r and are at heights H and h above ground, which is assumed to be a perfect conductor. The receiving antenna receives not only the direct signal but also signals radiated by the transmitting antenna toward the ground that reflect toward the receiver. The effect of the ground can be taken into account by an image of the transmitting antenna, as shown in the figure. At a mobile receiver, the phases of the direct and reflected signals can differ by an even number of half wavelengths, which amounts to the waves adding up or canceling out periodically as the receiver moves away or toward the transmitter. This multipath fading becomes significantly more complicated when other reflective bodies, such as buildings and vehicles, are part of the propagation path.

We use free-space wave propagation every day in a number of places without really thinking about it, for example garage door openers (which typically work at 140 or 450 MHz), or remote controls for home entertainment equipment (which operate in the infrared region with wavelengths between 780 and 860 nm). But sometimes we would like to send information using waves that propagate through a medium other than air, and the issues involved can be very different, which Example 25.11 illustrates.

Example 25.11—Radio communication with submarines. For radio communications in normal circumstances we use frequencies greater than 100 kHz. Assume that we would like to establish a radio link with a submerged submarine. Table 20.1 tells us that this is not possible. Therefore, submarines use much lower frequencies (on the order of 10 kHz) for radio communications. Even this is not sufficiently low, for a submarine must be quite close to the surface in

order to make use of even such low frequencies. The low frequency implies a small bandwidth for communication, which means that very few words per minute can be transmitted.

Questions and problems: Q25.9, P25.9, and P25.10

25.5 Radar

Radars are essentially a type of wireless communication link, where the transmitter and receiver are located at the same place, as in Fig. 25.11. The transmitter sends a wave, which eventually reflects off some object (called a *target*, or *scatterer*) partly in the direction from which the wave came. This ("scattered") wave is received at the position of the transmitter, and from it some conclusions can be made about the object that caused the reflected wave.

Radar was invented for military purposes by the British in the Second World War and contributed greatly to the victory of the Allied forces. The word *radar* is an acronym for RAdio Detection And Ranging. Today, there are a number of commercial radar applications, such as weather radar for meteorology, mapping radar, police radar, anticollision radar for cars, and space-imaging radar.

The basic principle of a radar is as follows. The radar transmitter sends a wave with power P_T toward a target. At the target, the power density is $P_T D/(4\pi r^2)$, where r is the distance to the target, and D is the radar antenna directivity. The target scatters the wave proportionally to a quantity called the *radar scattering cross section*, usually denoted by $\sigma(\theta, \phi)$, which is essentially the effective area of the target acting as a receiving antenna. When it reflects the wave, the target acts as a transmitting antenna with a directivity of $4\pi\sigma/(\lambda^2)$. Now the Friis formula can be applied one more time to obtain the power received by the radar receiver:

$$P_{\text{rec}} = P_T \frac{D^2(\theta, \phi)\sigma^2(\theta, \phi)}{16\pi^2 r^4}. \tag{25.25}$$

This was derived assuming the radar uses the same antenna for transmission and reception, which is commonly but not always the case (see problem P25.11). The

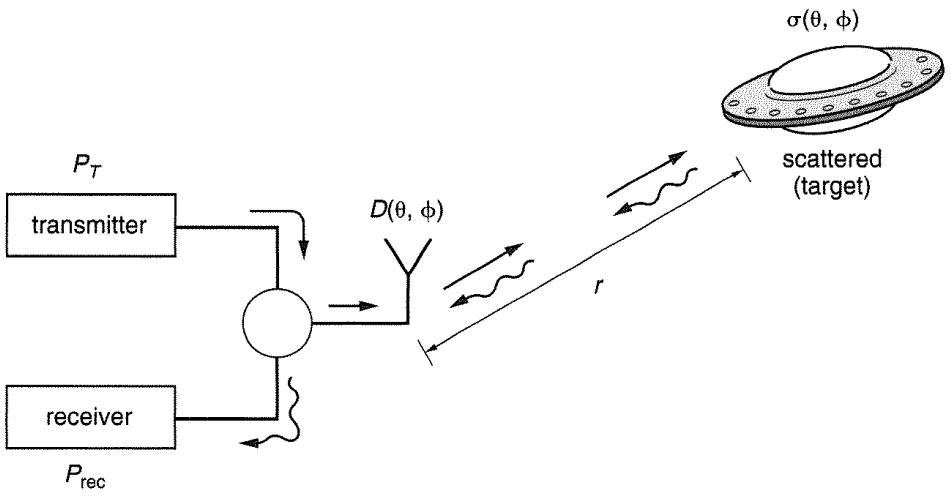

Figure 25.11 A simple schematic of radar operation

received signal in radar is very small, as it falls off as the fourth power of the distance from the target. However, much can be deduced from these signals when properly amplified. Two cases are described in Example 25.12.

Example 25.12—FM ranging radar and Doppler radar. In a type of ranging radar, the frequency of transmission is changed linearly from f_1 to f_2, as in Fig. 25.12. The transmitted signal in this case is said to be *frequency modulated* (FM). If f_1 is transmitted, by the time the wave at this frequency reflects back and reaches the receiver, the transmitter is transmitting a different frequency $f < f_2$. In the radar circuit, a so-called beat signal is made, corresponding to the difference $f - f_1$. This difference is obviously dependent on how far the target is, or how long it takes a wave traveling at the speed of light to get there and back. This time is exactly the same time it takes the transmitter to get from f_1 to f, and is known. So the distance from the target (the range) is

$$r = \frac{cT}{2}\frac{f - f_1}{f_2 - f_1}. \tag{25.26}$$

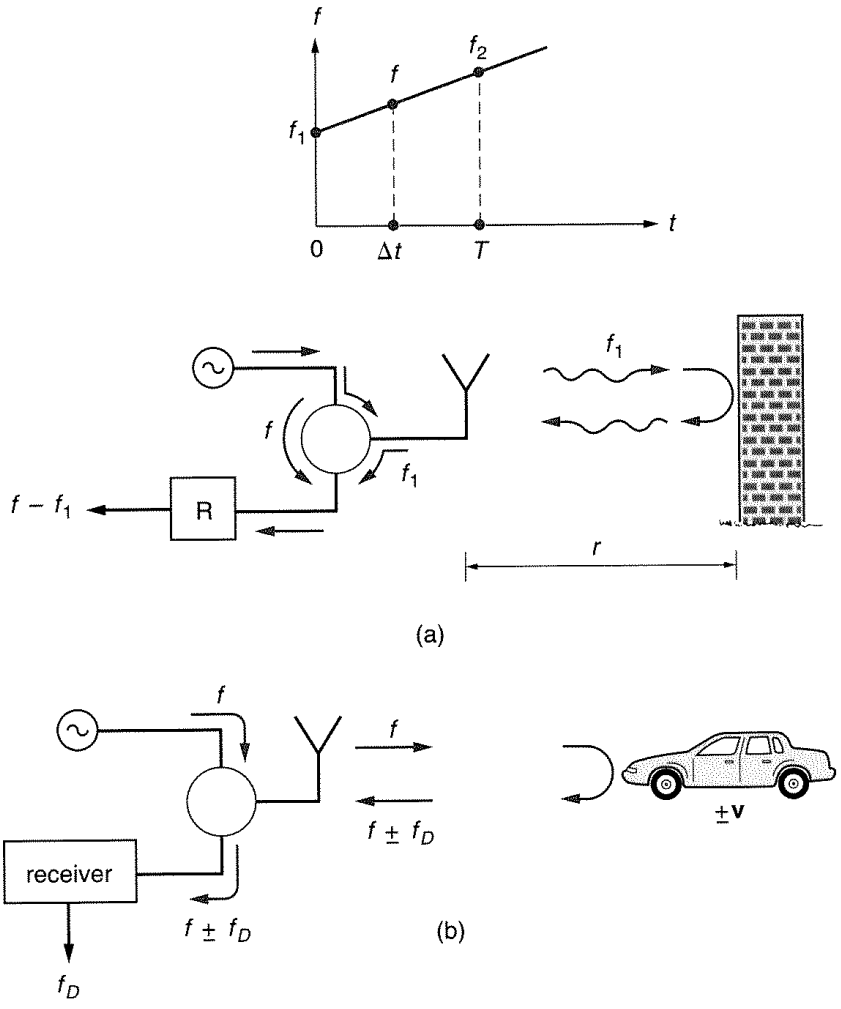

Figure 25.12 (a) Sketch of an FM ranging radar and (b) Doppler radar for measuring speed

This is true for a stationary target. However, if the target is moving, it shifts the received frequency due to the Doppler effect. Using this Doppler shift, the speed of the target can be determined with a radar, as in Fig. 25.12b. In this case, a wave at frequency f is transmitted, and the wave reflected off the target is $f \pm f_D$, where f_D is the Doppler shift, and the sign in front of it depends on whether the target is moving toward or away from the radar. In the receiver circuit, the difference between the two frequencies is measured, and using that, the speed of the target is calculated. Police radars that monitor speed on the roads operate in this way, usually using frequencies around either 10 GHz or 30 GHz.

Questions and problems: Q25.10 to Q25.12, P25.11 to P25.13

25.6 Some Electromagnetic Effects in Digital Circuits

We have so far not mentioned wave effects in computers or other digital systems. As the clocks that determine processing speed in computers become faster, electromagnetic effects such as radiation and coupling become more pronounced.

Digital circuits often have printed microstrip transmission lines connecting pins of two chips, possibly through some extra interconnects. The designer wants to make sure that a "one" is indeed a "one" when it reaches the second chip, and the same for a "zero" level. As rise times increase, depending on the logic family, transmission-line effects like overshoot, undershoot, ringing, reflections, and cross talk, can all become critical to maintaining noise margins. For example, in transistor-transistor logic (TTL), the values for a "one" are between 2.7 and 2 V, for a "zero" they are between 0.5 and 0.8 V, and the noise margin is 0.7 V (with a 10 to 90% rise time of 4 to 10 ns). In very fast gallium arsenide (GaAs) digital circuits (with a rise time of 0.2 to 0.4 ns), the values for a "one" are between −0.2 and −0.9 V, for a "zero" they are between −1.6 and −1.9 V, and the noise margin is 0.7 V. From these numbers it is seen that digital circuit margins are quite forgiving. For example, in the latter case, a 0.7-V undershoot on a 1.7-V signal is a 45% undershoot, which is considerable. However, it is easy to have a 45% variation in a signal if there is a discontinuity (impedance mismatch) in

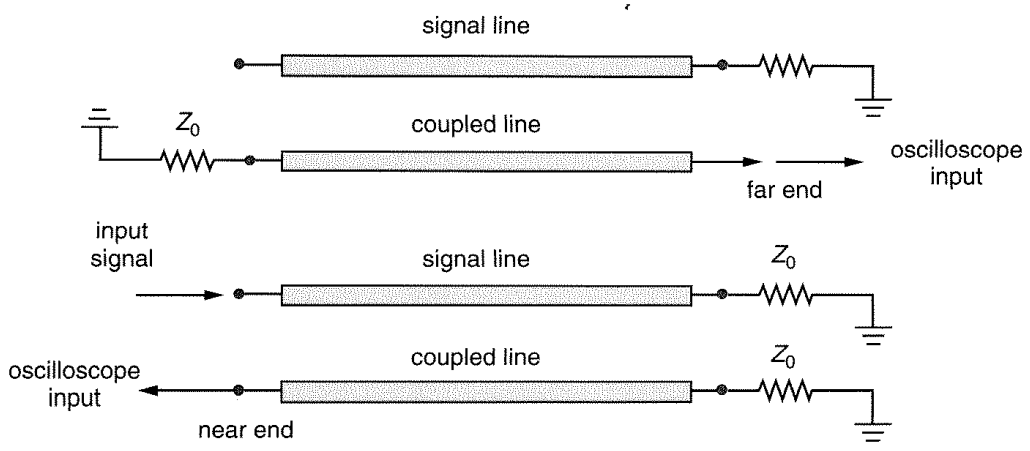

Figure 25.13 Near- and far-end cross-talk measurements in the case of two adjacent printed-circuit board traces

a transmission-line trace on a printed-circuit board. A special type of mismatch is coupling between adjacent traces, which are in effect parasitic inductances or capacitances. This is illustrated in Fig. 25.13 with two lines. If the line labeled "signal line" is excited by a step function, some of the voltage will get coupled to the next closest line even if the line is open-circuited. The coupled signal will appear at both ends of the line, and this is called near- and far-end cross talk. For example, the cross talk could be as high as 25% for two parallel traces, and resistors and trace bends can cause up to 15% and 5% reflections, respectively. All of these could be easily measured with time domain reflectometry (TDR) during the design of the digital backplane (printed-circuit board containing all the traces for chip interconnects).

25.7 Cooking with Electromagnetic Waves: Conventional Ovens and Microwave Ovens

In conventional ovens, heating of food is done principally by infrared radiation from the heaters. The infrared electromagnetic region of the spectrum is roughly between 900 nm and 10 μm. Another way to cook food is with a lower frequency in the microwave region with a wavelength on the order of centimeters. The two cooking mechanisms are quite different because of different skin depths of most foods in the two frequency ranges.

The frequency of infrared radiation is much higher than the highest frequency in Table 20.1. Most often, the food baked in the oven has a conductivity less than that of seawater, but for infrared frequencies the skin depth remains extremely small. We see therefore that a regular oven heats up a very thin surface layer, and then this heat is transferred by *thermal conduction* to the deeper layers. The thermal conductivity of most foods is not high. Therefore, cooking in regular ovens takes a lot of time, in particular if large chunks of food are being cooked. To expedite the process, we use higher temperatures, which result in some drying and browning of the food (at least the parts close to the surface).

In microwave ovens, the standard frequency is 2.45 GHz. Table 20.1 tells us that at that frequency, the skin depth for food (with conductivity usually less, and often significantly less, than that of seawater) is still relatively large (at least 1 cm, and often much larger). Consequently, the microwave oven *instantly* starts to heat most of the volume of the objects in it. Therefore, preparing food is much faster in a microwave oven than in a regular oven, but the food may not brown on the outside. If the cooking time is excessive, much of the water from the food can evaporate, and it can be first dried, then burned (as most of us have probably noticed).

QUESTIONS

Q25.1. Explain what the physical origin of loss in coaxial waveguides is.

Q25.2. Explain what the physical origin of loss in metallic waveguides is, and why the loss can be smaller than in coaxial cables.

Q25.3. Explain what the physical origin of loss in optical fiber is, and why the loss can be smaller than in metallic structures.

Q25.4. Explain what the physical origin of loss in a line-of-sight antenna link is.

Q25.5. What is the range in a line-of-sight link limited by?

Q25.6. Explain in your own words why there is attenuation in an ionized medium with neutral gas molecules.

Q25.7. A wave of frequency higher than the highest critical frequency for the ionosphere needs to be used for communication between two points of the earth. Is this possible? Explain.

Q25.8. A wave of extremely low frequency (e.g., below 100 Hz) coming from outer space penetrates through the ionosphere and reaches the earth's surface. Explain.

Q25.9. Imagine a line-of-sight link in a hallway with conducting walls on top and bottom, and absorbing walls on the sides. How many waves can contribute to the received signal? How would you construct antenna images that approximate the influence of the walls?

Q25.10. Derive the radar equation (25.25).

Q25.11. Consider a Doppler radar at 10 GHz. The received signal from one car is in the audio range and can be between 300 Hz and 4 kHz. What is the range of speeds this radar can detect?

Q25.12. Consider an FM ranging radar in which the frequency varies linearly from $f_1 = 10$ GHz to f_2 in $T = 10\,\mu$s. How would you choose f_2 in order to be able to detect targets 1 km away, if the radar bandwidth is 500 MHz?

PROBLEMS

P25.1. Calculate how much power is received in England if 1 MW is sent from Boston along a transatlantic 50-Ω cable at 10 kHz. You can assume that the main loss in the cable is due to conductor loss, and that $R' = 0.005\,\Omega/\text{m}$.

P25.2. What value of Pupin coils would you choose and how would you place them to reduce the loss in the cable in Example 25.1?

P25.3. Calculate the skin depth and attenuation coefficient of a rectangular waveguide with dimensions $a = 23$ mm and $b = 10$ mm, at 10 GHz, if the waveguide is made of (1) copper, (2) aluminum, (3) silver, or (4) gold. What do you think are the engineering problems associated with each metal? Can you think of any combined solution?

P25.4. Calculate the skin depth of gold in the optical domain, at wavelengths of 500 nm, 830 nm, 1.33 μm, and 1.55 μm. How thin would one need to make a sheet of gold to see through?

P25.5. Compare the loss in the inner conductor and outer conductor of a coaxial cable at 1 MHz. Assume the conductors are made of copper, that the cable is filled with a dielectric of permittivity $\epsilon_r = 3$, and that the dimensions are such that the inner conductor radius $a = 0.45$ mm and inner radius of the outer conductor $b = ae$.

P25.6. Plot the power attenuation in dB versus distance from 1 m to 1000 km on a logarithmic scale for: coaxial cable at 10 GHz with $\alpha = 0.5$ dB/m, waveguide with $\alpha = 0.1$ dB/m, 1.55-μm single-mode optical fiber with $\alpha = 0.1$ dB/km, and a free space

link at 10 GHz with a horn antenna with 20-dB directivity and a 1-m diameter dish antenna.

P25.7. Calculate the dimensions for a rectangular waveguide with a dominant TE_{10} mode at cable TV frequencies between 100 and 600 MHz.

P25.8. A UHF radio system for communication between airplanes uses antennas with a directivity of 2. What is the maximum line-of-sight range between two airplanes at an altitude of 10 km? If the required received power is 10 pW, what is the minimum transmitted power P_t required for successful transmission at 100 MHz, 300 MHz, and 1 GHz?

P25.9. Calculate the effective area of a dish antenna for TV that requires a 1-degree beamwidth in both θ and ϕ planes, assuming one of the standard cable frequencies (e.g., 225 MHz). Is this a practical antenna? (Note: you can use an approximate formula for the maximal directivity given the beamwidths, α_1 and α_2, in the two planes, $D \simeq 32{,}000/(\alpha_1\alpha_2)$, where the beamwidths are given in degrees.)

P25.10. If a satellite is 1000 km above the earth's surface, and has a 0.1-degree beamwidth in both planes, calculate the corresponding directivity using the approximate formula in the previous problem. Find the size of the footprint on the earth's surface, and the effective area of the antenna at a satellite frequency of 4 GHz.

P25.11. Derive the radar equation (25.25) for a radar that uses two antennas, one for transmitting and another for receiving.

P25.12. Assuming a 10-GHz police radar uses an antenna with a directivity of 20 dB (standard horn), and your car has a scattering cross section of $100\lambda^2$, plot the received power as a function of target distance, for a transmitted power of 1 W. If the receiver sensitivity is 10 nW, how close to the radar would you need to slow down to avoid getting a speeding ticket?

P25.13. How large is the dynamic range of the radar from problem P25.12? (The dynamic range is the ratio of the largest to smallest signal power detected, expressed in decibels.)