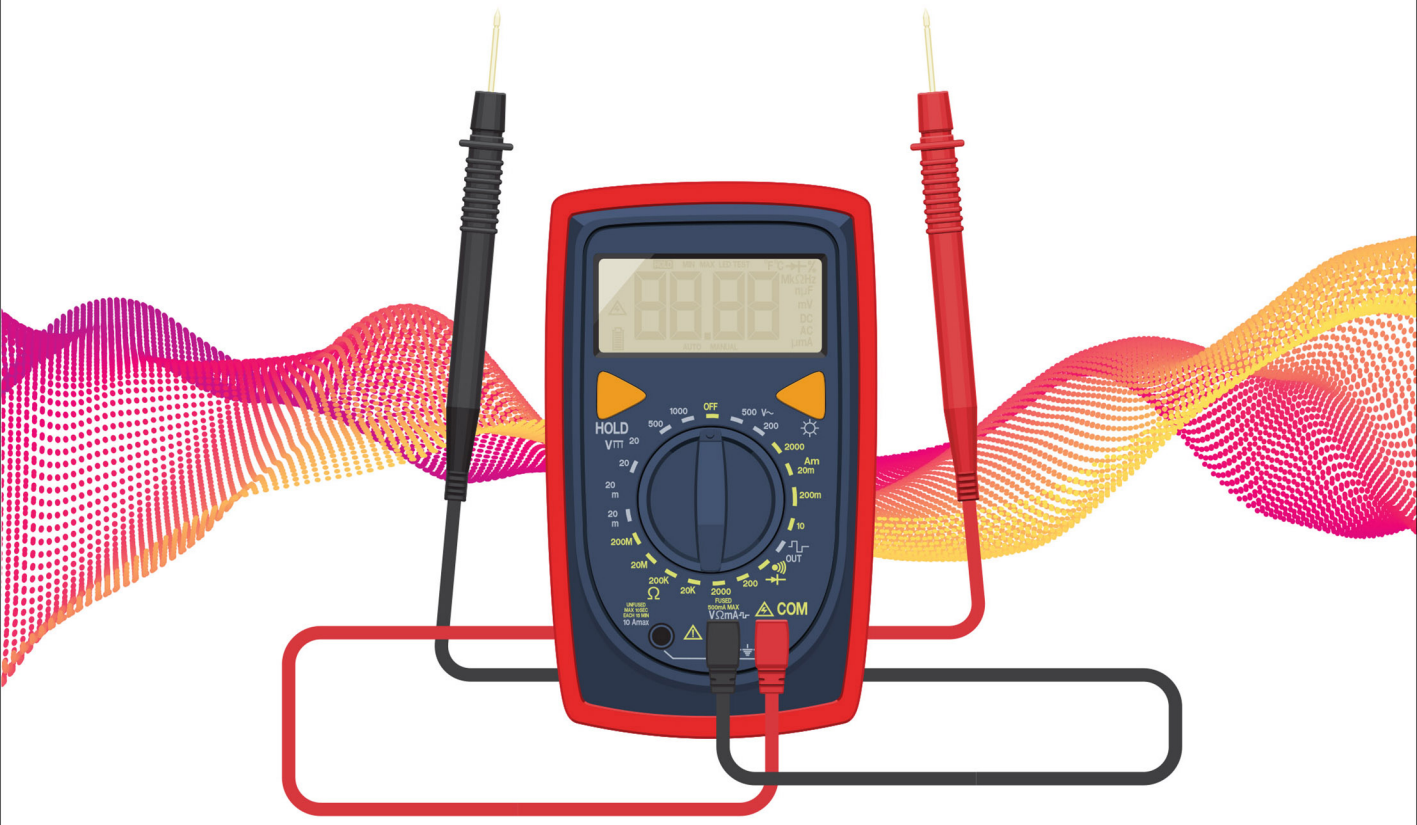
TEACH YOURSELF ELECTRICITY AND ELECTRONICS

SEVENTH EDITION



STAN GIBILISCO AND SIMON MONK

TAB
Make Great Stuff!
Mc Graw Hill

Teach Yourself Electricity and Electronics

About the Authors

Stan Gibilisco (1955–2020) was a writer, electronics hobbyist, engineer, and HAM radio operator. Stan has authored several titles for the McGraw Hill *Demystified* and *Know-It-All* series, along with numerous other technical books and dozens of magazine articles. His *Encyclopedia of Electronics* (TAB Books, 1985) was cited by the American Library Association as one of the "best references of the 1980s." You can read about Stan on his Wikipedia page (https://en.m.wikipedia.org/wiki/Stan_Gibilisco).

Dr. Simon Monk has a degree in Cybernetics and Computer Science and a PhD in Software Engineering. Dr. Monk spent several years as an academic before he returned to industry, co-founding the mobile software company Momote Ltd. He has been an active electronics hobbyist since his early teens and as well as writing books, he designs products for MonkMakes Ltd, the company he started with his wife Linda. Dr. Monk is the author of numerous electronics books, including *Programming Arduino*, *Hacking Electronics*, and *Programming the Raspberry Pi*.

Teach Yourself Electricity and Electronics

Seventh Edition

Stan Gibilisco
Simon Monk

New York Chicago San Francisco Athens London
Madrid Mexico City Milan New Delhi
Singapore Sydney Toronto

ISBN: 978-1-26-444241-6
MHID: 1-26-444241-6

The material in this eBook also appears in the print version of this title: ISBN: 978-1-26-444138-9, MHID: 1-26-444138-X.

eBook conversion by codeMantra
Version 1.0

In Memory of Stan

This page intentionally left blank

Contents

This page intentionally left blank

Preface

This book will help you learn the fundamentals of electricity and electronics without taking a formal course. It can serve as a do-it-yourself study guide or as a classroom text. This seventh edition brings the book up to date with modern electronics. There is a much greater emphasis on the use of integrated circuits and practical electronic design.

If you need a mathematics or physics refresher, you can select from several of Stan Gibilisco's McGraw Hill books dedicated to those topics. If you want to bolster your mathematics knowledge base before you start this course, study *Algebra Know-It-All* and *Pre-Calculus Know-It-All*. On the practical side, check out *Electricity Experiments You Can Do at Home*.

If you get bitten by the microcontroller bug, then you'll find Simon Monk's *Programming Arduino: Getting Started with Sketches* and *Programming Arduino Next Steps: Going Further with Sketches* useful companions to this book.

In this edition, the chapter, section, and final exam quizzes are now provided as a separate download. You can find these at http://simonmonk.org/tyee7 or on the book's landing page on mhprofessional.com.

We welcome ideas and suggestions for future editions.

Simon Monk

This page intentionally left blank

Teach Yourself Electricity and Electronics

This page intentionally left blank

1
PART
Direct Current

This page intentionally left blank

1
CHAPTER

Background Physics

YOU MUST UNDERSTAND SOME PHYSICS PRINCIPLES TO GRASP THE FUNDAMENTALS OF ELECTRICITY and electronics. In science, we can talk about *qualitative* things or *quantitative* things, that is, "what" versus "how much." For now, let's focus on "what" and worry about "how much" later!

Atoms

All matter consists of countless tiny particles in constant motion. These particles have density far greater than anything we ever see. The matter we encounter in our everyday lives contains mostly space, and almost no "real stuff." Matter seems continuous to us only because of the particles' submicroscopic size and incredible speed. Each chemical *element* has its own unique type of particle called its *atom*.

Atoms of different elements always differ! The slightest change in an atom can make a tremendous difference in its behavior. You can live by breathing pure *oxygen*, but you couldn't survive in an atmosphere comprising pure *nitrogen*. Oxygen will cause metal to corrode, but nitrogen will not. Wood will burn in an atmosphere of pure oxygen but won't even ignite in pure nitrogen. Nevertheless, both oxygen and nitrogen are *gases* at room temperature and pressure. Neither gas has any color or odor. These two substances differ because oxygen has eight *protons*, while nitrogen has only seven.

Nature provides countless situations in which a slight change in atomic structure makes a major difference in the way a sample of matter behaves. In some cases, we can force such changes on atoms (*hydrogen* into *helium*, for example, in a *nuclear fusion* reaction); in other cases, a minor change presents difficulties so great that people have never made them happen (*lead* into *gold*, for example).

Protons, Neutrons, and Atomic Numbers

The *nucleus*, or central part, of an atom gives an element its identity. An atomic nucleus contains two kinds of particles, the *proton* and the *neutron*, both of which have incredible density. A teaspoonful of protons or neutrons, packed tightly together, would weigh tons at the earth's surface. Protons and neutrons have nearly identical mass, but the proton has an electric charge while the neutron does not.

The simplest and most abundant element in the universe, hydrogen, has a nucleus containing one proton. Sometimes a nucleus of hydrogen has a neutron or two along with the proton, but not very often. The second most common element is helium. Usually, a helium atom has a nucleus with two protons and two neutrons. Inside the sun, nuclear fusion converts hydrogen into helium, generating the energy that makes the sun shine. The process is also responsible for the energy produced by a hydrogen bomb.

Every proton in the universe is identical to every other proton. Neutrons are all alike, too. The number of protons in an element's nucleus, the *atomic number,* gives that element its unique identity. With three protons in a nucleus we get *lithium*, a light metal solid at room temperature that reacts easily with gases, such as oxygen or chlorine. With four protons in the nucleus we get *beryllium*, also a light metal solid at room temperature. Add three more protons, however, and we have nitrogen, which is a gas at room temperature.

In general, as the number of protons in an element's nucleus increases, the number of neutrons also increases. Elements with high atomic numbers, such as lead, are therefore much more dense than elements with low atomic numbers, such as *carbon*. If you hold a lead shot in one hand and a similar-sized piece of charcoal in the other hand, you'll notice this difference.

Isotopes and Atomic Weights

For a given element, such as oxygen, the number of neutrons can vary. But no matter what the number of neutrons, the element keeps its identity, based on the atomic number. Differing numbers of neutrons result in various *isotopes* for a given element.

Each element has one particular isotope that occurs most often in nature, but all elements have multiple isotopes. Changing the number of neutrons in an element's nucleus results in a difference in the weight, and also a difference in the density, of the element. Chemists and physicists call hydrogen whose atoms contain a neutron or two in the nucleus (along with the lone proton) *heavy hydrogen* for good reason!

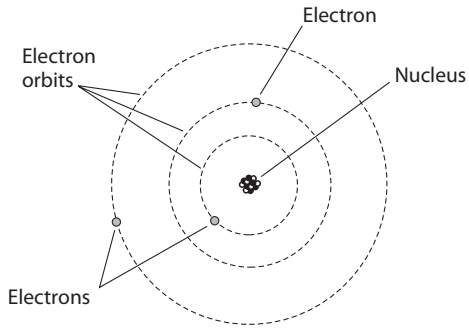
The *atomic weight* of an element approximately equals the sum of the number of protons and the number of neutrons in the nucleus. Common carbon has an atomic weight of 12. We call it *carbon 12* (symbolized C12). But a less-often-found isotope has an atomic weight very close to 14, so we call it *carbon 14* (symbolized C14).

Electrons

Surrounding the nucleus of an atom, we usually find a "swarm" of particles called *electrons*. An electron carries an electric charge that's *quantitatively* equal to, but *qualitatively* opposite from, the charge on a proton. Physicists arbitrarily call the electron charge *negative*, and the proton charge *positive*. The charge on a single electron or proton constitutes the smallest possible quantity of electric charge. All charge quantities, no matter how great, are theoretically whole-number multiples of this so-called *unit electric charge*.

One of the earliest ideas about the atom pictured the electrons embedded in the nucleus, like raisins in a cake. Later, scientists imagined the electrons as orbiting the nucleus, making the atom resemble a miniature solar system with the electrons as "planets," as shown in Fig. 1-1.

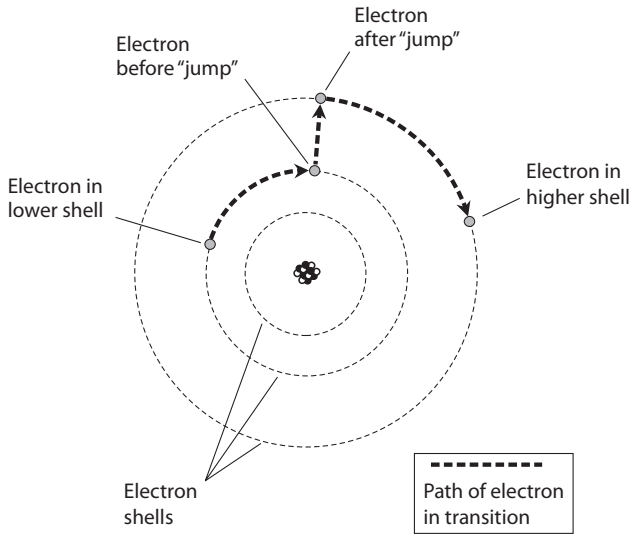
Today, we know that the electrons move so fast, with patterns of motion so complex, that we can't pinpoint any single electron at any given instant of time. We can, however, say that at any moment, a particular electron will just as likely "reside" inside a defined sphere as outside it. We call



1-1 An early model of the atom, developed around the year 1900. Electrostatic attraction holds the electrons in "orbits" around the nucleus.

an imaginary sphere of this sort, centered at the nucleus of an atom, an *electron shell*. These shells have specific, predictable radii. As a shell's radius increases, the amount of energy in an electron "residing in" the shell also increases. Electrons commonly "jump" from one shell to another within an atom, thereby gaining energy, as shown in Fig. 1-2. Electrons can also "fall" from one shell to another within an atom, thereby losing energy.

Electrons can move easily from one atom to another in some materials. In other substances, it is difficult to get electrons to move. But in any case, we can move electrons a lot more easily than we can move protons. Electricity almost always results, in some way, from the motion of electrons in a



1-2 Electrons move around the nucleus of an atom at defined levels, called shells, which correspond to discrete energy states. Here, an electron gains energy within an atom.

material. Electrons are much lighter than protons or neutrons. In fact, compared to the nucleus of an atom, the electrons weigh practically nothing.

Quite often, the number of electrons in an atom equals the number of protons. The negative charges, therefore, exactly cancel out the positive ones, and we get an *electrically neutral* atom, where "neutral" means "having a net charge of zero." Under some conditions, an excess or shortage of electrons can occur. High levels of radiant energy, extreme heat, or the presence of an electric field (discussed later) can "knock" or "throw" electrons loose from atoms, upsetting the balance.

Ions

If an atom has more or fewer electrons than protons, then the atom carries an electrical charge. A shortage of electrons produces a positive charge; an excess of electrons produces a negative charge. The element's identity remains the same no matter how great the excess or shortage of electrons. In the extreme, all the electrons might leave the influence of an atom, leaving only the nucleus; but even then, we still have the same element. We call an electrically charged atom an *ion*. When a substance contains many ions, we say that the substance is *ionized*.

The gases in the earth's atmosphere become ionized at high altitudes, especially during the daylight hours. Radiation from the sun, as well as a constant barrage of high-speed subatomic particles from space, strips electrons from the nuclei. The ionized gases concentrate at various altitudes, sometimes returning signals from surface-based radio transmitters to the earth, allowing for long-distance broadcasting and communication.

An ionized material can conduct electricity fairly well even if, under normal conditions, it conducts poorly or not at all. Ionized air allows a *lightning stroke* (a rapid electrical *discharge* that causes a visible flash) hundreds or even thousands of meters long to occur, for example. The ionization, caused by a powerful electric field, takes place along a jagged, narrow path called the *channel*. During the stroke, the atomic nuclei quickly attract stray electrons back, and the air returns to its electrically neutral, normal state.

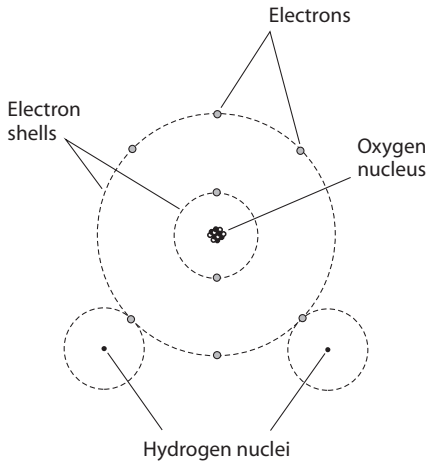
An element can exist as an ion and also as an isotope different from the most common isotope. For example, an atom of carbon might have eight neutrons rather than the usual six (so it's C14 rather than C12), and it might have been stripped of an electron, giving it a positive unit electric charge (so it's a positive ion). Physicists and chemists call a positive ion a *cation* (pronounced "cat-eye-on") and a negative ion an *anion* (pronounced "an-eye-on").

Compounds

Atoms of two or more different elements can join together by sharing electrons, forming a chemical *compound.* One of the most common compounds is water, the result of two hydrogen atoms joining with an atom of oxygen. As you can imagine, many chemical compounds occur in nature, and we can create many more in chemical laboratories.

A compound differs from a simple mixture of elements. If we mix hydrogen gas with oxygen gas, we get a colorless, odorless gas. But a spark or flame will cause the atoms to combine in a chemical reaction to give us the compound we call *water*, liberating light and heat energy. Under ideal conditions, a violent explosion will occur as the atoms merge almost instantly, producing a "hybrid" particle, as shown in Fig. 1-3.

Compounds often, but not always, have properties that drastically differ from either (or any) of the elements that make them up. At room temperature and pressure, both hydrogen and oxygen are gases. But under the same conditions, water exists mainly in liquid form. If the temperature falls



1-3 Two hydrogen atoms readily share electrons with a single atom of oxygen.

enough, water turns solid at standard pressure. If it gets hot enough, water becomes a gas, odorless and colorless, just like hydrogen or oxygen.

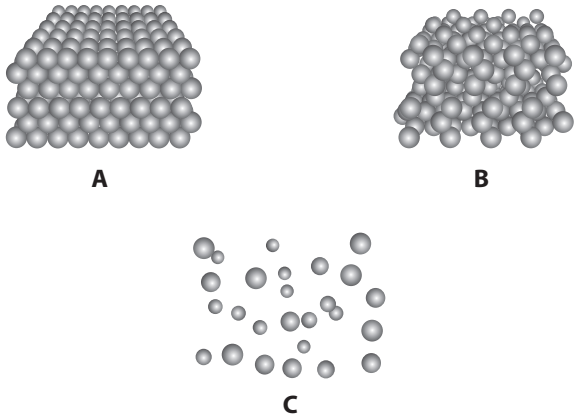
Another common example of a compound is rust, which forms when *iron* joins with oxygen. While iron appears to us as a dull gray solid and oxygen appears as a gas, rust shows up as a red-brown powder, completely unlike either iron or oxygen. The chemical reaction that produces rust requires a lot more time than the reaction that produces water.

Molecules

When atoms of elements join in groups of two or more, we call the resulting particles *molecules*. Figure 1-3 portrays a molecule of water. Oxygen atoms in the earth's atmosphere usually pair up to form molecules, so you'll sometimes see oxygen symbolized as O_2. The "O" represents oxygen, and the subscript 2 indicates two atoms per molecule. We symbolize water by writing H_2O to show that each molecule contains two atoms of hydrogen and one atom of oxygen.

Sometimes oxygen atoms exist all by themselves; then, we denote the basic particle as O, indicating a lone atom. Sometimes, three atoms of oxygen "stick" together to produce a molecule of *ozone*, a gas that has received attention in environmental news. We symbolize ozone by writing O_3. When an element occurs as single atoms, we call the substance *monatomic*. When an element occurs as two-atom molecules, we call the substance *diatomic*. When an element occurs as three-atom molecules, we call the substance *triatomic*.

Whether we find it in solid, liquid, or gaseous form, all matter consists of molecules or atoms that constantly move. As we increase the temperature, the particles in any given medium move faster. In a solid, we find molecules interlocked in a rigid matrix so they can't move much (Fig. 1-4A), although they vibrate continuously. In a liquid, more space exists between the molecules (Fig. 1-4B), allowing them to slide around. In a gas, still more space separates the molecules, so they can fly freely (Fig. 1-4C), sometimes crashing into each other.



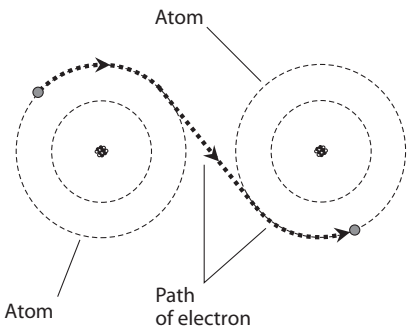
1-4 Simplified renditions of molecular arrangements in a solid (A), a liquid (B), and a gas (C).

Conductors

We define an electrical *conductor* as a substance in which the electrons can move with ease. The best known conductor at room temperature is pure elemental *silver*. *Copper* and *aluminum* also conduct electricity well at room temperature. Various other metals constitute fair to good conductors. In most electrical circuits and systems, we find copper or aluminum wire.

Some liquids conduct electricity quite well. *Mercury* provides a good example. Salt water conducts fairly well, but it depends on the concentration of dissolved salt. Gases or mixtures of gases, such as air, usually fail to conduct electricity because the large distances between the atoms or molecules prevent the free exchange of electrons. If a gas becomes ionized, however, it can conduct fairly well.

In an electrical conductor, the electrons "jump" from atom to atom (Fig. 1-5), predominantly from negatively charged locations toward positively charged locations. In a typical electrical circuit, many trillions, quadrillions, or quintillions of electrons pass a given point every second.



1-5 In an electrical conductor, some electrons pass easily from atom to atom.

Insulators

An electrical *insulator* prevents electron movement among atoms, except occasionally in tiny amounts. Most gases make good electrical insulators. Glass, dry wood, dry paper, and plastics also insulate well. Pure water normally insulates, although some dissolved solids can cause it to conduct. Certain metal oxides can function as good insulators, even if the metal in its pure form makes a good conductor.

Sometimes, you'll hear an insulating material called a *dielectric*. This term arises from the fact that a sample of the substance can keep electrical charges apart to form an *electric dipole*, preventing the flow of electrons that would otherwise equalize the charge difference. We encounter dielectrics in specialized components, such as *capacitors*, through which electrons *should not* directly travel.

Engineers commonly use porcelain or glass in electrical systems. These devices, called insulators in the passive rather than the active sense, are manufactured in various shapes and sizes for different applications. You can see them on utility lines that carry high *voltage*. The insulators hold the wire up without risking a *short circuit* with a metal tower or a *bleedoff* (slow discharge) through a salt-water-soaked wooden pole.

If we try hard enough, we can force almost any electrical insulator to let electrons move by forcing ionization to occur. When electrons are stripped away from their atoms, they can roam more or less freely. Sometimes a normally insulating material gets charred, or melts down, or gets perforated by a spark. Then it loses its insulating properties, and electrons can move through it.

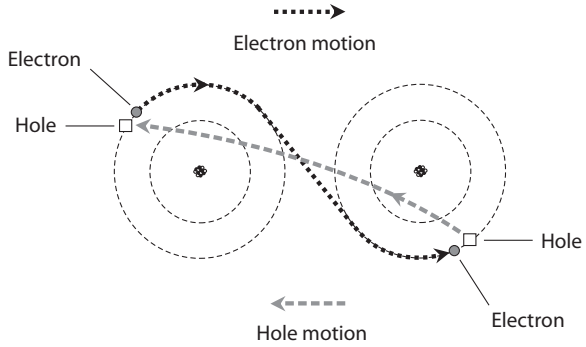
Resistors

Some substances, such as carbon, allow electrons to move among atoms *fairly* well. We can modify the conductivity of such materials by adding impurities such as clay to a carbon paste, or by winding a long, thin strand of the material into a coil. When we manufacture a component with the intent of giving it a specific amount of conductivity, we call it a *resistor*. These components allow us to limit or control the rate of electron flow in a device or system. As the conductivity improves, the *resistance* decreases. As the conductivity goes down, the resistance goes up. Conductivity and resistance vary in *inverse proportion*.

Engineers express resistance in units called *ohms*. The higher the resistance in ohms, the more opposition a substance offers to the movement of electrons. For wires, the resistance is sometimes specified in terms of *ohms per unit length* (foot, meter, kilometer, or mile). In an electrical system, engineers strive to minimize the resistance (or *ohmic value*) because resistance converts electricity into heat, reducing the *efficiency* that the engineers want and increasing the *loss* that they don't want.

Semiconductors

In a *semiconductor*, electrons flow easily under some conditions, and with difficulty under other conditions. In their pure form, some semiconductors carry electrons almost as easily as good conductors, while other semiconductors conduct almost as poorly as insulators. But semiconductors differ fundamentally from plain conductors, insulators, or resistors. In the manufacture of a semiconductor device, chemists treat the materials so that they conduct well some of the time, and poorly some of the time—and we can control the conductivity by altering the conditions. We find semiconductors in *diodes*, *transistors*, and *integrated circuits*.



1-6 In a sample of semiconductor material, the holes travel in a direction opposite the electron motion.

Semiconductors include substances, such as *silicon*, *selenium*, or *gallium,* that have been "doped" by the addition of *impurities,* such as *indium* or *antimony*. Have you heard of *gallium-arsenide diodes*, *metal-oxide transistors*, or *silicon rectifiers*? Electrical conduction in these materials occurs as a result of the motion of electrons, but the physical details of the process are rather complicated. Sometimes engineers speak of the movement of *holes* rather than electrons. A hole is a sort of electron deficiency. You might think of it as a place where an electron normally belongs, but for some reason it's missing. Holes travel opposite to the flow of electrons, as shown in Fig. 1-6.

When electrons make up most of the *charge carriers* in a substance, we have an *N-type semiconductor*. When most of the charge carriers are holes, we have a *P-type semiconductor*. A sample of P-type material passes some electrons, and a sample of N-type material carries some holes. We call the more abundant charge carrier the *majority carrier*, and the less abundant one the *minority carrier*.

Current

Whenever charge carriers move through a substance, an electric *current* exists. We express and measure current indirectly in terms of the number of electrons or holes passing a single point in one second. Electric current flows rapidly through any conductor, resistor, or semiconductor. Nevertheless, the charge carriers actually move at only a small fraction of the speed of light in a vacuum.

A great many charge carriers go past any given point in one second, even in a system carrying relatively little current. In a household electric circuit, a 100-watt (100-W) light bulb draws about *six quintillion* (6 followed by 18 zeroes or 6×10^{18}) charge carriers per second. Even the smallest bulb carries *quadrillions* (numbers followed by 15 zeros) of charge carriers every second. Most engineers find it inconvenient to speak of current in terms of charge carriers per second, so they express current in *coulombs per second* instead. We might think of a coulomb as an "engineer's superdozen"—approximately 6,240,000,000,000,000,000 (6.24×10^{18}) electrons or holes. When 1 coulomb (1 C) of charge carriers passes a given point per second, we have an *ampere*, the standard unit of electric current. A 60-W bulb in your desk lamp draws about half an ampere (0.5 A). A typical electric utility heater draws 10 A to 12 A.

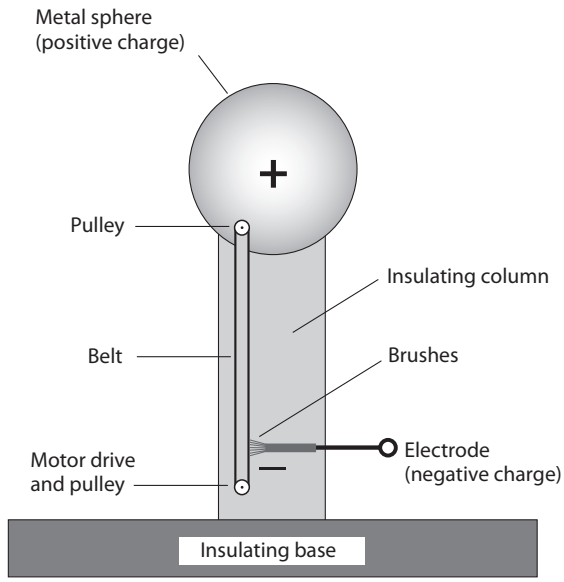
When a current flows through a resistance—always the case because even the best conductors have finite, nonzero resistance—we get heat. Sometimes we observe light as well. Old-fashioned *incandescent lamps* are deliberately designed so that the currents through their filaments produce visible light.

Static Electricity

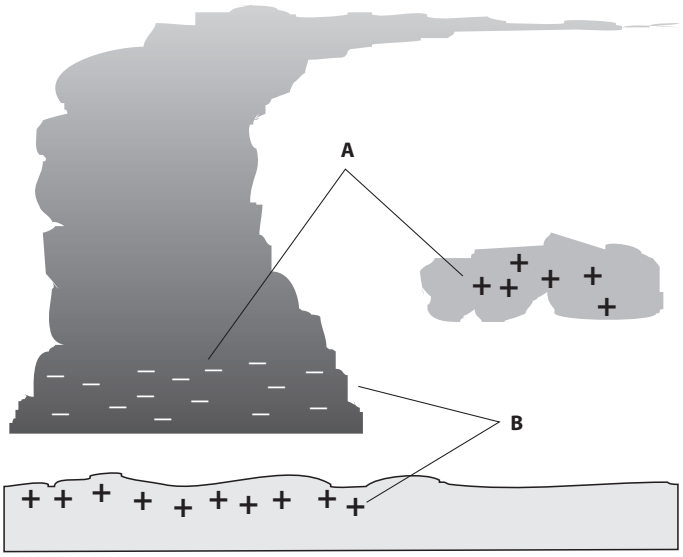
When you walk on a carpeted floor while wearing hard-soled shoes, an excess or shortage of electrons can develop on your body, creating *static electricity*. It's called "static" because the charge carriers don't flow—until you touch a metallic object connected to earth ground or to some large fixture. Then an abrupt discharge occurs, accompanied by a spark, a snapping or popping noise, and a startling sensation.

If you acquire a much greater charge than you do under ordinary circumstances, your hair will stand on end because every strand will repel every other as they all acquire a static charge of the same polarity. When the discharge takes place, the spark might jump a centimeter or more. Then it will more than startle you; you could actually get hurt. Fortunately, charge buildups of that extent rarely, if ever, occur with ordinary carpet and shoes. However, a device called a *Van de Graaff generator* (Fig. 1-7), found in physics labs, can cause a spark several centimeters long. Use caution if you work around these things. They can be dangerous.

Lightning provides the most spectacular example of the effects of static electricity on this planet. Lightning strokes commonly occur between clouds, and between clouds and the ground. The stroke is preceded by a massive static charge buildup. Figure 1-8 illustrates *cloud-to-cloud* (A) and *cloud-to-ground* (B) electric dipoles caused by weather conditions. In the scenario shown at B, the positive charge in the earth follows along beneath a storm cloud.



1-7 Simplified illustration of a Van de Graaff generator. This machine can create a charge sufficient to produce a spark several centimeters long.



1-8 Electrostatic charges can build up between clouds (A) or between a cloud and the earth's surface (B).

Electromotive Force (Voltage)

Charge carriers can move in an orderly fashion only if they experience a well-defined directional force in the form of a "push" or a "pull." This force can result from a buildup of static electric charges, as in the case of a lightning stroke. When the charge builds up, attended by *positive polarity* (shortage of electrons) in one place and *negative polarity* (excess of electrons) in another place, a powerful *electromotive force* (EMF) exists. We express and measure EMF in units called *volts*.

Ordinary household electricity has an effective EMF, or *voltage*, of between 110 volts (110 V) and 130 V; usually it's about 117 V. In the United States and most other countries, a new, fully charged car battery has an EMF of very close to 12.6 V. The static charge that you acquire when walking on a carpet with hard-soled shoes on a dry afternoon can reach several thousand volts. Before a discharge of lightning, millions of volts exist.

An EMF of 1 V, across a component having a resistance of 1 ohm, will cause a current of 1 A to flow through that component. In a DC circuit, the current (in amperes) equals the voltage (in volts) divided by the resistance (in ohms). This fact forms the cornerstone for a classic relationship in electricity called *Ohm's law*. If we double the voltage across a component whose resistance remains constant, then the current through that component doubles. If we keep the voltage constant but double the resistance, then the current goes down by half. We'll examine Ohm's law more closely later in this course.

Electromotive force can exist without any flow of current, producing static electricity, as we've seen. However, an EMF without current also exists between the two wires of an electric lamp when the switch is off. An EMF without current exists between the terminals of a common *flashlight cell* when we don't connect it to anything. Whenever we have an EMF between two points, an electric

current will flow if we provide a conductive path between those points. Voltage, or EMF, is sometimes called *electric potential* or *potential difference* for this reason. An EMF has the potential (that is, the ability) to move charge carriers, given the right conditions.

A huge EMF doesn't necessarily drive a lot of current through a conductor or resistance. Think of your body after you've spent some time walking around on the carpet. Although the EMF might seem deadly in terms of sheer magnitude (thousands of volts), relatively few coulombs of charge carriers accumulate on your body. In relative terms, not that many electrons flow through your finger when you touch an external object. That's why you don't get a severe shock. However, if plenty of coulombs are available, then even a modest EMF, such as 117 V (typical of a household utility outlet), can drive a lethal current through your body. That's why it's dangerous to repair an electrical device when it's connected to a source of power. The utility plant can deliver an unlimited number of coulombs.

Non-Electrical Energy

In scientific experiments, we often observe phenomena that involve energy in non-electrical form. Visible light provides an excellent example. A light bulb converts electricity into radiant energy that we can see. This fact motivated people like Thomas Edison to work with electricity, making discoveries and refining devices that make our lives convenient today. We can also convert visible light into electricity. A *photovoltaic cell* (also called a *solar cell*) works this sort of magic.

Light bulbs always give off heat as well as light. In fact, incandescent lamps actually give off more energy as heat than as light. You've probably had experience with electric heaters, designed for the purpose of changing electrical energy into heat energy. This "heat" is actually a form of radiant energy called *infrared* (IR), which resembles visible light, except that IR has a longer *wavelength* and you can't see the rays.

We can convert electricity into *radio waves*, *ultraviolet* (UV) rays, and *X rays*. These tasks require specialized devices such as radio transmitters, *mercury-vapor lamps*, and *electron tubes*. Fast-moving protons, neutrons, electrons, and atomic nuclei also constitute non-electrical forms of energy.

When a conductor moves in a magnetic field, electric current flows in that conductor. This effect allows us to convert mechanical energy into electricity, obtaining an *electric generator*. Generators can also work backwards, in which case we have an *electric motor* that changes electricity into mechanical energy.

A magnetic field contains energy of a unique kind. The science of *magnetism* is closely related to electricity. The oldest and most universal source of magnetism is the *geomagnetic field* surrounding the earth, which arises as a result of the alignment of iron atoms in the core of the planet.

A changing magnetic field creates a fluctuating electric field, and a fluctuating electric field produces a changing magnetic field. This phenomenon, called *electromagnetism*, makes it possible to send wireless signals over long distances. The electric and magnetic fields keep producing one another over and over again through space.

Dry cells, *wet cells*, and *batteries* convert *chemical energy* into electrical energy. In an automotive battery, for example, acid reacts with metal electrodes to generate a potential difference. When we connect the poles of the battery to a component having finite resistance, current flows. Chemical reactions inside the battery keep the current going for a while, but the battery eventually runs out of energy. We can restore the chemical energy to a lead-acid automotive battery (and certain other types) by driving current through it for a period of time, but some batteries (such as most ordinary flashlight cells and lantern batteries) become useless when they run out of chemical energy.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

2
CHAPTER

Electrical Units

LET'S LEARN ABOUT THE STANDARD UNITS THAT ENGINEERS USE IN DIRECT-CURRENT (DC) CIRCUITS. Many of these principles apply to common utility alternating-current (AC) systems as well.

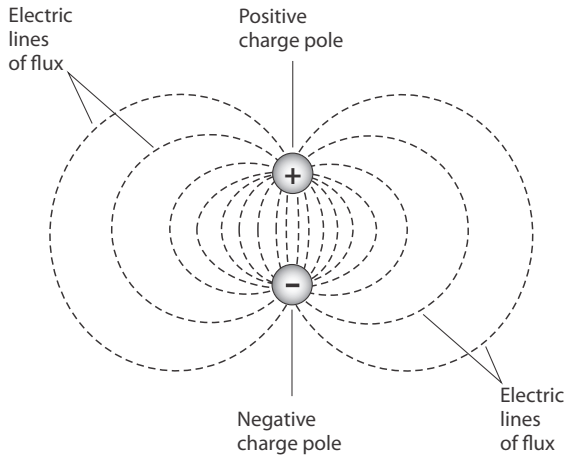
The Volt

In Chap. 1, you learned about the volt, the standard unit of electromotive force (EMF), or potential difference. An accumulation of electrostatic charge, such as an excess or shortage of electrons, always occurs when we have a potential difference between two points or objects. A power plant, an electrochemical reaction, light rays striking a semiconductor chip, and other phenomena can also produce voltages. We can get an EMF when we move an electrical conductor through a fixed magnetic field, or when we surround a fixed electrical conductor with a fluctuating magnetic field.

A potential difference between two points, called *poles*, invariably produces an *electric field*, represented by *electric lines of flux,* as shown in Fig. 2-1. We call such a pair of electrically charged poles an *electric dipole*. One pole carries relatively positive charge, and the other pole carries relatively negative charge. The positive pole always has fewer electrons than the negative pole. Note that the electron numbers are relative, not absolute! An electric dipole can exist even if both poles carry surplus electrons, or if both poles suffer from electron deficiencies, relative to some external point of reference having an absolutely neutral charge.

The abbreviation for volt (or volts) is V. Sometimes, engineers use smaller units. The *millivolt* (mV) equals 0.001 V. The *microvolt* (μV) equals 0.000001 V. Units larger than the volt also exist. One kilovolt (kV) represents 1000 V. One *megavolt* (MV) equals 1,000,000 V, or 1000 kV.

In an everyday dry cell, the poles maintain a potential difference somewhere between 1.2 and 1.7 V. In an automotive battery, it's in the range of 12 V to 14 V. In household AC utility wiring, the potential difference alternates polarity and maintains an effective value of approximately 117 V for electric lights and most small appliances, and 234 V for washing machines, ovens, or other large appliances. In some high-power radio transmitters, the EMF can range in the thousands of volts. The largest potential differences on our planet—upwards of 1 MV—build up in thunderstorms, sandstorms, and violent erupting volcanoes.



2-1 Electric lines of flux always exist near poles of electric charge.

The existence of a voltage always means that *charge carriers*, which are mostly electrons in a conventional circuit, will travel between the charge poles if we provide a decent path for them to follow. Voltage represents the driving force, or "pressure," that impels charge carriers to move. If we hold all other factors constant, a high voltage will make the charge carriers flow in greater quantity per unit of time, thereby producing a larger electrical current than a low voltage. But that statement oversimplifies the situation in most practical systems, where "all other factors" rarely "hold constant"!

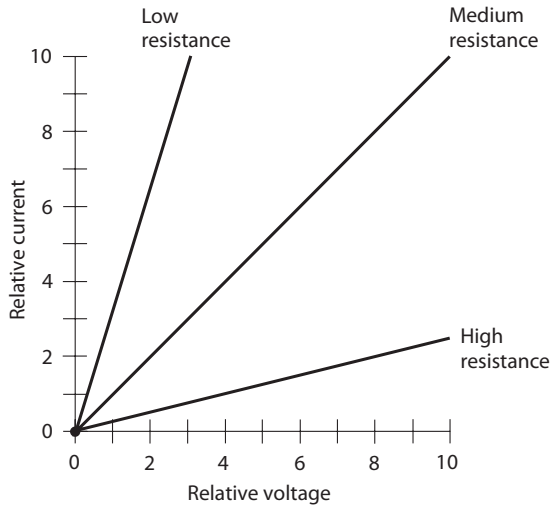
Current Flow

If we provide a conducting or semiconducting path between two poles having a potential difference, charge carriers flow in an attempt to equalize the charge between the poles. This current continues for as long as the path remains intact, and as long as a charge difference exists between the poles.

Sometimes the charge difference between two electric poles decreases to zero after a current has flowed for a while. This effect takes place in a lightning stroke, or when you touch a radiator after shuffling around on a carpet. In these instances, the charge between the poles equalizes in a fraction of a second. In other cases, the charge takes longer to dissipate. If you connect a piece of wire directly between the positive and negative poles of a dry cell, the cell "runs out of juice" after a few minutes. If you connect a light bulb across the cell to make a "flashlight," the charge difference may take an hour or two to get all the way down to zero.

In household electric circuits, the charge difference never equalizes unless a power failure occurs. Of course, if you short-circuit an AC electrical outlet (don't!), the fuse or breaker will blow or trip, and the charge difference will immediately drop to zero. But if you put a standard utility light bulb at the outlet, the charge difference will continue to exist at "full force" even as the current flows. The power plant can maintain a potential difference of 117 V across a lot of light bulbs indefinitely.

Have you heard that the deadly aspect of electricity results from current, not voltage? Literally, that's true, but the statement plays on semantics. You could also say "It's the heat, not the fire, that burns you." Okay! But a deadly current can arise only in the presence of a voltage sufficient to drive a certain amount of current through your body. You don't have to worry about deadly currents flowing between your hands when you handle a 1.5-V dry cell, even though, in theory, such a cell



2-2 Relative current as a function of relative voltage for low, medium, and high resistances.

could produce currents strong enough to kill you if your *body resistance* were much lower. You're safe when handling flashlight cells, but you've got good reason to fear for your life around household utility circuits. A voltage of 117 V, or even worse, the 220 V used in much of the world, can easily pump enough current through your body to electrocute you.

It all goes back to Ohm's law. In an electric circuit whose conductance (or resistance) never varies, the current is directly proportional to the applied voltage. If you double the voltage, you double the current. If you cut the voltage in half, the current goes down by half. Figure 2-2 shows this relationship as a graph in general terms. Here, we assume that the *power supply* can always provide as many charge carriers per unit of time as we need.

The Ampere

Current expresses the rate at which charge carriers flow past a fixed point per unit of time. The standard unit of current is the *ampere*, which represents one coulomb (6,240,000,000,000,000,000, or 6.24×10^{18}) of charge carriers flowing past a given point every second.

An ampere is a comparatively large amount of current. The abbreviation is A. Often, you'll want to express current in terms of *milliamperes,* abbreviated mA, where 1 mA = 0.001 A. You'll also sometimes hear of *microamperes* (μA), where 1 μA = 0.000001 A or 0. 001 mA. You might even encounter *nanoamperes* (nA), where 1 nA = 0.000000001 A = 0.001 μA.

A current of a few milliamperes will give you a rude electrical shock. About 50 mA will jolt you severely, and 100 mA can kill you if it flows through your heart. An ordinary utility light bulb draws 0.5 A to 1 A of current in a household utility circuit. An electric iron draws approximately 10 A; an entire household normally uses between 10 A and 100 A, depending on the size of the house and the kinds of appliances it has, and also on the time of day, week, or year.

The amount of current that flows in an electrical circuit depends on the voltage, and also on the resistance. In some electrical systems, extremely large currents, say 1000 A, can flow. You'll get a current like this if you place a metal bar directly across the output terminals of a massive electric generator. The bar has an extremely low resistance, and the generator can drive many coulombs of charge carriers through the bar every second. In some semiconductor electronic devices, a few

nanoamperes will suffice to allow for complicated processes. Some electronic clocks draw so little current that their batteries last as long as they would if you left them on the shelf.

Resistance and the Ohm

Resistance quantifies the opposition that a circuit imposes against the flow of electric current. You can compare resistance to the *reciprocal* of the diameter of a garden hose (where conductance compares to the actual diameter). For metal wire, this analogy works pretty well. Small-diameter wire has higher resistance than large-diameter wire made of the same metal.

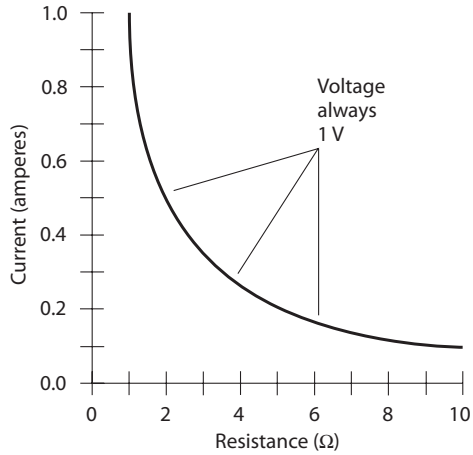
The standard unit of resistance is the *ohm*, pronounced to rhyme with "gnome," written as an upper-case Greek letter omega (Ω). You'll also hear about *kilohms* (symbolized k or kΩ), where 1 k = 1000 Ω, or about 1 *megohm* (symbolized M or MΩ), where 1 M = 1,000,000 Ω or 1000 k.

Electric wire is sometimes rated for *resistance per unit length.* The standard unit for this purpose is the Ω *per foot* (Ω/ft) or the Ω *per meter* (Ω/m). You might also come across the unit *ohm per kilometer* (ohm/km). Table 2-1 shows the resistance per unit of length for various common sizes of solid copper wire at room temperature as a function of the wire size, as defined by a scheme known as the *American Wire Gauge* (AWG).

When we place a potential difference of 1 V across a component whose resistance equals 1 Ω, assuming that the power supply can deliver an unlimited number of charge carriers, we get a current of 1 A. If we double the resistance to 2 Ω, the current decreases to 0.5 A. If we cut the resistance by a factor of 5 to get only 0.2 Ω, the current increases by the same factor, from 1 A to 5 A. The current flow, for a constant voltage, varies in *inverse proportion* to the resistance. Figure 2-3 shows the current, through components of various resistances, given a constant potential difference of 1 V.

Table 2-1. Approximate resistance per unit of length in ohms per kilometer (Ω/km) at room temperature for solid copper wire as a function of the wire size in American Wire Gauge (AWG)

Wire size, AWG #	Ω/km
2	0.52
4	0.83
6	1.3
8	2.7
10	3.3
12	5.3
14	8.4
16	13
18	21
20	34
22	54
24	86
26	140
28	220
30	350



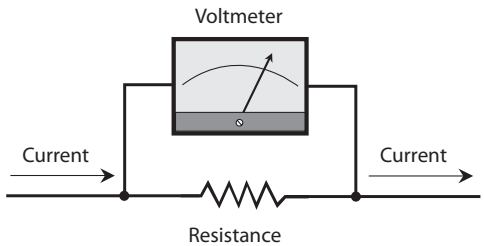
2-3 Current as a function of resistance through an electric device for a constant voltage of 1 V.

Whenever an electric current flows through a component, a potential difference appears across that component. If the component has been deliberately manufactured to exhibit a certain resistance, we call it a *resistor*. Figure 2-4 illustrates this effect. In general, the potential difference arises in direct proportion to the current through the resistance. Engineers take advantage of this effect when they design electronic circuits, as you'll learn later in this book.

Electrical circuits always have some resistance. No such thing as a perfect conductor (an object with mathematically zero resistance) exists in the "real world." When scientists cool certain metals down to temperatures near *absolute zero*, the substances lose practically all of their resistance, so that current can flow around and around for a long time. This phenomenon is called *superconductivity*. But nothing can ever become an *absolutely perfect* conductor.

Just as a *perfectly* resistance-free substance cannot exist in the real world, we'll never encounter an *absolutely infinite* resistance, either. Even dry air conducts electric current to some extent, although the effect is usually so small that scientists and engineers can ignore it. In some electronic applications, engineers select materials based on how "nearly infinite" their resistance appears; but when they say that, they exploit a figure of speech. They really mean to say that the resistance is so gigantic that we can consider it "infinite" for all practical purposes.

In electronics, the resistance of a component often varies, depending on the conditions under which that component operates. A transistor, for example, might have high resistance some of the time, and low resistance at other times. High/low resistance variations can take place thousands, millions, or billions of times each second. In this way, oscillators, amplifiers, and digital devices function in radio receivers and transmitters, telephone networks, digital computers, and satellite links (to name just a few applications).



2-4 Whenever current passes through a component having resistance, a voltage exists across that component.

Conductance and the Siemens

Electricians and engineers sometimes talk about the *conductance* of a material, rather than about its resistance. The standard unit of conductance is the *siemens,* abbreviated S. When a component has a conductance of 1 S, its resistance equals 1 Ω. If we double the resistance of a component, its conductance drops to half the former value. If we halve the resistance, we double the conductance. Conductance in siemens always equals the reciprocal of resistance in ohms, as long as we confine our attention to one component or circuit at a time.

If we know the resistance of a component in ohms, we can get the conductance in siemens; we simply divide 1 by the resistance. If we know the conductance in siemens, we can get the resistance in ohms; we divide 1 by the conductance. In calculations and equations, engineers denote resistance by writing an italicized, uppercase letter *R*, and conductance by writing an italicized, uppercase letter *G*. If we express *R* in ohms and *G* in siemens, then

$$G = 1/R$$

and

$$R = 1/G$$

In "real-world" electrical and electronic circuits, you'll often use units of conductance much smaller than the siemens. A resistance of 1 k represents a conductance of one *millisiemens* (1 mS). If we encounter a component whose resistance equals 1 M, its conductance is one *microsiemens* (1 μS). You'll sometimes hear about *kilosiemens* (kS) or *megasiemens* (MS), representing resistances of 0.001 Ω and 0.000001 Ω, respectively. Short lengths of heavy wire have conductance values in the range of kilosiemens. A heavy, solid copper or silver rod might exhibit a conductance in the megasiemens range.

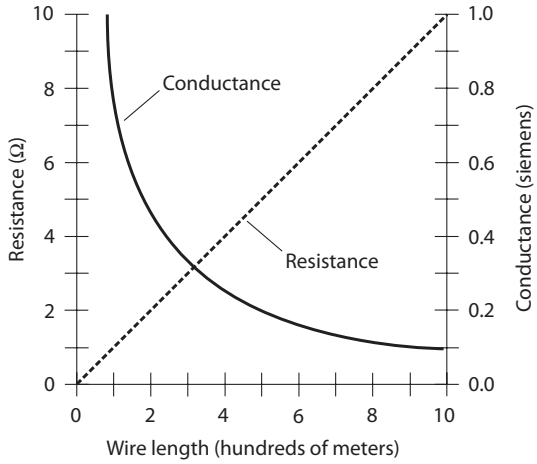
If a component has a resistance of 50 Ω, its conductance equals 1/50 S or 0.02 S. We can also call this quantity 20 mS. Now imagine a piece of wire with a conductance of 20 S. Its resistance equals 1/20 Ω or 0.05 Ω. You won't often hear or read the term "milliohm" in technical conversations or papers, but you might say that an 0.05-Ω length of wire has a resistance of 50 milliohms, and you'd be technically correct.

When you want to determine the *conductivity* of a component, circuit, or system, you must exercise caution or you might end up calculating the wrong value. If wire has a resistance per unit length of 10 Ω/km, you can't say that it has a conductivity of 1/10, or 0.1, S/km. A 1-km span of such wire does indeed have a conductance of 0.1 S, but a 2-km span of the same wire has a resistance of 20 Ω because you have twice as much wire. That's not twice the conductance, but half the conductance, of the 1-km span. If you say that the conductivity of the wire is 0.1 S/km, then you might be tempted to say that 2 km of the wire has 0.2 S of conductance. That would be a mistake! Conductance decreases with increasing wire length.

Figure 2-5 illustrates the resistance and conductance values for various lengths of wire having a resistance per unit length of 10 Ω/km.

Power and the Watt

Whenever we drive an electrical current through a resistive component, the temperature of that component rises. We can measure the intensity of the resulting *heat* in units called *watts* (symbolized W), representing *power*. (As a variable quantity in equations, we denote power by writing *P*.) Power can manifest itself in various forms such as mechanical motion, radio waves, visible light, or noise.



2-5 Resistance and conductance for various lengths of wire having a resistivity of 10 Ω/km.

But we'll always find heat (in addition to any other form of power) in an electrical or electronic device, because no "real-world" system operates with 100-percent efficiency. Some power always goes to waste, and this waste shows up mainly as heat.

Look again at Fig. 2-4. A certain potential difference appears across the resistor, although the illustration does not reveal the actual voltage. A current flows through the resistor; again, the diagram doesn't tell us the value. Suppose that we call the voltage across the resistor V and the current through the resistor I, expressed in volts (V) and amperes (A), respectively. If we let P represent the power in watts dissipated by the resistor, then

$$P = VI$$

If the voltage V across the resistance is caused by two flashlight cells in series, giving us 3 V, and if the current I through the resistance (a flashlight bulb, perhaps) equals 0.2 A, then V = 3 V and $I = 0.2$ A, and we can calculate the power P as

$$P = VI = 3 \times 0.2 = 0.6 \text{ W}$$

Suppose the voltage equals 220 V, giving rise to a current of 400 mA. To calculate the power, we must convert the current into amperes: 400 mA = 400/1000 A = 0.400 A. Then we have

$$P = VI = 220 \times 0.400 = 88.0 \text{ W}$$

You will often hear about *milliwatts* (mW), *microwatts* (μW), *kilowatts* (kW), and *megawatts* (MW). By now, you should know what these units represent when you see the prefixes. Otherwise, you can refer to Table 2-2, which lists the common *prefix multipliers* for physical units.

Once in a while, you'll want to take advantage of the power equation to find a current through a component or a voltage across a component. In that case, you can use the variant

$$I = P/V$$

to find current, or

$$V = P/I$$

Table 2-2. Prefix multipliers from 0.000000000001 (trillionths, or units of 10^{-12}) to 1,000,000,000,000 (trillions, or units of 10^{12})

Prefix	Symbol	Multiplier
pico-	p	0.000000000001 (or 10^{-12})
nano-	n	0.000000001 (or 10^{-9})
micro-	μ	0.000001 (or 10^{-6})
milli-	m	0.001 (or 10^{-3})
centi-	c	0.01 (or 10^{-2})
deci-	d	0.1 (or 10^{-1})
kilo-	k	1000 (or 10^{3})
mega-	M	1,000,000 (or 10^{6})
giga-	G	1,000,000,000 (or 10^{9})
tera-	T	1,000,000,000,000 (or 10^{12})

to find the voltage. Always convert to standard units (volts, amperes, and watts) before performing calculations with any of these formulas. Otherwise, you risk getting an answer that's too large or small by one or more *orders of magnitude* (powers of 10)!

A Word about Notation

Sometimes, symbols and abbreviations appear in italics, and sometimes they don't. We'll encounter subscripts often, and sometimes even the subscripts are italicized. Following are some rules that apply to notation in electricity and electronics.

- We never italicize the abbreviations for units such as volts (V), amperes (A), and watts (W).
- We never italicize the abbreviations for objects or components such as resistors (R), batteries (B), capacitors (C), and inductors (L).
- We never italicize the abbreviations for quantifying prefixes such as kilo- (k), micro- (μ), mega- (M), or nano- (n).
- Labeled points in drawings might or might not be italicized. It doesn't matter, as long as a diagram remains consistent within itself. We might call a point either P or *P*, for example.
- We always italicize the symbols for mathematical constants and variables such as time (*t*), the speed of light in a vacuum (*c*), velocity (*v*), and acceleration (*a*).
- We always italicize the symbols for electrical quantities such as voltage (*V*), current (*I*), resistance (*R*), and power (*P*).
- We never italicize numeric subscripts. We might denote a certain resistor as R_2, but never as R_{2}; we might denote a certain amount of current as I_4, but never as I_{4}.
- For non-numeric subscripts, the same rules apply as for general symbols.

Once in a while we'll see the same symbol italicized in one place and not in another—even within a single diagram or discussion! We might, for example, talk about "resistor number 3" (symbolized R_3), and then later in the same paragraph talk about its value as "resistance number 3" (symbolized R_3). Still later, we might talk about "the *n*th resistor in a combination of resistors" (R_n) and then "the *n*th resistance in a combination of resistances" (R_n).

Energy and the Watt-Hour

Have you heard the terms "power" and "energy" used interchangeably, as if they mean the same thing? Well, they don't! The term *energy* expresses power dissipated over a certain period of time. Conversely, the term *power* expresses the instantaneous rate at which energy is expended at a particular moment in time.

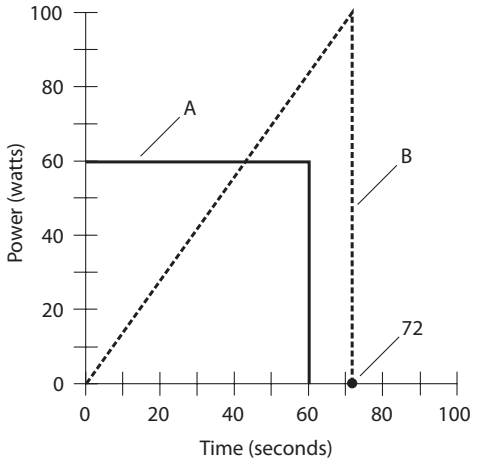
Physicists measure energy in units called *joules*. One joule (1 J) technically equals a *watt-second*, the equivalent of 1 W of power dissipated for 1 s of time (1 W · s or 1 Ws). In electricity, you'll more often encounter the *watt-hour* (symbolized W · h or Wh) or the *kilowatt-hour* (symbolized kW · h or kWh). As their names imply, a watt-hour represents the equivalent of 1 W dissipated for 1 h, and 1 kWh represents the equivalent of 1 kW of power dissipated for 1 h.

An energy quantity of 1 Wh can manifest itself in infinitely many ways. A light bulb rated at 60 W consumes 60 Wh in 1 h, the equivalent of a watt-hour per minute (1 Wh/min). A lamp rated at 100 W consumes 1 Wh in 1/100 h, or 36 s. Whenever we double the power, we halve the time required to consume 1 Wh of energy. But in "real-world" scenarios, the rate of power dissipation rarely remains constant. It can change from moment to moment in time.

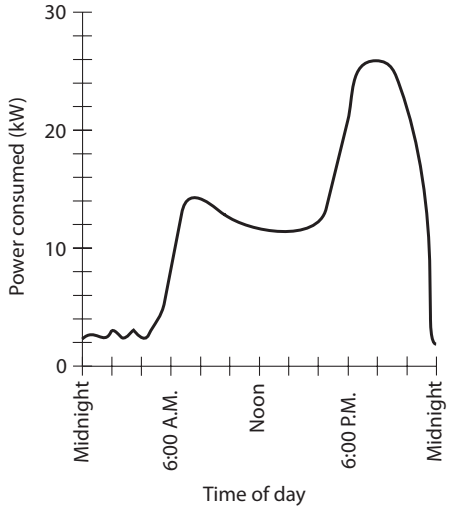
Figure 2-6 illustrates two hypothetical devices that consume 1 Wh of energy. Device A uses its power at a constant rate of 60 W, so it consumes 1 Wh in 1 min. The power consumption rate of device B varies, starting at zero and ending up at more than 60 W. How do we know that this second device really consumes 1 Wh of energy? To figure that out, we must determine the *area under the graph*. In this case, the graph encloses a simple triangle. We recall from our basic geometry courses that the area of a triangle equals half the base length times the height. Device B receives power for 72 s, or 1.2 min; that's $1.2/60 = 0.02$ h. The area under the graph is therefore $1/2 \times 100 \times 0.02 = 1$ Wh.

When you calculate energy values, you must always keep in mind the units with which you work. In the example of Fig. 2-6, you would use the watt-hour, so you must multiply watts by hours. If you multiply watts by minutes or seconds, you'll get the wrong kind of unit in your answer—and sometimes a "unit" that doesn't have any technical definition at all!

Often, graphs of power versus time show up as complex curves, not "neat" figures, such as rectangles or triangles. Consider the graph of power consumption in your home, as a function of time, over the course of a hypothetical day. That graph might resemble the curve in Fig. 2-7. Obviously, you won't find a simple formula to calculate the area under this curve, but you can use another



2-6 Two devices that consume 1 Wh of energy. Device A dissipates a constant amount of power as time passes. Device B dissipates an increasing amount of power as time passes.



2-7 Graph showing the amount of power consumed by a hypothetical household, as a function of the time of day.

scheme to determine the total energy consumed by your household over a period of time. You can employ a special meter that measures electrical energy in kilowatt-hours (kWh).

Every month, the power company sends a representative, or employs a wireless device, to record the number of kilowatt-hours that your meter displays. The power company then subtracts the reading taken the previous month from the current value. A few days later, you get a bill for the month's energy usage. The "power meter" (a misnomer, because it's really an energy meter) automatically keeps track of total consumed energy, without anybody having to go through high-level mathematical calculations to find the areas under irregular curves, such as the graph of Fig. 2-7.

Other Energy Units

The joule, while standard among scientists, isn't the only energy unit that exists. You'll occasionally encounter the *erg*, a tiny unit equivalent to 0.0000001 of a joule. Some scientists use the erg in laboratory experiments involving small amounts of expended energy.

Most folks have heard or read about the *British thermal unit (Btu)*, equivalent to 1055 joules. People use the Btu to define the cooling or heating capacity of air-conditioning equipment. To cool your room from 85 to 78°F, you need a certain amount of energy, perhaps best specified in Btu. If you plan to have an air conditioner or furnace installed in your home, an expert will determine the size of the unit that best suits your needs. That person might tell you how "powerful" the unit should be, in terms of its ability to heat or cool in British thermal units per hour (Btu/h).

Physicists also use, in addition to the joule, a unit of energy called the *electron-volt* (eV). It's a minuscule unit indeed, equal to only 0.00000000000000000016 joule (you can count 18 zeroes after the decimal point and before the 1). Physicists represent this number as 1.6×10^{-19}. A single electron in an electric field of 1 V gains 1 eV of energy. Atomic physicists rate *particle accelerators* (or, informally, "atom smashers") in terms of *megaelectron-volts* (MeV, where 1 MeV = 1,000,000 eV), or *gigaelectron-volts* (GeV, where 1 GeV = 1000 MeV), or *teraelectron-volts* (TeV, where 1 TeV = 1000 GeV) of energy capacity.

Another energy unit, employed to denote mechanical work, is the *foot-pound* (ft-lb). It's the amount of "labor" needed to elevate a weight of one pound (1 lb) straight upward by a distance of one foot (1 ft), not including any friction. One foot-pound equals 1.356 J.

Table 2-3. Conversion factors between joules and various other energy units

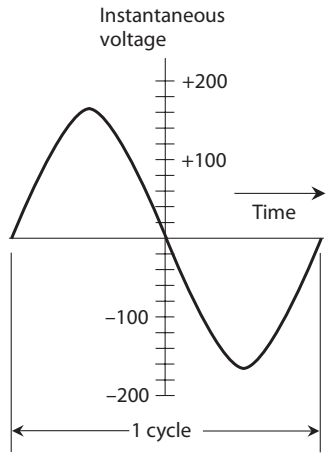
Unit	To convert energy in the unit at left to joules, multiply by	To convert energy in joules to the unit at left, multiply by
British thermal units (Btu)	1055	0.000948
electron-volts (eV)	1.6×10^{-19}	6.2×10^{18}
ergs	0.0000001 (or 10^{-7})	10,000,000 (or 10^{7})
foot-pounds (ft-lb)	1.356	0.738
watt-hours (Wh)	3600	0.000278
kilowatt-hours (kWh)	3,600,000 (or 3.6×10^{6})	0.000000278 (or 2.78×10^{-7})

Table 2-3 summarizes all of the energy units described here, along with conversion factors to help you change from any particular unit to joules or vice-versa. The table includes watt-hours and kilowatt-hours. In electricity and electronics, you'll rarely need to concern yourself with any energy unit other than these two.

Alternating Current and the Hertz

Direct current (DC) always flows in the same direction, but household utility current reverses direction at regular intervals. In the United States (and much of the world), the current direction reverses once every 1/120 second, completing one full cycle every 1/60 second. In some countries, the direction reverses every 1/100 second, taking 1/50 second to go through a complete cycle. When we encounter a periodically reversing current flow of this sort, we call it *alternating current* (AC).

Figure 2-8 shows a common "117-V" utility AC wave as a graph of voltage versus time. If you're astute, you'll see that the maximum positive and negative voltages don't equal 117 V. Instead, they come close to 165 V. The *effective voltage* for an AC wave usually differs from the *instantaneous maximum*, or *peak*, voltage. For the waveform shown in Fig. 2-8, the effective value is approximately



2-8 One cycle of utility alternating current (AC). The instantaneous voltage is the voltage at any particular instant in time. The peak voltages are approximately plus and minus 165 V.

0.707 times the peak value (the theoretically exact multiplication factor equals the reciprocal of the square root of 2). Conversely, the peak value is approximately 1.414 times the effective value (theoretically the factor equals the square root of 2).

The *hertz* (symbolized *Hz*) is the basic unit of AC *frequency*. One hertz represents one complete cycle per second. Because a typical utility AC cycle repeats itself every 1/60 second, we say that the wave has a *frequency* of 60 Hz. In the United States, 60 Hz is the standard frequency for AC. In much of the rest of the world, however, it's 50 Hz.

In wireless communications, you'll hear about *kilohertz* (kHz), *megahertz* (MHz), and *gigahertz* (GHz). These units relate to each other as follows:

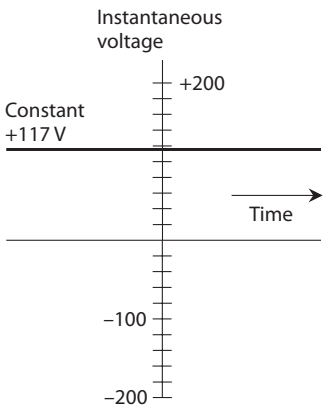
- 1 kHz = 1000 Hz = 10^3 Hz
- 1 MHz = 1000 kHz = 1,000,000 Hz = 10^6 Hz
- 1 GHz = 1000 MHz = 1,000,000 kHz = 1,000,000,000 Hz = 10^9 Hz

Usually, but not always, the waves have shapes like the one shown in Fig. 2-8. Engineers and technicians call it a *sine wave* or a *sinusoid*.

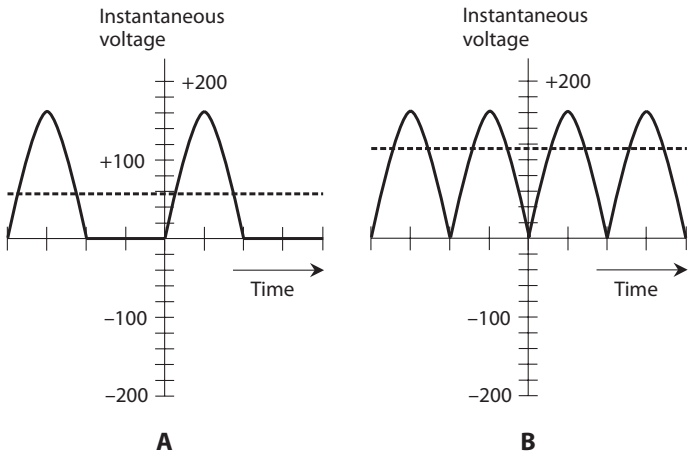
Rectification and Pulsating Direct Current

Batteries and other sources of DC produce constant voltage, which we can graphically portray by plotting a straight, horizontal line on a coordinate grid, showing voltage as a function of time. Figure 2-9 shows a representation of DC. For pure DC, the peak voltage equals the effective voltage. In some systems that derive their power from sources other than batteries, the instantaneous DC voltage fluctuates rapidly with time. This situation exists, for example, if we pass the sinusoid of Fig. 2-8 through a *rectifier* circuit, which allows the current to flow in only one direction.

Rectification changes AC into DC. To obtain rectification, we can use a device called a *diode*. When we rectify an AC wave, we can either cut off or invert one-half of the AC wave to get *pulsating DC* output. Figure 2-10 illustrates two different waveforms of pulsating DC. In the waveform at A, we simply remove the negative (bottom) half of the cycle. In the situation at B, we invert the negative portion of the wave, making it positive instead—a "mirror image" of its former self. Figure 2-10A shows *half-wave rectification*; it involves only half the waveform. Figure 2-10B illustrates *full-wave rectification*, in which both halves of the waveform contribute to the output. In the



2-9 A representation of pure direct current (DC).



2-10 At A, half-wave rectification of common utility AC. At B, full-wave rectification of common utility AC. Effective voltages are shown by the dashed lines.

output of a full-wave rectifier, all of the original current still flows, even though it doesn't alternate as the input does.

The effective value, compared with the peak value, for pulsating DC depends on whether we apply half-wave or full-wave rectification to an AC wave. In Figs. 2-10A and 2-10B, the effective voltages appear as dashed lines, and the instantaneous voltages show up as solid curves. The instantaneous voltage changes from *instant* to *instant* in time (that's where the term comes from).

In Fig. 2-10B, the effective voltage equals $2^{-1/2}$ (roughly 0.707) times the peak voltage, just as with ordinary AC. The direction of current flow, for many kinds of devices, doesn't make any difference. But in Fig. 2-10A, half of the wave has been lost, cutting the effective value in half so that it's only $2^{-1/2}/2$ (approximately 0.354) times the peak voltage.

In household "wall-outlet" AC for powering-up conventional appliances in the United States, we observe a peak EMF of about 165 V, and an effective EMF of about 117 V. If we subject this electricity to full-wave rectification, both the peak and the effective EMFs remain at these values. If we put such a wave through a half-wave rectifier, the peak EMF remains the same, but the effective output EMF drops to about 58.5 V.

Stay Safe!

For all "intents and purposes," one rule applies concerning safety around electrical apparatus. Never forget it, even for a moment. One careless move can kill anyone.

Warning!

If you have any doubts about whether or not you can safely work with a device, assume that you *cannot*. In that case, have a professional electrician work on it.

Household electricity, with an effective EMF of about 117 V (and often twice that for large electrical appliances and in many places outside the United States), is more than sufficient to kill you if it drives current through your chest cavity. Certain devices, such as spark coils, can produce lethal currents even from an automotive battery. Consult the American Red Cross or your electrician concerning what types of circuits, procedures, and devices are safe, and what kinds aren't.

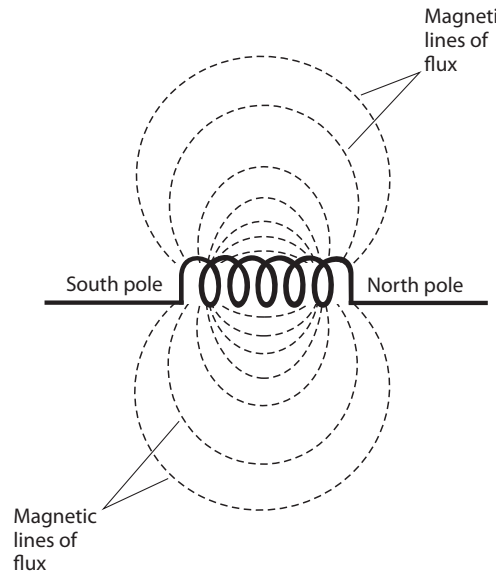
Magnetism

Whenever an electric current flows—that is, whenever charge carriers move—a *magnetic field* appears in the vicinity. In a straight wire that carries electrical current, *magnetic lines of flux* surround the wire in circles, with the wire at the center. (The lines of flux aren't physical objects, but they offer a convenient way to represent the magnetic field.) You'll sometimes hear or read about a certain number of flux lines per unit cross-sectional area, such as 100 lines per square centimeter. That terminology expresses, informally, the relative intensity of the magnetic field.

Magnetic fields arise whenever the atoms of certain materials align themselves. Iron is the most familiar element with this property. The atoms of iron in the earth's core have become aligned to some extent, a complex phenomenon caused by the rotation of our planet and its motion with respect to the magnetic field of the sun. The magnetic field surrounding the earth gives rise to fascinating effects, such as the concentration of charged particles that you see as the *aurora borealis* during a "solar storm."

When you wind a piece of wire into a tight coil, the resulting magnetic flux takes a shape similar to the flux field surrounding the earth. Two well-defined *magnetic poles* develop, as shown in Fig. 2-11. You can increase the intensity of such a field by placing a special core inside the coil. Iron, steel, or some other material that can be readily magnetized works very well for this purpose. We call such substances *ferromagnetic materials*.

A ferromagnetic core doesn't increase the total *quantity* of magnetism in and around a coil, but it can produce a more *intense* field. This is the principle by which an electromagnet works. It also



2-11 Magnetic flux lines around a current-carrying coil of wire. The flux lines converge at the magnetic poles.

facilitates the operation of electrical transformers for utility current. Technically, magnetic lines of flux emerge from north poles and converge toward south poles. Therefore, the magnetic field "flows" from the north end of a coil or bar magnet to the south end, following the lines of flux through the surrounding space.

Magnetic Units

We can express the overall quantity of a magnetic field in units called *webers*, abbreviated Wb. One weber is mathematically equivalent to one volt-second (1 V · s). For weaker magnetic fields, we can use a smaller unit called the *maxwell* (symbolized Mx). One maxwell equals 0.00000001 Wb, or 0.01 microvolt-second (0.01 μV · s).

We can express the *flux density* of a magnetic field in terms of webers or maxwells per square meter or per square centimeter. A flux density of one weber per square meter (1 Wb/m^2) represents one *tesla* (1 T). One *gauss* (1 G) equals 0.0001 T, or one maxwell per square centimeter (1 Mx/cm^2).

In general, as the electric current through a wire increases, so does the flux density near the wire. A coiled wire produces a greater flux density for a given electrical current than a single, straight wire. As we increase the number of turns in a coil of a specific diameter that carries a constant current, the flux density in and around the coil increases.

Sometimes, engineers specify magnetic field strength in *ampere-turns* (At). The ampere-turn quantifies a phenomenon called *magnetomotive force*. A one-turn wire loop, carrying 1 A of current, produces a magnetomotive force of 1 At. Doubling the number of turns with constant current doubles the magnetomotive force. Doubling the current for a constant number of turns also doubles the magnetomotive force. If you have 10 A flowing in a 10-turn coil, the magnetomotive force equals 10 × 10, or 100 At. If you have 100 mA flowing in a 200-turn coil, the magnetomotive force equals 0.1 × 200, or 20 At. (Remember that 100 mA = 0.1 A.)

Once in a while, you might hear or read about a unit of magnetomotive force called the *gilbert* (Gb). One gilbert equals approximately 0.796 At. Conversely, 1 At equals approximately 1.26 Gb.

Tech Tidbit

A DC-carrying coil's magnetomotive force depends only on the current through the coil and the number of turns that it has. That's it! Nothing else makes any difference.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

3
CHAPTER

Measuring Devices

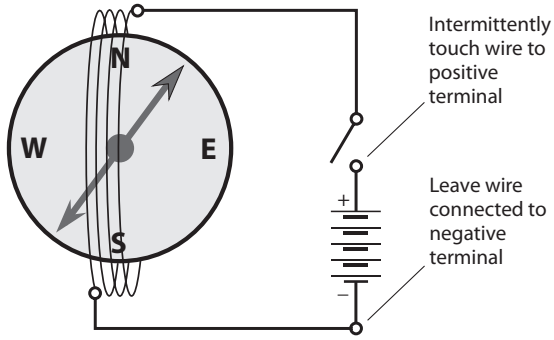
LET'S LOOK AT THE INSTRUMENTS THAT ENGINEERS USE TO MEASURE ELECTRICAL QUANTITIES. Some measuring devices work because electric and magnetic fields produce forces proportional to the field intensity. Some meters determine electric current by measuring the amount of heat produced as charge carriers move through a medium with a known resistance. Some meters have small motors whose speed depends on the measured quantity of charge carriers. But, the most common types of measurement device convert analog measurements into a digital form (for processing) before displaying them.

Electromagnetic Deflection

Early experimenters noticed that an electric current produces a magnetic field. They could detect this field by placing a magnetic compass near a wire. The compass pointed toward magnetic north when the wire carried no current. But when the experimenters drove DC through the wire by connecting it to the terminals of a battery, the compass needle deflected toward the east or west. The extent of the deflection depended on the distance between the compass and the wire, and also on how much current the wire carried. When scientists first observed this effect, they called it *electromagnetic deflection*, and that's still a good descriptive term today.

Experimenters tried various compass-and-wire arrangements to find out how much they could force the compass needle to rotate. The experimenters also wanted to make the device as sensitive as possible. When scientists wrapped the wire in a coil around the compass, as shown in Fig. 3-1, they got a device that could detect small currents. They called this effect *galvanism*, and they called the coil-around-a-compass device a *galvanometer*. The extent of any given galvanometer's needle displacement increased with increasing current. The experimenters had almost reached their goal of building a meter that could quantitatively measure current, but one final challenge remained: calibrate the galvanometer somehow.

You can make a galvanometer at home. Buy a cheap compass, about two feet of insulated bell wire, and a large 6-V lantern battery. Wind the wire around the compass four or five times, as shown in Fig. 3-1, and align the compass so that the needle points along the wire turns when the wire is disconnected from the battery. Make sure that the compass lies flat on a horizontal surface, such as a table or desk. Then connect one end of the wire to the negative (−) terminal of the battery. Touch

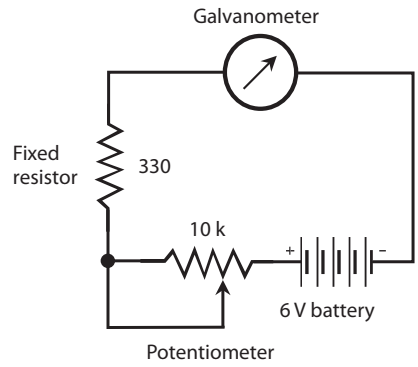


3-1 A simple galvanometer. The magnetic compass must lie flat.

the other end to the positive (+) terminal for a moment, and watch the compass needle. Don't leave the wire connected to the battery for more than a few seconds at a time.

You can buy a *resistor* and a *linear-taper potentiometer* at an electronics retail outlet and set up an experiment that shows how galvanometers measure current. For a 6-V lantern battery, the fixed resistor should have a value of at least 330 Ω and should be rated to dissipate at least ¼ W of power. The potentiometer should have a maximum value of 10 k. Connect the resistor and the potentiometer in series between one end of the bell wire and one terminal of the battery, as shown in Fig. 3-2. Short-circuit the center contact of the potentiometer to one of its end contacts. Use the resulting two terminals in the circuit.

When you adjust the potentiometer, the compass needle should deflect more or less, depending on the current through the wire. As the resistance decreases, the current increases, and so does the number of degrees by which the needle deflects. You can vary the current by changing the potentiometer setting. You can reverse the direction of needle deflection by reversing the battery polarity. Early experimenters calibrated their galvanometers by referring to the "degrees" scale around the perimeter of the compass, and generating graphs of degrees versus amperes. They calculated the theoretical currents in amperes by dividing the known voltage by the known resistance, taking advantage of Ohm's law, about which you'll learn more in the next chapter.



3-2 A circuit for demonstrating how a galvanometer indicates relative current. Resistances are in ohms; k indicates kilohms.

Electrostatic Deflection

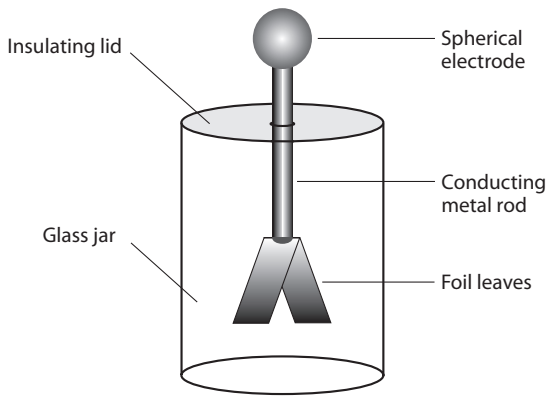
Electric fields produce forces, just as magnetic fields do. Have you noticed this effect when your hair stands on end in dry, cold weather? If you live in a place where the winters get severe, shuffle around on a rug while wearing hard-soled shoes next January, and see if you can get your hair to stand up. Have you heard that people's hair stands up just before a lightning bolt hits nearby? Sometimes it does, but not always.

The *electroscope* is a common physics lab device for demonstrating electrostatic force. It consists of two foil leaves, attached to a conducting rod, placed in a sealed container so that air currents can't disturb the leaves (Fig. 3-3). When a charged object comes near, or touches, the contact at the top of the rod, the leaves stand apart from each other because the leaves become flooded with like electric charges—either an excess or a deficiency of electrons—and "like poles always repel." The extent to which the leaves stand apart depends on the amount of electric charge. It's difficult to measure this deflection and correlate it with charge quantity; electroscopes don't make good meters. But electrostatic forces can operate against tension springs or magnets, allowing engineers to build sensitive, accurate *electrostatic meters*.

An electrostatic meter can quantify alternating (or AC) electric charges as well as direct (or DC) charges. This property gives electrostatic meters an advantage over electromagnetic meters, such as the galvanometer. If you connect a source of AC to the coil of the galvanometer device portrayed in Fig. 3-1, the current in one direction pulls the meter needle one way, the current in the other direction pushes the needle the opposite way, and the opposing forces alternate so fast that the needle doesn't have time to deflect noticeably in either direction! But if you connect a source of AC to an electrostatic meter, the plates repel whether the charge is positive or negative at any given instant in time. The alternations make no difference in the direction of the force.

Most electroscopes aren't sensitive enough to show much deflection with ordinary 117-V utility AC. Don't try connecting 117 V to an electroscope anyway, however. That electricity can present an electrocution hazard if you bring it out to points where you can come into physical contact with it.

An electrostatic meter has another useful property. The device does not draw any current, except a tiny initial current needed to put a charge on the plates. Sometimes, an engineer or experimenter doesn't want a measuring device connected to a significant amount of current because so-called *current drain* affects the behavior of the circuit under test. Galvanometers, by contrast, always need some current to produce an indication.



3-3 An electroscope can detect the presence of an electrostatic charge.

If you have access to a laboratory electroscope, try charging it up with a glass rod after rubbing the rod against a dry cloth. When you pull the rod away from the electroscope, the foil leaves will remain standing apart. The charge "sits there" on the foil, trapped! If the electroscope drew any current, the leaves would fall back together again, just as the galvanometer compass needle returns to magnetic north the instant you take the wire away from the battery.

Thermal Heating

Whenever current flows through a substance with a finite, nonzero resistance, the temperature of that substance rises. The extent of the temperature increase depends on the current; for any particular sample, more current generates more heat. If we choose a metal or alloy with known physical properties, tailor a wire from that alloy to a certain length and diameter, use a sensitive, accurate thermometer to measure the wire's temperature, and place the entire assembly inside a thermally insulated package, we end up with a *hot-wire meter*. This device allows us to measure AC as well as DC because the heating doesn't depend on the direction of current flow. Hot-wire meters allow for the measurement of AC at frequencies up to several gigahertz.

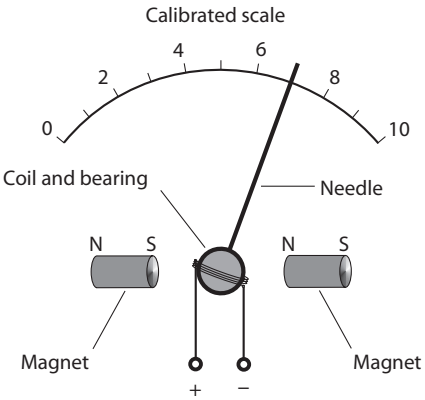
We can take advantage of a variation of the hot-wire principle by placing two different metals (called *dissimilar metals*) into direct contact with each other, forming a boundary called a *junction*. The junction heats up when current flows through it. Engineers call this effect the *thermocouple principle*. As with the hot-wire meter, we can use a thermometer to measure the extent of the heating. The thermocouple principle works in reverse as well. When we apply heat to a thermocouple, it generates DC, which we can measure with a galvanometer. This effect allows us to build an *electronic thermometer*.

Ammeters

A magnetic compass, surrounded by a coil of wire, makes an effective but temperamental current-measuring meter. The compass must lie flat on a horizontal surface. We must align the coil with the compass needle under no-current conditions. We must rotate the compass so that the needle points at the "N" on the scale (that is, 0° *magnetic azimuth*) under no-current conditions. All of these restrictions add up to quite an annoyance for experimenters working in labs among complex electronic systems. You'll hardly ever see a compass galvanometer in a professional engineer's or technician's workshop.

The external magnetic field for a galvanometer need not come from the earth. A permanent magnet, placed near or inside the meter, can provide the necessary magnetic field. A nearby magnet supplies a far stronger magnetic force than does the earth's magnetic field (or *geomagnetic field*), allowing for the construction of a meter that can detect much weaker currents than an old-fashioned galvanometer can. We can orient such a meter in any direction, and slant it any way we want, and it will always work the same way. We can attach the coil directly to the meter pointer, and suspend the pointer from a spring bearing in the field of the magnet. This type of metering scheme, called the *D'Arsonval movement,* has existed for more than a century. Some metering devices still employ it. Figure 3-4 illustrates the functional principle of a D'Arsonval current-measuring meter.

We can fabricate a variation of the D'Arsonval movement by attaching the meter needle to a permanent magnet, and winding the coil in a fixed form around the magnet. Current in the coil produces a magnetic field, which in turn generates a force if the coil and magnet line up correctly with respect to each other. This scheme works okay, but the mass of the permanent magnet results



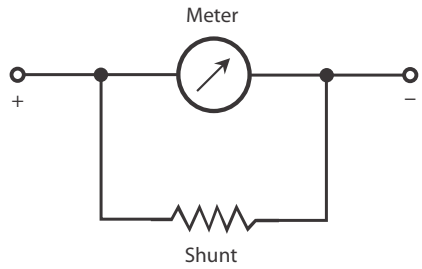
3-4 Functional drawing of a D'Arsonval movement for measuring current.

in a slower needle response, and the meter is more prone to *overshoot* than the true D'Arsonval movement. In overshoot, the inertia of the magnet's mass, once overcome by the magnetic force, causes the needle to fly past the actual point for the current reading, and then to wag back and forth a couple of times before coming to rest in the right place.

Yet another alternative avails itself: We can substitute an *electromagnet* in place of the permanent magnet in a D'Arsonval meter assembly. The electromagnet works with the same current that flows in the coil attached to the meter needle. This arrangement gets rid of the need for a massive, permanent magnet inside the meter. It also eliminates the possibility that the meter sensitivity will change over time if the strength of the permanent magnet deteriorates. Such a demise can result from exposure to heat, severe mechanical vibration, or the mere passage of years.

The sensitivity of any D'Arsonval type meter depends on the amount of current needed to produce a certain force inside the device. That force, in turn, depends on the strength of the permanent magnet (if the meter uses a permanent magnet) and the number of turns in the coil. As the strength of the magnet and/or the number of coil turns increases, the amount of current necessary to produce a given force goes down. In an electromagnet type D'Arsonval meter, the combined number of coil turns affects the sensitivity. If we hold the current constant, the force increases in direct proportion to the number of coil turns. The more magnetomotive force the coils produce, the greater the needle deflection for a given amount of current, and the less current it takes to cause a certain amount of needle movement. The most sensitive D'Arsonval current meters can detect a microampere or two. The amount of current for *full-scale deflection* (the needle goes all the way up without banging against the stop pin) can be as little as about 50 μA in commonly available microammeters.

Sometimes we want an ammeter that will allow for a wide range of current measurements. We can't easily change the full-scale deflection of a meter because that task would require altering the number of coil turns and/or the strength of the magnet inside the assembly. However, all ammeters have a certain amount of *internal resistance* (even though in a well-designed ammeter, the internal resistance is extremely low; in an ideal one, it would be zero). If we connect a resistor, having the same internal resistance as the meter, in parallel with the meter, the resistor will draw half the current while the meter draws the other half. Then it will take twice the current through the assembly to deflect the meter to full scale, as compared with the meter alone. By choosing a resistor of a specific value, we can increase the full-scale deflection of any ammeter by a fixed, convenient factor, such as 10, or 100, or 1000. The resistor must be able to carry the necessary current without



3-5 We can connect a resistor, called a meter shunt, across a current-detecting meter to reduce the meter's sensitivity.

overheating. The resistor might have to deal with practically all of the current flowing through the meter/resistor combination, leaving the meter to carry only 1/10, or 1/100, or 1/1000 of the current. We call a resistor in this application a *shunt* (Fig. 3-5).

Voltmeters

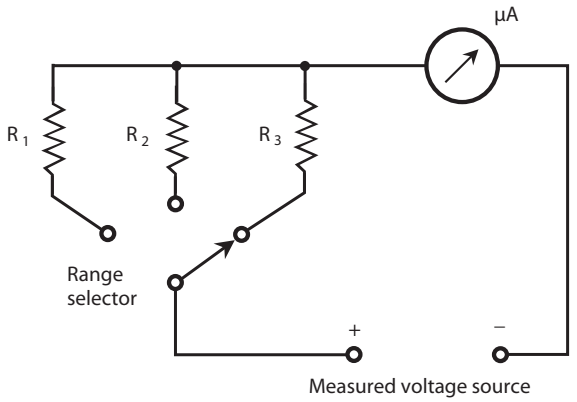
Current, as we have seen, consists of a flow of charge carriers. Voltage, also called electromotive force (EMF) or potential difference, is manifest as "electric pressure" that makes current possible. Given a circuit having a constant resistance, the current through the circuit varies in direct proportion to the voltage across the circuit.

Early experimenters saw that they could use ammeters to measure voltage indirectly. An ammeter acts as a constant-resistance circuit (although the resistance is low). If you connect an ammeter directly across a source of voltage such as a battery, the meter needle deflects. In fact, a milliammeter needle will probably "hit the pin" if you connect it right across a battery; the meter might even suffer permanent damage. (Never connect milliammeters or microammeters directly across voltage sources!) An ammeter, perhaps with a range of 0-10 A, might not deflect to full scale if you place it across a battery, but the meter coil will rapidly drain the battery. Some batteries, such as automotive lead-acid cells, can rupture or explode under these conditions.

Ammeters, as we've learned, have low internal resistance. That's because they're intended for connection in *series* with other parts of a circuit. A circuit under test should "see" an ammeter as a short circuit—ideally like a piece of copper wire—so the meter won't affect the operation of the circuit. Ammeters aren't meant to be connected directly across a source of voltage! However, if you place a large resistor in series with an ammeter, and then connect the combination across a battery or other type of power supply, you no longer have a short circuit. The ammeter will give an indication that varies in direct proportion to the source voltage. The smaller the full-scale reading of the ammeter, the larger the resistance needed to get a meaningful indication on the meter. With a microammeter and a gigantic resistance in series, you can construct a *voltmeter* that will draw only a little current from the source.

You can tailor a voltmeter to have various full-scale ranges by switching different values of resistance in series with the microammeter, as shown in Fig. 3-6. The meter exhibits high internal resistance because the resistors have large ohmic values. As the supply voltage increases, so does the meter's internal resistance because the necessary series resistance increases as the voltage increases.

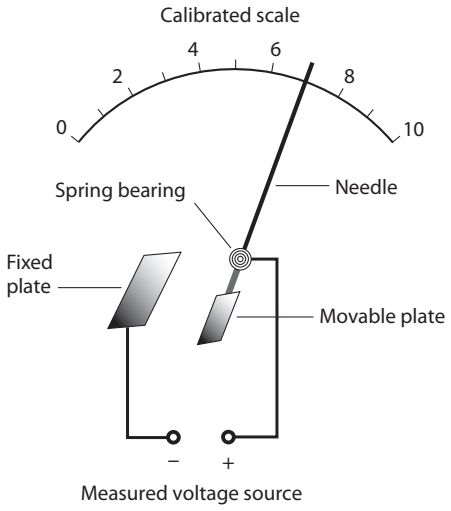
A voltmeter should have high internal resistance; the higher the better. (An ideal voltmeter would have infinite internal resistance.) You don't want the meter to draw significant current from the power source; ideally it wouldn't draw any current at all. You don't want circuit behavior to



3-6 A simple circuit using a microammeter (μA) to measure DC voltage.

change, even a tiny bit, when you connect or disconnect the meter. The less current a voltmeter draws, the less it affects the behavior of anything that operates from the power supply.

An alternative type of voltmeter uses electrostatic deflection, rather than electromagnetic deflection, to produce its readings. Remember that electric fields produce forces, just as magnetic fields do. Therefore, a pair of electrically charged plates attract or repel each other. An *electrostatic voltmeter* takes advantage of the attractive force between two plates having opposite electric charge, or having a large potential difference. Figure 3-7 portrays the functional mechanics of an electrostatic voltmeter. It constitutes, in effect, a sensitive, calibrated electroscope. The device draws essentially no current from the power supply. Nothing but air exists between the plates, and air constitutes a nearly perfect electrical insulator. A properly designed electrostatic meter can indicate AC voltage as well as DC voltage. However, the construction tends to be fragile, and mechanical vibration can influence the reading.



3-7 Functional drawing of an electrostatic voltmeter movement.

Ohmmeters

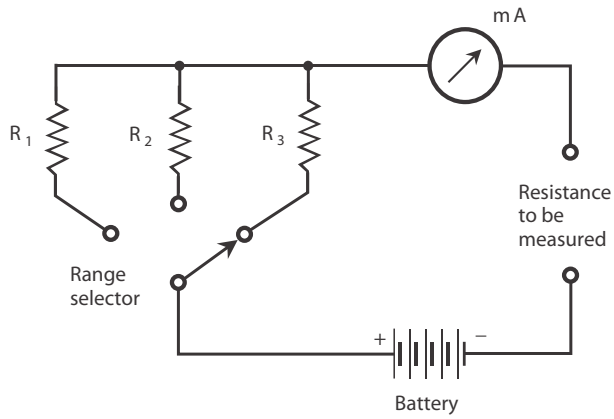
The current through a circuit depends on the resistance, as long as we hold all other factors constant. This principle provides us with a way to measure the DC resistance of a component, device, or circuit.

We can construct an *ohmmeter* by placing a milliammeter or microammeter in series with a set of fixed, switchable resistances and a battery that provides a known, constant voltage, as shown in Fig. 3-8. If we select the resistances carefully, we can get the meter to give us indications in ohms over practically any metering range we want. A typical ohmmeter can quantify resistances from less than 1 Ω to several tens of megohms. We assign the zero point on the meter scale the value of *infinity ohms*, theoretically describing a perfect insulator. The value of the series resistance sets the full-scale meter point to a certain minimum resistance, such as 1 Ω, 10 Ω, 100 Ω, 1 k, 10 k, and so on.

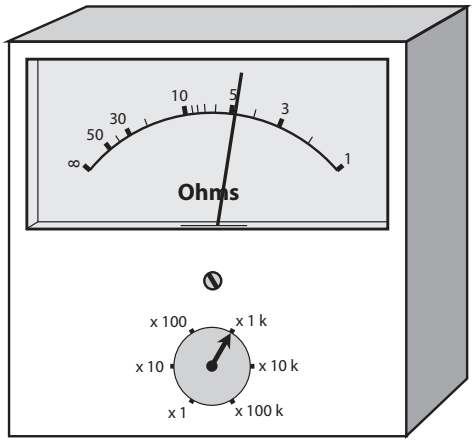
An analog ohmmeter with a D'Arsonval meter "reads backwards." The maximum resistance corresponds to the left-hand end of the meter scale, and the resistance decreases as we move toward the right on the scale. Engineers must calibrate an ohmmeter at the site of manufacture, or else in a well-equipped electronics lab. A slight error in the values of the series resistors can cause gigantic errors in measured resistance. The resistors must have precise *tolerances*. In other words, they must exhibit values close to what the manufacturer claims, to within a fraction of 1 percent if possible. For an ohmmeter to work right, its internal battery must provide a precise, constant, and predictable voltage.

An ohmmeter built from a milliammeter or microammeter always has a *nonlinear scale*. The increments vary in size, depending on where you look on the scale. In most ohmmeters, the graduations "squash together" toward the "infinity" end of the scale. Because of this nonlinearity, you might find it difficult to read the meter for high values of resistance unless you select the optimum meter range by switching the appropriate resistance in series with the meter device.

To use such a meter, connect an ohmmeter in a circuit with the meter set for the highest resistance range first. Then switch the range down until the meter needle comes to rest in a readable



3-8 A circuit using a milliammeter (mA) to measure DC resistance.



3-9 An ohmmeter, in this case showing about 4.7 × 1 k, or 4700 Ω.

part of the scale. After taking down the actual meter reading from the scale, multiply that number by the appropriate amount, as indicated on the range switch. Figure 3-9 shows an ohmmeter reading. The meter itself indicates approximately 4.7, and the range switch says 1 k. This combination indicates a resistance of about 4.7 k, or 4700 Ω.

Ohmmeters give inaccurate readings if a potential difference exists between the points in the circuit to which the meter is connected. The external voltage either adds to, or subtracts from, the ohmmeter's internal battery voltage. Sometimes, in this type of situation, an ohmmeter might tell you that a circuit has "more than infinity" ohms! The needle will hit the pin at the left end of the scale. When you use an ohmmeter to measure DC resistance, you must always make certain that no voltage exists between the points where you intend to connect the meter terminals. You can easily check for such "distracting voltage" using a voltmeter before you use the ohmmeter. If you observe a voltage between those points, you'll probably have to power-down the entire circuit before you try to measure the resistance.

Digital Multimeters

All the measurement devices that have been discussed so far are useful background to what has become the most common measurement device in an electronics lab, that is, the digital multimeter or DMM.

A DMM like the one shown in Fig. 3-10 can measure voltage and current, both AC and DC, to a high degree of precision and accuracy. In addition, you can also select other useful measurement features, such as resistance, capacitance (we will meet this later), frequency, as well as special modes for testing other electronic components such as diodes and even measuring temperature, using a special probe.

It's often said that you shouldn't learn electric guitar on a Fender Stratocaster. Similarly, if you are looking for your first multimeter, don't go for the top of the range. Spending $20 or $30 will get you as very capable measurement device.

DMMs combine analog measurement techniques with digital processing and a microcontroller (tiny computer) along with a digital display (usually LCD) on which the measurements are shown. The input to a DMM uses a high-impedance amplifier, to ensure that the meter itself has very little

3-10 A low-cost digital multimeter.

effect on the quantity that you are trying to measure. Some DMMs have a USB or bluetooth interface so that you can log your readings over time onto your computer. Most DMMs have an input impedance of 10 MΩ or more.

Despite the obvious advantages of DMMs many people still retain an analog multimeter because sometimes the movement of the needle on an electromechanical meter can tell you more about what is going on with something you are measuring than a series of rapidly changing numbers can.

Frequency Counters

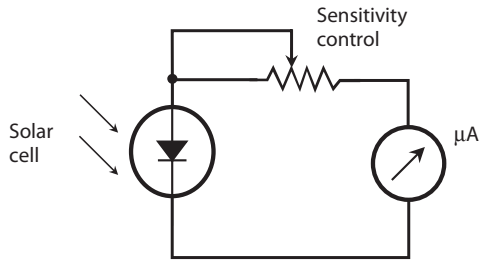
A DMM will often have the ability to measure fairly low frequencies. There are also specialized instruments called frequency counters that operate over a greater range.

A *frequency counter* measures the frequency of an AC wave by counting pulses or cycles, in a manner similar to the way a utility meter counts the number of turns of a motor. The frequency counter works electronically, without any moving parts. It can keep track of thousands, millions, or billions of pulses per second, and it shows the rate on a numeric digital display.

The accuracy of the frequency counter depends on the *lock-in time*, which can vary from a fraction of a second to several seconds. A typical frequency counter allows the user to select from lock-in times of 0.1 second, 1 second, or 10 seconds. Increasing the lock-in time by a factor of 10 causes the accuracy to increase by one additional digit. Modern frequency counters can provide readings accurate to six, seven, or eight digits. Sophisticated lab devices can show frequency to 10 digits or more.

Other Meter Types

Following are brief descriptions of some less common types of meters that you'll occasionally encounter in electricity and electronics.



3-11 A simple light meter with a potentiometer to adjust the sensitivity.

Light Meters

Photographers commonly measure the intensity of visible light rays using a *light meter*, technically called an *illuminometer*. You can make an illuminometer by connecting a microammeter to a solar cell (technically called a *photovoltaic cell*), using a *potentiometer* (variable resistor) to control the meter sensitivity, as shown in Fig. 3-11. More sophisticated devices use DC amplifiers, similar to the type found in a FETVM, to enhance the sensitivity and to allow for several different ranges of readings.

Solar cells don't respond to light at exactly the same wavelengths as human eyes. An engineer would say that the *sensitivity-versus-wavelength function* of a typical solar cell differs from the sensitivity-versus-wavelength function of the human eye. We can overcome this problem by placing a specially designed color filter in front of the solar cell, so that the solar cell becomes sensitive to the same wavelengths, in the same proportions, as our eyes. Illuminometer manufacturers calibrate their products at the factory so that the meter displays visible-light intensity in standard illumination units, such as *lumens* or *candela*.

With appropriate modification, we can use a meter such as the one diagrammed in Fig. 3-11 to roughly determine the intensity of *infrared* (IR) or *ultraviolet* (UV) rays. Various specialized photovoltaic cells exhibit their greatest sensitivity at non-visible wavelengths, including IR and UV.

Oscilloscopes

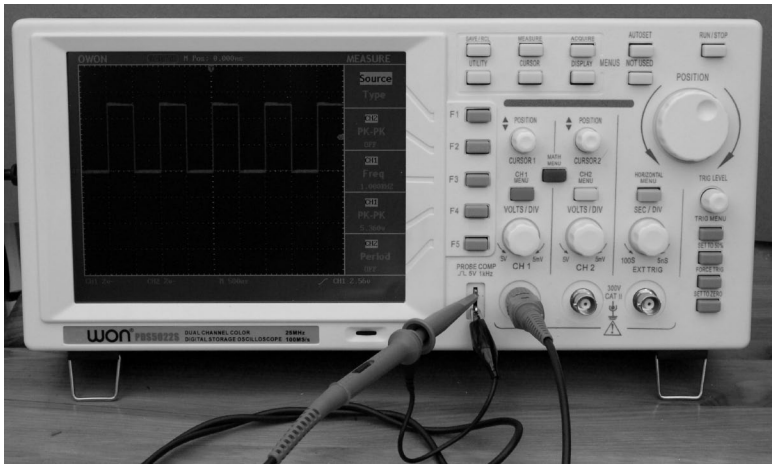
Another graphic metering device, popular with electronics engineers, is the *oscilloscope*, which measures and records quantities that *oscillate* (vary periodically) at rates of hundreds, thousands, or millions of times per second.

An old-fashioned oscilloscope created a "graph" by throwing a beam of electrons at a phosphor screen. A *cathode-ray tube* (CRT), similar to the kind in a television set, was employed. However, modern oscilloscopes like the one shown in Fig. 3-12 digitize the readings and then display them on what amounts to a computer screen. Digital oscilloscopes are generally lower cost than a comparable CRT-based oscilloscope. And indeed CRT-based oscilloscopes are something of a dying breed.

Digital oscilloscopes can normally display two different signals at the same time and even low-cost oscilloscopes will operate at frequencies up to the tens of MHz. Digital storage oscilloscopes (DSS) also retail readings in digital memory, so that you can examine signals at your leisure, after recoding them.

Oscilloscopes are useful for observing and analyzing the shapes of signal waveforms, and also for measuring peak signal levels (rather than only the effective levels). We can use an oscilloscope to indirectly measure the frequency of a waveform. The horizontal scale of an oscilloscope shows time, and the vertical scale shows the instantaneous signal voltage. We can also use an oscilloscope to indirectly measure power or current, by placing a known value of resistance across the input terminals.

Technicians and engineers develop a sense of what a signal waveform "ought to look like." Then they can often ascertain, by observing the oscilloscope display, whether or not the circuit under test is behaving the way it should.



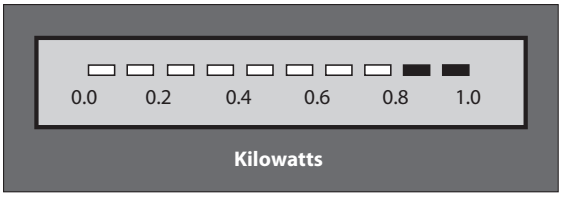
3-12 A digital oscilloscope.

Bar-Graph Meters

A cheap, simple kind of meter comprises a string of *light-emitting diodes* (LEDs) or an LCD along with a digital scale to indicate approximate levels of current, voltage, or power. This type of meter, like a digital meter, has no moving parts. To some extent, it offers the relative-reading feeling you get with an analog meter. Figure 3-13 shows a bar-graph meter designed to indicate the power output, in kilowatts, for a radio transmitter. This meter can follow along fairly well with fluctuations in the reading. In this example, the meter indicates about 0.8 kW, or 800 W.

The chief drawback of bar-graph meters is the fact that most of them don't give precise readings; they can only approximate. For this reason, engineers and technicians rarely use bar-graph meters in a laboratory environment. In addition, the individual LEDs or LCDs in some bar-graph meters flicker intermittently on and off when the signal level "falls between" two values given by the bars. Viewers find this phenomenon distracting or irritating.

3-13 A bar-graph meter. In this case, the indication is about 80 percent of full-scale, representing 0.8 kW or 800 W.



Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

4
CHAPTER

Direct-Current Circuit Basics

IN THIS CHAPTER, YOU'LL LEARN MORE ABOUT CIRCUIT DIAGRAMS, WHICH ENGINEERS AND TECHNICIANS CALL *schematic diagrams,* or simply *schematics,* because they detail the *schemes* for designing and assembling circuits. You'll also learn more about how current, voltage, resistance, and power interact in simple DC circuits.

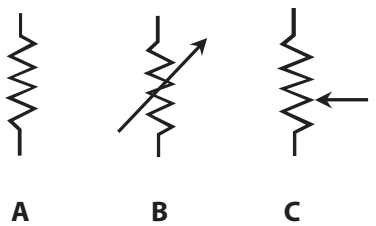
Schematic Symbols

When we want to denote an *electrical conductor* such as a wire, we draw a straight, solid line either horizontally across, or vertically up and down, the page. We can "turn a corner" when we draw a conductor line, but we should strive to minimize the total number of "corners" in a diagram. By following this convention, we can keep schematics neat and ensure that they're easy for others to read.

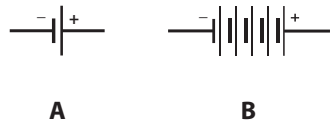
When two conductor lines cross, assume that they *do not* connect at the crossing point unless you see a heavy, solid dot where they meet. Whenever you draw a "connecting dot," make it clearly visible, no matter how many conductors meet at the junction.

We portray a *resistor* by drawing a "zig-zag," as shown in Fig. 4-1A. We portray a two-terminal variable resistor or *potentiometer* by drawing a "zig-zag" with an arrow through it (Fig. 4-1B). We portray a three-terminal potentiometer by drawing a "zig-zag" with an arrow pointing sideways at it (Fig. 4-1C).

We symbolize an *electrochemical cell* by drawing two parallel lines, one longer than the other (Fig. 4-2A). The longer line represents the positive (+) terminal, while the shorter line represents the negative (−) terminal. We symbolize a *battery*, which is a combination of two or more cells in



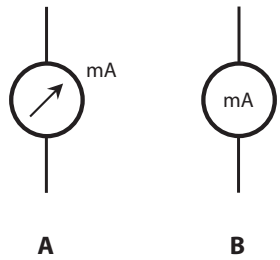
4-1 Schematic symbols for a fixed resistor (A), a two-terminal variable resistor (B), and a three-terminal potentiometer (C).



4-2 Schematic symbols for a single electrochemical cell (A) and a multiple-cell electrochemical battery (B).

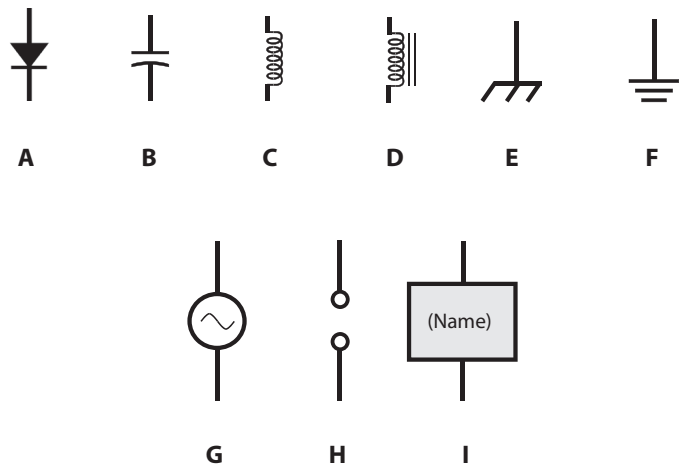
series, by drawing several parallel lines, alternately long and short (Fig. 4-2B). As with the cell, the longer end line represents the positive terminal and the shorter end line represents the negative terminal.

We can portray meters as circles. Sometimes the circle has an arrow inside it, and the meter type, such as mA (milliammeter) or V (voltmeter) is written alongside the circle, as shown in Fig. 4-3A. Sometimes the meter type is denoted inside the circle, and no arrow appears (Fig. 4-3B). It doesn't matter which way you draw meters in your schematics, as long as you keep the style consistent throughout your work.



4-3 Meter symbols can have the designator either outside the circle (A) or inside (B). In this case, both symbols represent a milliammeter (mA).

Some other common symbols include the *incandescent lamp*, the *capacitor*, the *air-core coil*, the *iron-core coil*, the *chassis ground*, the *earth ground*, the *AC source*, the set of *terminals*, and the *black box* (specialized component or device), a rectangle with the designator written inside. These symbols appear in Fig. 4-4.

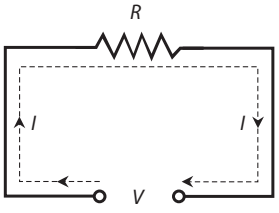


4-4 Schematic symbols for a diode (A), a fixed capacitor (B), a fixed inductor with air core (C), fixed inductor with laminated-iron core (D), chassis ground (E), earth ground (F), a signal generator or source of alternating current (G), a pair of terminals (H), and a specialized component or device (I).

Circuit Simplification

We can simplify most DC circuits to three major components: a voltage source, a set of conductors, and a resistance, as shown in Fig. 4-5. We call the source voltage *V*, the current *I*, and the resistance *R*. The standard units for these components are the volt (V), the ampere (A), and the Ω, respectively. Italicized letters represent mathematical variables (voltage, current, and resistance in this case). Non-italicized characters represent abbreviations for physical units.

We already know that a relationship exists between the voltage, current, and resistance in a DC circuit. If one of these parameters changes, then one or both of the others will also change. If we make the resistance smaller, the current will get larger. If we increase the voltage, the current will also increase. If the current in the circuit increases, the voltage across the resistor will increase. Ohm's law comprises a simple set of formulas defining the relationship between these three quantities.



4-5 Basic elements of a DC circuit with voltage *V*, current *I*, and resistance *R*.

Ohm's Law

Scientists gave Ohm's law its name in honor of *Georg Simon Ohm*, a German physicist who (according to some historians) first expressed it in the 1800s. To calculate the voltage when we know the current and the resistance, use the formula

$$V = IR$$

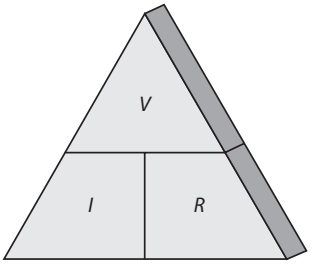
To calculate the current when we know the voltage and the resistance, use

$$I = V/R$$

To calculate the resistance when we know the voltage and the current, use

$$R = V/I$$

You need only remember one of these formulas to derive the other two. You can arrange the three variables geometrically into an *Ohm's law triangle,* as shown in Fig. 4-6. When you want to find the formula for a particular parameter, cover up its symbol and read the positions of the others.



4-6 The Ohm's law triangle showing voltage *V*, current *I*, and resistance *R*, expressed in volts, amperes, and ohms, respectively.

If you want Ohm's law to produce the correct results, you must use the proper units. Under most circumstances, you'll want to use the *standard units* of volts, amperes, and ohms. If you use volts, milliamperes (mA), and ohms, or if you use kilovolts (kV), microamperes (μA), and megohms (M), you can't expect to get the right answers. If you see initial quantities in units other than volts, amperes, and ohms, you should convert to these standard units before you begin your calculations. After you've done all the arithmetic, you can convert the individual units to whatever you like. For example, if you get 13,500,000 Ω as a calculated resistance, you might prefer to call it 13.5 MΩ. But in the calculation, you should use the number 13,500,000 (or 1.35×10^7) and stay with units of ohms.

Current Calculations

In order to determine the current in a circuit, we must know the voltage and the resistance, or be able to deduce them. Figure 4-7 illustrates a generic circuit with a variable DC generator, a voltmeter, some wire, an ammeter, and a potentiometer.

Problem 4-1

Suppose that the DC generator in Fig. 4-7 produces 36 V and we set the potentiometer to a resistance of 18 Ω. What's the current?

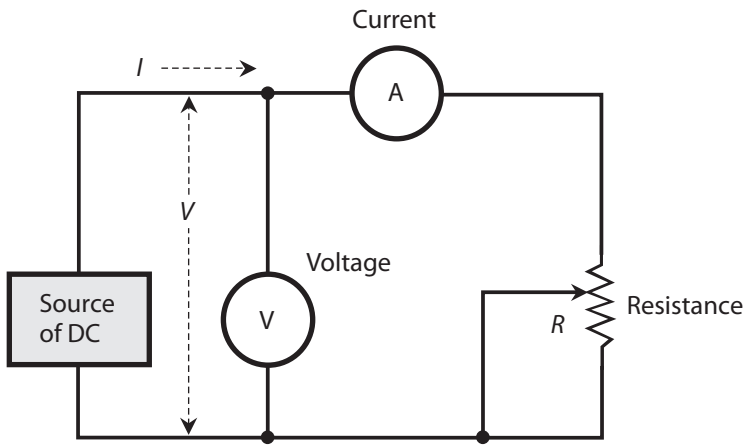
Solution

Use the formula $I = V/R$. Plug in the values for V and R in volts and ohms, getting

$$I = V/R = 36/18 = 2.0 \text{ A}$$

Problem 4-2

Imagine that the DC generator in Fig. 4-7 produces 72 V and the potentiometer is set to 12 k. What's the current?



4-7 A circuit for doing calculations with Ohm's law.

Solution

First, convert the resistance to ohms, getting 12 k = 12,000 Ω. Then input the values in volts and ohms to get

$$I = V/R = 72/12{,}000 = 0.0060 \text{ A} = 6.0 \text{ mA}$$

Problem 4-3

Suppose that we adjust the DC generator in Fig. 4-7 so that it produces 26 kV, and we adjust the potentiometer so that it has a resistance of 13 M. What's the current?

Solution

First, change the resistance value from 13 M to 13,000,000 Ω. Then change the voltage value from 26 kV to 26,000 V. Finally, plug the voltage and resistance into the Ohm's law formula, getting

$$I = V/R = 26{,}000/13{,}000{,}000 = 0.0020 \text{ A} = 2.0 \text{ mA}$$

Voltage Calculations

We can use Ohm's law to calculate the DC voltage between two points when we know the current and the resistance.

Problem 4-4

Suppose we set the potentiometer in Fig. 4-7 to 500 Ω, and we measure the current as 20 mA. What's the DC voltage?

Solution

Use the formula $V = IR$. First, convert the current to amperes: 20 mA = 0.020 A. Then multiply the current by the resistance to obtain

$$V = IR = 0.\ 020 \times 500 = 10 \text{ V}$$

Problem 4-5

We set the potentiometer in Fig. 4-7 to 2.33 k, and we get 250 mA of current. What's the voltage?

Solution

Before doing any arithmetic, we convert the resistance and current to ohms and amperes. A resistance of 2.33 k equals 2330 Ω, and a current of 250 mA equals 0.250 A. Now we can calculate the voltage as

$$V = IR = 0.250 \times 2{,}330 = 582.5 \text{ V}$$

We can round this result up to 583 V.

Problem 4-6

We set the potentiometer in Fig. 4-7 to get 1.25 A of current, and we measure the resistance as 203 Ω. What's the voltage?

Solution

These values are both in standard units. Input them directly to get

$$V = IR = 1.25 \times 203 = 253.75 \text{ V} = 254 \text{ V}$$

We can (and should) round off because we can't justify a result that claims more precision than the data that we start with.

The Rule of Significant Figures

Competent engineers and scientists go by the *rule of significant figures*, also called the *rule of significant digits*. After completing a calculation, we always round the answer off to the *least* number of digits given in the input data numbers.

If we follow this rule in Problem 4-6, we must round off the answer to three significant digits, getting 254 V. That's because the resistance (203 Ω) is specified only to that level of accuracy. If the resistance were given as 203.0 Ω, then we would again round off the answer to 254 V. If the resistance were given as 203.00 Ω, then we could still state the answer to only three significant digits, getting 254 V because we know the current only to three significant digits.

This rule takes some "getting-used-to" if you haven't known about it or practiced it before. But after a while, you'll use it "automatically" without giving it much thought.

Resistance Calculations

We can use Ohm's law to calculate the resistance between two points when we know the voltage and the current.

Problem 4-7

If the voltmeter in Fig. 4-7 reads 12 V and the ammeter shows 2.0 A, what's the resistance of the potentiometer?

Solution

Use the formula $R = V/I$. We can plug in the values directly because they're expressed in volts and amperes. It works out as

$$R = V/I = 12/2.0 = 6.0 \ \Omega$$

Problem 4-8

What's the value of the resistance in Fig. 4-7 if the current equals 24 mA and the voltage equals 360 mV?

Solution

First, convert to amperes and volts, obtaining $I = 0.024$ A and $V = 0.360$ V. Then plug the numbers into the Ohm's law equation to get

$$R = V/I = 0.360/0.024 = 15\ \Omega$$

Problem 4-9

Suppose that the ammeter in Fig. 4-7 reads 175 μA and the voltmeter indicates 1.11 kV. What's the resistance?

Solution

Convert to amperes and volts, getting $I = 0.000175$ A and $V = 1110$ V. Then input these numbers, rounding off to get

$$R = V/I = 1110/0.000175 = 6{,}342{,}857\ \Omega = 6.34\ \text{M}$$

Power Calculations

We can calculate the power P in a DC circuit, such as the one in Fig. 4-7, using the formula

$$P = VI$$

If we aren't given the voltage directly, we can calculate it if we know the current and the resistance. Recall the Ohm's law formula for obtaining voltage:

$$V = IR$$

If we know I and R but we don't know V, we can get the power P as

$$P = VI = (IR)I = I^2R$$

If we know only V and R but don't know I, we can restate I as

$$I = V/R$$

Then we can substitute into the voltage-current power formula to obtain

$$P = VI = V(V/R) = V^2/R$$

Problem 4-10

Suppose that the voltmeter in Fig. 4-7 reads 15 V and the ammeter shows 70 mA. How much power does the potentiometer dissipate?

Solution

Use the formula $P = VI$. First, convert the current to amperes, getting $I = 0.070$ A. (The last 0 counts as a significant digit.) Then multiply by 15 V, getting

$$P = VI = 15 \times 0.070 = 1.05\ \text{W}$$

The input data only has two significant digits, while this answer, as it stands, has three. Rounding up gives 1.1 A. That's the number we should use.

Problem 4-11

If the resistance in the circuit of Fig. 4-7 equals 470 Ω and the voltage source delivers 6.30 V, what's the power dissipated by the potentiometer?

Solution

We don't have to do any unit conversions. Plug in the values directly and then do the arithmetic to get

$$P = V^2/R = 6.30 \times 6.30 \,/\, 470 = 0.0844 \text{ W} = 84.4 \text{ mW}$$

Problem 4-12

Suppose that the resistance in Fig. 4-7 is 33 k and the current is 756 mA. What's the power dissipated by the potentiometer?

Solution

We can use the formula $P = I^2R$ after converting to ohms and amperes: $R = 33{,}000$ and $I = 0.756$. Then calculate and round off to get

$$P = 0.756 \times 0.756 \times 33{,}000 = 18{,}861 \text{ W} = 18.9 \text{ kW}$$

Obviously, a common potentiometer can't dissipate that much power! Most potentiometers are rated at 1 W or so.

Problem 4-13

How much voltage would we need to drive 60.0 μA through 33.0 k?

Solution

These input numbers both have three significant figures because the zeros on the far right are important. (Without them, you'd only have two significant figures in your values.) Use Ohm's law to find the voltage after converting to amperes and ohms, obtaining

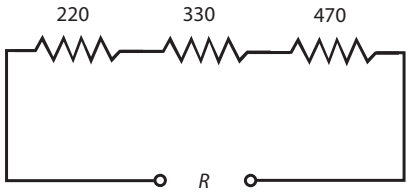
$$V = IR = 0.0000600 \times 33{,}000 = 1.98 \text{ V}$$

Resistances in Series

When we connect two or more resistances in series, their ohmic values add up to get the total (or *net*) resistance.

Problem 4-14

We connect three resistors in series with individual resistances of 220 Ω, 330 Ω, and 470 Ω, as shown in Fig. 4-8. What's the net resistance of the combination?



4-8 Three resistors in series. Illustration for Problem 4-14. All resistances are expressed in Ω.

Solution

Because we know all the values in ohms, we can add without doing any unit conversions to get

$$R = 220 + 330 + 470 = 1020\ \Omega = 1.02\ \text{k}$$

That's the pure theory. But when we build a real-life circuit, the exact resistances depend on the component *tolerances*: how much we should expect the actual values to vary, as a result of manufacturing quirks, from the values specified by the vendor.

Resistances in Parallel

We can evaluate resistances in parallel by considering them as *conductances* instead. Engineers express conductance in units called *siemens*, symbolized S. (The word "siemens" serves both in the singular and the plural sense). When we connect conductances in parallel, their values add up, just as resistances add up in series. If we change all the ohmic values to siemens, we can add these figures up and convert the result back to ohms.

Engineers use the uppercase, italic letter *G* to symbolize conductance as a parameter or mathematical variable. The conductance in siemens equals the reciprocal of the resistance in ohms. We can express this fact using two formulas, assuming that neither *R* nor *G* ever equals zero:

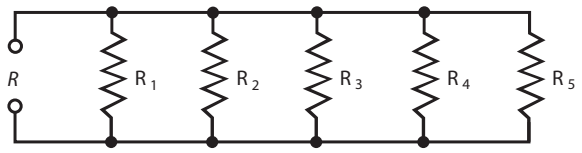
$$G = 1/R$$

and

$$R = 1/G$$

Problem 4-15

Consider five resistors in parallel. Call the resistors R_1 through R_5, and call the total resistance *R*, as shown in Fig. 4-9. Suppose that the individual resistors have values of $R_1 = 10\ \Omega$, $R_2 = 20\ \Omega$, $R_3 = 40\ \Omega$, $R_4 = 50\ \Omega$, and $R_5 = 100\ \Omega$. What's the total resistance *R* of this parallel combination? (Note that we should not italicize R when it means "resistor" as a physical object, but we should italicize *R* when it means "resistance" as in a mathematical variable.)



4-9 Five resistors R_1 through R_5, connected in parallel, produce a net resistance *R*. Illustration for Problems 4-15 and 4-16.

Solution

To solve this problem, start by converting the resistances to conductances by taking their reciprocals. We'll get

$$G_1 = 1/R_1 = 1/10 = 0.10 \text{ S}$$
$$G_2 = 1/R_2 = 1/20 = 0.050 \text{ S}$$
$$G_3 = 1/R_3 = 1/40 = 0.025 \text{ S}$$
$$G_4 = 1/R_4 = 1/50 = 0.020 \text{ S}$$
$$G_5 = 1/R_5 = 1/100 = 0.0100 \text{ S}$$

When we add these numbers, we obtain

$$G = 0.10 + 0.050 + 0.025 + 0.020 + 0.0100 = 0.205 \text{ S}$$

The total resistance, rounded to two significant figures, turns out as

$$R = 1/G = 1/0.205 = 4.9\ \Omega$$

We can calculate the net resistance of a parallel combination directly, but the arithmetic can get messy. Refer again to Fig. 4-9. The resistances combine according to the formula

$$R = 1/(1/R_1 + 1/R_2 + 1/R_3 + 1/R_4 + 1/R_5)$$

Once in a while, you'll encounter a situation where you have multiple resistances in parallel and their values are all equal. In a case of that sort, the total resistance equals the resistance of any one component divided by the number of components. For example, two 80-Ω resistors combine in parallel to yield a net resistance of $80/2 = 40\ \Omega$; four of the same resistors combine in parallel to produce $80/4 = 20\ \Omega$; five of them combine in parallel to give you $80/5 = 16\ \Omega$.

Problem 4-16

We have five resistors R_1 through R_5 connected in parallel, as shown in Fig. 4-9. Suppose that each one of the resistances is 1.800 k. What's the total resistance, *R*, of this combination?

Solution

Here, we can convert the resistances to 1800 Ω and then divide by 5 to get

$$R = 1800/5 = 360.0\ \Omega$$

We're entitled to four significant figures here because we know the input value as stated, 1.800 k, to that many digits. We can treat the divisor 5 as *exact*, accurate to however many significant digits we want because the arrangement contains *exactly five* resistors.

Division of Power

When we connect sets of resistors to a source of voltage, each resistor draws some current. If we know the voltage, we can figure out how much current the entire set demands by calculating the net resistance of the combination, and then considering the combination as a single resistor.

If the resistors in the network all have the same ohmic value, the power from the source divides up equally among them, whether we connect the resistors in series or in parallel. For example, if we have eight identical resistors in series with a battery, the network consumes a certain amount of power, each resistor bearing 1/8 of the load. If we rearrange the circuit to connect the resistors in parallel with the same battery, the network *as a whole* dissipates more power than it does when the resistors are in series, but each *individual* resistor handles 1/8 of the total power, just as when they're in series.

If the resistances in a network do not all have identical ohmic values, then some resistors dissipate more power than others.

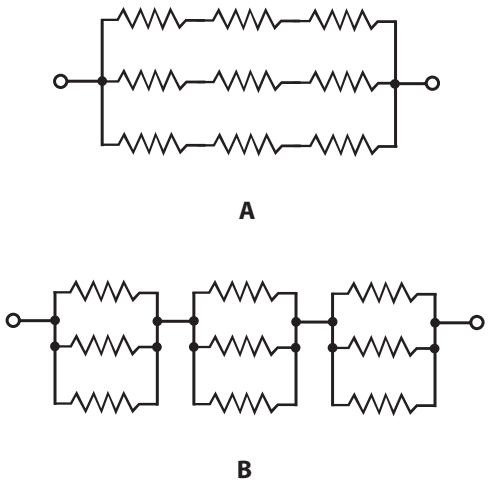
Resistances in Series-Parallel

We can connect sets of resistors, all having identical ohmic values, in parallel sets of series networks, or in series sets of parallel networks. In either case, we get a *series-parallel network* that can greatly increase the total power-handling capacity of the network over the power-handling capacity of a single resistor.

Sometimes, the total resistance of a series-parallel network equals the value of any one of the resistors. This happens if the components are all identical, and are arranged in a network called an *n-by-n* (or $n \times n$) *matrix*. That means when n is a whole number, we have n series-connected sets of n resistors with all the sets connected in parallel (Fig. 4-10A), or else n parallel-connected sets of n resistors, all connected in series (Fig. 4-10B). In practice, these arrangements yield identical results.

A series-parallel array of $n \times n$ resistors, all having identical ohmic values and identical power ratings, has n^2 times the power-handling capability of any resistor by itself. For example, a 3×3 series-parallel matrix of 2 W resistors can handle up to $3^2 \times 2 = 9 \times 2 = 18$ W. If we have a 10×10 array of 1/2 W resistors, then it can dissipate up to $10^2 \times 1/2 = 50$ W. Simply multiply the power-handling capacity of each individual resistor by the total number of resistors in the matrix.

The above-described scheme works if, *but only if*, all of the resistors have identical ohmic values and identical power-dissipation ratings. If the resistors have values and/or ratings that differ (even slightly), one of the components might draw more current than it can withstand, so it will burn out.



4-10 Two examples of series-parallel resistance matrices. At A, sets of series resistors join in parallel. At B, sets of parallel resistances join in series. These examples show symmetrical *n*-by-*n* matrices, where $n = 3$.

Then the current distribution in the network will change, increasing the likelihood that a second resistor will fail. We can end up with a chain reaction of component destruction!

If we need a resistor that can handle 50 W and a certain series-parallel network will handle 75 W, that's fine. But we shouldn't "push our luck" and expect to get away with a network that will handle only 48 W in the same application. We should allow some extra tolerance, say 10 percent over the minimum rating. If we expect the network to dissipate 50 W, we should build it to handle 55 W, or a bit more. We don't have to engage in "overkill," however. We'll waste resources if we build a network that can handle 500 W when we only expect it to cope with 50 W—unless that's the only convenient combination we can cobble together with resistors we have on hand.

Non-symmetrical series-parallel networks, made up from identical resistors, can increase the power-handling capability over that of a single resistor. But in these cases, the total resistance differs from the value of any individual resistor. To obtain the overall power-handling capacity, we can always multiply the power-handling capacity of any individual resistor by the total number of resistors, whether the network is symmetrical or not—again, *if and only if,* all the resistors have identical ohmic values and identical power-dissipation ratings.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

5
CHAPTER

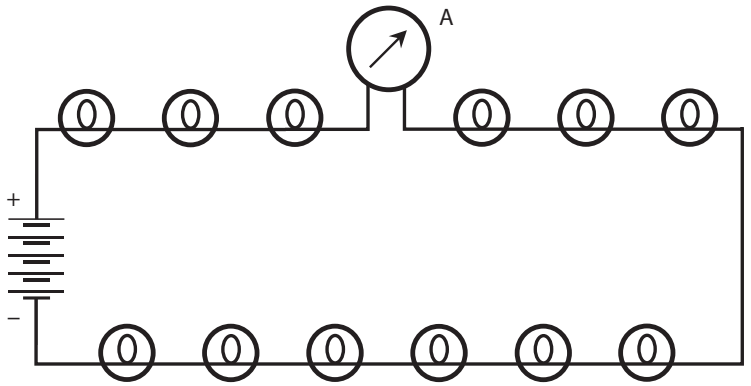
Direct-Current Circuit Analysis

IN THIS CHAPTER, YOU'LL LEARN MORE ABOUT DC CIRCUITS AND HOW THEY BEHAVE UNDER VARIOUS conditions. These principles apply to most AC utility circuits as well.

Current through Series Resistances

If you are old enough to remember the strings of incandescent light bulbs used in festive lighting before LEDs took over, this scenario will explain the annual ritual of finding out which bulbs had failed since the lights were used a year previously. In a series circuit such as a string of light bulbs connected to a DC battery (Fig. 5-1), the current at any given point equals the current at any other point. In this situation, we connect an ammeter A between two of the bulbs. If we move the ammeter to any other point in the current path, it will indicate the same current. This uniformity of current holds true in any series DC circuit, no matter what the components are, and regardless of whether or not they all have the same resistance.

If the bulbs in Fig. 5-1 had different resistances, the current would still be the same at every point in the circuit, but some of the bulbs would consume more power than others. That scenario



5-1 Light bulbs in series, with an ammeter (A) in the circuit.

would likely present a problem; some of the lights would burn brightly and others would hardly glow at all. In a series circuit, even a slight discrepancy in the resistances of the individual components can cause major irregularity in the distribution of power.

Now suppose that one of the bulbs in Fig. 5-1 burns out. The entire string goes dark. We short out the faulty bulb's socket, hoping to get the lights working again. They do, but the current through the chain increases because its overall resistance goes down. Each remaining bulb carries a little more current than it should. Pretty soon, another bulb will burn out because of the excessive current. If we then replace it with a second short circuit, the current will rise still further. We should not be surprised if another bulb blows out almost right away thereafter.

Voltages across Series Resistances

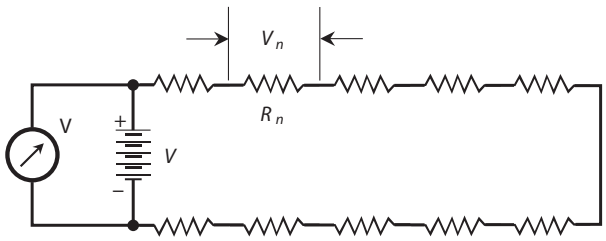
Because the bulbs in the string of Fig. 5-1 all have identical resistances, the potential differences across them are all the same. If we have 12 identical series-connected bulbs in a 120 V circuit, each bulb gets 1/12 of the total, or 10 V. This even distribution of voltage will continue to hold true even if we replace all the bulbs with brighter or dimmer ones, as long as we make sure that all the bulbs in the string are identical.

Examine the schematic diagram of Fig. 5-2. Each resistor carries the same current, whether or not their resistances are all the same. Each resistance R_n has a potential difference V_n across it, equal to the product of the current and the resistance of that particular resistor. The voltages V_n appear in series, so they add up arithmetically. The voltages across all the resistors add up to the supply voltage V. If they did not, a "phantom voltage" would exist somewhere, adding or taking away the unaccounted-for voltage. But that can't happen! Voltage can't come out of nowhere, nor can it vanish into "thin air."

Look at this situation another way. The voltmeter V in Fig. 5-2 shows the voltage V of the battery because we've connected the meter across the battery. The voltmeter V also shows the sum of the voltages V_n across the set of resistances because V is connected across the whole combination. The meter says the same thing whether we think of it as measuring the battery voltage V or as measuring the sum of the voltages V_n across the series combination of resistances. Therefore, V equals the sum of all the voltages V_n.

If you want to calculate the voltage V_n across any particular resistance R_n in a circuit like the one in Fig. 5-2, remember the Ohm's law formula for finding voltage in terms of current and resistance. When you adapt that formula for this situation, you get

$$V_n = IR_n$$



5-2 Analysis of voltages in a series circuit.

where V_n represents the potential difference in volts across the particular resistor, I represents the current in amperes through the whole circuit (and also, therefore, through the resistor of interest), and R_n represents the particular resistor's value in ohms. To determine the current I, you must know the total resistance R (the sum of all the resistances) and the supply voltage V. Then you can use the formula

$$I = V/R$$

If you're astute, you'll notice that you can substitute the value of I from the second of the foregoing formulas into the first one, getting

$$V_n = (V/R)\,R_n = V(R_n/R)$$

This new formula reveals an interesting fact. Each resistor in a series circuit receives voltage in direct proportion to the total supply voltage, and also in direct proportion to the ratio of its resistance to the total resistance. Engineers and technicians exploit these proportional relationships to build circuits called *voltage dividers*.

Problem 5-1

Figure 5-2 shows 10 resistors in series. Suppose that five of the resistors have values of 20 Ω, and the other five have values of 30 Ω. Further suppose that the battery provides 25 V DC. How much potential difference exists across any one of the 20-Ω resistors? How much potential difference exists across any one of the 30-Ω resistors?

Solution

Let's find the total resistance R of the entire series combination, so that we can calculate the current I on the basis of R and the battery voltage V. Once we know the current, we can find the voltage across any individual resistor. We have a total resistance of

$$R = (20 \times 5) + (30 \times 5) = 100 + 150 = 250\ \Omega$$

The current at any point in the circuit is therefore

$$I = V/R = 25/250 = 0.10\ \text{A}$$

If we let $R_n = 20\ \Omega$, we have

$$V_n = IR_n = 0.10 \times 20 = 2.0\ \text{V}$$

and if we let $R_n = 30\ \Omega$, we have

$$V_n = IR_n = 0.10 \times 30 = 3.0\ \text{V}$$

Let's verify the fact that the voltages across all of the resistors add up to the supply voltage. We have five resistors with 2.0 V across each, for a total of 10 V; we have five resistors with 3.0 V across each, for a total of 15 V. The sum of the voltages across the resistors is therefore

$$V = 10 + 15 = 25\ \text{V}$$

Problem 5-2

In the circuit of Fig. 5-2, as described in Problem 5-1 and its solution, what will happen to the voltages across the resistances if we short-circuit three of the 20-Ω resistors and two of the 30-Ω resistors?

Solution

We've replaced three of the 20-Ω resistors with short circuits and two of the 30-Ω resistors with short circuits, leaving us, in effect, with two 20-Ω resistors and three 30-Ω resistors in series. Now we have

$$R = (20 \times 2) + (30 \times 3) = 40 + 90 = 130\ \Omega$$

The current is therefore

$$I = V/R = 25/130 = 0.19\ \text{A}$$

The voltage V_n across any of the "unshorted" 20-Ω resistances R_n is

$$IR_n = 0.19 \times 20 = 3.8\ \text{V}$$

The voltage V_n across any of the "unshorted" 30-Ω resistances R_n is

$$IR_n = 0.19 \times 30 = 5.7\ \text{V}$$

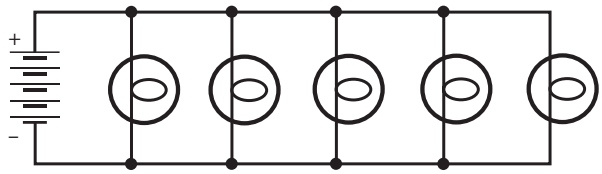
Checking the total voltage, we add and round off to two significant figures, getting

$$V = (2 \times 3.8) + (3 \times 5.7) = 7.6 + 17.1 = 25\ \text{V}$$

Voltage across Parallel Resistances

Imagine a set of light bulbs connected in parallel with a DC battery (Fig. 5-3). If one bulb burns out, we'll have an easy time correcting the problem. In the parallel configuration, only the bad bulb goes dark, so we can identify it immediately. Parallel circuits have another advantage over series circuits. Variations in the resistance of one element have no effect on the power that the other elements receive.

In a parallel circuit, the voltage across each component equals the supply or battery voltage. The current drawn by any particular component depends only on the resistance of that component, and



5-3 Light bulbs in parallel.

not on the resistances of any of the others. In this sense, the components in a parallel-wired circuit operate independently, as opposed to the series-wired circuit in which they interact.

If any one branch of a parallel circuit opens up, is disconnected, or is removed, the conditions in the other branches don't change. If we add new branches to a parallel circuit, assuming the power supply can handle the increased current demand, conditions in previously existing branches remain as they were. Parallel circuits exhibit better overall stability than series circuits. That's why engineers wire nearly all utility circuits in parallel.

Currents through Parallel Resistances

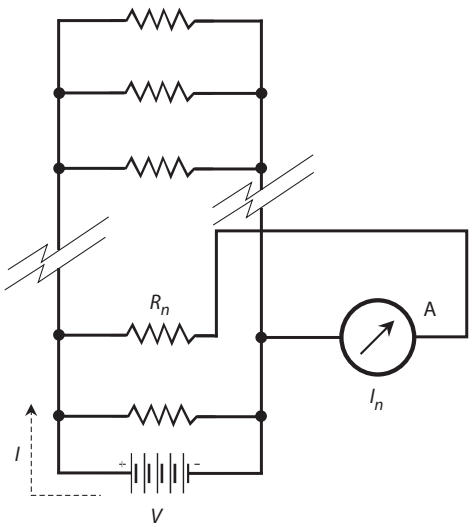
Figure 5-4 illustrates a generic circuit with resistors wired in parallel. Let's call the resistance values R_n, the total circuit resistance R, and the battery voltage V. We can measure the current I_n that flows through any particular branch n, containing resistance R_n, using an ammeter A.

In a parallel circuit such as the one in Fig. 5-4, the sum of all the currents I_n equals the total current I drawn from the power source. In a parallel circuit, the current divides among the individual components, in a manner similar to the way that voltage divides among the components in a series circuit.

Have you noticed that we portray the direction of current flow in Fig. 5-4 as going outward from the positive battery terminal? Don't let this confuse you. Electrons, which constitute most of the charge carriers in an ordinary wire, flow out of the negative terminal of a battery and toward the positive terminal. However, scientists consider *theoretical current*, more often called *conventional current*, as a flow of electricity from positive to negative.

Problem 5-3

Suppose that the battery in Fig. 5-4 delivers 24 V and we have 10 resistors, each with a value of 100 Ω, in the parallel circuit. What's the total current, I, drawn from the battery?



5-4 Analysis of current in a parallel circuit.

Solution

First, let's find the total resistance. Because all of the resistors have the same value, we can divide $R_n = 100$ by 10 to get $R = 10\ \Omega$. Then we can calculate the current as

$$I = V/R = 24/10 = 2.4\ \text{A}$$

Problem 5-4

In the circuit of Fig. 5-4 and described in Problem 5-3 and its solution, we have connected the ammeter A to indicate the current flowing through one particular resistor R_n. What does the meter show?

Solution

We must determine the current I_n that flows in any given branch of the circuit. The voltage equals 24 V across every branch, and every resistor has a value of $R_n = 100\ \Omega$. We can determine the ammeter reading using Ohm's law, obtaining

$$I_n = V/R_n = 24/100 = 0.24\ \text{A}$$

Because we have a parallel circuit, all of the branch currents I_n should add up to the total current I. Each of the 10 identical branches carries 0.24 A, so the total current is

$$I = 0.24 \times 10 = 2.4\ \text{A}$$

This result agrees with the solution to Problem 5-3, as we would expect.

Problem 5-5

We connect three resistors in parallel across a battery that supplies $V = 12$ V. The resistance values are $R_1 = 24\ \Omega$, $R_2 = 48\ \Omega$, and $R_3 = 60\ \Omega$. These resistances carry currents I_1, I_2, and I_3, respectively. What is the current I_2 through R_2?

Solution

Ohm's law allows us to solve this problem as if R_2 were the only resistance in the circuit. We need not worry about the fact that it's part of a parallel combination; the other branches don't affect I_2. Therefore

$$I_2 = V/R_2 = 12/48 = 0.25\ \text{A}$$

Problem 5-6

How much current does the entire parallel combination of resistors, as described in Problem 5-5, draw from the battery?

Solution

We can approach this problem in two different ways. The first method involves finding the total resistance R of the parallel combination, and then calculating the total current I based on R. The second method involves calculating the individual currents I_1, I_2, and I_3 through R_1, R_2, and R_3, respectively, and then adding them up.

Using the first method, we begin by changing the resistances R_n into conductances G_n. This gives us

$$G_1 = 1/R_1 = 1/24 = 0.04167 \text{ S}$$
$$G_2 = 1/R_2 = 1/48 = 0.02083 \text{ S}$$
$$G_3 = 1/R_3 = 1/60 = 0.01667 \text{ S}$$

When we add these values together, we see that the net conductance of the parallel combination is $G = 0.07917$ S. The net resistance is the reciprocal of this value, or

$$R = 1/\text{G} = 1/0.07917 = 12.631\ \Omega$$

We can use Ohm's law to find

$$I = V/R = 12/12.631 = 0.95 \text{ A}$$

We kept some extra digits throughout the calculation, rounding off at the end of the process, in order to minimize *cumulative rounding error*.

Now let's try the second method. We calculate the currents through the individual resistors using Ohm's law as follows:

$$I_1 = V/R_1 = 12/24 = 0.5000 \text{ A}$$
$$I_2 = V/R_2 = 12/48 = 0.2500 \text{ A}$$
$$I_3 = V/R_3 = 12/60 = 0.2000 \text{ A}$$

When we add these current values up, we get a total current of

$$I = I_1 + I_2 + I_3 = 0.\ 5000 + 0.2500 + 0.2000 = 0.95 \text{ A}$$

Again, we kept a few extra digits during the calculation, rounding off to two significant digits at the end.

Power Distribution in Series Circuits

When you want to calculate the power dissipated by resistors connected in series, you can calculate the total current I and then determine the power P_n dissipated by any one of the resistances R_n, using the formula

$$P_n = I^2 R_n$$

The total power (in watts) dissipated in a series combination of resistances equals the sum of the wattages dissipated in each resistance.

Problem 5-7

Imagine a DC series circuit with a power supply that provides 150 V and three resistances $R_1 = 200\ \Omega$, $R_2 = 400\ \Omega$, and $R_3 = 600\ \Omega$. How much power does R_2 dissipate?

Solution

First, let's find the current drawn from the battery by the whole circuit. The resistors are connected in series, so the total resistance is

$$R = 200 + 400 + 600 = 1200\ \Omega$$

According to Ohm's law, the current is

$$I = 150/1200 = 0.125 \text{ A}$$

The power dissipated by R_2 is

$$P_2 = I^2R_2 = 0.125 \times 0.125 \times 400 = 6.25 \text{ W}$$

Problem 5-8

Calculate the total dissipated power P in the circuit of Problem 5-7 using two different methods.

Solution

First, let's figure out the power dissipated by each resistance separately, and then add the figures. We know the power P_2 from the solution to Problem 5-7. We calculated the current as 0.125 A. Now we can calculate using the power formula based on currents and resistances to get

$$P_1 = I^2R_1 = 0.125 \times 0.125 \times 200 = 3.125 \text{ W}$$

and

$$P_3 = I^2R_3 = 0.125 \times 0.125 \times 600 = 9.375 \text{ W}$$

Adding P_1, P_2, and P_3 yields a total power of $P = 3.125 + 6.25 + 9.375 = 18.75$ W, which we should round off to 18.8 W because we have input data accurate to only three significant figures.

The second method involves finding the total series resistance and then calculating the power from that value and the current value we calculated in Problem 5-7. The net resistance of the series combination is $R = 1200\ \Omega$, so

$$P = I^2R = 0.125 \times 0.125 \times 1200 = 18.75 \text{ W}$$

We once again round this value to 18.8 W.

Power Distribution in Parallel Circuits

When we connect resistances in parallel, the currents in the individual resistors can, and often do, differ. The currents will turn out identical, if and only if, all of the resistances are identical. But the voltage across any given resistor always equals the voltage across any other resistor. We can find the power P_n dissipated by any particular resistance R_n using the formula

$$P_n = V^2/R_n$$

where V represents the voltage of the supply. In a parallel DC circuit, just as in a series circuit, the total dissipated wattage equals the sum of the wattages dissipated by the individual resistances.

Problem 5-9

Suppose that a DC circuit contains three resistances $R_1 = 22\ \Omega$, $R_2 = 47\ \Omega$, and $R_3 = 68\ \Omega$ connected in parallel across a battery that supplies a voltage of $V = 3.0$ V. Find the power dissipated by each resistance.

Solution

We can start by finding V^2, the square of the supply voltage. We'll need to use this figure several times:

$$V^2 = 3.0 \times 3.0 = 9.0$$

Resistance R_1 dissipates a power of

$$P_1 = 9.0/22 = 0.4091 \text{ W}$$

which we can round off to 0.41 W. Resistance R_2 dissipates a power of

$$P_2 = 9.0/47 = 0.1915 \text{ W}$$

which we can round off to 0.19 W. Resistance R_3 dissipates a power of

$$P_3 = 9.0/68 = 0.1324 \text{ W}$$

which we can round off to 0.13 W.

Problem 5-10

Find the total consumed power of the resistor circuit in Problem 5-9 using two different methods.

Solution

The first method involves adding the wattages P_1, P_2, and P_3 from the solution to Problem 5-9. If we use the four-significant-digit values, we get

$$P = 0.4091 + 0.1915 + 0.1324 = 0.7330 \text{ W}$$

which rounds off to 0.73 W. The second method involves finding the net resistance R of the parallel combination, and then calculating the power from the net resistance and the battery voltage. (As an exercise, calculate the net resistance yourself.) The net resistance works out to $R = 12.28\ \Omega$, accurate to four significant figures. Now we can calculate the total dissipated wattage as

$$P = V^2/R = 9.0/12.28 = 0.7329 \text{ W}$$

which we can round off to 0.73 W.

It's the Law!

In electricity and electronics, DC circuit analysis always follows certain axioms, or *laws*. The following rules merit your best efforts at memorization.

- In a series circuit, the current is the same at every point.
- In a parallel circuit, the voltage across any resistance equals the voltage across any other resistance, or across the whole set of resistances.

- In a series circuit, the voltages across all the resistances add up to the supply voltage.
- In a parallel circuit, currents through resistances always add up to the total current drawn from the power supply.
- In a series or parallel circuit, the total dissipated wattage equals the sum of the wattages dissipated in the individual resistances.

Now, let's get acquainted with two of the most famous laws that govern DC circuits. These rules have broad scope, allowing us to analyze complicated series-parallel DC networks.

Kirchhoff's First Law

The physicist *Gustav Robert Kirchhoff* (1824–1887) conducted research and experiments before anyone understood much about how electric currents flow. Nevertheless, he used common sense to deduce two important properties of DC circuits.

Kirchhoff reasoned that in any DC circuit, the current going into any point ought to equal the current coming out of that point. This fact, Kirchhoff thought, must hold true no matter how many branches lead into or out of the point. Figure 5-5 shows two examples of this principle.

Examine Fig. 5-5A. At point *X*, the total current *I* going in equals $I_1 + I_2$, the total current coming out. At point *Y*, the total current $I_2 + I_1$ going in equals *I*, the total current coming out. Now examine Fig. 5-5B. At point *Z*, the total current $I_1 + I_2$ going in equals $I_3 + I_4 + I_5$, the total current coming out. We've just seen two examples of *Kirchhoff's First Law*. We can also call it *Kirchhoff's Current Law* or the *current-conservation principle*.

Problem 5-11

Refer to Fig. 5-5A. Suppose that all three resistors have values of 100 Ω, and that $I_1 = 2.0$ A, while $I_2 = 1.0$ A. What is the battery voltage?

Solution

First, let's find the total current *I* that the whole network of resistors demands from the battery. That's an easy task; we simply add the branch currents to get

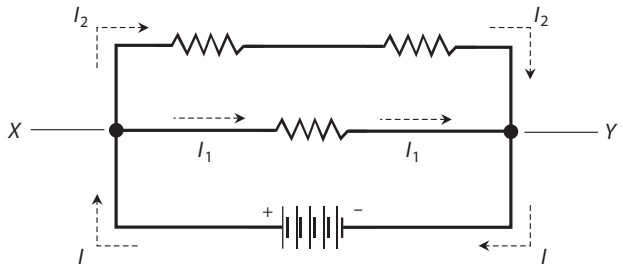
$$I = I_1 + I_2 = 2.0 + 1.0 = 3.0 \text{ A}$$

Next, let's find the resistance of the entire network. The two 100-Ω resistances in series add to give us a net resistance of 200 Ω, which appears in parallel with a third resistance of 100 Ω. You can do the calculations and find that the net resistance *R* across the battery is 66.67 Ω. Then we can use Ohm's law to calculate

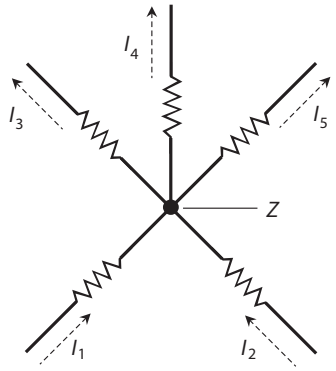
$$V = IR = 3.0 \times 66.67 = 200 \text{ V}$$

Problem 5-12

In the circuit of Fig. 5-5B, imagine that both resistors below point *Z* have values of 100 Ω, and all three resistors above point *Z* have values of 10.0 Ω. Suppose that 500 mA flows through each 100-Ω resistor. How much current flows through any one of the 10.0-Ω resistors? How much potential difference appears across any one of the 10.0-Ω resistors?



A



B

5-5 Kirchhoff's First Law. At A, the total current flowing into point *X* or point *Y* equals the total current flowing out of that point. That is, $I = I_1 + I_2$. At B, the total current flowing into point *Z* equals the current flowing out of point *Z*. That is, $I_1 + I_2 = I_3 + I_4 + I_5$.

Solution

The total current into point *Z* equals 500 mA + 500 mA = 1.00 A. According to Kirchhoff's First Law, this same current emerges from point *Z*, dividing equally three ways among the resistors because their values are all identical (10.0 Ω). The current through any one of those resistors is therefore 1.00/3 A = 0.333 A. We can now determine the potential difference *V* across any one of the 10.0 Ω resistances *R* using Ohm's law, getting

$$V = IR = 0.333 \times 10.0 = 3.33 \text{ V}$$

Kirchhoff's Second Law

The sum of all the voltages, as you go around any DC circuit from some fixed point and return there from the opposite direction, and taking polarity into account, always equals zero. Does this notion seem counterintuitive? If so, let's try to "get into Kirchhoff's head" and figure out what he was thinking.

When Kirchhoff wrote his second law, he must have realized that potential differences can't appear out of nowhere, nor can they vanish into nothingness. If you go around a DC circuit (no matter how complicated) and then return to the point from which you started, the potential

difference between the starting and finishing points *must* equal zero because those two points coincide. It doesn't matter what happens as you go around. The voltage relative to the starting point might rise and fall along the path; it might go positive, then negative, and then positive again. But in the end, all those voltages add up to nought—the potential difference between a point and itself. We call this principle *Kirchhoff's Second Law*, *Kirchhoff's Voltage Law*, or the *voltage-conservation principle*.

If we ignore polarity, then the sum of the voltages across all the individual resistances in a complex DC circuit always adds up to the supply voltage. Kirchhoff's Second Law expands on this principle, taking into account the fact that the polarity of the potential difference across each resistance opposes the polarity of the power supply. Figure 5-6 illustrates an example of Kirchhoff's Second Law. The polarity of the battery voltage V opposes the polarity of the sum of the voltages $V_1 + V_2 + V_3 + V_4$ across the individual resistors. Therefore, when we take polarity into account, we have

$$V + V_1 + V_2 + V_3 + V_4 = 0$$

Problem 5-13

Refer to Fig. 5-6. Suppose that the four resistors have values of $R_1 = 50\ \Omega$, $R_2 = 60\ \Omega$, $R_3 = 70\ \Omega$, and $R_4 = 80\ \Omega$. Also suppose that a current of $I = 500$ mA flows through the circuit. What are the voltages V_1, V_2, V_3, and V_4 across the individual resistors? What's the battery voltage V?

Solution

First, let's find the voltages V_1, V_2, V_3, and V_4 across each of the resistors, using Ohm's law. We should convert the current to amperes, so $I = 0.500$ A. For R_1, we have

$$V_1 = IR_1 = 0.500 \times 50 = 25\ \text{V}$$

For R_2,

$$V_2 = IR_2 = 0.500 \times 60 = 30\ \text{V}$$

For R_3,

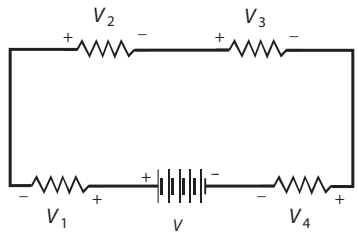
$$V_3 = IR_3 = 0.500 \times 70 = 35\ \text{V}$$

For R_4,

$$V_4 = IR_4 = 0.500 \times 80 = 40\ \text{V}$$

The battery voltage equals the sum

$$V = V_1 + V_2 + V_3 + V_4 = 25 + 30 + 35 + 40 = 130\ \text{V}$$



5-6 Kirchhoff's Second Law. The sum of the voltages across the resistances is equal to, but has opposite polarity from, the supply voltage. Therefore, $V + V_1 + V_2 + V_3 + V_4 = 0$.

Problem 5-14

In the situation shown by Fig. 5-6, suppose that the battery provides $V = 20$ V. Suppose the resistors labeled with voltages V_1, V_2, V_3, and V_4 have ohmic values in the ratio 1:2:3:4, respectively. What's the voltage V_3?

Solution

This problem doesn't provide any information about the current, nor does it tell us the exact resistances. But we need not know these things to solve for V_3. Regardless of the actual ohmic values, the ratio $V_1:V_2:V_3:V_4$ will always turn out the same as long as the resistances remain in the ratio 1:2:3:4. We can, therefore, "plug in" any ohmic values we want for the values of the resistors, as long as they exist in that ratio.

Let R_n be the resistance across which the voltage is V_n, where n can range from 1 to 4. Now suppose that the resistances are as follows:

- The resistance across which V_1 occurs is $R_1 = 1.0\ \Omega$
- The resistance across which V_2 occurs is $R_2 = 2.0\ \Omega$
- The resistance across which V_3 occurs is $R_3 = 3.0\ \Omega$
- The resistance across which V_4 occurs is $R_4 = 4.0\ \Omega$

These resistances follow the required ratio. The total resistance is

$$R = R_1 + R_2 + R_3 + R_4 = 1.0 + 2.0 + 3.0 + 4.0 = 10\ \Omega$$

We can calculate the current through the entire series combination as

$$I = V/R = 20/10 = 2.0\ \text{A}$$

We can now calculate the potential difference V_3, which appears across the resistance R_3, with Ohm's law, obtaining

$$V_3 = IR_3 = 2.0 \times 3.0 = 6.0\ \text{V}$$

Voltage Division

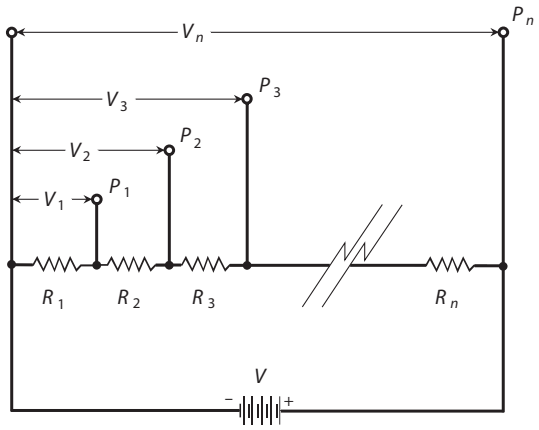
Resistances in series produce ratios of voltages, as we've already seen. We can tailor these ratios to meet certain needs by building *voltage-divider networks*.

When we want to design and build a voltage divider network, we should make the resistors' ohmic values as small as possible without imposing too much current demand on the battery or power supply. When we put the voltage divider to use with a real-world circuit, we don't want the *internal resistance* of that circuit to upset the operation of the divider. The voltage divider "fixes" the intermediate voltages most effectively when the resistance values are as small as the current-delivering capability of the power supply will allow.

Figure 5-7 illustrates the principle of voltage division. The individual resistances are R_1, R_2, R_3, and so on, all the way up to R_n. The total resistance is

$$R = R_1 + R_2 + R_3 + \cdots + R_n$$

The battery voltage equals V, so the current I in the circuit equals V/R. At the various points P_1, P_2, P_3, ..., P_n, let's call the potential differences relative to the negative battery terminal V_1, V_2, V_3, and



5-7 General arrangement for a voltage-divider circuit.

so on, all the way up to V_n. The last voltage V_n equals the battery voltage, V, as we can see by looking at the diagram. All the other voltages are less than V, and ascend in succession. Mathematically, we write this fact as

$$V_1 < V_2 < V_3 < \ldots < V_n = V$$

where the mathematical symbol < means "is less than."

The voltages at the various points increase according to the sum total of the resistances up to each point, in proportion to the total resistance, multiplied by the supply voltage. Therefore, the following equations hold true:

$$V_1 = VR_1/R$$
$$V_2 = V(R_1 + R_2)/R$$
$$V_3 = V(R_1 + R_2 + R_3)/R$$

and so on. This process continues for each of the voltages at points all the way up to

$$V_n = V(R_1 + R_2 + R_3 + \cdots + R_n)/R = VR/R = V$$

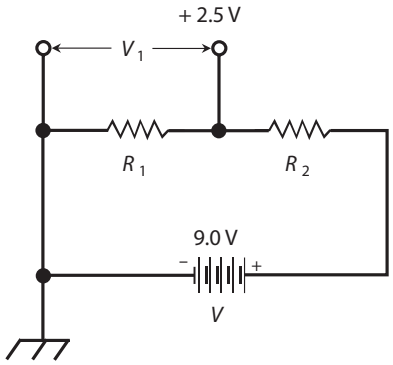
Problem 5-15

Imagine that we've constructed an electronic device to do a specific task. Its battery supplies a voltage of $E = 9.0$ V. We connect its negative terminal to *common ground* (also called *chassis ground*). We want to build a voltage-divider network to obtain a terminal point where the DC voltage equals +2.5 V with respect to common ground. Provide an example of a pair of DC resistance values that we can connect in series so that +2.5 V appears at the point between them, as long as we don't connect any external circuit to the network.

Solution

Figure 5-8 illustrates the network. There exist infinitely many different combinations of resistances that will do our job. As long as we don't connect the voltage divider to any external circuit, we will always observe that, at the terminal point between the resistors,

$$R_1/(R_1 + R_2) = V_1/V$$



5-8 A voltage divider network that derives 2.5-V DC from a 9.0-V DC source.

Let's choose a total value of

$$R_1 + R_2 = 100\ \Omega$$

In this case, we want to obtain a voltage $V_1 = +2.5$ V at the terminal point between the resistors. That means we want to have

$$V_1/V = 2.5/9.0 = 0.28$$

We must make the ratio of R_1 to the overall resistance equal to the voltage ratio, so that

$$R_1/(R_1 + R_2) = 0.28$$

We've chosen to construct our network such that the total resistance, $R_1 + R_2$, equals 100 Ω. Substituting 100 for $(R_1 + R_2)$ in the above formula, we get

$$R_1/100 = 0.28$$

which solves to

$$R_1 = 28\ \Omega$$

so therefore

$$R_2 = 100 - 28 = 72\ \Omega$$

In a practical situation, we'll want to choose the smallest possible value for R. This might be more or less than 100 Ω, depending on the nature of the circuit and the current-delivering capability of the battery. The *actual resistance values* don't determine the voltage at the terminal point; their *ratio* does.

Problem 5-16

How much current I, in milliamperes, flows through the series resistances in the situation described in Problem 5-15 and its solution?

Solution

Using the Ohm's law formula for current in terms of resistance, we obtain

$$I = V/(R_1 + R_2) = 9.0/100 = 0.090\ \text{A} = 90\ \text{mA}$$

Problem 5-17

Imagine that we want the voltage-divider network of Fig. 5-8 to draw 600 mA to ensure the proper operation of the device that we connect across R_1. Suppose that the battery provides $V = 9.00$ V. We want +2.50 V to appear at the terminal point between the resistors. What values for R_1 and R_2 should we use?

Solution

Let's calculate the total resistance first, using Ohm's law. Converting 600 mA to amperes, we get $I = 0.600$ A. Then

$$R_1 + R_2 = V/I = 9.00/0.600 = 15.0\ \Omega$$

We want our resistance values to exist in the ratio

$$R_1/R_2 = 2.50/9.00 = 0.280$$

We should therefore choose

$$R_1 = 0.280 \times 15.0 = 4.20\ \Omega$$

and

$$R_2 = 15.0 - 4.2\ 0 = 10.8\ \Omega$$

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

6
CHAPTER

Resistors

ALL ELECTRICAL COMPONENTS, DEVICES, AND SYSTEMS EXHIBIT RESISTANCE. IN PRACTICE, PERFECT conductors don't exist. You've seen some examples of circuits containing components designed to reduce or limit the current. We call these components *resistors*.

Purpose of the Resistor

Resistors play diverse roles in electrical and electronic equipment, despite the fact that their only direct action constitutes interference with the flow of current. Common applications include the following:

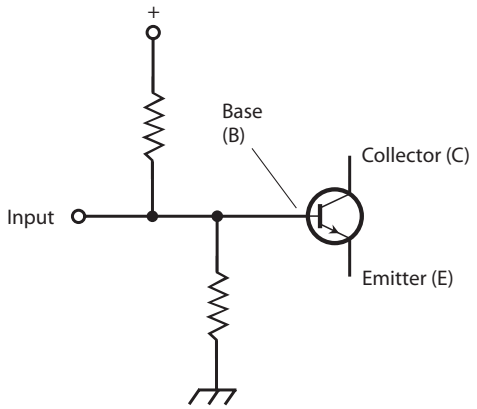
- Voltage division
- Biasing
- Current limiting
- Power dissipation
- Bleeding off charge
- Impedance matching

Voltage Division

You've learned how to build voltage dividers using resistors. The resistors dissipate some power in doing this job, but the resulting potential differences ensure that an external circuit or system operates properly. For example, a well-engineered voltage divider allows an amplifier to function efficiently, reliably, and with a minimum of distortion.

Biasing

In a *bipolar transistor*, or a *field-effect transistor*, the term *bias* means that we deliberately apply a certain DC voltage to one point on the circuit relative to another, or relative to electrical ground. Networks of resistors can accomplish this function.



6-1 A pair of resistors can act as a voltage divider in a transistor circuit.

A radio transmitting amplifier works with a different bias than an oscillator or a low-level receiving amplifier. Sometimes we must build low-resistance voltage dividers to bias a transistor; in some cases a single resistor will do.

Figure 6-1 shows a bipolar transistor that obtains its bias from a pair of resistors in a voltage-divider configuration. We'll learn about the transistor electrodes, called the *emitter* (E), the *base* (B), and the *collector* (C), later in this course.

Current Limiting

One common use of a resistor to limit current is when using an LED. LEDs have a forward voltage that does not vary much on how much current is flowing through the LED. Typically for a red LED this might be 1.8 V. So, if you were to connect such an LED directly to say a 5-V supply, the LED would draw too much current, overheat, and burn out.

Figure 6-2 shows a current-limiting resistor connected in series with an LED so that, following Kirchhoff's current law, whatever current flows through the resistor will also flow through the LED. Since the LED forward voltage and the supply voltage are fixed at 1.8 V and 5 V respectively, we can pick a value of resistor to set the LED current (typically 20 mA).

$$R = V/I = (5 - 1.8)/0.02 = 160\ \Omega$$

Power Dissipation

In some applications, we want a resistor to dissipate power as heat. Such a resistor might constitute a "dummy" component, so that a circuit "sees" the resistor mimic the behavior of something more complicated.

6-2 A resistor to limit the current through an LED.

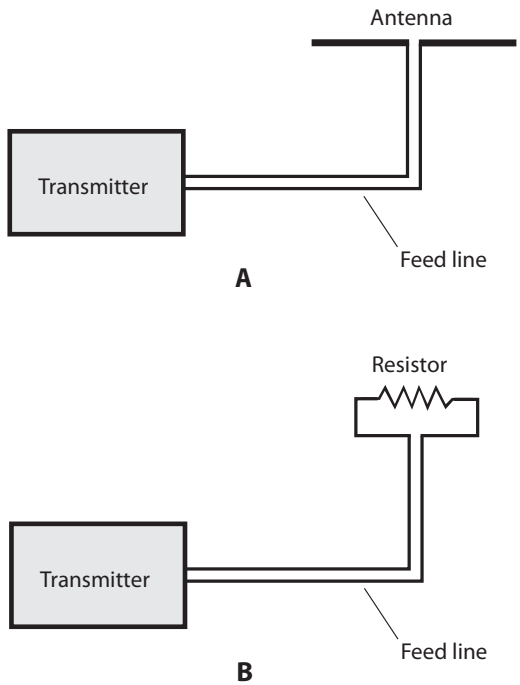
When testing a radio transmitter, for example, we can install a massive resistor in place of the antenna (Fig. 6-3). This engineering trick allows us to test the transmitter for long periods at high power levels without interfering with on-the-air communications. The transmitter output heats the resistor without radiating any signal. However, the transmitter "sees" the resistor as if it were a real antenna—and a perfect one, too, if the resistor has the correct ohmic value.

We might take advantage of a resistor's power-dissipating ability at the input of a power amplifier, such as the sort used in hi-fi audio equipment. Sometimes the circuit *driving* the amplifier (supplying its input signal) produces too much power. A resistor, or network of resistors, can dissipate the excess power so that the amplifier doesn't receive too much input signal. In any type of amplifier, *overdrive* (an excessively strong input signal) can cause distortion, inefficiency, and other problems.

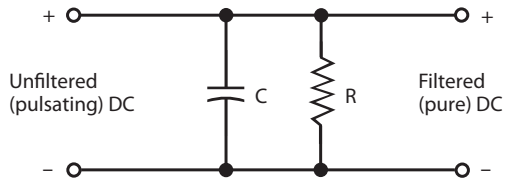
Bleeding Off Charge

A high-voltage, DC power supply employs capacitors (sometimes along with other components) to smooth out the current pulsations, known as *ripple*. These *filter capacitors* acquire an electric charge and store it for a while. In some power supplies, filter capacitors can hold the full output voltage of the supply, say something like 750 V, for a long time even after we power the whole system down. Anyone who attempts to repair or test such a power supply can receive a deadly shock from this voltage.

If we connect a *bleeder resistor* in parallel with each individual filter capacitor in a power supply, the resistors will drain the capacitors' stored charge, sparing personnel who service or test the power supply the risk of electrocution. In Fig. 6-4, the bleeder resistor R should have a value high



6-3 At A, a radio transmitter connected to a real antenna. At B, the same transmitter connected to a resistive "dummy" antenna.



6-4 A bleeder resistor (R) connected across the filter capacitor (C) in a power supply.

enough so that it doesn't interfere with the operation of the power supply, but low enough so that it will discharge the capacitor C in a short time after power-down.

Even if a power supply has bleeder resistors installed, the wise engineer or technician, wearing heavy, insulated gloves, will short out all filter capacitors, using a screwdriver or other metal tool with an insulated handle, after power-down and before working on the circuit. Even if the supply has bleeder resistors, they can take awhile to get rid of the residual charge. Besides that, bleeder resistors sometimes fail!

Impedance Matching

We encounter a more sophisticated application for resistors in the *coupling* between two amplifiers, or in the input and output circuits of amplifiers. In order to produce the greatest possible amplification, the *impedances* between the output of a given amplifier and the input of the next must precisely agree. The same holds true between a source of signal and the input of an amplifier. The principle also applies between the output of an amplifier and a *load,* whether that load is a speaker, a headset, or whatever. We might think of impedance as the AC "big brother" of DC resistance. You'll learn about impedance in Part 2 of this book.

Fixed Resistors

In your engineering adventures, you'll find that resistors are the component that you will use most often.

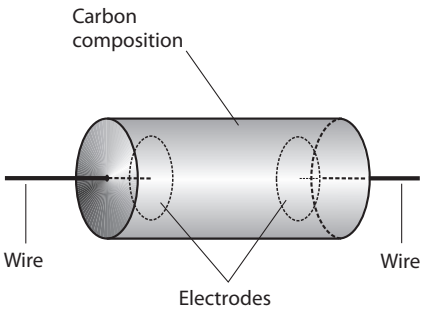
In the following discussion on fixed resistors, the figures depict the cylindrical design of so-called through-hole components, designed, to be soldered to copper pads on the underside of a circuit board. These days, most resistors are so-called "chip resistors" or surface mount device (SMD) resistors. These tiny resistors are generally "film" resistors and work the same way as their through-hole relatives.

Here are the most common types of resistor.

Carbon-Composition Resistors

The cheapest method of making a resistor involves mixing powdered carbon (a fair electrical conductor) with a nonconductive solid or paste, pressing the resulting clay-like "goo" into a cylindrical shape, inserting wire leads in the ends, and then letting the whole mass harden (Fig. 6-5). The resistance of the final product depends on the ratio of carbon to the nonconducting material, and also on the physical distance between the wire leads. This process yields a *carbon-composition resistor*.

You'll find carbon-composition resistors in a wide range of ohmic values. This kind of resistor is *nonreactive*, meaning that it introduces almost pure resistance into the circuit, and essentially no *inductive reactance* or *capacitive reactance*. (You'll learn more about both types of reactance later in this book.) This property makes carbon-composition resistors useful in the construction of radio receivers and transmitters, where the slightest extraneous reactance can cause trouble.



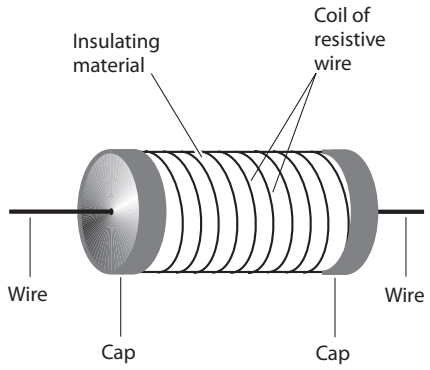
6-5 Construction of a carbon-composition resistor.

Carbon-composition resistors dissipate power in proportion to their physical size and mass. Most of the carbon-composition resistors you see in electronics stores can handle ¼ W or ½ W. You can also find ⅛ W units for use in miniaturized, low-power circuitry, and 1 W or 2 W units for circuits that require electrical ruggedness. Occasionally, you'll see a carbon-composition resistor with a power rating, such as 50 or 60 W, but not often.

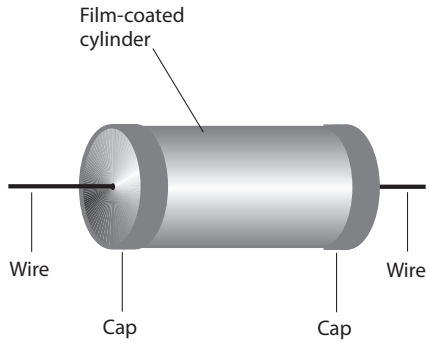
Wirewound Resistors

We can also obtain resistance with a length of wire made from poorly conducting material. The wire can take the form of a coil wound around a cylindrical form, as shown in Fig. 6-6. The resistance depends on how well the wire conducts, on its diameter or *gauge*, and on its total stretched-out length. When we construct a component in this fashion, we have a *wirewound resistor*.

Wirewound resistors usually have low to moderate values of resistance. You'll find wirewound resistors whose ohmic values fall within a very narrow range, sometimes a fraction of 1 percent either way of the quoted value. We say that such components have *close tolerance* or *tight tolerance*. Some wirewound resistors can handle large amounts of power. That's their best asset. On the downside, wirewound resistors always exhibit some inductive reactance because of their coiled geometry, making them a poor choice for use in situations where high-frequency AC or radio-frequency (RF) current flows.



6-6 Construction of a wirewound resistor.



6-7 Construction of a film-type resistor.

Film-Type Resistors

We can apply carbon paste, resistive wire, or some mixture of ceramic and metal to a cylindrical form as a film, or thin layer, to obtain a specific resistance. When we do this, we get a *carbon-film resistor* or *metal-film resistor*. Superficially, the component looks like a carbon-composition resistor, but the construction differs (Fig. 6-7). Film resistors are by far the most common type of resistor.

The cylindrical form consists of an insulating substance, such as porcelain, glass, or thermoplastic. The film can be deposited on this form by various methods, and the value tailored as desired. Metal-film resistors can be manufactured to extremely close tolerances. Film-type resistors usually have low to medium-high resistance.

Film-type resistors, like carbon-composition resistors, have little or no inductive reactance—a big asset in high-frequency AC applications. However, film-type resistors generally can't handle as much power as carbon-composition or wirewound types of comparable physical size.

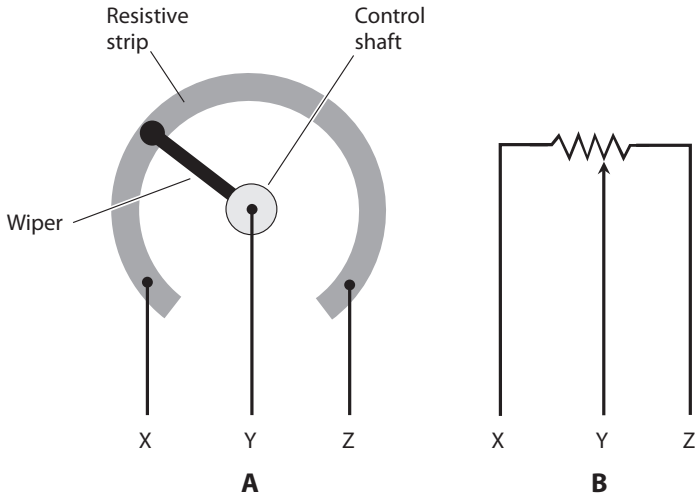
The Potentiometer

Figure 6-8A illustrates the construction geometry of a *potentiometer*, which acts as a variable resistor. Figure 6-8B shows the schematic symbol. A resistive strip, similar to that found on film-type fixed resistors, forms approximately ¾ of a circle (an arc of 270°), with terminals connected to either end. This strip exhibits a fixed value of resistance. To obtain the variable resistance, a sliding contact, attached to a rotatable shaft and bearing, goes to a third (middle) terminal. The resistance between the middle terminal and either end terminal can vary from zero up to the resistance of the whole strip. Most potentiometers can handle only low levels of current, at low to moderate voltages. You'll encounter two major designs in your electronics work: the *linear-taper potentiometer* and the *audio-taper potentiometer*.

Linear-Taper Potentiometer

A linear-taper potentiometer uses a strip of resistive material with constant density all the way around. As a result, the resistance between the center terminal and either end terminal changes at a steady rate as the control shaft rotates. Engineers usually prefer linear-taper potentiometers in electronic test instruments. Linear-taper potentiometers also exist in some consumer electronic devices.

Suppose that a linear-taper potentiometer has a value of 0 to 270 Ω. In most units, the shaft rotates through a total angular displacement of about 270°. The resistance between the center and



6-8 Simplified functional drawing of a rotary potentiometer (A), and the schematic symbol (B).

one end terminal increases along with the number of angular degrees that the shaft turns away from that end terminal. The resistance between the center and the other end terminal equals 270 minus the number of degrees that the control shaft subtends with respect to that terminal. The resistance between the middle terminal and either end terminal, therefore, constitutes a *linear function* of the angular shaft position.

Audio-Taper Potentiometer

In some applications, linear taper potentiometers don't work well. The volume control of a radio receiver or hi-fi audio amplifier provides a good example. Humans perceive sound intensity according to the *logarithm* of the actual sound power, not in direct proportion to the actual power. If you use a linear-taper potentiometer to control the volume (or *gain*) of a radio receiver or audio system, the sound volume (as you hear it) will change slowly in some parts of the control range, and rapidly in other parts of the control range. The device will work, but not in a user-friendly fashion.

An audio-taper potentiometer, if properly selected and correctly installed, can compensate for the way in which people perceive sound levels. The resistance between the center and either end terminal varies according to a *nonlinear function* of the angular shaft position. Some engineers call this type of device a *logarithmic-taper potentiometer* or *log-taper potentiometer* because the function of resistance versus angular displacement follows a logarithmic curve. As you turn the shaft, the sound intensity seems to increase in a linear manner, even though the actual power variation is logarithmic.

Slide Potentiometer

A potentiometer can employ a straight strip of resistive material rather than a circular strip so that the control moves up and down, or from side to side, in a straight line. This type of variable resistor, called a *slide potentiometer*, finds applications in hi-fi audio *graphic equalizers*, as gain controls in some amplifiers, and in other applications where operators prefer a straight-line control movement to a rotating control movement. Slide potentiometers exist in both linear-taper and audio-taper configurations.

Resistor Specifications

When we select a resistor for a particular application, we must obtain a unit that has the correct properties, or *specifications*. Here are some of the most important specifications to watch for.

Ohmic Value

In theory, a resistor can have any ohmic value from the lowest possible (such as a shaft of solid silver) to the highest (dry air). In practice, we'll rarely find resistors with values less than about 0.1 Ω or more than about 100 M.

Not every possible ohmic value is available for resistors. Instead they are available in various ranges. The most common range is called E12, beacuse there are 12 values available for each power of 10. In the case of E12, these are:

{1.0, 1.2, 1.5, 1.8, 2.2, 2.7, 3.3, 3.9, 4.7, 5.6, 6.8, 8.2}

As well as the E12 range, there are also E24, E48, E96, and even E192 ranges. But it is rare to need a more specific value of resistor than that available on the E12 range.

Tolerance

The E12 range above represents standard resistance values available in tolerances of plus or minus 10 percent (±10%). This means that the resistance might be as much as 10% more or less than the indicated amount. In the case of a 470-Ω resistor, for example, the value can be larger or smaller than the rated value by as much as 47 Ω, and still adhere to the rated tolerance. That's a range of 423 to 517 Ω.

In practice, modern manufacturing techniques mean that tolerances are actually much tighter and more likely to be ±2% even for an E12 resistor.

Engineers calculate resistor tolerance figures on the basis of the *rated* resistance, not the measured resistance. For example, we might test a "470-Ω" resistor and find it to have an actual resistance of 427 Ω; this discrepancy would still put the component within ±10% of the specified value. But if we test it and find it to have a resistance of 420 Ω, its actual value falls outside the rated range, so it constitutes a "reject."

For applications requiring exceptional precision, resistors exist that boast tolerances tighter than ±0.2%. We might need a resistor of such quality in a circuit or system where a small error can make a big difference.

Power Rating

A manufactured resistor always bears a specification that tells us how much power it can safely dissipate. The dissipation rating indicates *continuous duty*, which means that the component can dissipate a certain amount of power constantly and indefinitely.

We can calculate how much current a given resistor can handle using the formula for power P (in watts) in terms of current I (in amperes) and resistance R (in ohms), as follows:

$$P = I^2R$$

With algebra, we can change this formula to express the maximum allowable current in terms of the power dissipation rating and the resistance, as follows:

$$I = (P/R)^{1/2}$$

where the ½ power represents the square root.

We can effectively multiply the power rating for a given resistor by connecting identical units in series-parallel matrices of 2 × 2, 3 × 3, 4 × 4, or larger. If we need a 47-Ω, 45-W resistor but we have only a lot of 47-Ω, 1-W resistors available, we can connect seven sets of seven resistors in parallel (a 7 × 7 series-parallel matrix) and get a 47-Ω resistive component that can handle up to 7 × 7 W, or 49 W.

Resistor power dissipation ratings, like the ohmic values, are specified with a margin for error. A good engineer never tries to "push the rating" and use, say, a ¼-W resistor in a situation where it will need to draw 0.27 W. In fact, good engineers usually include their own safety margin, in addition to that offered by the vendor. Allowing a 10% safety margin, for example, we should never demand that a ¼-W resistor handle more than about 0.225 W, or expect a 1-W resistor to dissipate more than roughly 0.9 W.

The only routes a resistor has for dissipating its heat are air circulating around it, and to some extent transfer of heat to the circuit board that the resistor is soldered to. So, if the resistor is in a poorly ventilated enclosure further allowance should be made.

Temperature Compensation

All resistors change value when the temperature rises or falls dramatically. Because resistors dissipate power by design, they get hot in operation. Sometimes the current that flows through a resistor does not rise high enough to appreciably heat the component. But in some cases it does, and the heat can cause the resistance to change. If this effect becomes great enough, a sensitive circuit will behave differently than it did when the resistor was still cool. In the worst-case scenario, an entire device or system can shut down because of a single "temperamental" resistor.

Resistor manufacturers do various things to prevent problems caused by resistors changing value when they get hot. In one scheme, resistors are specially manufactured so that they don't appreciably change value when they heat up. We call these components *temperature compensated*. As you might expect, a temperature-compensated resistor can cost several times as much as an ordinary resistor.

Rather than buy a single temperature-compensated resistor, we can employ a single resistor or a series-parallel matrix of resistors with a power rating several times higher than we ever expect the component to dissipate. This technique, called *over-engineering*, keeps the resistor or matrix from reaching temperatures high enough to significantly change the resistance. Alternatively, we might take several resistors, say five of them, each with five times the intended resistance, and connect them all in parallel. Or we can take several resistors, say four of them, each with about ¼ of the intended resistance, and connect them in series.

Whatever trick we employ to increase the power-handling capability of a component, we should never combine resistors with different ohmic values or power ratings into a single matrix. If we try that, then one of them might end up taking most of the load while the others "loaf," and the combination will perform no better than the single hot resistor we started with. Whenever we want to build a "resistor gang" to handle high current or keep cool under load, we should always procure a set of *identical* components.

Are You Astute?

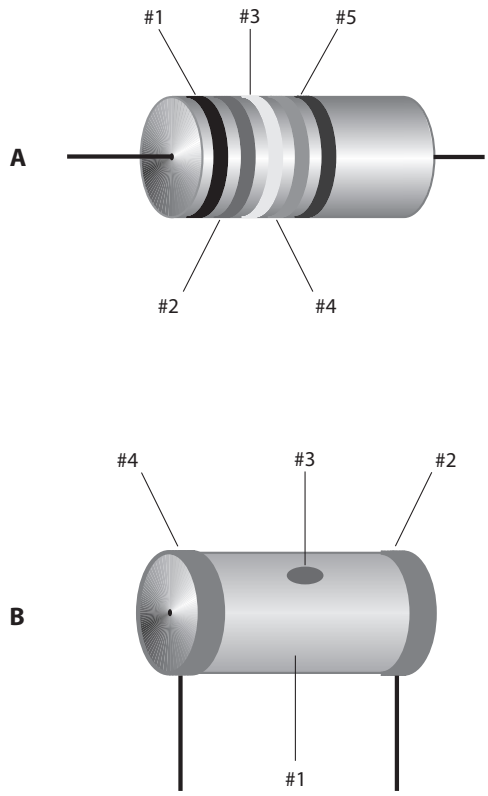
You might ask, "If we want to build our own temperature-compensated resistor, can we use two resistors with half (or twice) the value we need, but with *opposite* resistance-versus-temperature characteristics, and connect them in series or parallel?" That's an excellent question. If we can find two such resistors, the component whose resistance decreases with increasing temperature (that is, the one that has a *negative temperature coefficient*) will partially or totally "undo" the thermal problem

caused by the component whose resistance goes up with increasing temperature (the one that has a *positive temperature coefficient*). This scheme can sometimes work. Unfortunately, we'd likely spend more time trying to find two "ideally mismatched" resistors than we'd spend by resorting to "brute force" and building a series-parallel matrix.

The Color Code for Resistors

Most cylindrical through-hole resistors have *color bands* that indicate their values and tolerances. You'll see three, four, or five bands around carbon-composition resistors and film resistors. Other resistors have enough physical bulk to allow for printed numbers that tell us the values and tolerances straightaway.

On resistors with *axial leads* (wires that come straight out of both ends), the first, second, third, fourth, and fifth bands are arranged as shown in Fig. 6-9A. On resistors with *radial leads* (wires that come off the ends at right angles to the axis of the component body), the colored regions are arranged as shown in Fig. 6-9B. The first two regions represent single digits 0 through 9, and the third region represents a multiplier of 10 to some power. (For the moment, don't worry about the fourth and fifth regions.) Table 6-1 indicates the numerals corresponding to various colors.



6-9 At A, locations of color-code bands on a resistor with axial leads. At B, locations of color code designators on a resistor with radial leads.

Table 6-1. Color code for the first three bands on fixed resistors. See text for discussion of the fourth and fifth bands

Color of band	Numeral (first and second bands)	Multiplier (third band)
Black	0	1
Brown	1	10
Red	2	100
Orange	3	1000 (1 k)
Yellow	4	10^4 (10 k)
Green	5	10^5 (100 k)
Blue	6	10^6 (1 M)
Violet	7	10^7 (10 M)
Gray	8	10^8 (100 M)
White	9	10^9 (1000 M or 1 G)

Suppose that you find a resistor with three bands: yellow, violet, and red, in that order. You can read as follows, from left to right, referring to Table 6-1:

- Yellow = 4
- Violet = 7
- Red = ×100

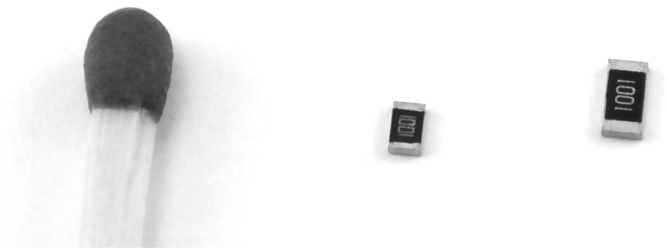
You conclude that the rated resistance equals 4700 Ω, or 4.7 k. As another example, suppose you find a resistor with bands of blue, gray, and orange. You refer to Table 6-1 and determine that

- Blue = 6
- Gray = 8
- Orange = ×1000

This sequence tells you that the resistor is rated at 68,000 Ω, or 68 k.

If a resistor has a fourth colored band on its surface (#4 as shown in Fig. 6-9A or B), then that band tells you the tolerance. A silver band indicates ±10%. A gold band indicates ±5%.

SMD resistors like the ones shown in Fig. 6-10 do not have color stripes, but rather numbers printed in very tiny writing. Both resistors are marked 1001. Like a four-digit color code, this represents 100 followed by one extra zero, making the resistance value 1000 Ω or 1 k.



6-10 SMD resistors with match for scale.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

7
CHAPTER

Cells and Batteries

IN ELECTRICITY, WE CALL A UNIT SOURCE OF DC ENERGY A *CELL*. WHEN WE CONNECT TWO OR MORE cells in series, parallel, or series-parallel, we obtain a *battery*. Numerous types of cells and batteries exist, and inventors keep discovering more.

Electrochemical Energy

Early in the history of electricity science, physicists noticed that when metals came into contact with certain chemical solutions, a potential difference sometimes appeared between the pieces of metal. These experimenters had discovered the first *electrochemical cells*.

A piece of lead and a piece of lead dioxide immersed in an acid solution (Fig. 7-1) acquire a persistent potential difference. In the original experiments, scientists detected this voltage by connecting a galvanometer between the pieces of metal. A resistor in series with the galvanometer prevents excessive current from flowing, and keeps acid from "boiling out" of the cell. Nowadays, of course, we can use a laboratory voltmeter to measure the potential difference.

If we draw current from a cell, such as the one shown in Fig. 7-1, for a long time by connecting a resistor between its terminals, the current will gradually decrease, and the electrodes will become coated. Eventually, all the *chemical energy* in the acid will have turned into *electrical energy* and dissipated as *thermal energy* in the resistor and the cell's own chemical solution, escaping into the surrounding environment in the form of *kinetic energy*.

Single Use and Rechargeable Cells and Batteries

Some electrical cells, once their chemical energy has been used up, must be thrown away. We call such a device a *primary cell*. Other kinds of cells, such as the lead-acid type, can get their chemical energy back again by means of *recharging*. Such a cell constitutes a *secondary cell*.

Primary cells include the ones you usually put in a flashlight, in a transistor radio, and in various other consumer devices. They use dry *electrolyte* (conductive chemical) pastes along with metal electrodes, and go by names, such as *dry cell*, *zinc-carbon cell*, or *alkaline cell*. When you encounter a shelf full of "batteries" in a department store, you'll see primary cells that go by names, such as

AAA batteries, *D batteries*, *camera batteries*, and *watch batteries*. (These are actually cells, not true batteries.) You'll also see 9-V *transistor batteries* and large 6-V *lantern batteries*.

You can also find rechargeable cells in consumer stores. They cost several times as much as ordinary dry cells, and the requisite charging unit also costs a few dollars. But if you take care of rechargeable cells, you can use them hundreds of times, and they'll pay for themselves several times over.

The battery in your car or truck consists of several secondary cells connected in series. These cells recharge from the *alternator* (a form of generator) or from an external charging unit. A typical *automotive battery* has cells like the one in Fig. 7-1. You should never short-circuit the terminals of such a battery or connect a load to it that draws a large amount of current because the acid (sulfuric acid) can erupt out of the battery container. Serious skin and eye injuries can result. In fact, it's a bad idea to short-circuit any cell or battery because it can rupture and damage surrounding materials, wiring, and components. Some "shorted-out" cells and batteries can heat up enough to catch on fire.

In an electric car, the battery will consist of hundreds or even thousands of rechargeable cells arranged both in series and parallel to make a battery that can store very large amounts of energy.

Cells in Series and Parallel

When we want to make a battery from two or more electrochemical cells, we should always use cells having the same chemical composition and the same physical size and mass. In other words, all the cells in the set should be identical! Assuming we heed that principle, we can generalize as follows.

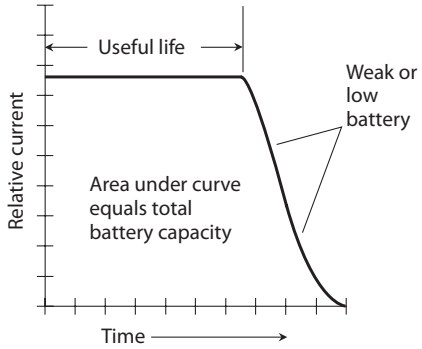
- When we connect cells in series, the *no-load output voltage* (when we don't make the cells deliver any current) multiplies by the number of cells, while the *maximum deliverable current* (when we make the cells produce as much current as they can) equals the maximum deliverable current from only one cell.
- When we connect cells in parallel, the no-load output voltage of the whole set equals the no-load output voltage of only one cell, while the maximum deliverable current from the set multiplies according to the number of cells.

Storage Capacity

Engineers commonly work with two units of electrical energy: the watt-hour (Wh) and the kilowatt-hour (kWh), as we've learned. Any electrochemical cell or battery has a certain amount of electrical energy that we can "extract" before it "runs out of juice." We can quantify that energy in terms of watt-hours or kilowatt-hours. Some engineers express the capacity of a cell or battery of known voltage in units called *ampere-hours* (Ah).

As an example, a battery with a rating of 2 Ah can provide 2 A for 1 h, or 1 A for 2 h, or 100 mA for 20 h. Infinitely many possibilities exist, as long as the product of the current (in amperes) and the usage time (in hours) equals 2. Practical usage limitations are the *shelf life* at one extreme, and the *maximum deliverable current* at the other. We define shelf life as the length of time the battery will last if we never use it at all; this time period might be several years. We define the maximum deliverable current as the highest amount of current that the battery can provide at any moment without suffering a significant decrease in the output voltage because of its own *internal resistance*.

Small cells have storage capacity ratings of a few milliampere-hours (mAh) up to 100 or 200 mAh. Medium-sized cells can supply 500 mAh to 1 Ah. Large automotive batteries can provide upwards of 50 Ah. The energy capacity in watt-hours equals the ampere-hour capacity multiplied by the battery voltage. For a cell or battery having a particular chemical composition, the storage



7-1 A flat discharge curve, considered ideal.

capacity varies directly in proportion to the physical volume of the device. A cell whose volume equals 20 cubic centimeters (cm^3), therefore, has twice the total energy storage capacity of a cell having the same chemical makeup, but that has a volume of only 10 cm^3.

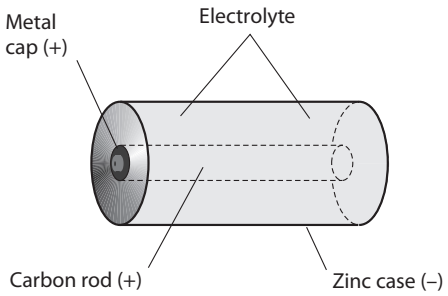
An *ideal cell* or *ideal battery* (a theoretically perfect one) delivers a constant current for a while, and then the current drops fast (Fig. 7-1). Some types of cells and batteries approach this level of perfection, which we represent graphically as a *flat discharge curve*. Most cells and batteries are imperfect, and some are far from the ideal, delivering current that declines steadily from the start. When the deliverable current under constant load has tailed off to about half of its initial value, we say that the cell or battery has become "weak." At this time, we should replace it. If we allow such a cell or battery to run down until the current goes to zero, we call it "dead." The area under the curve in Fig. 7-1 represents the total capacity of the cell or battery in ampere-hours.

"Grocery Store" Cells and Batteries

The cells you see in retail stores provide approximately 1.5-V DC, and are available in sizes known as AAA (very small), AA (small), C (medium large), and D (large). You can also find batteries that deliver 6- or 9-V DC.

Zinc-Carbon Cells

Figure 7-2 is a "translucent" drawing of a *zinc-carbon cell*. The zinc forms the case, which serves as the negative electrode. A carbon rod constitutes the positive electrode. The electrolyte comprises a



7-2 Construction of a zinc-carbon electrochemical cell.

paste of manganese dioxide and carbon. Zinc-carbon cells don't cost very much. They work well at moderate temperatures, and in applications where the current drain is moderate to high. They don't perform well in extreme cold or extreme heat.

Alkaline Cells

The *alkaline cell* has granular zinc as the negative electrode, potassium hydroxide as the electrolyte, and an element called a *polarizer* as the positive electrode. The construction resembles that of the zinc-carbon cell. An alkaline cell can work at lower temperatures than a zinc-carbon cell can. The alkaline cell also lasts longer in most electronic devices. It's the cell of choice for use in transistor radios, calculators, and portable cassette players. The shelf life exceeds that of a zinc-carbon cell. As you might expect, it costs more than a zinc-carbon cell of comparable physical size.

9-V Batteries

A *transistor battery* consists of six tiny zinc-carbon or alkaline cells connected in series and enclosed in a small box-shaped case. Each cell supplies 1.5 V, so the battery supplies 9 V. Even though these batteries have more voltage than individual cells, the energy capacity is less than that of a single-size C or D cell. The electrical energy that we can get from a cell or battery varies in direct proportion to the amount of chemical energy stored in it—and that, in turn, is a direct function of the *volume* (physical size) of the cell or the *mass* (quantity of chemical matter) of the cell. Cells of size C or D have more volume and mass than a transistor battery does, and therefore, contain more stored energy for the same chemical composition. We can find transistor batteries in low-current electronic devices, such as remote-control garage-door openers, television (TV) and hi-fi remote-control units, and electronic calculators.

Lithium Batteries

In recent years, Lithium cell technology has led to a new wave of both single-use and rechargeable batteries becoming available to power the new generations of mobile phones, laptops, and electric vehicles.

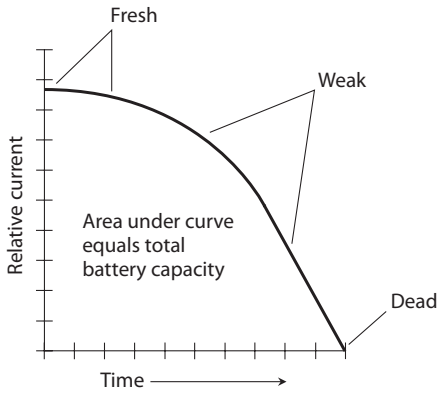
Single-Use Lithium Cells and Batteries

Lithium cells gained popularity in the early 1980s. We can find several variations in the chemical makeup of these cells. They all contain lithium, a light, highly reactive metal. Lithium cells typically supply 1.5 V to 3.5 V, depending on the chemistry used in manufacture. These cells, like all other cells, can be stacked to make batteries.

Lithium batteries originally found application as memory backup power supplies for electronic microcomputers. Lithium cells and batteries have superior shelf life. They can last for years in very-low-current applications, such as memory backup or the powering of a digital *liquid-crystal-display* (LCD) watch or clock. These cells also provide high-energy capacity per unit of volume or mass.

Lithium-Polymer (LiPo) Cells and Batteries

LiPo cells produce a voltage of 3.7 V rising to 4.35 V (fully charged) and are, therefore, often used as a single cell or a 7.4-V battery of two cells. Larger batteries are also found in laptops and other high-power devices.



7-3 A declining discharge curve.

The energy density of a LiPo cell is much higher than any of the other readily available battery technology. For this reason, they have become the technology of choice for most applications in which a rechargeable battery is needed.

You must use care when charging LiPo cells, as they are prone to catching fire if overcharged. Overdischarging can easily destroy the battery. Some cells include a built-in IC that automatically prevents overcharging or undercharging. In a LiPo battery with two or more cells, each cell should be charged separately, using a special balanced charger.

Lead-Acid Batteries

You've seen the basic configuration for a lead-acid cell, which has a solution of sulfuric acid, along with a lead electrode (negative) and a lead-dioxide electrode (positive). These cells are rechargeable.

Automotive batteries comprise series-connected sets of lead-acid cells having a free-flowing liquid acid. You can't tip such a battery on its side, or turn it upside-down, without running the risk of having some of the acid electrolyte spill out. Some lead-acid batteries have semisolid electrolytes; they find applications in consumer electronic devices, notebook computers, and *uninterruptible power supplies* (UPSs) that can keep a desktop computer running for a few minutes if the utility power fails.

A large lead-acid battery, such as the one in your car or truck, can store several tens of ampere-hours. The smaller ones, like those in a UPS, have less capacity but more versatility. Their main attributes include the fact you can use and recharge them many times, they don't cost much money, and you don't have to worry about the irregular discharge characteristics that some rechargeable cells and batteries have.

The heavy weight and toxicity of lead are gradually reducing the usage of lead-acid batteries as lighter and more environment-friendly cell technologies are becoming cheaper.

Nickle Metal Hydride Cells and Batteries

Nickel-based batteries are available in packs of cells. You can sometimes plug these packs directly into consumer equipment. In other cases, the batteries actually form part of the device housing. All nickel-based cells are rechargeable. You can put them through hundreds or even thousands of *charge/discharge cycles* if you take good care of them.

Configurations and Applications

Nickel-based cells come in various sizes and shapes. *Cylindrical cells* look like ordinary dry cells. You'll find *button cells* in cameras, watches, memory backup applications, and other places where miniaturization matters. *Flooded cells* find application in heavy-duty electronic and electromechanical systems; some of these can store 1000 Ah or more. *Spacecraft cells* are manufactured in airtight, thermally protected packages that can withstand the rigors of a deep-space environment.

Most orbiting satellites endure total darkness for approximately half the time, and bask in direct sunlight the other half of the time. (The rare exception is the satellite with a carefully prescribed orbit that keeps it above the *gray line*, or the zone of surface sunrise or sunset. Such a satellite "sees" the sun all the time.) *Solar panels* can operate while the satellite receives sunlight, but during the times that the earth eclipses the sun, electrochemical batteries must power the electronic equipment on the satellite. The solar panels can charge the electrochemical battery, in addition to powering the satellite, for the "daylight" half of each orbit.

Precautions

Never discharge nickel-based cells all the way until they "totally die." If you make that mistake, you can cause the polarity of a cell, or of one or more cells in a battery, to permanently reverse, ruining the device for good.

Nickel-based cells and batteries sometimes exhibit a bothersome characteristic called *memory* or *memory drain*. If you use such a device repeatedly, and you allow it to discharge to the same extent with every cycle, it seems to lose most of its capacity and "die too soon." You can sometimes "cure" a nickel-based cell or battery of this problem by letting it run down until it stops working properly, recharging it, running it down again, and repeating the cycle numerous times. In stubborn cases, you'll want to buy a new cell or battery instead of spending a lot of time trying to rejuvenate the old one.

Nickel-based cells and batteries work best if used with charging units that take several hours to fully replenish the charge. So-called *high-rate* or *quick* chargers are available, but some of these can force too much current through a cell or battery. It's best if the charger is made especially for the cell or battery type you use.

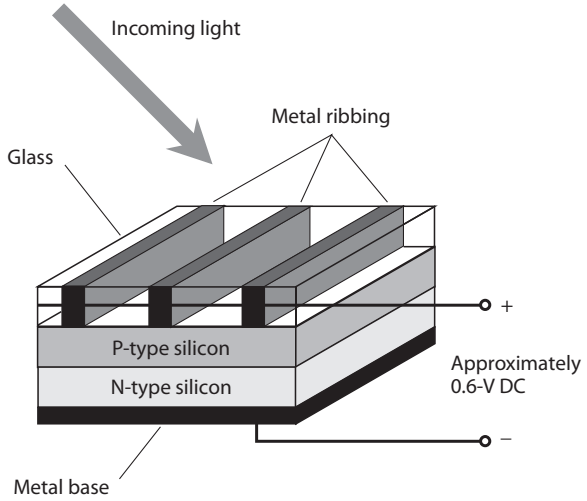
Photovoltaic Cells and Batteries

The *photovoltaic* (PV) *cell*, also called a *solar cell*, differs fundamentally from the electrochemical cell. A PV cell converts visible light, infrared (IR) rays, and/or ultraviolet (UV) rays directly into DC electricity.

Construction and Performance

Figure 7-4 shows the basic internal construction of a photovoltaic cell. A flat semiconductor *P-N junction* forms the active region within the device. It has a transparent housing so that radiant energy can directly strike the *P-type* silicon. The metal ribs, forming the positive electrode, are interconnected by means of tiny wires. The negative electrode consists of a metal backing, called the *substrate*, placed in contact with the *N-type* silicon.

Most silicon-based solar cells provide about 0.6 V DC in direct sunlight. If the current demand is low, muted sunlight or artificial lamps can produce the full output voltage from a solar cell. As the current demand increases, the cell must receive more intense illumination to produce its full output



7-4 Construction of a silicon photovoltaic (PV) cell.

voltage. A maximum limit exists to the current that a solar cell can deliver, no matter how bright the light. To obtain more current than that, we must connect multiple cells in parallel.

When we connect numerous photovoltaic cells in series-parallel, we obtain a *solar panel*. A large solar panel might consist of, say, 50 parallel sets of 20 series-connected cells. The series connection boosts the voltage, and the parallel connection increases the current-delivering ability. Sometimes, you'll find multiple solar panels connected in series or parallel to make vast arrays.

Practical Applications

Solar cells have become cheaper and more efficient in recent years, as researchers increasingly look to them as an alternative energy source. Solar panels are widely used in earth-orbiting satellites and interplanetary spacecraft. The famed Mars rovers could never have worked without them. Some alternative-energy enthusiasts have built systems that use solar panels in conjunction with rechargeable batteries, such as the lead-acid or nickel-cadmium types, to provide power independent of the commercial utilities.

A completely independent solar/battery power system is called a *stand-alone* system. It employs large solar panels, large-capacity lead-acid batteries, a *power inverter* to convert the DC into AC, and a sophisticated charging circuit. Obviously, these systems work best in environments where most days are sunny! The maximum deliverable power in full sunlight depends on the surface area of the panel.

Solar cells, either alone or supplemented with rechargeable batteries, can be connected into a home electric system in an *interactive* arrangement with the electric utilities. When the solar power system can't provide for the needs of the household all by itself, the utility company can make up for the shortage. Conversely, when the solar power system supplies more than enough for the needs of the home, the utility company can buy the excess energy from the consumer.

Fuel Cells

In the late 1900s, a new type of electrochemical power device, called the *fuel cell*, emerged. Many scientists and engineers believe that fuel cells hold promise as an alternative energy source to help offset our traditional reliance on coal, oil, and natural gas.

Hydrogen Fuel

The most talked-about fuel cell during the early years of research and development became known as the *hydrogen fuel cell*. As its name implies, it derives electricity from hydrogen. The hydrogen combines with oxygen (it oxidizes) to form energy and water. The hydrogen fuel cell produces no pollution and no toxic by-products. When a hydrogen fuel cell "runs out of juice," we need nothing more than a new supply of hydrogen to get it going again; its oxygen comes from the earth's atmosphere.

Instead of literally burning, the hydrogen in a fuel cell oxidizes in a controlled fashion and at a much lower temperature. The *proton exchange membrane* (PEM) *fuel cell* is one of the most widely used. A PEM hydrogen fuel cell generates approximately 0.7-V DC under no-load conditions. In order to obtain higher voltages, we can connect multiple PEM fuel cells in series. A series-connected set of fuel cells technically forms a battery, but engineers call it a *stack*.

Commercial manufacturers provide fuel-cell stacks in various sizes. A stack having roughly the size and weight of a book-filled travel suitcase can power a subcompact electric car. Smaller cells, called *micro fuel cells*, can provide DC to run devices that have historically operated from conventional cells and batteries. These include portable radios, lanterns, and notebook computers.

Other Fuels

Fuel cells can use energy sources other than hydrogen. Almost anything that will combine with oxygen to form energy can work. *Methanol*, a form of alcohol, is easier to transport and store than hydrogen because methanol exists as a liquid at room temperature. *Propane* powers some fuel cells. It can be stored in tanks for barbecue grills and some rural home heating systems. Still other fuel cells operate from *methane*, also known as natural gas. Theoretically, any combustible material will work: even oil or gasoline!

Some scientists object to the use of any energy source that employs so-called *fossil fuels* on which society has acquired a heavy dependence. To some extent we can dismiss these opinions as elitist, but in another sense, we must acknowledge a practical concern: Our planet has a finite supply of fossil fuels, the demand for which will grow for decades to come, especially in developing countries. Today's exotic energy alternative might become tomorrow's fuel of choice in the developed nations. The harder we try to make it so, the sooner it can happen.

A Promising Technology

As of this writing, fuel cells have not replaced conventional electrochemical cells and batteries in common applications, mainly because of the high cost. Hydrogen holds the honors as the most abundant and simplest chemical element in the universe, and it produces no toxic by-products when we liberate its stored energy. Hydrogen might, therefore, seem ideal as the choice for use in fuel cells. But storage and transport of hydrogen has proven difficult and expensive, especially for fuel cells and stacks intended for systems not affixed to permanent pipelines.

An interesting scenario, suggested by one of my physics teachers in the 1970s, involves piping hydrogen gas through lines already designed to carry methane. Some infrastructure modification would be necessary to safely handle hydrogen, which escapes through small cracks and openings more easily than methane. But hydrogen, if obtained at reasonable cost and in abundance, could power large fuel-cell stacks in households and businesses. Power inverters could convert the DC from such a stack to utility AC. A typical home power system of this sort would easily fit into a small room or a corner of the basement.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

8
CHAPTER

Magnetism

A *MAGNETIC FIELD* ARISES WHEN ELECTRIC CHARGE CARRIERS MOVE. CONVERSELY, WHEN AN ELECTRICAL conductor moves in a magnetic field, current flows in that conductor.

Geomagnetism

The earth has a core consisting largely of iron, heated to the extent that some of it liquefies. As the earth rotates on its axis, the iron in the core flows in convection patterns, generating the *geomagnetic field* that surrounds our planet and extends thousands of kilometers into space.

Earth's Magnetic Poles and Axis

The geomagnetic field has poles, just as an old-fashioned bar magnet does. On the earth's surface, these poles exist in the arctic and antarctic regions, but they are displaced considerably from the *geographic poles* (the points where the earth's axis intersects the surface). The *geomagnetic lines of flux* converge or diverge at the *geomagnetic poles*. The *geomagnetic axis* that connects the geomagnetic poles tilts somewhat with respect to the *geographic axis* on which the earth rotates.

Charged subatomic particles from the sun, constantly streaming outward through the solar system, distort the geomagnetic field. This so-called *solar wind* "blows" the geomagnetic field out of symmetry. On the side of the earth facing the sun, the lines of flux compress. On the side of the earth opposite the sun, the lines of flux dilate. Similar effects occur in other planets, notably Jupiter, that have magnetic fields. As the earth rotates, the geomagnetic field "dances" into space in the direction facing away from the sun.

The Magnetic Compass

Thousands of years ago, observant people noticed the presence of the geomagnetic field, even though they didn't know exactly what caused it. Certain rocks, called *lodestones*, when hung by strings, always orient themselves in a generally north-south direction. Long ago, seafarers and explorers correctly attributed this effect to the presence of a "force" in the air. The reasons for this phenomenon

remained unknown for centuries, but adventurers put it to good use. Even today, a *magnetic compass* makes a valuable navigation aid. It can work when more sophisticated navigational devices, such as a *Global Positioning System* (GPS), fails.

The geomagnetic field interacts with the magnetic field around a compass needle, which comprises a small bar magnet. This interaction produces force on the compass needle, causing it to align itself parallel to the geomagnetic lines of flux in the vicinity. The force operates not only in a horizontal plane (parallel to the earth's surface), but also vertically in most locations. The vertical force component vanishes at the *geomagnetic equator*, a line running around the globe equidistant from both geomagnetic poles, so the force there is perfectly horizontal. But as the *geomagnetic latitude* increases, either toward the north or the south geomagnetic pole, the *magnetic force* pulls up and down on the compass needle more and more. We call the extent of the vertical force component at any particular place the *geomagnetic inclination*. Have you noticed this when using a magnetic compass? One end of the needle dips a little toward the compass face, while the other end tilts upward toward the glass.

Because the earth's geomagnetic axis and geographic axis don't coincide, the needle of a magnetic compass usually points somewhat to the east or west of true geographic north. The extent of the discrepancy depends on our surface location. We call the angular difference between geomagnetic north (north according to a compass) and geographic north (or *true north*) the *geomagnetic declination*.

Magnetic Force

As children, most of us discovered that magnets "stick" to some metals. Iron, nickel, a few other elements, and alloys or solid mixtures containing any of them constitute *ferromagnetic materials*. Magnets exert force on these metals. Magnets do not generally exert force on other metals unless those metals carry electric currents. Electrically insulating substances never "attract magnets" under normal conditions.

Cause and Strength

When we bring a *permanent magnet* near a sample of ferromagnetic material, the atoms in the material line up to a certain extent, temporarily magnetizing the sample. This atomic alignment produces a magnetic force between the atoms of the sample and the atoms in the magnet. Every single atom acts as a tiny magnet; when they act in concert with one another, the whole sample behaves as a magnet. Permanent magnets always attract samples of ferromagnetic material.

If we place two permanent magnets near each other, we observe a stronger magnetic force than we do when we bring either magnet near a sample of ferromagnetic material. The mutual force between two rod-shaped or bar-shaped permanent magnets is manifest as attraction if we bring two opposite poles close together (north-near-south or south-near-north) and repulsion if we bring two like poles into proximity (north-near-north or south-near-south). Either way, the force increases as the distance between the ends of the magnets decreases.

Some *electromagnets* produce fields so powerful that no human can pull them apart if they get "stuck" together, and no one can bring them all the way together against their mutual repulsive force. (We'll explore how electromagnets work later in this chapter.) Industrial workers can use huge electromagnets to carry heavy pieces of scrap iron or steel from place to place. Other electromagnets can provide sufficient repulsion to suspend one object above another, an effect known as *magnetic levitation*.

Electric Charge Carriers in Motion

Whenever the atoms in a sample of ferromagnetic material align to any extent rather than existing in a random orientation, a magnetic field surrounds the sample. A magnetic field can also result from the motion of electric *charge carriers*. In a wire, electrons move in incremental "hops" along the conductor from atom to atom. In a permanent magnet, the movement of orbiting electrons occurs in such a manner that an *effective current* arises.

Magnetic fields can arise from the motion of charged particles through space, as well as from the motion of charge carriers through a conductor. The sun constantly ejects protons and helium nuclei, both of which carry positive electric charges. These particles produce effective currents as they travel through space. The effective currents in turn generate magnetic fields. When these fields interact with the geomagnetic field, the subatomic particles change direction and accelerate toward the geomagnetic poles.

When an eruption on the sun, called a *solar flare,* occurs, the sun ejects far more charged subatomic particles than usual. As these particles approach the geomagnetic poles, their magnetic fields, working together, disrupt the geomagnetic field, spawning a *geomagnetic storm*. Such an event causes changes in the earth's upper atmosphere, affecting "shortwave radio" communications and producing the *aurora borealis* ("northern lights") and *aurora australis* ("southern lights"), well-known to people who dwell at high latitudes. If a geomagnetic storm reaches sufficient intensity, it can interfere with wire communications and electric power transmission at the surface.

Lines of Flux

Physicists consider magnetic fields to comprise *flux lines*, or *lines of flux*. The intensity of the field depends on the number of flux lines passing at right angles through a region having a certain cross-sectional area, such as a centimeter squared (cm^2) or a meter squared (m^2). The flux lines are not actual material fibers, but their presence can be shown by means of a simple experiment.

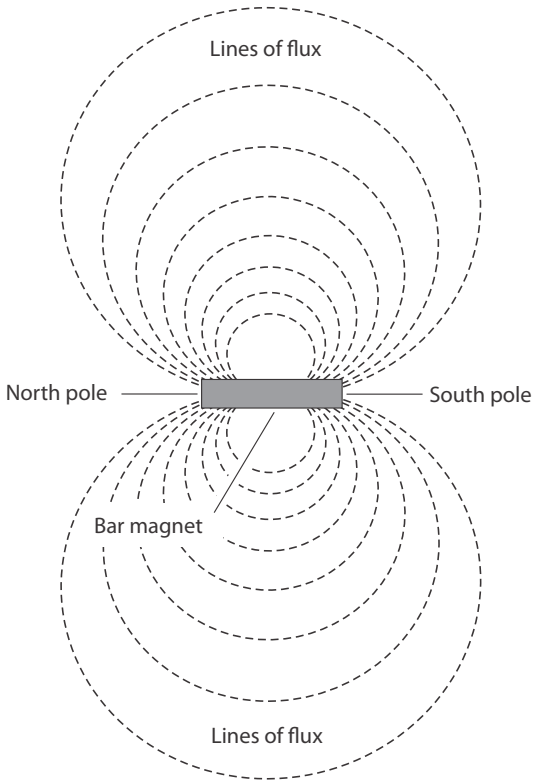
Have you seen the classical demonstration in which iron filings lie on a sheet of paper, and then the experimenter holds a permanent magnet underneath the sheet? The filings arrange themselves in a pattern that shows, roughly, the "shape" of the magnetic field in the vicinity of the magnet. A bar magnet has a field whose lines of flux exhibit a characteristic pattern (Fig. 8-1).

Another experiment involves passing a current-carrying wire through the paper at a right angle. The iron filings bunch up in circles centered at the point where the wire passes through the paper. This experiment shows that the lines of flux around a straight, current-carrying wire form concentric circles in any plane passing through the wire at a right angle. The center of every "flux circle" lies on the wire, which constitutes the path along which the charge carriers move (Fig. 8-2).

Polarity

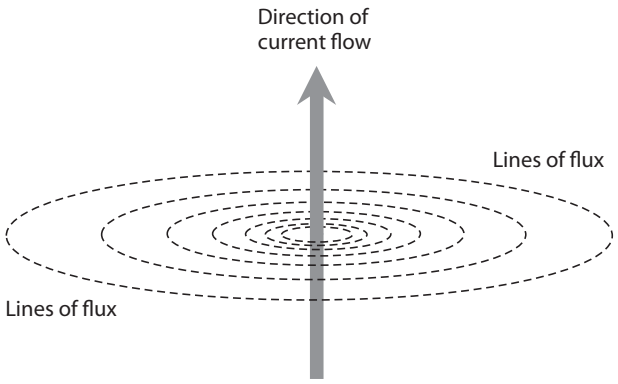
A magnetic field has a specific orientation at any point in space near a current-carrying wire or a permanent magnet. The flux lines run parallel with the direction of the field. Scientists consider the magnetic field to begin, or originate, at a *north pole*, and to end, or terminate, at a *south pole*. These poles do not correspond to the geomagnetic poles; in fact, they're the opposite! The north geomagnetic pole is in reality a south pole because it attracts the north poles of magnetic compasses. Similarly, the south geomagnetic pole is really a north pole because it attracts the south poles of compasses. In the case of a permanent magnet, we can usually (but not always) tell where the magnetic poles are located. Around a current-carrying wire, the magnetic field revolves endlessly.

A charged electric particle (such as a proton) hovering in space forms an *electric monopole*, and the electric flux lines around it aren't closed. A positive charge does not have to mate with a negative



8-1 Magnetic flux around a bar magnet.

charge. The electric flux lines around any stationary, charged particle run outward in all directions for a theoretically infinite distance. But magnetic fields behave according to stricter laws. Under normal circumstances, all magnetic flux lines form closed loops. In the vicinity of a magnet, we can always find a starting point (the north pole) and an ending point (the south pole). Around a current-carrying wire, the loops form circles.



8-2 Magnetic flux produced by charge carriers traveling in a straight line.

Magnetic Dipoles

You might at first suppose that the magnetic field around a current-carrying wire arises from a monopole, or that no poles exist. The concentric flux circles don't seem to originate or terminate anywhere. But you can assign originating and terminating points to those circles, thereby defining a *magnetic dipole*—a pair of opposite magnetic poles in close proximity.

Imagine that you hold a flat piece of paper next to a current-carrying wire, so that the wire runs along one edge of the sheet. The magnetic circles of flux surrounding the wire pass through the sheet of paper, entering one side and emerging from the other side, so you have a "virtual magnet." Its north pole coincides with the face of the paper sheet from which the flux circles emerge. Its south pole coincides with the opposite face of the sheet, into which the flux circles plunge.

The flux lines in the vicinity of a magnetic dipole always connect the two poles. Some flux lines appear straight in a local sense, but in the larger sense they always form curves. The greatest magnetic field strength around a bar magnet occurs near the poles, where the flux lines converge or diverge. Around a current-carrying wire, the greatest field strength occurs near the wire.

Magnetic Field Strength

Physicists and engineers express the overall magnitude, or quantity, of a magnetic field in units called *webers*, symbolized Wb. We can employ a smaller unit, the *maxwell* (Mx), for weak fields. One weber equals 100,000,000 (10^8) maxwells. Therefore,

$$1 \text{ Wb} = 10^8 \text{ Mx}$$

and

$$1 \text{ Mx} = 10^{-8} \text{ Wb}$$

The Tesla and the Gauss

If you have a permanent magnet or an electromagnet, you might see its "strength" expressed in terms of webers or maxwells. But more often, you'll hear or read about units called *teslas* (T) or *gauss* (G). These units define the concentration, or intensity, of the magnetic field as its flux lines pass at right angles through flat regions having specific cross-sectional areas.

The *flux density*, or number of "flux lines per unit of cross-sectional area," forms a more useful expression for magnetic effects than the overall quantity of magnetism. In equations, we denote flux density using the letter *B*. A flux density of one tesla equals one weber per meter squared (1 Wb/m^2). A flux density of one gauss equals one maxwell per centimeter squared (1 Mx/cm^2). As things work out, the gauss equals 0.0001 (10^{-4}) tesla, so we have the relations

$$1 \text{ G} = 10^{-4} \text{ T}$$

and

$$1 \text{ T} = 10^4 \text{ G}$$

If you want to convert from teslas to gauss (not gausses!), multiply by 10^4. If you want to convert from gauss to teslas, multiply by 10^{-4}.

Quantity versus Density

If the distinctions between webers and teslas, or between maxwells and gauss, confuse you, think of an ordinary light bulb. Suppose that a lamp emits 15 W of visible-light power. If you enclose

the bulb completely, then 15 W of visible light strike the interior walls of the chamber, regardless of the size of the chamber. But this notion doesn't give you a useful notion of the brightness of the light. You know that a single bulb produces plenty of light if you want to illuminate a closet, but nowhere near enough light to illuminate a gymnasium. The important consideration is the number of watts *per unit of area*. When you say that a bulb gives off so-many watts of light *overall*, it's like saying that a magnet has a magnetic quantity of so-many webers or maxwells. When you say that the bulb produces so-many watts of light *per unit of area*, it's like saying that a magnetic field has a flux density of so-many teslas or gauss.

Magnetomotive Force

When we work with wire loops, *solenoidal* (helical) coils, and rod-shaped electromagnets, we can quantify a phenomenon called *magnetomotive force* with a unit called the *ampere-turn* (At). This unit describes itself well: the number of amperes flowing in a coil or loop, times the number of turns that the coil or loop contains.

If we bend a length of wire into a loop and drive 1 A of current through it, we get 1 At of magnetomotive force inside the loop. If we wind the same length of wire (or any other length) into a 50-turn coil and keep driving 1 A of current through it, the resulting magnetomotive force increases by a factor of 50, to 50 At. If we then reduce the current in the 50-turn loop to 1/50 A or 20 mA, the magnetomotive force goes back down to 1 At.

Sometimes, engineers employ a unit called the *gilbert* to express magnetomotive force. One gilbert (1 Gb) equals approximately 0.7958 At. The gilbert represents a slightly smaller unit than the At does. Therefore, if we want to determine the number of ampere-turns when we know the number of gilberts, we should multiply by 0.7958. To determine the number of gilberts when we know the number of ampere-turns, we should multiply by 1.257.

Magnetomotive force does not depend on core material or loop diameter. Even if we place a metal rod in a solenoidal coil, the magnetomotive force will not change if the current through the wire remains the same. A tiny 100-turn air-core coil carrying 1 A produces the same magnetomotive force as a huge 100-turn air-core coil carrying 1 A. Magnetomotive force depends *only* on the current and the number of turns.

Flux Density versus Current

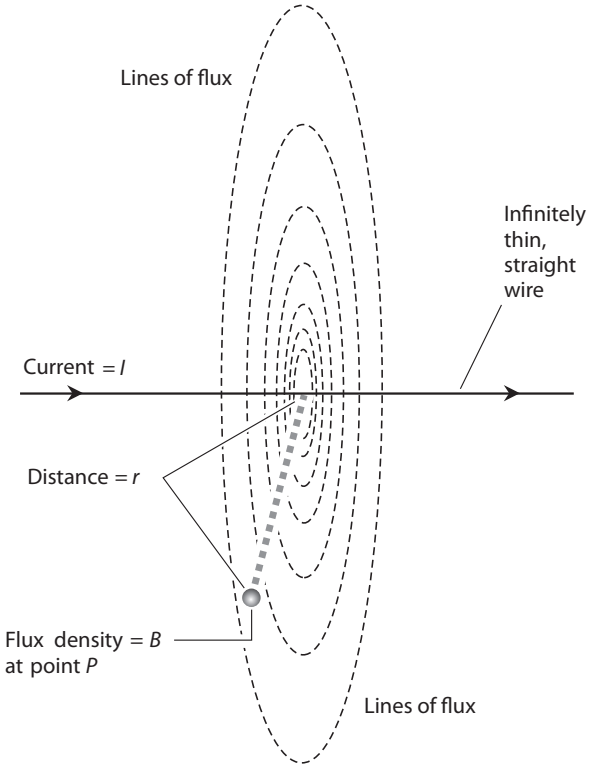
In a straight wire carrying a steady, direct current and surrounded by air or a vacuum, we observe the greatest flux density near the wire, and diminishing flux density as we get farther away from the wire. We can use a simple formula to express magnetic flux density as a function of the current in a straight wire and the distance from the wire.

Imagine an infinitely thin, absolutely straight, infinitely long length of wire (that's the ideal case). Suppose that the wire carries a current of I amperes. Represent the flux density (in teslas) as B. Consider a point P at a distance r (in meters) from the wire, measured in a plane perpendicular to the wire, as shown in Fig. 8-3. We can find the flux density at point P using the formula

$$B = 2 \times 10^{-7} I / r$$

We can consider the value of the constant, 2×10^{-7}, mathematically exact to any desired number of significant figures.

Of course, we'll never encounter a wire with zero thickness or infinite length. But as long as the wire thickness constitutes a small fraction of r, and as long as the wire is reasonably straight near point P, this formula works quite well in most applications.



8-3 The magnetic flux density varies inversely with the distance from a wire carrying constant current.

Problem 8-1

What is the flux density B_t in teslas at a distance of 200 mm from a straight, thin wire carrying 400 mA of DC?

Solution

First, we must convert all quantities to units in the International System (SI). Thus, we have $r = 0.200$ m and $I = 0.400$ A. We can input these values directly into the formula for flux density to obtain

$$B_t = 2.00 \times 10^{-7} \times 0.400 / 0.200 = 4.00 \times 10^{-7}\ \text{T}$$

Problem 8-2

In the above-described scenario, what is the flux density B_g (in gauss) at point P?

Solution

To figure this out, we must convert from teslas to gauss, multiplying the result in the solution to Problem 8-1 by 10^4 to get

$$B_g = 4.00 \times 10^{-7} \times 10^4 = 4.00 \times 10^{-3}\ \text{G}$$

Electromagnets

The motion of electrical charge carriers always produces a magnetic field. This field can reach considerable intensity in a tightly coiled wire having many turns and carrying a large current. When we place a ferromagnetic rod called a *core* inside a coil, as shown in Fig. 8-4, the magnetic lines of flux concentrate in the core, and we have an electromagnet. Most electromagnets have cylindrical cores. The length-to-radius ratio can vary from extremely low (fat pellet) to extremely high (thin rod). Regardless of the length-to-radius ratio, the flux produced by current in the wire temporarily magnetizes the core.

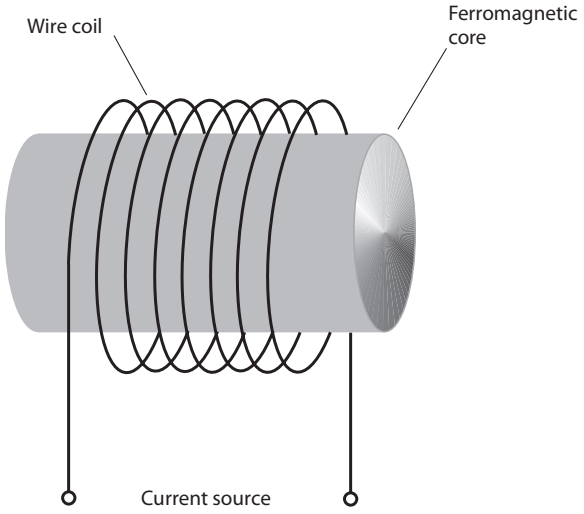
Direct-Current Types

You can build a DC electromagnet by wrapping a couple of hundred turns of insulated wire around a large iron bolt or nail. You can find these items in any good hardware store. You should test the bolt for ferromagnetic properties while you're still in the store, if possible. (If a permanent magnet "sticks" to the bolt, then the bolt is ferromagnetic.) Ideally, the bolt should measure at least ⅜ inch (approximately 1 cm) in diameter and at least 6 inches (roughly 15 cm) long. You must use insulated wire, preferably made of solid, soft copper.

Wind the wire at least several dozen (if not 100 or more) times around the bolt. You can layer the windings if you like, as long as they keep going around in the same direction. Secure the wire in place with electrical or masking tape. A 6-V battery consisting of four AA cells can provide plenty of DC to operate the electromagnet. If you like, you can connect two or more such batteries in parallel to increase the current delivery. Never leave the coil connected to the battery for more than a few seconds at a time.

Warning!

Do not use a lead-acid automotive battery for this experiment. The near short-circuit produced by an electromagnet can cause the acid from such a battery to boil out, resulting in serious injury.



8-4 A simple electromagnet.

All DC electromagnets have defined north and south poles, just as permanent magnets have. The magnetic field exists only as long as the coil carries current. When you remove the power source, the magnetic field nearly vanishes. A small amount of *residual magnetism* remains in the core after the current stops flowing in the coil, but this field has minimal intensity.

Alternating-Current Types

Do you suspect that you can make an electromagnet extremely powerful if, rather than using a lantern battery for the current source, you plug the ends of the coil directly into an AC utility outlet? In theory, you can do this. But don't! You'll expose yourself to the danger of electrocution, expose your house to the risk of electrical fire, and most likely cause a fuse or circuit breaker to open, killing power to the device anyway. Some buildings lack the proper fuses or circuit breakers to prevent excessive current from flowing through the utility wiring in case of an overload. If you want to build and test a safe AC electromagnet, my book, *Electricity Experiments You Can Do at Home* (McGraw-Hill, 2010), offers instructions for doing it.

Some commercially manufactured electromagnets operate from 60-Hz utility AC. These magnets "stick" to ferromagnetic objects. The polarity of the magnetic field reverses every time the direction of the current reverses, producing 120 fluctuations, or 60 complete north-to-south-to-north polarity changes, every second. In addition, the instantaneous intensity of the magnetic field varies along with the AC cycle, reaching alternating-polarity peaks at 1/120-second intervals and nulls of zero intensity at 1/120-second intervals. Any two adjacent peaks and nulls occur 1/4 cycle, or 1/240 second, apart.

If you bring a permanent magnet or DC electromagnet near either "pole" of an AC electromagnet, no net force results from the AC electromagnetism itself because equal and opposite attractive and repulsive forces occur between the alternating magnetic field and any steady external field. But the permanent magnet or the DC electromagnet attracts the core of the AC electromagnet, whether the AC device carries current or not.

Tech Tidbit

With any electromagnet, a DC source, such as a battery, always produces a magnetic field that maintains the same polarity at all times. An AC source always produces a magnetic field that periodically reverses polarity.

Problem 8-3

Suppose that we apply 80-Hz AC to an electromagnet instead of the standard 60 Hz. What will happen to the interaction between the alternating magnetic field and a nearby permanent magnet or DC electromagnet?

Solution

Assuming that the behavior of the core material remains the same, the situation at 80 Hz will not change from the 60-Hz case. In theory, the AC frequency makes no difference in the behavior of an AC electromagnet. In practice, however, the magnetic field weakens at high AC frequencies because the AC electromagnet's *inductance* tends to impede the flow of current. This so-called *inductive reactance* depends on the number of coil turns, and also on the characteristics of the ferromagnetic core.

Magnetic Materials

Ferromagnetic substances cause magnetic lines of flux to bunch together more tightly than they exist in free space. A few materials cause the lines of flux to dilate compared with their free-space density. We call these substances *diamagnetic*. Examples of such materials include wax, dry wood, bismuth, and silver. No diamagnetic material reduces the strength of a magnetic field by anywhere near the factor that ferromagnetic substances can increase it. Usually, engineers use diamagnetic objects to keep magnets physically separated while minimizing the interaction between them.

Permeability

Permeability expresses the extent to which a ferromagnetic material concentrates magnetic lines of flux relative to the flux density in a vacuum. By convention, scientists assign a permeability value of 1 to a vacuum. If we have a coil of wire with an air core and we drive DC through the wire, then the flux inside the coil is about the same as it would be in a vacuum. Therefore, the permeability of air equals almost exactly 1. (Actually it's a tiny bit higher, but the difference rarely matters in practice.)

If we place a ferromagnetic core inside the coil, the flux density increases, sometimes by a large factor. By definition, the permeability equals that factor. For example, if a certain material causes the flux density inside a coil to increase by a factor of 60 compared with the flux density in air or a vacuum, that material has a permeability of 60. Diamagnetic materials have permeability values less than 1, but never very much less. Table 8-1 lists the permeability values for some common substances.

Retentivity

When we subject a substance, such as iron, to a magnetic field as intense as it can handle, say by enclosing it in a wire coil carrying high current, some residual magnetism always remains after the

Table 8-1. Permeability Values for Some Common Materials

Substance	Permeability (approx.)
Air, dry, at sea level	1.0
Alloys, ferromagnetic	3000–1,000,000
Aluminum	Slightly more than 1
Bismuth	Slightly less than 1
Cobalt	60–70
Iron, powdered and pressed	100–3000
Iron, solid, refined	3000–8000
Iron, solid, unrefined	60–100
Nickel	50–60
Silver	Slightly less than 1
Steel	300–600
Vacuum	1.0 (exact, by definition)
Wax	Slightly less than 1
Wood, dry	Slightly less than 1

current stops flowing in the coil. *Retentivity*, also known as *remanence*, quantifies the extent to which a substance "memorizes" a magnetic field imposed on it.

Imagine that we wind a wire coil around a sample of ferromagnetic material and then drive so much current through the coil that the magnetic flux inside the core reaches its maximum possible density. We call this condition *core saturation*. We measure the flux density in this situation, and get a figure of B_{max} (in teslas or gauss). Now suppose that we remove the current from the coil, and then we measure the flux density inside the core again, obtaining a figure of B_{rem} (in teslas or gauss, as before). We can express the retentivity B_r of the core material as a ratio according to the formula

$$B_r = B_{rem} / B_{max}$$

or as a percentage using the formula

$$B_{r\%} = 100\ B_{rem} / B_{max}$$

As an example, suppose that a metal rod can attain a flux density of 135 G when enclosed by a current-carrying coil. Imagine that 135 G represents the maximum possible flux density for that material. (For any substance, such a maximum always exists, unique to that substance; further increasing the coil current or number of turns will not magnetize it any further.) Now suppose that we remove the current from the coil, and 19 G remain in the rod. Then the retentivity B_r is

$$B_r = 19/135 = 0.14$$

As a percentage,

$$B_{r\%} = 100 \times 19/135 = 14\%$$

Certain ferromagnetic substances exhibit high retentivity, and therefore, make excellent permanent magnets. Other ferromagnetic materials have poor retentivity. They can sometimes work okay as the cores of electromagnets, but they don't make good permanent magnets.

If a ferromagnetic substance has low retentivity, it can function as the core for an AC electromagnet because the polarity of the magnetic field in the core follows along closely as the current in the coil alternates. If the material has high retentivity, the material acts "magnetically sluggish" and has trouble following the current reversals in the coil. Substances of this sort don't work well in AC electromagnets.

Problem 8-4

Suppose that we wind a coil of wire around a metal core to make an electromagnet. We find that by connecting a variable DC source to the coil, we can drive the magnetic flux density in the core up to 0.500 T but no higher. When we shut down the current source, the flux density inside the core drops to 500 G. What's the retentivity of this core material?

Solution

First, let's convert both flux density figures to the same units. We recall that $1\ T = 10^4$ G. Therefore, the flux density in gauss is $0.500 \times 10^4 = 5{,}000$ G when the current flows in the coil, and 500 G after we remove the current. "Plugging in" these numbers gives us the ratio

$$B_r = 500/5{,}000 = 0.100$$

or the percentage

$$B_{r\%} = 100 \times 500/5{,}000 = 100 \times 0.100 = 10.0\%$$

Permanent Magnets

Industrial engineers can make any suitably shaped sample of ferromagnetic material into a permanent magnet. The strength of the magnet depends on two factors:

- The retentivity of the material used to make it
- The amount of effort put into magnetizing it

The manufacture of powerful permanent magnets requires an alloy with high retentivity. The most "magnetizable" alloys derive from specially formulated mixtures of aluminum, nickel, and cobalt, occasionally including trace amounts of copper and titanium. Engineers place samples of the selected alloy inside heavy wire coils carrying high, continuous DC for an extended period of time.

You can magnetize any piece of iron or steel. Some technicians use magnetized tools when installing or removing screws from hard-to-reach places in computers, wireless transceivers, and other devices. If you want to magnetize a tool, stroke its metal shaft with the end of a powerful bar magnet several dozen times. But beware: Once you've imposed residual magnetism in a tool, it will remain magnetized to some extent forever!

Flux Density inside a Long Coil

Consider a long, helical coil of wire, commonly known as a *solenoid*, having n turns in a single layer. Suppose that it measures s meters in length, carries a steady direct current of I amperes, and has a ferromagnetic core of permeability μ. Assuming that the core has not reached a state of saturation, we can calculate the flux density B_t (in teslas) inside the material using the formula

$$B_t = 4\pi \times 10^{-7}\ \mu nI/s \approx 1.2566 \times 10^{-6}\ \mu nI/s$$

If we want to calculate the flux density B_g (in gauss), we can use the formula

$$B_g = 4\pi \times 10^{-3}\ \mu nI/s \approx 0.012566\ \mu nI/s$$

Problem 8-5

Imagine a DC electromagnet that carries a certain current. It measures 20 cm long, and has 100 turns of wire. The core, which has permeability $\mu = 100$, has not reached a state of saturation. We measure the flux density inside it as $B_g = 20$ G. How much current flows in the coil?

Solution

Let's start by ensuring that we use the proper units in our calculation. We're told that the electromagnet measures 20 cm in length, so we can set $s = 0.20$ m. The flux density equals 20 G. Using algebra, we can rearrange the second of the above formulas so that it solves for I. We start with

$$B_g = 0.012566\ \mu nI/s$$

Dividing through by I, we get

$$B_g/I = 0.012566\ \mu n/s$$

When we divide both sides by B_g, we obtain

$$I^{-1} = 0.012566\ \mu n/(sB_g)$$

Finally, we take the reciprocal of both sides to get

$$I = 79.580\ sB_g / (\mu n)$$

Now we can input the numbers from the statement of the problem. We calculate

$$\begin{aligned} I &= 79.580\ sB_g / (\mu n) = 79.580 \times 0.20 \times 20 / (100 \times 100) \\ &= 79.580 \times 4.0 \times 10^{-4} = 0.031832\ \text{A} = 31.832\ \text{mA} \end{aligned}$$

We should round this result off to 32 mA.

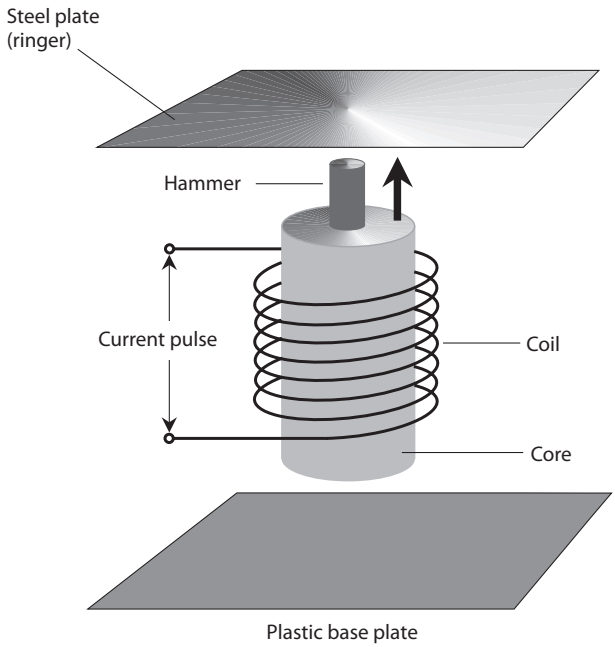
Magnetic Machines

Electrical relays, bell ringers, electric "hammers," and other mechanical devices make use of the principle of the solenoid. Sophisticated electromagnets, sometimes in conjunction with permanent magnets, allow us to construct motors, meters, generators, and other electromechanical devices.

The Chime

Figure 8-5 illustrates a *bell ringer*, also known as a *chime*. Its solenoid comprises an electromagnet. The ferromagnetic core has a hollow region in the center along its axis, through which a steel rod, called the *hammer*, passes. The coil has many turns of wire, so the electromagnet produces a high flux density if a substantial current passes through the coil.

When no current flows in the coil, gravity holds the rod down so that it rests on the plastic base plate. When a pulse of current passes through the coil, the rod moves upward at high speed. The magnetic field "wants" the ends of the rod, which has the same length as the core, to align with the



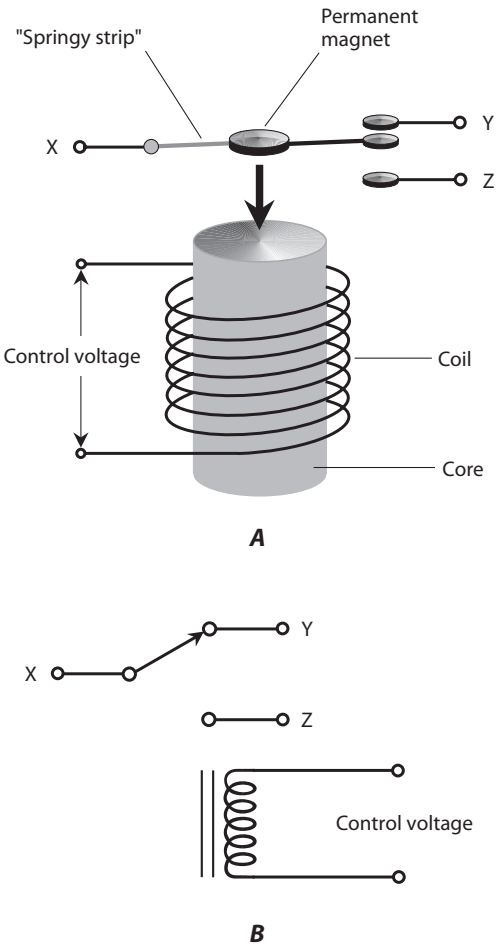
8-5 A bell ringer, also known as a chime.

ends of the core. But the rod's upward momentum causes it to pass through the core and strike the ringer. Then the steel rod falls back to its resting position, allowing the ringer to reverberate.

The Electromechanical Relay

We can't always locate switches near the devices they control. For example, suppose that you want to switch a communications system between two different antennas from a station control point 50 meters away. Wireless antenna systems carry high-frequency AC (the radio signals) that must remain within certain parts of the circuit. A *relay* makes use of a solenoid to allow remote-control switching.

Figure 8-6A illustrates a simple relay, and Fig. 8-6B shows the schematic diagram for the same device. A movable lever, called the *armature*, is held to one side (upward in this diagram) by a spring when no current flows through the coil. Under these conditions, terminal X contacts terminal Y, but X does not contact Z. When a sufficient current flows in the coil, the armature moves to the other side (downward in this illustration), disconnecting terminal X from terminal Y, and connecting X to Z.



8-6 Simplified drawing of a relay (at A) and the schematic symbol for the same relay (at B).

A *normally closed relay* completes the circuit when no current flows in the coil, and breaks the circuit when coil current flows. ("Normal," in this sense, means the absence of coil current.) A *normally open relay* does the opposite, completing the circuit when coil current flows, and breaking the circuit when coil current does not flow. The relay shown in Fig. 8-6 can function either as a normally open relay or a normally closed relay, depending on which contacts we select. It can also switch a single line between two different circuits.

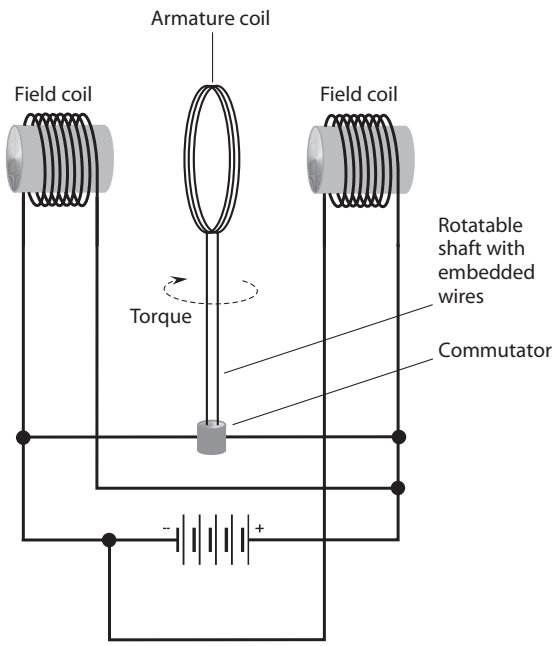
These days, engineers install relays primarily in circuits and systems that must handle large currents or voltages. In applications in which the currents and voltages remain low to moderate, electronic semiconductor switches, which have no moving parts, offer better performance and reliability than relays.

The DC Motor

Magnetic fields can produce considerable mechanical forces. We can harness these forces to perform useful work. A *DC motor* converts DC into rotating mechanical energy. In this sense, a DC motor constitutes a specialized *electromechanical transducer*. Such devices range in size from *nanoscale* (smaller than a bacterium) to *megascale* (larger than a house). Nanoscale motors can circulate in the human bloodstream or modify the behavior of internal body organs. Megascale motors can pull trains along tracks at hundreds of kilometers per hour.

In a DC motor, we connect a source of electricity to a set of coils, producing magnetic fields. The attraction of opposite poles, and the repulsion of like poles, is switched in such a way that a constant torque (rotational force) results. As the coil current increases, so does the torque that the motor can produce—and so does the energy it takes to operate the motor at a constant speed.

Figure 8-7 illustrates a DC motor in simplified form. The *armature coil* rotates along with the motor shaft. A set of two coils called the *field coil* remains stationary. Some motors use a pair of



8-7 Simplified drawing of a DC motor.

permanent magnets instead of a field coil. Every time the shaft completes half a rotation, the *commutator* reverses the current direction in the armature coil, so the shaft torque continues in the same angular direction. The shaft's *angular* (rotational) *momentum* carries it around so that it doesn't "freeze up" at the points in time when the current reverses direction.

Electric Generator

The construction of an *electric generator* resembles that of an electric motor, although the two devices function in the opposite sense. We might call a generator a specialized *mechano-electrical transducer* (although I've never heard anybody use that term). Some generators can also operate as motors; we call such devices *motor-generators*.

A typical generator produces AC when a coil rotates in a strong magnetic field. We can drive the shaft with a gasoline-powered motor, a turbine, or some other source of mechanical energy. Some generators employ commutators to produce pulsating DC output, which we can *filter* to obtain pure DC for use with precision equipment.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

2
PART
Alternating Current

This page intentionally left blank

9
CHAPTER

Alternating-Current Basics

WE CAN EXPRESS DC IN TERMS OF TWO VARIABLES: DIRECTION (POLARITY) AND INTENSITY (AMPLITUDE). If we want to gain a full understanding of alternating current (AC), we must work a little harder.

Definition of AC

For most of us, the term "AC" is associated with the idea of AC outlets and high voltage. While AC outlets are indeed AC, what we really mean in this and the following few chapters is mostly concerned with low-voltage AC.

Direct current (DC) doesn't really change: a resistor is powered and dissipating heat or it isn't, but the key feature of AC is that it is concerned with signals that are changing. Capacitors and inductors are components that only really do anything useful when we start to look at changing voltages and currents.

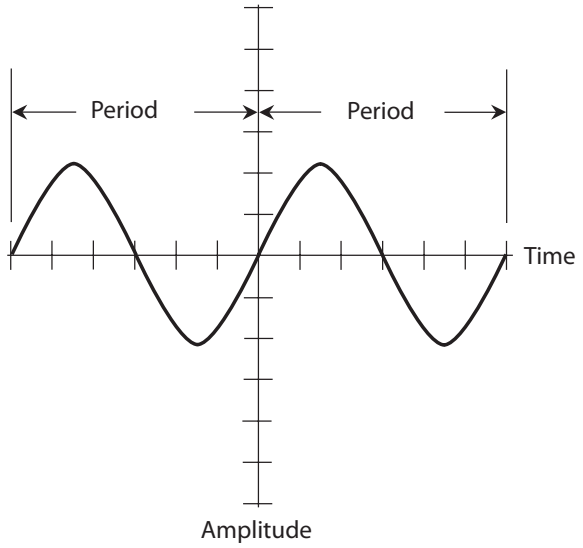
DC or voltage has polarity that remains constant as time passes. Although the amplitude (the number of amperes, volts, or watts) can fluctuate from moment to moment, the charge carriers always flow in the same direction at any point in the circuit, and the charge poles always keep the same relative orientation.

In AC, the charge-carrier flow and the polarity reverse at regular intervals. The *instantaneous amplitude* (the amplitude at any given instant in time) of AC usually varies because of the repeated reversal of polarity. But in some cases, the amplitude remains constant, even though the polarity keeps reversing. The repeating change in polarity makes AC fundamentally different from DC.

Most of electronics is concerned with AC and it is essential to understand AC to understand how things like audio amplifiers and radios work. Just as you have learned how to use Ohm's law with DC, there are some mathematical models that help us understand the behaviors of AC.

Period and Frequency

In a *periodic AC wave*, the kind that we'll work with in this chapter (and throughout the rest of this book), the function of *instantaneous amplitude versus time* repeats, and the same pattern recurs indefinitely. The length of time between one complete iteration of the wave pattern, or one *cycle*,



9-1 A sine wave. The period represents the time it takes for one cycle to complete itself.

and the next iteration constitutes the *period* of the wave. Figure 9-1 shows two complete cycles, and therefore, two periods, of a simple AC wave. In theory, the period of a wave can range from a tiny fraction of a second to millions of years. When we want to express the period of an AC wave in seconds, we can denote it by writing T.

In the "olden days," scientists and engineers specified AC frequency in *cycles per second*, abbreviated *CPS*. They expressed medium and high frequencies in *kilocycles*, *megacycles*, or *gigacycles*, representing thousands, millions, or billions (thousand-millions) of cycles per second. Nowadays, we express frequency in *hertz*, abbreviated Hz. The hertz and the cycle per second refer to exactly the same thing, so 1 Hz = 1 CPS, 10 Hz = 10 CPS, and so on. We can express medium and high frequencies in *kilohertz* (kHz), *megahertz* (MHz), or *gigahertz* (GHz), where

$$1 \text{ kHz} = 1000 \text{ Hz}$$

$$1 \text{ MHz} = 1000 \text{ kHz} = 1{,}000{,}000 \text{ Hz} = 10^6 \text{ Hz}$$

$$1 \text{ GHz} = 1000 \text{ MHz} = 1{,}000{,}000{,}000 \text{ Hz} = 10^9 \text{ Hz}$$

Sometimes we'll need an even bigger unit, the *terahertz* (THz), to specify AC frequency. A frequency of 1 THz constitutes a trillion (1,000,000,000,000 or 10^{12}) hertz. Electrical currents rarely attain such frequencies, although some forms of *electromagnetic radiation* do.

The frequency of an AC wave in hertz, denoted f, equals the reciprocal of the period T in seconds. Mathematically, we have

$$f = 1/T$$

and

$$T = 1/f$$

Some AC waves contain all their energy at a single frequency. We call such waves *pure*. But often, an AC wave contains energy components at multiples of the main, or *fundamental*, frequency. Components can sometimes also exist at frequencies that don't seem to bear any logical relation to the fundamental. Once in a while, we'll encounter a *complex AC wave* that contains energy at hundreds, thousands, or even infinitely many different component frequencies.

The Sine Wave

In its simplest form, AC has a *sine-wave*, or *sinusoidal*, nature. In a sine wave, the direction of the current reverses at regular intervals, and the current-versus-time curve follows the graph that we get when we plot the trigonometric *sine function* on a coordinate grid. Figure 9-1 shows the general shape of a sine wave—two complete cycles of it.

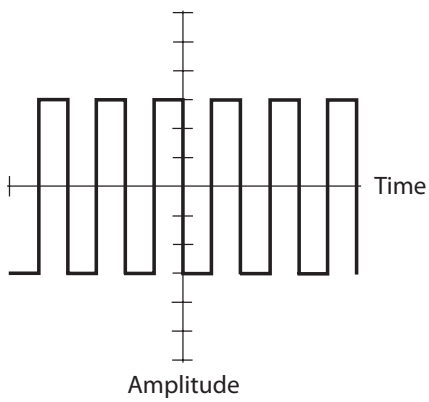
Whenever an AC wave contains all of its energy at a single frequency, it will exhibit a perfectly sinusoidal shape when we graph its amplitude as a function of time. Conversely, if an AC wave constitutes a perfect sinusoid, then all of its energy exists at a single frequency.

In practice, a wave might look like a sinusoid when displayed on an oscilloscope, even though it actually contains some imperfections too small to see. Nevertheless, a pure, single-frequency AC wave not only looks sinusoidal when we look at it on an oscilloscope screen, but it actually constitutes a *flawless* sinusoid.

Square Waves

As seen on an oscilloscope display, a *square wave* looks, well, square or rectangular. There is no smooth transition between the positive and negative cycles, the change is (at least theoretically) instant.

True square waves have equal negative and positive peaks, so the absolute amplitude of the wave never varies. Half of the time it's $+x$, and the other half of the time it's $-x$ (where x represents a



9-2 A square wave.

certain fixed number of amperes or volts). Some squared-off waves appear "lopsided" because the negative and positive amplitudes differ. Still others remain at positive polarity longer than they remain at negative polarity (or vice-versa). They constitute examples of *asymmetrical square waves*, more properly called *rectangular waves*.

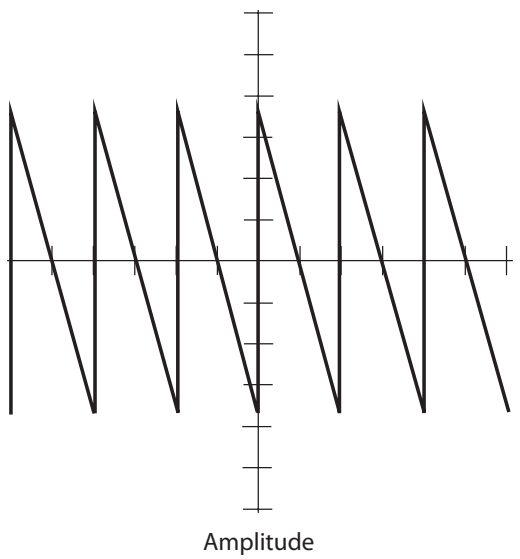
Sawtooth Waves

Some AC waves rise and/or fall in straight, sloping lines, as seen on an oscilloscope screen. The slope of the line indicates how fast the amplitude changes. We call them *sawtooth waves* because they resemble the "teeth" on a saw blade. Various electronic test devices and sound synthesizers can generate sawtooth waves with diverse frequencies and amplitudes.

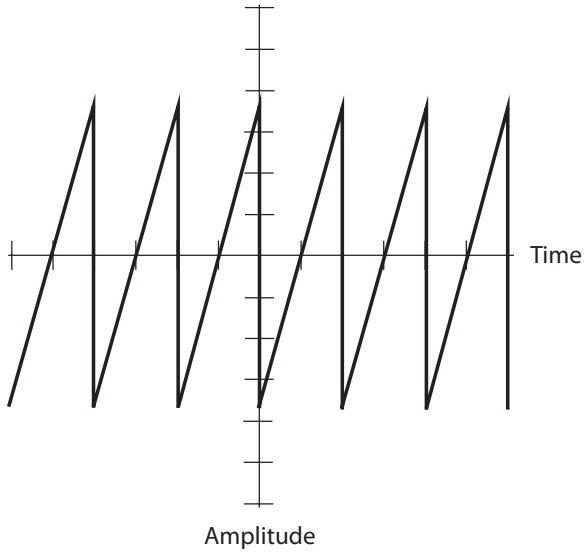
Figure 9-3 shows a sawtooth wave in which the positive-going slope (called the *rise*) is essentially instantaneous as in a square or rectangular wave, but the negative-going slope (called the *decay*) is not so steep. The period of the wave equals the time between points at identical positions on two successive pulses.

Another form of sawtooth wave exhibits a defined, finite rise and an instantaneous decay. Engineers call it a *ramp* because each individual cycle looks like an incline going upwards (Fig. 9-4). Engineers use ramps in scanning circuits for old-fashioned television receivers and oscilloscopes. The ramp tells the electron beam to move, or *trace*, at constant speed from left to right across the *cathode-ray-tube* (CRT) screen during the rise. Then the ramp retraces, or brings the electron beam back, instantaneously during the decay so the beam can begin the next trace across the screen.

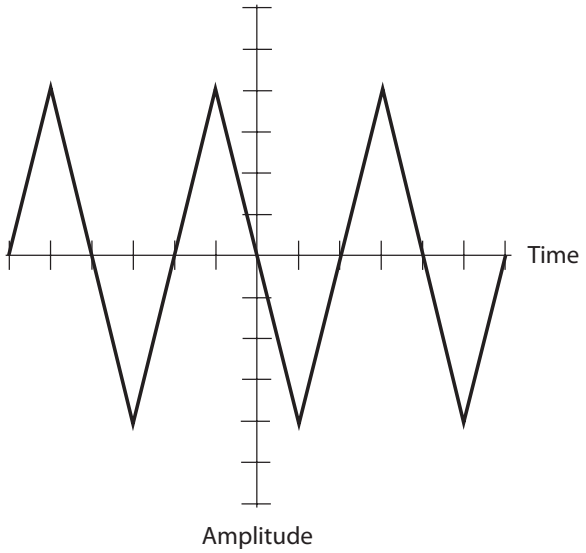
Sawtooth waves can have rise and decay slopes in an infinite number of different combinations. Figure 9-5 shows a common example. In this case, neither the rise nor the decay occurs instantaneously; the rise time equals the decay time. As a result, we get a so-called *triangular wave*.



9-3 A sawtooth wave with a fast rise and a slow decay.



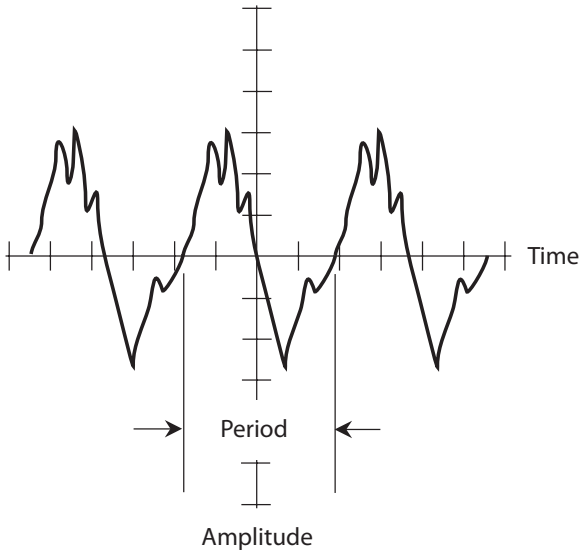
9-4 A sawtooth wave with a slow rise and a fast decay.



9-5 A triangular wave with equal rise and decay rates.

Complex Waveforms

As long as a wave has a definite period, and provided that the polarity keeps switching back and forth between positive and negative, it constitutes AC, no matter how complicated the actual shape of the waveform appears. Figure 9-6 shows an example of a complex AC wave. It has a definable period, and therefore, a definable frequency. The period equals the time between two points on succeeding wave repetitions.



9-6 A complex waveform.

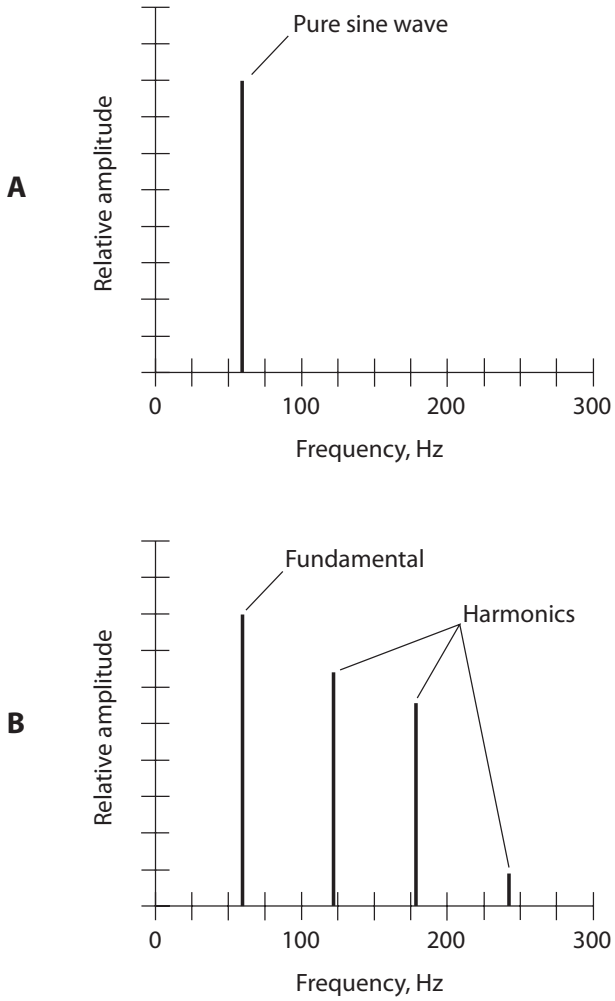
With some waves, we'll find it difficult or impossible to ascertain the period. This sort of situation can occur when a wave has two components of the same amplitude. Such a wave exhibits a multifaceted *frequency spectrum*; it contains equal amounts of energy at two different frequencies, so we can't decide whether to think about the part with the shorter period (the higher-frequency component) or the part with the longer period (the lower-frequency component).

Frequency Spectrum

An oscilloscope usually shows a graph of amplitude as a function of time. Time appears on the horizontal axis and represents the *independent variable* or *domain* of the function. But suppose you want to see the amplitude of a complex signal as a function of frequency, rather than as a function of time? You can do it with the help of an instrument called a spectrum analyzer—although many digital oscilloscopes can also operate as spectrum analyzers. Its horizontal axis shows frequency as the independent variable, ranging from some adjustable minimum frequency (at the extreme left) to some adjustable maximum frequency (at the extreme right).

An AC sine wave, as displayed on a spectrum analyzer, appears as a single *pip,* or vertical line, as shown in Fig. 9-7A. The wave concentrates all of its energy at a single frequency. But many, if not most, AC waves contain *harmonic* energy along with energy at the fundamental frequency. A harmonic is a secondary wave that occurs at a whole-number multiple of an AC wave's fundamental frequency. For example, if we have an AC wave whose fundamental frequency equals 60 Hz, then harmonics can exist at 120 Hz, 180 Hz, 240 Hz, and so on. The 120-Hz wave constitutes the *second harmonic*, the 180-Hz wave represents the *third harmonic*, the 240-Hz wave constitutes the *fourth harmonic*, and so on.

In general, if a wave has a frequency equal to n times the fundamental (where n equals some whole number), then we call that wave the *nth harmonic*. Figure 9-7B illustrates a wave's

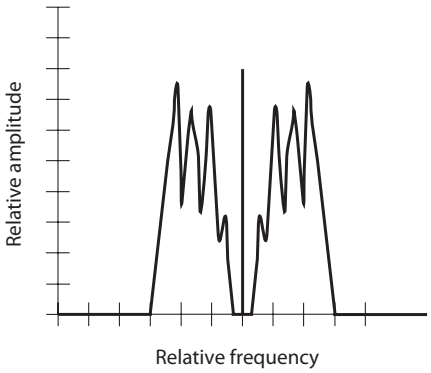


9-7 At A, a spectral diagram of a pure, 60-Hz sine wave. At B, a spectral diagram of a 60-Hz wave with three harmonics.

fundamental pip along with several harmonic pips as they would look on the display screen of a spectrum analyzer.

Square waves and sawtooth waves contain harmonic energy in addition to energy at the fundamental frequency. Other waves that contain a lot of harmonic energy can get more complicated. The exact shape of a wave depends on the amount of energy in the harmonics, and the way in which this energy is distributed among them.

Irregular waves can have any imaginable frequency distribution. Figure 9-8 shows a spectral (frequency-domain) display of an *amplitude-modulated* (AM) voice radio signal. Much of the energy concentrates at the frequency shown by the vertical line. That line portrays the signal's *carrier frequency*. We can also see evidence of energy near, but not exactly at, the carrier frequency. That part of the signal contains the voice information, technically called the signal's *intelligence*.



9-8 A spectral diagram of a modulated radio signal.

Fractions of a Cycle

Engineers break the AC cycle down into small parts for analysis and reference. We can compare a complete cycle to a single revolution around a circle, especially when we work with pure sine waves.

Degrees

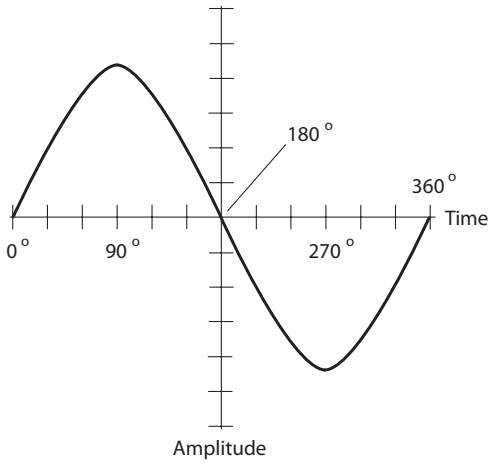
Engineers commonly divide an AC cycle into 360 equal increments called *degrees* or *degrees of phase*, symbolized by the standard degree symbol like the one used for temperature (°). We assign 0° to the point in the cycle where the wave magnitude is zero and positive-going. We give the same point on the next cycle the value 360°. In between these two extremes, we have values such as the following:

- The point ⅛ of the way through the cycle corresponds to 45°.
- The point ¼ of the way through the cycle corresponds to 90°.
- The point ⅜ of the way through the cycle corresponds to 135°.
- The point ½ of the way through the cycle corresponds to 180°.
- The point ⅝ of the way through the cycle corresponds to 225°.
- The point ¾ of the way through the cycle corresponds to 270°.
- The point ⅞ of the way through the cycle corresponds to 315°.

You can doubtless imagine other points, or calculate them. You can multiply the fractional part of the cycle by 360° to get the number of degrees to which a particular point corresponds. Conversely, you can divide the number of degrees by 360 to get the fractional part of the cycle. Figure 9-9 illustrates a perfect sine wave cycle and the various degree points we encounter as we proceed through the cycle from beginning to end.

Radians

As an alternative to the degree scheme, we can break up a wave cycle into 2π equal parts, where π (pi) represents the circumference-to-diameter ratio of a circle. This constant equals approximately 3.1416. A *radian* (rad) of phase equals roughly 57.296°. Physicists often express the frequency of a wave in *radians per second* (rad/s) rather than in hertz. Because a complete 360° cycle comprises 2π radians, the *angular frequency* of a wave, in radians per second, equals 2π (approximately 6.2832) times the frequency in hertz.



9-9 We can divide a single wave cycle into 360 equal degrees.

Phase Difference

Even if two AC waves have exactly the same frequency, they can produce different effects if they exist "out of sync" with each other. This phenomenon occurs in vivid fashion when AC waves combine to produce a third, or *composite*, wave. We can observe several "factoids" about combinations of pure sine waves, meaning AC waves that maintain constant frequency and that have equal (although opposite) positive and negative peak amplitudes.

- If two pure AC sine waves have the same frequency and the same intensity but differ in phase by 180° (½ cycle), then they precisely cancel each other out, and we don't observe any signal at all.
- If two pure AC sine waves have the same frequency and the same intensity, and if they coincide in phase (*phase coincidence*), then the composite wave has the same frequency and the same phase, but twice the intensity, of either wave alone.
- If two pure AC sine waves have the same frequency but different intensities, and if they differ in phase by 180°, then the composite signal has the same frequency as the originals, an intensity equal to the difference between the two, and a phase that coincides with the stronger of the two.
- If two pure AC sine waves have the same frequency and different intensities, and if they coincide in phase, then the composite wave has the same frequency and the same phase as the originals, and an intensity equal to the sum of the two.
- If two pure AC sine waves have the same frequency but differ in phase by some odd amount, such as 75° or 2.1 rad, then the resulting signal has the same frequency as the originals, but does not necessarily have the same wave shape, the same intensity, or the same phase as either original. As you can imagine, an infinite variety of such cases can occur.

In the United States, most household electricity from wall outlets consists of a 60-Hz sine wave with only one phase component. However, most electric utility companies send the energy over long distances as a combination of three separate 60-Hz waves, each differing by 120° of phase, which corresponds to ⅓ of a cycle. We call this mode *three-phase AC*. Each of the three waves carries ⅓ of the total power.

Caution!

We can define the relative phase between two AC waves if—*but only if*—they have precisely the same frequency. If their frequencies differ even slightly, then we can't define the phase of one wave relative to the other. That's because one wave's train of cycles keeps "catching and passing" the other wave's train of cycles.

Expressions of Amplitude

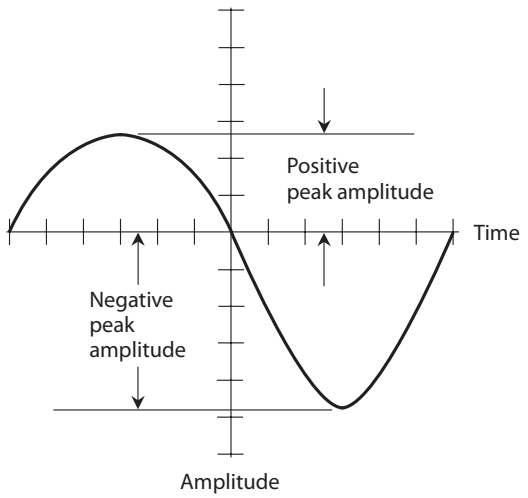
Depending on the quantity that we measure, we might specify the amplitude of an AC wave in amperes (for current), in volts (for voltage), or in watts (for power). In any case, we must remain aware of the time frame in which we measure or express the amplitude. Do we want to look at the amplitude of a wave at a single point in time, or do we want to express the amplitude as some quantity that does not depend on time?

Instantaneous Amplitude

The *instantaneous amplitude* of an AC wave equals the amplitude at some precise moment, or instant, in time. This value constantly changes as the wave goes through its cycle. The manner in which the instantaneous amplitude varies depends on the waveform. We can represent instantaneous amplitude values as individual points on the graphical display of a waveform.

Peak Amplitude

The *peak* (pk) *amplitude* of an AC wave is the maximum extent, either positive or negative, that the instantaneous amplitude attains. In many situations, the *positive peak amplitude* (pk+) and the *negative peak amplitude* (pk−) of an AC wave represent exact mirror images of each other. But sometimes things get more complicated. Figure 9-9 shows a wave in which the positive peak amplitude equals the negative peak amplitude, the only difference being in the polarity. Figure 9-10 shows a "lopsided" wave with different positive and negative peak amplitudes.



9-10 A wave with unequal positive and negative peak amplitudes.

In rigorous terms, we can define the positive peak amplitude of a waveform as the positive (upward) displacement from the horizontal axis to the *maximum* instantaneous voltage amplitude point on the graph (the *crest* of the waveform). Conversely, we can define the negative peak amplitude as the negative (downward) displacement from the horizontal axis to the *minimum* instantaneous amplitude point (the *trough* of the waveform).

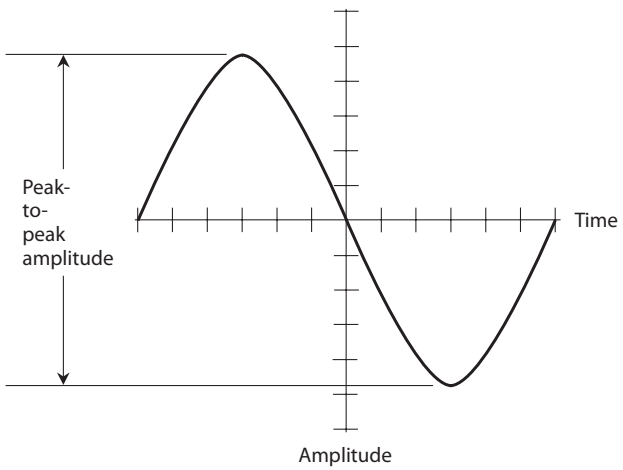
Peak-to-Peak Amplitude

The *peak-to-peak* (pk-pk) *amplitude* of a wave equals the net difference between the positive peak amplitude and the negative peak amplitude, as shown in Fig. 9-11. When the positive and negative peak amplitudes of an AC wave have equal extent and opposite polarity (as they do in Fig. 9-9, for example), then the peak-to-peak amplitude equals exactly 2 times the positive peak amplitude, and exactly −2 times the negative peak amplitude. However, we can't make such simplistic statements for a "lopsided" wave (as we see in Fig. 9-10, for example).

Root-Mean-Square (RMS) Amplitude

Often, we'll want to quantify or express the *effective amplitude* of an AC wave. The effective amplitude of an AC wave constitutes the voltage, current, or power that a DC source would have to produce in order to cause the same general effect as the AC wave does. When someone tells you that a wall outlet provides 117 V, they usually mean 117 effective volts. Effective voltage usually differs from the peak voltage or the peak-to-peak voltage.

The most common expression for effective AC wave intensity is the *root-mean-square* (RMS) amplitude. This terminology reflects how we "operate on" the AC waveform. We start by squaring all the instantaneous amplitude values, point-by-point, to make everything positive. Then we average the resulting "all-positive" wave over one full cycle. Finally we take the square root of that average. The mathematics gets advanced, but the RMS scheme reflects real-world wave behavior. In practice, electronic metering devices do the "calculations" for us, providing direct RMS readings.



9-11 Peak-to-peak (pk-pk) amplitude of a sine wave.

Sine-Wave Value Basics

We can state five simple rules that apply to pure sine waves with equal and opposite positive and negative peak amplitudes:

1. The RMS value is approximately 0.707 times the positive peak value or −0.707 times the negative peak value.
2. The RMS value is approximately 0.354 times the peak-to-peak value.
3. The positive peak value is approximately 1.414 times the RMS value, and the negative peak value is approximately −1.414 times the RMS value.
4. The peak-to-peak value is approximately 2.828 times the RMS value.
5. The average value is always zero (the sum of the positive and negative peak values, taking the polarity signs into account).

We'll often specify RMS amplitude when talking about utility AC, radio-frequency (RF) AC, and audio-frequency (AF) AC signals.

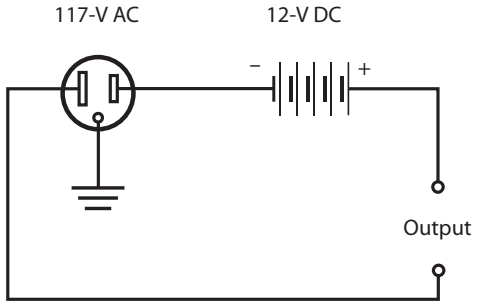
Other RMS Values

Non-sine waves follow different rules for RMS amplitude. In the case of a perfect square wave, for example, the RMS value equals the positive or negative peak value, which in turn equals half the peak-to-peak value. For sawtooth and irregular waves, the relationship between the RMS value and the peak value depends on the exact shape of the wave.

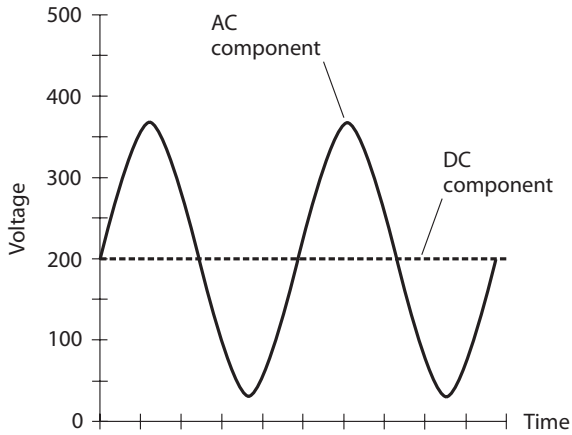
> **Tech Tidbit**
>
> We can't work out the average voltage of an irregular wave unless we know the exact waveform. That calculation usually needs a computer and a sophisticated waveform analyzer.

Superimposed DC

Sometimes a wave has components of both AC and DC. We can get an AC/DC combination by connecting a DC voltage source, such as a battery, in series with an AC voltage source, such as the utility mains. Figure 9-12 shows an example. Imagine connecting a 12-V battery in series with the wall outlet. Imagine it—but don't do it!



9-12 Connection of a DC source in series with an AC source.



9-13 Waveform resulting from a 117-V AC sine-wave source connected in series with a +200-V DC source.

Warning!
Do not try this experiment.
The battery could rupture, and its chemical paste or fluid could injure you.

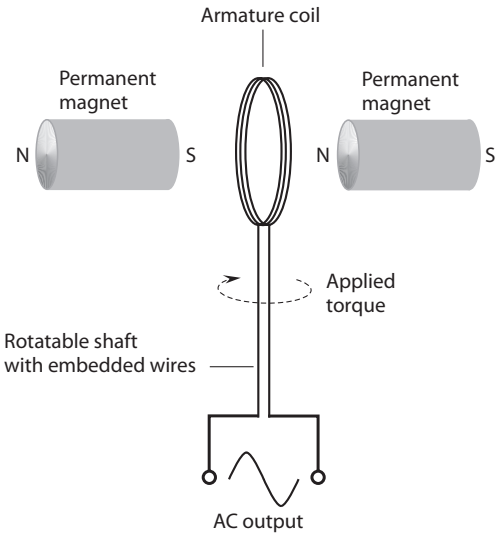
The AC wave is displaced either positively or negatively by 12 V, depending on the polarity of the battery. The voltage combination results in a sine wave at the output, but one peak exceeds the other peak by 24 V (twice the battery voltage).

Any AC wave can have a DC component superimposed. If the DC component exceeds the peak value of the AC wave, then fluctuating, or pulsating, DC will result. This would happen, for example, if a +200-V DC source were connected in series with the output of a common utility AC outlet, which has peak voltages of approximately ±165 V. Pulsating DC would appear, with an average value of +200 V but with instantaneous values much higher and lower (Fig. 9-13).

The Generator

We can generate AC by rotating a coil of wire in a magnetic field, as shown in Fig. 9-14. An AC voltage appears between the ends of the wire coil. The AC voltage that a generator can produce depends on the strength of the magnetic field, the number of turns in the wire coil, and the speed at which the coil rotates. The AC frequency depends only on the speed of rotation. Normally, for utility AC, this speed equals 3600 revolutions per minute (r/min), or 60 complete revolutions per second (RPS), so the AC output frequency equals 60 Hz.

When we connect a *load*, such as a light bulb or an electric space heater, to an AC generator, we will experience difficulty turning the generator shaft, as compared to *no-load conditions* (nothing connected to the output). As the amount of electrical power demanded from a generator increases, so does the mechanical power required to drive it. That's why we can't expect to connect a generator to a stationary bicycle and pedal an entire city into electrification. We can never get something for nothing. The electrical power that comes out of a generator can never exceed the mechanical power driving it. In fact, some energy always goes to waste, mainly as heat in the generator. Your



9-14 Functional diagram of an AC generator.

legs might generate enough power to run a small radio or television set, but nowhere near enough to provide electricity for a city.

The *efficiency* of a generator equals the ratio of the electrical output power to the mechanical driving power, both measured in the same units (such as watts or kilowatts), multiplied by 100 to get a percentage. No generator reaches 100 percent efficiency in the "real world," but a good one can come fairly close.

At power plants, massive turbines drive the electric generators. Heated steam, under pressure, forces the turbines to rotate. This steam is derived from natural sources of energy such as fossil-fuel combustion, nuclear reactions, or heat from deep inside the earth. In some power plants, moving water directly drives the turbines. In still other facilities, wind drives them. Any of these energy sources, properly harnessed, can provide tremendous mechanical power, explaining why power plants can produce megawatts of electrical power.

Why AC and Not DC?

Do you wonder why the electric utility companies produce AC instead of DC? Well, AC may seem more complicated than DC in theory, but in practice AC has proven simpler to implement when we want to provide electricity to a large number of people. Electricity in the form of AC lends itself to voltage transformation, but in the form of DC it does not. Electrochemical cells produce DC directly, but they can't supply large populations. To serve millions of consumers, we need the immense power of falling or flowing water, the ocean tides, wind, fossil fuels, controlled nuclear reactions, or geothermal heat. All of these energy sources can drive turbines that turn AC generators.

Technology continues to advance in the realm of solar-electric energy. These days a significant part of our electricity might come from *photovoltaic* power plants. These would generate DC. We could obtain high voltages by connecting giant arrays of solar panels in series.

Thomas Edison favored DC over AC for electrical power transmission before the electric infrastructure had been designed and developed. His colleagues argued that AC would work better. But Edison knew something that his contemporaries apparently preferred to ignore, if they knew it at all. At extremely high voltages, DC travels more efficiently over long distances than AC does. Long lengths of

wire exhibit less *effective resistance* (also called *ohmic loss*) with DC than with AC, and less energy goes to waste in the form of magnetic fields surrounding the wire. Interestingly, "interconnects" such as the link from mainland Europe to the UK to allow sharing of electrical energy operate at very-high-voltage DC rather than AC. So perhaps Edison could claim a belated victory.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

10
CHAPTER

Inductance

IN THIS CHAPTER, YOU'LL LEARN ABOUT ELECTRICAL COMPONENTS THAT OPPOSE THE FLOW OF AC BY storing energy as magnetic fields. We call these devices *inductors*, and we call their action *inductance*. Inductors usually comprise wire coils, but even a length of wire or cable can form an inductor.

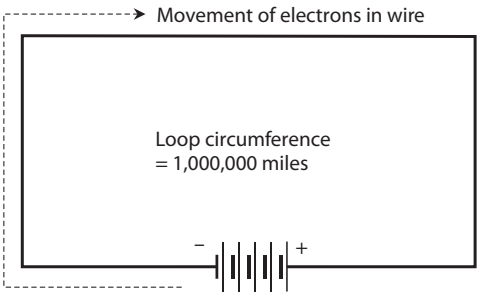
Along with, the much more frequently used, capacitors, inductors only do something interesting when used with AC.

The Property of Inductance

Imagine a wire 1,000,000 miles (about 1,600,000 km) long. Imagine that we make this wire into a huge loop, and then we connect its ends to the terminals of a battery, as shown in Fig. 10-1, driving current through the wire.

If we used a short wire for this experiment, the current would begin to flow immediately, and it would attain a level limited only by the resistance in the wire and the resistance in the battery. But because we have an extremely long wire, the electrons require some time to work their way from the negative battery terminal, around the loop, and back to the positive terminal. Therefore, it will take some time for the current to build up to its maximum level.

The magnetic field produced by the loop will start out small, during the first few moments when current flows in only part of the loop. The field will build up as the electrons get around the



10-1 We can use a huge, imaginary loop of wire to illustrate the principle of inductance.

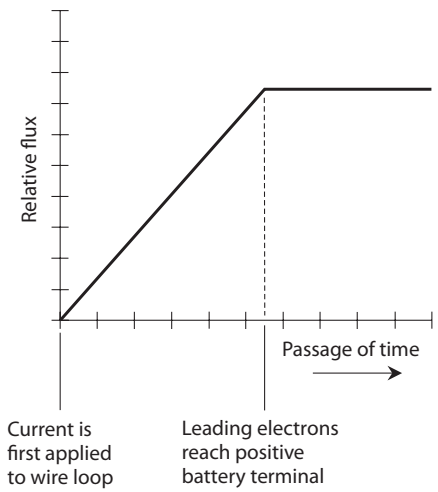
loop. Once the electrons reach the positive battery terminal so that a steady current flows around the entire loop, the magnetic field quantity will attain its maximum and level off, as shown in Fig. 10-2. At that time, we'll have a certain amount of energy stored in the magnetic field. The amount of stored energy will depend on the *inductance* of the loop, which depends on its overall size. We symbolize inductance, as a property or as a mathematical variable, by writing an italicized, uppercase letter *L*. Our loop constitutes an inductor. To abbreviate "inductor," we write an uppercase, non-italicized letter L.

Obviously, we can't make a wire loop measuring anywhere near 1,000,000 miles in circumference. But we can wind fairly long lengths of wire into compact coils. When we do that, the magnetic flux for a given length of wire increases compared with the flux produced by a single-turn loop, increasing the inductance. If we place a ferromagnetic rod called a *core* inside a coil of wire, we can increase the flux density and raise the inductance even more.

We can attain values of *L* many times greater with a ferromagnetic core than we can get with a similar-sized coil having an air core, a solid plastic core, or a solid dry wooden core. (Plastic and dry wood have permeability values that differ little from air or a vacuum; engineers occasionally use these materials as coil cores or "forms" in order to add structural rigidity to the windings without significantly changing the inductance.) The current that an inductor can handle depends on the diameter of the wire. But the value of *L* also depends on the number of turns in the coil, the diameter of the coil, and the overall shape of the coil.

If we hold all other factors constant, the inductance of a helical coil increases in direct proportion to the number of turns of wire. Inductance also increases in direct proportion to the diameter of the coil. If we "stretch out" a coil having a certain number of turns and a certain diameter while holding all other parameters constant, its inductance goes down. Conversely, if we "squash up" an elongated coil while holding all other factors constant, the inductance goes up.

Under normal circumstances, the inductance of a coil (or any other type of device designed to function as an inductor) remains constant regardless of the strength of the signal we apply. In this context, "abnormal circumstances" refer to an applied signal so strong that the inductor wire melts, or the core material heats up excessively. Good engineering sense demands that such conditions should never arise in a well-designed electrical or electronic system.



10-2 Relative magnetic flux in and around a huge loop of wire connected to a current source, as a function of time.

The Unit of Inductance

When we first connect a battery across an inductor, the current builds up at a rate that depends on the inductance. The greater the inductance, the slower the rate of current buildup for a given battery voltage. The unit of inductance quantifies the ratio between the rate of current buildup and the voltage across an inductor. An inductance of one *henry* (1 H) represents a potential difference of one volt (1 V) across an inductor within which the current increases or decreases at the rate of one ampere per second (1 A/s).

The henry constitutes a huge unit of inductance. You won't often see an inductor this large, although some power-supply filter chokes have inductances up to several henrys. Usually, engineers and technicians express inductances in *millihenrys* (mH), *microhenrys* (μH), or *nanohenrys* (nH). The units relate as follows:

$$1 \text{ mH} = 0.001 \text{ H} = 10^{-3} \text{ H}$$
$$1 \text{ μH} = 0.001 \text{ mH} = 10^{-6} \text{ H}$$
$$1 \text{ nH} = 0.001 \text{ μH} = 10^{-9} \text{ H}$$

Small coils with few turns of wire produce small inductances, in which the current changes quickly and the induced voltages are small. Large coils with ferromagnetic cores, and having many turns of wire, have high inductances in which the current changes slowly and the induced voltages are large. The current from a battery, building up or dying down through a high-inductance coil, can give rise to a large potential difference between the end terminals of the coil—many times the voltage of the battery itself. Spark coils, such as those used in internal combustion engines, take advantage of this principle. That's why large coils present a deadly danger to people ignorant of the wiles of inductance!

Inductors in Series

Imagine that we place two or more current-carrying inductors in close proximity and connect them in series. As long as the magnetic fields around those inductors don't interact, their inductances add exactly as resistances in series do. The total inductance, also called the *net inductance*, equals the sum of the individual inductances. We must use the same size unit for all the inductors if we want this simple rule to work.

Suppose that we have inductances L_1, L_2, L_3, …, L_n connected in series. As long as the magnetic fields of the inductors don't interact—that is, as long as no *mutual inductance* exists—we can calculate the total inductance L using the formula

$$L = L_1 + L_2 + L_3 + \cdots + L_n$$

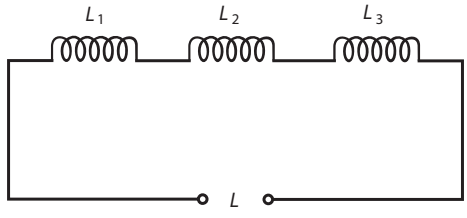
Problem 10-1

Imagine three inductances L_1, L_2, and L_3 connected in series, as shown in Fig. 10-3. Suppose that no mutual inductance exists, and each inductance equals 40.0 mH. What's the total inductance L?

Solution

We simply add up the values to obtain

$$L = L_1 + L_2 + L_3 = 40.0 + 40.0 + 40.0 = 120 \text{ mH}$$



10-3 Inductances connected in series.

Problem 10-2

Consider three inductors having no mutual inductance, with values of $L_1 = 20.0$ mH, $L_2 = 55.0$ μH, and $L_3 = 400$ nH. What's the total inductance L, in millihenrys, of these components if we connect them in series, as shown in Fig. 10-3?

Solution

First, let's convert all the inductances to the same units. Microhenrys will do! In that case, we have

$$L_1 = 20.0 \text{ mH} = 20{,}000 \text{ μH}$$
$$L_2 = 55.0 \text{ μH}$$
$$L_3 = 400 \text{ nH} = 0.400 \text{ μH}$$

The total inductance equals the sum of these values, or

$$L = 20{,}000 + 55.0 + 0.400 = 20{,}055.4 \text{ μH}$$

After we convert to millihenrys, we get 20.0554 mH, which we can round off to 20.1 mH.

Inductors in Parallel

If no mutual inductance exists among two or more parallel-connected inductors, their values add like resistances in parallel. Consider several inductances L_1, L_2, L_3, …, L_n connected in parallel. We can calculate the net inductance L, with the formula

$$\begin{aligned} L &= 1 / (1/L_1 + 1/L_2 + 1/L_3 + \cdots + 1/L_n) \\ &= (1/L_1 + 1/L_2 + 1/L_3 + \cdots + 1/L_n)^{-1} \end{aligned}$$

As with inductances in series, we must make sure that all the units agree.

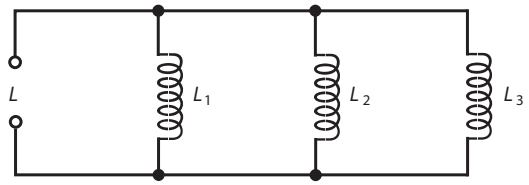
Problem 10-3

Imagine three inductances L_1, L_2, and L_3 connected in parallel as shown in Fig. 10-4. Suppose that no mutual inductance exists, and each inductance equals 40 mH. What's the total inductance L?

Solution

According to the formula defined above, we have

$$\begin{aligned} L &= 1 / (1/L_1 + 1/L_2 + 1/L_3) = 1 / (1/40 + 1/40 + 1/40) \\ &= 1 / (3/40) = 40/3 = 13.333 \text{ mH} \end{aligned}$$



10-4 Inductances connected in parallel.

We should round this figure off to 13 mH because the original inductance values extend to only two significant digits.

Problem 10-4

Imagine four inductances $L_1 = 75.0$ mH, $L_2 = 40.0$ mH, $L_3 = 333$ μH, and $L_4 = 7.00$ H, all connected in parallel with no mutual inductance. What's the net inductance L?

Solution

Let's use henrys as the standard unit. Then we have

$$L_1 = 0.0750 \text{ H}$$
$$L_2 = 0.0400 \text{ H}$$
$$L_3 = 0.000333 \text{ H}$$
$$L_4 = 7.00 \text{ H}$$

When we plug these values into the parallel-inductance formula, we obtain

$$\begin{aligned} L &= 1/(1/0.0750 + 1/0.0400 + 1/0.000333 + 1/7.00) \\ &= 1/(13.33 + 25.0 + 3003 + 0.143) \\ &= 1/3041.473 = 0.00032879 \text{ H} = 328.79 \text{ μH} \end{aligned}$$

We should round this figure off to 329 μH. That's only a little less than the value of the 333-μH inductor all by itself!

Interaction among Inductors

In real-world circuits, we usually observe some mutual inductance between or among *solenoidal* (cylindrical or helical) coils. The magnetic fields extend significantly outside such coils, and mutual effects are difficult to avoid. The same holds true between nearby lengths of wire, especially at high AC frequencies. Sometimes, mutual inductance has no detrimental effect, but in some situations it does. We can minimize mutual inductance between coils by using *toroidal* (donut-shaped) windings instead of solenoidal windings. We can minimize mutual inductance between wires by *shielding* them—that is, by insulating them and then wrapping them with grounded sheets or braids of metal. The most common shielded wire takes a form known as *coaxial cable*.

Coefficient of Coupling

The *coefficient of coupling,* symbolized k, quantifies the extent to which two inductors interact, that is, whether their magnetic fields reinforce or oppose each other. We specify k as a number ranging from 0 (no interaction) to 1 (the maximum possible interaction).

Two coils separated by a huge distance have the minimum possible coefficient of coupling, which is zero ($k = 0$); two coils wound on the same form, one right over the other, exhibit the maximum possible coefficient of coupling ($k = 1$). As we bring two inductors closer and hold all other factors constant, k increases.

We can multiply k by 100 and add a percent-symbol (%) to express the coefficient of coupling as a percentage, defining the range $k_{\%} = 0\%$ to $k_{\%} = 100\%$.

Mutual Inductance

Engineers symbolize the *mutual inductance* between two inductors by writing an uppercase italic M. We can express this quantity in the same units as inductance: henrys, millihenrys, microhenrys, or nanohenrys. For any two particular inductors, M depends on the inductance values and the coefficient of coupling.

When we have two inductors with values of L_1 and L_2 (both expressed in the same size units) and with a coefficient of coupling equal to k, we can calculate the mutual inductance by multiplying the inductances, taking the square root of the result, and then multiplying by k. Mathematically, we have

$$M = k\,(L_1 L_2)^{1/2}$$

where the ½ power represents the positive square root.

Effects of Mutual Inductance

Mutual inductance can either increase or decrease the net inductance of a pair of series connected coils, compared with the condition of zero mutual inductance. The magnetic fields around the coils either reinforce or oppose each other, depending on the phase relationship of the AC applied to them. If the two AC waves (and thus the magnetic fields they produce) coincide in phase, the inductance increases as compared with the condition of zero mutual inductance. If the two waves oppose in phase, the net inductance decreases relative to the condition of zero mutual inductance.

When we have two inductors connected in series and we observe *reinforcing* mutual inductance, we can calculate the total inductance L with the formula

$$L = L_1 + L_2 + 2M$$

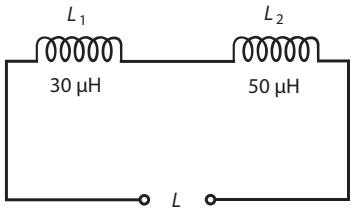
where L_1 and L_2 represent the inductances, and M represents the mutual inductance, all in the same size units. When we have two inductors connected in series and we observe *opposing* mutual inductance, we can calculate the total inductance L with the formula

$$L = L_1 + L_2 - 2M$$

where, again, L_1 and L_2 represent the values of the individual inductors, and M represents the mutual inductance, all in the same size units.

Problem 10-5

Imagine that we connect two coils, having inductances of 30 μH and 50 μH, in series so that their fields reinforce, as shown in Fig. 10-5. Suppose that the coefficient of coupling equals 0.500. What's the net inductance?



10-5 Illustration for Problem 10-5.

Solution

First, we must derive M from k. According to the formula for this purpose, we have

$$M = 0.500\ (50 \times 30)^{1/2} = 19.4\ \mu\text{H}$$

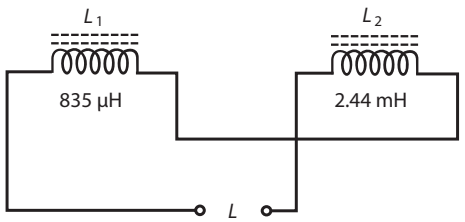
Now we can calculate the total inductance, getting

$$L = L_1 + L_2 + 2M = 30 + 50 + 38.8 = 118.8\ \mu\text{H}$$

which we should round off to 120 µH.

Problem 10-6

Imagine two coils with inductances of $L_1 = 835\ \mu\text{H}$ and $L_2 = 2.44$ mH. We connect them in series so that their magnetic fields oppose each other with a coefficient of coupling equal to 0.922, as shown in Fig. 10-6. What's the net inductance?



10-6 Illustration for Problem 10-6.

Solution

We know the coil inductances in different units. Let's use microhenrys for our calculations, so we have $L_1 = 835\ \mu\text{H}$ and $L_2 = 2440\ \mu\text{H}$. Now we can calculate M from k, obtaining

$$M = 0.922\ (835 \times 2440)^{1/2} = 1316\ \mu\text{H}$$

Finally, we calculate the total inductance as

$$L = L_1 + L_2 - 2M = 835 + 2440 - 2632 = 643\ \mu\text{H}$$

Air-Core Coils

The simplest inductors (besides plain, straight lengths of wire) are coils of insulated or enameled wire. You can wind a coil on a hollow cylinder made of plastic or other non-ferromagnetic material, forming an *air-core coil*. In practice, the attainable inductance for such coils can range from a few nanohenrys up to about 1 mH. The frequency of an applied AC signal does not affect the

inductance of an air-core coil, but as the AC frequency increases, smaller and smaller values of inductance produce significant effects.

An air-core coil made of heavy-gauge wire, and having a large radius, can carry high current and can handle high voltages. Air dissipates almost no energy as heat, so air makes an efficient core material even though it has low permeability. For these reasons, air-core coil designs represent an excellent choice for the engineer who wants to build high-power RF transmitters, amplifiers, or tuning networks. However, air-core coils take up a lot of physical space in proportion to the available inductance, especially when designed to handle high currents and voltages.

Ferromagnetic Cores

Inductor manufacturers crush samples of ferromagnetic material into dust and then bind the powder into various shapes, providing cores that greatly increase the inductance of a coil having a given number of turns. Depending on the mixture used, the flux density can increase by a factor of up to about 1,000,000 (10^6). A physically small coil can thereby acquire a large inductance if we put a *powdered-iron core* inside it. Powdered-iron cores work well from the middle audio frequencies (AF) to well into the radio-frequency (RF) range.

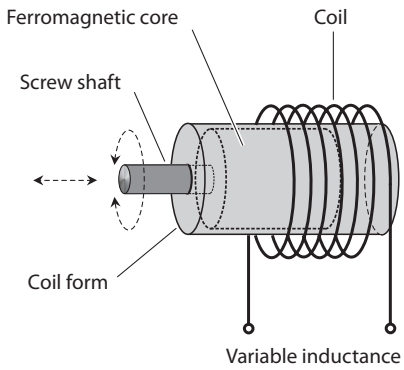
Core Saturation

If a powdered-iron-core coil carries more than a certain amount of current, the core will *saturate*. When an inductor core operates in a state of *saturation*, the ferromagnetic material holds as much magnetic flux as it possibly can. Any further increase in the coil current will not produce an increase in the core's magnetic flux. In practical systems, this effect causes decreasing inductance with coil currents that exceed the critical value. In extreme cases, saturation can cause a coil to waste considerable power as heat, making the coil *lossy*.

Permeability Tuning

If you've ever looked inside an old analog radio receiver, you might well have seen small metal box-like components with a hole in the top, where a screwdriver can be used to tune an inductor.

We can *tune* (vary the inductance of) a solenoidal coil without changing the number of turns if we slide a ferromagnetic core in and out of it. Because moving the core in and out of a coil changes the effective permeability within the coil, some engineers call this practice *permeability tuning*. We can precisely control the in/out core position by attaching the core to a screw shaft, as shown in Fig. 10-7. As we rotate the shaft clockwise, the core enters the coil, so the inductance increases. As we turn the shaft counterclockwise, the core moves out of the coil, so the inductance decreases.



10-7 We can accomplish permeability tuning by moving a ferromagnetic core in and out of a solenoidal coil.

Toroids

When we want to construct a coil with a ferromagnetic core, we don't have to wind the wire on a rod-shaped core. We can use another core geometry called the *toroid*, whose shape resembles that of a donut. Figure 10-8 illustrates how we can wind a coil on a ferromagnetic toroid core.

Toroidal coils offer at least three advantages over solenoidal ones. First, it takes fewer turns of wire to get a certain inductance with a toroid, as compared with a solenoid. Second, we can make a toroid physically smaller for a given inductance and current-carrying capacity. Third, essentially all of the magnetic flux in a toroid remains within the core material. This phenomenon practically eliminates unwanted mutual inductance between a toroid and other components near it.

Toroidal coils have limitations. We can't permeability-tune a toroidal coil because the core constitutes a full circle, always "serving" the entire coil. Most people find toroidal coils harder to wind than solenoidal ones, especially when the winding must have a large number of turns. In some situations, we might actually *want* mutual inductance to exist between or among physically separate coils; if we use a toroid, we must wind both coils on the same core to make this happen.

Pot Cores

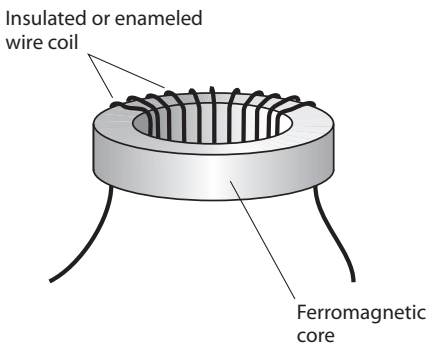
An alternative exists to the toroidal geometry for confining magnetic flux. We can surround a loop-shaped coil of wire with a ferromagnetic shell, as shown in Fig. 10-9, obtaining a *pot core*. A typical pot core has two halves, inside one of which we wind the coil. The wires emerge through small holes or slots.

Pot cores have advantages similar to those of toroids. The shell prevents the magnetic flux from extending outside the physical assembly, so mutual inductance between the coil and anything else in its vicinity is always zero. We can get far more inductance with a pot core than we can with a solenoidal coil of comparable physical size. In fact, pot cores work even better than toroids if we want to obtain a large inductance in a small space.

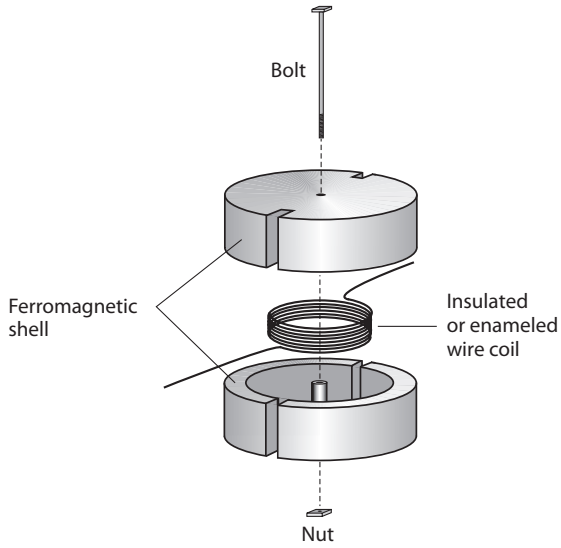
Pot-core coils can prove useful over the full AF range, even the lowest-frequency extreme (approximately 20 Hz). They don't function very well at frequencies above a few hundred kilohertz. Because of their geometry, we can't permeability-tune them. If we hold all other factors constant, then the inductance increases as the shell permeability increases. It doesn't matter, within reason, how strong or weak the signal is.

Inductors for RF Use

The RF spectrum ranges from a few kilohertz to well above 100 GHz. At the low end of this range, inductors generally use ferromagnetic cores. As the frequency increases, cores having low



10-8 A toroidal coil surrounds a donut-shaped ferromagnetic core.



10-9 Exploded view of a pot core. We wind the coil inside the ferromagnetic shell.

permeability find favor. We'll commonly see toroids in RF systems designed for use at frequencies up through about 30 MHz. Above 30 MHz or so, we'll more often see air-core coils.

Transmission-Line Inductors

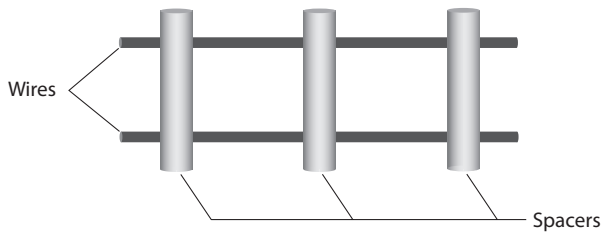
At frequencies about 100 MHz, we can make an inductor from a length of *transmission line*, rather than from a wire loop or coil. Most transmission lines exist in either of two geometries, the *parallel-wire* type or the *coaxial* type.

Parallel-Wire Line

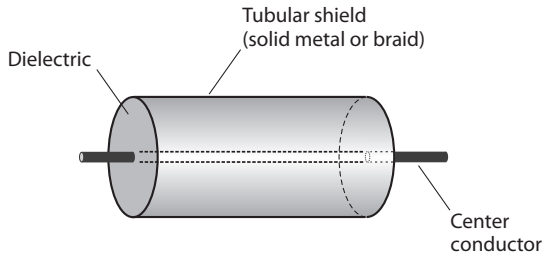
A parallel-wire transmission line consists of two wires running alongside each other with constant spacing (Fig. 10-10). Polyethylene rods, molded at regular intervals to the wires, keep the spacing between the wires constant. A solid or "windowed" web of polyethylene can serve the same purpose. The substance separating the wires constitutes the *dielectric* of the transmission line.

Coaxial line

A coaxial transmission line contains a wire *center conductor* surrounded by a tubular braid or pipe called the *shield* (Fig. 10-11). Solid polyethylene *beads* (which resemble tiny toroids in shape), or a continuous hose-like length of foamed or solid polyethylene, separates the center conductor from the shield, maintains the spacing between them, and acts as a dielectric.



10-10 Parallel-wire transmission line. The spacers consist of sturdy insulating material.



10-11 Coaxial transmission line. The dielectric material keeps the center conductor along the axis of the tubular shield.

Line Inductance

Short lengths of transmission line behave as inductors, provided that the line length remains less than 90° (¼ of a wavelength), and as long as we connect the line conductors *directly together* at the far end. We must use a little trickery if we want to design such an inductor. It will work only over a certain range of frequencies.

If f represents the frequency in megahertz, then we can calculate ¼ wavelength in *free space* (a vacuum), expressed in centimeters (s_{cm}), using the formula

$$s_{cm} = 7500/f$$

The length of a "real-world" quarter-wavelength transmission line is shortened from the free-space quarter wavelength because the dielectric reduces the speed at which the RF energy travels along the line. In practice, ¼ of an *electrical wavelength* along a transmission line can range anywhere from about 0.66 (or 66%) of the free-space quarter wavelength to 0.95 (or 95%) of the free-space quarter wavelength. Engineers call this "shortening factor" the *velocity factor* of the line because it represents the speed of RF waves in the line divided by the speed of RF waves in free space (the speed of light). If we let v represent the velocity factor of a particular transmission line, then the above formula for the length of a quarter-wave line, in centimeters, becomes

$$s_{cm} = 7500v/f$$

Very short lengths of line—a few electrical degrees—produce small values of inductance. As the length approaches ¼ wavelength, the inductance increases.

Transmission line inductors behave differently than coils in one important way. The inductance of a coil varies little, if at all, with changes in the frequency. But the value of a transmission-line inductor varies *drastically* as the frequency goes up or down. At first, the inductance increases as the frequency increases. As we approach a certain "critical" frequency, the inductance grows arbitrarily large, "approaching infinity." At the critical frequency, where the line measures precisely ¼ electrical wavelength from end to end, the line (as long as its conductors remain shorted out at the far end) acts like an open circuit at the signal-input end.

If we continue to increase the frequency of the applied AC signal so that the line's electrical length exceeds ¼ wavelength, the line acts as a *capacitor* rather than as an inductor. (You'll learn about capacitors in the next chapter.) The line continues to act as a capacitor up to the frequency at which it measures ½ electrical wavelength. Then, when the frequency reaches a second critical point at which the line measures precisely ½ electrical wavelength from end to end, the length of line behaves as a short circuit; conditions at the signal end mimic those at the far end. If we increase the frequency indefinitely, the line will act as an inductor again, then as an open circuit, then as

a capacitor again, then as a short circuit, then as an inductor again, and so on. Each critical (or transition) frequency will exist when the electrical length of the line exactly equals a whole-number multiple of ¼ wavelength.

Stray Inductance

Any length of wire, no matter how short and no matter what the frequency, exhibits at least a little inductance. As with a transmission line, the inductance of any fixed-length wire increases as the frequency increases.

In wireless communications equipment, the inductance of, and among, wires can constitute a problem. A circuit might generate its own signal (oscillate) even if that's the last thing we want it to do. A receiver might respond to signals that we don't want it to intercept. A transmitter can send out signals on unauthorized frequencies. The frequency response of any circuit can change in unpredictable ways, degrading the performance of the equipment. Sometimes the effects of so-called *stray inductance* remain small or negligible; in other scenarios, stray inductance can cause serious malfunctions or even *catastrophic system failure*.

If we want to minimize stray inductance, we can use coaxial cables between and among sensitive circuits or components. We must connect the shield of each cable section to the *common ground* of the apparatus. In some systems, we might even have to enclose individual circuits in tight metal boxes to *electrically isolate* them.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

11
CHAPTER

Capacitance

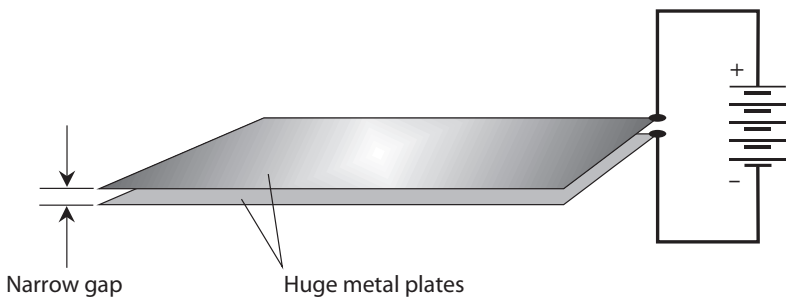
ELECTRICAL RESISTANCE SLOWS THE FLOW OF AC OR DC CHARGE CARRIERS (USUALLY ELECTRONS) BY "brute force." Inductance impedes the flow of AC charge carriers by storing the energy as a magnetic field. *Capacitance* impedes the flow of AC charge carriers by storing the energy as an *electric field*.

Like inductors, capacitors come into their own with AC. However, where many battery-powered electronics may have no inductors in them at all, it is very rare to find a circuit that doesn't include a few capacitors.

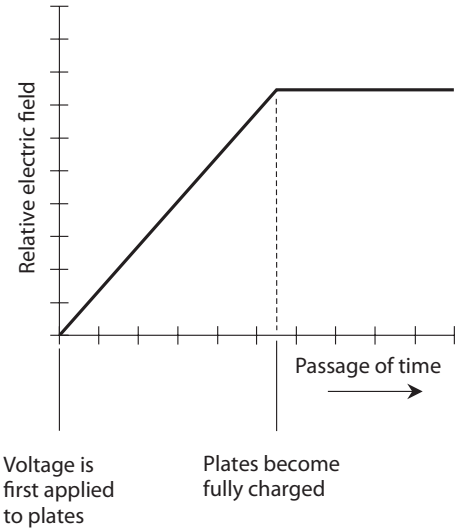
The Property of Capacitance

Imagine two gigantic, flat, thin sheets of metal that conduct electricity well. Suppose that they have equal surface areas the size of the state of Nebraska. We place them one above the other, keep them parallel to each other, and separate them by a few centimeters of space. If we connect these two sheets of metal to the terminals of a battery, as shown in Fig. 11-1, the sheets will charge electrically, one with positive polarity and the other with negative polarity.

If the plates were small, they would both charge up almost instantly, attaining a relative voltage equal to the voltage of the battery. However, because the plates are so large and massive, it will take a little time for the negative one to "fill up" with extra electrons, and it will take an equal amount of



11-1 A hypothetical capacitor comprising two huge conducting plates.



11-2 Relative electric field intensity, as a function of time, between two metal plates connected to a voltage source.

time for the other one to have its excess electrons "drained out." Eventually, the voltage between the two plates will equal the battery voltage, and an electric field will exist in the space between them.

The electric field will start out small immediately after we connect the battery because the plates can't charge up right away. The charge will increase over a period of time; the rate of increase will depend on the plates' surface areas and on the spacing between them. Figure 11-2 portrays the intensity of this electric field as a function of time. We define *capacitance* as the ability of plates such as these, and of the space between them, to store electrical energy. As a quantity or variable, we denote capacitance by writing an uppercase, italic letter *C*.

Simple Capacitors

Obviously, we can't physically construct a capacitor of the dimensions described above! But we can place two sheets or strips of thin metal foil together, keeping them evenly separated by a thin layer of nonconducting material such as paper or plastic, and then roll up the assembly to obtain a large mutual surface area in a small physical volume. When we do this, the electric field quantity and intensity between the plates can get large enough so that the device exhibits considerable capacitance. Alternatively, we can take two separate sets of several plates and mesh them together with thin layers of insulating ceramic material between them.

When we place a layer of solid *dielectric* material between the plates of a capacitor, the electric flux concentration increases several times—perhaps many times—without our having to increase the surface areas of the plates. In this way, we can get a physically small component to exhibit a large capacitance. The voltage that such a capacitor can handle depends on the thickness of the metal sheets or strips, on the spacing between them, and on the type of dielectric material that we use to build the component. In general, capacitance varies in direct proportion to the mutual surface area of the conducting plates or sheets, but inversely according to the separation between the conducting sheets. We can summarize these relations in four statements:

1. If we maintain constant spacing between the sheets and increase their mutual area, the capacitance goes up.

2. If we maintain constant spacing between the sheets and decrease their mutual area, the capacitance goes down.
3. If we maintain constant mutual area and move the sheets closer together, the capacitance goes up.
4. If we maintain constant mutual area and move the sheets farther apart, the capacitance goes down.

The capacitance of a particular component also depends on the *dielectric constant* of the material between the metal sheets or plates. Dielectric constants are represented as numbers. By convention, scientists and engineers assign a dielectric constant of exactly 1 to a vacuum. If we take a capacitor in which a vacuum exists between the metal sheets or plates and then fill up all of the space with a material whose dielectric constant equals k, then the capacitance will increase by a factor of k. Dry air has a dielectric constant of almost exactly 1 (a little more, but rarely worth our worry). Table 11-1 lists the dielectric constants for several common substances.

These days the majority of capacitors used in circuits are tiny surface mount devices (SMDs) that use a ceramic compound with a very high dielectric constant such as barium titanate.

The Unit of Capacitance

When we connect a battery between the plates of a capacitor, the potential difference between the plates builds up at a rate that depends on the capacitance. The greater the capacitance, the slower the rate of change of the voltage in the plates. The standard unit of capacitance quantifies the ratio between the current that flows and the rate of voltage change between the plates as the plates charge up. A capacitance of one *farad* (1 F) represents a current flow of 1 A while the voltage increases at the rate of 1 V/s. A capacitance of 1 F also results in 1 V of potential difference for an electric charge of 1 coulomb (1 C).

Table 11-1. Dielectric constants for several media that are used in capacitors. Except for air and a vacuum, these values are approximate

Substance	Dielectric constant (approx.)
Air, dry, at sea level	1.0
Glass	4.8–8.0
Mica	4.0–6.0
Mylar	2.9–3.1
Paper	3.0–3.5
Plastic, hard, clear	3.0–4.0
Polyethylene	2.2–2.3
Polystyrene	2.4–2.8
Polyvinyl chloride	3.1–3.3
Porcelain	5.3–6.0
Quartz	3.6–4.0
Strontium titanate	300–320
Barium titanate	1000–2000
Teflon	2.0–2.2
Titanium oxide	160–180
Vacuum	1.0

The farad represents a huge unit of capacitance. With the exception of a special class of capacitor called a "super-capacitor" or "ultra-capacitor" you won't see a real-world capacitor with a value of 1 F. Engineers express capacitance values in terms of the *microfarad* (μF), the *nanofarad* (nF), and the *picofarad* (pF), where

$$1\ \mu\text{F} = 0.000001\ \text{F} = 10^{-6}\ \text{F}$$
$$1\ \text{nF} = 0.001\ \mu\text{F} = 10^{-9}\ \text{F}$$
$$1\ \text{pF} = 0.001\ \text{nF} = 10^{-12}\ \text{F}$$

The *millifarad*, which theoretically represents 0.001 F or 10^{-3} F, has never gained common usage.

Hardware manufacturers can produce physically small components that have fairly large capacitance values. Conversely, some capacitors with small values take up large physical volumes. The physical size of a capacitor, if all other factors are held constant, is proportional to the voltage that it can handle. The higher the rated voltage, the bigger the component.

Capacitors in Series

We rarely observe any mutual interaction among capacitors. We don't have to worry about *mutual capacitance* very often, the way we have to think about mutual inductance when working with wire coils.

When we connect two or more capacitors in series, their values combine just as resistances combine in parallel, assuming that no mutual capacitance exists among the components. If we connect two capacitors of the same value in series, the net capacitance equals half the capacitance of either component alone. In general, if we have several capacitors connected in series, we observe a net capacitance smaller than that of any of the individual components. As with resistances and inductances, we should always use the same size units when we calculate the net capacitance of any combination.

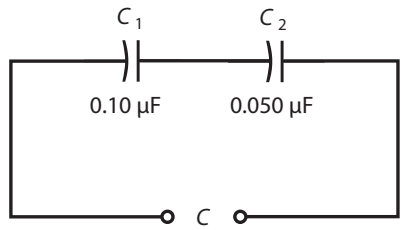
Consider several capacitors with values C_1, C_2, C_3, …, C_n connected in series. We can find the net capacitance C using the formula

$$C = 1 / (1/C_1 + 1/C_2 + 1/C_3 + \cdots + 1/C_n)$$

If we connect two or more capacitors in series, and if one of them has a value *many times* smaller than the values of all the others, then the net capacitance equals the *smallest* capacitance for most practical purposes.

Problem 11-1

Suppose that two capacitances, $C_1 = 0.10\ \mu\text{F}$ and $C_2 = 0.050\ \mu\text{F}$, appear in series, as shown in Fig. 11-3. What's the net capacitance?



11-3 Capacitors in series. Illustration for Problem 11-1.

Solution

Let's use microfarads as the unit for our calculations. Using the above formula, we first find the reciprocals of the individual capacitances, getting

$$1/C_1 = 10$$

and

$$1/C_2 = 20$$

We add these numbers to obtain the reciprocal of the net series capacitance:

$$1/C = 10 + 20 = 30$$

Finally, we take the reciprocal of C^{-1} to obtain

$$C = 1/30 = 0.033\ \mu\text{F}$$

Problem 11-2

Imagine that we connect two capacitors with values of 0.0010 μF and 100 pF in series. What's the net capacitance?

Solution

First, let's convert both capacitances to microfarads. A value of 100 pF represents 0.000100 μF, so $C_1 = 0.0010\ \mu\text{F}$ and $C_2 = 0.000100\ \mu\text{F}$. The reciprocals are

$$1/C_1 = 1000$$

and

$$1/C_2 = 10{,}000$$

Now we can calculate the reciprocal of the series capacitance as

$$1/C = 1000 + 10{,}000 = 11{,}000$$

Therefore

$$C = 1/11{,}000 = 0.000091\ \mu\text{F}$$

We can also state this capacitance as 91 pF.

Problem 11-3

Suppose that we connect five 100-pF capacitors in series. What's the total capacitance?

Solution

If we have n capacitors in series, all of the same value, then the net capacitance C equals $1/n$ of the capacitance of any one of the components alone. In this case we have five 100-pF capacitors in series, so we get a net capacitance of

$$C = 100/5 = 20.0\ \text{pF}$$

Why the Curved Lines?

Do you wonder why the capacitor symbols in Fig. 11-3 (and everywhere else that they appear in this book) consist of one straight line and one curved line? Engineers commonly use this notation

because, in many situations, one end of the capacitor either connects directly to, or faces toward, a *common ground* point (a neutral point, with a reference voltage of zero). In the circuit of Fig. 11-3, no common ground exists, so it doesn't matter which way the capacitors go. However, later in this book, we'll encounter some circuits in which the capacitor orientation does matter.

Capacitors in Parallel

Capacitances in parallel add like resistances in series. The net capacitance equals the sum of the individual component values, as long as we use the same units all the way through our calculations.

Suppose that we connect capacitors C_1, C_2, C_3, …, C_n in parallel. As long as we observe no mutual capacitance between the components, we can calculate the net capacitance C with the formula

$$C = C_1 + C_2 + C_3 + \cdots + C_n$$

If we parallel-connect two or more capacitors, and if one of them has a value *many times* larger than the values of all the others, then for most practical purposes the net capacitance equals the *largest* capacitance.

Problem 11-4

Imagine three capacitors connected in parallel, having values of $C_1 = 0.100$ μF, $C_2 = 0.0100$ μF, and $C_3 = 0.00100$ μF, as shown in Fig. 11-4. What's the net parallel capacitance?

Solution

We simply add the values up because all of the capacitances are expressed in the same size units (microfarads). We have

$$C = 0.100 + 0.0100 + 0.00100 = 0.11100 \text{ μF}$$

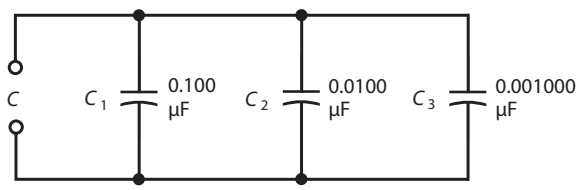
We can round off this result to $C = 0.111$ μF.

Problem 11-5

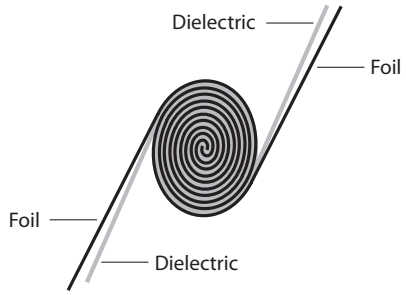
Suppose that we connect two capacitors in parallel, one with a value of 100 μF and one with a value of 100 pF. What's the net capacitance?

Solution

In this case, we can say straightaway that the net capacitance is 100 μF for practical purposes. The 100-pF capacitor has a value that's only 1/1,000,000 of the capacitance of the 100-μF component. The smaller capacitance contributes essentially nothing to the net capacitance of this combination.



11-4 Capacitors in parallel. Illustration for Problem 11-4.



11-5 Cross-sectional drawing of a capacitor made of two foil sheets rolled up with dielectric material between them.

Fixed Capacitors

A *fixed capacitor* has a value that we can't adjust, and that (ideally) does not vary when environmental or circuit conditions change. Several common types of fixed capacitors have been used for decades.

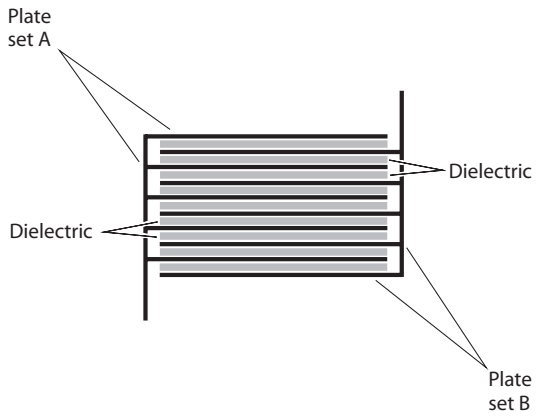
Paper Capacitors

In the early days of electronics, capacitors were commonly made by placing paper, soaked with mineral oil, between two strips of foil, rolling the assembly up (Fig. 11-5), attaching wire leads to the two pieces of foil, and enclosing the rolled-up foil and paper in an airtight cylindrical case. *Paper capacitors* can still sometimes be found in older electronic equipment. They have values ranging from about 0.001 μF to 0.1 μF, and can handle potential differences of up to about 1000 V.

Mica Capacitors

Mica comprises a naturally occurring, solid, transparent mineral substance that flakes off in thin sheets. It makes an excellent dielectric for capacitors. *Mica capacitors* can be manufactured by alternately stacking metal sheets and layers of mica, or by applying silver ink to sheets of mica. The metal sheets are wired together into two meshed sets, forming the two terminals of the capacitor, as shown in Fig. 11-6.

Mica capacitors have low loss, so they exhibit high efficiency as long as we don't subject them to excessive voltage. Mica-capacitor voltage ratings can range from a few volts (with thin mica sheets)



11-6 Cross-sectional drawing of a capacitor made of two meshed sets of several metal plates, separated by layers of dielectric material.

up to several thousand volts (with thick mica sheets and heavy-gauge metal plates). Mica capacitors occupy large physical volume in proportion to their capacitance. Mica capacitors work well in wireless receivers and transmitters. Values range from a few tens of picofarads up to approximately 0.05 μF.

Ceramic Capacitors

Ceramic materials work well as dielectrics. Sheets of metal are stacked alternately with wafers of ceramic to make these capacitors. A ceramic capacitor can have anywhere from one layer to dozens of layers. The geometry of Fig. 11-6 can serve as a general illustration. Ceramic, like mica, has low loss and allows for high efficiency.

Multilayer ceramic capacitors (MLCCs) are the most commonly used general-purpose capacitors in modern designs. Improvements in technology mean that they can span a vast range of capacitances from a few pF to many μF and still be extremely compact in size.

For small values of capacitance, we need only one disk-shaped layer of ceramic material; we can glue two metal plates to the disk, one on each side, to obtain a *disk-ceramic capacitor*. To get larger capacitance values, we can stack layers of metal and ceramic, connecting alternate layers of metal together as the electrodes. The smallest single-layer ceramic capacitors exhibit only a few picofarads of capacitance. Large MLCCs can have values ranging into the hundreds of microfarads.

Plastic-Film Capacitors

Plastics make good dielectrics for the manufacture of capacitors. *Polyethylene* and *polystyrene* are commonly used. The method of manufacture resembles that for paper capacitors. Stacking methods work with plastic. The geometries can vary, so we'll find *plastic-film capacitors* in various shapes. Capacitance values range from about 50 pF to several tens of microfarads. Most often, we'll encounter values from approximately 0.001 μF to 10 μF. Plastic capacitors function well in electronic circuits at all frequencies, and at low to moderate voltages. They exhibit good efficiency, although not as high as that of mica-dielectric or air-dielectric capacitors.

Aluminum Electrolytic Capacitors

All of the above-mentioned types of capacitors provide relatively small values of capacitance. They are also *nonpolarized*, meaning that we can connect them into a circuit in either direction (in some cases the vendor offers a recommendation as to which side should go to signal ground). An *electrolytic* capacitor provides approximately 1 μF up to thousands of microfarads, but we *must* connect it in the proper direction if we want it to work properly. An electrolytic capacitor constitutes a *polarized* component.

Component manufacturers assemble *electrolytic capacitors* by rolling up multiple layers of aluminum foil strips separated by paper saturated with an *electrolyte* liquid. The electrolyte conducts electric current. When DC flows through the component, the aluminum oxidizes because of chemical interaction with the electrolyte. The oxide layer does not conduct, and therefore, forms the dielectric for the capacitor. The layer is extremely thin, yielding high capacitance per unit of volume. Electrolytic capacitors can have values up to thousands of microfarads, and some can handle thousands of volts. These capacitors are most often seen in audio amplifiers and DC power supplies.

For large values of capacitance (>10 μF) electrolytics are lower cost than other technologies, but this does come at the cost of a short lifespan. Because they are often used as smoothing capacitors, where constant charging and discharging at high rates leads to heating and eventual breakdown of the capacitor. Electronics repairers will often check the electrolytics first if a piece of equipment stops working.

Tantalum Capacitors

Another type of electrolytic capacitor uses tantalum rather than aluminum. The tantalum can comprise foil like the aluminum in a conventional electrolytic capacitor. The tantalum can also take the form of a porous pellet, the irregular surface of which provides a large area in a small volume. An extremely thin oxide layer forms on the tantalum.

Tantalum capacitors have high reliability and excellent efficiency. We'll often see them used in military and aerospace environments—or anywhere where technicians find servicing inconvenient or impossible—because these devices almost never fail. Tantalum capacitors have values similar to those of aluminum electrolytics, and they work well in audio and digital circuits as replacements for aluminum types.

They do however cost more than aluminum electrolytics and if abused by overvoltage can fail in a somewhat explosive manner.

Transmission-Line Capacitors

In Chap. 10, we learned that sections of transmission line, cut to lengths shorter than ¼ electrical wavelength and shorted out at the far end, can function as inductors. Such line sections will act as capacitors if, instead of shorting the far end, we leave it open. The capacitance of such a transmission-line section increases with length until we get to ¼ electrical wavelength, when it behaves like a short circuit at the input end. Transmission-line capacitors are frequency-sensitive, just as transmission-line inductors are.

Semiconductor-Based Capacitors

Later in this book, you'll learn about *semiconductors*. These materials revolutionized electrical and electronic circuit design during the twentieth century. Today, nearly all electronic systems consists mainly of semiconductor-based components.

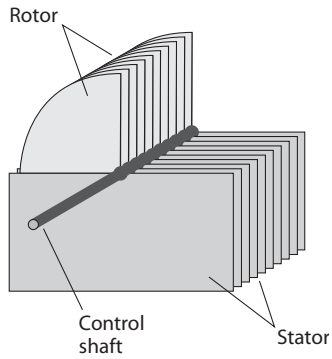
Manufacturers can employ semiconducting materials to build capacitors. A semiconductor *diode* conducts current in one direction, and refuses to conduct in the other direction. When a voltage appears across a diode so that it does not conduct, the diode acts as a capacitor. The capacitance varies depending on how much of this *reverse voltage* we apply to the diode. The greater the reverse voltage, the smaller the capacitance. This phenomenon makes the diode act as a *variable capacitor*. Some diodes are especially manufactured for this role. Their capacitances fluctuate rapidly along with pulsating DC. We call them *varactor diodes* or simply *varactors*.

Capacitors can be "etched" into the semiconductor materials of an *integrated circuit* (IC), also called a *chip*, as miniature varactors. A tiny capacitor can also be "etched" into an IC by sandwiching an oxide layer between two thin layers that conduct well. Most ICs look like little boxes with protruding metal prongs, which provide the electrical connections to external circuits and systems.

Semiconductor capacitors usually have small values of capacitance. They always have microscopic dimensions, and they can handle only low voltages. The advantages of semiconductor-based capacitors include miniaturization, and an ability (in the case of the varactor) to change in value at a rapid rate.

Variable Capacitors

We can vary the value of a capacitor at will by adjusting the mutual surface area between the plates, or by changing the spacing between the plates. Two main types of variable capacitors (besides varactors) exist: the *air variable capacitor* and the *trimmer capacitor*. We'll occasionally encounter a less common type known as a *coaxial capacitor*.



11-7 Simplified drawing of an air-variable capacitor.

Air-Variable Capacitors

We can assemble an air-variable capacitor by connecting two sets of metal plates so that they mesh, and by affixing one set to a rotatable shaft. The rotatable set of plates constitutes the *rotor*, and the fixed set constitutes the *stator*. We'll find this type of component in "vintage" radio receivers (particularly those that used *vacuum tubes* rather than semiconductor components), and in high-power wireless antenna tuning networks. Figure 11-7 illustrates an air-variable capacitor.

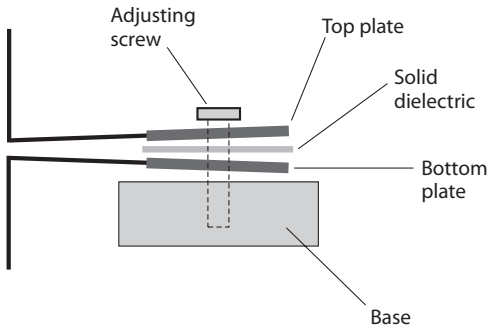
Air variables have maximum capacitance that depends on the number of plates in each set, and also on the spacing between the plates. Common maximum values range from 50 to 500 pF; occasionally we'll find an air variable that can go up to 1000 pF. Minimum values are generally on the order of a few picofarads. The voltage-handling capability depends on the spacing between the plates. Some air variables can handle several kilovolts at high AC frequencies. But the biggest advantage of air as a dielectric material is the fact that it has low loss—comparable to that of a vacuum.

We'll find air variables mainly in wireless equipment, designed to work at frequencies above approximately 500 kHz. These components offer high efficiency and excellent *thermal stability* (meaning that their values don't appreciably change with wild fluctuations in the ambient temperature). Although air variables technically lack polarization, we'll usually want to connect the rotor plates, along with the control shaft, to the metal chassis or circuit board perimeter, which constitutes the *common ground*.

Trimmer Capacitors

When we don't need to change the value of a capacitor often, we can use a *trimmer capacitor* in place of the more expensive, and bulkier, air-variable capacitor. A trimmer consists of two plates, mounted on a ceramic base and separated by a sheet of solid dielectric. We can vary the spacing between the plates with an adjusting screw as shown, in Fig. 11-8. Some trimmers contain two interleaved sets of multiple plates, alternating with dielectric layers to increase the capacitance.

We can connect a trimmer capacitor in parallel with an air-variable capacitor, facilitating exact adjustment of the latter component's minimum capacitance. Some air-variable capacitors have trimmers built in to serve this purpose. Typical maximum values for trimmers range from a few picofarads up to about 200 pF. They handle low to moderate voltages, exhibit excellent efficiency, and are free of polarizing characteristics.



11- 8 Cross-sectional drawing of a trimmer capacitor.

Coaxial Capacitors

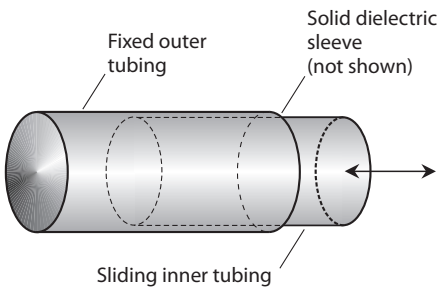
We can use two telescoping sections of metal tubing to build a so-called *coaxial capacitor* (Fig. 11-9). The device works because of the variable effective surface area between the inner and outer tubing sections. A sleeve of plastic dielectric separates the sections of tubing, allowing us to adjust the capacitance by sliding the inner section in or out of the outer section. Coaxial capacitors work well at high-frequency AC applications, particularly in wireless antenna tuners. Their values range from a few picofarads up to approximately 100 pF.

Capacitor Specifications

When you seek a capacitor for a particular application, you should look for a component that has the proper specifications for the job you have in mind. Following are two especially significant capacitor specifications.

Tolerance

Component manufacturers rate capacitors according to how nearly we can expect their values to match the quoted capacitance. We call this rating the *tolerance*. The most common tolerance for fixed capacitors is ±10 percent; some capacitors are rated at ±5 percent, or even ±1 percent. As the tolerance figure decreases, we can expect the actual component value to more closely match the quoted value. For example, capacitor rated at 100 pF ±10 percent can range from 90 pF to 110 pF. But if the tolerance equals ±1 percent, the manufacturer guarantees that the capacitance will not stray outside the range of 99 pF to 101 pF.



11-9 Simplified drawing of a coaxial variable capacitor.

Problem 11-6

Suppose that we find a capacitor rated at 0.10 μF ±10 percent. What's the guaranteed range of capacitance?

Solution

First, we multiply 0.10 by 10 percent to get the plus-or-minus variation. When we carry out that arithmetic, we get $0.10 \times 0.10 = 0.010$ μF. Then we add and subtract this value from the quoted capacitance to get the capacitance range, obtaining a minimum possible value of $0.10 - 0.010 = 0.09$ μF and a maximum possible value of $0.10 + 0.010 = 0.11$ μF.

Temperature Coefficient

Some capacitors increase in value as the temperature increases. These components have a *positive temperature coefficient*. Other capacitors decrease in value as the temperature rises; these exhibit a *negative temperature coefficient*. Some capacitors are specially manufactured (at considerable cost) so that their values remain constant over a certain temperature range. Within this span of temperatures, such capacitors have *zero temperature coefficient*.

Engineers commonly specify the temperature coefficient of a component in terms of *percent per degree Celsius* (%/°C). Sometimes, we can connect a capacitor with a negative temperature coefficient in series or parallel with a capacitor having a positive temperature coefficient, and the two opposite effects more or less nullify each other over a limited range of temperatures. In other instances, we can employ a capacitor with a positive or negative temperature coefficient to cancel out, or at least minimize, the effects of temperature on other components in a circuit, such as inductors and resistors.

Interelectrode Capacitance

Any two pieces of conducting material in close proximity can act as a capacitor. Often, such *interelectrode capacitance* is so small—a couple of picofarads or less—that we don't have to worry about it. In utility and audio-frequency (AF) circuits, interelectrode capacitance rarely poses any trouble, but it can cause problems in radio-frequency (RF) systems. The risk of trouble increases as the frequency increases.

The most common consequences of excessive interelectrode capacitance are *feedback* and unwanted changes in the characteristics of a circuit with variations in the operating frequency. We can minimize the interelectrode capacitance in an electronic device or system by keeping the interconnecting wires as short as possible within each individual circuit, by using shielded cables to connect circuits to each other, and by enclosing the most sensitive circuits in metal housings.

Equivalent Series Resistance

In an ideal world, capacitors would have only capacitance and no inductance or resistance. In practice, capacitors have an equivalent series resistance or ESR. This is value of a notional resistor in series with one of the capacitor leads and an electrode. The ESR will be one of the factors to take into consideration when selecting a capacitor for a particular job.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

12
CHAPTER

Phase

IN AN AC WAVE, EVERY FULL CYCLE REPLICATES EVERY OTHER FULL CYCLE; THE WAVE REPEATS INDEFINITELY. In this chapter, you'll learn about the simplest possible shape (or waveform) for an AC disturbance. We call it a *sine wave* or *sinusoid*.

Not More Math!

The next few chapters are going to introduce a fair amount of mathematics. It's not always obvious that this is how you use math:

1. Describe a real world things using a convenient mathematical notation (algebra, calculus, complex numbers, polar coordinates, etc.)
2. Manipulate that mathematical model in an entirely abstract way, using the rules of math in order to make predictions.
3. Interpret the manipulated mathematical model back into real world terms.

People often struggle with math because they fail to suspend reality during stage 2 and want to know what it "means." This is an unnecessary distraction as it might not mean anything.

To take a really simple example of Ohm's law, and a practical problem of what current flows through a 100 Ω resistor when there is a voltage of 5 V across it.

Stage 1. Express the problem in math: $V = IR$ so $5 = I \times 100$
Stage 2. Apply algebra and solving for the unknown I we get: $I = 5/100 = 50$ mA
Stage 3. Interpret: 50 mA flows through the resistor

All this, without actually having to put together a real voltage source, resistor, and ammeter and start measuring things. It's a theoretical result.

This use of math in this book is about to become more complicated, as the situation that we are going to model (AC) requires more powerful tools than simple algebra. In particular, we are going to use polar coordinates and complex numbers to help us model how components behave with AC.

Instantaneous Values

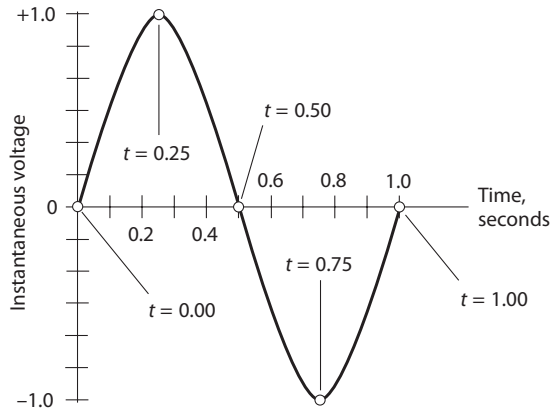
When we graph its instantaneous amplitude as a function of time, an AC sine wave has the characteristic shape shown in Fig. 12-1. This illustration shows how the graph of the function $y = \sin x$ looks on an (x,y) coordinate plane. (The abbreviation *sin* stands for *sine* in trigonometry.) Imagine that the peak voltages equal +1.0 V and −1.0 V. Further imagine that the period equals exactly one second (1.0 s) so the wave has a frequency of 1.0 Hz. Let's say that the wave begins at time $t = 0.0$ s. In this scenario, each cycle begins every time the value of t "lands on" a whole number. At every such instant, the voltage is zero and *positive-going*.

If you freeze time at, say, $t = 446.00$ s and then measure the instantaneous voltage, you'll find that it equals 0.0 V. Looking at the diagram, you can see that the instantaneous voltage will also equal 0.0 V every so-many-and-a-*half* seconds, so it will equal 0.0 V at, say, $t = 446.50$ s. But instead of getting more positive at the "second-and-a-half" instants, the voltage trends negative. If you freeze time at so-many-and-a-*quarter* seconds, say, $t = 446.25$ s, the instantaneous voltage will equal +1.0 V. The wave will rest exactly at its positive peak. If you stop time at so-many-and-*three-quarter* seconds, say, $t = 446.75$ s, the instantaneous voltage will rest exactly at its negative peak, −1.0 V. At intermediate time points, such as so-many-and-*three-tenths* seconds, the voltage will have intermediate values.

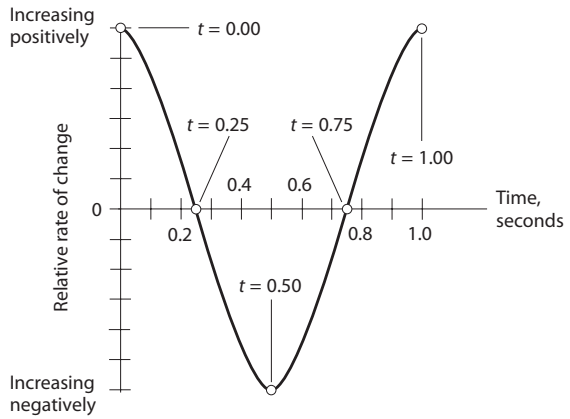
Rate of Change

Figure 12-1 reveals the fact that the instantaneous voltage sometimes increases and sometimes decreases. *Increasing*, in this context, means "getting more positive," and *decreasing* means "getting more negative." In the situation shown by Fig. 12-1, the most rapid increases in voltage occur when $t = 0.00$ s and $t = 1.00$ s. The most rapid voltage decrease takes place when $t = 0.50$ s. When $t = 0.25$ s, and also when $t = 0.75$ s, the instantaneous voltage neither increases nor decreases. But these "unchanging voltages" exist only for vanishingly small instants in time.

Let n equal some positive whole number of seconds. No matter what whole number we choose for n, the situation at $t = n.25$ s appears the same as it does for $t = 0.25$ s. Also, for $t = n.75$ s, things appear the same as they are when $t = 0.75$ s. The single cycle shown in Fig. 12-1 represents every possible condition of an AC sine wave having a frequency of 1.0 Hz and peak values of +1.0 V and −1.0 V. The entire wave cycle repeats for as long as AC continues to flow in the circuit, assuming that we don't change the voltage or the frequency.



12-1 A sine wave with a period of 1 second has a frequency of 1 Hz.



12-2 A sine wave representing the rate of change in the instantaneous voltage of the wave in Fig. 12-1.

Now imagine that you want to observe the *instantaneous rate of change* in the voltage of the wave in Fig. 12-1, as a function of time. A graph of this function turns out as a sine wave, too—but it appears displaced to the left of the original wave by ¼ of a cycle. If you plot the instantaneous rate of change of a sine wave as a function of time (Fig. 12-2), you get the *derivative* of the waveform. The derivative of a sine wave turns out as a *cosine wave* because the mathematical derivative (in calculus) of the sine function equals the cosine function. The cosine wave has the same shape as the sine wave, but the *phase* differs by ¼ of a cycle.

Circles and Vectors

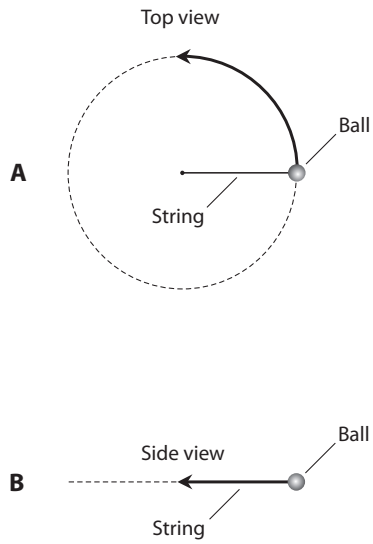
An AC sine wave represents the most efficient manner in which an electrical quantity can alternate. It has only one frequency component. All the wave energy concentrates into a single, smooth, swinging variation—a single frequency. When an AC wave has this characteristic, we can represent its fluctuations by comparing it to the motion of an object that follows a circular orbit at a constant speed around a fixed central point.

Circular Motion

Imagine that you revolve a ball around and around at the end of a string, at a rate of one revolution per second (1 r/s) so that the ball describes a horizontal circle in space, as shown in Fig. 12-3A. If a friend stands some distance away, with his or her eyes in the plane of the ball's path, she sees the ball oscillating back and forth, as shown in Fig. 12-3B, with a frequency of 1 Hz. That's one complete cycle per second because you cause the ball to "orbit" at 1 r/s.

If you graph the position of the ball, as seen by your friend, with respect to time, you'll get a sine wave, as shown in Fig. 12-4. This wave has the same fundamental shape as all sine waves. One sine wave might exhibit a greater distance between the peaks (peak-to-peak amplitude) than another, and one sine wave might appear "stretched out" lengthwise (wavelength) more than another. But every sine wave has the same general nature as every other. If we multiply or divide the peak-to-peak amplitude and/or the wavelength of any sine wave by the right numbers, we can make that sine wave fit exactly along the curve of any other sine wave. A *standard sine wave* has the following function in the (x,y) coordinate plane:

$$y = \sin x$$

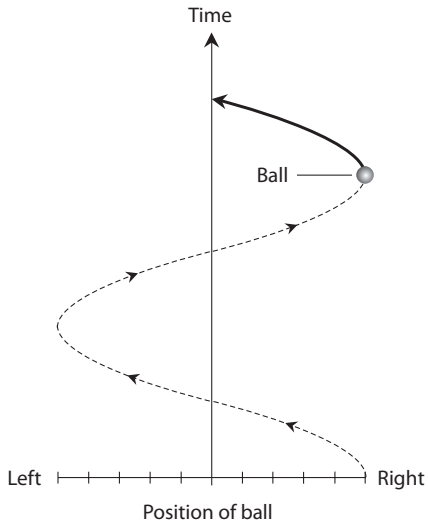


12-3 A revolving ball and string as seen from above (A) and from the side (B).

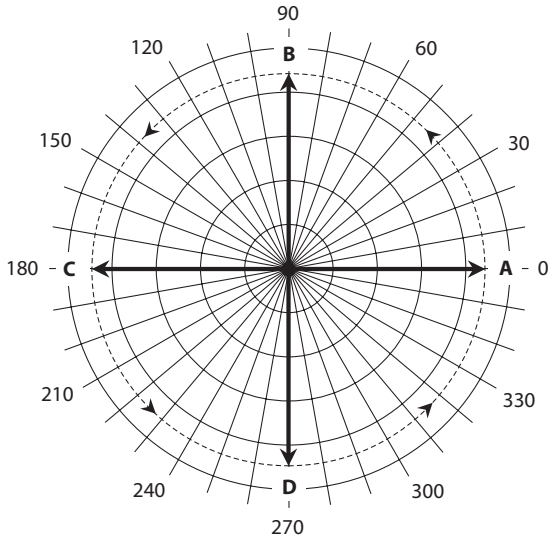
In the situation of Fig. 12-3A, you might make the string longer or shorter. You might whirl the ball around faster or slower. These changes would alter the peak-to-peak amplitude and/or the frequency of the sine wave graphed in Fig. 12-4. But you can always portray a sine wave as the equivalent of constant circular revolution. Mathematicians and engineers call this trick the *circular-motion model* of a sine wave.

The Rotating Vector

In Chap. 9, you learned about *degrees of phase*. If you wondered why we discussed phase in terms of angles going around a circle, you should have a better grasp of the idea now! A circle contains 360 angular degrees (360°), as you know from your courses in basic geometry. Points along a sine wave intuitively correspond to angles, or positions, around a circle.



12-4 Position of ball (horizontal axis) as seen from the side, graphed as a function of time (vertical axis).



12-5 Rotating-vector representation of a sine wave. Vector **A** portrays the start of the cycle (0°); vector **B** portrays the wave ¼ of the way through the cycle (90°); vector **C** portrays the wave halfway through the cycle (180°); vector **D** portrays the wave ¾ of the way through the cycle (270°). The vector length never changes.

Figure 12-5 shows how we can use a *rotating vector* to represent a sine wave in a system of *polar coordinates*. In the polar coordinate system, we plot a point according to its distance (called the *radius*) from the *origin* (center of the graph) and its angle expressed counterclockwise from "due east" (called the *direction*). Compare this system with the more traditional system of *rectangular coordinates*, where we plot a point according to its horizontal and vertical displacement from the origin.

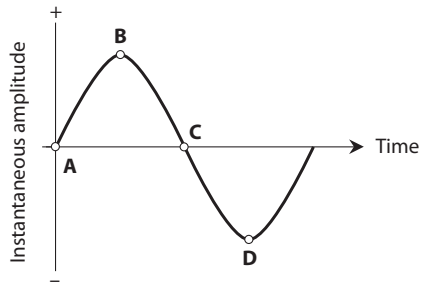
A *vector* constitutes a mathematical quantity with two independent properties, called *magnitude* (also called *length* or *amplitude*) and *direction* (or *angle*). Vectors lend themselves perfectly to polar coordinates. In the circular-motion model of Fig. 12-5, we can note the following specific situations:

- Vector **A** has a direction angle of 0°; it portrays the instant at which the wave amplitude equals zero and increases positively.
- Vector **B** points "north," representing the 90° phase angle, at which the wave has attained its maximum positive amplitude.
- Vector **C** points "west," representing a phase angle of 180°, the instant when the wave has gone back to zero amplitude while growing more negative.
- Vector **D** points "south," representing 270° of phase, the instant at which the wave has attained its maximum negative amplitude.
- When the vector has rotated counterclockwise through a full circle (360°), it once again becomes vector **A**, and the wave begins its next cycle.

If you're astute, you'll notice that while the vector's direction constantly changes, its length always remains the same.

Vector "Snapshots"

Figure 12-6 shows the four points, on a sine wave, representing the *instantaneous vectors* **A, B, C,** and **D** from Fig. 12-5. Think of these four points as "snapshots" of the wave vector as it rotates



12-6 The four points for the vector model of Fig. 12-5, shown in the standard amplitude-versus-time graphical manner for a sine wave.

counterclockwise at a constant *angular speed* that corresponds to one revolution per cycle of the wave. If the wave has a frequency of 1 Hz, the vector revolves at a rate of 1 r/s. We can increase or decrease the frequency and still use this model. If the wave has a frequency of 100 Hz, the speed of the vector will equal 100 r/s, or a revolution every 0.01 s. If the wave has a frequency of 1 MHz, then the speed of the vector will equal 1,000,000 r/s (10^6 r/s), and it will go once around the circle every 0.000001 s (10^{-6} s).

The peak amplitude (either positive or negative without the sign) of a pure AC sine wave corresponds to the length of its vector in the circular-motion model. Therefore, the peak-to-peak amplitude corresponds to twice the length of the vector. As the amplitude increases, the vector gets longer. In Fig. 12-5, we portray time during an individual cycle as an angle going counterclockwise from "due east."

In Fig. 12-5 and all other circular-model sine-wave vector diagrams, the vector length never changes, although it constantly rotates counterclockwise at a steady angular speed. The frequency of the wave corresponds to the speed at which the vector rotates. As the wave frequency increases, so does the vector's rotational speed. Just as the sine-wave's vector length remains independent of its rate of rotation, the amplitude of a sine wave is independent of its frequency.

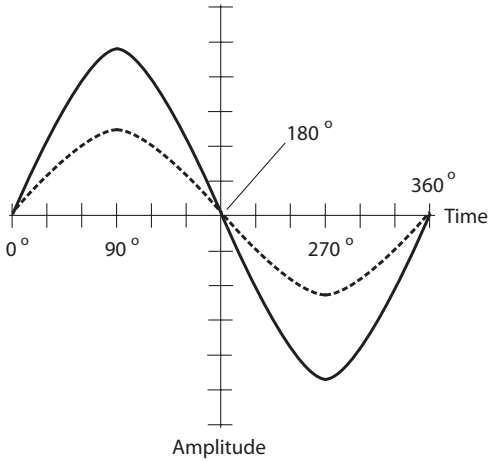
Expressions of Phase Difference

The *phase difference*, also called the *phase angle*, between two sine waves can have meaning only when those two waves have the same frequency. If the frequencies differ, even by the slightest amount, the relative phase constantly changes, and we can't specify a value for it. In the following discussions of phase angle, let's assume that the two waves always have identical frequencies.

Phase Coincidence

The term *phase coincidence* means that two waves begin at exactly the same moment. They "line up perfectly." Figure 12-7 illustrates a situation of this sort for two sine waves having different amplitudes. The phase difference in this case equals 0°. Alternatively, we could say that the phase difference equals some whole-number multiple of 360°, but engineers and technicians rarely speak of any phase angle of less than 0° or more than 360°.

If two sine waves exist in phase coincidence, and if neither wave has any DC superimposed on it, then the resultant wave constitutes a sine wave with positive peak (pk+) or negative peak (pk−) amplitudes equal to the sum of the positive and negative peak amplitudes of the composite waves. The phase of the resultant wave coincides with the phases of the two composite waves.



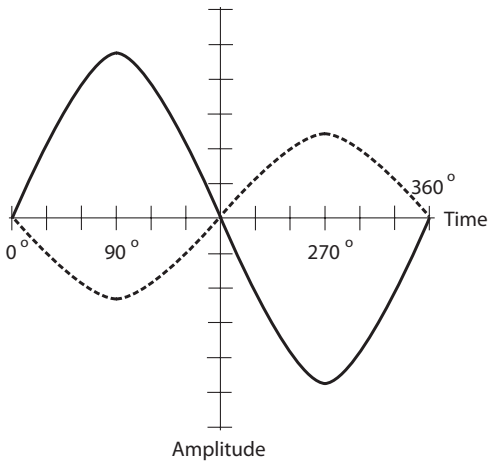
12-7 Two sine waves in phase coincidence.

Waves ½ Cycle out of Phase

When two sine waves begin exactly ½ cycle (180°) apart, we get a situation like the one shown in Fig. 12-8. In this case, engineers sometimes say that the waves are *out of phase*, although this expression constitutes an imprecise statement because someone might take it to mean a phase difference other than 180°.

If two sine waves have the same amplitudes and exist 180° out of phase, and if neither wave has DC superimposed, they cancel each other out because the instantaneous amplitudes of the two waves are equal and opposite at every moment in time.

If two sine waves have different amplitudes and exist 180° out of phase, and if neither wave has DC superimposed, then the resultant wave is a sine wave with positive or negative peak amplitudes equal to the difference between the positive and negative peak amplitudes of the composite waves. The phase of the resultant wave coincides with the phase of the stronger composite.



12-8 Two sine waves that differ in phase by 180°.

Phase Opposition

Any perfect sine wave without superimposed DC has the property that, if we shift its phase by precisely 180°, we get a result identical to what we get by "flipping the original wave upside-down" (inverting it), a condition called *phase opposition*. Not all waveforms have this property. Perfect square waves do, but most rectangular and sawtooth waves don't, and irregular waveforms almost never do.

In most *nonsinusoidal* waves (waves that do not follow the sine or cosine function's graph), a phase shift of 180° *does not* yield the same result as "flipping the wave upside down." Never forget the conceptual difference between a 180° phase shift and a state of phase opposition!

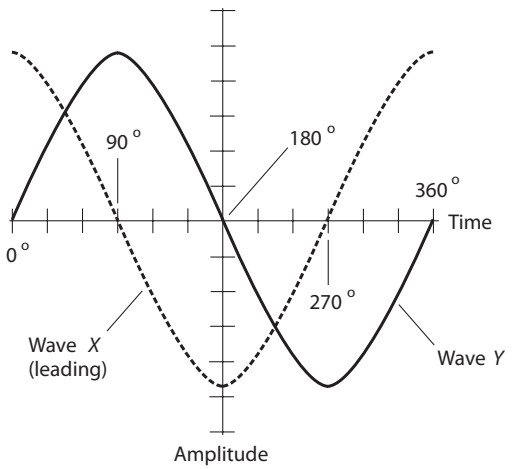
Intermediate Phase Differences

Two perfect sine waves having the same frequency can differ in phase by any amount from 0° (phase coincidence), through 90° (*phase quadrature*, meaning a difference of a quarter of a cycle), through 180°, through 270° (phase quadrature again), and finally back to 360° (phase coincidence again).

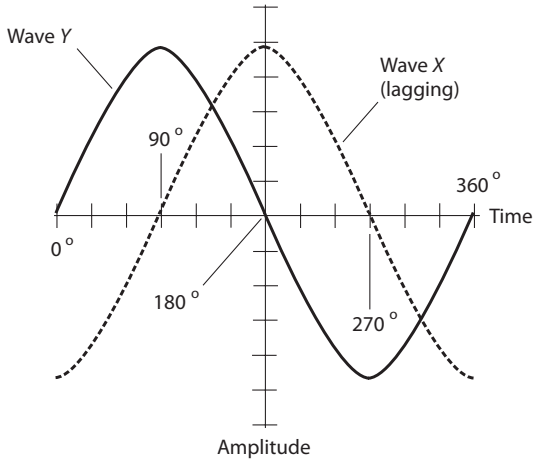
Leading Phase

Imagine two sine waves, called wave *X* and wave *Y*, with identical frequency. If wave *X* begins a fraction of a cycle *earlier* than wave *Y*, then we say that wave *X leads* wave *Y* in phase. For this situation to hold true, wave *X* must begin its cycle less than 180° before wave *Y*. Figure 12-9 shows wave *X* leading wave *Y* by 90°.

When a particular wave *X* (the dashed line in Fig. 12-9) leads another wave *Y* (the solid line), then wave *X* lies to the *left* of wave *Y* on the time axis by some distance less than ½ wavelength. In a time-domain graph or display, displacement to the left represents earlier moments in time, and displacement to the right represents later moments in time; time "flows" from left to right.



12-9 Wave *X* leads wave *Y* by 90° of phase (¼ of a cycle).



12-10 Wave *X* lags wave *Y* by 90° of phase (¼ of a cycle).

Lagging Phase

Now imagine that some sine wave *X* begins its cycle more than 180° (½ cycle) but less than 360°(a full cycle) before wave *Y* starts. In this situation, we can imagine that wave *X* starts its cycle *later* than wave *Y* by some value between 0° and 180°. Then we say that wave *X lags* wave *Y*. Figure 12-10 shows wave *X* lagging wave *Y* by 90°. When a particular wave *X* (the dashed line in Fig. 12-10) lags another wave *Y* (the solid line), then wave *X* lies to the *right* of wave *Y* on the time axis by some distance less than ½ wavelength.

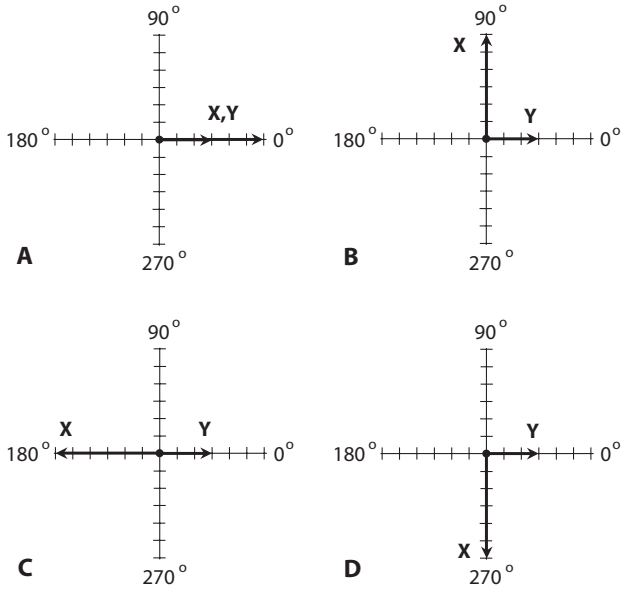
Vector Diagrams of Relative Phase

Suppose that a sine wave *X* leads a sine wave *Y* by, say, *q* degrees (where *q* represents a positive angle less than 180°). In this situation, we can draw the two waves as vectors, with vector **X** oriented *q* degrees *counterclockwise* from vector **Y**. In the opposite sense, if a sine wave *X* lags a sine wave *Y* by *q* degrees, then vector **X** appears to point in a direction going *clockwise* from vector **Y** by *q* degrees. If two waves *X* and *Y* coincide in phase, then their vectors **X** and **Y** point in the same direction. If two waves *X* and *Y* occur 180° out of phase, then their vectors **X** and **Y** point in opposite directions.

Figure 12-11 shows four phase relationships between two sine waves *X* and *Y* that have the same frequency but different amplitudes, as follows:

1. At A, wave *X* exists in phase with wave *Y*, so vectors **X** and **Y** line up.
2. At B, wave *X* leads wave *Y* by 90°, so vector **X** points in a direction 90° counterclockwise from vector **Y**.
3. At C, waves *X* and *Y* exist 180° apart in phase, so vectors **X** and **Y** point in opposite directions.
4. At D, wave *X* lags wave *Y* by 90°, so vector **X** points in a direction 90° clockwise from vector **Y**.

In all of these examples, we should imagine that the vectors both rotate *counterclockwise* at a continuous, steady rate as time passes, always maintaining the same angle *with respect to each other*,



12-11 Vector representations of phase difference. At A, waves *X* and *Y* are in phase. At B, wave *X* leads wave *Y* by 90°. At C, wave *X* and wave *Y* are 180° out of phase. At D, wave *X* lags wave *Y* by 90°. We represent time as counterclockwise rotation of both vectors **X** and **Y** at a constant angular speed.

and always staying at the same lengths. If the frequency in hertz equals *f*, then the pair of vectors rotates together, counterclockwise, at an angular speed of *f*, expressed in complete, full-circle rotations per second (r/s).

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

13
CHAPTER

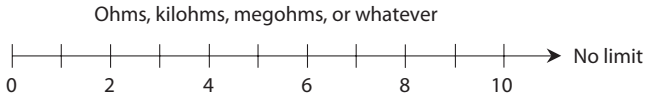
Inductive Reactance

IN DC ELECTRICAL CIRCUITS, THE CURRENT, VOLTAGE, RESISTANCE, AND POWER RELATE ACCORDING to simple equations. The same equations work for AC circuits, provided that the components merely dissipate energy, and never store or release it. If a component stores or releases energy in an AC system, we say that the component has *reactance*. When we mathematically combine a component's reactance and resistance, we get an expression of the component's *impedance*, which fully quantifies how that component opposes, or *impedes*, the flow of AC.

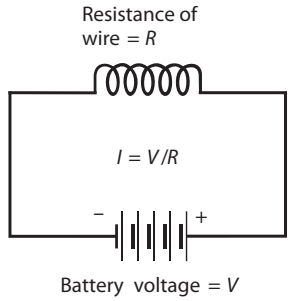
Inductors and Direct Current

We can express DC resistance (in ohms, kilohms, megohms, or whatever other unit we want) as a number ranging from 0 (representing a perfect conductor) to extremely large values (representing poor conductors). Scientists call resistance a *scalar* quantity because we can portray its values as points on a *half-line,* or *ray,* having a one-dimensional *scale,* as shown in Fig. 13-1. However, when we add inductance to a circuit that already contains resistance and then drive AC through that circuit, things get more complicated.

Imagine that you have a supply wire that conducts electricity well. If you wind a length of the wire into a coil to make an inductor, and then you connect the coil to a battery or other source of DC (Fig. 13-2), the wire draws a small amount of current at first. But the current quickly becomes large, no matter how you configure the wire. You might wind it into a single-turn loop, let it lie in a mess on the floor, or wrap it around a wooden stick. In any case, you'll get a current equal to $I = V/R$, where I represents the current (in amperes), V represents the DC source voltage (in volts), and R represents the DC resistance of the wire (in ohms).



13-1 We can represent resistance values along a half-line or ray.



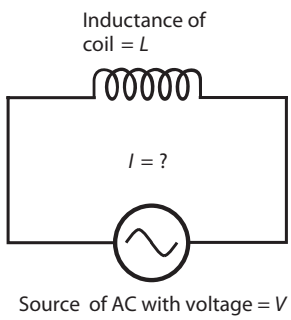
13-2 An inductor connected across a source of DC.

You can make an *electromagnet* by passing DC through a coil wound around an iron rod. You'll still observe a large, constant current in the coil, just as you would if the coil had no iron core. In a practical electromagnet, the coil heats up as some of the electrical energy dissipates in the wire; not all of the electrical energy contributes to the magnetic field. If you increase the DC source voltage and also increase the ability of the source to produce large currents, the wire in the coil will heat up more. Ultimately, if you increase the source voltage enough, and if it can deliver unlimited current, the wire will heat to the melting point.

Inductors and Alternating Current

Now suppose that you change the voltage source across the coil from DC to *pure AC* (that is, AC having no DC component), as shown in Fig. 13-3. Imagine that you can vary the frequency of the AC from a few hertz to hundreds of hertz, then kilohertz, then megahertz. At low frequencies, you'll see a large current in the coil, just as you did with the source of DC. But the coil exhibits a certain amount of inductance, and it takes a little time for current to establish itself in the coil. Depending on how many turns the coil has, and on whether the core consists of air or a ferromagnetic material, you'll reach a point, as you steadily increase the frequency, when the coil starts to get "sluggish." The current won't have time to fully establish itself in the coil before the AC polarity reverses.

At sufficiently high AC frequencies, the current through a coil will have trouble following the changes in the instantaneous voltage across it. Just as the coil starts to "think" that it can fully conduct, the AC voltage wave will pass its peak, go back to zero, and then "try" to pull the current the other way. In effect, this "sluggishness" will cause the coil to oppose the current in much the same way as a plain resistor would. As you raise the AC frequency, the coil's opposition to current will increase. Eventually, if you keep on increasing the frequency, the coil will fail to acquire



13-3 An inductor connected across a source of AC.

a significant current flow before the voltage polarity reverses. The coil will then act like a resistor with a high ohmic value.

With respect to AC, an inductor functions as a frequency-dependent resistor. We use the term *inductive reactance* to describe the opposition that the coil offers to AC. We express, or measure, inductive reactance in ohms. Inductive reactance can vary as resistance does, from almost nothing (a short piece of wire) to a few ohms (a small coil) to kilohms or megohms (coils having many turns, or coils with ferromagnetic cores operating at high AC frequencies). We can portray inductive reactance values along a half-line, just as we do with resistance. The numerical values on the half-line start at zero and increase without limit.

Reactance and Frequency

Inductive reactance constitutes one of two forms of reactance. (We'll examine the other form in the next chapter.) In mathematical expressions, we symbolize reactance in general as X, and we symbolize inductive reactance as X_L.

If the frequency of an AC source equals f (in hertz) and the inductance of a coil equals L (in henrys), then we can calculate the inductive reactance X_L (in ohms) using the formula

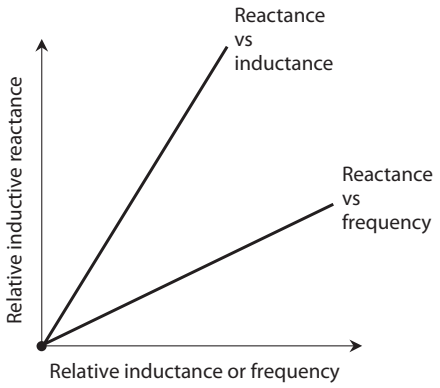
$$X_L = 2\pi fL \approx 6.2832\, fL$$

This same formula applies if we specify the frequency f in kilohertz and the inductance L in millihenrys. It also applies if we express f in megahertz and L in microhenrys. If we quantify frequency in thousands, we must quantify inductance in thousandths; if we quantify frequency in millions, we must quantify inductance in millionths.

Inductive reactance increases in a linear manner with (that is, in direct proportion to) increasing AC frequency, so the function of X_L versus f shows up as a straight line when we plot its graph on a rectangular coordinate plane. Inductive reactance also increases linearly with inductance, so the function of X_L versus L also appears as a straight line on a rectangular graph. Summarizing:

- If we hold L constant, then X_L varies in direct proportion to f.
- If we hold f constant, then X_L varies in direct proportion to L.

Figure 13-4 illustrates these relations on a generic rectangular coordinate grid.



13-4 Inductive reactance varies in direct proportion to the inductance at a fixed frequency, and in direct proportion to the frequency for a fixed value of inductance.

Problem 13-1

Suppose that a coil has an inductance of 0.400 H, and the frequency of the AC passing through it equals 60.0 Hz. What's the inductive reactance?

Solution

Using the above formula, we can calculate and round off to three significant figures, getting

$$X_L = 6.2832 \times 60.0 \times 0.400 = 151\ \Omega$$

Problem 13-2

How much inductive reactance will the above-described coil have if the power supply comprises a battery that supplies pure DC?

Solution

Because DC has a frequency of zero, we'll observe no inductive reactance at all. We can verify this by calculating

$$X_L = 6.2832 \times 0 \times 0.400 = 0\ \Omega$$

Inductance has no practical effect with pure DC. The coil will exhibit a little bit of DC resistance because no wire constitutes a perfect electrical conductor, but that's not the same thing as AC reactance!

Problem 13-3

If a coil has an inductive reactance of 100 Ω at a frequency of 5.00 MHz, what's its inductance?

Solution

In this case, we must plug numbers into the formula and solve for the unknown L. Let's start out with the equation

$$100 = 6.2832 \times 5.00 \times L = 31.416\ L$$

Because we know the frequency in megahertz, the inductance will come out in microhenrys (μH). We can divide both sides of the above equation by 31.416 and then round off to three significant figures, getting

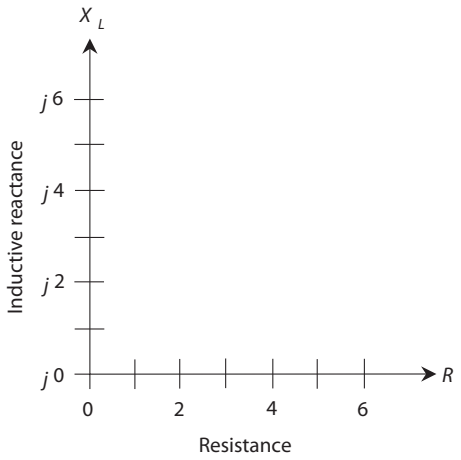
$$L = 100 / 31.416 = 3.18\ \mu\text{H}$$

The RX_L Quarter-Plane

In a circuit containing both resistance and inductance, we can't use a straight-line scale to portray the circuit's behavior with AC that varies in frequency. We must orient separate resistance and reactance rays perpendicular to each other to make a coordinate system, as shown in Fig. 13-5. Resistance appears on the horizontal axis, increasing as we move to the right. Inductive reactance appears on the vertical axis, increasing as we go upward. We call this grid the *resistance-inductive-reactance* (RX_L) *quarter-plane*.

What Are Those Little *j*'s For?

You're bound to wonder what the lowercase italic letters j represent in front of all the reactance numbers in Fig. 13-5. Engineers use the symbol j to represent a mathematical quantity called the



13-5 The RX_L quarter-plane for inductive reactance (X_L) and resistance (R).

unit imaginary number. It's the positive square root of −1. (If you didn't already know that negative numbers can have square roots, you do now!) Electrical engineers call the positive square root of −1 the *j operator*. When we multiply *j* by itself over and over, we get the following four-way repeating sequence of quantities:

$$j \times j = -1$$
$$j \times j \times j = -j$$
$$j \times j \times j \times j = 1$$
$$j \times j \times j \times j \times j = j$$
$$j \times j \times j \times j \times j \times j = -1$$
$$j \times j \times j \times j \times j \times j \times j = -j$$
$$j \times j \times j \times j \times j \times j \times j \times j = 1$$
$$j \times j \times j \times j \times j \times j \times j \times j \times j = j$$
$$j \times j \times j \times j \times j \times j \times j \times j \times j \times j = -1$$
$$\downarrow$$
and so on, forever

When we multiply *j* by an ordinary number (that is, a *real number*), such as 2 or ½ or 7.764958, we get an *imaginary number*. All of the points on the vertical scale in Fig. 13-5 represent imaginary numbers. When we add an imaginary number to a real number, we get a *complex number*. All of the points in the entire quarter-plane of Fig. 13-5 represent complex numbers.

By the way, mathematicians use the letter *i* instead of *j*. But *i* would much too easily confused with *I* for current, and so in electronics we use *j*.

Complex Impedance

Each point on the RX_L quarter-plane corresponds to a unique *complex-number impedance* (or simply *complex impedance*). Conversely, each complex impedance value corresponds to a unique point on the quarter-plane. We express a complete impedance value Z, containing resistance and inductive reactance, on the RX_L quarter-plane in the form

$$Z = R + jX_L$$

where R represents the resistance (in ohms) and X_L represents the inductive reactance (also in ohms).

Some RX_L Examples

Suppose that we have a pure resistance, say $R = 5\ \Omega$. In this case, the complex impedance equals $Z = 5 + j0$. We can plot it at the point $(5, j0)$ on the RX_L quarter-plane. If we have a pure inductive reactance, such as $X_L = 3\ \Omega$, then the complex impedance equals $Z = 0 + j3$, and its point belongs at $(0, j3)$ on the RX_L quarter-plane. Engineers sometimes incorporate both resistance and inductive reactance into electronic circuit designs. Then we encounter complex impedance values, such as $Z = 2 + j3$ or $Z = 4 + j1.5$. Figure 13-6 shows graphical representations of the four complex impedances mentioned in this paragraph.

Approaching the RX_L Extremes

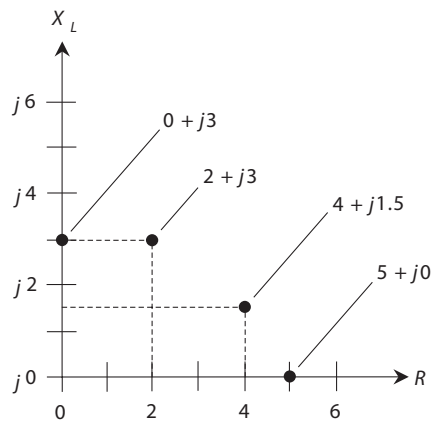
All practical coils have some resistance because no real-world wire conducts current perfectly. All resistors have a tiny bit of inductive reactance; all electrical components have wires called *leads* at each end, and any length of wire (even a straight one) exhibits some inductance. Therefore, in an AC circuit, we'll never encounter a mathematically perfect pure resistance, such as $5 + j0$, or a mathematically perfect pure reactance, such as $0 + j3$. We can approach these ideals, but we can never actually attain them (except in quiz and test problems).

How RX_L Points Move

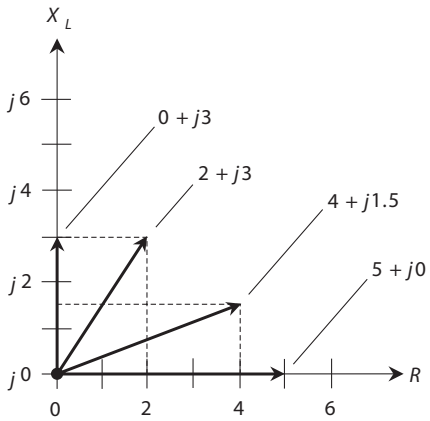
Always remember that the values for X_L represent *reactances* (expressed in ohms), and not *inductances* (expressed in henrys). In an RX_L circuit, the reactance varies with the AC frequency, even if the inductance value never changes at all. Changing the frequency produces the graphical effect of making the points move in the RX_L quarter-plane. The points go vertically upward as the AC frequency increases, and downward as the AC frequency decreases. If the AC frequency goes all the way down to zero, thereby resulting in DC, the inductive reactance vanishes, and we're left with only a little bit of resistance, representing the DC ohmic loss in the inductor.

Some RX_L Impedance Vectors

Engineers sometimes represent points in the RX_L quarter-plane as vectors. Expressing a point in the RX_L quarter-plane as a vector gives that point a unique magnitude and a unique direction. Figure 13-6 shows four different points, each one represented by a certain distance to the right of the *origin*



13-6 Four points in the RX_L quarter-plane.



13-7 Four vectors in the RX_L quarter-plane, corresponding to the points shown in Fig. 13-6.

point $(0, j0)$ that corresponds to the complex number $0 + j0$, and a certain distance upward from the origin. The first number in each complex sum represents the resistance R, and the second number represents the inductive reactance X_L. The RX_L combination constitutes a two-dimensional quantity. We can't define RX_L combinations as single numbers (scalar quantities) because any given RX_L combination possesses two quantities that can vary independently.

You can depict points, such as those shown in Fig. 13-6, by drawing straight rays from the origin out to those points. Then you can think of the rays instead of the points, with each ray having a certain length, or magnitude, and a certain direction, or angle counterclockwise from the resistance axis. These rays constitute *complex impedance vectors* (Fig. 13-7). When you think of complex impedances as vectors instead of mere points, you take advantage of a mathematical tool that can help you evaluate how AC circuits work under various conditions.

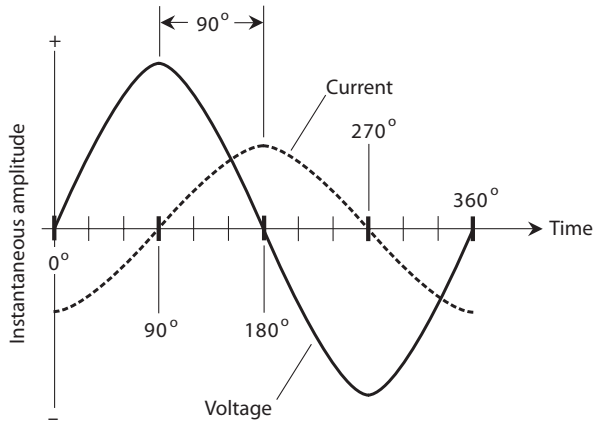
Current Lags Voltage

When we place an AC voltage source across an inductor and then power up the source so that the instantaneous voltage starts to increase (either positive or negative) from zero, it takes a fraction of a cycle for the current to follow. Later in the AC wave cycle, when the voltage starts decreasing from its maximum peak (either positive or negative), it again takes a fraction of a cycle for the current to follow. The instantaneous current can't quite keep up with the instantaneous voltage, as it does in a pure resistance, so we observe that in a circuit containing inductive reactance, the current *lags* (follows behind) the voltage. In some situations, this lag constitutes only a tiny fraction of an AC cycle, but it can range all the way up to ¼ of a cycle (90° of phase).

Pure Inductive Reactance

Suppose that we place an AC voltage source across a coil of wire made from an excellent conductor such as copper. Then we adjust the frequency of the AC source to a value high enough so that the inductive reactance X_L greatly exceeds the resistance R (by a factor of millions, say). In this situation, the coil acts as an essentially pure inductive reactance, and the current lags ¼ of a cycle (90°) behind the voltage for all intents and purposes, as shown in Fig. 13-8.

At low AC frequencies, we need a gigantic inductance if we want the current lag to approach 90°. As the AC frequency increases, we can get away with smaller inductances. If we could find some



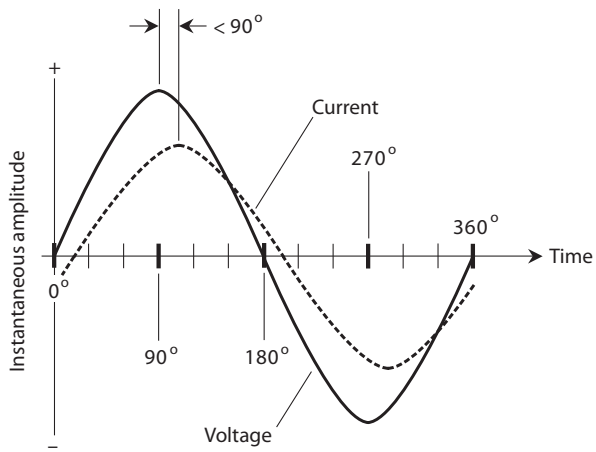
13-8 In a pure inductive reactance, the current lags the voltage by 90°.

wire that had no resistance whatsoever, and if we wound a coil with this wire, then the current would lag the voltage by exactly 90°, regardless of the AC frequency, and regardless of the coil size. In that case, we'd have an *ideal inductor* or a *pure inductive reactance*. No such thing exists in the "real world," but when the value of X_L greatly exceeds the value of R, the vector in the RX_L quarter-plane points almost exactly straight up along the X_L axis. The vector subtends an angle of just about 90° from the R axis.

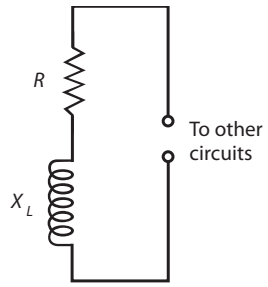
Inductive Reactance with Resistance

When the resistance in a resistance-inductance (RL) circuit is significant compared with the inductive reactance, the current lags the voltage by something less than 90°, as shown in the example of Fig. 13-9. If R is small compared with X_L, the current lag equals almost 90°, but as R gets larger relative to X_L, the lag decreases.

The value of R in an RL circuit can increase relative to X_L if we deliberately place a pure resistance in series with the inductance. It can also happen because the AC frequency gets so low that X_L



13-9 In a circuit with inductive reactance and resistance, the current lags the voltage by less than 90°.



13-10 Schematic representation of a circuit containing resistance and inductive reactance.

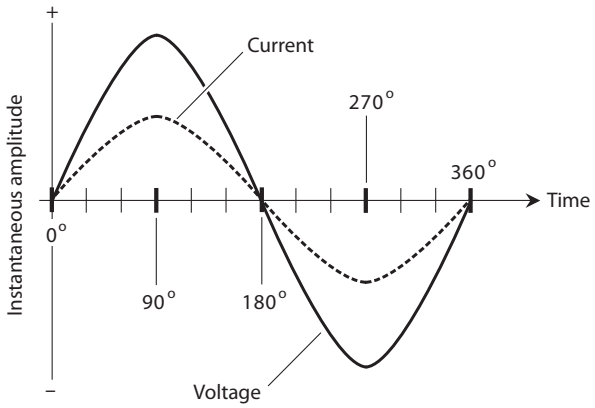
decreases until it reaches values comparable to the loss resistance R in the coil winding. In either case, we can schematically represent the circuit as an inductor in series with a resistor (Fig. 13-10).

If we know the values of X_L and R, we can find the *angle of lag*, also called the *RL phase angle* (or simply the *phase angle* if we know that we're dealing with resistance and inductance), by plotting the point $R + jX_L$ on the RX_L quarter-plane, drawing the vector from the origin out to that point, and then measuring the angle of the vector, counterclockwise from the resistance axis. We can use a protractor to measure this angle, or we can compute its value using trigonometry.

Actually, we don't have to know the actual values of X_L and R in order to find the angle of lag. All we need to know is their ratio. For example, if $X_L = 5\ \Omega$ and $R = 3\ \Omega$, we get the same phase angle as we do if $X_L = 50\ \Omega$ and $R = 30\ \Omega$, or if $X_L = 200\ \Omega$ and $R = 120\ \Omega$. The angle of lag turns out the same for any values of X_L and R in the ratio 5:3.

Pure Resistance

As the resistance in an RL circuit becomes large with respect to the inductive reactance, the angle of lag gets small. The same thing happens if the inductive reactance gets small compared with the resistance. When R exceeds X_L by a large factor, the vector in the RX_L quarter-plane lies almost on the R axis, going "east," or to the right. The phase angle in this case is close to 0°. The current flows nearly in phase with the voltage fluctuations. In a pure resistance with no inductance whatsoever, the current would follow along exactly in phase with the voltage, as shown in Fig. 13-11. A pure resistance doesn't store and release energy as an inductive circuit does. It acquires and relinquishes all of its energy immediately, so no current lag occurs.



13-11 In a circuit with pure resistance (no reactance), the current tracks right along in phase with the voltage.

How Much Lag?

If you know the ratio of the inductive reactance to the resistance (X_L/R) in an *RL* circuit, then you can find the phase angle. Of course, you can also find the phase angle if you know the actual values of X_L and R.

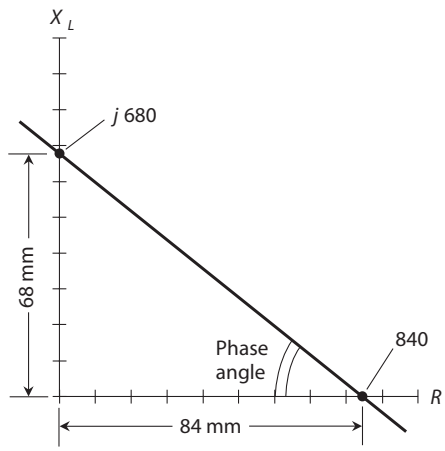
Pictorial Method

You can draw an RX_L quarter-plane on a piece of paper and then use a ruler and a protractor to find a phase angle in most *RL* situations. First, using the ruler and a sharp pencil, draw a straight line a little more than 100-mm long, going from left to right. Then with the protractor, construct a line off the left end of this first line, going vertically upwards. Make this line at least 100-mm long. The horizontal line, or the one going to the right, constitutes the R axis of a coordinate system. The vertical line, or the one going upwards, forms the X_L axis.

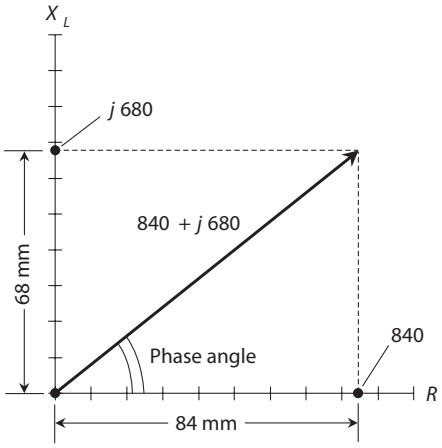
If you know the values of X_L and R, divide them down or multiply them up so they're both between 0 and 100. For example, if $X_L = 680\ \Omega$ and $R = 840\ \Omega$, you can divide them both by 10 to get $X_L = 68$ and $R = 84$. Plot these points lightly by making hash marks on the vertical and horizontal lines you've drawn. The R mark in this example should lie 84 mm to the right of the origin, and the X_L mark should lie 68 mm up from the origin.

Next, draw a line connecting the two hash marks, as shown in Fig. 13-12. This line will run at a slant, and will form a triangle along with the two axes. Your hash marks, and the origin of the coordinate system, form the three *vertices* (corner points) of a *right triangle*. We call the triangle "right" because one of its angles constitutes a *right angle* (90°). Measure the angle between the slanted line and the R axis. Extend one or both of the lines if necessary in order to get a good reading on the protractor. This angle will fall somewhere between 0° and 90°. It represents the phase angle in the *RL* circuit.

You can find the complex impedance vector, $R + jX_L$, by constructing a rectangle using the origin and your two hash marks as three of the rectangle's four vertices, and drawing new horizontal and vertical lines to complete the figure. The vector will show up as the diagonal of this rectangle (Fig. 13-13). The angle between this vector and the R axis will represent the phase angle. It should have the same measure as the angle of the slanted line relative to the R axis in Fig. 13-12.



13-12 Pictorial method of finding phase angle in a circuit containing resistance and inductive reactance.



13-13 Another pictorial method of finding phase angle in a circuit containing resistance and inductive reactance. This method shows the impedance vector.

Trigonometric Method

If you have a scientific calculator that can find the *Arctangent* of a number (also called the *inverse tangent* and symbolized either as Arctan or $\tan^{-1}$), you can determine the phase angle more precisely than the pictorial method allows. Given the values of X_L and R, the phase angle equals the Arctangent of their ratio. We symbolize phase angle as a variable with the lowercase Greek letter phi (pronounced "fie" or "fee" and written ϕ). Expressed mathematically, the phase angle is

$$\phi = \tan^{-1} (X_L/R)$$

or

$$\phi = \text{Arctan } (X_L/R)$$

If you use a calculator app on a computer, you can find a number's Arctangent by setting the calculator program to work in the scientific mode, entering the number, hitting the key or checking the box marked "inv," and finally hitting the "tan" key.

Problem 13-4

Suppose that the inductive reactance in an *RL* circuit equals 680 Ω and the resistance equals 840 Ω. What's the phase angle?

Solution

The ratio X_L/R equals 680/840. A calculator will display this quotient as something like 0.8095 followed by some more digits. Find the Arctangent of this number. You should get 38.99 and some more digits. You can round this off to 39.0°.

Problem 13-5

Suppose that an *RL* circuit operates at a frequency of 1.0 MHz with a resistance of 10 Ω and an inductance of 90 µH. What's the phase angle? What does this result tell you about the nature of this *RL* circuit at this frequency?

Solution

First, find the inductive reactance using the formula

$$X_L = 6.2832\,fL = 6.2832 \times 1.0 \times 90 = 565\ \Omega$$

Then find the ratio

$$X_L/R = 565/10 = 56.5$$

The phase angle equals Arctan 56.5, which, rounded to two significant figures, comes out to be 89°. Now you know that this *RL* circuit contains an almost pure inductive reactance because the phase angle is close to 90°. Therefore, you know that the resistance contributes little to the behavior of this *RL* circuit at 1.0 MHz.

Problem 13-6

What's the phase angle for the above circuit at a frequency of 10 kHz? With that information, what can you say about the behavior of the circuit at 10 kHz?

Solution

You must calculate X_L all over again for the new frequency. You can use megahertz as your unit of frequency because megahertz work in the formula with microhenrys. A frequency of 10 kHz equals 0.010 MHz. Calculating, you get

$$X_L = 6.2832\,fL = 6.2832 \times 0.010 \times 90 = 5.65\ \Omega$$

Calculating the ratio of inductive reactance to resistance, you get

$$X_L/R = 5.65/10 = 0.565$$

The phase angle at the new frequency equals Arctan 0.565, which, rounded to two significant figures, turns out as 29°. This angle is not close to either 0° or 90°. Therefore, you know that at 10 kHz, the resistance and the inductive reactance both play significant roles in the behavior of the *RL* circuit.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

14
CHAPTER

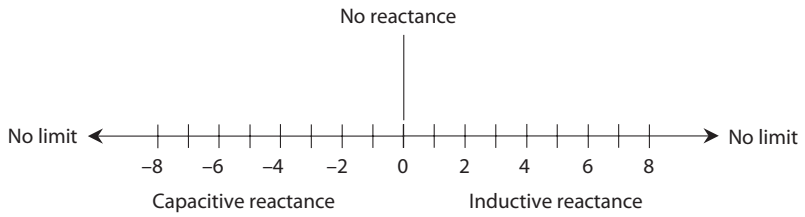
Capacitive Reactance

CAPACITIVE REACTANCE ACTS AS THE NATURAL COUNTERPART OF INDUCTIVE REACTANCE. WE CAN represent it graphically as a ray that goes in a negative direction. When we join the capacitive-reactance and inductive-reactance rays at their end points (both of which correspond to a reactance of zero), we get a complete number line, as shown in Fig. 14-1. This line depicts all possible values of reactance because any nonzero reactance must be either inductive or capacitive.

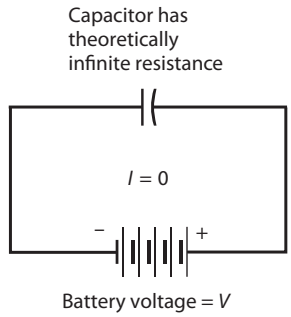
Capacitors and Direct Current

Imagine two huge, flat metal plates, both of which constitute excellent electrical conductors. If we connect them to a source of DC, as shown in Fig. 14-2, they draw a large amount of current while they become electrically charged. But as the plates reach equilibrium, the charging current goes down to zero.

If we increase the voltage of the battery or power supply, we eventually reach a point at which sparks jump between the plates of our capacitor. Ultimately, if the power supply can deliver the necessary voltage, this sparking, or *arcing*, becomes continuous. Under these conditions, the pair of plates no longer acts like a capacitor at all. When we place excessive voltage across a capacitor, the dielectric can't provide electrical separation between the plates. We call this undesirable condition *dielectric breakdown*.



14-1 We can represent inductive and capacitive reactance values as points along a number line.



14-2 A capacitor connected across a source of DC.

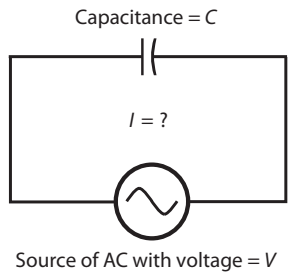
In an air-dielectric or vacuum-dielectric capacitor, dielectric breakdown manifests itself as a temporary affair, rarely causing permanent damage to the component. The device operates normally after we reduce the voltage so that the arcing stops. However, in capacitors made with solid dielectric materials, such as mica, paper, polystyrene, or tantalum, dielectric breakdown can burn or crack the dielectric, causing the component to conduct current even after we reduce the voltage. If a capacitor suffers this sort of damage, we must remove it from the circuit, discard it, and replace it with a new one.

Capacitors and Alternating Current

Now suppose that we change the voltage source from DC to AC (Fig. 14-3). Imagine that we can adjust the frequency of this AC from a low initial value of a few hertz, up to hundreds of hertz, then to many kilohertz, megahertz, or gigahertz.

At first, the voltage between the plates follows the voltage of the power source as the AC polarity alternates. The plates can charge up quickly if they have small surface areas and/or if a lot of space exists between them, but they can't charge instantaneously. As we increase the frequency of the applied AC, we reach a point at which the plates can't charge up very much before the AC polarity reverses. Just as the plates begin to get a good charge, the AC passes its peak and starts to discharge them. As we raise the frequency still further, the set of plates acts increasingly like a short circuit. Eventually, if we keep raising the AC frequency, the period of the wave becomes much shorter than the charge/discharge time, and current flows in and out of the plates just as fast as it would if the plates were removed altogether and replaced with a plain piece of wire.

Capacitive reactance quantifies the opposition that a capacitor offers to AC. We express and measure capacitive reactance in ohms, just as we do with inductive reactance or pure resistance.



14-3 A capacitor connected across a source of AC.

But capacitive reactance, by convention, has negative values rather than positive ones. Capacitive reactance, denoted X_C in mathematical formulas, can vary from near zero (when the plates are gigantic and close together, and/or the frequency is very high) to a few negative ohms, to many negative kilohms or megohms.

Capacitive reactance varies with frequency. It gets *larger negatively* as the AC frequency goes down, and *smaller negatively* as the applied AC frequency goes up. This behavior runs contrary to what happens with inductive reactance, which gets *larger positively* as the frequency goes up. Sometimes, nontechnical people talk about capacitive reactance in terms of its *absolute value*, with the minus sign removed. Then we might say that X_C increases as the frequency goes down, or that X_C decreases as the frequency goes up. Nevertheless, we'll work with negative X_C values and stick with that convention. That way, the mathematics will give us the most accurate representation of how AC circuits behave when they contain capacitance.

Capacitive Reactance and Frequency

In a purely theoretical sense, capacitive reactance "mirrors" inductive reactance. In a geometrical or graphical sense, X_C constitutes a continuation of X_L into negative values, something like the extensions of the Celsius or Fahrenheit temperature scales to values "below zero."

If we specify the frequency of an AC source (in hertz) as f, and if we specify the capacitance of a component (in farads) as C, then we can calculate the capacitive reactance using the formula

$$X_C = -1/(2\pi fC) \approx -1/(6.2832fC)$$

This formula also works if we input f in megahertz and C in microfarads (μF). It could even apply for values of f in kilohertz (kHz) and values of C in *millifarads* (mF)—but you'll almost never see capacitances expressed in millifarads.

For Quick Reference

1 microfarad = 1 μF = 10^{-6} F
1 nanofarad = 1 nF = 10^{-9} F
1 picofarad = 1 pF = 10^{-12} F

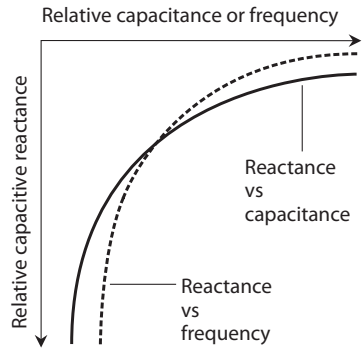
The function X_C versus f appears as a curve when we graph it in rectangular coordinates. The curve contains a *singularity* at $f = 0$; it "blows up negatively" as the frequency approaches zero. The function of X_C versus C also appears as a curve that attains a singularity at $C = 0$; it "blows up negatively" as the capacitance approaches zero. Summarizing:

- If we hold C constant, then X_C varies inversely with the negative of f.
- If we hold f constant, then X_C varies inversely with the negative of C.

Figure 14-4 illustrates these relations as graphs on a rectangular coordinate plane.

Read the Signs and Watch the Mess!

The arithmetic for dealing with capacitive reactance can give you trouble if you're not careful. You have to work with reciprocals, so the numbers can get awkward. Also, you have to watch those



14-4 Capacitive reactance is negatively, and inversely, proportional to capacitance. Capacitive reactance is also negatively, and inversely, proportional to frequency.

negative signs. You can easily forget to include a minus sign when you should (I've done that more than once), and you might insert a negative sign when you shouldn't. The signs are critical when you want to draw graphs describing systems containing reactance. A minus sign tells you that you're working with capacitive reactance rather than inductive reactance.

Problem 14-1

Suppose that a capacitor has a value of 0.00100 μF at a frequency of 1.00 MHz. What's the capacitive reactance?

Solution

We can apply the foregoing formula directly because we know the input data in microfarads (millionths) and in megahertz (millions):

$$X_C = -1/(6.2832 \times 1.00 \times 0.00100) = -1/(0.0062832)$$
$$= -159\ \Omega$$

Problem 14-2

What will happen to the capacitive reactance of the above-described capacitor if the frequency decreases to zero, so that the power source provides DC rather than AC?

Solution

In this case, we'll get an expression with 0 in the denominator if we plug the numbers into the capacitive-reactance formula, yielding a meaningless quantity. We might say, "The reactance of a capacitor at DC equals negative infinity," but a mathematician would wince at a statement like that. We'd do better to say, "We can't define the reactance of a capacitor at DC, but it doesn't matter because reactance applies only to AC circuits."

Problem 14-3

Suppose that a capacitor has a reactance of −100 Ω at a frequency of 10.0 MHz. What's its capacitance?

Solution

In this problem, we must put the numbers in the formula and then use algebra to solve for the unknown C. Let's start with the equation

$$-100 = -1/(6.2832 \times 10.0 \times C)$$

Dividing through by -100 and then multiplying through by C, we obtain

$$C = 1/(628.32 \times 10.0) = 1/6283.2 = 0.00015915$$

which rounds to $C = 0.000159$ μF. Because we input the frequency value in megahertz, this capacitance comes out in microfarads. We might also say that $C = 159$ pF, remembering that $1 \text{ pF} = 0.000001 \text{ μF} = 10^{-6}$ μF.

The RX_C Quarter-Plane

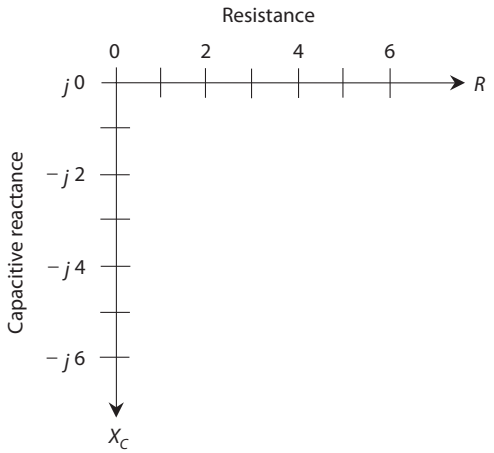
In a circuit containing resistance along with capacitive reactance, the characteristics work in two dimensions, in a way that "mirrors" the situation with the RX_L quarter-plane. We can place resistance and capacitive-reactance half-lines end-to-end at right angles to construct an RX_C *quarter-plane*, as shown in Fig. 14-5. We plot resistance values horizontally, with increasing values toward the right. We plot capacitive reactance values vertically, with increasingly negative values as we move downward. We can denote complex impedances Z containing both resistance and capacitance in the form

$$Z = R + jX_C$$

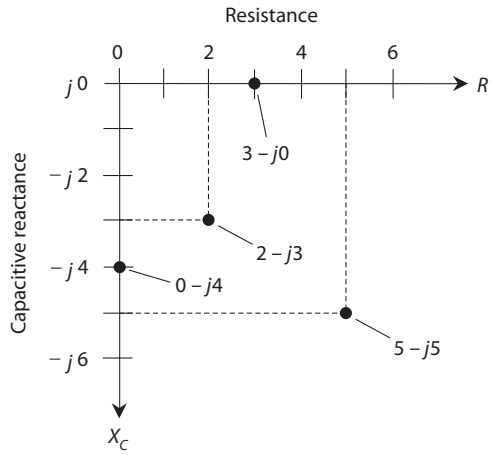
keeping in mind that values of X_C never go into positive territory.

Some RX_C Examples

If we have a pure resistance, say $R = 3\ \Omega$, then the complex-number impedance equals $Z = 3 + j0$, which corresponds to the point $(3, j0)$ on the RX_C quarter-plane. If we have a pure capacitive



14-5 The RX_C quarter-plane for capacitive reactance (X_C) and resistance (R).



14-6 Four points in the RX_C quarter-plane.

reactance, say $X_C = -4\ \Omega$, then the complex impedance equals $Z = 0 + j(-4)$, which we can write more simply as $Z = 0 - j4$ and plot at the point $(0,-j4)$ on the RX_C quarter-plane. The points representing $Z = 3 + j0$ (which we can also express as $Z = 3 - j0$ because the two values are identical) and $Z = 0 - j4$, along with two others, appear on the RX_C quarter-plane in Fig. 14-6.

Approaching the RX_C Extremes

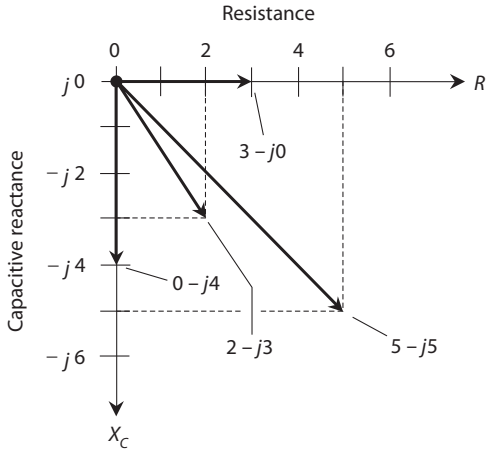
In practical circuits, all capacitors exhibit some *leakage conductance*. If the frequency goes to zero—the source produces DC—a tiny current will inevitably flow because no real-world dielectric material constitutes a perfect electrical insulator (not even a vacuum). Some capacitors have almost no leakage conductance, but none are completely free of it. Conversely, all electrical conductors have a little capacitive reactance, simply because they occupy physical space. Therefore, we'll never see a mathematically pure conductor of AC, either. The impedances $Z = 3 - j0$ and $Z = 0 - j4$ both represent theoretical idealizations.

How RX_C Points Move

Remember that the values for X_C indicate reactance values, not capacitance values. Reactance varies with the frequency in an RX_C circuit. If we raise or lower the AC frequency that we apply to a particular capacitor, the value of X_C changes. Increasing the AC frequency causes X_C to get *smaller negatively* (closer to zero). Reducing the AC frequency causes X_C to get *larger negatively* (farther from zero, or lower down on the RX_C quarter-plane). If the frequency drops all the way to zero, the capacitive reactance drops off the bottom of the quarter-plane and loses meaning. Then we have two plates or sets of plates holding opposite electrical charges, but no "action" unless or until we discharge the component.

Some RX_C Impedance Vectors

We can represent points in the RX_C quarter-plane as vectors, just as we do in the RX_L quarter-plane. Figure 14-6 shows four different points, each one represented by a certain distance to the right of, and/or below, the origin (corresponding to the complex impedance $0 - j0$). The first number in each value represents the resistance R and the second number represents the capacitive reactance X_C.



14-7 Four vectors in the RX_C quarter-plane, corresponding to the points shown in Fig. 14-6.

The RX_C combination constitutes a two-dimensional quantity. We can depict points, such as those shown in Fig. 14-6, by drawing vectors from the origin out to those points, as shown in Fig. 14-7.

Capacitors and DC Revisited

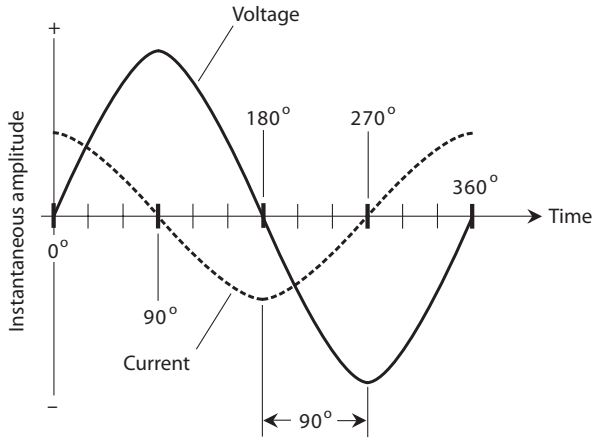
If the plates of a practical capacitor have large surface areas, are placed close together, and are separated by a good solid dielectric, we will experience a sudden, dramatic bit of "action" when we discharge the component. A massive capacitor can hold enough charge to electrocute an unsuspecting person who comes into contact with its terminals. The well-known scientist and American statesman *Benjamin Franklin* wrote about an experience of this sort with a "home-brewed" capacitor called a *Leyden jar*, which he constructed by placing metal foil sheets inside and outside a glass bottle and then connecting a high-voltage battery to them for a short while. After removing the battery, Franklin came into contact with both foil sheets at the same time and described the consequent shock as a "blow" that knocked him to the floor. Luckily for himself and the world, he survived. If you ever encounter a Leyden jar, treat it with the respect that it deserves, however "innocent" it might look. If you get careless, it can kill you!

Current Leads Voltage

When we drive AC through a capacitor and the instantaneous current starts to increase (in either direction), it takes a fraction of a cycle for the voltage between the plates to follow. Once the current starts decreasing from its maximum peak (in either direction) in the cycle, it again takes a fraction of a cycle for the voltage to follow. The instantaneous voltage can't keep up with the instantaneous current, as it does in a pure resistance. Therefore, in a circuit containing capacitive reactance as well as resistance, the voltage lags the current in phase. A more often-used expression for this phenomenon says that the current *leads* the voltage.

Pure Capacitive Reactance

Suppose that we connect an AC voltage source across a capacitor. Imagine that the frequency is low enough, and/or the capacitance is small enough, so the absolute value of the capacitive reactance X_C greatly exceeds the resistance R (by a factor of millions, say). In this situation, the current leads



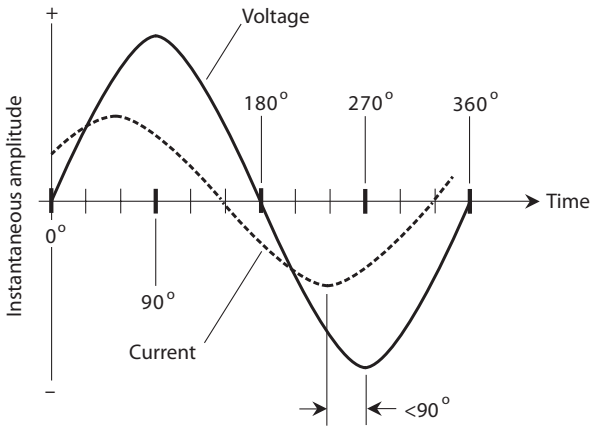
14-8 In a pure capacitive reactance, the current leads the voltage by 90°.

the voltage by just about 90° (Fig. 14-8). We have an essentially pure capacitive reactance, so the vector in the RX_C plane points almost exactly straight down, at an angle of almost exactly −90° with respect to the R axis.

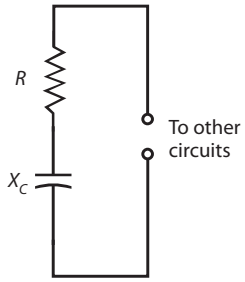
Capacitive Reactance and Resistance

When the resistance in a resistance-capacitance circuit compares favorably with the absolute value of the capacitive reactance, the current leads the voltage by an angle of less than 90° (Fig. 14-9). If R is small compared with the absolute value of X_C, the difference equals almost 90°. As R gets larger, or as the absolute value of X_C becomes smaller, the phase angle decreases. If R becomes much larger than the absolute value of X_C, the phase angle approaches 0°. We call a circuit containing resistance and capacitance an *RC circuit*.

The value of R in an RC circuit might increase relative to the absolute value of X_C because we add resistance deliberately into a circuit. Or, it might happen because the frequency becomes so high that the absolute value of X_C drops to a value comparable to the loss resistance in the circuit conductors. In either case, we can represent the circuit as a resistance R in series with a capacitive reactance X_C (Fig. 14-10).



14-9 In a circuit with capacitive reactance and resistance, the current leads the voltage by less than 90°.



14-10 Schematic representation of a circuit containing resistance and capacitive reactance.

If we know the values of X_C and R, we can find the *angle of lead*, also called the *RC phase angle* (or simply the *phase angle* if we know that we're dealing with resistance and capacitance), by plotting the point for $R - jX_C$ on the RX_C plane, drawing the vector from the origin out to that point, and then measuring the angle of the vector clockwise from the R axis. We can use a protractor to measure this angle, as we did in the previous chapter for *RL* phase angles, or we can use trigonometry to calculate the angle.

As with *RL* circuits, we need to know only the ratio of X_C to R to determine the phase angle. For example, if $X_C = -4\ \Omega$ and $R = 7\ \Omega$, you'll get the same angle as with $X_C = -400\ \Omega$ and $R = 700\ \Omega$, or with $X_C = -16\ \Omega$ and $R = 28\ \Omega$. The phase angle is the same whenever the ratio of X_C to R equals $-4{:}7$.

Pure Resistance

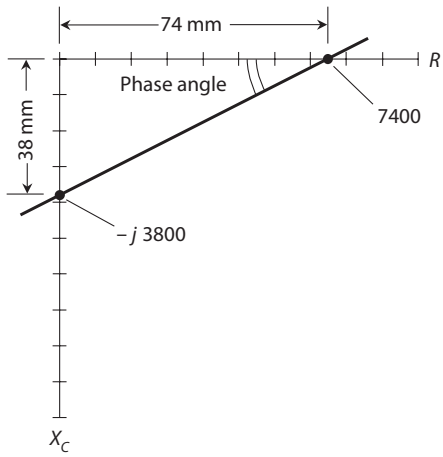
As the resistance in an *RC* circuit grows large compared with the absolute value of the capacitive reactance, the angle of lead grows smaller. The same thing happens if the absolute value of X_C gets small compared with the value of R. When R greatly exceeds the absolute value of X_C (regardless of their actual values), the vector in the *RC* plane points almost along the R axis. Then the *RC* phase angle is close to 0°. The voltage comes nearly into phase with the current. The plates of the capacitor do not come anywhere near getting fully charged with each cycle. The capacitor "passes the AC" with very little loss, as if it were shorted out. Nevertheless, it still has an extremely high value of X_C for any AC signals at much lower frequencies that might happen to exist across it at the same time. (Engineers put this property of capacitors to use in electronic circuits when they want to let high-frequency AC signals pass through a particular point while blocking signals at DC and at low AC frequencies.)

How Much Lead?

If you know the ratio of the capacitive reactance to the resistance X_C/R in an *RC* circuit, you can find the phase angle. Of course, you can find this angle if you know the precise values, too.

Pictorial Method

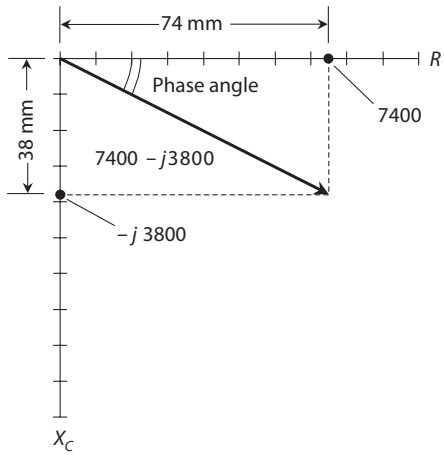
You can use a protractor and a ruler to find phase angles for *RC* circuits, just as you did with *RL* circuits in the previous chapter, as long as the angles aren't too close to 0° or 90°. First, draw a line somewhat longer than 100 mm, going from left to right on the paper. Then, use the protractor to construct a line going somewhat more than 100 mm vertically downward, starting at the left end of the horizontal line. The horizontal line forms the R axis of an RX_C quarter-plane. The line going down constitutes the X_C axis.



14-11 Pictorial method of finding phase angle in a circuit containing resistance and capacitive reactance.

If you know the actual values of X_C and R, divide or multiply them by a constant, chosen to make both values fall between −100 and 100. For example, if $X_C = -3800\ \Omega$ and $R = 7400\ \Omega$, divide them both by 100, getting −38 and 74. Plot these points on the lines. The X_C point should lie 38 mm below the intersection point between your two axes. The R point should lie 74 mm to the right of the intersection point. Next, draw a line connecting the two points, as shown in Fig. 14-11. This line will lie at a slant, and will form a triangle along with the two axes. Therefore, you'll get a right triangle, with the right angle at the origin of the quarter-plane. Using a protractor, measure the angle between the slanted line and the R axis. Extend the lines, if necessary, to get a good reading on the protractor. This angle will fall somewhere between 0° and 90°. Multiply this reading by −1 to get the RC phase angle. That is, if the protractor shows 27°, the RC phase angle equals −27°.

You can draw the actual vector by constructing a rectangle using the origin and your two points, making new perpendicular lines to complete the figure. The diagonal of this rectangle represents the vector, which runs out from the origin (Fig. 14-12). The angle between the R axis and this vector,



14-12 Another pictorial method of finding phase angle in a circuit containing resistance and capacitive reactance. This method shows the impedance vector.

multiplied by −1, gives you the phase angle. It has the same measure as the angle of the slanted line that you constructed in the process portrayed in Fig. 14-11.

Trigonometric Method

Using trigonometry, you can determine the *RC* phase angle more precisely than the pictorial method allows. Given the values of X_C and R, the *RC* phase angle equals the Arctangent of their ratio. We symbolize the phase angle in *RC* circuits by writing the lowercase Greek letter ϕ, just as we do in *RL* circuits. The formula is

$$\phi = \text{Arctan } (X_C/R)$$

When doing problems of this kind, remember to use the *capacitive reactance* values for X_C, and not the capacitance values. Also, use the actual value for X_C (a negative number) and not the absolute value of X_C (a positive number). If you know the capacitance but not the reactance, you must use the formula for X_C in terms of capacitance and frequency, and then calculate the phase angle. You should get angles that come out smaller than 0° but larger than −90°.

Don't Get Confused about the Angle!

By convention, phase angles in *RC* circuits always range from 0° down to −90°. This contrasts *RC* phase angles to *RL* phase angles, which always range from 0° up to 90°. You can avoid confusion about phase angles by remembering a simple rule: The phase angle always has the same sign (positive or negative) as the reactance.

Problem 14-4

Suppose that the capacitive reactance in an *RC* circuit equals –3800 Ω and the resistance equals 7400 Ω. What's the phase angle?

Solution

You can determine the ratio of the capacitive reactance to the resistance, getting

$$X_C/R = -3800/7400$$

The calculator display should show you something like –0.513513513. Find the Arctangent of this number, getting a phase angle of –27.18111109° on the calculator display. Round this result off to –27.18°.

Problem 14-5

Suppose that we operate an *RC* circuit at a frequency of 3.50 MHz. It has a resistance of 130 Ω and a capacitance of 150 pF. What's the phase angle to the nearest degree?

Solution

First, find the capacitive reactance for a capacitor of 150 pF at 3.50 MHz. Convert the capacitance to microfarads, getting $C = 0.000150$ μF. Remember that *micro*farads go with *mega*hertz. You'll get

$$\begin{aligned} X_C &= -1/(6.2832 \times 3.50 \times 0.000150) = -1/0.00329868 \\ &= -303\ \Omega \end{aligned}$$

Now you can find the ratio

$$X_C/R = -303/130 = -2.33$$

The phase angle equals the Arctangent of –2.33, which works out to be – 67° to the nearest degree.

Problem 14-6

What's the phase angle in the above-described circuit if you increase the frequency to 8.10 MHz?

Solution

You need to find the new value for X_C because it will change as a result of the frequency change. Calculating, you get

$$X_C = -1/(6.2832 \times 8.10 \times 0.000150) = -1/0.007634$$
$$= -131\ \Omega$$

The ratio X_C/R in this case equals –131/130, or –1.008. The phase angle equals the Arctangent of –1.008, which rounds off to –45°.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

15
CHAPTER

Impedance and Admittance

IN THIS CHAPTER, WE'LL DEVELOP A "RIGOROUS" WORKING MATHEMATICAL MODEL FOR COMPLEX impedance. We'll also learn about *admittance*, which quantifies how well AC circuits allow (or admit) the flow of current, rather than restraining (or impeding) it.

Imaginary Numbers Revisited

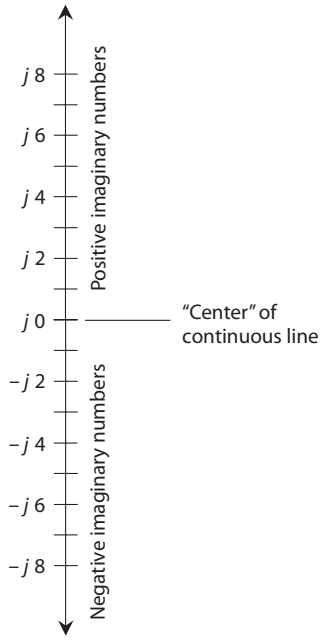
As we learned in Chap. 13, the engineering symbol j represents the unit imaginary number, technically defined as the positive square root of -1. Let's review this concept, because some people find it difficult to believe that negative numbers can have square roots. When we multiply j by itself, we get -1.

The term "imaginary" comes from the notion that j is somehow "less real" than the so-called *real numbers*. That's not true! All numbers are "unreal" in the sense that they're all abstract, however we classify them. You know that, if you've ever taken a course in number theory.

Actually, j isn't the "only" square root of -1. A negative square root of -1 also exists; it equals $-j$. When we multiply either j or $-j$ by itself, we get -1. (Pure mathematicians often denote these same numbers as i or $-i$.) The *set of imaginary numbers* comprises all possible real-number multiples of j. Examples include:

- $j \times 4$, which we write as $j4$
- $j \times 35.79$, which we write as $j35.79$
- $j \times (-25.76)$, which we write as $-j25.76$
- $j \times (-25{,}000)$, which we write as $-j25{,}000$

We can multiply j by any real number and portray it as a point on a line. If we do that for all the real numbers, we get an *imaginary-number line* (Fig. 15-1). We orient the imaginary number line vertically, at a right angle to the horizontal real number line, when we want to graphically render real and imaginary numbers at the same time. In electronics, real numbers represent resistances. Imaginary numbers represent reactances.



15-1 The imaginary-number line.

Complex Numbers Revisited (in Detail)

When we add a real number and an imaginary number, we get a *complex number*. In this context, the term "complex" does not mean "complicated"; a better word might be "composite." Examples include:

- The sum of 4 and $j5$, which equals $4 + j5$
- The sum of 8 and $-j7$, which equals $8 - j7$
- The sum of -7 and $j13$, which equals $-7 + j13$
- The sum of -6 and $-j87$, which equals $-6 - j87$

To completely portray the set of complex numbers in graphical form, we need a two-dimensional *coordinate plane*.

Adding and Subtracting Complex Numbers

When we want to add one complex number to another, we add the real parts and the complex parts separately and then sum up the total. For example, the sum of $4 + j7$ and $45 - j83$ works out as

$$(4 + 45) + j(7 - 83) = 49 + j(-76) = 49 - j76$$

Subtracting complex numbers involves a little trickery because we can easily confuse our signs. We can avoid this confusion by converting the difference to a sum. For example, we can find the

difference $(4 + j7) - (45 - j83)$ if we first multiply the second complex number by -1 and then add the result, obtaining

$$(4 + j7) - (45 - j83) = (4 + j7) + [-1(45 - j83)]$$
$$= (4 + j7) + (-45 + j83) = [4 + (-45)] + j(7 + 83)$$
$$= -41 + j90$$

Alternatively, you can subtract the real and imaginary parts separately, and then combine the result back into an imaginary number to get your final answer. Subtracting the negative of a quantity is the same as adding that quantity. Working out the above difference without converting to a sum, you'll get

$$(4 + j7) - (45 - j83) = (4 - 45) + j[7 - (-83)]$$
$$= -41 + j(7 + 83) = -41 + j90$$

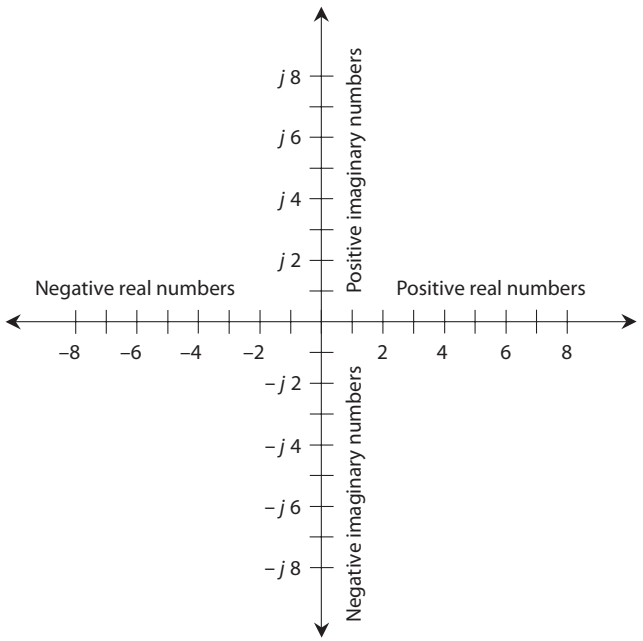
Multiplying Complex Numbers

When we want to multiply one complex number by another, we should treat them both as sums of number pairs—that is, as *binomials*. If we have four real numbers a, b, c, and d, then

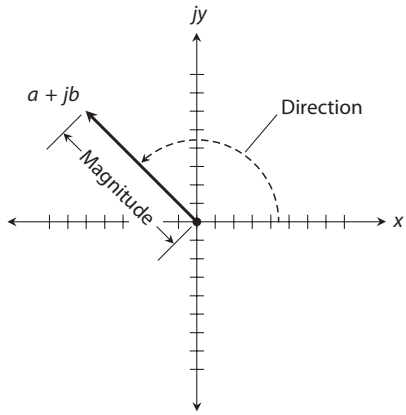
$$(a + jb)(c + jd) = ac + jad + jbc + j^2 bd$$
$$= (ac - bd) + j(ad + bc)$$

The Complex Number Plane

We can construct a complete *complex-number plane* by taking the real and imaginary number lines and placing them together, at right angles, so that they intersect at the zero points, 0 and $j0$. Figure 15-2



15-2 The complex number plane.



15-3 Magnitude and direction of a vector in the complex number plane.

illustrates the arrangement, which gives us a *rectangular coordinate plane,* just like the ones that people use to graph everyday relations, such as temperature versus time.

Complex Number Vectors

Engineers sometimes represent complex numbers as vectors in the coordinate plane. This gives each complex number a unique *magnitude* and a unique *direction*. The magnitude of the vector for a complex number $a + jb$ equals the distance of the point (a, jb) from the origin $(0, j0)$. We represent the vector direction as the angle, expressed going counterclockwise from the positive real-number axis. Figure 15-3 illustrates how this scheme works.

Absolute Value

The *absolute value* of a complex number $a + jb$ equals the length, or magnitude, of its vector in the complex plane, measured from the origin $(0, j0)$ to the point (a, jb). Let's break this scenario down into three cases.

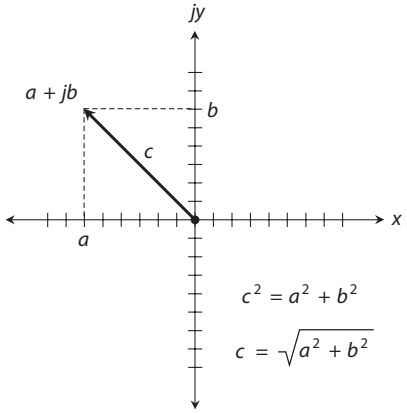
1. For a *pure real number* $a + j0$, the absolute value equals a, if a is positive. If a is negative, the absolute value of $a + j0$ equals $-a$.
2. For a *pure imaginary number* $0 + jb$, the absolute value equals b, if b (a real number) is positive. If b is negative, the absolute value of $0 + jb$ equals $-b$.
3. If the number $a + jb$ is neither a pure real nor a pure imaginary number, we must use a formula to find its absolute value. First, we square both a and b. Then we add those two squares. Finally, we take the square root of the sum of the squares to get the length c of the vector representing $a + jb$. Figure 15-4 shows the geometry of this method.

Problem 15-1

Find the absolute value of the complex number $-22 - j0$.

Solution

We have a pure real number in this case. Actually, $-22 - j0$ is the very same complex number as $-22 + j0$, because $-j0 = j0$. The absolute value equals $-(-22)$, or 22.



15-4 Calculation of the absolute value (length) of a vector. Here, we represent the vector length as *c*.

Problem 15-2

Find the absolute value of $0 - j34$.

Solution

This quantity is a pure imaginary number where $b = -34$ because $0 - j34 = 0 + j(-34)$. The absolute value equals –(–34), or 34.

Problem 15-3

Find the absolute value of $3 - j4$.

Solution

In this case, we have $a = 3$ and $b = -4$. Using the formula described above and shown in Fig. 15-4, we get

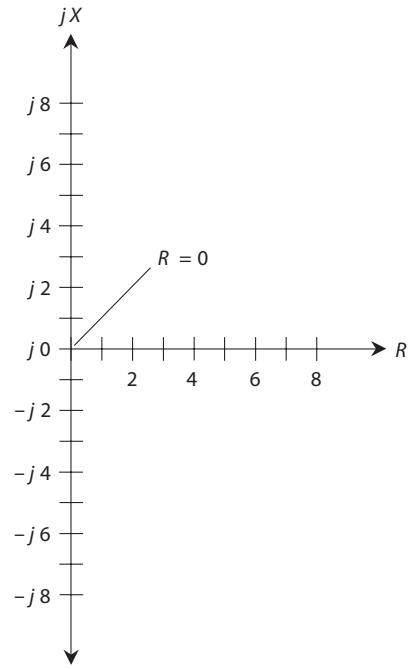
$$[3^2 + (-4)^2]^{1/2} = (9 + 16)^{1/2} = 25^{1/2} = 5$$

The *RX* Half-Plane

Recall the quarter-plane for resistance R and inductive reactance X_L from Chap. 13. This region corresponds to the upper-right quadrant of the complex number plane shown in Fig. 15-2. Similarly, the quarter-plane for resistance R and capacitive reactance X_C corresponds to the lower-right quadrant of the complex number plane of Fig. 15-2. We represent resistances as nonnegative real numbers. We represent reactances as imaginary numbers.

No Negative Resistance

Strictly speaking, negative resistance can't exist because we can't have anything better than a perfect conductor. In some cases, we might treat a source of DC, such as a battery, as if it constitutes a "negative resistance." Once in a while, we might encounter a device in which the current drops as the applied voltage increases, producing a "reversed" resistance-behavior phenomenon that some engineers call "negative resistance." But for most practical applications, the resistance can never go



15-5 The complex impedance half-plane, also called the resistance-reactance (*RX*) half-plane.

"below zero." We can, therefore, remove the negative axis, along with the upper-left and lower-left quadrants, of the complex-number plane, obtaining an *RX half-plane,* as shown in Fig. 15-5. This system provides a complete set of coordinates for depicting complex impedance.

"Negative Inductors" and "Negative Capacitors"

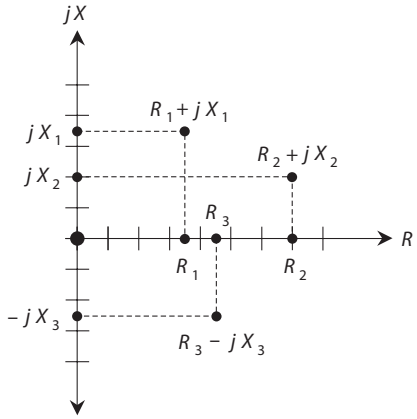
Capacitive reactance X_C is effectively an extension of inductive reactance X_L into the realm of negatives. Capacitors act like "negative inductors." We can also say that inductors act like "negative capacitors" because the negative of a negative number equals a positive number. Reactance, in general, can vary from extremely large negative values, through zero, to extremely large positive values.

Complex Impedance Points

Imagine the point representing $R + jX$ moving around in the *RX* half-plane, and imagine where the corresponding points on the axes lie. We can locate these points by drawing dashed lines from the point $R + jX$ to the R and X axes, so that the dashed lines intersect the axes at right angles. Figure 15-6 shows several examples.

Now think of the points for R and X moving toward the right and left, or up and down, on their axes. Imagine what happens to the point representing $R + jX$ in various scenarios. This exercise will give you an idea of how impedance changes as the resistance and reactance vary in an AC electrical circuit.

Resistance constitutes a one-dimensional phenomenon. Reactance also manifests itself as one-dimensional. To completely define complex impedance, however, we must use two dimensions. The *RX* half-plane meets this requirement. Remember that the resistance and the reactance can vary independently of one another!



15-6 Some points in the *RX* half-plane, showing their resistance and reactance components.

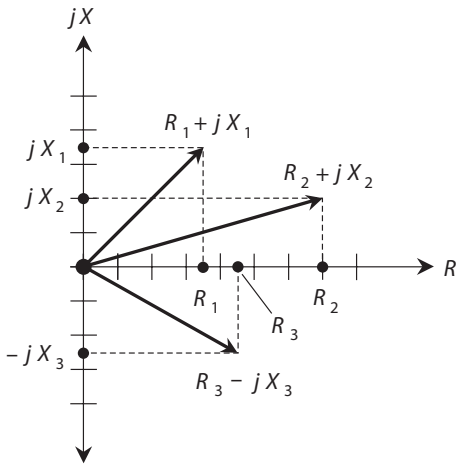
Complex Impedance Vectors

We can represent any impedance $R + jX$ as a complex number of the form $a + jb$. We simply let $R = a$ and $X = b$. Now we can envision how the impedance vector changes as we vary either the resistance R or the reactance X, or both, independently. If X remains constant, an increase in R causes the complex impedance vector to grow longer. If R remains constant and X_L gets larger, the vector grows longer. If R stays the same but X_C gets larger negatively, the vector once again grows longer. Figure 15-7 illustrates the vectors corresponding to the points from Fig. 15-6.

Absolute-Value Impedance

You'll occasionally read or hear that the "impedance" of some device or component equals a certain number of "ohms." For example, in audio electronics, you'll encounter things like "8-Ω" speakers and "600-Ω" amplifier inputs. How, you ask, can manufacturers quote a single number for a quantity that needs two dimensions for its complete expression? That's a good question. Two answers exist.

First, specifications, such as "8 Ω" for a speaker or "600 Ω" for an amplifier input, refer to *purely resistive impedances*, also known as *nonreactive impedances*. Thus, the "8-Ω" speaker



15-7 Vectors representing the points shown in Fig. 15-6.

really has a complex impedance of $8 + j0$, and the "600-Ω" input circuit is designed to operate with a complex impedance at, or near, $600 + j0$.

Second, you can talk about the length of the impedance vector (the absolute value of the complex impedance), calling this length a certain number of ohms. If you try to define impedance this way, however, you risk ambiguity and confusion because you can find infinitely many different vectors of a given length in the *RX* half-plane.

Sometimes, engineers and technicians write the uppercase, italic letter Z in place of the word "impedance," so you'll read expressions such as "$Z = 50$ Ω" or "$Z = 300$ Ω nonreactive." In this context, if no specific impedance is given, "$Z = 8$ Ω" can theoretically refer to $8 + j0$, $0 + j8$, $0 - j8$, or any other complex impedance whose point lies on a half-circle centered at the coordinate origin and having a radius of 8 units, as shown in Fig. 15-8.

Problem 15-4

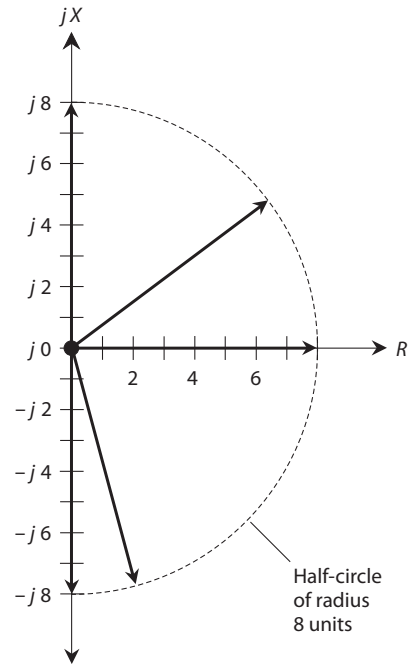
Name seven different complex impedances that the expression "$Z = 10$ Ω" might mean.

Solution

We can easily name three such impedance values, each consisting of a pure reactance or a pure resistance, as follows:

$$Z_1 = 0 + j10$$
$$Z_2 = 10 + j0$$
$$Z_3 = 0 - j10$$

These impedances represent pure inductance, pure resistance, and pure capacitance, respectively.



15-8 Some vectors representing an absolute-value impedance of 8 Ω. Infinitely many such vectors exist in theory, all of which terminate on the dashed circle.

A right triangle can exist having sides in a ratio of 6:8:10 units. We know this fact from basic coordinate geometry, because $6^2 + 8^2 = 10^2$. Therefore, we can have the following impedances, all of which have an absolute value of "10 Ω":

$$Z_4 = 6 + j8$$
$$Z_5 = 6 - j8$$
$$Z_6 = 8 + j6$$
$$Z_7 = 8 - j6$$

Conductance

In an AC circuit, electrical *conductance* behaves exactly as it does in a DC circuit. We symbolize conductance (as a variable in equations) by writing an uppercase italic letter *G*. We express the relationship between conductance and resistance as the two formulas

$$G = 1/R$$

and

$$R = 1/G$$

The standard unit of conductance is the *siemens*, abbreviated as the uppercase non-italic letter S. In the above formulas, we get *G* in siemens if we input *R* in ohms, and vice versa. As the conductance increases, the resistance decreases, and more current flows for a fixed applied voltage. Conversely, as *G* decreases, *R* goes up, and less current flows when we apply a fixed voltage.

Susceptance

Sometimes we'll come across the term *susceptance* in reference to AC circuits. We symbolize this quantity (as a variable in equations) by writing an uppercase italic letter *B*. Susceptance is the reciprocal of reactance, and it can occur in either the capacitive form or the inductive form. If we symbolize *capacitive susceptance* as B_C and *inductive susceptance* as B_L, then

$$B_C = 1/X_C$$

and

$$B_L = 1/X_L$$

The Reciprocal of *j*

All values of *B* theoretically contain the *j* operator, just as do all values of *X*. But when it comes to finding reciprocals of quantities containing *j*, things get tricky. The reciprocal of *j* actually equals its negative! That is,

$$1/j = -j$$

and

$$1/(-j) = j$$

As a result of these properties of j, the sign reverses whenever you find a susceptance value in terms of a reactance value. When expressed in terms of j, inductive susceptance is negative imaginary, and capacitive susceptance is positive imaginary—just the opposite situation from inductive reactance and capacitive reactance.

Imagine an inductive reactance of 2 Ω. We express this in imaginary terms as $j2$. To find the inductive susceptance, we must find $1/(j2)$. Mathematically, we convert this expression to a real-number multiple of j by breaking it down in steps as follows:

$$1/(j2) = (1/j)(1/2) = (1/j)0.5 = -j0.5$$

Now imagine a capacitive reactance of 10 Ω. We express this quantity in imaginary terms as $-j10$. To find the capacitive susceptance, we must find $1/(-j10)$. Here's how we can convert it to the product of j and a real number:

$$1/(-j10) = (1/-j)(1/10) = (1/-j)0.1 = j0.1$$

To find an imaginary value of susceptance in terms of an imaginary value of reactance, we must first take the reciprocal of the real-number part of the expression, and then multiply the result by –1.

Problem 15-5

Suppose that we have a capacitor of 100 pF at a frequency of 3.10 MHz. What's the capacitive susceptance B_C?

Solution

First, let's find X_C by using the formula for capacitive reactance. We have

$$X_C = -1/(6.2832\,fC)$$

Note that 100 pF = 0. 000100 μF. Therefore,

$$\begin{aligned} X_C &= -1/(6.2832 \times 3.10 \times 0.000100) = -1/0.00195 \\ &= -513\ \Omega \end{aligned}$$

The imaginary value of X_C equals $-j513$. The susceptance B_C equals $1/X_C$, so we have

$$B_C = 1/(-j513) = j0.00195\ \text{S}$$

The siemens quantifies susceptance, just as it defines conductance. Therefore, we can state the foregoing result as 0.00195 S of capacitive susceptance.

General Formula for B_C

We can now see that the general formula for capacitive susceptance in siemens, in terms of frequency in hertz and capacitance in farads, is

$$B_C = 6.2832\,fC$$

This formula also works for frequencies in megahertz and capacitance values in microfarads.

Problem 15-6

Suppose an inductor has $L = 163$ μH at a frequency of 887 kHz. What's the inductive susceptance B_L?

Solution

Note that 887 kHz = 0.887 MHz. We can calculate X_L from the formula for inductive reactance as follows:

$$X_L = 6.2832\,fL = 6.2832 \times 0.887 \times 163 = 908\ \Omega$$

The imaginary value of X_L equals $j908$. The susceptance B_L equals $1/X_L$. It follows that

$$B_L = -1/(j908) = -j0.00110\ \text{S}$$

We can state this result as -0.00110 S of inductive susceptance.

General Formula for B_L

The general formula for inductive susceptance in siemens, in terms of frequency in hertz and inductance in henrys, is

$$B_L = -1/(6.2832\,fL)$$

This formula also works for frequencies in kilohertz and inductance values in millihenrys, and for frequencies in megahertz and inductance values in microhenrys.

Admittance

Real-number conductance and imaginary-number susceptance combine to form *complex admittance*, symbolized (as a variable in equations) as the uppercase italic letter *Y*. Admittance provides a complete expression of the extent to which a circuit allows AC to flow. As the absolute value of complex impedance gets larger, the absolute value of complex admittance becomes smaller, in general. Huge impedances correspond to tiny admittances, and vice versa.

Complex Admittance

We can express admittance in complex form, just as we can do with impedance. However, we'd better keep careful track of which quantity we're talking about! We can avoid confusion if we take care to employ the correct symbol. We can get a complete expression of admittance by taking the complex composite of conductance and susceptance. Engineers usually write complex admittance in the form

$$Y = G + jB$$

when the susceptance is positive (capacitive), and in the form

$$Y = G - jB$$

when the susceptance is negative (inductive).

The "Parallel Advantage"

In Chaps. 13 and 14, we worked with series *RL* and *RC* circuits. Did you wonder, at that time, why we ignored parallel circuits in those discussions? We had a good reason: Admittance works far

better than impedance when we want to mathematically analyze parallel AC circuits. In parallel AC circuits, resistance and reactance combine to make a mathematical mess. But conductance and susceptance add directly together in parallel circuits, yielding admittance. We'll analyze parallel *RL* and *RC* circuits in the next chapter.

The *GB* Half-Plane

We can portray complex admittance on a coordinate grid similar to the complex-impedance (*RX*) half-plane. We get a half-plane, not a complete plane because no such thing as negative conductance exists in the "real world." (We can't have a component that conducts worse than not at all!) We plot conductance values along a horizontal *G* axis. We plot susceptance along a vertical *B* axis. Figure 15-9 shows several points on the *GB half-plane*.

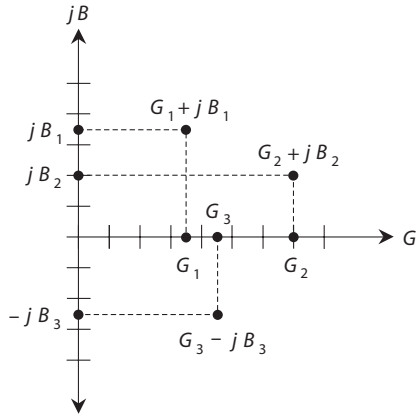
It's Inside-Out

Superficially, the *GB* half-plane looks identical to the *RX* half-plane. But mathematically, the two couldn't differ more! The *GB* half-plane is "inside-out" with respect to the *RX* half-plane. The center, or origin, of the *GB* half-plane represents the point at which no conductance exists for DC or for AC. It represents the *zero-admittance point* rather than the *zero-impedance point*. In the *GB* half-plane, the origin corresponds to a perfect open circuit. In the *RX* half-plane, the origin represents a perfect short circuit.

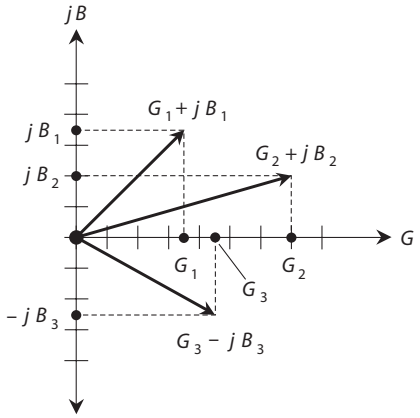
As you move out toward the right ("east") along the *G*, or conductance, axis of the *GB* half-plane, the conductance improves, and the current gets greater. When you move upward ("north") along the *jB* axis from the origin, you have ever-increasing positive (capacitive) susceptance. When you go downward ("south") along the *jB* axis from the origin, you encounter increasingly negative (inductive) susceptance.

Vector Representation of Admittance

We can denote specific complex admittance values as vectors, just as we can do with complex impedance values. Figure 15-10 shows the points from Fig. 15-9 as complex admittance vectors. Given a fixed applied AC voltage, long vectors in the *GB* half-plane generally indicate large currents, and short vectors indicate small currents.



15-9 Some points in the *GB* half-plane, along with their conductance and susceptance components.



15-10 Vectors representing the points shown in Fig. 15-9.

Imagine a point moving around on the *GB* half-plane. Think of the vector getting longer and shorter, and changing direction as well. Vectors pointing generally "northeast," or upward and to the right, correspond to conductances and capacitances in parallel. Vectors pointing in a more or less "southeasterly" direction, or downward and to the right, portray conductances and inductances in parallel.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

16
CHAPTER

Alternating-Current Circuit Analysis

IN THIS CHAPTER, WE WILL BRING TOGETHER WHAT WE HAVE LEARNT SEPARATELY ABOUT REACTANCE in inductors and capacitors and look at analyzing AC circuits that contain resistors, capacitors, and inductors.

When you see an AC circuit that contains coils and/or capacitors, you can envision a complex-number half-plane, either *RX* (resistance-reactance) or *GB* (conductance-susceptance). The *RX* half-plane applies to series circuit analysis. The *GB* half-plane applies to parallel circuit analysis.

Complex Impedances in Series

In any situation where resistors, coils, and capacitors are connected in series, each component has an impedance that we can represent as a vector in the *RX* half-plane. The vectors for resistors remain constant regardless of the frequency. But the vectors for coils and capacitors vary as the frequency increases or decreases.

Pure Reactances

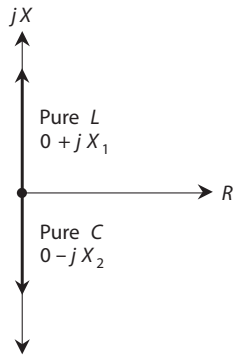
Pure inductive reactance (X_L) and capacitive reactance (X_C) simply add together when we connect coils and capacitors in series. That is,

$$X = X_L + X_C$$

In the *RX* half-plane, their vectors add, but because these vectors point in exactly opposite directions—inductive reactance upward and capacitive reactance downward, as shown in Fig. 16-1—the resultant sum vector inevitably points either straight up or straight down, unless the reactances are equal and opposite, in which case their sum equals the zero vector.

Problem 16-1

Suppose that we connect a coil and a capacitor in series with $jX_L = j200$ and $jX_C = -j150$. What's the net reactance?



16-1 We can represent pure inductance and pure capacitance as reactance vectors that point straight up and down.

Solution

We add the values to get

$$jX = j200 + (-j150) = j(200 - 150) = j50$$

Problem 16-2

Suppose that we connect a coil and capacitor in series with $jX_L = j30$ and $jX_C = -j110$. What's the net reactance?

Solution

Again, we add the values, obtaining

$$jX = j30 + (-j110) = j(30 - 110) = -j80$$

Problem 16-3

Suppose that we connect a coil of $L = 5.00$ μH and a capacitor of $C = 200$ pF in series, and then drive an AC signal through the combination at $f = 4.00$ MHz. What's the net reactance?

Solution

First, we calculate the reactance of the inductor at 4.00 MHz to get

$$\begin{aligned} jX_L &= j6.2832fL \\ &= j(6.2832 \times 4.00 \times 5.00) = j125.664 \end{aligned}$$

Next, we calculate the reactance of the capacitor at 4.00 MHz (noting that 200 pF = 0.000200 μF) to get

$$\begin{aligned} jX_C &= -j[1/(6.2832fC)] \\ &= -j[1/(6.2832 \times 4.00 \times 0.000200)] = -j198.943 \end{aligned}$$

Finally, we add the reactances and round off to three significant figures, obtaining

$$jX = j125.664 + (-j198.943) = -j73.3$$

Beware of Cumulative Rounding Errors!

Do you wonder why we carried through some extra digits during the course of the preceding calculations, rounding off only at the very end of the process? The extra digits in the intermediate steps reduce our risk of ending up with a so-called *cumulative rounding error*. In situations where repeated rounding introduces many small errors, the final calculation can end up a full digit or two off, even after rounding to the appropriate number of significant figures. Let's use this precaution, when necessary, in all of our future calculations.

Problem 16-4

What's the net reactance of the above-described combination at $f = 10.0$ MHz?

Solution

First, we calculate the reactance of the inductor at 10.0 MHz to get

$$\begin{aligned} jX_L &= j6.2832fL \\ &= j(6.2832 \times 10.0 \times 5.00) = j314.16 \end{aligned}$$

Next, we calculate the reactance of the capacitor at 10.00 MHz to get

$$\begin{aligned} jX_C &= -j[1/(6.2832fC)] \\ &= -j[1/(6.2832 \times 10.0 \times 0.000200)] = -j79.58 \end{aligned}$$

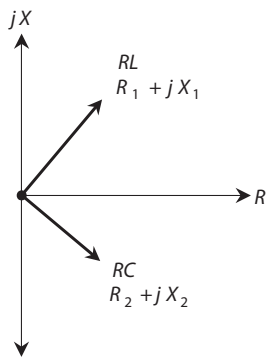
Finally, we add the reactances and round off, obtaining

$$jX = j314.16 + (-j79.58) = j235$$

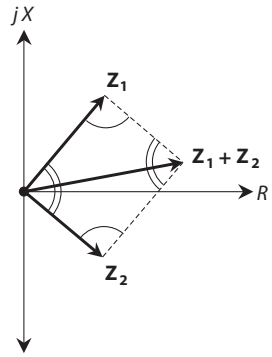
Adding Impedance Vectors

Whenever the resistance in a series circuit reaches values that are significant compared with the reactance, the impedance vectors no longer point straight up and straight down. Instead, they run off toward the "northeast" (for the inductive part of the circuit) and "southeast" (for the capacitive part). Figure 16-2 shows an example of this condition.

When two impedance vectors don't lie along a single line, we must use *vector addition* to get the correct net impedance vector. Figure 16-3 shows the geometry of vector addition. We construct



16-2 When we have resistance along with reactance, the impedance vectors are neither vertical nor horizontal.



16-3 Parallelogram method of complex-impedance vector addition.

a *parallelogram*, using the two vectors $Z_1 = R_1 + jX_1$ and $Z_2 = R_2 + jX_2$ as two adjacent sides of the figure. The diagonal of the parallelogram constitutes the vector representing the net complex impedance. In a parallelogram, pairs of opposite angles always have equal measures. These equalities are indicated by the pairs of single and double arcs in Fig. 16-3.

Formula for Complex Impedances in Series

Consider two complex impedances, $Z_1 = R_1 + jX_1$ and $Z_2 = R_2 + jX_2$. If we connect these impedances in series, we can represent the net impedance Z as the vector sum

$$Z = (R_1 + jX_1) + (R_2 + jX_2) = (R_1 + R_2) + j(X_1 + X_2)$$

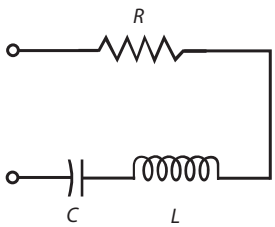
The resistance and reactance components add separately. Remember that inductive reactances are positive imaginary while capacitive reactances are negative imaginary!

Series *RLC* Circuits

With an inductance, capacitance, and resistance in series (Fig. 16-4), you can imagine the resistance R as belonging entirely to the coil, if you want to take advantage of the above formulas. Then you have only two vectors to add (instead of three), when you calculate the impedance of the series *RLC* circuit. Mathematically, the situation works out as follows:

$$Z = (R + jX_L) + (0 + jX_C) = R + j(X_L + X_C)$$

Again, remember that X_C is never positive! So, although these general formulas contain addition symbols exclusively, you must add in a negative value (the equivalent of subtraction) when you include a capacitive reactance.



16-4 A series resistance-inductance-capacitance (*RLC*) circuit.

Problem 16-5

Suppose that we connect a resistor, a coil, and a capacitor in series with $R = 50\ \Omega$, $X_L = 22\ \Omega$, and $X_C = -33\ \Omega$. What's the net impedance Z?

Solution

We can consider the resistor as part of the coil, obtaining $50 + j22$ and $0 - j33$. Adding these gives the resistance component of $50 + 0 = 50$, and the reactive component of $j22 - j33 = -j11$. Therefore, $Z = 50 - j11$.

Problem 16-6

Consider a resistor, a coil, and a capacitor in series with $R = 600\ \Omega$, $X_L = 444\ \Omega$, and $X_C = -444\ \Omega$. What's the net impedance, Z?

Solution

Again, we can imagine the resistor as part of the inductor. Then the complex impedance vectors are $600 + j444$ and $0 - j444$. Adding these, the resistance component equals $600 + 0 = 600$, and the reactive component equals $j444 - j444 = j0$. We have a net impedance $Z = 600 + j0$, a purely resistive impedance.

Series Resonance

When a series-connected *RLC* circuit has zero net reactance at a certain frequency, we say that the circuit exhibits *series resonance* at that frequency.

Problem 16-7

Consider a resistor, a coil, and a capacitor connected in series. The resistor has a value of 330 Ω, the capacitance equals 220 pF, and the inductance equals 100 μH. We operate the circuit at 7.15 MHz. What's the complex impedance?

Solution

First, we calculate the inductive reactance. Remember that

$$X_L = 6.2832fL$$

and that megahertz and microhenrys go together in the formula. We multiply to obtain

$$jX_L = j(6.2832 \times 7.15 \times 100) = j4492$$

Next, we calculate the capacitive reactance using the formula

$$X_C = -1/(6.2832fC)$$

We can convert 220 pF to microfarads, obtaining $C = 0.000220$ μF. Then we have

$$jX_C = -j[1/(6.2832 \times 7.15 \times 0.000220)] = -j101$$

Now, we can lump the resistance and the inductive reactance together, so one of the impedances becomes $330 + j4492$. The other impedance equals $0 - j101$. Adding these gives us

$$Z = 330 + j4492 - j101 = 330 + j4391$$

We can justify only three significant digits of accuracy here, so we might want to state this result as $Z = 330 + j4.39\text{k}$ (remembering that "k" stands for "kilohms").

Problem 16-8

Suppose that we connect a resistor, a coil, and a capacitor in series. The resistance equals 50.0 Ω, the inductance equals 10.0 μH, and the capacitance equals 1000 pF. We operate the circuit at 1592 kHz. What's the complex impedance?

Solution

First, let's calculate the inductive reactance. Note that 1592 kHz = 1.592 MHz. Plugging in the numbers, we obtain

$$jX_L = j(6.2832 \times 1.592 \times 10.0) = j100$$

Next, we calculate the capacitive reactance. Let's convert picofarads to microfarads, and use megahertz for the frequency. Then we have

$$\begin{aligned} jX_C &= -j[1/(6.2832 \times 1.592 \times 0.001000)] \\ &= -j100 \end{aligned}$$

When we put the resistance and the inductive reactance together into a single complex number, we get $50.0 + j100$. We know that the capacitor's impedance is $0 - j100$. Adding the two complex numbers, we get

$$Z = 50.0 + j100 - j100 = 50.0 + j0$$

Our circuit exhibits a pure resistance of 50.0 Ω at 1592 kHz.

Complex Admittances in Parallel

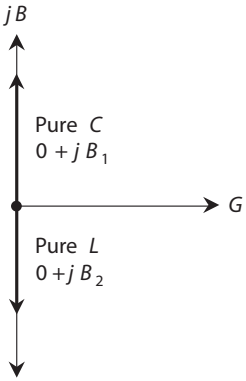
When you see resistors, coils, and capacitors in parallel, remember that each component, whether a resistor, an inductor, or a capacitor, has an admittance that you can represent as a vector in the *GB* half-plane. The vectors for pure conductances remain constant, even as the frequency changes. But the vectors for the coils and capacitors vary with frequency.

Pure Susceptances

Pure inductive susceptance (B_L) and capacitive susceptance (B_C) add together when coils and capacitors appear in parallel. That is,

$$B = B_L + B_C$$

Remember that B_L is never positive, and B_C is never negative; we must invert the "sign scenario" with susceptance values as compared with reactance values.



16-5 We can represent pure capacitance and pure inductance as susceptance vectors that point straight up and down.

In the *GB* half-plane, pure jB_L and jB_C vectors add. Because such vectors always point in exactly opposite directions—capacitive susceptance upward and inductive susceptance downward—the sum, jB, points either straight up or straight down, as shown in Fig. 16-5, unless the susceptances happen to be equal and opposite, in which case they cancel and the result equals the zero vector.

Problem 16-9

Consider a coil and capacitor connected in parallel with $jB_L = -j0.05$ and $jB_C = j0.08$. What's the net susceptance?

Solution

We add the values to get

$$jB = jB_L + jB_C = -j0.05 + j0.08 = j0.03$$

Problem 16-10

Suppose that we connect a coil and capacitor in parallel with $jB_L = -j0.60$ and $jB_C = j0.25$. What's the net susceptance?

Solution

Again, we add the values to obtain

$$jB = -j0.60 + j0.25 = -j0.35$$

Problem 16-11

Suppose that we connect a coil of $L = 6.00$ μH and a capacitor of $C = 150$ pF in parallel and drive a signal through them at $f = 4.00$ MHz. What's the net susceptance?

Solution

First, we calculate the susceptance of the inductor at 4.00 MHz, obtaining

$$\begin{aligned} jB_L &= -j[1/(6.2832fL)] \\ &= -j[1/(6.2832 \times 4.00 \times 6.00)] = -j0.00663144 \end{aligned}$$

Next, we calculate the susceptance of the capacitor (converting its value to microfarads) at 4.00 MHz, getting

$$\begin{aligned} jB_C &= j(6.2832fC) \\ &= j(6.2832 \times 4.00 \times 0.000150) = j0.00376992 \end{aligned}$$

Finally, we add the inductive and capacitive susceptances and round off to three significant figures, ending up with

$$jB = -j0.00663144 + j0.00376992 = -j0.00286$$

Problem 16-12

What's the net susceptance of the above parallel-connected inductor and capacitor at a frequency of $f = 5.31$ MHz?

Solution

First, we calculate the susceptance of the inductor at 5.31 MHz to get

$$jB_L = -j[1/(6.2832 \times 5.31 \times 6.00)] = -j0.00499544$$

Next, we calculate the susceptance of the capacitor (converting its value to microfarads) at 5.31 MHz, getting

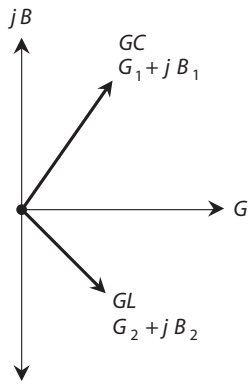
$$jB_C = j(6.2832 \times 5.31 \times 0.000150) = j0.00500457$$

Finally, we add the inductive and capacitive susceptances and round off to three significant figures, obtaining

$$jB = -j0.00499544 + j0.00500 = j0.00$$

Adding Admittance Vectors

When the conductance is significant in a parallel circuit containing inductance and capacitance, the admittance vectors don't point straight up and down. Instead, they run off toward the "northeast" (for the capacitive part of the circuit) and "southeast" (for the inductive part), as shown in Fig. 16-6.



16-6 When we have conductance along with susceptance, admittance vectors are neither vertical nor horizontal.

You've seen how vectors add in the *RX* half-plane. In the *GB* half-plane, things work in the same way. The net admittance vector equals the sum of the component admittance vectors.

Formula for Complex Admittances in Parallel

Imagine that we connect two admittances $Y_1 = G_1 + jB_1$ and $Y_2 = G_2 + jB_2$ in parallel. We can find the net admittance Y as the complex-number sum

$$Y = (G_1 + jB_1) + (G_2 + jB_2) = (G_1 + G_2) + j(B_1 + B_2)$$

Parallel *RLC* Circuits

When we connect a coil, capacitor, and resistor in parallel (Fig. 16-7), we can think of the resistance as a *conductance* in siemens (symbolized S), which equals the reciprocal of the value in ohms. If we consider the conductance as part of the inductor, we have only two complex numbers to add, rather than three, when finding the admittance of a parallel *RLC* circuit. We can use the formula

$$Y = (G + jB_L) + (0 + jB_C) = G + j(B_L + B_C)$$

Again, we must remember that B_L is never positive! So, although the formulas here have addition symbols in them, we actually subtract a value for the inductive susceptance by adding in a negative number.

Problem 16-13

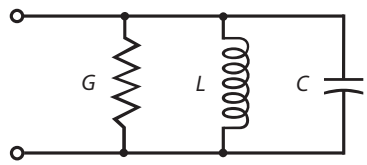
Suppose that we connect a resistor, a coil, and a capacitor in parallel. The resistor has a conductance of $G = 0.10$ S. The susceptances are $jB_L = -j0.010$ and $jB_C = j0.020$. What's the complex admittance of this combination?

Solution

Let's consider the resistor as part of the coil, so we have two complex admittances in parallel: $0.10 - j0.010$ and $0.00 + j0.020$. Adding these values "part-by-part" gives us a conductance component of $0.10 + 0.00 = 0.10$ and a susceptance component of $-j0.010 + j0.020 = j0.010$. Therefore, the complex admittance equals $0.10 + j0.010$.

Problem 16-14

Consider a resistor, a coil, and a capacitor in parallel. The resistor has a conductance of $G = 0.0010$ S. The susceptances are $jB_L = -j0.0022$ and $jB_C = j0.0022$. What is the complex admittance of this combination?



16-7 A parallel resistance-inductance-capacitance (*RLC*) circuit. Here, *G* represents conductance (the reciprocal of resistance), so we can just as well call this a *conductance-inductance-capacitance* (*GLC*) circuit.

Solution

Again, let's consider the resistor as part of the coil. Then the complex admittances are $0.0010 - j0.0022$ and $0.0000 + j0.0022$. Adding these, we get a conductance component of $0.0010 + 0.0000 = 0.0010$ and a susceptance component of $-j0.0022 + j0.0022 = j0.0000$. Thus, the admittance equals $0.0010 + j0.0000$, a pure conductance.

Parallel Resonance

When a parallel *RLC* circuit lacks net susceptance at a certain frequency, we have a condition called *parallel resonance* at that frequency.

Problem 16-15

Suppose that we connect a resistor, a coil, and a capacitor in parallel. The resistor has a value of 100 Ω, the capacitance is 200 pF, and the inductance is 100 μH. We operate the circuit at a frequency of 1.00 MHz. What's the net complex admittance?

Solution

First, let's calculate the inductive susceptance. We recall the formula and plug in the numbers. Megahertz and microhenrys go together, so we have

$$\begin{aligned} jB_L &= -j[1/(6.2832fL)] \\ &= -j[1/(6.2832 \times 1.00 \times 100)] = -j0.00159155 \end{aligned}$$

Next, we calculate the capacitive susceptance. We can convert 200 pF to 0.000200 μF, so we have

$$\begin{aligned} jB_C &= j(6.2832fC) \\ &= j(6.2832 \times 1.00 \times 0.000200) = j0.00125664 \end{aligned}$$

We can consider the conductance, which equals $1/100 = 0.0100$ S, and the inductive susceptance together so that one of the parallel-connected admittances equals $0.0100 - j0.00159155$. The other admittance is $0 + j0.00125664$. When we add these complex numbers and then round off the susceptance coefficient to three significant figures, we get

$$\begin{aligned} Y &= 0.0100 - j0.00159155 + j0.00125664 \\ &= 0.0100 - j0.000335 \end{aligned}$$

Problem 16-16

Consider a resistor, a coil, and a capacitor in parallel. The resistance is 10.00 Ω, the inductance is 10.00 μH, and the capacitance is 1000 pF. The frequency is 1592 kHz. What's the complex admittance?

Solution

First, let's calculate the inductive susceptance. We can convert the frequency to megahertz; 1592 kHz = 1.592 MHz. Now we plug in the numbers to get

$$jB_L = -j[1/(6.2832 \times 1.592 \times 10.00)] = -j0.00999715$$

Next, we calculate the capacitive susceptance. We can convert 1000 pF to 0.001000 μF, so we obtain

$$jB_C = j(6.2832 \times 1.592 \times 0.001000) = j0.01000285$$

Finally, we consider the conductance, which equals 1/10.00 = 0.1000 S, and the inductive susceptance in a single component, so that one of the parallel-connected admittances is $0.1000 - j0.00999715$. The other is $0 + j0.01000285$. Adding these complex numbers and then rounding off the susceptance coefficient to four significant figures, we obtain

$$\begin{aligned} Y &= 0.1000 - j0.00999715 + j0.01000285 \\ &= 0.1000 - j0.000 \end{aligned}$$

Converting Complex Admittance to Complex Impedance

As we've seen, the *GB* half-plane looks like the *RX* half-plane, although mathematically they differ. We can convert a quantity from a complex admittance $G + jB$ to a complex impedance $R + jX$ using the formulas

$$R = G/(G^2 + B^2)$$

and

$$X = -B/(G^2 + B^2)$$

If we know the complex admittance, we should find the resistance and reactance components individually, using the above formulas. Then we can assemble the two components into the complex impedance $R + jX$.

Problem 16-17

Suppose that a circuit has an admittance $Y = 0.010 - j0.0050$. What's the complex impedance, assuming the frequency never varies?

Solution

In this case, $G = 0.010$ S and $B = -0.0050$ S. We determine $G^2 + B^2$ as follows:

$$\begin{aligned} G^2 + B^2 &= 0.010^2 + (-0.0050)^2 \\ &= 0.000100 + 0.000025 = 0.000125 \end{aligned}$$

Knowing this common denominator, we can calculate R and X as

$$R = G/0.000125 = 0.010/0.000125 = 80\ \Omega$$

and

$$X = -B/0.000125 = 0.0050/0.000125 = 40\ \Omega$$

Our circuit has a complex impedance of $Z = 80 + j40$.

Putting It All Together

When we encounter a parallel circuit containing resistance, inductance, and capacitance, and we want to determine the complex impedance of the combination, we should go through the following steps in order:

- Calculate the conductance G of the resistor.
- Calculate the susceptance B_L of the inductor.
- Calculate the susceptance B_C of the capacitor.
- Determine the net susceptance $B = B_L + B_C$.
- Determine the quantity $G^2 + B^2$.
- Compute R in terms of G and B using the appropriate formula.
- Compute X in terms of G and B using the appropriate formula.
- Write down the complex impedance as the sum $R + jX$.

Problem 16-18

Suppose that we connect a resistor of 10.0 Ω, a capacitor of 820 pF, and a coil of 10.0 μH in parallel. We operate the circuit at 1.00 MHz. What's the complex impedance?

Solution

Let's proceed according to the steps described above, leaving in some extra digits in the susceptance figures and then rounding off to three significant figures at the end of the process, as follows:

- Calculate $G = 1/R = 1/10.0 = 0.100$.
- Calculate $B_L = -1/(6.2832fL) = -1/(6.2832 \times 1.00 \times 10.0) = -0.0159155$.
- Calculate $B_C = 6.2832fC = 6.2832 \times 1.00 \times 0.000820 = 0.00515222$. (We must remember to convert the capacitance from picofarads to microfarads.)
- Calculate $B = B_L + B_C = -0.0159155 + 0.00515222 = -0.0107633$.
- Define $G^2 + B^2 = 0.100^2 + (-0.0107633)^2 = 0.0101158$.
- Calculate $R = G/0.0101158 = 0.100/0.0101158 = 9.89$.
- Calculate $X = -B/0.0101158 = 0.0107633/0.0101158 = 1.06$.
- The complex impedance equals $R + jX = 9.89 + j1.06$.

Problem 16-19

Suppose that we connect a resistor of 47.0 Ω, a capacitor of 500 pF, and a coil of 10.0 μH in parallel. What's their complex impedance at 2.25 MHz?

Solution

We proceed in the same fashion as we did when solving Problem 16-18, leaving in some extra digits in the conductance and susceptance figures until the end, as follows:

- Calculate $G = 1/R = 1/47.0 = 0.0212766$.
- Calculate $B_L = -1/(6.2832fL) = -1/(6.2832 \times 2.25 \times 10.0) = -0.00707354$.
- Calculate $B_C = 6.2832fC = 6.2832 \times 2.25 \times 0.000500 = 0.0070686$. (We convert the capacitance to microfarads.)
- Calculate $B = B_L + B_C = -0.00707354 + 0.0070686 = 0.00000$.

- Define $G^2 + B^2 = 0.0212766^2 + 0.00000^2 = 0.00045269$.
- Calculate $R = G/0.00045269 = 0.0212766/0.00045269 = 47.000$.
- Calculate $X = -B/0.00045269 = 0.00000/0.00045269 = 0.00000$.
- The complex impedance is $R + jX = 47.000 + j0.00000$. When we round both values off to three significant figures, we get $47.0 + j0.00$. This complex-number quantity represents a pure resistance equal to the value of the resistor in the circuit.

Reducing Complicated *RLC* Circuits

Sometimes we'll see circuits with several resistors, capacitors, and/or coils in series and parallel combinations. We can always reduce such a circuit to an equivalent series or parallel *RLC* circuit that contains one resistance, one capacitance, and one inductance.

Series Combinations

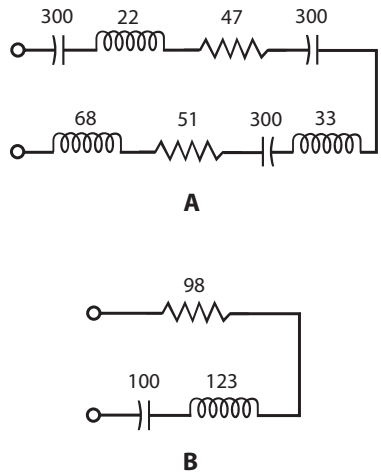
Resistances in series simply add. Inductances in series also add. Capacitances in series combine in a more complicated way, which you learned earlier. If you don't remember the formula, it is

$$C = 1/(1/C_1 + 1/C_2 + \cdots + 1/C_n)$$

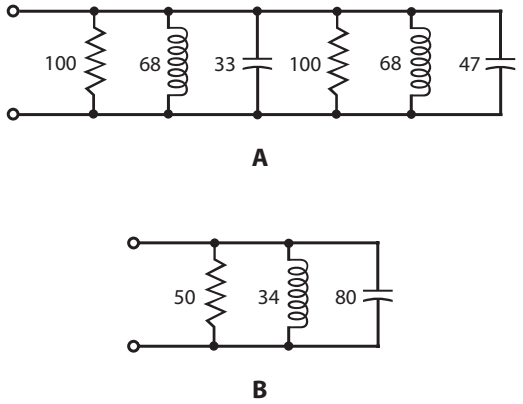
where C_1, C_2, …, and C_n represent the individual capacitances, and C represents the net capacitance of the series combination. Figure 16-8A shows an example of a "complicated" series *RLC* circuit. Figure 16-8B shows the equivalent circuit with one resistance, one capacitance, and one inductance.

Parallel Combinations

Resistances and inductances combine in parallel just as capacitances combine in series. Capacitances in parallel simply add up. Figure 16-9A shows an example of a "complicated" parallel *RLC* circuit. Figure 16-9B shows the equivalent circuit with one resistance, one capacitance, and one inductance.



16-8 At A, a "complicated" series circuit containing multiple resistances and reactances. At B, the same circuit simplified. Resistances are in ohms; inductances are in microhenrys (μH); capacitances are in picofarads (pF).



16-9 At A, a "complicated" parallel circuit containing multiple resistances and reactances. At B, the same circuit simplified. Resistances are in ohms; inductances are in microhenrys (μH); capacitances are in picofarads (pF).

"Nightmare" Circuits

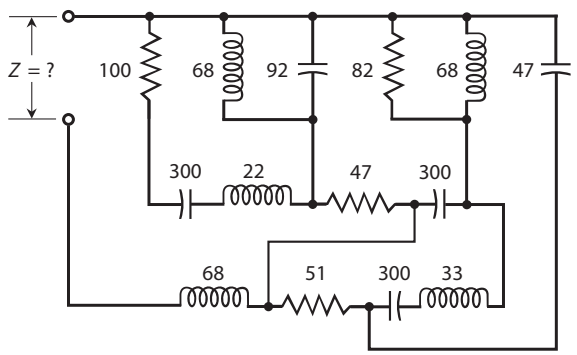
Imagine an *RLC* circuit like the one shown in Fig. 16-10. How would you find the net complex impedance at, say, 8.54 MHz? You'll rarely encounter circuits such as this one in practical applications—or if you do, no one will likely ask you to *calculate* the net impedance at any particular frequency. But you can rest assured that, given a frequency, a complex impedance does exist, no matter how complicated the circuit might be.

A true electronics "geek" could use a computer to work out the theoretical complex impedance of a circuit, such as the one in Fig. 16-10, at a specific frequency. In practice, however, an engineer might take an *experimental* approach by building the circuit, connecting a signal generator to its input terminals, and measuring the resistance R and the reactance X at the frequency of interest using a lab instrument called an *impedance bridge*.

Ohm's Law for Alternating Current

We can state Ohm's law for DC circuits as a simple relationship between three variables: the current I (in amperes), the voltage V (in volts), and the resistance R (in ohms). Here are the formulas, in case you don't recall them:

$$V = IR$$
$$I = V/R$$
$$R = V/I$$



16-10 A series-parallel "nightmare circuit" containing multiple resistances and reactances. Resistances are in ohms; inductances are in microhenrys (μH); capacitances are in picofarads (pF).

In AC circuits containing no reactance, these same formulas apply, as long as we work with root-mean-square (RMS) voltages and currents. If you need to refresh your memory concerning the meaning of RMS, refer back to Chap. 9.

Purely Resistive Impedances

When the impedance Z in an AC circuit contains no reactance, all of the current and voltage exist through and across a pure resistance R. In that case, we can express Ohm's law in terms of the three formulas

$$V = IZ$$
$$I = V/Z$$
$$Z = V/I$$

where $Z = R$, and the values I and V represent the RMS values for the current and voltage.

Complex Impedances

When you want to determine the relationship between current, voltage, and resistance in an AC circuit that contains reactance along with the resistance, things get quite interesting. Recall the formula for the square of the absolute-value impedance in a series *RLC* circuit:

$$Z^2 = R^2 + X^2$$

This equation tells us that

$$Z = (R^2 + X^2)^{1/2}$$

so Z equals the length of the vector $R + jX$ in the complex impedance plane. This formula applies only for series *RLC* circuits.

The square of the absolute-value impedance for a parallel *RLC* circuit, in which the resistance equals R and the reactance equals X, is defined as

$$Z^2 = R^2X^2/(R^2 + X^2)$$

We can calculate Z, the absolute-value impedance, directly as

$$Z = [R^2X^2/(R^2 + X^2)]^{1/2}$$

The ½ power of a quantity represents the positive square root of that quantity.

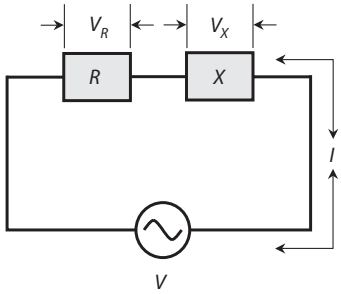
Problem 16-20

Imagine that a series *RX* circuit (shown by the generic block diagram of Fig. 16-11) has a resistance of $R = 50.0\ \Omega$ and a capacitive reactance of $X = -50.0\ \Omega$. If we apply 100 V RMS AC to this circuit, what's the RMS current?

Solution

First, let's calculate the complex impedance using the above formula for series circuits, going to a few extra digits to prevent cumulative rounding errors in later calculations. We have

$$\begin{aligned} Z &= (R^2 + X^2)^{1/2} = [50.0^2 + (-50.0)^2]^{1/2} \\ &= (2500 + 2500)^{1/2} = 5000^{1/2} = 70.71068\ \Omega \end{aligned}$$



16-11 A series circuit containing resistance and reactance. Illustration for Problems 16-20 through 16-23.

which we can round off to 70.7 Ω. Now we can calculate the current, using the original "un-rounded" figure for Z, as

$$I = V/Z = 100 / 70.71068 = 1.414214 \text{ A RMS}$$

which we can round off to 1.41 A RMS.

Problem 16-21

What are the RMS AC voltages across the resistance and the reactance, respectively, in the circuit described in Problem 16-20?

Solution

The Ohm's law formulas for DC will work here. We determined the current, going to several extra digits, as $I = 1.414214$ A RMS, so the voltage drop V_R across the resistance is

$$V_R = IR = 1.414214 \times 50.0 = 70.7107 \text{ V RMS}$$

which rounds off to 70.7 V RMS. The voltage drop V_X across the reactance is

$$V_X = IX = 1.41 \times (-50.0) = -70.7107 \text{ V RMS}$$

which rounds off to −70.7 V RMS. Do you see why we included all the extra digits in our intermediate calculations? If we'd used the rounded figure for the current (1.41 A RMS) in the foregoing calculations, we'd have gotten 70.5 and −70.5 V RMS for the answers we just found. That provides a great example of how cumulative rounding errors can mislead us if we're not careful!

Signs and Phase with RMS Values

Here's an important note in regards to the signs (plus or minus) in RMS figures. When we deal with RMS values, minus signs have no meaning. An RMS value can never be negative, so the minus sign in this result constitutes a mathematical artifact. We can consider the voltage across the reactance as 70.7 V RMS, equal in magnitude to the voltage across the resistance. *But the phase is different*, and that's what the minus sign tells us!

Note that the voltages across the resistance and the reactance (a capacitive reactance in the above-described case because it's negative) don't add up to 100 V RMS, which appears across the whole circuit. This phenomenon occurs because, in an AC circuit containing resistance and reactance, we always observe a difference in phase between the voltage across the resistance and the voltage across the reactance. The voltages across the components add up to the applied voltage *vectorially*, but not always *arithmetically*.

Problem 16-22

Suppose that a series *RX* circuit (Fig. 16-11) has $R = 10.0\ \Omega$ and $X = 40.0\ \Omega$. The applied voltage is 100 V RMS AC. What's the current?

Solution

First, we calculate the complex impedance using the above formula for series circuits, obtaining

$$\begin{aligned} Z &= (R^2 + X^2)^{1/2} = [10.0^2 + (40.0)^2]^{1/2} \\ &= (100 + 1600)^{1/2} = 1700^{1/2} = 41.23106\ \Omega \end{aligned}$$

When we use Ohm's law to calculate the current, we get

$$I = V/Z = 100/41.23106 = 2.425356 \text{ A RMS}$$

which rounds off to 2.43 A RMS.

Problem 16-23

What are the RMS AC voltages across the resistance and the reactance, respectively, in the circuit described in Problem 16-22?

Solution

Knowing the current (and using our "un-rounded" result), we can calculate the voltage across the resistance as

$$V_R = IR = 2.425356 \times 10.0 = 24.25356 \text{ V RMS}$$

which rounds off to 24.3 V RMS. The voltage across the reactance is

$$V_X = IX = 2.425356 \times 40.0 = 97.01424 \text{ V RMS}$$

which rounds off to 97.0 V RMS. If we take the arithmetic sum $V_R + V_X$, we get $24.25356 + 97.01424 = 121.2678$ V RMS, which rounds off to 121.3 V RMS, as the total voltage across *R* and *X*. Again, this value comes out different from the actual applied voltage. The simple DC rule does not work here, for the same reason it didn't work in the scenario of Problem 16-21. The AC voltage across the resistance doesn't add up arithmetically with the AC voltage across the reactance because the two AC waves differ in phase.

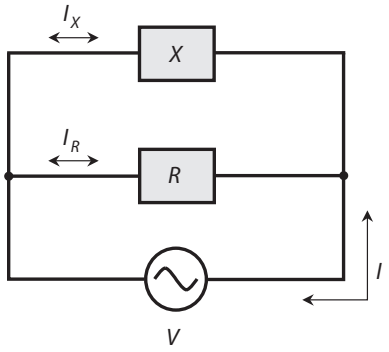
Problem 16-24

Suppose that a parallel *RX* circuit (shown by the generic block diagram of Fig. 16-12) has $R = 30.0\ \Omega$ and $X = -20.0\ \Omega$. We supply the circuit with $V = 50.0$ V RMS. What's the total current drawn from the AC supply?

Solution

First, we find the absolute-value impedance, remembering the formula for parallel circuits and going to a few extra digits to avoid cumulative rounding errors. We get

$$\begin{aligned} Z &= [R^2X^2/(R^2 + X^2)]^{1/2} \\ &= \{30.0^2 \times (-20.0)^2/[30.0^2 + (-20.0)^2]\}^{1/2} \\ &= [900 \times 400/(900 + 400)]^{1/2} = (360{,}000/1300)^{1/2} \\ &= 277^{1/2} = 16.64332\ \Omega \end{aligned}$$



16-12 A parallel circuit containing resistance and reactance. Illustration for Problems 16-24 and 16-25.

The total current is therefore

$$I = V/Z = 50 / 16.64332 = 3.004208 \text{ A RMS}$$

which rounds off to 3.00 A RMS.

Problem 16-25

What are the RMS currents through the resistance and the reactance, respectively, in the circuit described in Problem 16-24?

Solution

The Ohm's law formulas for DC will work here. We can calculate the current through the resistance as

$$I_R = V/R = 50.0 / 30.0 = 1.67 \text{ A RMS}$$

For the current through the reactance, we calculate

$$I_X = V/X = 50.0 / (-20.0) = -2.5 \text{ A RMS}$$

As before, we can neglect the minus sign when we think in terms of RMS, so we can call this current 2.5 A RMS. Note that if we directly add the current across the resistance to the current across the reactance, we don't get 3.00 A, the actual total current. This effect takes place for the same reason that AC voltages don't add arithmetically in AC circuits that contain resistance and reactance. The constituent currents, I_R and I_X, differ in phase. Vectorially, they add up to 3.00 A RMS, but arithmetically, they don't.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

17
CHAPTER

Alternating-Current Power and Resonance

WHEN WE WANT TO OPTIMIZE HOW POWER "TRAVELS," OR CHANGES FORM, WE FACE A CHALLENGE. A phenomenon called *resonance* can play an important role in efficient power transfer and conversion, especially at high AC frequencies.

Forms of Power

Scientists and engineers define power as the rate at which energy is expended, radiated, or dissipated. This definition applies to mechanical motion, chemical effects, electricity, sound waves, radio waves, heat, infrared (IR), visible light, ultraviolet (UV), X rays, gamma rays, and high-speed subatomic particles.

Units of Power

The standard unit of power is the *watt*, abbreviated W and equivalent to a *joule per second* (J/s). Sometimes, engineers specify power in *kilowatts* (kW or thousands of watts), *megawatts* (MW or millions of watts), *gigawatts* (GW or billions of watts), or *terawatts* (TW or trillions of watts). In numerical terms,

- 1 kW = 1000 W
- 1 MW = 1,000,000 W
- 1 GW = 1,000,000,000 W
- 1 TW = 1,000,000,000,000 W

We can also express power in *milliwatts* (mW or thousandths of watts), *microwatts* (μW or millionths of watts), *nanowatts* (nW or billionths of watts), or *picowatts* (pW or trillionths of watts). In numerical terms,

- 1 mW = 0.001 W
- 1 μW = 0.000001 W
- 1 nW = 0.000000001 W
- 1 pW = 0.000000000001 W

Volt-Amperes

In DC circuits, and also in AC circuits having no reactance, we can define power as the product of the voltage V across a device and the current I through that device; that is,

$$P = VI$$

If we express V in volts and I in amperes, then P comes out in *volt-amperes* (VA). Volt-amperes translate directly into watts when no reactance exists in a circuit (Fig. 17-1).

Volt-amperes, also called *VA power* or *apparent power*, can take various forms. A resistor converts electrical energy into heat energy at a rate that depends on the value of the resistance and the current through it. A light bulb converts electricity into light and heat. A radio antenna converts high-frequency AC into radio waves. A loudspeaker or headset converts low-frequency AC into sound waves. A microphone converts sound waves into low-frequency AC. Power figures in these forms give us measures of the intensity of the heat, light, radio waves, sound waves, or AC electricity.

Instantaneous Power

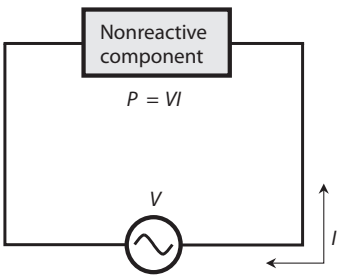
Engineers usually think of electrical power in RMS (see Ch. 9) terms. But for VA power, peak values are sometimes used instead. If the AC constitutes a perfect sine wave with no DC component, the peak current (in either direction) equals 1.414 times the RMS current, and the peak voltage (of either polarity) equals 1.414 times the RMS voltage. If the current and the voltage follow along exactly in phase, then the product of their peak values equals twice the product of their RMS values.

In a reactance-free, sine-wave AC circuit, we see instants in time when the VA power equals twice the effective power. At other points in time, the VA power equals zero; at still other moments, the VA power falls somewhere between zero and twice the effective power level (Fig. 17-2). We call this constantly changing power level, measured or expressed at any particular point in time, the *instantaneous power*. In some situations, such as with an amplitude-modulated (AM) radio signal, the instantaneous power varies in a complicated fashion.

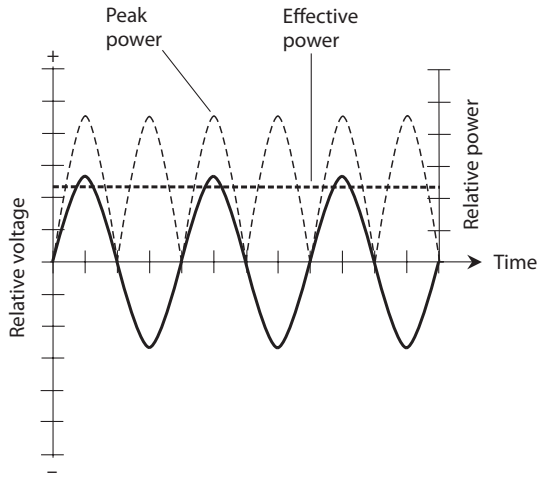
Reactive or Imaginary Power

In a pure resistance, the rate of energy expenditure per unit of time (or *true power*) equals the VA power (also known as *apparent power*). But when reactance exists, the VA power exceeds the power manifested as heat, light, radio waves, or whatever. The apparent power is greater than the true power. We call the difference *reactive power* or *imaginary power* because it exists only within the reactive part of the circuit, represented by the imaginary-number part of the complex-number impedance.

Inductors and capacitors store energy and release it a fraction of a cycle later, over and over, for as long as AC flows. This phenomenon, like true power, manifests itself as the rate at which energy



17-1 When an AC component contains no reactance, the power P is the product of the voltage V across the component and the current I through the component.



17-2 Peak versus effective power for a sine wave. The left-hand vertical scale shows relative voltage. The right-hand vertical scale shows relative power. The solid curve represents the voltage as a function of time. The light and heavy dashed waves show peak and effective power, respectively, as functions of time.

changes from one form to another. But rather than existing in a form that we can employ in some practical way, imaginary power can only go in and out of "storage." We can't use it for anything. The storage/release cycle repeats along with the actual AC cycle.

True Power Doesn't Travel

If you connect a radio transmitter to a cable that runs to an antenna, you might say you're "feeding power" through the cable to the antenna. Even experienced radio-frequency (RF) engineers and technicians sometimes say that. However, true power always involves a *change in form*, such as from electrical current and voltage into radio waves, or from sound waves into heat. Power does not actually *travel* from place to place. It simply *happens* in one spot or another.

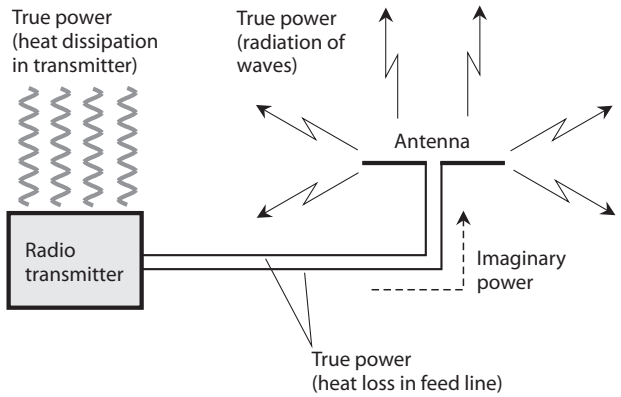
In a radio antenna system, some true power dissipates in a wasteful manner, ending up as heat in the transmitter amplifiers and in the transmission (feed) line, as shown in Fig. 17-3. Obviously, a competent RF engineer seeks to minimize the extent of this waste. The useful dissipation of true power occurs when the imaginary power, in the form of electric and magnetic fields, gets to the antenna, where it changes form and emerges from the antenna as *electromagnetic (EM) waves*.

If you do much work with wireless transmitting antenna systems, you'll hear or read expressions, such as "forward power" and "reflected power," or "power goes from this amplifier to these speakers." You can talk or write in such terms if you like, but keep in mind that the notion can sometimes lead to mistaken conclusions. For example, you might get the idea that an antenna system works more or less efficiently than it actually does.

Reactance Consumes No Power

A pure inductance or a pure capacitance can't dissipate any power. A pure reactance can only store energy and then give it back to the circuit a fraction of a cycle later. In real life, the dielectrics or wires in coils and capacitors dissipate some power as heat, but ideal components wouldn't do that. A capacitor, as we've learned, stores energy as an electric field. An inductor stores energy as a magnetic field.

A reactive component causes AC to shift in phase, so that the current does not follow in step with the voltage, as it would in a reactance-free circuit. In a circuit containing inductive reactance,



17-3 True power and imaginary power in a radio transmitter and antenna system.

the current lags the voltage by up to 90°, or one-quarter of a cycle. In a circuit with capacitive reactance, the current leads the voltage by up to 90°.

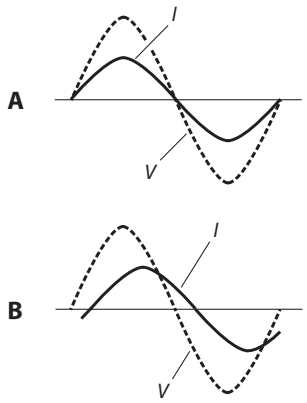
In a purely resistive circuit, the voltage and current precisely follow each other so that they combine in the most efficient possible way (Fig. 17-4A). But in a circuit containing reactance, the voltage and current do not follow exactly along with each other (Fig. 17-4B) because of the phase difference. In that case, the product of the voltage and current (the VA or apparent power) exceeds the actual energy expenditure (the true power).

Power Parameters

In an AC circuit containing nonzero resistance and nonzero reactance, we can summarize the relationship between true power P_T, apparent (VA) power P_{VA}, and imaginary (reactive) power P_X in terms of the formula

$$P_{VA} = (P_T^2 + P_X^2)^{1/2}$$

where $P_T < P_{VA}$ and $P_X < P_{VA}$. If no reactance exists, then $P_{VA} = P_T$ and $P_X = 0$. Engineers strive to minimize, and if possible eliminate, the reactance in power-transmission systems.



17-4 At A, the current *I* and voltage *V* are in phase in a nonreactive AC circuit. At B, *I* and *V* are not in phase when reactance exists.

Power Factor

In an AC circuit, we call the ratio of the true power to the VA power, P_T/P_{VA}, the *power factor* (*PF*). In the ideal case in which we have no reactance, $P_T = P_{VA}$, so $PF = 1$. If the circuit contains reactance but lacks resistance or conductance (zero or infinite resistance), then $P_T = 0$, so $PF = 0$.

When a *load*, or a circuit in which we want power to dissipate or change form, contains both resistance and reactance, then *PF* falls between 0 and 1. We can also express the power factor as a percentage $PF_{\%}$ between 0 and 100. If we know P_T and P_{VA}, then we can calculate *PF* as

$$PF = P_T/P_{VA}$$

and $PF_{\%}$ as

$$PF_{\%} = 100P_T/P_{VA}$$

When a load has nonzero, finite resistance and nonzero, finite reactance, then some of the power dissipates as true power, and some of the power gets "rejected" by the load as imaginary power.

We can determine the power factor in an AC circuit that contains reactance and resistance in two ways:

1. Find the cosine of the phase angle
2. Calculate the ratio of the resistance to the absolute-value impedance

Cosine of Phase Angle

Recall that in a circuit having reactance and resistance, the current and the voltage do not follow along exactly in phase. The phase angle (ϕ) constitutes the extent, expressed in degrees, to which the current and the voltage differ in phase. In a pure resistance, $\phi = 0°$. In a pure reactance, $\phi = +90°$ (if the net reactance is inductive) or $\phi = -90°$ (if the net reactance is capacitive). We can calculate the power factor as

$$PF = \cos \phi$$

and

$$PF_{\%} = 100 \cos \phi$$

Problem 17-1

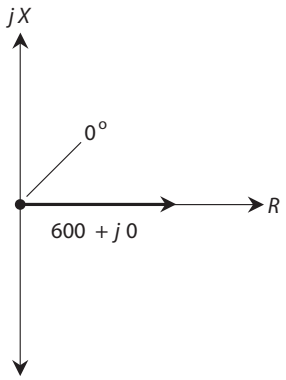
Suppose that a circuit comprises a pure resistance of 600 Ω with no reactance whatsoever. What's the power factor?

Solution

Without doing any calculations, you can sense that $PF = 1$ because $P_{VA} = P_T$ in a pure resistance. It follows that $P_T/P_{VA} = 1$. But you can also look at this situation by noting that the phase angle equals 0° because the current follows in phase with the voltage. Using a calculator, you'll find that $\cos 0° = 1$. Therefore, $PF = 1 = 100\%$. Figure 17-5 illustrates the *RX* half-plane vector for this situation. Remember that you should express the phase angle with respect to the *R* axis.

Problem 17-2

Suppose that a circuit contains a pure capacitive reactance of −40 Ω, but no resistance. What's the power factor?



17-5 Vector diagram showing the phase angle for a purely resistive impedance of $600 + j0$. The R and jX scales are relative.

Solution

Here, the phase angle equals $-90°$, as shown in the *RX* half-plane vector diagram of Fig. 17-6. A calculator will tell you that $\cos -90° = 0$. Therefore, $PF = 0$, and $P_T/P_{VA} = 0 = 0\%$. None of the power is true; it's all reactive.

Problem 17-3

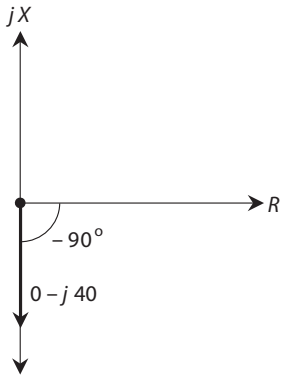
Consider a circuit that contains a resistance of 50 Ω and an inductive reactance of 50 Ω, connected in series. What's the power factor?

Solution

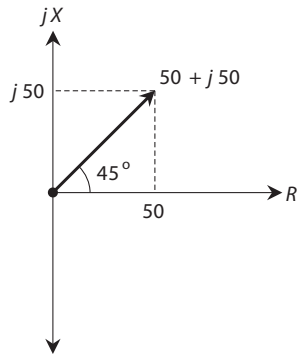
The phase angle equals 45° (Fig. 17-7). The resistance and the reactance vectors have equal lengths, forming two sides of a right triangle. The complex impedance vector constitutes the *hypotenuse* (longest side) of the right triangle. To determine the power factor, you can use a calculator to find $\cos 45° = 0.707$, so you know that $P_T/P_{VA} = 0.707 = 70.7\%$.

The Ratio *R*/*Z*

We can calculate the power factor in an *RX* circuit by finding the ratio of the resistance R to the absolute-value impedance Z. Figure 17-7 provides an example. A right triangle is formed by the



17-6 Vector diagram showing the phase angle for a purely capacitive impedance of $0 - j40$. The R and jX scales are relative.



17-7 Vector diagram showing the phase angle for a complex impedance of $50 + j50$. The R and jX scales are relative.

resistance vector R (the base), the reactance vector jX (the height), and the absolute-value impedance Z (the hypotenuse). The cosine of the phase angle equals the ratio of the base length to the hypotenuse length, or R/Z.

Problem 17-4

Suppose that a series circuit has an absolute-value impedance Z of 100 Ω, with a resistance R of 80 Ω. What's the power factor?

Solution

You can set up and calculate the ratio

$$PF = R/Z = 80/100 = 0.8 = 80\%$$

It doesn't matter whether the net reactance in this circuit happens to be capacitive or inductive.

Problem 17-5

Consider a series circuit with an absolute-value impedance of 50 Ω, purely resistive. What's the power factor?

Solution

Here, $R = Z = 50$ Ω. Therefore

$$PF = R/Z = 50/50 = 1 = 100\%$$

Problem 17-6

Consider a circuit with a resistance of 50 Ω and a capacitive reactance of −30 Ω. What's the power factor? Use the cosine method.

Solution

Remember the formula for phase angle in terms of reactance and resistance:

$$\phi = \text{Arctan } (X/R)$$

where X represents the reactance and R represents the resistance. Therefore

$$\phi = \text{Arctan } (-30/50) = \text{Arctan } (-0.60) = -31°$$

The power factor equals the cosine of this angle, so

$$PF = \cos (-31°) = 0.86 = 86\%$$

Problem 17-7

Consider a series circuit with a resistance of 30 Ω and an inductive reactance of 40 Ω. What's the power factor? Use the R/Z method.

Solution

First, find the absolute-value impedance using the formula for series circuits:

$$Z = (R^2 + X^2)^{1/2}$$

where R represents the resistance and X represents the net reactance. Plugging in the numbers, you get

$$Z = (30^2 + 40^2)^{1/2} = (900 + 1600)^{1/2} = 2500^{1/2} = 50\ \Omega$$

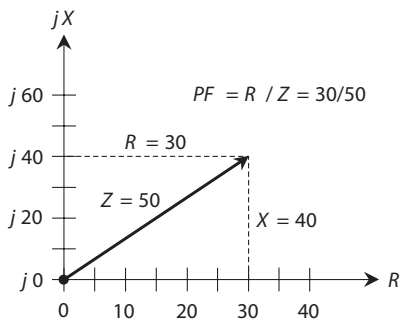
Now you can calculate the power factor as

$$PF = R/Z = 30/50 = 0.60 = 60\%$$

You can graph this situation as a 30:40:50 right triangle (Fig. 17-8).

How Much of the Power Is True?

The above formulas allow you to figure out, given the resistance, reactance, and VA power, how many watts constitute true, or real power, and how many watts constitute reactive, or imaginary power. Engineers must consider this situation when working with RF equipment because some RF wattmeters display VA power rather than true power. When reactance exists along with the resistance in a circuit or system, such wattmeters give artificially high readings.



17-8 Illustration for Problem 17-7. The vertical and horizontal scale increments differ; this is a common practice in graphs.

Problem 17-8

Suppose that a circuit has 50 Ω of resistance and 30 Ω of inductive reactance in series. A wattmeter shows 100 W, representing the VA power. What's the true power?

Solution

We must determine the power factor to figure out the answer to this question. First, we calculate the phase angle as

$$\phi = \text{Arctan}\ (X/R) = \text{Arctan}\ (30/50) = 31°$$

The power factor equals the cosine of the phase angle, so

$$PF = \cos 31° = 0.86$$

The formula for the power factor in terms of true and VA power is

$$PF = P_T/P_{VA}$$

We can rearrange this formula to solve for true power, obtaining

$$P_T = PF \times P_{VA}$$

When we plug in $PF = 0.86$ and $P_{VA} = 100$, we get

$$P_T = 0.86 \times 100 = 86 \text{ W}$$

Problem 17-9

Suppose that a circuit has a resistance of 1000 Ω in *parallel* with a capacitance of 1000 pF. We operate the circuit at a frequency of 100 kHz. If a wattmeter designed to read VA power shows 88.0 W, what's the true power?

Solution

This problem is rather complicated because the components appear in parallel. To begin, let's make sure that we have the units in agreement so the formulas work right. We can convert the frequency f to megahertz, getting $f = 0.100$ MHz. We can convert the capacitance to microfarads, getting $C = 0.001000$ μF. From the previous chapter, we remember the formula for capacitive susceptance, and calculate it for this situation as

$$\begin{aligned} B_C &= 6.2832fC \\ &= 6.2832 \times 0.100 \times 0.001000 = 0.00062832 \text{ S} \end{aligned}$$

The conductance of the resistor, G, equals the reciprocal of the resistance, R, so

$$G = 1/R = 1/1000 = 0.001000 \text{ S}$$

Now let's use the formulas for calculating resistance and reactance in terms of conductance and susceptance in parallel circuits. First, we find the resistance as

$$\begin{aligned} R &= G/(G^2 + B^2) \\ &= 0.001000/(0.001000^2 + 0.00062832^2) \\ &= 0.001000/0.0000013948 = 716.95\ \Omega \end{aligned}$$

Next, we find the reactance as

$$\begin{aligned} X &= -B/(G^2 + B^2) \\ &= -0.00062832/0.0000013948 = -450.47\ \Omega \end{aligned}$$

again rounded to four significant figures. Now we can calculate the phase angle as

$$\begin{aligned} \phi &= \text{Arctan}\ (X/R) \\ &= \text{Arctan}\ (-450.47/716.95) = \text{Arctan}\ (-0.62831) = -32.142^\circ \end{aligned}$$

We calculate the power factor as

$$PF = \cos \phi = \cos\ (-32.142^\circ) = 0.84673$$

The VA power P_{VA} is given as 88.0 W. Therefore, we can calculate the true power, rounding off to three significant figures (because that's the extent of the accuracy of our input data), getting

$$\begin{aligned} P_T &= PF \times P_{VA} = 0.84673 \times 88.0 \\ &= 74.5\ \text{W} \end{aligned}$$

Power Transmission

Consider a radio broadcast or communications station. The transmitter produces high-frequency AC. We want to get the signal efficiently to an antenna located some distance from the transmitter. This process involves the use of an *RF transmission line*, also known as a *feed line*. The most common type is coaxial cable. Alternatively, we can use two-wire line, also called parallel-wire line, in some antenna systems. At ultra-high and microwave frequencies, another kind of transmission line, known as a *waveguide*, is often employed.

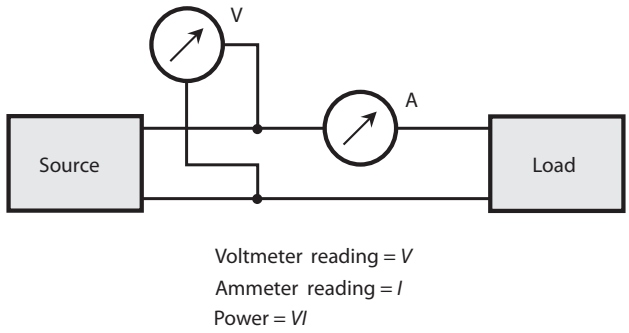
Loss: the Less, the Better!

The overriding challenge in the design and construction of any *power transmission* system lies in minimizing the loss. Power wastage occurs almost entirely as heat in the transmission line conductors and dielectric, and in objects near the line. Some loss can take the form of unwanted EM radiation from the line. Loss also occurs in transformers. Power loss in an electrical system is analogous to the loss of usable work produced by friction in a mechanical system.

In an ideal power-transmission system, all of the power constitutes VA power; that is, all of the power occurs as AC in the conductors and an alternating voltage between them. We don't want power in a transmission line or transformer to exist in the form of true power because that situation translates into heat loss, radiation loss, or both. True power dissipation or radiation should always take place in the load, usually at the opposite end of the transmission line from the source.

Power Measurement in a Transmission Line

In an AC transmission line, we can measure power by placing an AC voltmeter between the conductors and an AC ammeter in series with one of the conductors, as shown in Fig. 17-9. We should place both meters at the same point on the line (even though they don't look that way in this diagram!). In that case the power P (in watts) equals the product of the RMS voltage V (in volts) and the RMS current I (in amperes). We can use this technique in any transmission line at any AC



17-9 Power measurement in a transmission line. Ideally, we should measure the voltage and the current at the same physical point on the line.

frequency, from the 60 Hz of a utility system to many gigahertz in some wireless communications systems. However, when we measure power this way, we don't necessarily get an accurate indication of the true power dissipated by the load at the end of the line.

Recall that any transmission line has an inherent *characteristic impedance*. The value of this parameter Z depends on the diameters of the line conductors, the spacing between the conductors, and the type of dielectric material that separates the conductors. If the load constitutes a pure resistance R containing no reactance, and if $R = Z_o$, then the power indicated by the voltmeter/ammeter scheme will equal the true power dissipated by the load—provided that we place the voltmeter and ammeter directly across the load, at the end of the line opposite the source.

If the load constitutes a pure resistance that differs from the characteristic impedance of the line (that is, $R \neq Z_o$), then the voltmeter and ammeter will not give an indication of the true power. Also, if the load contains any reactance along with the resistance, the voltmeter/ammeter method will fail to give us an accurate reading of the true power, no matter what the resistance.

Impedance Mismatch

If we want a power transmission system to perform at its best, then the load impedance must constitute a pure resistance equal to the characteristic impedance of the line. When we don't have this ideal state of affairs, the system has an *impedance mismatch*.

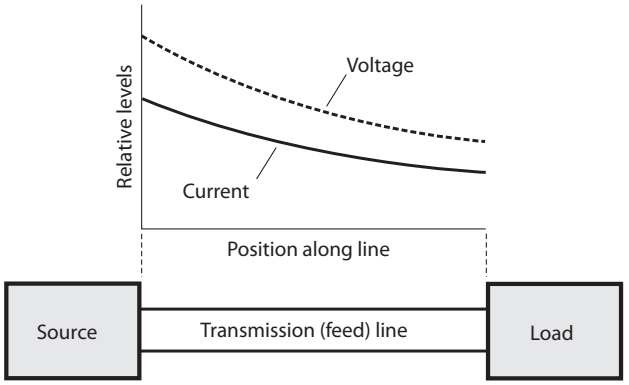
Small impedance mismatches can usually—but not always—be tolerated in power transmission systems. In very-high-frequency (VHF), ultra-high-frequency (UHF), and microwave wireless transmitting systems, even a small impedance mismatch between the load (antenna) and the line can cause excessive power losses in the line.

We can usually get rid of an impedance mismatch by installing a *matching transformer* between the transmission line and the load. We can also correct impedance mismatches in some situations by deliberately placing a reactive component (inductor or capacitor) in series or parallel with the load to cancel out any existing load reactance.

Loss in a Mismatched Line

When we terminate a transmission line in a pure resistance R equal to the characteristic impedance Z_o of the line, then the RMS current I and the RMS voltage V remain constant all along the line, provided that the line has no ohmic loss and no dielectric loss. In such a situation, we have

$$R = Z_o = V/I$$



17-10 Along a matched transmission line, the voltage-to-current ratio V/I holds constant everywhere, although the actual values of V and I decrease with increasing distance from the source.

where we express R and Z_o in ohms, V in volts RMS, and I in amperes RMS. Of course, no transmission line is perfectly *lossless*. In a "real-world" transmission line, the current and voltage gradually decrease as a signal makes its way from the source to the load. Nevertheless, if the load constitutes a pure resistance equal to the characteristic impedance of the line, the current and voltage remain *in the same ratio* at all points along the line, as shown in Fig. 17-10.

Standing Waves

If a transmission line and its load aren't perfectly matched, then the current and voltage alternately rise and fall as they move along the line. We call the maxima and minima *loops* and *nodes*, respectively. At a maximum-current point (a *current loop*), the voltage reaches its minimum (a *voltage node*). Conversely, at a maximum-voltage point (a *voltage loop*), the current attains its minimum (a *current node*).

If we graph the current and voltage loops and nodes along a mismatched transmission line as functions of the position on the line, we see wavelike patterns that remain fixed over time. These orderly patterns of current and voltage don't move in either direction along the line; they simply "stand there." For this reason, engineers call them *standing waves*.

Losses Caused by Standing Waves

When a transmission line contains standing waves, we observe a corresponding pattern in the extent of the line loss, as follows:

- At current loops, the loss in the line conductors reaches a maximum.
- At current nodes, the loss in the line conductors reaches a minimum.
- At voltage loops, the loss in the line dielectric reaches a maximum.
- At voltage nodes, the loss in the line dielectric reaches a minimum.

It's tempting to suppose that all of the loss variations ought to average out in a mismatched line, so that the excess loss in some places gets "paid back" in the form of reduced loss in other places. But things don't work that way. Overall, the losses in a mismatched line always exceed the losses in a perfectly matched line. The extra loss increases as the mismatch gets worse. We call it *transmission-line mismatch loss* or *standing-wave loss*; it occurs as heat dissipation, so it constitutes true power. Any true power that heats up a transmission line goes to waste because it never reaches the load. We want the true power to end up in the load—if not all of it, then as much of it as possible.

As the extent of the mismatch in a power-transmission system increases, so does the loss caused by the current and voltage loops in the standing waves. The more loss a line has when perfectly matched, the more loss a given amount of mismatch will cause. Standing-wave loss also increases as the frequency increases, if we hold all other factors constant.

Line Overheating

A severe mismatch between the load and the transmission line can cause another problem besides lost power: physical damage to, or destruction of, the line.

Suppose that a certain transmission line can effectively function with a radio transmitter that generates up to 1 kW of power, assuming that the line and the load are perfectly matched. If a severe mismatch exists and you try to feed 1 kW into the same line, the extra current at the current loops can heat the conductors to the point at which the dielectric material melts and the line shorts out. It's also possible for the voltage at the voltage loops to cause arcing between the line conductors. This arcing can perforate and/or burn the dielectric, ruining the line.

When we have no choice but to operate an RF transmission line with a significant impedance mismatch, we must refer to *derating functions* in order to determine how much power the line can safely handle. Manufacturers of prefabricated lines, such as coaxial cable, can (or should) provide this information.

Resonance

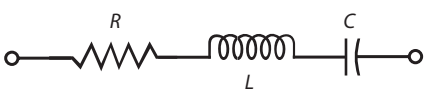
We observe resonance in an AC circuit when capacitive and inductive reactance both exist, but they have equal and opposite values so that they cancel each other out. We saw a couple of examples of this phenomenon in Chap. 16. Let's explore it in more detail.

Series Resonance

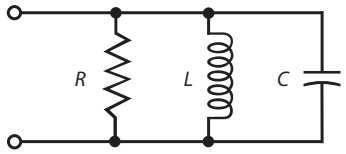
Capacitive reactance X_C and inductive reactance X_L can have equal magnitudes although they produce opposite effects. In any circuit containing some inductance and some capacitance, there exists a frequency at which $X_L = -X_C$. This condition constitutes resonance, and we symbolize the frequency at which it occurs by writing f_o. In a simple *LC* circuit, we observe only one such frequency; in some circuits involving transmission lines or antennas, we observe many such frequencies. In that case, we call the lowest frequency at which resonance occurs the *fundamental resonant frequency*, symbolized f_o.

You'll recognize Fig. 17-11 as a schematic diagram of a series *RLC* circuit. If we apply a variable-frequency AC signal to the end terminals, we'll find that at one particular "critical" frequency, $X_L = -X_C$. This phenomenon will always occur if *L* and *C* are both finite and nonzero. The "critical" frequency represents f_o for the circuit. At f_o, the effects of capacitive reactance and inductive reactance cancel out so that the circuit appears as a pure resistance, with a value theoretically equal to *R*. We call this condition *series resonance*.

If $R = 0$ (a short circuit), then Fig. 17-11 represents a *series LC circuit*, and the impedance at resonance theoretically equals $0 + j0$. Under these conditions, the circuit offers no opposition to the



17-11 A series *RLC* circuit.



17-12 A parallel *RLC* circuit.

flow of alternating current at f_o. Of course, "real-world" series *LC* circuits always contain at least a small amount of resistance so that a little loss occurs in the coil and capacitor; the real-number part of the complex impedance doesn't *exactly* equal 0. If the coil and capacitor have high quality and exhibit minimal loss, however, we can usually say that $R = 0$ "for all intents and purposes."

Parallel Resonance

Figure 17-12 shows a generic parallel *RLC* circuit. In this situation, we can think of the resistance R as a conductance G, with $G = 1/R$. For that reason, some people might say that we should call this arrangement a parallel *GLC* circuit. We can get away with either term as long as we include the word "parallel"!

At one particular frequency f_o, the inductive susceptance B_L will exactly cancel the capacitive susceptance B_C, so we have $B_L = -B_C$. This condition always occurs for some applied AC signal frequency f_o, as long as the circuit contains finite, nonzero inductance and finite, nonzero capacitance. At f_o, the susceptances cancel each other out, leaving theoretically zero susceptance. The admittance through the circuit theoretically equals the conductance G of the resistor. We call this condition *parallel resonance*.

If the circuit contains no resistor, but only a coil and capacitor, we have a *parallel LC circuit*, and the admittance at resonance theoretically equals $0 + j0$. The circuit will offer a lot of opposition to alternating current at f_o, and the complex impedance will, in an informal sense, equal "infinity" (∞). In a practical "real-world" *LC* circuit, the coil and capacitor always have a little bit of loss, so the real-number part of the complex impedance isn't really infinite. However, if we use low-loss components, we can get real-number coefficients of many megohms or even gigohms for parallel *LC* circuits at resonance, so that we can say that $R = \infty$ "for all intents and purposes."

Calculating the Resonant Frequency

We can calculate the resonant frequency f_o, of a series *RLC* or parallel *RLC* circuit in terms of the inductance L in henrys and the capacitance C in farads, using the formula

$$f_o = 1 / [2\pi(LC)^{1/2}]$$

Considering $\pi = 3.1416$, we can simplify this formula to

$$f_o = 0.15915 / (LC)^{1/2}$$

You might remember from your basic algebra or precalculus course that the ½ power of a quantity represents the positive square root of that quantity. The foregoing formula will also work if you want to find f_o in megahertz (MHz) when you know L in microhenrys (μH) and C in microfarads (μF).

The Effects of *R* and *G*

Interestingly, the value of R or G does not affect the value of f_o in series or parallel *RLC* circuits. However, the presence of nonzero resistance in a series resonant circuit, or nonzero conductance in a

parallel resonant circuit, makes f_o less well-defined than it would be if the resistance or conductance did not exist. If we short out R in the circuit of Fig. 17-11 or remove R from the circuit of Fig. 17-12, we have LC circuits that exhibit the most well-defined possible *resonant responses*.

In a series RLC circuit, the resonant frequency response becomes more "broad" as the resistance increases, and more "sharp" as the resistance decreases. In a parallel RLC circuit, the resonant frequency response becomes more "broad" as the *conductance* increases (R gets smaller), and more "sharp" as the conductance decreases (R gets larger). In theory, the "sharpest" possible responses occur when $R = 0$ in a series circuit, and when $G = 0$ (that is, $R = \infty$) in a parallel circuit.

Problem 17-10

Find the resonant frequency of a series circuit with a 100-μH inductor and a 100-pF capacitor.

Solution

We should convert the capacitance to 0.000100 μF. Then we can find the product $LC = 100 \times 0.000100 = 0.0100$. When we take the square root of this, we get 0.100. Finally, we can divide 0.15915 by 0.100, getting $f_o = 1.5915$ MHz. We should round this off to 1.59 MHz.

Problem 17-11

Find the resonant frequency of a parallel circuit consisting of a 33-μH coil and a 47-pF capacitor.

Solution

Let's convert the capacitance to 0.000047 μF. Then we find the product $LC = 33 \times 0.000047 = 0.001551$. Taking the square root of this, we get 0.0393827. Finally, we divide 0.15915 by 0.0393827 and round off, getting $f_o = 4.04$ MHz.

Problem 17-12

Suppose that we want to design a circuit so that it exhibits $f_o = 9.00$ MHz. We have a 33.0-pF fixed capacitor available. What size coil will we need to obtain the desired resonant frequency?

Solution

Let's use the formula for the resonant frequency and plug in the values. Then we can use arithmetic to solve for L. We convert the capacitance to 0.0000330 μF and calculate in steps as follows:

$$
\begin{aligned}
f_o &= 0.15915 / (LC)^{1/2} \\
9.00 &= 0.15915 / (L \times 0.0000330)^{1/2} \\
9.00^2 &= 0.15915^2 / (0.0000330 \times L) \\
81.0 &= 0.025329 / (0.0000330 \times L) \\
81.0 \times 0.0000330 \times L &= 0.025329 \\
0.002673 \times L &= 0.025329 \\
L &= 0.025329 / 0.002673 \\
&= 9.48\ \mu H
\end{aligned}
$$

Problem 17-13

Suppose that we want to design an *LC* circuit with $f_o = 455$ kHz. We have a 100-µH in our "junk box." What size capacitor do we need?

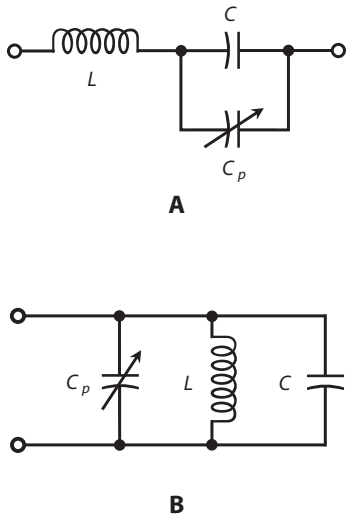
Solution

We should convert the frequency to 0.455 MHz. Then the calculation proceeds in the same way as with the preceding problem:

$$f_o = 0.15915 / (LC)^{1/2}$$
$$0.455 = 0.15915 / (100 \times C)^{1/2}$$
$$0.455^2 = 0.15915^2 / (100 \times C)$$
$$0.207025 = 0.025329 / (100 \times C)$$
$$0.207025 \times 100 \times C = 0.025329$$
$$20.7025 \times C = 0.025329$$
$$C = 0.025329 / 20.7025$$
$$= 0.00122 \text{ µF}$$

Adjusting the Resonant Frequency

In practical circuits, engineers often place variable inductors and/or variable capacitors in series or parallel *LC* circuits designed to function at resonance, thereby allowing for small errors in the actual resonant frequency (as opposed to the calculated value for f_o). We can design a circuit of this sort, called a *tuned circuit*, so that it exhibits a frequency slightly higher than f_o, and then install a *padding capacitance* (C_p) in parallel with the main capacitance C, as shown in Fig. 17-13. A padding capacitor is a small component with an adjustable value ranging from around 1 pF up to several picofarads, or several tens of picofarads. If the engineers want the circuit to offer a wider range of resonant frequencies, a variable capacitor, having a value ranging from a few picofarads up to several hundred picofarads, can serve the purpose.



17-13 Padding capacitors (C_p) allow limited adjustment of the resonant frequency in a series *LC* circuit (as shown at A), or in a parallel *LC* circuit (as shown at B).

Resonant Devices

Resonant circuits often consist of coils and capacitors in series or parallel, but other kinds of hardware also exhibit resonance.

Piezoelectric Crystals

Pieces of the mineral *quartz*, when cut into thin wafers and subjected to voltages, will vibrate at high frequencies. Because of the physical dimensions of such a *piezoelectric crystal*, these vibrations occur at a precise frequency f_o, and also at whole-number multiples of f_o. We call these multiples—$2f_o$, $3f_o$, $4f_o$, and so on—*harmonic frequencies* or simply *harmonics*. The frequency f_o constitutes the *fundamental frequency* or simply the *fundamental*. The fundamental, f_o, is the lowest frequency at which resonance occurs. Quartz crystals can act like *LC* circuits in electronic devices. A crystal exhibits an impedance that varies with frequency. The reactance equals zero at f_o and the harmonic frequencies.

Cavities

Lengths of metal tubing, cut to specific dimensions, exhibit resonance at very-high, ultra-high, and microwave radio frequencies. They work in much the same way as musical instruments resonate with sound waves. However, the waves take the form of EM fields rather than acoustic disturbances. Such *cavities*, also called *cavity resonators*, have reasonable physical dimensions at frequencies above about 150 MHz. We can get a cavity to work below this frequency, but we'll find it difficult to construct because of its great length and clumsiness. Like crystals, cavities resonate at a specific fundamental frequency f_o, and also at all of the harmonic frequencies.

MEMS Oscillators

A relatively new invention, MEMs (MicroElectronicMechanical System) oscillators are essentially chips with a cavity fabricated onto their silicon substrate along with a "sustaining" amplifier that keeps the oscilllation going. They may also conatin other circuitry such as a PLL (see Chap. 26) that allows the frequency of the oscillator to be programmed. Such devices are taking over from more conventional crystal oscillators because of their small size and low cost.

Sections of Transmission Line

When we cut a transmission line to any whole-number multiple of ¼ wavelength, it behaves as a resonant circuit. The most common length for a *transmission-line resonator* is ¼ wavelength, giving us a so-called *quarter-wave section*.

When we short-circuit a quarter-wave section at one end and apply an AC signal to the other end, the section acts like a parallel-resonant *LC* circuit, and it has an extremely high (theoretically infinite) resistive impedance at the fundamental resonant frequency f_o. When we leave one end of a quarter-wave section open and apply an AC signal to the other end, the section acts like a series-resonant *LC* circuit, and it has an extremely low (theoretically zero) resistive impedance at f_o. In effect, a quarter-wave section converts an AC short circuit into an AC open circuit and vice versa—at a specific frequency f_o.

The length of a quarter-wave section depends on the desired fundamental resonant frequency f_o. It also depends on how fast the EM energy travels along the line. We can define this speed in terms of a *velocity factor*, abbreviated *v*, expressed as a fraction or percentage of the speed of

light in free space. Manufacturers provide velocity factors for prefabricated transmission lines, such as coaxial cable or old-fashioned two-wire television "ribbon."

If the frequency in megahertz equals f_o and the velocity factor of a line equals v (expressed as a fraction), then we can calculate the length L_{ft} of a quarter-wave section of transmission line in feet using the formula

$$L_{ft} = 246v / f_o$$

If we know the velocity factor as a percentage $v_{\%}$, then the above formula becomes

$$L_{ft} = 2.46v_{\%} / f_o$$

If we know the velocity factor as a fraction v, then the length L_m of a quarter-wave section in meters is

$$L_m = 75.0v / f_o$$

For the velocity factor as a percentage $v_{\%}$, we have

$$L_m = 0.750v_{\%} / f_o$$

Note that we use L to stand for "length," not "inductance," in this context!

Antennas

Many types of antennas exhibit resonant properties. The simplest type of resonant antenna, and the only kind that we'll consider here, is the center-fed, half-wavelength *dipole antenna* (Fig. 17-14).

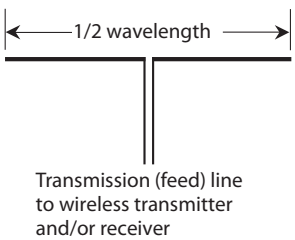
We can calculate the approximate length L_{ft}, in feet, for a dipole antenna at a frequency of f_o using the formula

$$L_{ft} = 467 / f_o$$

taking into account the fact that EM fields travel along a wire at about 95% of the speed of light. A straight, thin wire in free space, therefore, has a velocity factor of approximately $v = 0.95$. If we specify the approximate length of the half-wave dipole in meters as L_m, then

$$L_m = 143 / f_o$$

A half-wave dipole has a purely resistive impedance of about 73 Ω at its fundamental frequency f_o. But this type of antenna also resonates at harmonics of f_o. The dipole measures a full wavelength from end to end at $2f_o$; it measures 3/2 wavelength from end to end at $3f_o$; it measures two full wavelengths from end to end at $4f_o$, and so on.



17-14 The half-wave, center-fed dipole constitutes a simple and efficient antenna.

Radiation Resistance

At f_o and all of the *odd-numbered harmonics*, a dipole antenna behaves like a series resonant *RLC* circuit with a fairly low resistance. At all *even-numbered* harmonics, the antenna acts like a parallel resonant *RLC* circuit with a high resistance.

Does the mention of resistance in a half-wave dipole antenna confuse you? Maybe it should! Figure 17-14 shows no resistor. Where, you ask, does the resistance in a half-wave dipole come from? The answer gets into some rather esoteric EM-wave theory, and it brings to light an interesting property that all antennas have: *radiation resistance*. This parameter constitutes a crucial factor in the design and construction of all RF antenna systems.

When we connect a radio transmitter to an antenna and send out a signal, energy radiates into space in the form of radio waves from the antenna. Although no physical resistor exists anywhere in the antenna system, the radiation of radio waves acts just like power dissipation in a pure resistance. In fact, if we replace a half-wave dipole antenna with a 73-Ω nonreactive resistor that can safely dissipate enough power, we'll discover that a wireless transmitter connected to the opposite end of the line won't "know" the difference between that resistor and the dipole antenna.

Problem 17-14

How many feet long, to the nearest foot, is a quarter-wave section of transmission line at 7.1 MHz if the velocity factor equals 80%?

Solution

We can use the formula for the length L_{ft} of a quarter-wave section based on the velocity factor $v_{\%} = 80$, as follows:

$$\begin{aligned} L_{ft} &= 2.46\, v_{\%} / f_o \\ &= (2.46 \times 80) / 7.1 \\ &= 28 \text{ ft} \end{aligned}$$

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

18
CHAPTER

Transformers and Impedance Matching

WE CAN USE A TRANSFORMER TO OBTAIN THE OPTIMUM VOLTAGE FOR A CIRCUIT, DEVICE, OR SYSTEM. Transformers have various uses in electricity and electronics. For example, they can:

- Match the impedances between a circuit and a load.
- Match the impedances between two different circuits or devices.
- Provide DC isolation between circuits or devices while letting AC pass.
- Make balanced and unbalanced circuits, feed systems, and loads compatible.

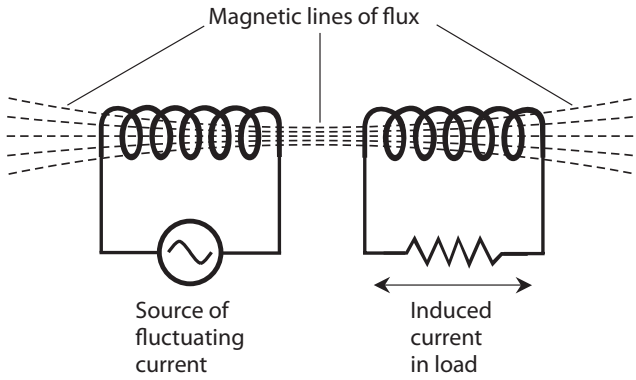
Transformers taking AC from a mains outlet and producing low voltage for operating low-voltage appliances used to be found everywhere. These days, those low-frequency bulky transformers are used quite rarely, but high-frequency lightweight transformers are found in switch mode power supplies and transformers are also used in Ethernet connectors as well as a host of more traditional uses.

Principle of the Transformer

When we place two wires near and parallel to each other and then drive a fluctuating current through one of them, a fluctuating current appears in the other, even though no direct physical connection exists between them. We call this effect *electromagnetic induction*. All AC transformers work according to this principle.

Induced Current and Coupling

If one wire carries sine-wave AC of a certain frequency, then the *induced current* shows up as sine-wave AC of the same frequency in the other wire. As we reduce the spacing between the two wires, keeping them straight and parallel at all times, the induced current increases for a given current in the first wire. If we wind the wires into coils (making certain that the wires are insulated or enameled so that they can't "short out" between the coil turns or to anything else nearby) and place the coils along a common axis, as shown in Fig. 18-1, we observe more induced current than we do if the



18-1 Magnetic lines of flux between two aligned coils, when one coil carries fluctuating or alternating current.

same wires run straight and parallel. We say that the *coupling* improves when we coil the wires up and place them along a common axis. We can improve the coupling (efficiency of induced-current transfer) still more if we wind one coil directly over the other.

Primary and Secondary

A transformer comprises two coils of insulated or enameled wire, along with the *core*, or *form*, on which we wind them. We call the first coil, through which we deliberately drive current, the *primary winding*. We call the second coil, in which the induced current appears, the *secondary winding*. Engineers and technicians usually call them simply the *primary* and the *secondary*.

When we apply AC to a primary winding, the coil currents are attended by potential differences between the coil ends, constituting the *primary voltage* and *secondary voltage*. In a *step-down transformer*, the primary voltage exceeds the secondary voltage. In a *step-up transformer*, the secondary voltage exceeds the primary voltage. Let's abbreviate the primary voltage as V_{pri} and the secondary voltage as V_{sec}. Unless otherwise stated, we specify effective (RMS) voltages when talking about the AC in a transformer.

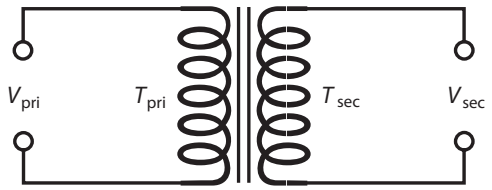
The windings of a transformer exhibit inductance because they're coils. The optimum inductance values for the primary and secondary depend on the frequency of operation, and also on the resistive components of the impedances of the circuits to which we connect the windings. As the frequency increases and the resistive component of the impedance remains constant, the optimum inductance decreases. If the resistive component of the impedance increases and the frequency remains constant, the optimum inductance increases.

Turns Ratio

We define the *primary-to-secondary turns ratio* in a transformer as the ratio of the number of turns in the primary T_{pri} to the number of turns in the secondary T_{sec}. We can denote this ratio as $T_{pri}:T_{sec}$ or T_{pri}/T_{sec}. In a transformer with optimum primary-to-secondary coupling (Fig. 18-2), we always find that

$$V_{pri}/V_{sec} = T_{pri}/T_{sec}$$

Stated in words, the primary-to-secondary voltage ratio equals the primary-to-secondary turns ratio.



18-2 The primary voltage (V_{pri}) and secondary voltage (V_{sec}) in a transformer depend on the number of turns in the primary winding (T_{pri}) versus the number of turns in the secondary winding (T_{sec}).

Problem 18-1

Suppose that a transformer has a primary-to-secondary turns ratio of exactly 9:1. We apply 117 V RMS AC across the primary terminals. Do we have a step-up transformer or a step-down transformer? How much voltage can we expect to see across the secondary?

Solution

This device is a step-down transformer. Let's use the above equation and solve for V_{sec}. We start with

$$V_{pri}/V_{sec} = T_{pri}/T_{sec}$$

We can plug in the values $V_{pri} = 117$ and $T_{pri}/T_{sec} = 9.00$ to obtain

$$117/V_{sec} = 9.00$$

Using a little bit of algebra, we can solve this equation to obtain

$$V_{sec} = 117 / 9.00 = 13.0 \text{ V RMS}$$

Problem 18-2

Consider a transformer with a primary-to-secondary turns ratio of exactly 1:9. The voltage across the primary equals 121.4 V RMS. Do we have a step-up transformer or a step-down transformer? What's the voltage across the secondary?

Solution

In this case, we have a step-up transformer. As before, we can plug in the numbers and solve for V_{sec}. We start with

$$V_{pri}/V_{sec} = T_{pri}/T_{sec}$$

Inputting $V_{pri} = 121.4$ and $T_{pri}/T_{sec} = 1/9.000$, we get

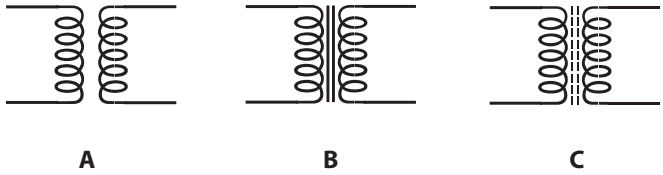
$$121.4/V_{sec} = 1/9.000$$

which solves to

$$V_{sec} = 9.000 \times 121.4 = 1093 \text{ V RMS}$$

Which Ratio Is Which?

Sometimes, when you read the specifications for a transformer, the manufacturer will quote the *secondary-to-primary turns ratio* rather than the primary-to-secondary turns ratio. You can denote the secondary-to-primary turns ratio as T_{sec}/T_{pri}. In a step-down transformer, T_{sec}/T_{pri} is less than 1, but in a step-up unit, T_{sec}/T_{pri} is greater than 1.



18-3 Schematic symbols for transformers. At A, air core. At B, laminated-iron core. At C, powdered-iron core.

When you hear someone say that a transformer has a certain "turns ratio" (say 10:1), you'd better make sure that you know which ratio they're talking about. Do they mean T_{pri}/T_{sec} or T_{sec}/T_{pri}? If you get it wrong, you'll miscalculate the secondary voltage by a factor equal to the *square* of the turns ratio! For example, in a transformer in which $T_{sec}/T_{pri} = 10{:}1$ and to which an input voltage of 25-V RMS AC is being provided, you don't want to think that you'll see only 2.5-V RMS AC across the secondary winding, when in fact you'll have to contend with 250-V RMS AC.

Ferromagnetic Cores

If we place a sample of ferromagnetic material within a pair of coils composing a transformer, the extent of coupling increases beyond what we get with an air core. However, some energy is invariably lost as heat in a ferromagnetic transformer core. Also, ferromagnetic cores limit the maximum frequency at which a transformer will work efficiently.

The schematic symbol for an air-core transformer looks like two inductor symbols placed back-to-back, as shown in Fig. 18-3A. If the transformer has a *laminated-iron* (layered-iron) core, we add two parallel lines to the schematic symbol (Fig. 18-3B). If the core consists of *powdered iron*, we break up the two parallel lines (Fig. 18-3C).

In transformers intended for use with 60-Hz utility AC, and also for low audio-frequency (AF) applications, sheets of an alloy called *silicon steel*, glued together in layers, are often employed as transformer cores. The silicon steel goes by the nickname *transformer iron*. The layering breaks up the electrical currents that tend to circulate in solid-iron cores. These so-called *eddy currents* flow in loops, serving no useful purpose, but they heat up the core, thereby wasting energy that we could otherwise obtain from the secondary winding. We can "choke off" eddy currents by breaking up the core into numerous thin, flat layers with insulation between them.

Transformer Loss

A rather esoteric form of loss, called *hysteresis loss*, occurs in all ferromagnetic transformer cores, but especially in laminated iron. Hysteresis is the tendency for a core material to act "sluggish" in accepting a fluctuating magnetic field. Air cores essentially never exhibit this type of loss. In fact, air has the lowest overall loss of any known transformer core material. Laminated cores exhibit high hysteresis loss above the AF range, so they don't work well above a few kilohertz.

At frequencies up to several tens of megahertz, powdered iron can serve as an efficient RF transformer core material. It has high magnetic permeability and concentrates the flux considerably. High-permeability cores minimize the number of turns needed in the coils, thereby minimizing the ohmic (resistive) loss that can take place in the wires.

At the highest radio frequencies (more than a couple of hundred megahertz), air represents the best overall choice as a transformer core material because of its low loss and low permeability.

Transformer Geometry

The properties of a transformer depend on the shape of its core, and on the way in which the wires surround the core. In electricity and electronics practice, you'll encounter several different types of *transformer geometry*.

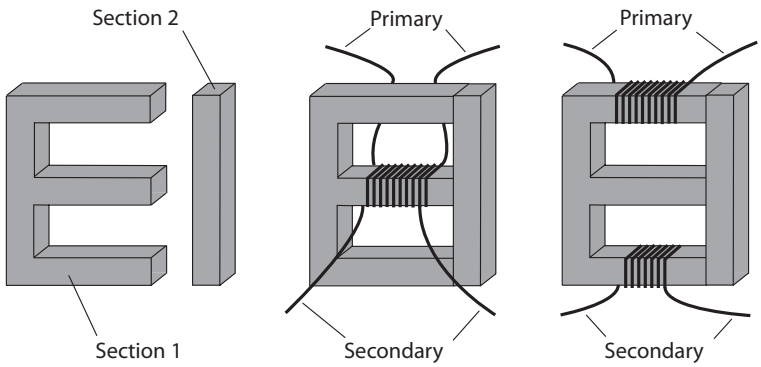
E Core

The *E core* gets its name from the fact that it has the shape of a capital letter E. A bar, placed at the open end of the E, completes the core assembly (Fig. 18-4A). We can wind a primary and secondary on an E core in either of two ways.

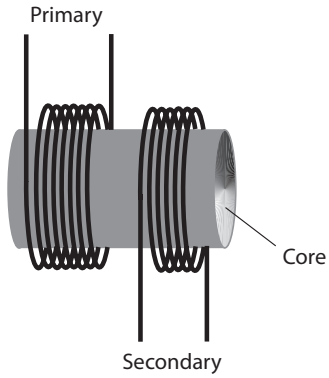
The simplest winding method involves winding both the primary coil and the secondary coil around the middle bar of the E, as shown in Fig. 18-4B. We call this scheme the *shell method* of transformer winding. It provides maximum coupling, but it also results in considerable capacitance between the primary and the secondary. Such *inter-winding capacitance* can sometimes be tolerated, but often it cannot. Another disadvantage of the shell geometry is the fact that, when we wind coils one on top of the other, the transformer can't handle very much voltage. High voltages cause *arcing* (sparking attended by unwanted current) between the windings, which can destroy the insulation on the wires, lead to permanent short circuits, and even set the transformer on fire.

If we don't want to use the shell method, we can employ the *core method* of transformer winding. In this scheme, we place one winding at the bottom of the E section, and the other winding at the top as shown in Fig. 18-4C. The coupling occurs by means of magnetic flux in the core. The core method results in lower inter-winding capacitance than we observe in a shell-wound transformer designed for the same voltage-transfer ratio because the windings are located physically farther apart. A core-wound transformer can handle higher voltages than a shell-wound transformer of the same physical size. Sometimes the center part of the E is left out of the core, in which case we have a transformer with an *O core* or a *D core*.

Shell-wound and core-wound transformers are commonly used at 60 Hz in electrical and electronic appliances and devices of all kinds. We can also find transformers of this sort in some older AF systems.



18-4 At A, a utility transformer E core, showing both sections. At B, the shell winding method. At C, the core winding method.



18-5 A solenoidal-core transformer.

Solenoidal Core

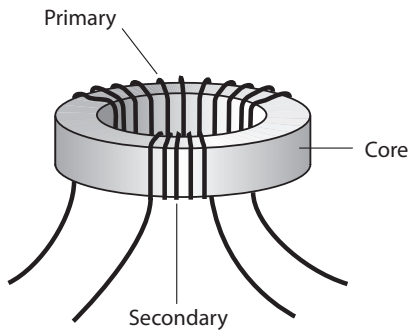
A pair of cylindrical coils, wound around a rod-shaped piece of powdered iron, can operate as an RF transformer, usually as a *loopstick antenna* in portable radio receivers and *radio direction-finding* (RDF) equipment. We can wind one of the coils directly over the other, or we can separate them, as shown in Fig. 18-5, to reduce the *interwinding capacitance* between the primary and secondary.

In a loopstick antenna, the primary winding intercepts the electromagnetic waves that carry the wireless signals. The secondary winding provides an optimum impedance match to the first amplifier stage, or *front end*, of the radio receiver or direction finder. We'll explore the use of transformers for impedance matching later in this chapter.

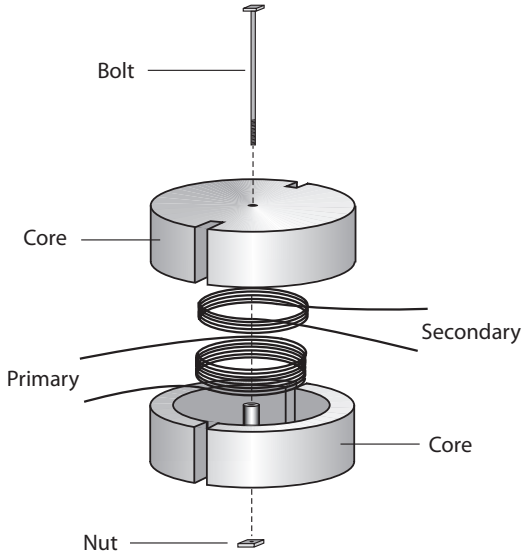
Toroidal Core

In recent decades, a donut-shaped transformer core called a *toroidal core* (or *toroid*) has become common for winding RF transformers. When we want to construct a *toroidal transformer*, we can wind the primary and secondary directly over each other, or we can wind them over different parts of the core, as shown in Fig. 18-6. As with other transformers, we observe more interwinding capacitance if we place the coils directly over each other than we do when we keep them physically separated.

In a toroidal inductor or transformer, practically all of the magnetic flux remains within the core material; almost none ventures outside. This property allows circuit designers to place toroids near other components without worrying about unintended mutual inductance. Also, a toroidal coil or transformer can be mounted directly on a metal chassis, and the external metal has no effect on transformer operation.



18-6 A toroidal-core transformer.



18-7 Exploded view of a pot-core transformer.

A toroidal core provides considerably more inductance per turn, for the same kind of ferromagnetic material, than a solenoidal core offers. We'll often see toroidal coils or transformers that have inductance values up to several hundred millihenrys.

Pot Core

A *pot core* takes the form of a ferromagnetic shell that completely surrounds a loop-shaped wire coil. The core is manufactured in two halves (Fig. 18-7). You wind the coil inside one of the shell halves, and then bolt the two shell halves together. In the resulting device, all of the magnetic flux remains confined to the core material.

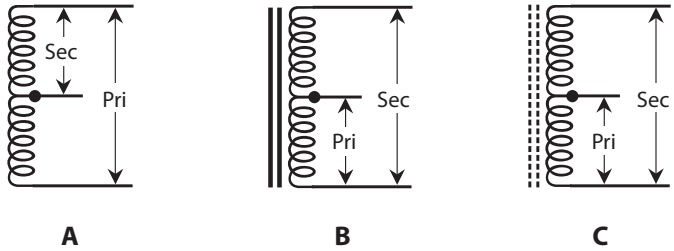
Like a toroid, a pot core is *self-shielding*. Essentially no magnetic coupling takes place between the windings and external components. You can use a pot core to wind a single, high-inductance coil. You can obtain inductance values of more than 1 H with a reasonable number of wire turns. However, you must wind the primary and secondary right next to each other; the geometry of the core prevents significant physical separation of the windings. Therefore, you'll always get a lot of interwinding capacitance.

Pot cores find diverse applications at AF, and also in the lowest-frequency part of the RF spectrum. You'll rarely, if ever, find these types of coils in high-frequency RF systems because other geometries can provide the necessary inductance values without the undesirable interwinding capacitance.

Autotransformer

In some situations, you might not need (or even want) DC isolation between the primary and secondary windings of a transformer. In a case of this sort, you can use an *autotransformer* that consists of a single, tapped winding. Figure 18-8 shows three autotransformer configurations:

- The unit at A has an air core, and operates as a step-down transformer.
- The unit at B has a laminated-iron core, and operates as a step-up transformer.
- The unit at C has a powdered-iron core, and operates as a step-up transformer.



18-8 Schematic symbols for autotransformers. At A, air core, step-down. At B, laminated iron core, step-up. At C, powdered iron core, step-up.

You'll sometimes find autotransformers in older radio receivers and transmitters. Autotransformers work well in impedance-matching applications. They also work as solenoidal loopstick antennas. Autotransformers are occasionally, but not often, used in AF applications and in 60-Hz utility wiring. In utility circuits, autotransformers can step the voltage down by a large factor, but they can't efficiently step voltages up by more than a few percent.

Power Transformers

Any transformer used in the 60-Hz utility line, intended to provide a certain RMS AC voltage for the operation of electrical circuits, constitutes a *power transformer*. Power transformers exist in a vast range of physical sizes, from smaller than a grapefruit to bigger than your living room.

At the Generating Plant

We'll find the largest transformers at locations where electricity is generated. Not surprisingly, high-energy power plants have bigger transformers that develop higher voltages than low-energy, local power plants have. These transformers handle extreme voltages and currents simultaneously.

When we want to transmit electrical energy over long distances, we must use high voltages. That's because, for a given amount of power ultimately dissipated by the loads, the current goes down as the voltage goes up. Lower current translates into reduced ohmic loss in the transmission line. Recall the formula for power in nonreactive circuits in terms of the current and the voltage:

$$P = VI$$

where P represents the power (in watts), V represents the voltage (in volts), and I represents the current (in amperes). If we can make the voltage 10 times larger, for a given power level, then the current goes down to only ⅒ as much. The ohmic losses in the wires vary in proportion to the *square* of the current. To understand why, remember that

$$P = I^2 R$$

where P represents the power (in watts), I represents the current (in amperes), and R represents the resistance (in ohms). Engineers can't do very much about the wire resistance or the power consumed by the loads in a large electrical grid, but the engineers can adjust the voltage, and thereby control the current.

Suppose that we increase the voltage in a power transmission line by a factor of 10, while the load at the end of the line draws constant power. The increase in voltage reduces the current to ⅒

of its previous value. As a result, we cut the ohmic loss to $(1/10)^2$, or 1/100, of its previous amount. We, therefore, enjoy a big improvement in the efficiency of the transmission line (at least in terms of the ohmic loss in the wires).

Now we know why regional power plants have massive transformers capable of generating hundreds of thousands—or even millions—of volts! Up to a certain limit, we get better results for our money if we use high RMS voltage than we do if we use heavy-gauge wire for long-distance utility transmission lines.

Along the Line

Extreme voltages work well for *high-tension* power transmission, but a 200-kV RMS AC electrical outlet would not interest the average consumer! The wiring in a high-tension system requires considerable precautions to prevent arcing (sparking) and short circuits. Personnel must remain at least several meters away from the wires to avoid electrocution.

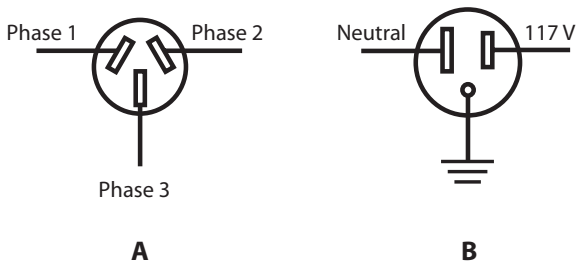
In a utility grid, medium-voltage power lines branch out from the major lines, and step-down transformers are used at the branch points. These lines fan out to lower-voltage lines, and step-down transformers are employed at these points, too. Each transformer must have windings heavy enough to withstand the product $P = VI$, the amount of VA power delivered to all the subscribers served by that transformer, at periods of peak demand.

Sometimes, such as during a heat wave, the demand for electricity rises above the normal peak level, "loading down" the circuit to the point that the voltage drops several percent. Then we have a *brownout*. If consumption rises further still, a dangerous current load appears at one or more intermediate power transformers. Circuit breakers in the transformers protect them from destruction by opening the circuit. Then we experience a *blackout*.

In most American homes, transformers step the voltage down to approximately 234-V RMS or 117-V RMS. Usually, 234-V RMS electricity appears in the form of three sine waves, called *phases*, each separated by 120°, and each appearing at one of the three slots in the outlet (Fig. 18-9A). We'll commonly see this system employed with heavy appliances, such as ovens, air conditioners, and washing machines. A 117-V RMS outlet, in contrast, supplies only one phase, appearing between two of the three slots in the outlet. The third opening should go directly to a substantial earth ground (Fig. 18-9B). This system is commonly used for basic household appliances, such as lamps, television sets, and computers.

In Electronic Equipment

Most consumer electronic systems have physically small power transformers. Most solid-state devices use low DC voltages, ranging from roughly 5 V to 50 V. For operation from 117-V RMS AC utility mains, therefore, such equipment requires either a step-down transformer as described here, or a switched-mode power supply as described in Chap. 24.



18-9 At A, an outlet for three-phase, 234-V RMS utility AC. At B, a conventional single-phase utility outlet for 117-V RMS utility AC.

In reality, the final transformation from the 117-V AC from the outlet on your wall to the now common 5-V USB connector of say a cellphone or Bluetooth speaker is only likely to directly use a step-down transformer as described here, if it only requires a very low operating power, or is an old piece of equipment. In high-powered AF or RF amplifiers, whose transistors can demand more than 1000 watts (1 kW) in some cases, the transformers require heavy-duty secondary windings, capable of delivering RMS currents of 90 A or more.

Warning!

Treat any voltage higher than 12 V as dangerous. The voltage in a television set or some ham radios can present an electrocution hazard even after you power the system down. Do not try to service such equipment unless you have the necessary training.

At Audio Frequencies

Audio-frequency (AF) power transformers resemble those employed for 60-Hz electricity, except that the frequency is somewhat higher (up to 20 kHz), and audio signals exist over a range, or *band*, of frequencies (20 Hz to 20 kHz) rather than at only one frequency.

Most AF transformers are constructed like miniature utility transformers. They have laminated E cores with primary and secondary windings wound around the cross bars, as shown in Fig. 18-4. Audio transformers can function in either the step-up or step-down mode, and are designed to match impedances rather than to produce specific output voltages.

Audio engineers strive to minimize the system reactance so that the absolute-value impedance, Z, is close to the resistance R for both the input and the output. For that ideal condition to exist, the reactance X must be zero or nearly zero. In the following discussion of impedance-matching transformers (both for AF and RF applications), let's assume that the impedances always constitute pure resistances of the form $Z = R + j0$.

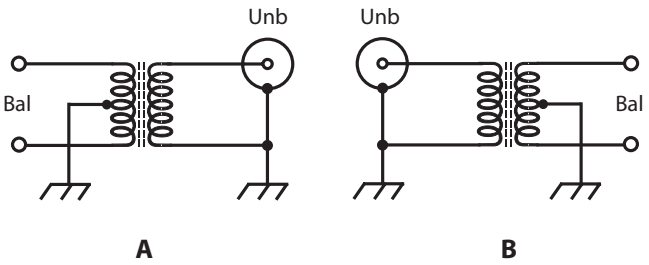
Isolation and Impedance Matching

Transformers can provide *isolation* between electronic circuits. While a transformer can and should provide *inductive coupling*, it should exhibit relatively little *capacitive coupling*. We can minimize the amount of capacitive coupling by using cores that minimize the number of wire turns needed in the windings, and by keeping the windings physically separated from each other (rather than overlapping).

Balanced and Unbalanced Loads and Lines

When we connect a device to a *balanced load*, we can reverse the terminals without significantly affecting the operating behavior. A plain resistor offers a good example of a balanced load. The two-wire antenna input in an old-fashioned analog television receiver provides another example. A *balanced transmission line* usually has two wires running alongside each other and separated by a constant physical distance, such as old-fashioned *TV ribbon*, also called *twinlead*.

An *unbalanced load* must be connected a certain way; we can't reverse the terminals. Switching the leads will result in improper circuit operation. In this sense, an unbalanced load resembles



18-10 At A, a balanced-to-unbalanced transformer. At B, an unbalanced-to-balanced transformer.

a polarized component, such as a battery, a diode, or an electrolytic capacitor. Many wireless antennas constitute unbalanced loads. Usually, unbalanced sources, transmission lines, and loads have one side connected to ground. The coaxial input of a television receiver is unbalanced; the shield (braid) of the cable connects to ground. An *unbalanced transmission line* usually comprises a coaxial cable of the sort that you see in a cable television system.

Normally, you can't connect an unbalanced line to a balanced load, or a balanced line to an unbalanced load, and expect to obtain optimum performance. However, a transformer can provide compatibility between these two types of systems. Figure 18-10A illustrates a *balanced-to-unbalanced transformer*. The balanced (input) side of the transformer has a grounded center tap. Figure 18-10B shows an *unbalanced-to-balanced transformer*. The balanced (output) side has a grounded center tap.

The turns ratio of a balanced-to-unbalanced transformer (also called a *balun*) or an unbalanced-to-balanced transformer (also known as an *unbal*) can equal 1:1, but that's not a requirement. If the impedances of the balanced and unbalanced parts of the systems are the same, then a 1:1 turns ratio works well. But if the impedances differ, the turns ratio should match the impedances. Shortly, we'll see how we can adjust the turns ratio of a transformer to convert a purely resistive impedance into another purely resistive impedance.

Transformer Coupling

Engineers sometimes use transformers between amplifier stages in electronic equipment to obtain a large *amplification factor* from the system. Part of the challenge in making a radio receiver or transmitter perform well involves getting the amplifiers to operate together in a stable manner. If too much feedback occurs, a series of amplifiers will oscillate, degrading or ruining the performance of the radio. Transformers that minimize the capacitance between the amplifier stages, while still transferring the desired signals, can help to prevent such oscillation.

Impedance-Transfer Ratio

The *impedance-transfer ratio* of a transformer varies according to the square of the turns ratio, and also according to the square of the voltage-transfer ratio. If the primary (source) and secondary (load) impedances are purely resistive and are denoted Z_{pri} and Z_{sec}, then

$$Z_{pri}/Z_{sec} = (T_{pri}/T_{sec})^2$$

and

$$Z_{pri}/Z_{sec} = (V_{pri}/V_{sec})^2$$

The inverses of these formulas, in which we express the turns ratio or voltage-transfer ratio in terms of the impedance-transfer ratio, are

$$T_{pri}/T_{sec} = (Z_{pri}/Z_{sec})^{1/2}$$

and

$$V_{pri}/V_{sec} = (Z_{pri}/Z_{sec})^{1/2}$$

Problem 18-3

Consider a situation in which we need a transformer to match an input impedance of 50.0 Ω, purely resistive, to an output impedance of 200 Ω, also purely resistive. What's the required turns ratio T_{pri}/T_{sec}?

Solution

The transformer must have a step-up impedance ratio of

$$Z_{pri}/Z_{sec} = 50.0/200 = 1/4.00$$

From the above information, we can calculate

$$T_{pri}/T_{sec} = (1/4.00)^{1/2} = 0.250^{1/2} = 0.5 = 1/2$$

Problem 18-4

Suppose that a transformer has a primary-to-secondary turns ratio of 9.00:1. The load, connected to the transformer output, constitutes a pure resistance of 8.00 Ω. What's the impedance at the primary?

Solution

The impedance-transfer ratio equals the square of the turns ratio. Therefore

$$Z_{pri}/Z_{sec} = (T_{pri}/T_{sec})^2 = (9.00/1)^2 = 9.00^2 = 81.0$$

We know that the secondary impedance Z_{sec} equals 8.00 Ω, so

$$Z_{pri} = 81.0 \times Z_{sec} = 81.0 \times 8.00 = 648\ \Omega$$

Radio-Frequency Transformers

Some RF transformers have primary and secondary windings, just like utility transformers. Others employ transmission-line sections. Let's look at these two types, which, taken together, account for most RF transformers in use today.

Wirewound Types

In the construction of wirewound RF transformers, we can use powdered-iron cores up to quite high frequencies. Toroidal cores work especially well because of their *self-shielding* characteristic (all of the magnetic flux stays within the core material). The optimum number of turns depends on the frequency, and also on the permeability of the core.

In high-power applications, air-core coils are often preferred. Although air has low permeability, it has negligible hysteresis loss, and will not heat up or fracture as powdered-iron cores sometimes do. However, some of the magnetic flux extends outside of an air-core coil, potentially degrading the performance of the transformer when it must function in close proximity to other components.

A major advantage of coil type transformers, especially when wound on toroidal cores, lies in the fact that we can get them to function efficiently over a wide band of frequencies, such as from 3.5 MHz to 30 MHz. A transformer designed to work well over a sizable frequency range is called a *broadband transformer*.

Transmission-Line Types

As you recall, any transmission line has a characteristic impedance, denoted as Z_o, that depends on the line construction. This property allows us to construct impedance transformers out of coaxial or parallel-wire line for operation at some radio frequencies.

Transmission-line transformers usually consist of quarter-wave sections. From the previous chapter, remember the formula for the length of a quarter-wave section:

$$L_{ft} = 246v/f_o$$

where L_{ft} represents the length of the section in feet, v represents the velocity factor expressed as a fraction or ratio, and f_o represents the operating frequency in megahertz. If we want to specify the length L_m in meters, then

$$L_m = 75v/f_o$$

Suppose that a quarter-wave section of line, with characteristic impedance Z_o, is terminated in a purely resistive impedance R_{out}. In this situation (Fig. 18-11), the impedance that appears at the input end of the line, R_{in}, also constitutes a pure resistance, and the following relations hold:

$$Z_o^2 = R_{in}\, R_{out}$$

and

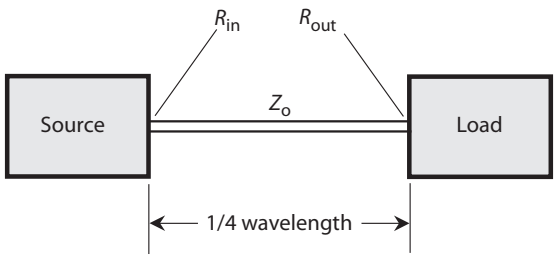
$$Z_o = (R_{in}\, R_{out})^{1/2}$$

We can rearrange the first formula to solve for R_{in} in terms of R_{out}, and vice versa:

$$R_{in} = Z_o^2 / R_{out}$$

and

$$R_{out} = Z_o^2 / R_{in}$$



18-11 A quarter-wave matching section of transmission line. The input impedance equals R_{in}, the output impedance equals R_{out}, and the characteristic impedance equals Z_o.

These equations hold true at the frequency f_o for which the line length measures ¼ wavelength. We can replace the word "wavelength" by the italic lowercase Greek letter lambda (λ), denoting the length of a quarter-wave section as $(¼)\lambda$ or 0.25λ.

Neglecting line losses, the above relations hold at all *odd harmonics* of f_o, that is, at $3f_o$, $5f_o$, $7f_o$, and so on, at which we have sections measuring 0.75λ, 1.25λ, 1.75λ, and so on. At other frequencies, a length of transmission line fails to function as a simple transformer. Instead, it behaves in a complex manner, the mathematical details of which would take us beyond the scope of this discussion.

Quarter-wave transmission-line transformers work well in antenna systems, especially at the higher frequencies (above a few megahertz) at which their dimensions become practical. A quarter-wave matching section should be constructed from an unbalanced line if the load is unbalanced, and from a balanced line if the load is balanced.

A disadvantage of quarter-wave sections arises from the fact that they work only at specific frequencies, depending on their physical length and on the velocity factor. But this shortcoming is often offset by the ease with which we can construct them, if we intend to use a piece of radio equipment at only one frequency, or at odd-numbered harmonics of that frequency.

Problem 18-5

Suppose an antenna has a purely resistive impedance of 100 Ω. We connect it to a ¼-wave section of 75-Ω coaxial cable. What's the impedance at the input end of the section?

Solution

Use the formula from above to calculate

$$R_{in} = Z_o^2 / R_{out} = 75^2 / 100 = 5625 / 100 = 56.25\ \Omega$$

We can round this result off to 56 Ω.

Problem 18-6

Consider an antenna known to have a purely resistive impedance of 300 Ω. You want to match it to the output of a radio transmitter designed to work into a 50.0-Ω pure resistance. What's the characteristic impedance needed for a quarter-wave matching section?

Solution

You can calculate it using the formula for that purpose, as follows:

$$Z_o = (R_{in}\, R_{out})^{1/2} = (300 \times 50.0)^{1/2} = 15{,}000^{1/2} = 122\ \Omega$$

Few, if any, commercially manufactured transmission lines have this particular characteristic impedance. Prefabricated lines come in standard Z_o values, and a perfect match might not be obtainable. In that case, you can try the closest obtainable Z_o. In this case, it would probably be 92 Ω or 150 Ω. If you can't find anything near the characteristic impedance needed for a quarter-wave matching section, then you're better off using a coil-type transformer.

What about Reactance?

When no reactance exists in an AC circuit using transformers, things stay simple. But often, especially in RF antenna systems, pure resistance doesn't occur naturally. We have to "force it" by

inserting inductors and/or capacitors to cancel the reactance out. The presence of reactance in a load makes a perfect match impossible with an impedance-matching transformer alone.

Inductive and capacitive reactances effectively oppose each other, and their magnitudes can vary. If a load presents a complex impedance $R + jX$, then we can cancel the reactance X by introducing an equal and opposite reactance $-X$ in the form of an inductor or capacitor connected in series with the load. This action gives us a pure resistance with a value equal to $(R + jX) - jX$, or simply R. The reactance-canceling component should always be placed at the point where the load connects to the line.

When we want to conduct wireless communications over a wide band of frequencies, we can place adjustable impedance-matching and reactance-canceling networks between the transmission line and the antenna. Such a circuit is called a *transmatch* or an *antenna tuner*. These devices not only match the resistive portions of the transmitter and load impedances, but they cancel reactances in the load. Transmatches are popular among amateur radio operators, who use equipment capable of operation from less than 2 MHz up to the highest known radio frequencies.

Optimum Location

Whenever we use a transformer or ¼-wave section to match the characteristic impedance of an RF feed line to the purely resistive impedance of a well-designed antenna, that transformer or section should be placed between the line and the antenna. We call this location the *feed point*. If we place the transformer or ¼-wave section anywhere else, it won't work at its best. In some cases, an improperly located transformer or ¼-wave section, even when tailored to the correct specifications, will make the overall impedance-mismatch situation worse!

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

3
PART
Basic Electronics

This page intentionally left blank

19
CHAPTER

Introduction to Semiconductors

SINCE THE 1960S WHEN THE TRANSISTOR BECAME COMMON IN CONSUMER DEVICES, *SEMICONDUCTORS* have acquired a dominating role in electronics. The term *semiconductor* arises from the ability of certain materials to conduct some of the time, but not all of the time. We can control the conductivity to obtain amplification, rectification, oscillation, signal mixing, switching, and other effects.

Developments in semiconductors have driven the move from discrete components into the use of integrated circuits and made possible the present-day consumer electronics.

The Semiconductor Revolution

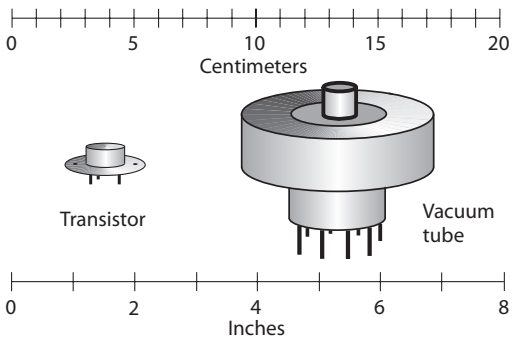
Decades ago, *vacuum tubes*, also known as *electron tubes*, were the only amplifying devices available for use in electronic systems. A typical tube (called a *valve* in the UK) ranged from the size of your thumb to the size of your fist. You'll still find tubes in some power amplifiers, microwave oscillators, and video display units.

Tubes generally required high voltage in order to operate efficiently. Even in modest radio receivers, it took at least 50-V DC, and more often 100-V DC or more, to get a tube to work. Such voltages mandated bulky, massive power supplies and created an electrical shock hazard.

Nowadays, a transistor of microscopic dimensions can perform the functions of a tube in most low-power electronic circuits. The power supply can comprise a couple of AA "flashlight cells" or a 9-V "transistor battery." Even in high-power applications, transistors have smaller dimensions and weigh less than tubes having similar signal-output specifications (Fig. 19-1).

Integrated circuits (ICs), hardly larger than individual transistors, can do the work of hundreds, thousands, or millions of vacuum tubes. You can find an excellent example of IC technology in personal computers and in the peripheral devices you use with them, such as displays, external disk drives, printers, and modems.

Vacuum tubes enjoy a few advantages over semiconductor devices, even today. We can momentarily exceed the voltage, current, or power rating of a vacuum tube, and the device will usually "forgive" us, while a semiconductor device in a similar application might "die" immediately. A few audio enthusiasts, especially in popular music bands, insist that amplifiers made with vacuum tubes produce better quality sound than similar amplifiers made with semiconductor devices.



19-1 A power-amplifier transistor (at left) has much smaller volume and mass than a vacuum tube of comparable signal-output capacity (right).

Semiconductor Materials

Various elements, compounds, and mixtures can function as semiconductors. Although the vast majority of semiconductors used in the electronics industry today are based on silicon, there are other semiconducting materials that are sometimes used in specialist applications.

Silicon

Silicon is an element with atomic number 14 and atomic weight 28. In its pure state, silicon appears as a light-weight metal similar to aluminum. Pure silicon conducts electric currents better than a dielectric material does, but not as well as most metallic conductors, such as silver, copper, or aluminum.

The earth's crust contains silicon in great abundance, and we can mine it from some crustal rocks and sand. In its natural state, silicon almost always exists "tied up" as compounds with other elements. Industrial vendors extract the pure element. Electronic-component manufacturers mix other substances, known as *impurities*, with silicon to give the semiconductor material specific properties. The resulting solids are cut into thin slices known as *chips* or *wafers*.

Gallium Arsenide

Another common semiconductor is the compound gallium arsenide. Engineers and technicians write its acronym-like chemical symbol, GaAs, and pronounce it aloud as "gas." If you hear about "gasfets" and "gas ICs," you're hearing about gallium-arsenide technology.

GaAs devices perform well at higher frequencies than silicon-based devices do because electric currents can travel faster through GaAs than through silicon compounds. GaAs devices are relatively immune to the direct effects of *ionizing radiation*, such as X rays and gamma rays.

You will also find GaAs used in photovoltaic cells, infrared LEDs, and laser diodes.

Selenium

Selenium is a chemical element whose electrical conductivity varies depending on the intensity of visible, ultraviolet (UV), or infrared (IR) radiation that strikes it. All semiconductor materials have this property, known as *photoconductivity*, to some degree, but in selenium, the effect is pronounced. For this reason, selenium constitutes an excellent choice for the manufacture of *photocells*. Selenium can also work well in certain types of *rectifiers* that convert AC to pulsating DC.

Selenium has exceptional *electrical ruggedness*, meaning that it can withstand short-lived electrical overloads, such as too much current or voltage. Selenium-based components can survive brief *transients*, or "spikes" of abnormally high voltage, better than components made with most other semiconductor materials.

Germanium

Pure elemental germanium constitutes a poor electrical conductor. It becomes a semiconductor only when we add impurities. In the early years of semiconductor technology, engineers used germanium-based components far more often than they do now. In fact, we'll rarely encounter a germanium device in any electronic system today. High temperatures, such as soldering tools generate, can destroy a germanium-based diode or transistor.

Metal Oxides

Certain metal oxides have properties that make them useful in the manufacture of semiconductor devices. When you hear about MOS (pronounced "moss") or CMOS (pronounced "sea moss") technology, you're hearing about *metal-oxide semiconductor* and *complementary metal-oxide semiconductor* devices, respectively.

Certain types of transistors, and many kinds of ICs, make use of MOS technology. In integrated circuits, MOS and CMOS construction allows for a large number of discrete components, such as resistors, inductors, diodes, and transistors, on a single chip. Engineers say that MOS/CMOS has *high component density*.

Metal-oxide components need almost no current in order to function. When we use a battery to power a small MOS-based device, that battery will last almost as long as it would if we let it sit on the shelf and didn't use it for anything. Most MOS-based devices can work at extreme speeds, allowing for operation at high frequencies in radio-frequency (RF) equipment, and facilitating the rapid switching that's important in today's computers.

All MOS and CMOS components suffer from one outstanding limitation: A discharge of static electricity can destroy one of them in an instant. We must use care when handling components of this type. Any technician working with MOS and CMOS components should wear a metal wrist strap that connects one wrist to a good earth ground, so static-electric buildup cannot occur.

This is especially important if you are handling them in a very low humidity environment.

Doping and Charge Carriers

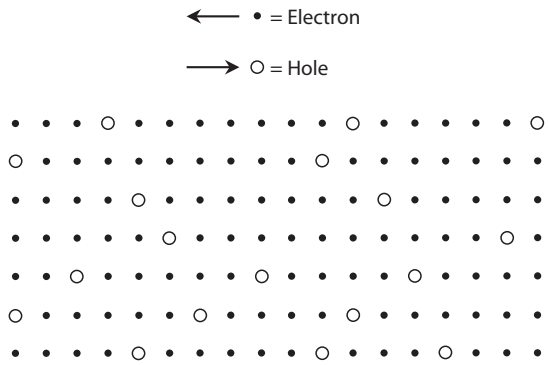
Impurities give a semiconductor material the properties that it needs to function as an electronic component. The impurities cause the material to conduct current in certain ways. When manufacturers add an impurity to a semiconductor element, they call the process *doping*. The impurity material itself is called a *dopant*.

Donor Impurities

When a dopant contains an inherent excess of electrons, we call it a *donor impurity*. Adding such a substance causes conduction mainly by means of electron flow, as in an ordinary metal such as copper. The excess electrons move from atom to atom when a potential difference exists between two different points in the material. Elements that serve as donor impurities include antimony, arsenic, bismuth, and phosphorus. A material with a donor impurity is called an *N-type semiconductor*, because an electron carries a unit of negative (N) electric charge.

Acceptor Impurities

If an impurity has an inherent deficiency of electrons, we call it an *acceptor impurity*. When we add an element, such as aluminum, boron, gallium, or indium, to a semiconductor element, the resulting material conducts mainly by means of *hole flow*. A *hole* comprises an "assigned spot" in an atom where an electron would exist under normal conditions, but in fact doesn't show up there.



19-2 Pictorial representation of hole flow. Solid black dots represent electrons, moving in one direction. Open circles represent holes, moving in the opposite direction.

A semiconductor with an acceptor impurity is called a *P-type semiconductor* because a hole has, in effect, a unit of positive (P) electric charge.

Majority and Minority Carriers

Charge carriers in semiconductor materials always constitute either electrons or holes. We never see "oddball" charge carriers, such as protons or helium nuclei, in electronic devices (although we do encounter them in high-energy physics). In any semiconductor substance, some of the current takes the form of electrons passed from atom to atom in a negative-to-positive direction, and some of the current occurs as holes that move from atom to atom in a positive-to-negative direction.

Sometimes electrons account for most of the current in a semiconductor. This situation exists if the material has donor impurities, that is, if it's of the N type. In other cases, holes account for most of the current. This phenomenon occurs when the material has acceptor impurities, making it P type. We call the more abundant, or dominating, charge carriers (either electrons or holes) the *majority carriers*. We call the less abundant ones the *minority carriers*. The ratio of majority to minority carriers can vary, depending on the exact chemical composition of the semiconductor material.

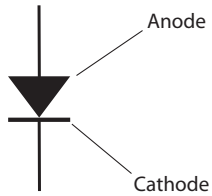
Figure 19-2 shows a simplified illustration of electron flow versus hole flow in a sample of N-type semiconductor material, in which the majority carriers constitute electrons and the minority carriers constitute holes. Each point location in the grid represents an atom. The solid black dots represent electrons. Imagine them moving from right to left as they "jump" from atom to atom. Small open circles represent holes. Imagine them moving from left to right as they "jump" from atom to atom. In the example, the positive battery or power-supply terminal (the "source of holes") would lie somewhere out of the picture toward the left, and the negative battery or power-supply terminal (the "source of electrons") would lie out of the picture toward the right.

The P-N Junction

Connecting a piece of semiconducting material, either P or N type, to a source of current can provide us with phenomena for scientific observations and experiments. But when the two types of material come into direct contact, the boundary between the P-type sample and the N-type sample, called the *P-N junction*, behaves in ways that make semiconductor materials truly useful.

The Semiconductor Diode

Figure 19-3 shows the schematic symbol for a *semiconductor diode*, formed by joining a piece of P-type material to a piece of N-type material. The N-type semiconductor is represented by the



19-3 Schematic symbol for a semiconductor diode.

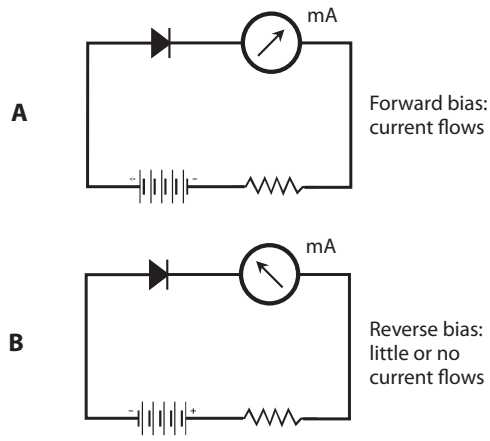
short, straight line in the symbol; we call it the *cathode*. The P-type semiconductor is represented by the arrow; we call it the *anode*. Electrons can move easily in the direction opposite the arrow, and holes can move easily in the direction in which the arrow points. Electrons normally do not move with the arrow, and holes normally do not move against the arrow.

If you connect a battery and a resistor in series with the diode, you'll get a current to flow if you connect the negative battery terminal to the diode's cathode and the positive terminal to the anode, as shown in Fig. 19-4A. A series-connected resistor eliminates the risk of diode destruction by excessive current. No current will flow if you reverse the battery polarity, as shown in Fig. 19-4B.

It takes a specific, well-defined minimum applied voltage for conduction to occur through a semiconductor diode in the situation shown by Fig. 19-4A. We call this threshold potential difference the *forward breakover voltage*. Depending on the type of semiconductor material, the forward breakover voltage for a particular diode can vary from about 0.3 V to 1 V. If the voltage across the P-N junction falls short of the forward breaker voltage, the diode will fail to conduct current, even when we connect it as shown in Fig. 19-4A. This effect, known as the *forward breakover effect* or the *P-N junction threshold effect*, allows us to build circuits to limit the maximum positive and/or negative peak voltages that signals can attain. We can also take advantage of this effect to construct a device called a *threshold detector*, in which a signal's positive or negative peak amplitude must equal or exceed a certain minimum in order to pass through.

How the Junction Works

When the N-type material has a negative voltage with respect to the P type (as in Fig. 19-4A) that exceeds the forward breakover voltage, electrons flow easily from N to P. The N-type semiconductor, which already has an excess of electrons, receives more; the P-type semiconductor, already



19-4 Series connection of a battery, a resistor, a current meter, and a diode. At A, forward bias results in a flow of current. At B, reverse bias results in no current.

"suffering" from a shortage of electrons, gets deprived of still more. The N-type material constantly "feeds" electrons to the P type in an "attempt" to create an electron balance, and the battery or power supply keeps "robbing" electrons from the P-type material in order to sustain the electron imbalance. Figure 19-5A illustrates this condition, known as *forward bias*. Current can flow through a forward-biased diode easily under these circumstances.

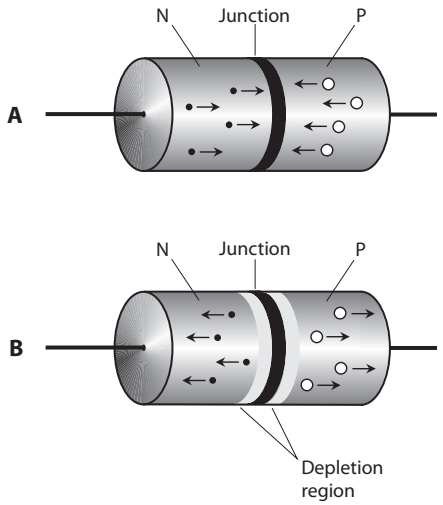
When we reverse the battery or DC power-supply polarity so that the N-type material acquires a positive voltage with respect to the P-type material, we have a condition called *reverse bias*. Electrons in the N-type material migrate toward the positive charge pole, away from the P-N junction. In the P-type material, holes drift toward the negative charge pole, also away from the P-N junction. The electrons constitute the majority carriers in the N-type material, and the holes are the majority carriers in the P-type material. The charge, therefore, disappears in the vicinity of the P-N junction, as shown in Fig. 19-5B. This "charge-free zone," where majority carriers are deficient, is called the *depletion region*. A shortage of majority carriers in any semiconductor substance means that the substance cannot conduct well, so a depletion region acts like an electrical insulator. This phenomenon reveals the reason why a semiconductor diode will not normally conduct when reverse-biased. A diode forms a "one-way current gate"—usually!

When the cathode and anode of a P-N junction have the same electrical potential as applied from an external source, we call the condition *zero bias*.

Junction Capacitance

Some P-N junctions can alternate between conduction (in forward bias) and non-conduction (in reverse bias) millions or billions of times per second. Other P-N junctions can't work so fast. The maximum switching speed depends on the capacitance at the P-N junction during conditions of reverse bias. As the *junction capacitance* of a diode increases, the highest frequency at which it can alternate between the conducting state and the non-conducting state decreases.

The junction capacitance of a diode depends on several factors, including the operating voltage, the type of semiconductor material, and the cross-sectional area of the P-N junction. If you examine Fig. 19-5B, you might get the idea that the depletion region, sandwiched between two semiconducting sections, can play a role similar to that of the dielectric in a capacitor. If so, you're right!



19-5 At A, forward bias of a P-N junction. At B, reverse bias of the same junction. Solid black dots represent electrons. White dots represent holes. Arrows indicate general directions of charge-carrier (hole or electron) movement.

A reverse-biased P-N junction forms a capacitor. Some semiconductor components, called *varactor diodes*, are manufactured with this property in mind.

We can vary the junction capacitance of a diode by changing the reverse-bias voltage because this voltage affects the width of the depletion region. As we increase the reverse voltage, the depletion region gets wider, and the capacitance goes down.

Avalanche Effect

Sometimes, a diode conducts even when reverse-biased. The greater the reverse bias voltage, the more like an electrical insulator a P-N junction gets—up to a point. But if the reverse bias reaches or exceeds a specific critical value, the voltage overcomes the ability of the junction to prevent the flow of current, and the junction conducts as if forward-biased. Engineers call this phenomenon the *avalanche effect* because conduction occurs in a sudden and massive way, something like an avalanche on a snowy mountain.

Avalanche effect does not damage a P-N junction (unless the applied reverse voltage is extreme). It's a temporary thing. The depletion region disappears while avalanche breakdown occurs. But when the voltage drops back below the critical value, the junction behaves normally again. If the bias remains reversed after the avalanche breakdown condition ends, the depletion region returns.

Some diodes are designed to take advantage of the avalanche effect. In other cases, avalanche effect limits the performance of a circuit. In a device called a *Zener diode*, designed specifically to regulate DC voltages, you'll hear or read about the *Zener voltage* specification. This value can range from a couple of volts to well over 100 V. Zener voltage is theoretically equivalent to avalanche voltage, but in the case of a Zener diode, the manufacturer tailors the semiconductor material so as to produce an exact, predictable avalanche voltage.

For *rectifier diodes* in power supplies, you'll hear or read about the *peak inverse voltage* (PIV) or *peak reverse voltage* (PRV) specification. That's the highest instantaneous reverse-bias voltage that we can expect the device to withstand without risking avalanche breakdown. In practical applications, rectifier diodes must have PIV ratings great enough so that the avalanche effect will never occur (or even come close to happening) during any part of the AC cycle.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

20
CHAPTER

Diode Applications

DIODES ARE THE ONE-WAY STREETS OF ELECTRONICS, ONLY ALLOWING CURRENT TO FLOW THROUGH them in one direction. This makes them incredibly useful devices and you will find them in all sorts of circuits.

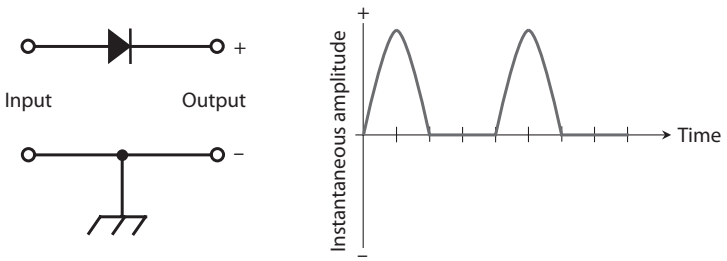
Rectification

A *rectifier diode* passes current in only one direction, as long as we don't exceed its specifications. This property makes the device useful for changing AC to DC (rectification). Generally speaking:

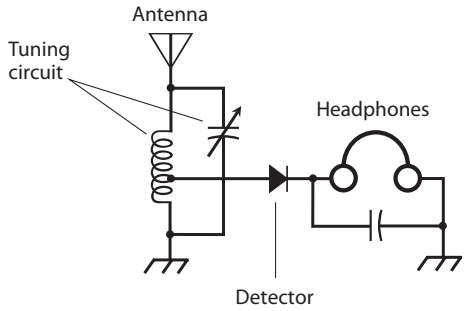
- When the cathode has a more negative charge than the anode, current flows.
- When the cathode has a more positive charge than the anode, current does not flow.

The constraints on this behavior are, as we've learned, the forward breakover and avalanche voltages.

Examine the circuit shown in Fig. 20-1A. Suppose that we apply a 60-Hz AC sine wave to the input terminals. During half of the cycle, the diode conducts, and during the other half, it doesn't. This behavior cuts off half of every cycle. Depending on which way we connect the diode, we can



20-1 At A, a half-wave rectifier circuit. At B, the output of the circuit shown at A when we apply an AC sine wave to the input.



20-2 Schematic diagram of a crystal-set radio receiver.

make it cut off either the positive half of the AC cycle or the negative half. Figure 20-1B is a graph of the output of the circuit shown at A.

The circuit and wave diagrams of Fig. 20-1 involve a *half-wave rectifier* circuit, which is the simplest possible rectifier. Simplicity constitutes the chief advantage of the half-wave rectifier over other rectifier circuits. You'll learn about various types of rectifier diodes and circuits in the next chapter.

Detection

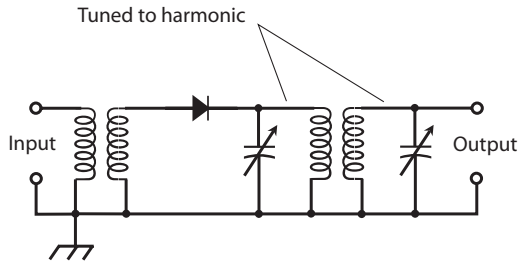
One of the earliest diodes, existing even before vacuum tubes, was made partly with semiconductor material. Known as a "cat's whisker," the device comprised a fine piece of wire in contact with a small fragment of the mineral *galena*. This contraption had the ability to act as a rectifier for extremely weak RF currents. When experimenters connected the "cat's whisker" in a configuration such as the circuit of Fig. 20-2, the resulting device could receive amplitude-modulated (AM) radio signals and produce audible output in the headset.

The galena fragment, sometimes called a "crystal," gave rise to the nickname *crystal set* for this primitive radio receiver. You can build a crystal set today using an RF diode, a coil, a tuning capacitor, a headset, and a long-wire antenna. The circuit needs no battery or other source of electrical power! If a broadcast station exists within a few miles of the antenna, the received signal alone produces enough audio to drive the headset. For ideal performance, the headset should be *shunted* with a capacitor whose value is large enough to "short out" residual RF current to ground, but not so large as to "short out" the audio signal. (When we say *shunted* in this context, we mean that the capacitor goes in parallel, or in *shunt*, with the headset.)

In the circuit of Fig. 20-2, the diode recovers the audio from the radio signal. We call this process *detection* or *demodulation*, and we call the whole circuit a *detector* or *demodulator*. If we want the detector to function, we must use a diode that has low junction capacitance, so that it can rectify at RF without acting like a capacitor. Some modern RF diodes resemble microscopic versions of the old "cat's whisker," enclosed in glass cases with axial leads.

Frequency Multiplication

When current passes through a diode, half of the cycle gets cut off, as shown in Fig. 20-1B. This "chopping-off" effect occurs from 60-Hz utility current through RF, regardless of the applied-signal frequency, as long as the diode capacitance remains small and as long as the reverse voltage remains below the avalanche threshold. The output wave from the diode looks much different than the input wave. We call this condition *nonlinearity*. Whenever a circuit exhibits nonlinearity, harmonics



20-3 A frequency-multiplier circuit using a semiconductor diode.

appear in the output. The harmonics show up as signals at integer multiples of the input frequency, as we learned in Chap. 9.

In situations where nonlinearity represents an undesirable state of affairs, engineers strive to make electronic circuits *linear*, so the output waveform has exactly the same shape as the input waveform (even if the amplitudes differ). But in some applications, we want a circuit to act in a nonlinear fashion—for example, when we intend to generate harmonics. We can deliberately introduce nonlinearity in a circuit to obtain *frequency multiplication*. Diodes work well for this purpose. Figure 20-3 illustrates a simple *frequency-multiplier* circuit. We tune the output *LC* circuits to the desired *n*th harmonic frequency, nf_o, rather than to the input or fundamental frequency, f_o.

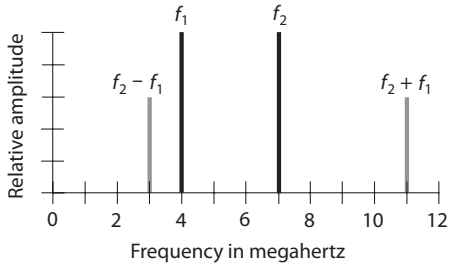
For a diode to work as a frequency multiplier in RF systems, it must be of a type that would also work well as a detector at the same frequencies. This means that the component should act like a rectifier, but not like a capacitor.

Signal Mixing

When we combine two waves having different frequencies in a nonlinear circuit, we get new waves at frequencies equal to the sum and difference of the frequencies of the input waves. Diodes can provide the nonlinearity that we need to make this happen.

Consider two AC signals with frequencies f_1 and f_2. Let's assign f_2 to the wave with the higher frequency, and f_1 to the wave with the lower frequency. If we combine these two signals in a nonlinear circuit, new waves result. One of the new waves has a frequency of $f_2 + f_1$, and the other has a frequency of $f_2 - f_1$. We call these sum and difference frequencies *beat frequencies*. We call the signals themselves *mixing products* or *heterodynes*. The heterodynes appear in the output along with the original signals at frequencies f_1 and f_2.

Figure 20-4 shows hypothetical input and output signals for a *mixer* circuit on a *frequency-domain* display. The amplitude (on the vertical scale or axis) constitutes a function of the frequency (on the horizontal scale or axis). Engineers see this sort of display when they look at the screen



20-4 Spectral (frequency-domain) illustration of signal mixing.

of a lab instrument known as a *spectrum analyzer*. In contrast, an ordinary oscilloscope displays amplitude (on the vertical scale or axis) as a function of time (on the horizontal scale or axis), so it provides a *time domain* display.

Switching

The ability of diodes to conduct currents when forward-biased and block currents when reverse-biased makes them useful for switching in some applications. Diodes can perform switching operations much faster than any mechanical device—up to millions or even billions of on/off operations per second.

One type of diode, made for use as an RF switch, has a special semiconductor layer sandwiched in between the P type and N type material. The material in this layer is called an *intrinsic* (or *I type*) *semiconductor*. The *intrinsic layer* (or *I layer*) reduces the capacitance of the diode, allowing the device to function effectively at higher frequencies than an ordinary diode can. A diode with an I type semiconductor layer sandwiched in between the P and N type layers is called a *PIN diode* (Fig. 20-5).

Direct-current bias, applied to one or more PIN diodes, allows us to effectively channel RF currents to desired points without using relays and cables. A PIN diode also makes a good RF detector, especially at very high frequencies.

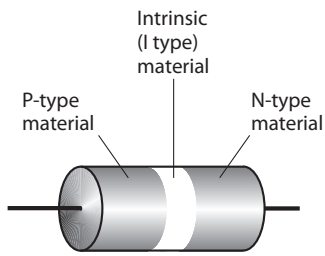
Voltage Regulation

Most diodes have an avalanche breakdown voltage much higher than the reverse-bias voltage ever gets. The value of the avalanche voltage depends on the internal construction of the diode, and on the characteristics of the semiconductor materials that compose it. *Zener diodes* are specially made to exhibit well-defined, constant avalanche voltages.

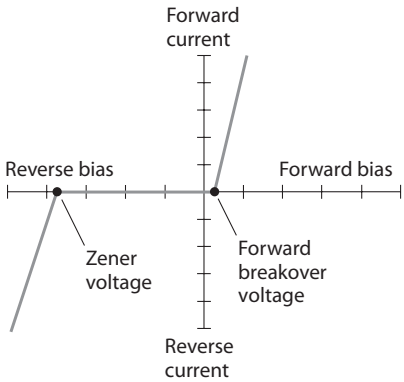
Suppose that a certain Zener diode has an avalanche voltage, also called the *Zener voltage*, of 50 V. If we apply a reverse bias to the P-N junction, the diode acts as an open circuit as long as the potential difference between the P- and N-type materials remains less than 50 V. But if the reverse-bias voltage reaches 50 V, even for a moment, the diode conducts. This phenomenon prevents the instantaneous reverse-bias voltage from exceeding 50 V.

Figure 20-6 shows a graph of the current through a hypothetical Zener diode as a function of the voltage. The Zener voltage shows up as an abrupt rise in the reverse current as the reverse-bias voltage increases (that is, as we move toward the left along the horizontal axis).

Figure 20-7 shows a simple Zener-diode voltage-regulator circuit. Note the polarity of the diode: we connect the cathode to the positive pole and the anode to the negative pole, opposite from the way we use a diode in a rectifier circuit. The series-connected resistor limits the current



20-5 The PIN diode has a layer of intrinsic (I type) semiconductor material at the P-N junction.



20-6 Current through a Zener diode as a function of the bias voltage.

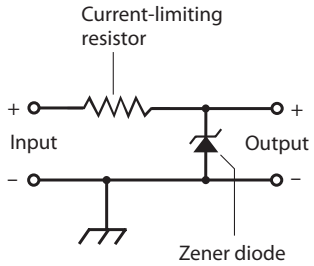
that can flow through the Zener diode. Without that resistor, the diode would conduct excessive current and burn out.

Amplitude Limiting

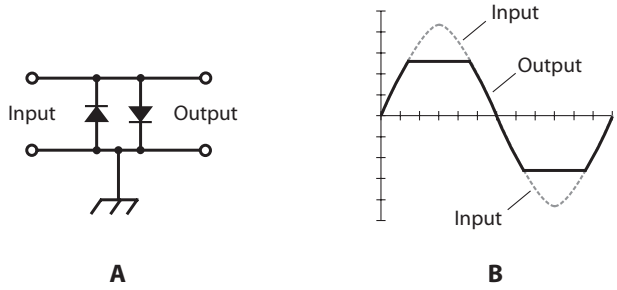
In Chap. 19, we learned that a forward-biased diode will not conduct until the voltage reaches or exceeds the forward breakover voltage. We can state a corollary to this principle: A diode will always conduct when the forward-bias voltage reaches or exceeds the forward breakover voltage. In a diode, the potential difference between the P- and N-type wafers remains fairly constant—roughly equal to the forward breakover voltage—as long as current flows in the forward direction. In the case of silicon diodes, this potential difference or *voltage drop* is approximately 0.6 V. For germanium diodes, the voltage drop is roughly 0.3 V, and for selenium diodes it's around 1 V.

We can take advantage of the "constant-voltage-drop" property of semiconductor diodes when we want to build a circuit to limit the amplitude of a signal. Figure 20-8A shows how we can connect two identical diodes back-to-back in parallel with the signal path to *limit*, or *clip*, the positive and negative peak voltages of an input signal. In this configuration, the peak voltages are limited to the forward breakover voltage of the diodes. Figure 20-8B shows the input and output waveforms of a typical clipped AC signal.

The downside of the *diode voltage-limiter* circuit, such as the one shown in Fig. 20-8A, is the fact that it introduces distortion when clipping occurs. This distortion might not cause a problem for the reception of digital signals, frequency-modulated (FM) signals, or analog signals that rarely reach the limiting voltage. But for amplitude-modulated (AM) signals with peaks that rise past the



20-7 Connection of a Zener diode for voltage regulation. The series-connected resistor limits the current to prevent destruction of the diode.



20-8 At A, connection of two diodes to act as an AC limiter. At B, illustration of sine-wave peaks cut off by the action of the diodes in an AC limiter.

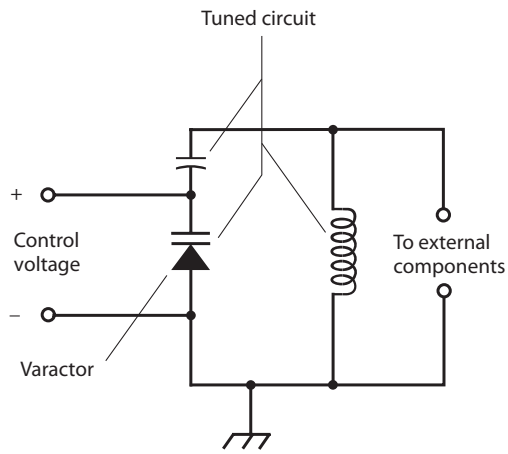
limiting voltage, *clipping distortion* can make voices difficult to understand, and it utterly ruins the sound quality of music!

Frequency Control

When we reverse-bias a diode, we observe a region at the P-N junction with dielectric (insulating) properties. As you know from the last chapter, this zone is called the depletion region because it has a shortage (depleted supply) of majority charge carriers. The width of the depletion region depends on several parameters, including the reverse-bias voltage.

As long as the reverse bias remains lower than the avalanche voltage, varying the bias affects the width of the depletion region. The fluctuating width in turn varies the junction capacitance. This capacitance, which is always on the order of only a few picofarads, varies inversely with the square root of the reverse-bias voltage—again, as long as the reverse bias remains less than the avalanche voltage. For example, if we quadruple the reverse-bias voltage, the junction capacitance drops to half; if we decrease the reverse-bias voltage by a factor of 9, then the junction capacitance increases by a factor of 3.

Although now largely replaced by VCO ICs, some diodes are manufactured especially for use as variable capacitors. Such a device is known as a varactor diode, as you learned in the last chapter. Varactors find their niche in a special type of circuit called a *voltage-controlled oscillator* (VCO). Figure 20-9 shows an example of a parallel-tuned *LC* circuit in a VCO, using a coil, a fixed capacitor,



20-9 Connection of a varactor diode in a tuned circuit.

and a varactor. The fixed capacitor, whose value should greatly exceed the capacitance of the varactor, keeps the coil from short-circuiting the control voltage across the varactor. The schematic symbol for a varactor diode has two lines on the cathode side, as opposed to one line in the symbol for a conventional diode.

Oscillation and Amplification

Under certain conditions, diodes can generate or amplify *microwave* RF signals—that is, signals at extremely high AC frequencies. Devices commonly employed for these purposes include the *Gunn diode*, the *IMPATT diode*, and the *tunnel diode*.

Gunn Diodes

A Gunn diode can produce from 100 mW to 1 W of RF power output, and is manufactured with gallium arsenide (GaAs). A Gunn diode oscillates because of the *Gunn effect*, named after the engineer *J. Gunn* who first observed it in the 1960s while working for the International Business Machines (IBM) Corporation.

A Gunn diode doesn't work like a rectifier, detector, or mixer. Instead, the oscillation takes place as a result of a quirk called *negative resistance*, in which an increase in the instantaneous applied voltage causes a decrease in the instantaneous current flow under specific conditions.

Gunn-diode oscillators are often tuned using varactor diodes. A Gunn-diode oscillator, connected directly to a horn-shaped antenna, gives us a device known as a *Gunnplexer*. Amateur-radio experimenters use Gunnplexers for low-power wireless communication at frequencies of 10 GHz and above.

IMPATT Diodes

The acronym *IMPATT* comes from the words *imp*act *a*valanche *t*ransit *t*ime. This type of diode, like the Gunn diode, works because of the negative resistance phenomenon. An *IMPATT diode* constitutes a microwave oscillating device like a Gunn diode, except that it's manufactured from silicon rather than gallium arsenide. An IMPATT diode can operate as an amplifier for a microwave transmitter that employs a Gunn-diode oscillator. As an oscillator, an IMPATT diode produces about the same amount of output power, at comparable frequencies, as a Gunn diode does.

Tunnel Diodes

Another type of diode that will oscillate at microwave frequencies is the *tunnel diode*, also known as the *Esaki diode*. Made from GaAs semiconductor material, the tunnel diode produces only enough power to function as a local oscillator in a microwave radio receiver or transceiver. Tunnel diodes work well as weak-signal amplifiers in microwave receivers because they generate very little unwanted noise. The low-noise characteristic is typical of GaAs devices.

Energy Emission

Some semiconductor diodes emit radiant energy when current passes through the P-N junction in a forward direction. This phenomenon occurs as electrons "fall" from higher to lower energy states within atoms.

LEDs and IREDs

Depending on the exact mixture of the semiconductors used in manufacture, visible light of almost any color can be produced by forward-biased diodes. Infrared and ultraviolet-emitting devices also exist. An *infrared-emitting diode* (IRED) produces energy at wavelengths slightly longer than those of visible red light.

The intensity of the radiant energy from an LED or IRED depends to some extent on the forward current. As the current rises, the brightness increases, but only up to a certain point. If the current continues to rise, no further increase in brilliance takes place, and we say that the LED or IRED is working in a state of *saturation*.

Digital Displays

Because LEDs can be made in various different shapes and sizes, they work well in digital displays. You've seen digital clock radios, hi-fi radios, calculators, and car radios that use LEDs. They make good indicators for "on/off," "a.m./p.m.," "battery low," and other conditions.

Many TV screens, computer monitors, and graphical displays of all shapes and sizes are now manufactured using a type of LED called an organic LED (OLED). OLEDs contain an organic electroluminescent layer that actually emits the light; their construction means that, unlike "regular" LEDs, they can be printed onto a substrate, even a flexible or curved surface. Hence, some smartphones have screens that curve around the body of the phone.

Communications

Both LEDs and IREDs work well in communications systems because we can modulate their intensity to carry information. When the current through the device is sufficient to produce output, but not so great as to cause saturation, the LED or IRED output follows along with rapid current changes. This phenomenon allows engineers to build circuits for transmitting digital signals over visible-light and IR energy beams. Modern telephone systems make use of modulated light transmitted through clear glass or plastic cables, a technology called *fiberoptics*.

Laser Diodes

Special LEDs and IREDs, known as *laser diodes*, produce *coherent radiation*. The rays from these diodes aren't the intense, parallel beams that most people imagine when they think about lasers. A laser LED or IRED generates a cone-shaped beam of low intensity. However, we can use lenses to focus the emission into a parallel beam. The resulting rays have some of the same properties as the beams from large lasers, including the ability to travel long distances with minimal decrease in intensity.

In fact, laser diodes of several Watts in power are now available that can be used in laser cutters and engravers.

Photosensitive Diodes

Most P-N junctions exhibit conductivity that varies with exposure to radiant energy, such as IR, visible light, and UV. Conventional diodes aren't normally affected by these rays because they're enclosed in opaque packages! Some *photosensitive diodes* have variable DC resistance that depends on the intensity of the visible, IR, or UV rays that strike their P-N junctions. Other types of diodes produce their own DC in the presence of radiant energy.

Silicon Photodiodes

A silicon diode, housed in a transparent case and constructed so that visible light can strike the barrier between the P and N type materials, forms a *silicon photodiode*. If we apply a reverse-bias voltage to the device at a certain level below the avalanche threshold, no current flows when the junction remains in darkness, but current flows when sufficient radiant energy strikes.

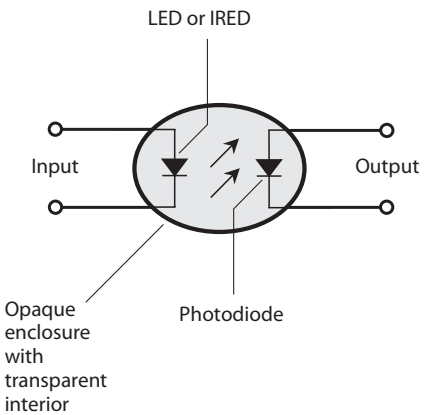
At constant reverse-bias voltage, the current varies in direct proportion to the intensity of the radiant energy, within certain limits. When radiant energy of variable intensity strikes the P-N junction of a reverse-biased silicon photodiode, the output current follows the light-intensity variations. This property makes silicon photodiodes useful for receiving modulated-light signals of the kind used in fiberoptic communications systems.

Silicon photodiodes exhibit greater sensitivity to radiant energy at some wavelengths than at others. The greatest sensitivity occurs in the *near infrared* part of the spectrum, at wavelengths slightly longer than the wavelength of visible red light.

The Optoisolator

An LED or IRED and a photodiode can be combined in a single package to construct a component called an *optoisolator*. This device, shown schematically in Fig. 20-10, creates a modulated-light signal and sends it over a small, clear gap to a receptor. An LED or IRED converts the electrical input signal to visible light or IR. A photodiode changes the visible light or IR back into an electrical signal, which appears at the output. An opaque enclosure prevents external light or IR from reaching the photodiode, ensuring that the photodiode receives only the radiation from the internal source.

When we want to transfer, or *couple*, an AC signal from one circuit to another, the two stages interact if we make the transfer through a direct electrical connection. The input impedance of a given stage, such as an amplifier, can affect the behavior of the circuits that "feed power" to it, leading to problems. Optoisolators overcome this effect because the coupling occurs optically rather than electrically. If the input impedance of the second circuit changes, the first circuit "sees" no change in the output impedance, which comprises only the impedance of the LED or IRED. The circuits' impedances, and the electrical effects thereof, are literally isolated from each other.



20-10 An optoisolator contains an LED or IRED at the input and a photodiode at the output with a transparent medium between them.

Photovoltaic Cells

A silicon diode, with no bias voltage applied, can generate DC all by itself if enough IR, visible, or UV energy strikes its P-N junction. We call this phenomenon the *photovoltaic effect*. Solar cells work because of this effect.

Photovoltaic cells are specially manufactured to have the greatest possible P-N junction surface area, thereby maximizing the amount of radiant energy that strikes the junction. A single silicon photovoltaic cell can produce about 0.6 V of DC electricity in daylight. The amount of current that it can deliver, and therefore, the amount of power it can provide, depends on the surface area of the junction.

We can connect photovoltaic cells in series-parallel combinations to provide power for solid-state electronic devices, such as portable radios. The DC from these arrays can charge batteries, allowing for use of the electronic devices when radiant energy is not available (for example, at night or on dark days). A large assembly of solar cells, connected in series-parallel, is called a *solar panel*.

The power produced by a solar panel depends on the power from each individual cell, the number of cells in the panel, the intensity of the radiant energy that strikes the panel, and the angle at which the rays hit the surface of the panel. Some solar panels can produce several kilowatts of electrical power when the midday sun's unobstructed rays arrive perpendicular to the surfaces of all the cells.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

21
CHAPTER

Bipolar Transistors

THE WORD *TRANSISTOR* IS A CONTRACTION OF "CURRENT-*TRANS*FERRING RES*ISTOR*." AS THE WORLD of electronics gets more digital and is more concerned with switching rather than analog behavior, bipolar transistors are not quite so ubiquitous as they once were. However, they are still a popular choice for many applications.

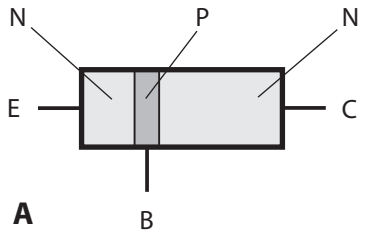
A *bipolar transistor* has two P-N junctions. Two configurations exist for bipolar transistors: a P-type layer between two N-type layers (called *NPN*), or an N-type layer between two P-type layers (called *PNP*).

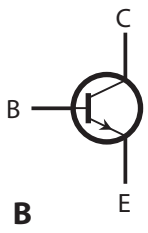
NPN versus PNP

Figure 21-1A is a simplified drawing of an *NPN bipolar transistor*, and Fig. 21-1B shows the symbol that engineers use for it in schematic diagrams. The P type, or center, layer constitutes the *base*. One of the N type semiconductor layers forms the *emitter*, and the other N-type layer forms the *collector*. We label the base, the emitter, and the collector B, E, and C, respectively.

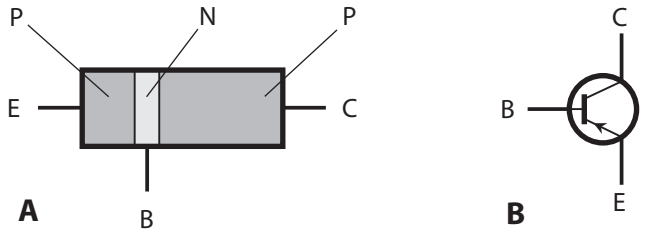
A *PNP bipolar transistor* has two P type layers, one on either side of a thin, N-type layer as shown in Fig. 21-2A. The schematic symbol appears at Fig. 21-2B. The N-type layer constitutes the base. One of the P-type layers forms the emitter, and the other P-type layer forms the collector. As with the NPN device, we label these electrodes B, E, and C.

We can tell from a schematic diagram whether the circuit designer means for a transistor to be NPN type or PNP type. Once we realize that the arrow always goes with the emitter, we can identify





21-1 At A, a simplified structural drawing of an NPN transistor. At B, the schematic symbol. We identify the electrodes as E = emitter, B = base, and C = collector.



21-2 At A, a simplified structural drawing of a PNP transistor. At B, the schematic symbol. We identify the electrodes as E = emitter, B = base, and C = collector.

the three electrodes without having to label them. In an NPN transistor, the arrow at the emitter points outward. In a PNP transistor, the arrow at the emitter points inward.

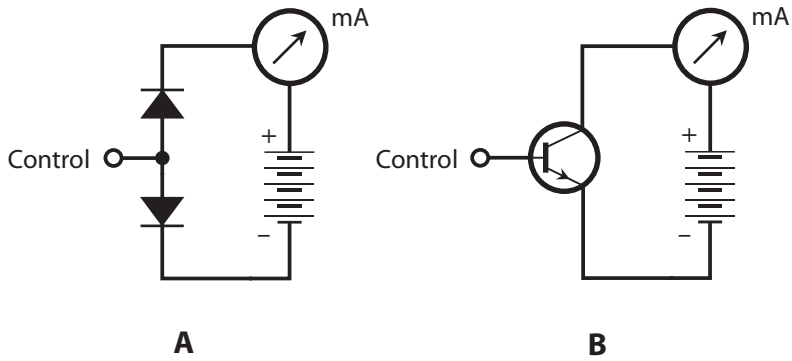
Generally, PNP and NPN transistors can perform the same electronic functions. However, they require different voltage polarities, and the currents flow in different directions. In many situations, we can replace an NPN device with a PNP device or vice versa, reverse the power-supply polarity, and expect the circuit to work with the replacement device if it has the appropriate specifications.

Biasing

For a while, let's imagine a bipolar transistor as two diodes connected in *reverse series* (that is, in series but in opposite directions). We can't normally connect two diodes together this way and get a working transistor, but the analogy works for *modeling* (technically describing) the behavior of bipolar transistors. Figure 21-3A shows a dual-diode NPN transistor model. The base is at the connection point between the two anodes. The cathode of one diode forms the device's emitter, while the cathode of the other diode forms the collector. Figure 21-3B shows the equivalent "real-world" NPN transistor circuit.

NPN Biasing

In an NPN transistor, we normally bias the device so that the collector voltage is positive with respect to the emitter. We illustrate this scheme by indicating the battery's polarity in Figs. 21-3A and 21-3B. Typical DC voltages for a transistor's power supply range between about 3 V and 50 V, meaning that the collector electrode and the emitter electrode differ in potential by 3 V to 50 V.



21-3 At A, the dual-diode model of a simple NPN circuit. At B, the actual transistor circuit.

In these diagrams, we label the base point "control" because the flow of current through the transistor depends on what happens at this electrode. Any change in the voltage that we apply to the base—either in the form of DC or AC—profoundly affects what happens inside the transistor, and also what happens in other components that we connect to it.

Zero Bias for NPN

Suppose that we connect an NPN transistor so that the base and the emitter are at the same voltage. We call this condition *zero bias* because the potential difference between the two electrodes equals 0 V. In this situation, the *emitter-base current*, often called simply the *base current* and denoted I_B, equals zero. The *emitter-base* (E-B) *junction*, which is a P-N junction, operates below its forward breakover voltage, preventing current from flowing between the emitter and the base. Zero bias also prevents any current from flowing between the emitter and the collector unless we inject an AC signal at the base to change things. Such a signal must, at least momentarily, attain a positive voltage equal to or greater than the forward breakover voltage of the E-B junction. When no current flows between the emitter and the collector in a bipolar transistor under no-signal conditions, we say that the device is operating in a state of *cutoff*.

Reverse Bias for NPN

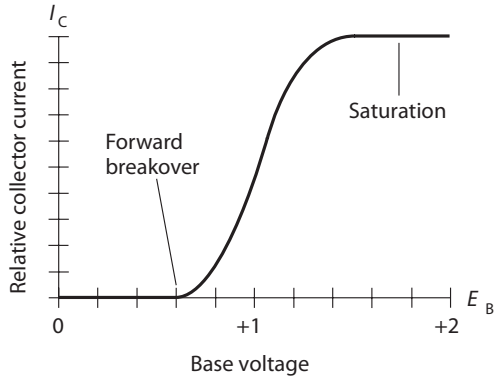
Imagine that we connect an extra battery between the base and the emitter in the circuit of Fig. 21-3B, with the polarity set such as to force the base voltage E_B to become negative with respect to the emitter. The presence of the new battery causes the E-B junction to operate in a state of reverse bias. No current flows through the E-B junction in this situation (as long as the new battery voltage is not so great that avalanche breakdown occurs at the E-B junction). In that sense, the transistor behaves in the same way when reverse-biased as it does when zero-biased. If we inject an AC signal at the base with the intent to cause a flow of current during part of the cycle, the signal must attain, at least momentarily, a positive voltage high enough to overcome the sum of the reverse bias voltage (produced by the battery) and the forward breakover voltage of the E-B junction. We'll have a harder time getting a reverse-biased transistor to conduct than we'll have getting a zero-biased transistor to conduct.

Forward Bias for NPN

Now suppose that we make E_B positive with respect to the emitter, starting at small voltages and gradually increasing. This situation gives us a state of *forward bias* at the E-B junction. If the forward bias remains below the forward breakover voltage, no current flows, either in the E-B junction or from the emitter to the collector. But when the base voltage E_B reaches and then exceeds the breakover point, the E-B junction conducts, and current starts to flow through the E-B junction.

The base-collector (B-C) junction of a bipolar transistor is normally reverse-biased. It will remain reverse-biased as long as E_B stays smaller than the supply voltage between the emitter and the collector. In practical transistor circuits, engineers commonly set E_B at a small fraction of the supply voltage. For example, if the battery between the emitter and the collector in Figs. 21-3A and 21-3B provides 12 V, the small battery between the emitter and the base might consist of a single 1.5-V cell. Despite the reverse bias of the B-C junction, a significant emitter-collector current, called *collector current* and denoted I_C, flows through the transistor once the E-B junction conducts.

In a real transistor circuit, such as the one shown in Fig. 21-3B, the meter reading will jump when we apply DC base bias to reach and then exceed the forward breakover voltage of the E-B junction. If we continue to increase the forward bias at the E-B junction, even a small rise in E_B, attended by a



21-4 Relative collector current (I_C) as a function of base voltage (E_B) for a hypothetical NPN silicon transistor.

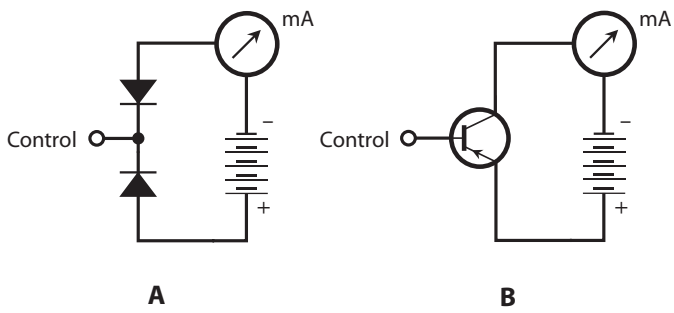
rise in the base current I_B, will cause a large increase in the collector current I_C. Figure 21-4 portrays the situation as a graph. Once current starts to flow in the collector, increasing E_B a tiny bit will cause the I_C to go up a lot! However, if E_B continues to rise, we'll eventually arrive at a voltage where the curve for I_C versus E_B levels off. Then we say that the transistor has reached a state of *saturation*. It's "running wide open," conducting as much as it possibly can, given a fixed potential difference between the collector and the emitter.

PNP Biasing

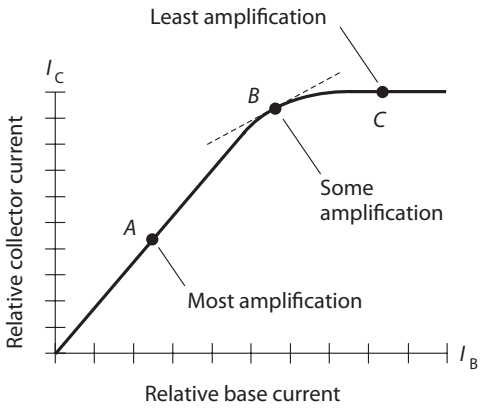
We can describe the situation inside a PNP transistor, as we vary the voltage of the small battery or cell between the emitter and the base, as a "mirror image" of the case for an NPN device. The diodes are reversed, the arrow points inward rather than outward in the transistor symbol, and all the polarities are reversed. The dual-diode PNP model, along with the "real-world" transistor circuit, appear in Fig. 21-5. We can repeat the foregoing discussion (for the NPN case) almost verbatim, except that we must replace every occurrence of the word "positive" with the word "negative." Qualitatively, the same things happen in the PNP device as in the NPN case.

Amplification

Because a small change in I_B causes a large variation in I_C when we set the DC bias voltages properly, a transistor can operate as a *current amplifier*. Engineers use several expressions to describe the current-amplification characteristics of bipolar transistors.



21-5 At A, the dual-diode model of a simple PNP circuit. At B, the actual transistor circuit.



21-6 Three different transistor bias points. We observe the most current amplification when we bias the device near the middle of the straight-line portion of the curve.

Collector Current versus Base Current

Figure 21-6 is a graph of the way the collector current I_C changes in a typical bipolar transistor as the base current I_B changes. We can see some points along this I_C versus I_B curve at which a transistor won't provide any current amplification. For example, if we operate the transistor in saturation (shown by the extreme upper right-hand part of the curve), the I_C versus I_B curve runs horizontally. In this zone, a small change in I_B causes little or no change in I_C. But if we bias the transistor near the middle of the "ramped-up" straight-line part of the curve in Fig. 21-6, the transistor will work as a current amplifier.

Whenever we want a bipolar transistor to amplify a signal, we must bias the device so that a small change in the current between the emitter and the base will result in a large change in the current between the emitter and the collector. The ideal voltages for E_B (the base bias) and E_C (the power-supply voltage) depend on the internal construction of the transistor, and also on the chemical composition of the semiconductor materials that make up its N-type and P-type sections.

Static Current Characteristics

We can describe the current-carrying characteristics of a bipolar transistor in simplistic terms as the *static forward current transfer ratio*. This parameter comes in two "flavors," one that describes the collector current versus the emitter current when we place the base at electrical ground (symbolized H_{FB}), and the other that describes the collector current versus the base current when we place the emitter at electrical ground (symbolized H_{FE}).

The quantity H_{FB} equals the ratio of the collector current to the emitter current at a given instant in time with the base grounded:

$$H_{FB} = I_C / I_E$$

For example, if an emitter current I_E of 100 mA results in a collector current I_C of 90 mA, then we can calculate

$$H_{FB} = 90/100 = 0.90$$

If $I_E = 100$ mA and $I_C = 95$ mA, then

$$H_{FB} = 95/100 = 0.95$$

The quantity H_{FE} equals the ratio of the collector current to the base current at a given instant in time with the emitter grounded:

$$H_{FE} = I_C / I_B$$

For example, if a base current I_B of 10 mA results in a collector current I_C of 90 mA, then we can calculate

$$H_{FE} = 90/10 = 9.0$$

If $I_B = 5.0$ mA and $I_C = 95$ mA, then

$$H_{FE} = 95/5.0 = 19$$

Alpha

We can describe the current variations in a bipolar transistor by dividing the *difference* in I_C by the *difference* in I_E that occurs when we apply a small signal to the emitter of a transistor with the base connected to electrical ground (or placed at the same potential difference as electrical ground). We call this ratio the *alpha*, symbolized as the lowercase Greek letter alpha (α). Let's abbreviate the words "the difference in" by writing *d*. Then mathematically, we can define

$$\alpha = dI_C / dI_E$$

We call this quantity the *dynamic current gain* of the transistor for the grounded-base situation. The alpha for any transistor is always less than 1, because whenever we apply a signal to the input, the base "bleeds off" at least a little current from the emitter before it shows up at the collector.

Beta

We get an excellent definition of current amplification for "real-world signals" when we divide the difference in I_C by the difference in I_B as we apply a small signal to the base of a transistor with the emitter at electrical ground. Then we get the dynamic current gain for the grounded-emitter case. We call this ratio the *beta*, symbolized as the lowercase Greek letter beta (β). Once again, let's abbreviate the words "the difference in" as *d*. Then we have

$$\beta = dI_C / dI_B$$

The beta for any transistor can exceed 1—and often does, greatly!—so this expression for "current gain" lives up to its name. However, under some conditions, we might observe a beta of less than 1. This condition can occur if we improperly bias a transistor, if we choose the wrong type of transistor for a particular application, or if we attempt to operate the transistor at a signal frequency that's far higher than the maximum frequency for which it is designed.

How Alpha and Beta Relate

Whenever base current flows in a bipolar transistor, we can calculate the beta in terms of the alpha with the formula

$$\beta = \alpha / (1 - \alpha)$$

and we can calculate the alpha in terms of the beta using the formula

$$\alpha = \beta / (1 + \beta)$$

With a little bit of algebra, we can derive these formulas from the fact that, at any instant in time, the collector current equals the emitter current minus the base current; that is,

$$I_C = I_E - I_B$$

"Real-World" Amplification

Let's look at Fig. 21-6 again. It's a graph of the collector current as a function of the base current (I_C versus I_B) for a hypothetical transistor. We can infer both H_{FE} and β from this graph. We can find H_{FE} at any particular point on the curve when we divide I_C by I_B at that point. Geometrically, the value of β at any given point on the curve equals the *slope* ("rise over run") of a *tangent line* at that point. On a two-dimensional coordinate grid, the tangent to a curve at a point constitutes the straight line that intersects the curve at that point without crossing the curve.

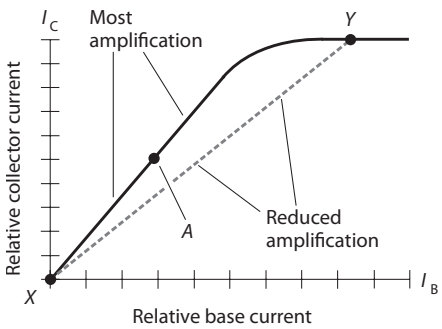
In Fig. 21-6, the tangent to the curve at point *B* appears as a dashed, straight line; the tangents to the curve at points *A* and *C* lie precisely along the curve (and, therefore, don't show up visually). As the slope of the line tangent to the curve increases, the value of β increases. Point *A* provides the highest value of β for this particular transistor, as long as we don't let the input signal get too strong. Points in the immediate vicinity of A provide good β values as well.

For small-signal amplification, point *A* in Fig. 21-6 represents a good bias level. Engineers would say that it's a favorable *operating point*. The β figure at point *B* is smaller (the curve slopes less steeply upward as we move toward the right) than the β figure at point *A*, so point *B* represents a less favorable operating point for small-signal amplification than point *A* does. At point *C*, we can surmise that $\beta = 0$ because the slope of the curve equals zero in that vicinity (it doesn't "rise" at all as we "run" toward the right). The transistor won't amplify weak signals when we bias it at point *C* or beyond.

Overdrive

Even when we bias a transistor so that it can produce the greatest possible current amplification (at or near point *A* in Fig. 21-6), we can encounter problems if we inject an AC input signal that's too strong. If the input-signal amplitude gets large enough, the transistor's operating point might move to or beyond point *B*, off the coordinate grid to the left, or both, during part of the signal cycle. In that case, the effective value of β will decrease. Figure 21-7 shows why this effect occurs. Points *X* and *Y* represent the instantaneous current extremes during the signal cycle in this particular case. Note that the slope of the line connecting points *X* and *Y* is less than the slope of the straight-line part of the curve at and near point *A*.

When our AC input signal is so strong that it drives the transistor to the extreme points *X* and *Y*, as shown in Fig. 21-7, a transistor amplifier introduces *distortion* into the signal, meaning that the



21-7 Excessive input reduces amplification.

output wave does not have the same shape as the input wave. We call this phenomenon *nonlinearity*. We can sometimes tolerate this condition, but often it's undesirable. Under most circumstances, we'll want our amplifier to remain *linear* (or to exhibit excellent *linearity*), meaning that the output wave has the same shape as (although probably stronger than) the input wave.

When the input signal to a transistor amplifier exceeds a certain critical maximum, we get a condition called *overdrive*. An overdriven transistor operates in or near saturation during part of the input signal cycle. Overdrive reduces the overall circuit efficiency, causes excessive collector current to flow, and can overheat the base-collector (B-C) junction. Sometimes overdrive can physically destroy a transistor.

Gain versus Frequency

A bipolar transistor exhibits an amplification factor (gain) that decreases as the signal frequency increases. Some bipolar transistors can amplify effectively at frequencies up to only a few megahertz. Other devices can work into the gigahertz range. The maximum operating frequency for a particular bipolar transistor depends on the capacitances of the P-N junctions inside the device. A low *junction capacitance* value translates into a high *maximum usable frequency*.

Expressions of Gain

You've learned about *current gain* expressed as a ratio. You'll also hear or read about *voltage gain* or *power gain* in amplifier circuits. You can express any gain figure as a ratio. For example, if you read that a circuit has a voltage gain of 15, then you know that the output signal voltage equals 15 times the input signal voltage. If someone tells you that the power gain of a circuit is 25, then you know that the output signal power equals 25 times the input signal power.

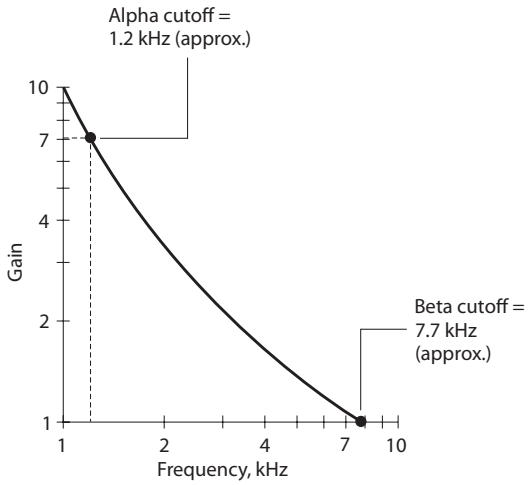
Alpha Cutoff

Suppose that we operate a bipolar transistor as a current amplifier, and we deliver an input signal to it at 1 kHz. Then we steadily increase the input-signal frequency so that the value of α declines. We define the *alpha cutoff frequency* of a bipolar transistor, symbolized f_α, as the frequency at which α decreases to 0.707 times its value at 1 kHz. (Don't confuse this use of the term "cutoff" with the state of "cutoff" that we get when we zero-bias or reverse-bias a transistor under no-signal conditions!) A transistor can have considerable gain at its alpha cutoff frequency. By looking at this specification for a particular transistor, we can get an idea of how rapidly it loses its ability to amplify as the frequency goes up.

Beta Cutoff Frequency

Imagine that we repeat the above-described variable-frequency experiment while watching β instead of α. We discover that β decreases as the frequency increases. We define the *beta cutoff frequency* (also called the *gain bandwidth product*) for a bipolar transistor, symbolized f_β or f_T, as the frequency at which β gets down to 1. If we try to make a transistor amplify above its beta cutoff frequency, we'll fail!

Figure 21-8 shows the alpha cutoff and beta cutoff frequencies for a hypothetical transistor on a graph of gain versus signal frequency. Note that the scales of this graph are not linear and the divisions are unevenly spaced. We call this type of plot a *log-log* graph because both scales are *logarithmic* rather than linear. The value on either scale increases in proportion to the *base-10 logarithm* of the distance from the origin, rather than varying in direct proportion to that distance.



21-8 Alpha cutoff and beta cutoff frequencies for a hypothetical transistor.

Common-Emitter Configuration

A bipolar transistor can be "wired up" in three general ways. We can ground the emitter for the signal, we can ground the base for the signal, or we can ground the collector for the signal. An often-used arrangement is the *common-emitter circuit*. "Common" means "grounded for the signal." Figure 21-9 shows the basic configuration.

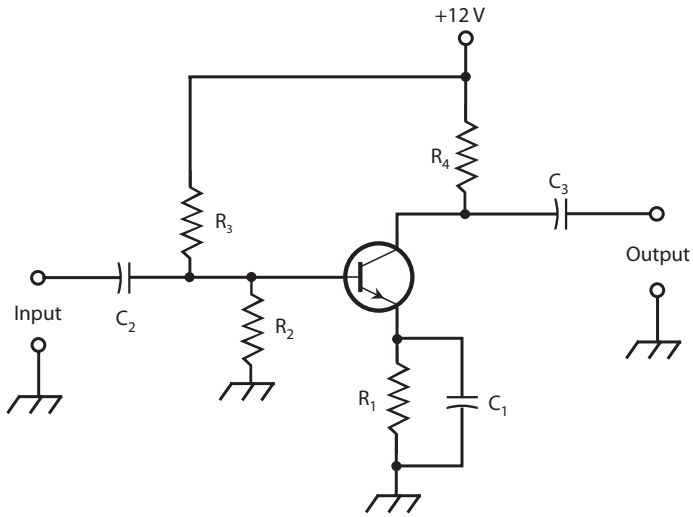
Even if a circuit point remains at ground potential for signals, it can have a significant DC voltage with respect to electrical ground. In the circuit shown, capacitor C_1 appears as a short circuit to the AC signal, so the emitter remains at *signal ground*. But resistor R_1 causes the emitter to attain and hold a certain positive DC voltage with respect to electrical ground (or a negative voltage, if we replace the NPN transistor with a PNP device). The exact DC voltage at the emitter depends on the resistance of R_1, and on the bias at the base. We set the DC base bias by adjusting the ratio of the values of resistors R_2 and R_3. The DC base bias can range from 0 V, or ground potential, to +12 V, which equals the power-supply voltage. Normally it's a couple of volts.

Capacitors C_2 and C_3 block DC to or from the external input and output circuits, while letting the AC signal pass. Resistor R_4 keeps the output signal from "shorting out" through the power supply. A signal enters the common-emitter circuit through C_2, so the signal causes the base current I_B to vary. The small fluctuations in I_B cause large changes in the collector current I_C. This current passes through resistor R_4, producing a fluctuating DC voltage across it. The AC component of this voltage passes unhindered through capacitor C_3 to the output.

The circuit of Fig. 21-9 represents the basis for many amplifier systems at all commonly encountered signal frequencies. The common-emitter configuration can produce more gain than any other arrangement. The output wave appears inverted (in phase opposition) with respect to the input wave. If the input signal constitutes a pure sine wave, then the common-emitter circuit shifts the signal phase by 180°.

Common-Base Configuration

As its name implies, the *common-base circuit* (Fig. 21-10) has the base at signal ground. The DC bias is the same as that for the common emitter circuit, but we apply the input signal at the emitter instead of at the base. This arrangement gives rise to fluctuations in the voltage across resistor R_1, causing

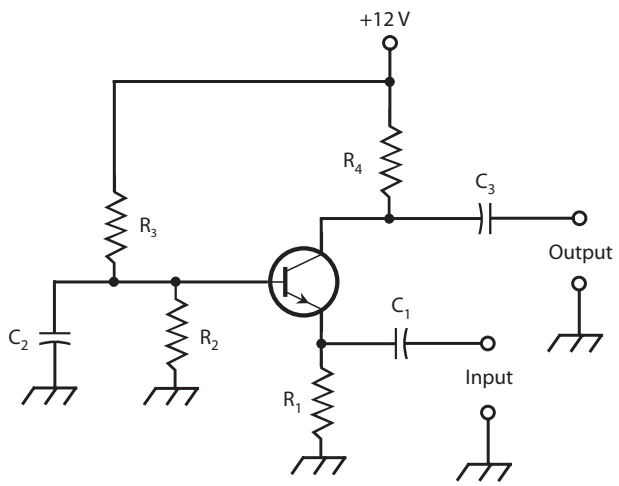


21-9 Common-emitter configuration for an NPN transistor circuit.

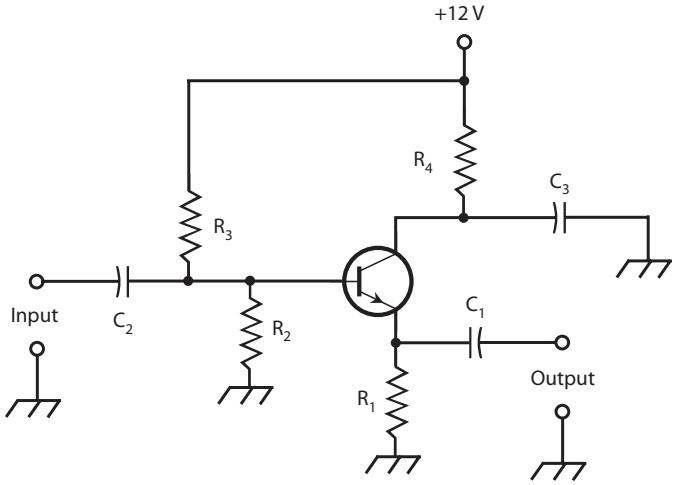
variations in I_B. These small current fluctuations produce large variations in the current through R_4. Therefore, amplification occurs. The output wave follows along in phase with the input wave.

The signal enters the transistor through capacitor C_1. Resistor R_1 keeps the input signal from "shorting out" to ground. Resistors R_2 and R_3 provide the base bias. Capacitor C_2 holds the base at signal ground. Resistor R_4 keeps the output signal from "shorting out" through the power supply. We get the output through capacitor C_3. The common-base circuit exhibits a relatively low input impedance, and provides somewhat less gain than a common-emitter circuit.

A common-base amplifier offers better *stability* than most common-emitter circuits do. By "better stability," we mean that the common-base circuit is less likely to *break into oscillation* (generate a signal of its own) as a result of amplifying some of its own output. The main reason for this "good behavior" is the low input impedance of the common-base circuit, which requires the input signal to offer significant power to drive the system into amplification. The common-base circuit isn't sensitive enough to get out of control easily!



21-10 Common-base configuration for an NPN transistor circuit.



21-11 Common-collector configuration, also known as an emitter-follower circuit, using an NPN transistor.

Sensitive amplifiers, such as optimally-biased common-emitter circuits, can pick up some of their own output as a result of stray capacitance between the input and output wires. This little bit of "signal leakage" can provide enough energy at the input to cause the whole circuit to "chase its own tail." When *positive* (in-phase) *feedback* gives rise to oscillation in an amplifier, we say that the amplifier suffers from *parasitic oscillation,* or *parasitics*, which can cause a radio transmitter to put out signals on unauthorized frequencies, or make a radio receiver stop working altogether.

Common-Collector Configuration

A *common-collector circuit* (Fig. 21-11) operates with the collector at signal ground. We apply the AC input signal at the transistor base, just as we do with the common-emitter circuit. The signal passes through C_2 onto the base. Resistors R_2 and R_3 provide the correct base bias. Resistor R_4 limits the current through the transistor. Capacitor C_3 keeps the collector at signal ground. Fluctuating DC flows through R_1, and a fluctuating voltage, therefore, appears across it. The AC part of this voltage passes through C_1 to the output. Because the output follows the emitter current, some engineers and technicians call this arrangement an *emitter-follower circuit*.

The output wave of a common-collector circuit appears exactly in phase with the input wave. The transistor exhibits a relatively high input impedance, while its output impedance remains low. For this reason, the common-collector circuit can take the place of a transformer when we want to match a high-impedance circuit to a low-impedance circuit or load. A well-designed emitter-follower circuit can function over a wider range of frequencies than a typical wirewound transformer can.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

22
CHAPTER

Field-Effect Transistors

THE BIPOLAR DEVICE ISN'T THE ONLY FORM OF TRANSISTOR THAT CAN SWITCH, AMPLIFY, OR OSCILLATE. A *field-effect transistor* (FET) can also do these things. Two main types of FET exist: the *junction FET* (JFET) and the *metal-oxide-semiconductor FET* (MOSFET). MOSFETs form the basis of digital electronics and are extremely well suited to the on-off switching that digital electronics requires. These are now the most commonly used type of transistor.

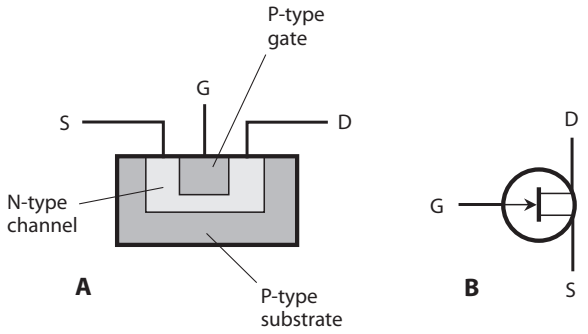
Principle of the JFET

In a JFET, an *electric field* within the device affects the amount of current that can flow through it. Charge carriers (electrons or holes) move from the *source* (S) electrode to the *drain* (D) electrode to produce a *drain current* I_D that normally equals the *source current*, I_S. The rate of flow of charge carriers—that is, the current—depends on the voltage at a control electrode called the *gate* (G). Fluctuations in *gate voltage* V_G cause changes in the current through the *channel*, the path that the charge carriers follow between the source and the drain. The current through the channel normally equals I_D. Under the right conditions, small fluctuations in V_G can cause large variations in I_D. This fluctuating drain current can, in turn, produce significant fluctuations in the voltage across an output resistance.

N-Channel versus P-Channel

Figure 22-1 is a simplified drawing of an *N-channel JFET* (at A) and its schematic symbol (at B). The N-type material forms the channel. The majority carriers in the channel are electrons, while the minority carriers are holes. We normally place the drain at a positive DC voltage with respect to the source, using an external power supply or battery. Just as NPN bipolar transistors are more commonly used than PNP transistors, N-channel MOSFETs are much more common than P-channel devices.

In an N-channel device, the gate consists of P-type material. Another section of P-type material, called the *substrate*, forms a boundary on the side of the channel opposite the gate. The voltage on the gate produces an electric field that interferes with the flow of charge carriers through the channel. As V_G becomes more negative, the electric field chokes an increasing amount of the current though the channel, so I_D decreases.



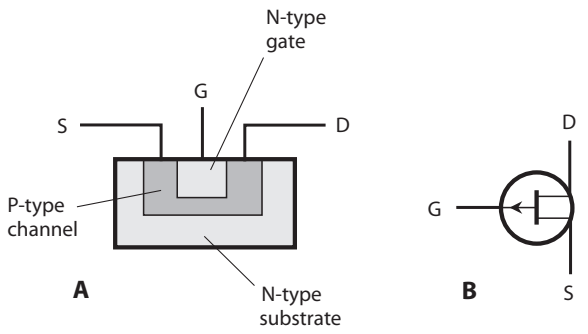
22-1 At A, a simplified structural drawing of an N-channel JFET. We identify the electrodes as S = source, G = gate, and D = drain. At B, the schematic symbol.

A *P-channel JFET* (Figs. 22-2A and B) has a channel of P-type semiconductor material. The majority charge carriers in the channel are holes, while the minority carriers are electrons. Using an external power supply or battery, we place the drain at a negative DC voltage with respect to the source. The more positive V_G gets, the more the electric field chokes off the current through the channel, and the smaller I_D becomes.

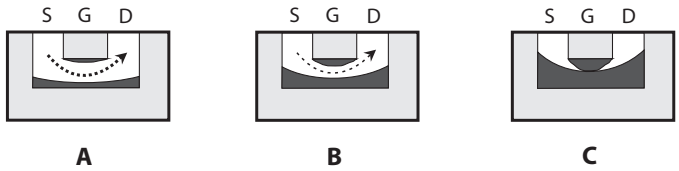
You can recognize the N-channel JFET in schematic diagrams by the presence of a small arrow pointing inward at the gate. You can recognize the P-channel JFET by the arrow pointing outward at the gate. Alternatively, you can tell the N-channel device from the P-channel device (in case the symbols lack arrows) by looking at the power-supply polarity. A positive drain indicates an N-channel JFET, and a negative drain indicates a P-channel JFET.

In electronic circuits, N-channel and P-channel devices can do the same kinds of things. The main difference lies in the power-supply or battery polarity. We can almost always replace an N-channel JFET with a P-channel JFET, reverse the polarity, and expect the circuit to work the same—assuming that the new device has the right specifications. Just as we'll find different kinds of bipolar transistors, so will we encounter various types of JFETs, each suited to a particular application. Some JFETs work well as weak-signal amplifiers and oscillators; others find their niche in power amplification; some work ideally as high-speed switches.

Field-effect transistors have certain advantages over bipolar transistors. Perhaps the most important asset arises from the fact that JFETs, in general, create less *internal noise* than bipolar transistors do. This property makes JFETs excellent for use as weak-signal amplifiers at very high or ultra-high radio frequencies; in general, they do better than bipolar transistors in this respect. Field-effect transistors exhibit high input impedance values—in some cases so high that they draw virtually no current, while nevertheless providing significant signal output.



22-2 At A, a simplified structural drawing of a P-channel JFET. At B, the schematic symbol.



22-3 At A, the depletion region (darkest area) is narrow, the channel (white area) is wide, and many charge carriers (heavy dashed line) flow. At B, the depletion region is wider, the channel is narrower, and fewer charge carriers flow. At C, the depletion region obstructs the channel, and no charge carriers flow.

Depletion and Pinchoff

The JFET works because the voltage at the gate generates an electric field that interferes, more or less, with the flow of charge carriers along the channel. Figure 22-3 shows a simplified rendition of the situation for an N-channel device.

As the drain voltage V_D increases, so does the drain current I_D, up to a certain level-off value. This property holds true as long as the gate voltage V_G remains constant, and doesn't get too large (negatively). However, as V_G becomes increasingly negative (Fig. 22-3A), a *depletion region* (shown as a solid dark area) begins to form in the channel. Charge carriers can't flow in the depletion region, so they must pass through a narrowed channel.

As V_G becomes more negative still, the depletion region widens, as shown in Fig. 22-3B. The channel narrows further, and the current through it declines some more. Ultimately, if the gate voltage becomes negative enough, the depletion region completely obstructs the flow of charge carriers, and the channel current drops to zero under no-signal conditions (Fig. 22-3C). We call this condition *pinchoff*. It's the equivalent of cutoff in a bipolar transistor.

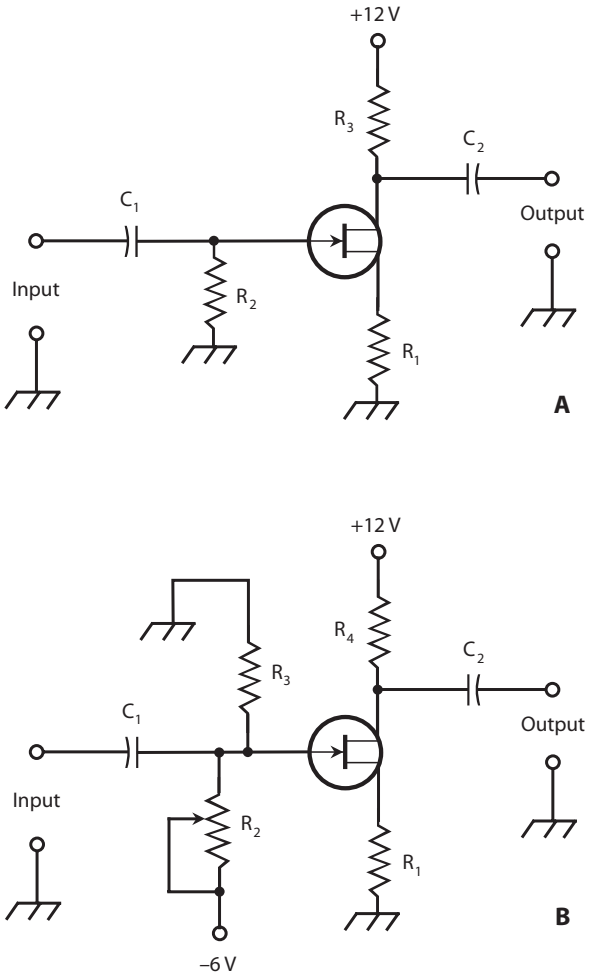
JFET Biasing

Figure 22-4 illustrates two biasing methods for N-channel JFET circuits. In Fig. 22-4A, we ground the gate through resistor R_2. The source resistor R_1 limits the current through the JFET. The drain current I_D flows through R_3, producing a voltage across R_3. That resistor also keeps the output signal from "shorting out" through the power supply or battery. The AC output signal passes through C_2 to the next circuit or load.

In Fig. 22-4B, we connect the gate through potentiometer R_2 to a source of voltage that's negative with respect to ground. Adjusting this potentiometer results in a variable negative gate voltage V_G between R_2 and R_3. Resistor R_1 limits the current through the JFET. The drain current I_D flows through R_4, producing a voltage across it. This resistor also keeps the output signal from "shorting out" through the power supply or battery. The AC output signal passes through C_2 to the next circuit or load.

In both circuits of Fig. 22-4, we provide the drain with a positive DC voltage relative to ground. For P-channel JFET circuits, simply reverse the polarities in Fig. 22-4, and replace the N-channel symbols with P-channel symbols.

Typical power-supply voltages in JFET circuits are comparable to those for bipolar transistor circuits. The voltage between the source and drain, abbreviated V_D, can range from about 3 V to 150 V DC; most often it's 6 to 12 V DC. The biasing arrangement in Fig. 22-4A works well for



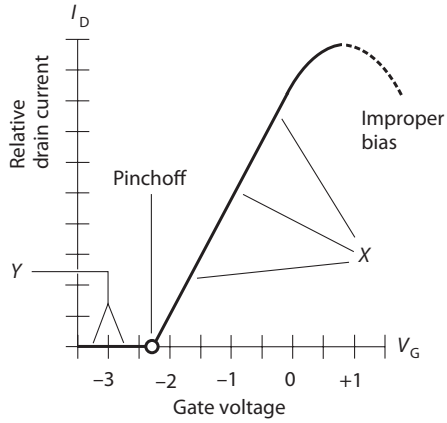
22-4 Two methods of biasing an N-channel JFET. At A, fixed gate bias; at B, variable gate bias.

weak-signal amplifiers, low-level amplifiers, and oscillators. The scheme at B works well in power amplifiers having substantial input signal amplitudes.

Amplification

Figure 22-5 shows a graph of I_D as a function of V_G for a hypothetical N-channel JFET. We assume that drain voltage V_D remains constant. When V_G is fairly large and negative, the JFET operates in a pinched-off state, so no current flows through the channel. As V_G gets less negative, the channel opens up, and I_D begins flowing. As V_G gets still less negative, the channel grows wider and I_D increases. As V_G approaches the point where the source-gate (S-G) junction (which constitutes a P-N junction) experiences forward breakover, the channel conducts as well as it possibly can; it's "wide open."

If V_G becomes positive enough so that the S-G junction goes past the forward-breakover point and conducts, some of the current in the channel "leaks out" through the gate. We rarely, if ever, want to see this state of affairs. We want the gate voltage to control the width of the channel and



22-5 Relative drain current (I_D) as a function of gate voltage (V_G) for a hypothetical N-channel JFET.

thereby control the current through it, but never to "suck" current out of it. Any current that flows out through the gate represents current that can't contribute to the output of the JFET.

Think of a JFET channel as a garden hose. When you want to reduce the flow of water at the output end of a hose, you can insert an adjustable valve somewhere along the length of the hose, or you can step down on the hose to pinch it narrower. You would not want to punch a hole in the hose to reduce the flow at the output because that action would let water go to waste.

The FET Amplifies Voltage

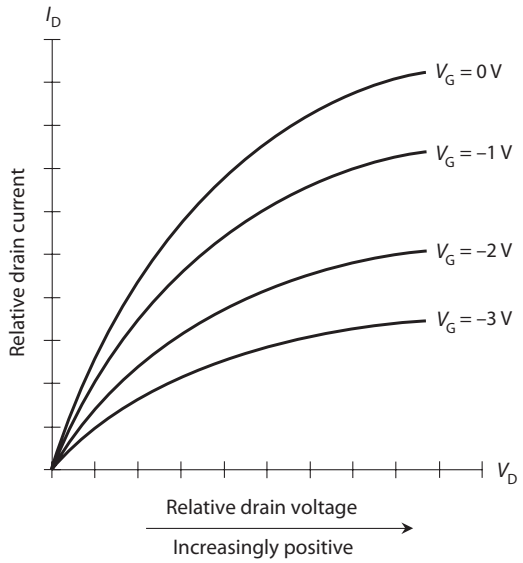
We can obtain the best amplification for weak signals when we set the no-signal gate voltage V_G so as to maximize the slope ("rise over run") of the drain-current versus gate-voltage (I_D versus V_G) curve. In Fig. 22-5, the range marked *X* shows the general region where this ideal condition exists. For power amplification, in which the JFET receives an input signal that's fairly strong to begin with, we'll often get the best results when we bias the JFET at or beyond pinchoff, in the range marked *Y*.

In either circuit shown in Fig. 22-4, I_D passes through the drain resistor. Small fluctuations in V_G cause large changes in I_D, and these variations in turn produce wide swings in the DC voltage across R_3 (in the circuit at A) or R_4 (in the circuit at B). The AC part of this voltage goes through capacitor C_2, and appears at the output as a signal of much greater AC voltage than that of the input signal at the gate. The JFET, therefore, acts as a *voltage amplifier*.

Drain Current versus Drain Voltage

Do you suspect that the drain current I_D, passing through the channel of a JFET, increases in a linear manner with increasing drain voltage V_D? This notion seems reasonable, but it's not what usually happens. Instead, I_D rises for a while as V_D increases steadily, and then I_D starts to level off as we increase V_D still more. We can plot I_D graphically as a function of V_D for various values of V_G under no-signal conditions. When we do that, we get a *family of characteristic curves*.

Figure 22-6 shows a family of characteristic curves for a hypothetical N-channel JFET. Engineers want to see graphs like this when choosing a JFET to serve in a specialized role, such as weak-signal amplification, oscillation, or power amplification. The graph of I_D versus V_G, one example of which appears in Fig. 22-5, is also an important specification that engineers consider. Characteristic curves portray DC behavior only; such curves are always derived under no-signal conditions.



22-6 A family of characteristic curves for a hypothetical N-channel JFET.

Transconductance

We learned in Chap. 22 that the beta of a bipolar transistor tells us how well the device can amplify a signal in a practical circuit. We can also call the bipolar beta the *dynamic current amplification*. In a JFET, engineers call its equivalent the *dynamic mutual conductance,* or *transconductance*.

Refer again to Fig. 22-5. Suppose that we set the gate voltage V_G at a certain value, resulting in a drain current I_D. If the gate voltage changes by a small amount dV_G, then the drain current will change by an increment dI_D. The transconductance g_{FS} equals the ratio of the change in the drain current to the change in the gate voltage, as follows:

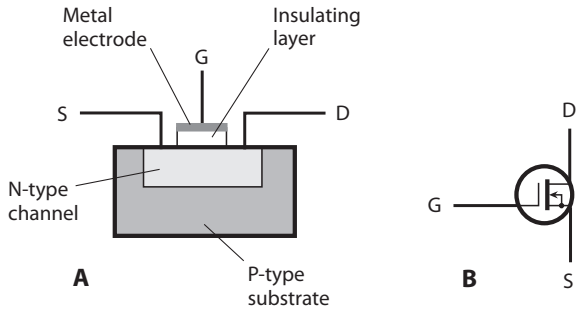
$$g_{FS} = dI_D / dV_G$$

The transconductance at any particular bias point translates to the slope ("rise over run") of a line tangent to the curve of Fig. 22-5 at that point.

As we can see from Fig. 22-5, the value of g_{FS} varies as we move along the curve. When we bias the JFET beyond pinchoff in the region marked *Y*, the slope of the curve equals zero; it's a horizontal line. In the range marked *Y*, we'll observe no fluctuation in I_D when V_G changes by small amounts. When there's a change in V_G, we'll see a change in I_D only when the channel conducts during at least part of the cycle of any input signal that we apply.

The region of greatest transconductance corresponds to the portion of the curve marked *X*, where the slope is the steepest. This region represents conditions under which we can derive the most gain from the device. Because this part of the curve constitutes a straight line, we can expect to get excellent linearity from any amplifier that we care to build with the JFET, provided that we keep the input signal from getting so strong that it drives the device outside of range *X* during any part of the cycle.

If we bias the JFET beyond the range marked *X*, the slope of the curve decreases, and we can't get as much amplification as we do in the range marked *X*. In addition, we can't expect the JFET to remain linear because the curve does not constitute a straight line in the "improper bias" range. If we keep on biasing the gate to greater and greater positive voltages, we get to the broken portion



22-7 At A, the functional structure of an N-channel MOSFET. At B, the schematic symbol. Electrodes are S = source, G = gate, and D = drain.

of the curve, arriving at the zone where the S-G junction goes past forward breakover and draws current out of the channel.

The MOSFET

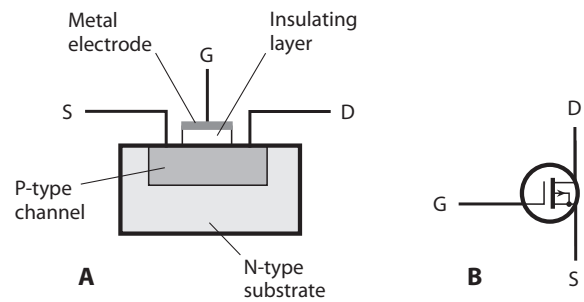
The acronym *MOSFET* (pronounced "*moss*-fet") stands for *metal-oxide-semiconductor field-effect transistor*. As with JFETs, two main types of MOSFET exist: The N-channel device and the P-channel device. Figure 22-7 shows a simplified cross-sectional drawing of an N-channel MOSFET along with the schematic symbol. The P-channel device and its symbol appear in Fig. 22-8.

Extremely High Input Impedance

When engineers originally developed the MOSFET, they called it an *insulated-gate field-effect transistor* or *IGFET*. That's still a pretty good term. The gate electrode is electrically insulated from the channel by a thin layer of dielectric material. As a result, the input impedance normally exceeds that of a JFET. In fact, the gate-to-source (G-S) resistance of a typical MOSFET compares favorably to the resistance of a capacitor of similar physical dimensions—it's "almost infinity"!

Because of its extreme G-S resistance, a MOSFET draws essentially no current, and therefore, essentially no power, from the input signal source. This property makes the MOSFET ideal for use in low-level and weak-signal amplifier circuits. Because the G-S "capacitor" is physically tiny, its capacitance is tiny as well, so the device can function quite well up to ultra-high radio frequencies (above 300 MHz).

MOSFETs are extremely sensitive to static electricity. This is not normally a problem for MOSFETs incorporated into integrated circuits because these include protection circuitry. But, when handling individual MOSFETs you should use an earthed wrist strap and earthed soldering iron.



22-8 At A, the functional structure of a P-channel MOSFET. At B, the schematic symbol.

Flexibility in Biasing

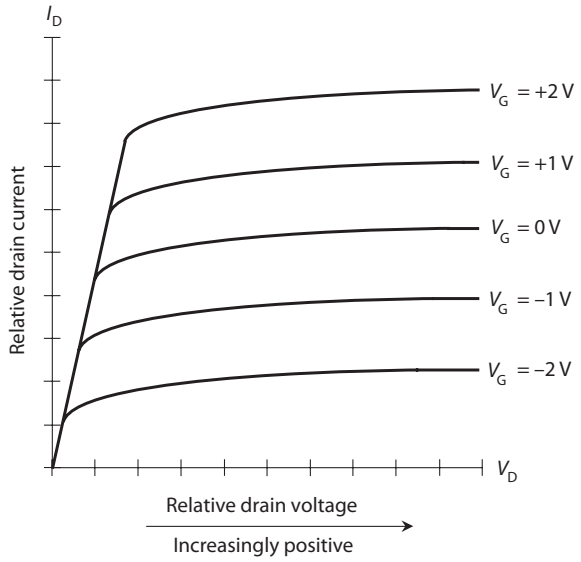
In electronic circuits, we can sometimes replace an N-channel JFET directly with an N-channel MOSFET, or a P-channel JFET with a P-channel MOSFET. However, in a few situations this simple substitution won't work. The characteristic curves for MOSFETs differ qualitatively from the characteristic curves for JFETs with similar amplifying characteristics.

The main difference between MOSFET behavior and JFET behavior arises from the fact that in a MOSFET, the source and the gate do not come together as a P-N junction. Instead, the two electrodes are physically separated by a "gap" of dielectric, so forward breakover can't occur. We can apply gate bias that's far more positive than +0.6 V to an N-channel MOSFET, or a lot more negative than −0.6 V to a P-channel MOSFET, and we'll never see current "leak" out through the gate as it would do in a JFET (unless, of course, we apply a voltage so great that *arcing* (sparking) takes place across the dielectric "gap").

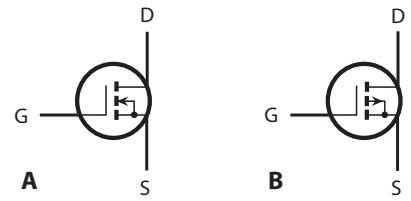
Figure 22-9 illustrates a family of characteristic curves for a hypothetical N-channel MOSFET under no-signal conditions. The horizontal axis portrays DC drain voltage, while the vertical axis portrays drain current. For any specific gate voltage, the drain current increases rapidly at first as we increase the drain voltage. However, as we continue to increase the drain voltage, the drain current rises at a slower rate, eventually leveling off. Once a particular I_D-versus-E_D curve has "flattened out," we can't increase the drain current any further by applying more DC drain voltage.

Depletion Mode versus Enhancement Mode

In the FET devices we've discussed so far, the channel normally conducts; as the depletion region grows, choking off the channel, the charge carriers must pass through a narrower and narrower path. We call this state of affairs the *depletion mode* for a field-effect transistor. A MOSFET can also function in the depletion mode. The drawings and schematic symbols of Figs. 22-7 and 22-8 show depletion-mode MOSFETs. The characteristic curves in Fig. 22-9 illustrate the behavior of a typical depletion-mode N-channel MOSFET. (To get the graphs for a P-channel device, reverse all the polarities.)



22-9 A family of characteristic curves for a hypothetical N-channel MOSFET.



22-10 Schematic symbols for enhancement-mode MOSFETs. At A, the N-channel device; at B, the P-channel device.

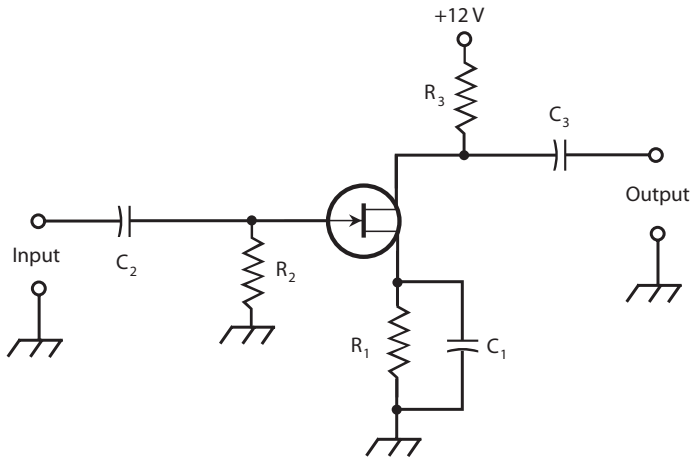
Metal-oxide semiconductor technology also allows an entirely different means of operation. An *enhancement-mode* MOSFET normally has a pinched-off channel. We must apply a bias voltage V_G to the gate so that a channel will form. If $V_G = 0$ in an enhancement-mode MOSFET, then $I_D = 0$ with no signal input. The enhancement-mode MOSFET, like its depletion-mode cousin, has an extremely low capacitance and high impedance between the gate and the channel. Enhancement-mode MOSFETs are in fact far more common than depletion mode.

In an N-channel enhancement-mode device, a *positive* voltage at the gate (with respect to the source) causes a conductive path to form in the channel. In the P-channel enhancement-mode device, we must apply a *negative* voltage at the gate in order to get the channel to conduct current. As the voltage increases, assuming the polarity is correct, the conductive channel grows wider, and the source-to-drain conductivity improves. This effect occurs, however, only up to a certain maximum current for a given constant DC drain voltage.

Figure 22-10 shows the schematic symbols for N-channel and P-channel enhancement-mode devices. Note that the vertical line is broken, rather than solid as it appears in the symbol for a depletion-mode MOSFET.

Common-Source Configuration

In a *common-source circuit*, we place the source at signal ground and apply the input signal to the gate. Figure 22-11 shows the general configuration for an N-channel JFET. The device could be an N-channel, depletion-mode MOSFET and the circuit arrangement would look the same. For an N-channel, enhancement-mode device, we would need an extra resistor between the gate and the



22-11 Common-source configuration. This diagram shows an N-channel JFET circuit.

positive power supply terminal. For P-channel devices, the schematics would be the same except that the supply would provide a negative, rather than a positive, voltage.

Capacitor C_1 and resistor R_1 place the source at signal ground, while elevating the source above ground for DC. The AC signal enters through C_2. Resistor R_2 adjusts the input impedance and provides bias for the gate. The AC signal passes out of the circuit through C_3. Resistor R_3 keeps the output signal from being shorted out through the power supply, while allowing for a positive voltage at the drain. This circuit can serve as the basic configuration for low-level RF amplifiers and oscillators.

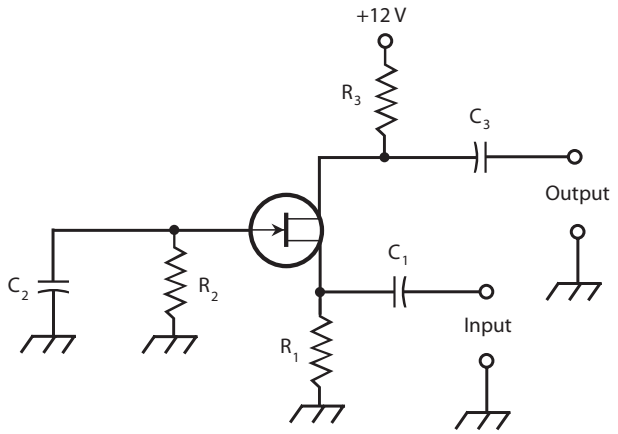
The common-source arrangement provides the greatest gain of the three FET circuit configurations. The output wave appears in phase opposition with respect to the input wave.

Common-Gate Configuration

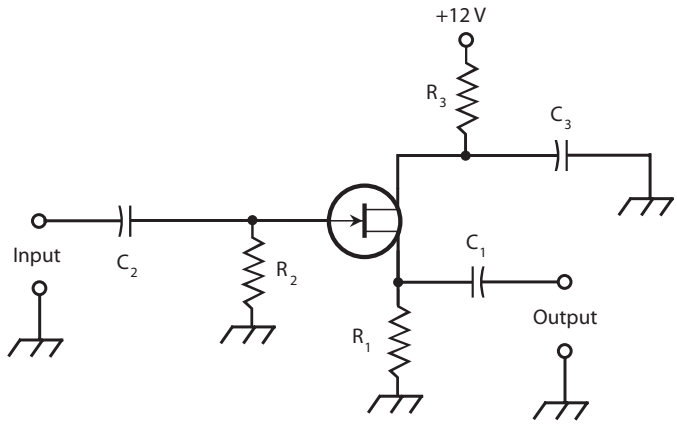
In the *common-gate circuit* (Fig. 22-12), we place the gate at signal ground and apply the input signal to the source. This diagram shows an N-channel JFET. For other types of FETs, the same considerations, as described above for the common-source circuit, apply. Enhancement-mode devices require an extra resistor between the gate and the positive supply terminal (or the negative terminal if the MOSFET is P-channel).

The DC bias for the common-gate circuit is basically the same as that for the common-source arrangement, but the signal follows a different path. The AC input signal enters through C_1. Resistor R_1 keeps the input signal from shorting out to ground. Gate bias is provided by R_1 and R_2. Capacitor C_2 places the gate at signal ground. (In some common-gate circuits, the gate goes directly to ground, and R_2 and C_2 are not used.) The output signal leaves the circuit through C_3. Resistor R_3 keeps the output signal from shorting out through the power supply, while still allowing the FET to receive the necessary DC voltage.

The common-gate arrangement produces less gain than its common-source counterpart, but it's far less likely to break into unwanted oscillation, making it a good choice for RF power-amplifier circuits. The output wave follows along in phase with the input wave.



22-12 Common-gate configuration. This diagram shows an N-channel JFET circuit.



22-13 Common-drain configuration, also known as a source follower. This diagram shows an N-channel JFET circuit.

Common-Drain Configuration

Figure 22-13 shows a *common-drain circuit*. We place the drain at signal ground. It is sometimes called a *source follower* because the output waveform follows the signal at the source. The FET is biased in the same way as for the common-source and common-gate circuits. In Fig. 22-13, we see an N-channel JFET, but any other kind of FET could be used, reversing the polarity for P-channel devices. Enhancement-mode MOSFETs would need an extra resistor between the gate and the positive supply terminal (or the negative terminal if the MOSFET is P-channel).

The input signal passes through C_2 to the gate. Resistors R_1 and R_2 provide gate bias. Resistor R_3 limits the current. Capacitor C_3 keeps the drain at signal ground. Fluctuating DC (the channel current) flows through R_1 as a result of the input signal; this causes a fluctuating DC voltage to appear across R_1. We take the output from the source, and its AC component passes through C_1.

The output wave of the common-drain circuit is in phase with the input wave. This scheme constitutes the FET equivalent of the bipolar common-collector (emitter follower) arrangement. The output impedance is low, so this type of circuit works quite well for matching a high input impedance to a low output or load impedance over a wide range of frequencies.

MOSFETs as Switches

MOSFETs can be made with very low on resistances (in the milliohm range) and very high on resistance (megaohms) making them ideal candidates for digital switches. The low on resistance means that a tiny MOSFET a few mm square can easily switch currents of several amperes and barely get warm.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

23
CHAPTER

Integrated Circuits

MOST *INTEGRATED CIRCUITS* (ICS), ALSO CALLED *CHIPS*, LOOK LIKE GRAY OR BLACK BOXES WITH metal terminals called *pins*. In schematic diagrams, engineers represent ICs as rectangles, usually with component designators printed inside, and with emerging lines (representing the pins in the actual device) leading to external components.

At the time of writing, the Digikey (one of the biggest component suppliers) lists around 720,000 different IC part numbers. There are specialized ICs for pretty much every electronic circuit or subsystem that you might want to design. The problem in finding an IC is normally that there are too many to choose from rather than there isn't one that does what you want.

An IC incorporates many transistors, capacitors, resistors, and diodes into a single device. There are ICs for pretty much any application you could think of, from audio amplifiers to entire microcontrollers.

Advantages of IC Technology

Integrated circuits have advantages over *discrete components* (individual transistors, diodes, capacitors, and resistors). The most important considerations follow.

Compactness

An obvious asset of IC design is economy of space. An IC is far more compact than an equivalent circuit made from discrete components. Integrated circuits allow for the construction of more sophisticated systems in smaller packages than discrete components do.

High Speed

The interconnections among internal IC components are physically tiny, making high switching speeds possible. As we increase the speed with which charge carriers can get from one component to another, we increase the number of computations that a system can do within a given span of time, and we reduce the time it takes for the system to perform complicated operations.

Low Power Consumption

Integrated circuits consume less power than equivalent discrete-component circuits. This advantage becomes a necessity in battery-operated systems. Because ICs use so little current, they produce less heat than their discrete-component equivalents, resulting in efficiency. The low-current feature also minimizes problems, such as frequency drift or intermittent failure, that can occur in equipment that gets hot with use.

Reliability

Integrated circuits fail less often, per component-hour of use, than systems built up from discrete components. The lower failure rate results from the fact that all component interconnections are sealed within the IC case, preventing the intrusion of dust, moisture, or corrosive gases. Therefore, ICs generally suffer less *downtime* (periods during which the equipment is out of service for repairs) than discrete-component systems do.

Price

Even ignoring the lower construction costs from using fewer components, an IC-based design will generally be a lot cheaper than its equivalent made from discrete components.

Modular Construction

Modern IC appliances use *modular construction*, in which individual ICs perform defined functions within a circuit board. The circuit board has a specific purpose. Repair technicians, using computers programmed with customized software, locate the faulty board, remove it, and replace it with a new one, getting the appliance back to the consumer in the shortest possible time.

Limitations of IC Technology

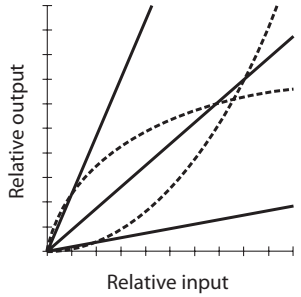
No technological advancement ever comes without a downside. Integrated circuits have limitations that engineers must consider when designing an electronic device or system.

Inductors Impractical

While some components are easy to fabricate onto chips, other components defy the IC manufacturing process. Inductors, except for components with extremely low values (in the nanohenry range), constitute a prime example. Devices using ICs must generally be designed to work with discrete inductors (coils) external to the ICs themselves. This constraint need not pose a problem, however. Resistance-capacitance (*RC*) circuits can do most things that inductance-capacitance (*LC*) circuits can do, and *RC* circuits can be etched onto an IC chip with no difficulty.

Unsuitable for Very High Power

The small size and low current consumption of ICs come with an inherent limitation. In general, when it comes to very high power applications above a few hundred watts, discrete power MOSFETs or other high-power transistors will be used. However, there are ICs designed to do most of the work and then drive external high-power transistors, keeping the component count low. As an example, 100 W plus audio power amplifier ICs are available that include tabs to which heatsinks can be attached.



23-1 In a linear IC, the relative output is a linear (straight-line) function of the relative input. The solid lines show examples of linear IC characteristics. The dashed curves show functions that aren't characteristic of properly operating linear ICs.

Linear ICs

A *linear IC* processes *analog signals* such as voices and music. The term "linear" arises from the fact that the amplification factor remains constant as the input amplitude varies. In technical terms, the output signal strength constitutes a *linear function* of the input signal strength, as shown by any of the three solid, straight lines in the graph of Fig. 23-1.

Operational Amplifier

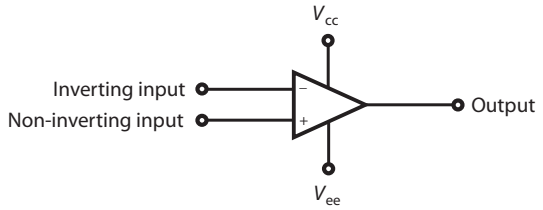
An *operational amplifier*, or *op amp*, is a specialized linear IC that consists of several transistors, resistors, diodes, and capacitors, all connected together so that high gain is possible over a wide range of frequencies. Some ICs contain two or more op amps, so you'll hear or read about *dual op amps* or *quad op amps*. Some ICs contain one or more op amps in addition to other circuits.

An op amp has two inputs, one *non-inverting* and one *inverting*, and one output. When a signal goes into the non-inverting input, the output wave emerges in phase coincidence with the input wave. When a signal goes into the inverting input, the output wave appears "upside-down" with respect to the input wave. An op amp has two power supply connections, one positive and one negative. Although op amps are also available that are designed to operate from a single supply voltage. The usual schematic symbol is a triangle (Fig. 23-2).

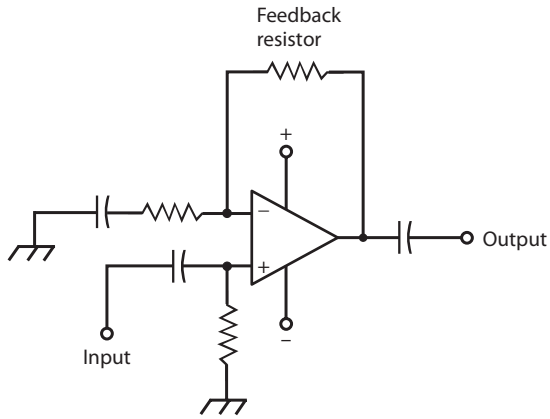
Op Amp Feedback and Gain

One or more external resistors determine the gain of an op amp. Normally, we place a resistor between the output and the inverting input to obtain a so-called *closed-loop configuration*. The feedback is negative (out of phase), causing the gain to remain lower than would hold true if no feedback existed. As we reduce the value of this resistor, the gain decreases because the negative feedback increases.

If we remove the feedback resistor, we get an *open-loop configuration*, in which the op amp produces its maximum rated gain. Figure 23-3 is a schematic diagram of a non-inverting closed-loop



23-2 Schematic symbol for an op amp. See text for discussion of connections.

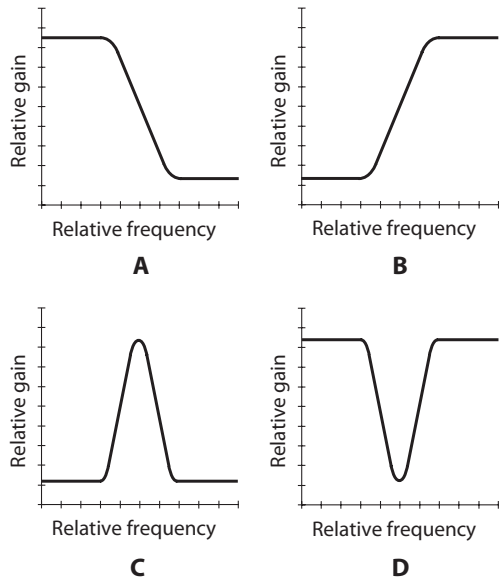


23-3 A closed-loop op amp circuit with negative feedback. If we remove the feedback resistor, we get an open-loop circuit.

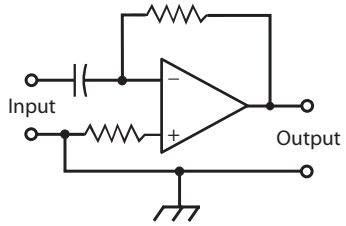
amplifier. Open-loop op amps sometimes exhibit instability, especially at low frequencies, breaking unexpectedly into oscillation. Open-loop op-amp circuits also generate significant internal noise, which can cause trouble in some applications.

If we install an *RC* combination in the feedback loop of an op amp, the gain depends on the input-signal frequency. Using specific values of resistance and capacitance, we can make a frequency-sensitive filter that provides any of four different characteristics, as shown in Fig. 23-4:

1. A *lowpass response* that favors low frequencies
2. A *highpass response* that favors high frequencies
3. A *resonant peak* that produces maximum gain at and near a single frequency
4. A *resonant notch* that produces maximum loss at and near a single frequency



23-4 Gain-versus-frequency response curves. At A, lowpass; at B, highpass; at C, resonant peak; at D, resonant notch.



23-5 A differentiator circuit that uses an op amp.

Op Amp Differentiator

A *differentiator* is a circuit whose instantaneous output amplitude varies in direct proportion to the rate at which the input amplitude changes. In the mathematical sense, the circuit mathematically *differentiates* the input signal wave function. Op amps lend themselves well to use as differentiators. Figure 23-5 shows an example.

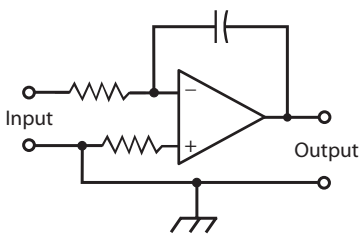
When the input to a differentiator is a constant DC voltage, the output equals zero (no signal). When the instantaneous input amplitude increases, the output is a positive DC voltage. When the instantaneous input amplitude decreases, the output is a negative DC voltage. If the input amplitude fluctuates periodically (a sine wave, for example), the output voltage varies according to the *instantaneous rate of change* (the mathematical *derivative*) of the instantaneous input-signal amplitude. We, therefore, observe an output signal with the same frequency as that of the input signal, although the waveform might differ.

In a differentiator circuit, a pure sine-wave input produces a pure sine-wave output, but the phase is shifted 90° to the left (¼ cycle earlier in time). Complicated input waveforms produce a wide variety of output waveforms.

Op Amp Integrator

An *integrator* is a circuit whose instantaneous output amplitude is proportional to the accumulated input signal amplitude as a function of time. The circuit mathematically *integrates* the input signal. In theory, the function of an integrator is the inverse, or opposite, of the function of a differentiator. Figure 23-6 shows how we can configure an op amp to obtain an integrator.

If we supply an integrator with an input signal waveform that fluctuates periodically, the output voltage varies according to the *integral*, or *antiderivative*, of the input voltage. This process yields an output signal with the same frequency as that of the input signal, although the waveform often differs. A pure sine-wave input produces a pure sine wave output, but the phase is shifted 90° to the right (¼ cycle later in time). Complex input waveforms can produce many types of output waveforms.



23-6 An integrator circuit that uses an op amp.

The performance of a practical integrator differs in one important way from the operation of a theoretically ideal integrator. If the mathematical integral of an input function yields an endlessly increasing output function, the actual output voltage rises to a certain maximum, either positive or negative, and then stays there. Obviously, we can't get an output voltage that increases forever without limit! The maximum voltage at the output of a real-world integrator never exceeds the power-supply or battery voltage.

Op Amp Buffer

An op amp naturally has a high input impedance and a low output impedance. If configured as a unity gain buffer, as shown in Fig. 23-7, then the output voltage will track the input voltage in a similar manner to using an NPN bipolar transistor in a common emitter configuration. But, in the case of the op-amp buffer there is no offset due to the base-emitter voltage. In the example of Fig. 23-7, a single supply op amp is used.

Linear Voltage Regulator

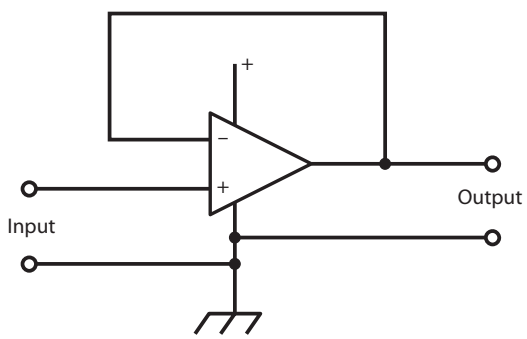
A *voltage regulator IC* acts to control the output voltage of a power supply. This feature is important with precision electronic equipment. Voltage-regulator ICs exist with various voltage and current ratings. Typical voltage regulator ICs have three terminals, ground, input, and output. Linear voltage regulators are very easy to use, but get hot when asked to regulate higher currents. In such situations, switching regulators are now more commonly used for their greater efficiency.

Timer

A *timer IC* is a specialized oscillator that produces a delayed output. We can tailor the delay time to suit the needs of a particular device. The delay is generated by counting the number of oscillator pulses; the length of the delay can be adjusted by means of external resistors and capacitors. Timer ICs find widespread application in digital frequency counters, where a precise time interval or "window" must be provided.

Comparator

A *comparator IC* has two inputs. It literally compares the voltages at the two inputs, which we call input A and input B. If the voltage at input A significantly exceeds the voltage at input B, the output equals about +5 V, giving us the logic 1, or high, state. If the voltage at input A is less than or



23-7 A unity gain buffer using an op-amp.

equal to the voltage at input B, we get an output voltage of about +2 V or less, yielding the logic 0, or low, state.

Voltage comparators are available for a variety of applications. Some can switch between low and high states at a rapid rate, while others are slow. Some have low input impedance, and others exhibit high input impedance. Some are intended for AF or low-frequency RF use; others work well in video or high-frequency RF applications.

Voltage comparators can actuate, or *trigger*, other devices, such as relays, alarms, and electronic switching circuits.

Digital ICs

A *digital IC* operates using two discrete states: high (logic 1) and low (logic 0). Digital ICs contain massive arrays of logic *gates* that perform logical operations at high speed.

Complementary-metal-oxide-semiconductor (CMOS) *logic*, pronounced "SEA-moss logic," employs both N type and P type silicon on a single chip—analogous to using both N-channel and P-channel MOSFETs in a discrete-component circuit. The advantages of CMOS technology include extremely low current drain, high operating speed, and immunity to external noise. Most digital ICs are built using CMOS technology.

All forms of MOS logic ICs require careful handling to prevent destruction by electrostatic discharges. The precautions are the same as those for handling discrete MOSFETs. All personnel who work with MOS ICs should "ground themselves" by wearing metal wrist straps connected to electrical ground, and by ensuring that the humidity in the lab does not get too low. When MOS ICs are stored, the pins should be pushed into conductive foam specifically manufactured for that purpose.

Component Density

In digital ICs, the number of transistors per chip is called the *component density*. It keeps increasing year by year. There's a practical limit to component density (imposed by the physical volumes of individual atoms), although some engineers think they'll be able to make single atoms perform multiple operations someday. In general, as IC component density increases, so does operating speed.

Small-Scale Integration

In *small-scale integration* (SSI), there are fewer than 10 transistors on a chip. These devices can carry the largest currents of any IC type because low component density translates into relatively large volume and mass per component. Small-scale integration finds application in voltage regulators and other moderate-power systems.

Medium-Scale Integration

In *medium-scale integration* (MSI), we have 10 to 100 transistors per chip. This density allows for considerable miniaturization, but does not constitute a high level of component density, relatively speaking, these days. An advantage of MSI is the fact that individual logic gates can carry fairly large currents in some applications. Both bipolar and MOS technologies can be adapted to MSI.

Large-Scale Integration

In *large-scale integration* (LSI), we have 100 to 1,000 transistors per semiconductor chip, a full *order of magnitude* (a factor of 10 times) more dense than MSI. Electronic wristwatches, single-chip calculators, *microcomputers*, and *microcontrollers* are examples of devices using LSI ICs.

Very-Large-Scale Integration

Very-large-scale integration (VLSI) devices have from 1,000 to 1,000,000 transistors per chip, up to three orders of magnitude more dense than LSI. High-end microcomputers, microcontrollers, and *memory chips* are made using VLSI.

Ultra-Large-Scale Integration

You might sometimes hear of *ultra-large-scale integration* (ULSI). Devices of this kind have more than 1,000,000 transistors per chip. The principal uses for ULSI technology include high-level computing, *supercomputing*, *robotics*, and *artificial intelligence* (AI). We should expect to see expanded use of ULSI technology in the future, particularly in medical devices.

IC Memory

Binary digital data, in the form of high and low states (logic 1 and 0), can be stored in memory chips that take a wide variety of physical forms. Some IC memory chips require a continuous source of backup voltage or they'll lose their data. Others can hold the data in the absence of backup voltage, in some cases for months or years. In electronic devices, we encounter two main types of memory: *random-access* and *read-only*.

Random-Access Memory

A *random-access memory* (RAM) chip stores binary data in *arrays*. The data can be *addressed* (selected) from anywhere in the matrix. Data is easily changed and stored back in RAM, in whole or in part. Engineers sometimes call a RAM chip *read/write memory*.

An example of the use of RAM is a word-processing computer file that you're actively working on. This paragraph, this chapter, and in fact, the whole text of this book was written in semiconductor RAM in small sections before being incrementally stored on the computer hard drive, and ultimately on external media.

There are two major categories of RAM: *dynamic RAM* (DRAM) and *static RAM* (SRAM). A DRAM chip contains transistors and capacitors, and data is stored as charges on the capacitors. The charge must be replenished frequently, or it will vanish through discharge. Replenishing is done automatically several hundred times per second. An SRAM chip uses a flip-flop to store the data. This arrangement gets rid of the need for constant replenishing of charge, but SRAM ICs require more elements than DRAM chips to store a given amount of data.

With any RAM chip, the data will vanish when we remove power unless we provide a means of *memory backup*. This is why such memories are called "volatile." The most common memory-backup scheme involves the use of a small cell or battery with a long shelf life. Modern IC memories need so little current to store their data that a *backup battery* lasts as long in the circuit as it would last sitting on a shelf doing nothing!

Non-volatile Memory

An *electrically erasable programmable read-only memory* (EEPROM) chip is a ROM device that we can reprogram by following a certain procedure. It's more difficult to rewrite data in EPROM than in RAM; It is much slower to erase or change a memory cell than RAM.

Flash memory is similar to EEPROM, in that it is non-volatile and is often used to contain the program code of microcontrollers.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

24
CHAPTER

Power Supplies

A *POWER SUPPLY* CONVERTS UTILITY AC TO PURE DC OF THE SORT WE GET FROM AN ELECTROCHEMICAL or solar battery. In this chapter, we will start by looking at the classical power supply design using a power transformer, rectification, smoothing, and finally regulation. But, then we will go on to look at how the compact, lightweight, but high power switched-mode power supplies work.

Power Transformers

We can categorize power transformers in two general ways: step-down or step-up. As you remember, the output, or secondary, voltage of a step-down transformer is lower than the input, or primary, voltage. The reverse holds true for a step-up transformer, where the output voltage exceeds the input voltage.

Most electronic devices need only a few volts to function. The power supplies for such equipment use step-down power transformers, with the primary windings connected to the utility AC outlets. The transformer's physical size and mass depend on the amount of current that we expect it to deliver. Some devices need only a small current at a low voltage. The transformer in a radio receiver, for example, can be physically small. A large amateur radio transmitter or hi-fi amplifier needs more current. The secondary windings of a transformer intended for those applications must consist of heavy-gauge wire, and the cores must have enough bulk to contain the large amounts of magnetic flux that the coils generate.

Transformer Ratings

Engineers rate power transformers according to the maximum output voltage and current they deliver. For a given unit, we'll often read or hear about the *volt-ampere* (VA) capacity, which equals the product of the nominal output voltage and maximum deliverable current.

A transformer with 12-V output, capable of providing up to 10 A of current, has a VA capacity of 12 V × 10 A, or 120 VA. The nature of power-supply filtering, which we'll learn about later in this chapter, makes it necessary for the power-transformer VA rating to significantly exceed the actual power in watts that the load consumes.

A high-quality, rugged power transformer, capable of providing the necessary currents and/or voltages, constitutes an integral and critical part of a well-engineered power supply. The transformer is usually the most expensive power-supply component to replace if it burns out, so we must always choose a transformer with the appropriate specifications when designing and building a power supply.

Rectifier Diodes

Rectifier diodes are available in various sizes, intended for different purposes. Most rectifier diodes comprise silicon semiconductor materials, so we call them *silicon rectifiers*. Some rectifier diodes are fabricated from selenium, so we call them *selenium rectifiers*. When we work with power-supply diodes, we must pay close attention to two specifications: the *average forward current* (I_o) rating and the *peak inverse voltage* (PIV) rating.

Average Forward Current

Electric current always produces some heat because every material medium offers at least a little bit of resistance. If we drive too much current through a diode, the resulting heat will destroy the P-N junction. When designing a power supply, we must use diodes with an I_o rating of at least 1.5 times the expected average DC forward current. If this current is 4.0 A, for example, the rectifier diodes should be rated at $I_o = 6.0$ A or more.

In a power supply that uses one or more rectifier diodes, the I_o flows through each individual diode. The current drawn by the *load* can, and usually does, differ from I_o. Also, note that I_o represents an *average* figure. The *instantaneous* forward current is another thing entirely, and can range up to 15 or 20 times I_o depending on the nature of the filtering circuit.

Some diodes have *heatsinks* to help carry heat away from the P-N junction. We can recognize a selenium rectifier by the appearance of its heatsink, which looks like a miniature version of an old-fashioned baseboard radiator built around a steam pipe.

We can connect two or more identical rectifier diodes in parallel to increase the current rating over that of an individual diode. When we do this, we should connect a small-value resistor in series with each diode to equalize the current. Every one of these resistors should have a value such that the voltage drop across it equals roughly 1 V under normal operating conditions.

Peak Inverse Voltage

The PIV rating of a diode tells us the maximum instantaneous reverse-bias voltage that it can withstand without avalanche breakdown. A well-designed power supply has diodes whose PIV ratings significantly exceed the peak AC input voltage. If the PIV rating is not great enough, the diode or diodes in a supply conduct current for part of the reverse cycle. This conduction degrades the efficiency of the power supply because the reverse current "bucks" the forward current.

We can connect two or more identical rectifier diodes in series to obtain a higher PIV rating than a single diode alone can provide. We'll sometimes see engineers take advantage of this technique when designing high-voltage power supplies, such as those needed for tube-type power amplifiers. High-value resistors, of about 500 ohms for each peak-inverse volt, are placed in parallel with each diode to distribute the reverse bias equally among the diodes. In addition, each diode is shunted by (connected in parallel with) a capacitor of approximately 0.01 μF.

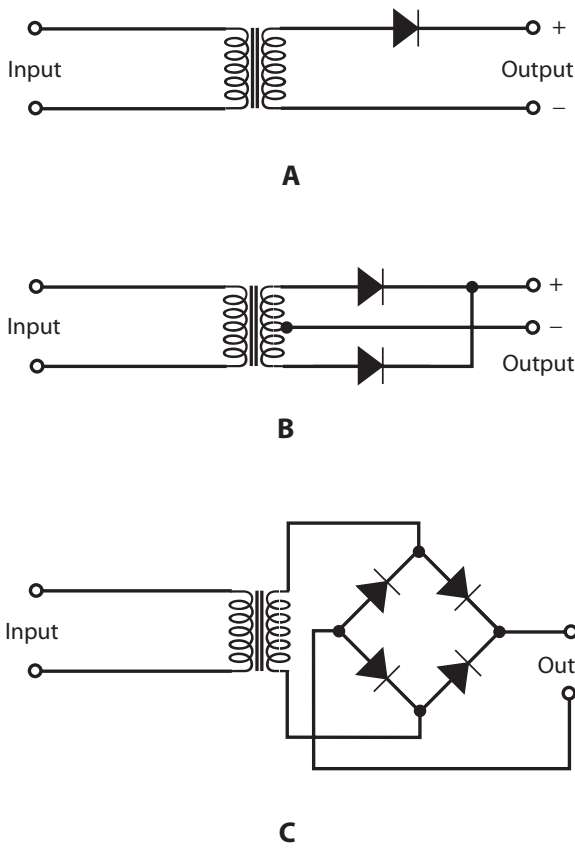
Half-Wave Circuit

The simplest rectifier circuit, called the *half-wave rectifier* (Fig. 24-1A), has a single diode that "chops off" half of the AC cycle. The effective output voltage from a power supply that uses a half-wave rectifier is much less than the peak transformer output voltage, as shown in Fig. 24-2A. The peak voltage across the diode in the reverse direction can range up to 2.8 times the applied RMS AC voltage.

Most engineers like to use diodes whose PIV ratings equal at least 1.5 times the maximum expected peak reverse voltage. Therefore, in a half-wave rectifier circuit, the diodes should be rated for at least 2.8 × 1.5, or 4.2, times the RMS AC voltage that appears across the secondary winding of the power transformer.

Half-wave rectification has shortcomings. First, the output is difficult to filter. Second, the output voltage can drop considerably when the supply must deliver high current. Third, half-wave rectification puts a strain on the transformer and diodes because it *pumps* them, meaning that the circuit works the diodes hard during half the AC cycle and lets them "loaf" during the other half.

Half-wave rectification will usually suffice when we want to design a power supply that will never have to deliver much current, or when the voltage can vary without affecting the behavior of



24-1 At A, a half-wave rectifier circuit. At B, a full-wave center-tap rectifier circuit. At C, a full-wave bridge rectifier circuit.

the equipment connected to it. The main advantage of a half-wave circuit is the fact that it costs less than more sophisticated circuits because it contains fewer parts.

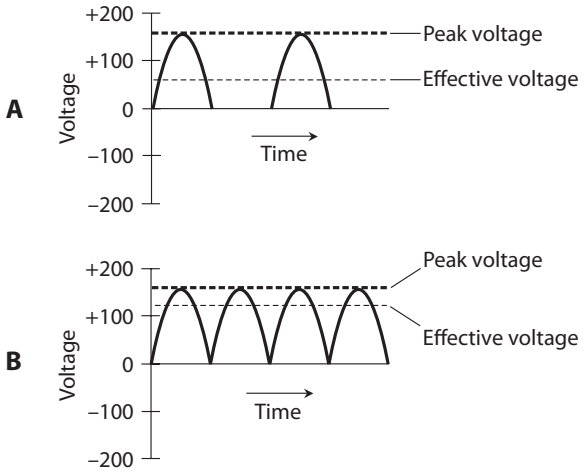
Full-Wave Center-Tap Circuit

We can take advantage of both halves of the AC cycle by means of *full-wave rectification*. A *full-wave center-tap rectifier* has a transformer with a connection called a *tap* at the center of the secondary winding (Fig. 24-1B). The tap connects directly to *electrical ground*, also called *chassis ground*. This arrangement produces voltages and currents at the ends of the secondary winding that oppose each other in phase. These two AC waves are individually half-wave rectified, cutting off one half of the cycle and then the other, repeatedly.

The effective output voltage from a power supply with a full-wave, center-tap rectifier (Fig. 24-2B) is greater, relative to the peak voltage, than the effective output voltage with the half-wave rectifier. The PIV across the diodes can, nevertheless, range up to 2.8 times the applied RMS AC voltage. Therefore, the diodes should have a PIV rating of at least 4.2 times the applied RMS AC voltage to ensure that they won't experience avalanche breakdown during any portion of the wave cycle.

The output of a full-wave center-tap rectifier is easier to filter than that of a half-wave rectifier, because the frequency of the pulsations in the DC (known as the *ripple frequency*) from a full-wave rectifier equals twice the ripple frequency of the pulsating DC from a half-wave rectifier, assuming identical AC input frequency in either situation. If you compare Fig. 24-2B with Fig. 24-2A, you will see that the full-wave-rectifier output is "closer to pure DC" than the half-wave rectifier output. Another advantage of a full-wave center-tap rectifier is the fact that it treats the transformer and diodes more "gently" than a half-wave rectifier does.

When we connect a load to the output of a power supply that uses a full-wave center-tap rectifier circuit, the voltage drops less than it does with the same load connected to a half-wave supply. However, because the transformer is more sophisticated, the full-wave center-tap circuit costs more than a half-wave circuit that delivers the same output voltage at the same rated maximum current.



24-2 At A, the output of a half-wave rectifier. At B, the output of a full-wave rectifier. Note the difference in how the effective voltages compare with the peak voltages.

Full-Wave Bridge Circuit

We can get full-wave rectification using a circuit known as a *full-wave bridge rectifier*, often called simply a *bridge*. Figure 24-1C shows a schematic diagram of a typical full-wave bridge circuit. The output waveform looks the same as the waveform that we get from the output of a full-wave center-tap circuit (Fig. 24-2B).

The effective output voltage from a power supply that uses a full-wave bridge rectifier is somewhat less than the peak transformer output voltage, as shown in Fig. 24-2B. The peak voltage across the diodes in the reverse direction equals about 1.4 times the applied RMS AC voltage. Therefore, each diode needs to have a PIV rating of at least 1.4×1.5, or 2.1, times the RMS AC voltage that appears at the transformer secondary in order to prevent avalanche breakdown from occurring during any part of the cycle.

The bridge circuit does not require a center-tapped transformer secondary. It uses the entire secondary winding on both halves of the wave cycle, so it makes more efficient use of the transformer than the full-wave center-tap circuit does. The bridge circuit also places less strain on the individual diodes than a half-wave or full-wave center-tap circuit does.

Power-Supply Smoothing

Most DC-powered devices need something more "pure" than the rough, pulsating DC that comes straight out of a rectifier circuit. We can eliminate, or at least minimize, the pulsations (ripple) in the rectifier output using a *power-supply filter*.

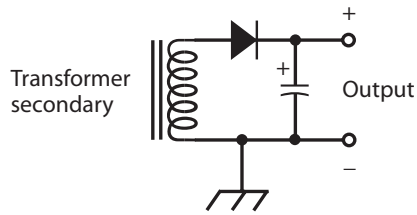
Capacitors Alone

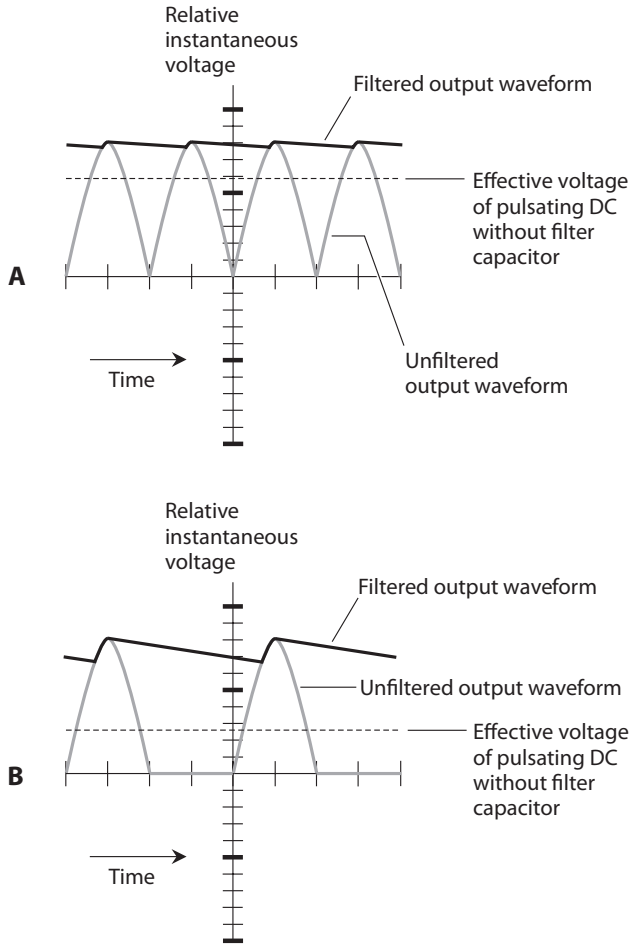
The simplest power-supply filter consists of one or more large-value capacitors, connected in parallel with the rectifier output, as shown in Fig. 24-3. A good component for this purpose is an *electrolytic capacitor*. This type of capacitor is *polarized*, meaning that we must connect it in the correct direction. Any given electrolytic capacitor also has a certain maximum rated working voltage. Pay attention to these particulars if you ever work with electrolytic capacitors!

Filter capacitors function by "trying" to maintain the DC voltage at its peak level, as shown in Fig. 24-4. The output of a full-wave rectifier (drawing A) lends itself more readily to this process than the output of a half-wave rectifier (drawing B). With a full-wave rectifier receiving a 60-Hz AC electrical input, the ripple frequency equals 120 Hz, but with a half-wave rectifier it's only 60 Hz. The filter capacitor, therefore, gets recharged twice as often with a full-wave rectifier as it does with a half-wave rectifier.

The two illustrations in Fig. 24-4 reveal the reason why a full-wave rectifier can produce more "pure" DC (for a given peak voltage and filter capacitance) than a half-wave rectifier can. The full-wave

24-3 We can use a large-value capacitor all by itself as a power-supply filter.





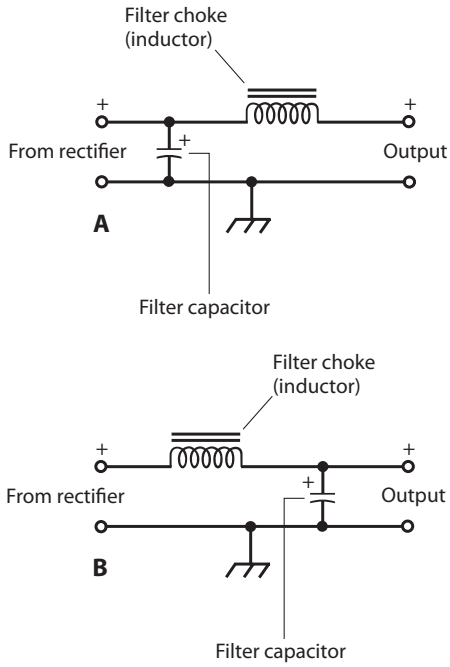
24-4 Filtering of ripple from a full-wave rectifier (A) and from a half-wave rectifier (B).

output gives the capacitor a "less bumpy ride," while the half-wave output lets the capacitor discharge more between each "refresh pulse."

Capacitors and Chokes

We can obtain enhanced ripple suppression by placing a large-value inductor in series with the rectifier output along with a large-value capacitor in parallel. When an inductor serves in this role, we call it a *filter choke*.

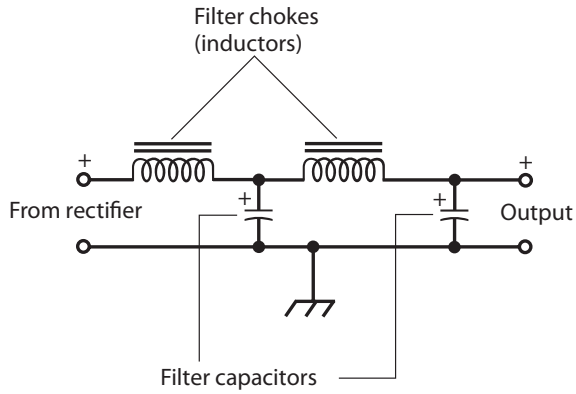
In a filter that uses a capacitor and an inductor, we can place the capacitor on the rectifier side of the choke to construct a *capacitor-input filter* (Fig. 24-5A). If we locate the filter choke on the rectifier side of the capacitor, we get a *choke-input filter* (Fig. 24-5B). Capacitor-input filtering works well when a power supply does not have to deliver much current. The output voltage, when the load is "light" (not much current is drawn), is higher with a capacitor-input filter than with a choke-input filter having identical input. If the supply needs to deliver large or variable amounts of current, however, a choke-input filter yields better performance because it produces a more stable output voltage for a wide variety of loads.



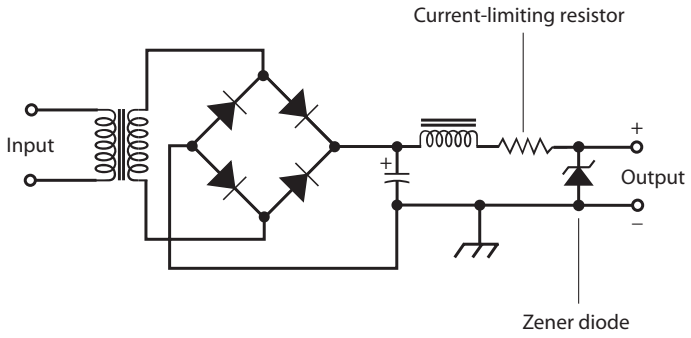
24-5 At A, a capacitor-input filter. At B, a choke-input filter.

If the DC output of a power supply must contain an absolute minimum of ripple, we can connect two or three capacitor/choke pairs in *cascade*. Figure 24-6 shows an example. Each inductor/capacitor pair constitutes a *section* of the filter. Multi-section filters can consist of capacitor-input or choke-input sections; the two types should never be mixed in the same filter.

In the example of Fig. 24-6, both capacitor/choke pairs are called *L sections* (not because of inductance, but because of their geometric shapes in the schematic diagram). If we eliminate the second capacitor, the filter becomes a *T section* (the inductors form the top of the T, and the capacitor forms the stem). If we move the second capacitor to the input and remove the second choke, the filter becomes a *pi section* (the capacitors form the pillars of an uppercase Greek letter pi, and the inductor forms the top).



24-6 Two choke-input filter sections in cascade.



24-7 A power supply with a Zener-diode voltage regulator in the output.

Voltage Regulation

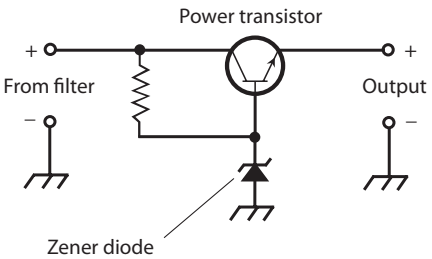
If we connect a Zener diode in parallel with the output of a power supply so that the component receives a reverse bias, the diode limits the output voltage. The diode must have an adequate power rating to prevent it from burning out. In addition, we must connect a resistance in series with it to limit the current. The limiting voltage depends on the particular Zener diode used. Zener diodes are available for any reasonable power-supply voltage.

Figure 24-7 is a diagram of a full-wave bridge DC power supply including a Zener diode for voltage regulation. Note the direction in which we connect the Zener diode: the arrow points from minus to plus. This polarity goes contrary to the orientation of a rectifier diode. We must take care to connect a Zener diode with the correct polarity, or it will burn out as soon as we apply power to the circuit!

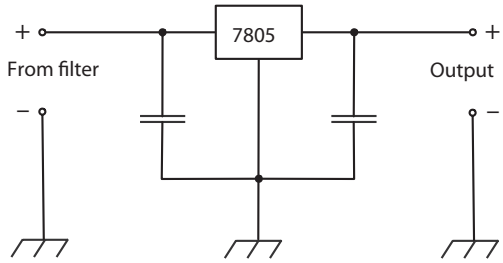
A simple Zener-diode voltage regulator, such as the one shown in Fig. 24-7, does not function effectively when we use the power supply with equipment that draws high current. The problem arises because the series resistor, essential to prevent destruction of the diode, creates a significant voltage drop when it carries more than a small amount of current. When we expect that a power supply will have to deliver a lot of current, we can employ a *power transistor* along with the Zener diode to obtain voltage regulation. Figure 24-8 shows an example. In this circuit, the resistor ensures proper operation of the transistor without causing a drop in the output voltage under high-current conditions.

Linear Voltage Regulator ICs

As with many common circuit functions, the problem of voltage regulation is now mostly dealt with using special purpose voltage regulator Integrated Circuits (ICs).



24-8 A voltage-regulator circuit using a Zener diode and a transistor.



24-9 An example of the use of a 7805 voltage regulator IC.

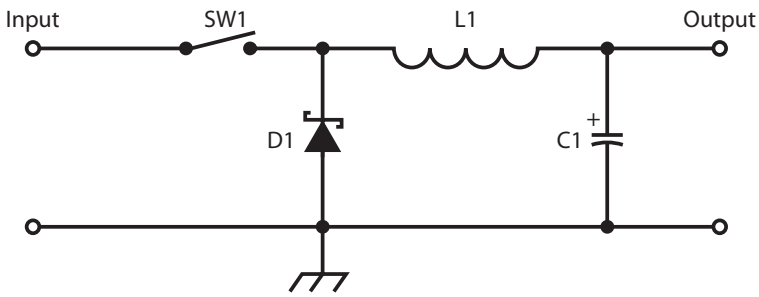
The popular 78XX family of voltage regulator ICs are three-pin devices looking like a power transistor. The last two digits of the 78XX IC's name denotes the regulated output voltage that the IC produces, so the 7812 has a 12-V output, the 7809 9 V, and the 7805 5 V. Internally, these ICs operate in much the same way as the power transistor and Zener diode shown in Fig. 24-8. However they also incorporate thermal protection so that if the IC package starts to get hot enough to damage the IC, it drops the output voltage and, hence, reduces the current until the device recovers.

Figure 24-9 shows how a 7805 might be used to provide a regulated 5-V output from an input of between 7 and 18 V. The IC is generally surrounded by a pair of capacitors. The capacitor at its input is needed only if the regulator is not situated close to the power supply filter. The output capacitor improves the stability of the regulation and the response of the regulator to sudden changes in current requirement. Typical values of input and output capacitors are 0.33 μF and 0.1 μF, respectively.

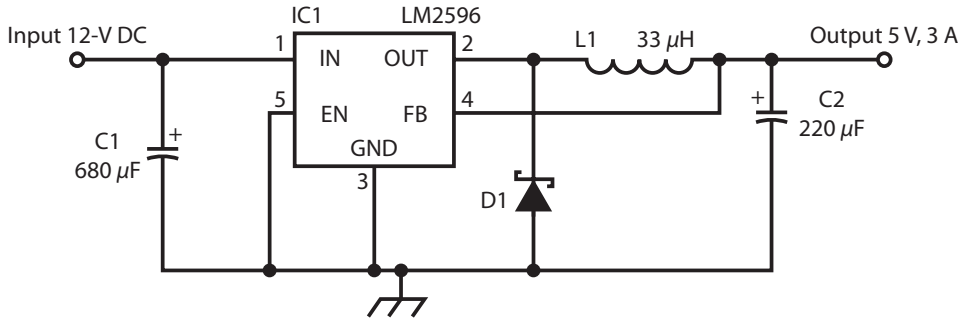
Switching Voltage Regulators

Linear voltage regulators "lose" the unwanted volts as heat. This makes them inefficient and is likely to get hot when asked to reduce more than a few volts at more than a few tens of milliamperes. A much more efficient approach is to use a switching regulator. The most common of these is called a "buck" converter. Figure 24-10 shows how a buck converter works.

SW1 is not a mechanical switch, but normally a MOSFET that is switched on and off with a high-speed PWM signal. Every time the switch closes, current starts to build in L1 and a voltage arises across L1 to oppose the input voltage. Energy is stored as a magnetic field in the inductor which is released through the load after smoothing by C1 when the switch is opened. The diode D1 then becomes forward biased allowing a path for the current to discharge through the inductor.



24-10 A switching "buck" converter.



24-11 A practical buck converter design.

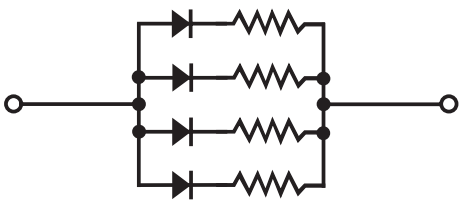
In practical circuits such as the one shown in Fig. 24-11, an IC is used that contains a MOSFET switch as well as a high-frequency oscillator and pulse control circuitry with feedback from the output voltage, that allows the buck converter to accurately regulate the voltage. Buck converter ICs are available at fixed voltages like the example of Fig. 24-11 using the LM2596 IC and also in versions where the output voltage is set by placing a pair of resistors acting as a voltage divider on the output voltage that supplies a reference voltage to a feedback pin on the IC.

The LM2596 contains the PWM circuitry and MOSFET switch to send variable duration pulses to its OUT pin. The FB (feedback) pin is connected to the DC output of the circuit, to regulate the voltage at 5 V. The EN (enable) pin sets the buck converter to a low-power standby mode if connected to IN, otherwise it should be connected to GND to enable the converter. The LM2596 operates at 150 kHz and can operate at 80 percent efficiency, resulting in only modest amounts of heat being generated.

Switched-Mode Power Supplies (SMPS)

The buck converter described in the previous section is great if you already have low voltage DC. But, with care, you can take this step even further and switch from 117-V AC supply. If you pick up your cellphone charger, you can tell just by its weight that there is not a regular transformer in there. A cellphone charger will contain a transformer, but it will be a small light-weight high-frequency transformer. In fact, most modern power sources for consumer products use switched-mode power supplies or SMPSs.

Figure 24-12 shows how an SMPS operates. Instead of using a transformer to reduce the AC voltage, the high-voltage AC input is rectified and filtered to produce high-voltage DC. This DC is then "chopped," or switched, at high frequency into a series of short pulses that feed an output transformer that steps down the voltage. The resulting low-voltage AC is then rectified and filtered



24-12 Block diagram for a SMPS.

into low-voltage DC. Because the transformer operates at such a high frequency, it can be made much smaller and lighter than a 60-Hz transformer.

DC voltage regulation is achieved by feedback from the DC output to the controller that alters the width of the pulses being chopped that, in turn, alters the DC output voltage.

To keep the DC output isolated from the high-voltage AC, the feedback path to the controller uses an optoisolator. This might seem like a lot of extra complexity, but the size, weight, and cost advantages of reducing the transformer size is enough to make SMPSs worthwhile. Generally an SMPS IC that includes the chopper and controller circuitry in a single package is used.

Equipment Protection

The output of a power supply should always remain free of sudden changes that can damage equipment or components, or interfere with their proper performance. To ensure the safety of personnel working around electrical and electronic systems, significant voltages must never appear on the external surfaces of a power supply, or on the external surfaces of any equipment connected to it.

Grounding

The best electrical ground for a power supply is the "third wire" ground provided in up-to-date AC utility circuits. When you examine a typical AC outlet in the United States, you'll see a "hole" shaped like an uppercase letter D turned on its side. That is—or should be—the electrical ground connection. The contacts inside this "hole" should go to a wire that ultimately terminates in a metal rod driven into the earth at the point where the electrical wiring enters the building. That connection constitutes an *earth ground*.

In older buildings, *two-wire AC systems* are common. You can recognize this type of system by noting the presence of two slots in the utility outlets without any ground "hole." Some of these systems obtain reasonable grounding by means of a scheme called *polarization*, in which the two slots have unequal length. The longer slot goes to the electrical ground. However, no two-wire system is as safe as a properly installed *three-wire AC system*, in which the ground connection is independent of both the outlet slots.

Unfortunately, the presence of a three-wire or polarized outlet system does not guarantee that an appliance connected to an outlet is well grounded. If the appliance has an electrical fault, or if the ground "holes" at the outlets weren't actually grounded by the people who originally installed the electrical utility system, a power supply can deliver unwanted voltages to the external surfaces of appliances and electronic devices. These voltages can present an electrocution hazard, and can also hinder the performance of equipment connected to the supply.

Warning!

All exposed metal surfaces of power supplies should be connected to the grounded wire of a three-wire electrical cord. Never defeat or remove the "third prong" of the plug. Always ensure that the electrical system in the building has been properly installed, so you don't work under the illusion that your system has a good ground when it actually does not. If you have any doubts about these matters, consult a professional electrician.

Transients

The AC that we observe at utility outlets is a sine wave with a constant voltage near 117-V RMS or 234-V RMS. However, in most household circuits, we occasionally observe voltage spikes, known as transients, that can attain positive or negative peak values of several thousand volts. Transients can result from sudden changes in the load in a utility circuit. A thundershower can produce transients throughout an entire town. Unless we take some measures to suppress them, transients can destroy the diodes in a power supply. Transients can also cause problems with sensitive electronic equipment, such as computers or microcomputer-controlled appliances.

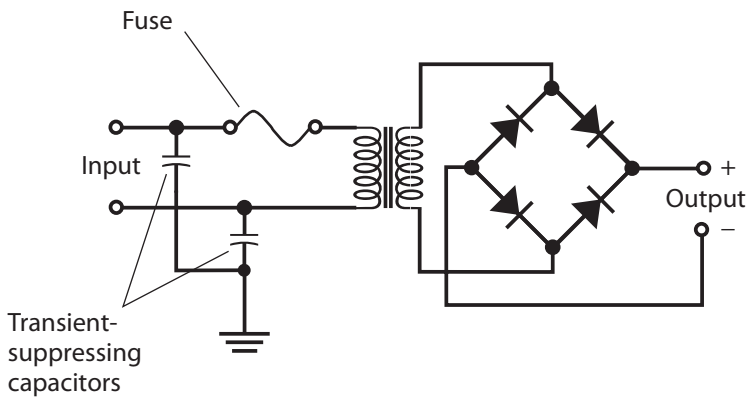
We can get rid of most common transients by connecting a small capacitor of about 0.01 μF, rated for 600 V or more, between each side of the transformer primary and electrical ground, as shown in Fig. 24-13. A *disk ceramic capacitor* (not an electrolytic capacitor) works well for this purpose. Disk ceramic capacitors have no polarity issues, so we can connect them in either direction and expect them to work well.

Commercially made transient suppressors are available. These devices use sophisticated methods to prevent sudden voltage spikes from reaching levels at which they can cause problems. It's a good idea to use transient suppressors with all sensitive electronic devices, including computers, hi-fi stereo systems, and television sets. In the event of a thundershower, the best way to protect such equipment is to physically unplug it from the wall outlets until the storm has passed.

Fuses

A *fuse* comprises a piece of soft wire that melts, breaking a circuit if the current exceeds a certain level. We should connect the fuse in series with the transformer primary, as shown in Fig. 24-13. A short circuit or overload anywhere in the power supply, or in equipment connected to it, will burn the fuse out. If a fuse blows out, we must replace it with another fuse having the same specifications. Fuses are rated in amperes (A). Thus, a 5-A fuse will carry up to 5 A before blowing out, and a 20-A fuse will carry up to 20 A, regardless of the power-supply voltage.

Fuses are available in two types: the *quick-break fuse* and the *slow-blow fuse*. A quick-break fuse consists of a straight length of wire or a metal strip. A slow-blow fuse usually has a spring inside along with the wire or strip. Quick-break fuses in slow-blow situations can burn out needlessly,



24-13 A full-wave bridge rectifier with transient-suppression capacitors and a fuse in the transformer primary circuit.

causing inconvenience. Slow-blow fuses in quick-break environments might not provide adequate protection to the equipment, letting excessive current flow for too long before blowing out.

Circuit Breakers

A *circuit breaker* performs the same function as a fuse, except that we can reset a breaker by turning off the power supply, waiting a little while, and then pressing a button or flipping a switch. Some breakers reset automatically when the equipment has been shut off for a certain length of time. Circuit breakers, like fuses, are rated in amperes.

If a fuse or breaker repeatedly burns out or trips, or if it blows or trips immediately after we replace or reset it, then a serious problem exists with the power supply or with the equipment connected to it. Burned-out diodes, a faulty transformer, and shorted filter capacitors in the supply can all cause trouble. A short circuit in the equipment connected to the supply, or the connection of a device in the wrong direction (polarity), can cause repeated fuse blowing or circuit-breaker tripping.

Never replace a fuse or breaker with a larger-capacity unit to overcome the inconvenience of repeated fuse/breaker blowing/tripping. Find the cause of the trouble, and repair the equipment as needed. The "penny in the fuse box" scheme can endanger equipment and personnel, and it increases the risk of fire in the event of a short circuit. Such an action can also cause extensive damage to power-supply components including diodes, transformers, or filter chokes.

> **Warning!**
>
> High-voltage power supplies can retain deadly voltages even after they've been switched off and unplugged. This danger exists because the filter capacitors retain their charge for a while, even in the absence of applied power. If you have any doubt about your ability to safely build or work with a power supply, leave it to a professional.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

25
CHAPTER

Amplifiers

AMPLIFICATION INVOLVES INCREASING THE POWER OF A SIGNAL, EITHER BY INCREASING ITS VOLTAGE, its current, or both. In this chapter we will look at various methods of achieving this, both using discrete transistors and ICs. But first, let's take a closer look at a topic that we've encountered briefly already: the decibel (dB) as an expression of relative signal strength.

The Decibel Revisited

We can consider amplitude increases as *positive gain* and amplitude decreases as *negative gain*. For example, if a circuit's output signal amplitude equals +6 dB relative to the input signal amplitude, then the output exceeds the input. If the output signal amplitude is −14 dB relative to the input, then the output is weaker than the input. In the first case, we say that the circuit has a gain of 6 dB. In the second case, we say that the circuit has a gain of −14 dB or a loss of 14 dB.

For Voltage

Consider a circuit with an RMS AC input voltage of V_{in} and an RMS AC output voltage of V_{out}, with both voltages expressed in the same units (volts, millivolts, microvolts, or whatever). Also suppose that the input and output impedances both constitute pure resistances of the same ohmic value. We can calculate the *voltage gain* of the circuit, in decibels, with the formula

$$\text{Gain (dB)} = 20 \log (V_{out}/V_{in})$$

In this equation, "log" stands for the *base-10 logarithm* or *common logarithm*. Scientists and engineers write the base-10 logarithm of a quantity x as "log x" or sometimes as "$\log_{10} x$." We can consider the coefficient 20 in the above equation to be an exact value when we make calculations, no matter how many significant digits we need.

Logarithms can have other bases besides 10, the most common of which is the *exponential constant*, symbolized as e and equal to approximately 2.71828. Scientists and engineers express the *base-e logarithm* (also called the *natural logarithm*) of a quantity x as "$\log_e x$" or "ln x." Without knowing all of the mathematical particulars concerning how *logarithmic functions* behave, we can calculate the logarithms of specific numbers with the help of a good scientific calculator. From now on, let's agree that when we say "logarithm" or write "log," we mean the base-10 logarithm.

Problem 25-1

Suppose that a circuit has an RMS AC input of 1.00 V and an RMS AC output of 14.0 V. How much gain, in decibels, does this circuit produce?

Solution

First, you must find the ratio V_{out}/V_{in}. Because $V_{out} = 14.\ 0$ V RMS and $V_{in} = 1.00$ V RMS, the ratio equals 14.0/1.00, or 14.0. Next, find the logarithm of 14.0. A calculator tells you that this quantity is quite close to 1.146128036. Finally, multiply this number by 20 and then round off to three significant figures, getting 22.9 dB.

Problem 25-2

Suppose that a circuit has an RMS AC input voltage of 24.2 V and an RMS AC output voltage of 19.9 V. What's the gain in decibels?

Solution

Find the ratio $V_{out}/V_{in} = 19.9/24.2 = 0.822314\ldots$. (The sequence of three dots, called an *ellipsis*, indicates extra digits introduced by the calculator. You can leave them in until the final round-off.) When you use a calculator to find the logarithm of this quantity, you get $\log 0.822314\ldots = -0.0849622\ldots$. The gain equals $20 \times (-0.0849622)$, which rounds off to -1.70 dB.

For Current

We can calculate current gain or loss figures in decibels in the same way as we calculate voltage gain or loss figures. If I_{in} represents the RMS AC input current and I_{out} represents the RMS AC output current (in the same units as I_{in}, such as amperes, milliamperes, microamperes, or whatever), then

$$\text{Gain (dB)} = 20 \log (I_{out}/I_{in})$$

For this formula to work, the input and output impedances must both comprise pure resistances, and the ohmic values must be identical.

For Power

We can calculate the *power gain* of a circuit, in decibels, by cutting the coefficient of the formula in half, from exactly 20 to exactly 10, accurate to as many significant digits as we want. If P_{in} represents the input signal power and P_{out} represents the output signal power (in the same units as P_{in}, such as watts, milliwatts, microwatts, kilowatts, or whatever), then

$$\text{Gain (dB)} = 10 \log (P_{out}/P_{in})$$

For this formula to work, the input and output impedances should both show up as pure resistances, but their ohmic values can differ.

Problem 25-3

Suppose that a power amplifier has an input of 5.72 W and an output of 125 W. What's the gain in decibels?

Solution

Find the ratio $P_{out}/P_{in} = 125/5.72 = 21.853146\ldots$. Then find the logarithm, obtaining log $21.853146\ldots = 1.339513\ldots$. Finally, multiply by 10 and round off to obtain a gain figure of $10 \times 1.339513\ldots = 13.4$ dB.

Problem 25-4

Suppose that an *attenuator* (a circuit designed deliberately to produce power loss) provides 10 dB power reduction. You supply this circuit with an input signal at 94 W. What's the output power?

Solution

An attenuation of 10 dB represents a gain of −10 dB. You know that $P_{in} = 94$ W, so P_{out} constitutes the unknown in the power gain formula. You must, therefore, solve for P_{out} in the equation

$$-10 = 10 \log (P_{out}/94)$$

First, divide each side by 10, getting

$$-1 = \log (P_{out}/94)$$

To solve this equation, you must take the *base-10 antilogarithm*, also known as the *antilog* or *inverse log*, of each side. The antilog function "undoes" the work of the log function. The antilog of a value x can be abbreviated as "antilog x." Pure mathematicians, as well as some scientists and engineers, denote it as "$\log^{-1} x$." Antilogarithms of specific numbers, like logarithms, can be determined with any good scientific calculator. (Function keys for the antilogarithm vary, depending on the particular calculator you use. You might have to enter the value and then hit an "Inv" key followed by a "log" key; you might enter the value and then hit a "10^x" key.) When you take the antilogarithm of both sides of the above equation, you get

$$\text{antilog } (-1) = \text{antilog } [\log (P_{out}/94)]$$

Working out the value on the left-hand side with a calculator, and noting on the right-hand side that the antilog "undoes" the work of the log, you get

$$0.1 = P_{out}/94$$

Multiplying each side by 94 tells you that

$$94 \times 0.1 = P_{out}$$

Finally, when you multiply out the left-hand side and transpose the left-hand and right-hand sides of the equation, you get the answer as

$$P_{out} = 9.4 \text{ W}$$

Decibels and Impedance

When determining the voltage gain (or loss) and the current gain (or loss) for a circuit in decibels, you can expect to get the same figure for both parameters only when the complex input impedance is *identical* to the complex output impedance. If the input and output impedances differ (either reactance-wise or resistance-wise, or both), then the voltage gain or loss generally differs from the current gain or loss.

Consider how transformers work. A step-up transformer, in theory, has voltage gain, but this voltage increase alone doesn't make a signal more powerful. A step-down transformer can exhibit theoretical current gain, but again, this current increase alone doesn't make a signal more powerful. In order to make a signal more powerful, a circuit must increase the signal *power*—the *product* of the voltage and the current!

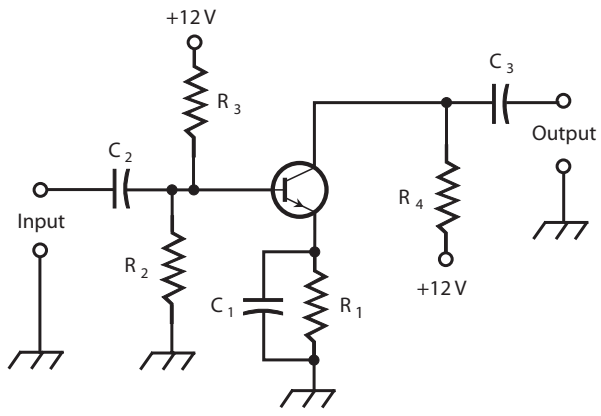
When determining power gain (or loss) for a particular circuit in decibels, the input and output impedances *do not* matter, as long as they're both free of reactance. In this sense, positive power gain always represents a real-world increase in signal strength. Similarly, negative power gain (or power loss) always represents a true decrease in signal strength.

Whenever we want to work out decibel values for voltage, current, or power, we should strive to get rid of all the reactance in a circuit, so that our impedances constitute pure resistances. Reactance "artificially" increases or decreases current and voltage levels; and while reactance theoretically consumes no power, it can have a profound effect on the behavior of instruments that measure power, and it can cause power to go to waste as unwanted heat dissipation.

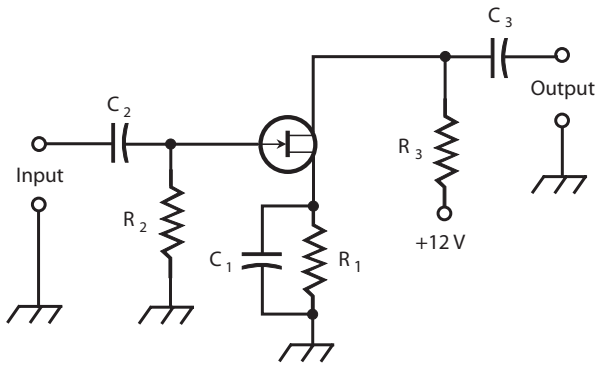
Basic Bipolar-Transistor Amplifier

In the previous chapters, you saw some circuits that use transistors. We can apply an input signal to some control point (the base, gate, emitter, or source), causing a much greater signal to appear at the output (usually the collector or drain). Figure 25-1 shows a circuit in which we've wired up an NPN bipolar transistor to serve as a *common-emitter amplifier*. The input signal passes through C_2 to the base. Resistors R_2 and R_3 provide the base bias. Resistor R_1 and capacitor C_1 allow for the emitter to maintain a constant DC voltage relative to ground, while keeping it grounded for the AC signal. Resistor R_1 also limits the current through the transistor. The AC output signal goes through capacitor C_3. Resistor R_4 keeps the AC output signal from shorting through the power supply.

In this amplifier, the optimum capacitance values depend on the design frequency of the amplifier, and also on the impedances at the input and output. In general, as the frequency and/or circuit impedance increase, we need less capacitance. At audio frequencies (abbreviated AF and representing frequencies from approximately 20 Hz to 20 kHz) and low impedances, the capacitors might have values as large as 100 μF. At radio frequencies (RF) and high impedances, values will normally equal only a fraction of a microfarad, down to picofarads at the highest frequencies and impedances.



25-1 A generic amplifier circuit with a bipolar transistor.



25-2 A generic amplifier circuit with a JFET.

The optimum resistor values also depend on the application. In the case of a weak-signal amplifier, typical values are 470 Ω for R_1, 4.7 k for R_2,10 k for R_3, and 4.7 k for R_4.

Basic FET Amplifier

Figure 25-2 shows an N-channel JFET hooked up as a *common-source amplifier*. The input signal passes through C_2 to the gate. Resistor R_2 provides the gate bias. Resistor R_1 and capacitor C_1 give the source a DC voltage relative to ground, while grounding it for signals. The output signal goes through C_3. Resistor R_3 keeps the output signal from shorting through the power supply.

A JFET exhibits high input impedance, so the value of C_2 should be small. If we use a MOSFET rather than a JFET, we'll get a higher input impedance, so C_2 will be smaller yet, sometimes 1 pF or less. The resistor values depend on the application. In some instances, we won't need R_1 and C_1 at all, and we can connect the source directly to ground. If we use resistor R_1, its optimum value depends on the input impedance and the bias needed for the FET. For a weak-signal amplifier, typical values are 680 Ω for R_1, 10 k for R_2, and 100 Ω for R_3.

Amplifier Classes

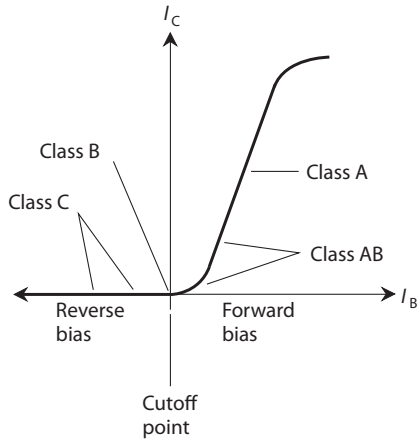
Engineers classify analog amplifier circuits according to the bias arrangement as *class A*, *class AB*, *class B*, and *class C*. Each class has its own special characteristics, and works best in its own unique set of circumstances. A specialized amplifier type, called *class D*, makes for very efficient power amplifiers.

The Class-A Amplifier

With the previously mentioned component values, the amplifier circuits in Figs. 25-1 and 25-2 operate in the class-A mode. This type of amplifier is *linear*, meaning that the output waveform has the same shape as (although a greater amplitude than) the input waveform.

When we want to obtain class-A operation with a bipolar transistor, we must bias the device so that, with no signal input, it operates near the middle of the straight-line portion of the I_C versus I_B (collector current versus base current) curve. Figure 25-3 shows this situation for a bipolar transistor. With a JFET or MOSFET, the bias must be such that, with no signal input, the device operates near the middle of the straight-line part of the I_D versus E_G (drain current versus gate voltage) curve as shown in Fig. 25-4.

When we use a class-A amplifier in the "real world," we must never allow the input signal to get too strong. An excessively strong input signal will drive the device out of the straight-line part of



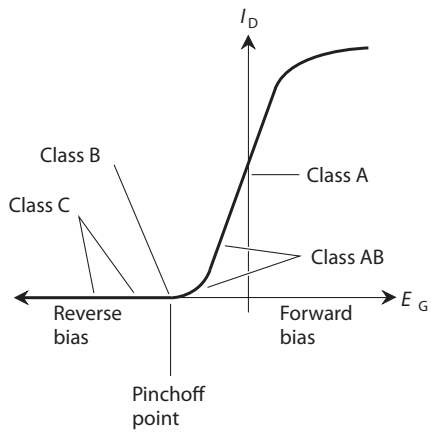
25-3 Classes of amplifier operation for a typical bipolar transistor.

the characteristic curve during part of the cycle. When this phenomenon occurs, the output waveform will no longer represent a faithful reproduction of the input waveform, and the amplifier will become *nonlinear*. In some types of amplifiers, we can tolerate (or even encourage) nonlinearity, but we'll want a class-A amplifier to operate in a linear fashion at all times.

Class-A amplifiers suffer from one outstanding limitation: The transistor draws current whether or not any input signal exists. The transistor must "work hard" even when no signal comes in. For weak-signal applications, the "extra work load" doesn't matter much; we must concern ourselves entirely with getting plenty of gain out of the circuit.

The Class-AB Amplifier

When we bias a bipolar transistor close to cutoff under no-signal conditions, or when we bias a JFET or MOSFET near pinchoff, the input signal always drives the device into the nonlinear part of the *operating curve*. (Figures 25-3 and 25-4 are examples of operating curves.) Then, by definition, we have a class-AB amplifier. Figures 25-3 and 25-4 show typical bias zones for class-AB amplifier operation. A small collector or drain current flows when no input signal exists, but it's less than the no-signal current in a class-A amplifier.



25-4 Classes of amplifier operation for a typical JFET.

Engineers sometimes specify two distinct modes of class-AB amplification. If the bipolar transistor or FET never goes into cutoff or pinchoff during any part of the input signal cycle, the amplifier works in *class* AB_1. If the device goes into cutoff or pinchoff for any part of the cycle (up to almost half), the amplifier operates in *class* AB_2.

In a class-AB amplifier, the output-signal wave doesn't have the same shape as the input-signal wave. But if the signal wave is *modulated*, such as in a voice radio transmitter, the waveform of the *modulating signal* comes out undistorted anyway. Therefore, class-AB operation can work very well for RF power amplifiers.

The Class-B Amplifier

When we bias a bipolar transistor exactly at cutoff or an FET exactly at pinchoff under zero-input-signal conditions, we cause an amplifier to function in the *class-B* mode. The operating points for class-B operation are labeled on the curves in Figs. 25-3 and 25-4. This scheme, like the class-AB mode, lends itself well to RF power amplification.

In class-B operation, no collector or drain current flows under no-signal conditions. Therefore, the circuit does not consume significant power unless a signal goes into it. (Class-A and class-AB amplifiers consume some DC power under no-signal conditions.) When we provide an input signal, current flows in the device during half of the cycle. The output signal waveform differs greatly from the input waveform. In fact, the wave comes out "half-wave rectified" as well as amplified.

You'll sometimes hear of class-AB or class-B "linear amplifiers," especially when you converse with amateur (ham) radio operators. In this context, the term "linear" refers to the fact that the amplifier doesn't distort the *modulation waveform* (sometimes called the *modulation envelope*), even though the *carrier waveform* is distorted because the transistor is not biased in the straight-line part of the operating curve under no-signal conditions.

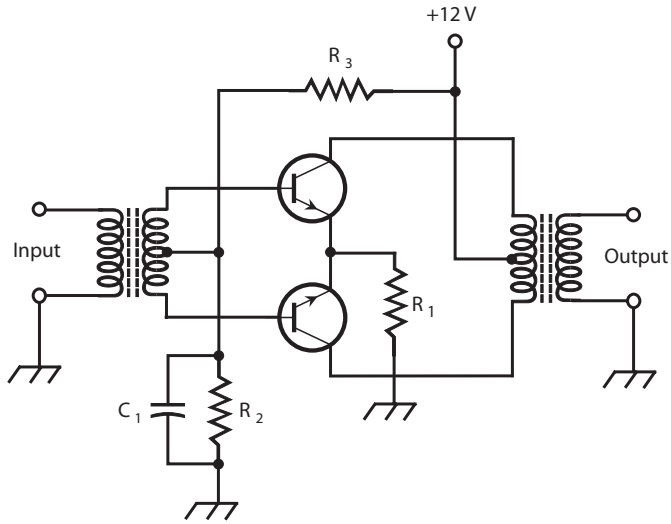
Class-AB_2 and class-B amplifiers draw power from the input signal source. Engineers say that such amplifiers require a certain amount of *drive* or *driving power* to function. Class-A and class-AB_1 amplifiers theoretically need no driving power, although the input signal must provide a certain amount of voltage to influence the behavior of the transistor.

The Class-B Push-Pull Amplifier

We can connect two bipolar transistors or FETs in a pair of class-B circuits that operate in tandem, one for the positive half of the input wave cycle and the other for the negative half. In this way, we can get rid of waveform distortion while retaining all the benefits of class-B operation. We call this type of circuit a *class-B push-pull amplifier*. Figure 25-5 shows an example using two NPN bipolar transistors.

Resistor R_1 limits the current through the transistors. Capacitor C_1 keeps the input transformer center tap at signal ground, while allowing for some DC base bias. Resistors R_2 and R_3 bias the transistors precisely at their cutoff points. For best results, the two transistors must be identical. Not only should their part numbers match, but we should pick them out *by experiment* for each amplifier circuit that we build, to ensure that the characteristics coincide as closely as possible.

Class-B push-pull circuits provide a popular arrangement for audio-frequency (AF) power amplification. Push-pull design offers the easy transistor workload of the class-B mode with the low-distortion, linear amplification characteristics of the class-A mode. However, a push-pull amplifier needs two center-tapped transformers, one at the input and the other at the output, a requirement that makes push-pull amplifiers more bulky and expensive than other types.



25-5 A class-B push-pull amplifier using NPN bipolar transistors.

All push-pull amplifiers share a unique and interesting quality: They "cancel out" the *even-numbered* harmonics in the output. This property offers an advantage in the design and operation of wireless transmitters because it gets rid of concerns about the second harmonic, which usually causes more trouble than any other harmonic. A push-pull circuit doesn't suppress *odd-numbered* harmonics any more than a "single-ended" circuit (one that employs a single bipolar transistor or FET) does.

The Class-C Amplifier

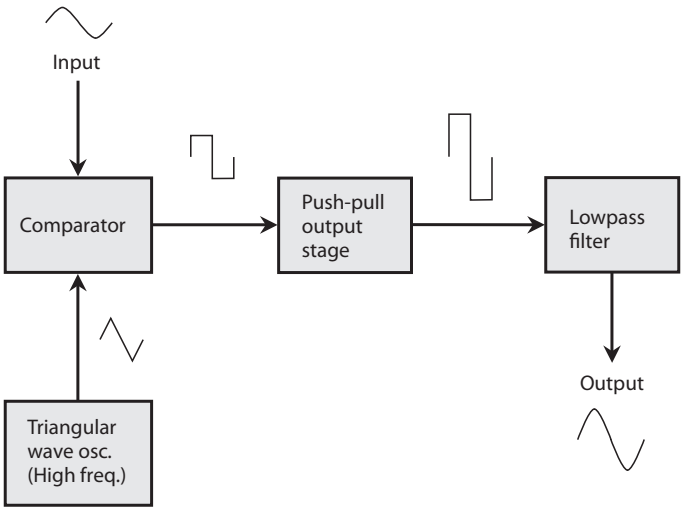
We can bias a bipolar transistor or FET past cutoff or pinchoff, and it will still work as a power amplifier (PA), provided that the drive is sufficient to overcome the bias during part of the cycle. We call this mode *class-C* operation. Figures 25-3 and 25-4 show no-signal bias points for class-C amplification.

Class-C amplifiers are nonlinear, even for amplitude-modulation (AM) envelopes. Therefore, engineers generally use class-C circuits only with input signals that are either "full-on" or "full-off." Such signals include old-fashioned *Morse code*, along with digital modulation schemes in which the frequency or phase (but not the amplitude) of the signal can vary, but the amplitude is always either zero or maximum.

A class-C amplifier needs a lot of driving power. It does not produce as much gain as other amplification modes. For example, we might need 300 W of signal drive to get 1 kW of signal power output from a class-C power amplifier. However, we get more "signal bang for the buck" in class C than we do with an amplifier working in any other mode. That is to say, we get optimum *efficiency* from the class-C scheme.

The Class-D Amplifier

A class-D amplifier differs radically from traditional amplifiers; its output transistors (generally MOSFETs) act in a digital way, always either on or off, as opposed to analog modes for the other classes. Class-D amplifiers use pulse-width modulation (PWM), which you'll learn about



25-6 Functional diagram of a class-D audio amplifier.

in Chap. 27 (Wireless Transmitters and Receivers) and Chap. 29 (Microcontrollers). Figure 25-6 shows the principle of class-D operation.

A class-D amplifier uses a comparator, which you learned about in Chap. 23 (Integrated Circuits) and a triangular-wave oscillator to convert an analog input signal to a train of pulses having different lengths. The pulse length varies in proportion to the instantaneous input-signal voltage. These pulses get amplified by a class-C-like push-pull stage. Then a lowpass filter, playing the role of a digital-to-analog converter, transforms the pulses back into an analog signal. (For low-quality audio applications with a loudspeaker output, the speaker can't respond fast enough to follow the pulse frequency, so it converts the digital signal to analog form without the need for a separate lowpass filter.)

To visualize the digitization process, imagine the input signal slowly changing compared to the output of the triangular-wave oscillator. Let's say that the input signal is, at first, stronger than the signal from the triangular-wave oscillator, so the comparator output is high. At some point, the input signal gets weaker than the triangular signal, and the comparator output goes low. The time that this transition takes depends on the voltage of the input signal. The higher the amplitude, the longer the pulse. The proportion of the time that the pulse remains high is called the *duty cycle*. The higher the duty cycle, the stronger the output signal.

When built with MOSFETs that have input impedances on the order of a few megohms, a class-D amplifier can supply several tens of watts of power without overheating. Class-D amplifiers almost always use ICs rather than discrete components, and have largely replaced analog amplifiers in consumer devices, such as cell phones, notebook computers, and tablet computers.

On the downside, class-D amplifiers cause some distortion. For hi-fi audio amplification, therefore, analog designs remain superior; class-A amplifiers are ideal for that purpose.

Efficiency in Power Amplifiers

An efficient power amplifier not only provides optimum output power with minimum heat generation and minimum strain on the transistors, but it conserves energy as well. These factors translate into reduced cost, reduced size and weight, and longer equipment life as compared with inefficient power amplifiers.

The DC Input Power

Suppose that you connect an ammeter or milliammeter in series with the collector or drain of an amplifier and the power supply. While the amplifier operates, the meter will show a certain reading. The reading might appear constant, or it might fluctuate with changes in the input signal level. The DC *collector input power* to a bipolar-transistor amplifier circuit equals the product of the collector current I_C and the collector voltage V_C. Similarly, for an FET, the *DC drain input power* equals the product of the drain current I_D and the drain voltage V_D. We can categorize DC input power figures as *average* or *peak* values. The following discussion involves only average power.

We can observe significant DC collector or drain input power even when an amplifier receives no input signal. A class-A circuit operates this way. In fact, when we apply an input signal to a class-A amplifier, the average DC collector or drain input power *does not change* compared to the value under no-signal conditions! In class AB_1 or class AB_2, we observe low current (and, therefore, low DC collector or drain input power) with zero input signal, and higher current (and, therefore, higher DC input power) when we apply a signal to the input. In classes B and C, we see no current (and, therefore, zero DC collector or drain input power) when no input signal exists. The current, and therefore, the DC input power, increase with increasing signal input.

We usually express or measure the DC collector or drain input power in watts, the product of amperes and volts, as long as the circuit has no reactance. We can express it in milliwatts for low-power amplifiers, or kilowatts (and, in extreme cases, megawatts) for high-power amplifiers.

The Signal Output Power

If we want to accurately determine the *signal output power* from an amplifier, we must employ a specialized AC wattmeter. The design of AF and RF wattmeters represents a sophisticated specialty in engineering. We can't connect an ordinary DC meter and rectifier diode to a power amplifier's output terminals and expect to get a true indication of the signal output power.

When a properly operating power amplifier receives no input signal, we never see any signal output, and therefore, the power output equals zero. This situation holds true for all classes of amplification. As we increase the strength of the input signal, we observe increasing signal output power up to a certain point. If we keep increasing the strength of the input signal (that is, the drive) past this point, we see little or no further increase in the signal output power.

We express or measure signal output power in watts, just as we do with DC input power. For very low-power circuits, we might want to specify the signal output power in milliwatts; for moderate- and high-power circuits we can express it in watts, kilowatts, or even megawatts.

Definition of Efficiency

Engineers define the *efficiency* of a power amplifier as the ratio of the signal output power to the DC input power. We can render this quantity either as a plain fraction (in which case it has a value between 0 and 1) or as a percentage (in which case it has a value between 0% and 100%). Let P_{in} represent the DC input power to a power amplifier. Let P_{out} represent the signal output power. We can quantify the efficiency (eff) as the ratio

$$\text{eff} = P_{out} / P_{in}$$

or as the percentage

$$\text{eff}_{\%} = 100\ P_{out} / P_{in}$$

Problem 25-5

Suppose that a bipolar-transistor amplifier has a DC input power of 120 W and a signal output power of 84 W. What's the efficiency in percent?

Solution

We should use the formula for amplifier efficiency $\text{eff}_{\%}$ expressed as a percentage. When we go through the arithmetic, we obtain

$$\text{eff}_{\%} = 100\ P_{\text{out}}/P_{\text{in}}$$
$$= 100 \times 84/120 = 100 \times 0.70 = 70\%$$

Problem 25-6

Suppose that the efficiency of an FET amplifier equals 0.600. If we observe 3.50 W of signal output power, what's the DC input power?

Solution

Let's plug in the given values to the formula for the efficiency of an amplifier expressed as a ratio, and then use simple algebra to solve the problem. We get

$$0.600 = 3.50/P_{\text{in}}$$

which solves to

$$P_{\text{in}} = 3.50 / 0.600 = 5.83 \text{ W}$$

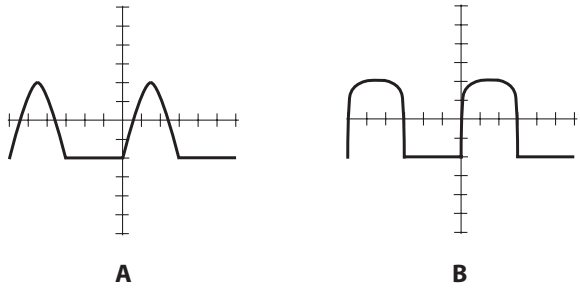
Efficiency versus Class

Class-A amplifiers have efficiency figures from 25% to 40%, depending on the nature of the input signal and the type of transistor used. A good class-AB_1 amplifier has an efficiency rating somewhere between 35% and 45%. A class-AB_2 amplifier, if well designed and properly operated, can exhibit an efficiency figure up to about 50%. Class-B amplifiers are typically 50% to 65% efficient. Class-C amplifiers can have efficiency levels as high as 75%.

Drive and Overdrive

In theory, class-A and class-AB_1 power amplifiers don't draw any power from the signal source to produce significant output power. This property constitutes one of the main advantages of these two modes. We need to provide only a certain AC signal voltage at the control electrode (the base, gate, emitter, or source) for class-A or class-AB_1 circuits to produce useful output signal power. Class-AB_2 amplifiers need some driving power to produce AC power output. Class-B amplifiers require more drive than class-AB_2 circuits do, and class-C amplifiers need the most driving power of all.

Whatever type of PA that we employ in a given situation, we must always make certain that we don't allow the driving signal to get too strong because it will produce an undesirable condition known as *overdrive*. When we force an amplifier to operate in a state of overdrive, we get excessive distortion in the output signal waveform. This distortion can adversely affect the modulation envelope. We can use an *oscilloscope* (or *scope*) to determine whether or not this type of distortion is



25-7 At A, an oscilloscope display of the signal output waveform from a properly operating class-B power amplifier. At B, a display showing distortion in the waveform caused by overdrive.

taking place in a particular situation. The scope gives us an instant-to-instant graphical display of signal amplitude as a function of time. We connect the scope to the amplifier output terminals to scrutinize the signal waveform. The output waveform for a particular class of amplifier always has a characteristic shape; overdrive causes a form of distortion known as *flat topping*.

Figure 25-7A illustrates the output signal waveform for a properly operating class-B amplifier. Figure 25-7B shows the output waveform for an overdriven class-B amplifier. Note that in drawing B, the waveform peaks appear blunted. This unwanted phenomenon shows up as distortion in the modulation on a radio signal, and also an excessive amount of signal output at harmonic frequencies. The efficiency of the circuit can be degraded, as well. The blunted wave peaks cause higher-than-normal DC input power but no increase in useful output power, resulting in below-par efficiency.

Audio Amplification

The circuits that we've examined so far have been generic but not application-specific. With capacitors of several microfarads, and when biased for class-A operation, these circuits offer good representations of audio amplifiers.

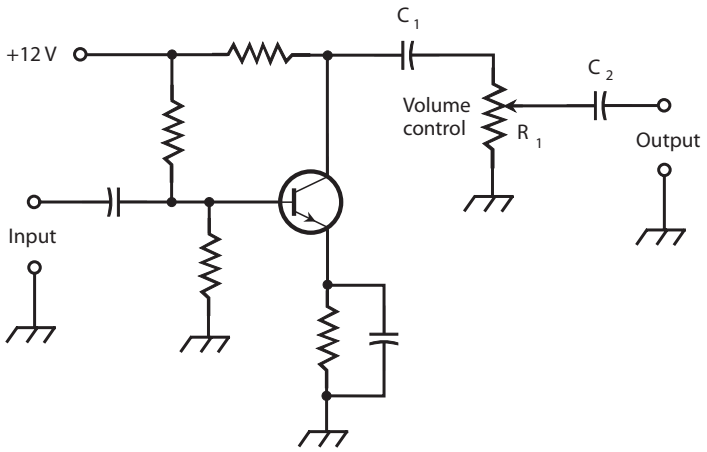
Frequency Response

High-fidelity (or hi-fi) audio amplifiers, of the kind used in music systems, must have more or less constant gain from 20 Hz to 20 kHz (at least), and preferably over a wider range such as 5 Hz to 50 kHz. Audio amplifiers for voice communications need to function well only between approximately 300 Hz and 3 kHz. In digital communications, audio amplifiers are designed to work over a narrow range of frequencies, sometimes less than 100 Hz wide.

hi-fi amplifiers usually contain resistance-capacitance (*RC*) networks that tailor the frequency response. These networks constitute *tone controls*, also called *bass* and *treble* controls. The simplest hi-fi amplifiers use a single control to provide adjustment of the tone. Sophisticated amplifiers have separate controls, one for bass and the other for treble. The most advanced hi-fi systems make use of *graphic equalizers*, having controls that affect the amplifier gain over several different frequency spans.

Volume Control

Audio amplifier systems usually consist of two or more *stages*. A stage has one bipolar transistor or FET (or a push-pull combination) along with resistors and capacitors. We can cascade two or more stages to get high gain. In one of the stages, we incorporate a *volume control*. The simplest volume control is a potentiometer that allows us to adjust the amplifier system gain without affecting its linearity.



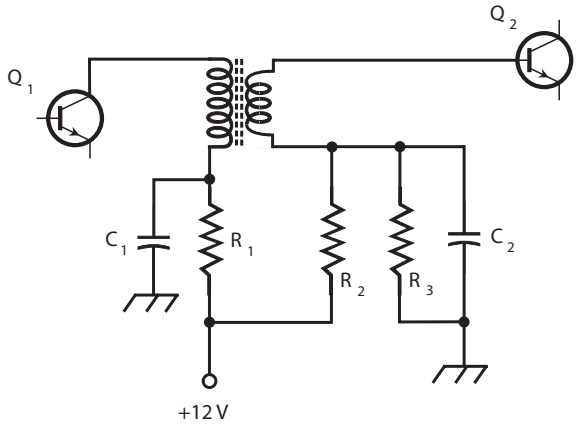
25-8 We can use a basic volume control (potentiometer R_1) to vary the gain in a low-power audio amplifier.

Figure 25-8 illustrates a basic volume control. In this amplifier, the gain through the transistor remains constant even as the input signal strength varies. The AC output signal passes through C_1 and appears across potentiometer R_1, the volume control. The *wiper* (indicated by the arrow) of the potentiometer "picks off" more or less of the AC output signal, depending on the position of the control shaft. Capacitor C_2 isolates the potentiometer from the DC bias of the following stage.

We should always place an audio volume control in a low-power stage. If we put the potentiometer at a high-power point, it will have to dissipate considerable power when we set the volume at a low level. High-power potentiometers, as you might guess, cost more than low-power ones, and they're harder to find. Even if we manage to obtain a potentiometer that can handle the strain, placing the volume control at a high-power point will cause the amplifier system to suffer from poor efficiency.

Transformer Coupling

We can use transformers to transfer (or *couple*) signals from one stage to the next in a cascaded amplifier system (also called an *amplifier chain*). Figure 25-9 illustrates *transformer coupling* between two amplifier circuits. Capacitors C_1 and C_2 keep the lower ends of the transformer windings at



25-9 An example of transformer coupling between amplifier stages.

signal ground. Resistor R_1 limits the current through the first transistor Q_1. Resistors R_2 and R_3 provide the base bias for transistor Q_2.

Transformer coupling costs more per stage than *capacitive coupling* does. However, transformer coupling can provide optimum signal transfer between amplifier stages. By selecting a transformer with the correct turns ratio, the output impedance of the first stage can be perfectly matched to the input impedance of the second stage, assuming that no reactance exists in either circuit.

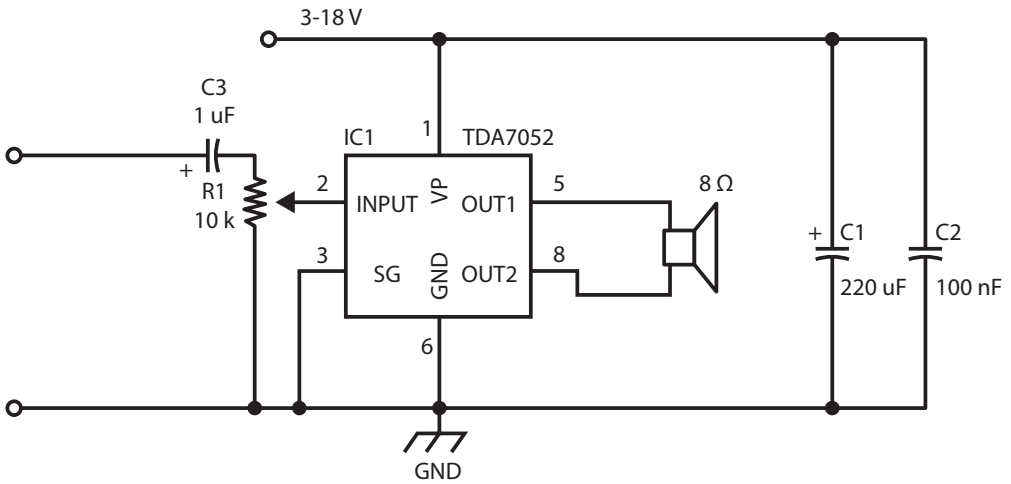
IC-Based Audio Amplifiers

It's a good thing to know about the different classes of amplifier and their pros and cons. However, your design can be kept simple by using an IC amplifier. For low-power amplifiers, of 1 W or less, linear audio amplifier ICs such as the popular TDA7052 are a good option. If you need more power and want to avoid the need for heatsinks for cooling, then a class-D amplifier IC is a good option.

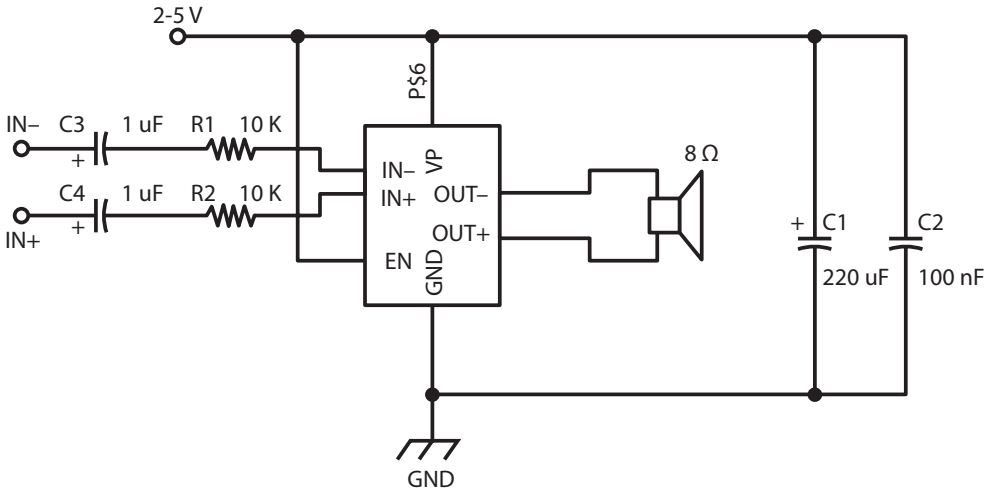
Figure 25-10 shows a typical application using the TDA7052 amplifier IC. The IC will directly drive an 8 Ω speaker in a push-pull arrangement. The only extra components required are the capacitor C1 providing AC coupling to the source of audio being amplified, potentiometer R1 that acts as a volume control and the capacitors C1 and C2. C2 should be placed as close as possible to the IC. This is called a decoupling capacitor and most ICs benefit from greater stability if they have a decoupling capacitor that can supply the fast moving currents that ICs require. C2 acts as a reservoir capacitor that can provide the power needed for the amplifier to drive its low impedance load.

Class-D audio power amplifiers are becoming the most common way to drive a loudspeaker. Digital ICs such as the PAM8302 provide almost all the components that are required. Figure 25-11 shows an example circuit using the PAM8302.

The PAM8302 has a fixed voltage gain of 24 dB and can supply up to 2.5 W into a load. It has differential inputs, amplifying the difference between IN+ and IN−. Usually IN− will share a common ground with the audio signal source.



25-10 A TDA7052 power amplifier example application.



25-11 A PAM8302 power amplifier example application.

There are many advantages to using an IC like the PAM8302 that would require a lot of extra components were you to provide the equivalent using discrete components:

- Low cost
- Low component count
- Short-circuit protection
- Thermal protection

Low-cost digital amplifier ICs like the PAM8302 generally have higher levels of distortion than linear designs and of course have a high frequency component that must be filtered out. However, when driving a low-cost speaker, the difference will probably not be noticed. Digital audio amplifier IC design has now progressed to the level where higher-end ICs can now perform as well as their linear cousins and are even finding their way into hi-fi audio amplifiers.

Radio-Frequency Amplification

The *RF spectrum* extends upward in frequency to well over 300 GHz. Sources disagree as to the exact low-frequency limit. Some texts put it at 3 kHz, some at 9 kHz, some at 10 kHz, and a few at the upper end of the AF range, usually defined as 20 kHz.

Weak-Signal versus Power Amplifiers

The *front end*, or first amplifying stage, of a radio receiver requires the most sensitive possible amplifier. The sensitivity depends on two factors: the gain (or amplification factor) and the *noise figure*, a measure of how well a circuit can amplify desired signals while generating a minimum of *internal noise*.

All semiconductor devices create internal noise as a consequence of charge-carrier movement. We call this phenomenon *electrical noise*. Internal noise can also result from the inherent motion of

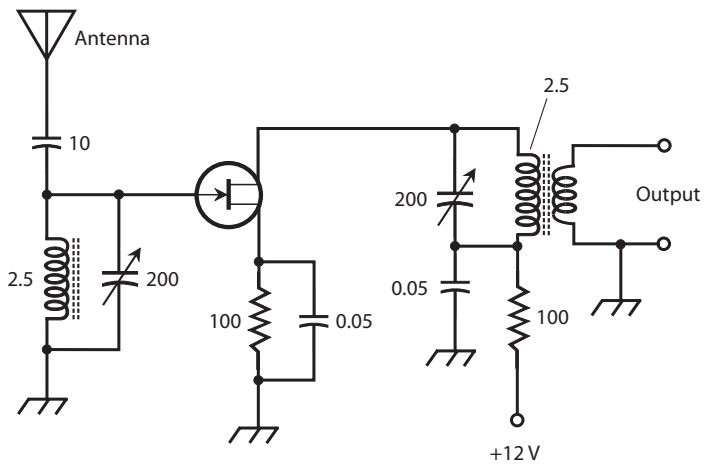
the molecules that comprise the semiconductor material; that's known as *thermal noise*. Random, rapid fluctuations in current can give rise to a third effect called *shot-effect noise*. In general, JFETs produce less electrical and shot-effect noise than bipolar transistors do. Gallium arsenide FETs, also called *GaAsFETs* (pronounced "gasfets"), are generally the "quietest" of all semiconductor devices.

As the operating frequency of a weak-signal amplifier increases, the noise figure gets increasingly important. That's because we observe less *external noise* at the higher radio frequencies than we do at the lower frequencies. External noise comes from the sun (*solar noise*), from outer space (*cosmic noise*), from thundershowers in the earth's atmosphere (*sferics*), from human-made internal combustion engines (*ignition noise*), and from various electrical and electronic devices (*appliance noise*). At a frequency of 1.8 MHz, the airwaves contain a great deal of external noise, and it doesn't make much difference if the receiver introduces a little noise of its own. But at 1.8 GHz, we see far less external noise, so receiver performance depends almost entirely on the amount of internally generated noise.

Tuned Circuits in Weak-Signal Amplifiers

Weak-signal amplifiers almost always take advantage of resonant circuits in the input, at the output, or both. This feature optimizes the amplification at the desired frequency, while helping to minimize noise on unwanted frequencies. Figure 25-12 is a schematic diagram of a typical tuned GaAsFET weak-signal RF amplifier designed for operation at about 10 MHz.

In some weak-signal RF amplifier systems, engineers use transformer coupling between stages and connect capacitors across the primary and/or secondary windings of the transformers. This tactic produces resonance at a frequency determined by the capacitance and the transformer winding inductance. If the set of amplifiers is intended for use at only one frequency, this method of coupling, called *tuned-circuit coupling*, enhances the system efficiency, but increases the risk that the stages will break into oscillation at the resonant frequency.



25-12 A tuned RF amplifier for use at approximately 10 MHz. Resistances are in ohms. Capacitances are in microfarads (μF) if less than 1, and in picofarads (pF) if more than 1. Inductances are in microhenrys (μH).

Broadband PAs

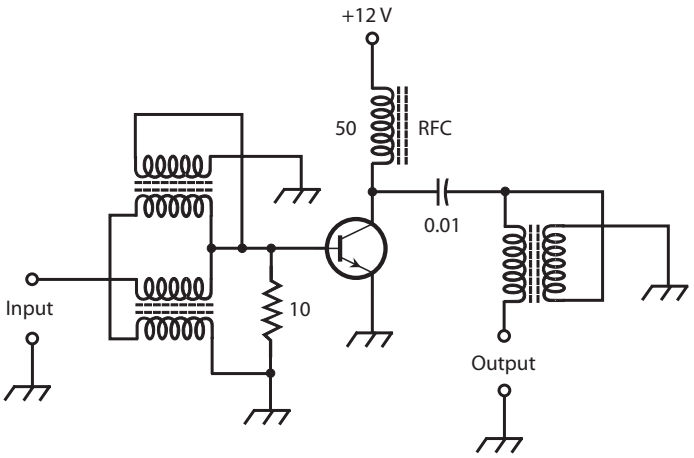
We can design RF power amplifiers to operate in either the *broadband* mode or the *tuned* mode. As these terms suggest, broadband amplifiers work over a wide range of frequencies without adjustment for frequency variations, while tuned amplifiers require adjustment of the resonant frequencies of internal circuits.

A broadband PA offers convenience because it does not require tuning within its design frequency range. The operator need not worry about critical adjustments, or bother altering the circuit parameters when changing the frequency. However, broadband PAs are, as a whole, less efficient than tuned PAs. Another disadvantage of broadband PAs is the fact that they will amplify any signal in the design frequency range, whether or not the operator wants that amplification to occur. For example, if some earlier stage in a radio transmitter oscillates at a frequency different from the intended signal frequency, and if this undesired signal falls within the design frequency range of the broadband PA, then that signal will undergo amplification along with the desired signal, producing unintended *spurious emission* from the transmitter.

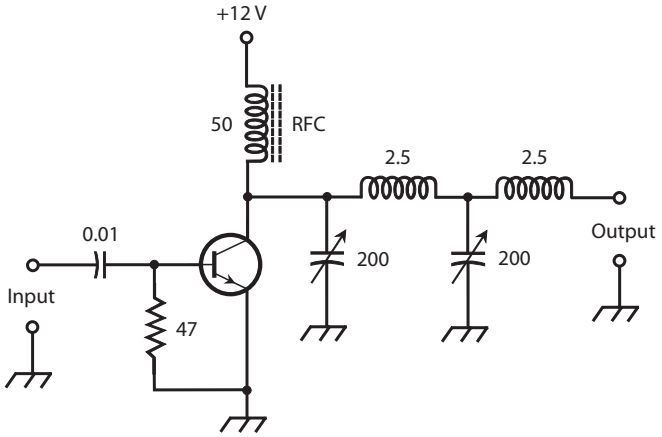
Figure 25-13 is a schematic diagram of a typical broadband PA using an NPN power transistor. This circuit can provide several watts of continuous RF power output over a range of frequencies from 1.5 MHz through 15 MHz. The transformers constitute a critical part of this circuit. They must work efficiently over a 10:1 range of frequencies. The 50-μH component labeled "RFC" is an *RF choke*, which passes DC and low-frequency AC while blocking high-frequency AC (that is, RF signals).

Tuned PAs

A tuned RF PA offers improved efficiency compared with broadband designs. The tuned circuits minimize the risk of spurious signals from earlier stages being amplified and transmitted over the air.



25-13 A broadband RF power amplifier, capable of producing a few watts output. Resistances are in ohms. Capacitances are in microfarads (μF). Inductances are in microhenrys (μH). The 50-μH component labeled "RFC" is an RF choke.



25-14 A tuned RF power amplifier, capable of producing a few watts of output. Resistances are in ohms. Capacitances are in microfarads (μF) if less than 1, and in picofarads (pF) if more than 1. Inductances are in microhenrys (μH).

Tuned PAs can work into a wide range of load impedances. In addition to a *tuning control*, or resonant circuit that adjusts the output of the amplifier to the operating frequency, the tuned amplifier incorporates a *loading control* that optimizes the signal transfer between the amplifier and the load (usually an antenna).

Tuned PAs have one significant limitation: The "tune-up" procedure (usually the adjustment of variable capacitors and/or variable inductors) takes time, and improper adjustment can result in damage to the transistor. If the tuning and/or loading controls aren't properly set, the amplifier efficiency will drop to near zero, while the DC power input remains high. Solid-state devices overheat quickly under these conditions because the excess power "has nowhere to go" except to dissipate itself as heat in the amplifier's components.

Figure 25-14 illustrates a tuned RF PA that can provide a few watts of useful power output at about 10 MHz. The transistor is the same type as the one used in the broadband amplifier of Fig. 25-13. The operator should adjust the tuning and loading controls (left-hand and right-hand variable capacitors, respectively) for maximum power output as indicated by an RF wattmeter.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

26
CHAPTER

Oscillators

ONE WAY TO THINK OF AN OSCILLATOR IS AS A SPECIALIZED AMPLIFIER WITH POSITIVE FEEDBACK. An alternative view of oscillators is as essentially a digital circuit that relies on a timing component such as a capacitor charging or a crystal resonating to produce a square wave. In this chapter, we will look at oscillators from both of those viewpoints.

Radio-frequency oscillators generate signals in a wireless broadcast or communications system. Audio-frequency oscillators find applications in music synthesizers, electronic sirens, security alarms, and electronic toys.

Positive Feedback

Feedback normally occurs either in phase with the input signal, or in phase opposition relative to the input signal. If we want to make an amplifier circuit oscillate, we must introduce some of its output signal back to the input in phase (that's called positive feedback). If we introduce some of the output signal back to the input in phase opposition, we have negative feedback that reduces the overall gain of the amplifier. Negative feedback is not always bad; engineers deliberately use it in some amplifiers to prevent unwanted oscillation.

The AC output signal wave from a common-emitter or common-source amplifier occurs in phase opposition with respect to the input signal wave. If you couple the collector to the base through a capacitor, you won't get oscillation. You must invert the phase in the feedback process if you want oscillation to occur. In addition, the amplifier must exhibit a certain minimum amount of gain, and the coupling from the output to the input must be substantial. The positive feedback path must be easy for a signal to follow. Many oscillators comprise common-emitter or common-source amplifier circuits with positive feedback.

The AC output signal wave from a common-base or common-gate amplifier is in phase with the input signal wave. You might, therefore, suppose that such circuits would make ideal candidates for oscillators. However, the common-base and common-gate circuits produce less gain than their common-emitter and common-source counterparts, so it's more difficult to make them oscillate. Common-collector and common-drain circuits are even worse in this respect because they have negative gain!

Feedback at a Single Frequency

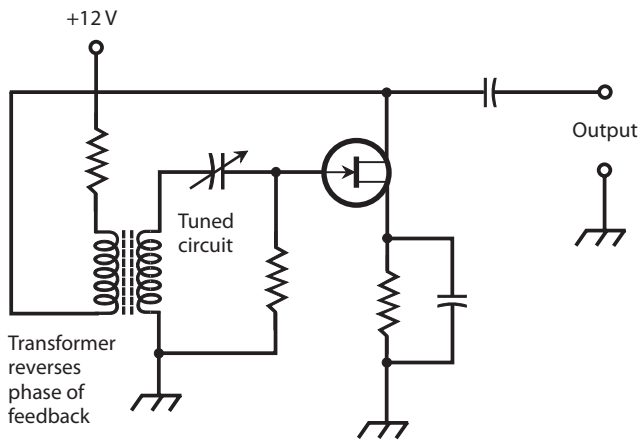
We can control the frequency of an oscillator using tuned, or resonant, circuits, usually consisting of inductance-capacitance (LC) or resistance-capacitance (RC) combinations. The LC scheme is common in radio transmitters and receivers; the RC method is more often used in audio work. The tuned circuit makes the feedback path easy for a signal to follow at one frequency, but difficult to follow at all other frequencies. As a result, oscillation takes place at a stable frequency, determined by the inductance and capacitance, or by the resistance and capacitance.

An Old-School Oscillator Circuit

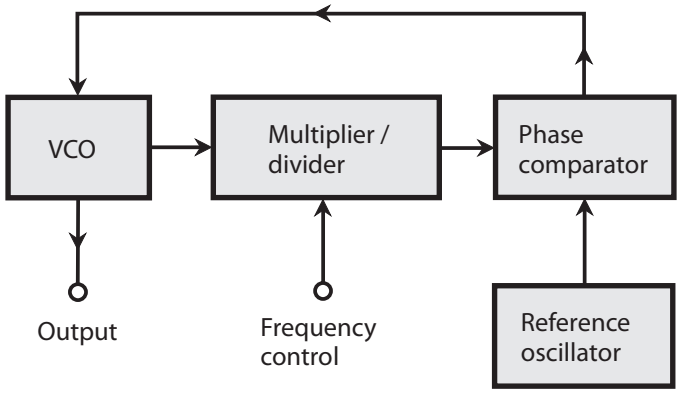
Many circuit arrangements can reliably produce oscillation. As an example, let's look at the Armstrong circuit. We can force a common-emitter or common-source class-A amplifier to oscillate by coupling the output back to the input through a transformer that reverses (inverts) the phase of the fed-back signal. The schematic diagram of Fig. 26-1 shows a common-source amplifier whose drain circuit is coupled to the gate circuit by means of a transformer. We control the frequency by adjusting a capacitor connected in series with the transformer secondary winding. The inductance of the transformer secondary, along with the capacitance, forms a resonant circuit that passes energy easily at one frequency, while attenuating (suppressing) the energy at other frequencies. Engineers call this type of circuit an Armstrong oscillator. We can substitute a bipolar transistor for the JFET, as long as we bias the device for class-A amplification.

The Voltage-Controlled Oscillator

We can adjust the frequency of an analog oscillator by connecting a varactor diode in the tuned LC circuit. Recall that a varactor, also called a varicap, is a semiconductor diode that functions as a variable capacitor when reverse-biased. As the reverse-bias voltage increases, the junction capacitance decreases, provided that we don't apply so much voltage that avalanche breakdown occurs.



26-1 An Armstrong oscillator using an N-channel JFET. This circuit constitutes a common-source amplifier with positive feedback through a tuned circuit.



26-2 Block diagram of a phase-locked loop (PLL).

As you might expect, dedicated ICs are also available that can control the output frequency of an oscillator using a control voltage.

The Phase-Locked Loop

One clever type of oscillator that makes good use of a voltage-controlled oscillator (VCO) is the phase-locked loop (PLL). The PLL makes use of a circuit called a frequency synthesizer. The output of a VCO passes through a programmable multiplier/divider, a digital circuit that divides and/or multiplies the VCO frequency by integral (whole number) values that we can freely choose. As a result, the output frequency can equal any rational-number multiple of the base frequency. We can, therefore, adjust a well-designed PLL circuit in small digital increments over a wide range of frequencies. Figure 26-2 is a block diagram of a PLL.

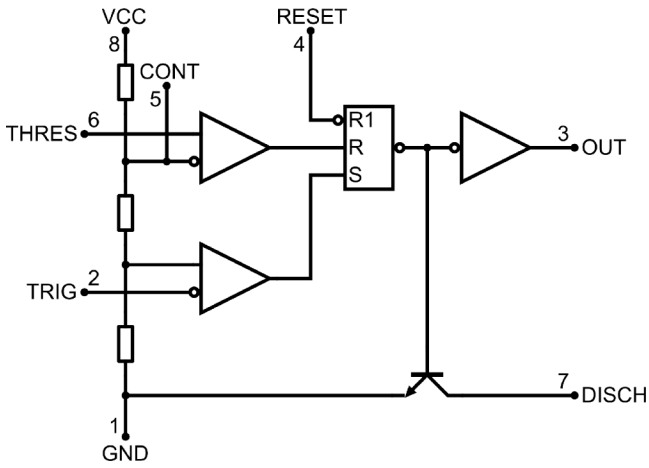
The output frequency of the multiplier/divider remains "locked," by means of a phase comparator, to the signal from a crystal-controlled reference oscillator. As long as the output from the multiplier/divider stays exactly on the reference oscillator frequency, the two signals remain exactly in phase, and the output of the phase comparator equals zero (that is, 0 V DC). If the VCO frequency begins to gradually increase or decrease (a phenomenon known as oscillator drift), the output frequency of the multiplier/divider also drifts, although at a different rate. Even a frequency change of less than 1 Hz causes the phase comparator to produce a DC error voltage. This error voltage is either positive or negative, depending on whether the VCO has drifted higher or lower in frequency. We apply the error voltage to the VCO, its frequency changes in a direction opposite to that of the drift, creating a DC feedback circuit that maintains the VCO frequency at a precise value. Engineers call it a loop circuit that locks the VCO onto a particular frequency by means of phase sensing—hence the expression phase-locked.

The key to the stability of the PLL lies in the fact that the reference oscillator employs crystal or resonator control.

As well as being used in analog circuits such as radio transmitters and receivers, PLLs are also widely used to generate clock signals in digital circuits.

Integrated Circuit Oscillators and Timers

There are a wide variety of oscillator ICs, including VCOs and PLLs. You can also find oscillators designed to have very long periods to be used as timers.



26-3 The NE555 oscillator/timer IC (Wikipedia).

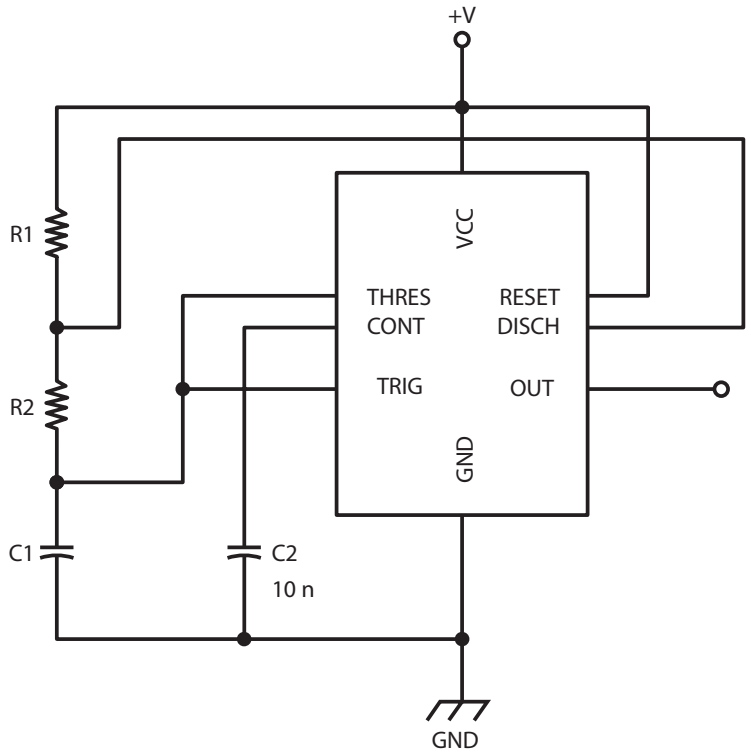
For most modern designs, a low-cost microcontroller might be used in place of a timer IC, especially as most microcontrollers have an internal resonator that can provide quite accurate timing without any external components. The downside of this approach is that, if you use a microcontroller you have to flash a program onto it.

The NE555 Timer IC

The NE555 timer IC (often just called a 555 timer) is of historical significance as one of the most manufactured ICs in history. These ICs cost just a few cents. The popularity of this IC is largely due to its great flexibility. It can be used as an oscillator, but also as a one-shot timer, that produces a single pulse of a certain duration. Figure 26-3 shows how a 555 timer is organized.

The positive supply voltage (VCC) uses a series of three resistors as a voltage divider that provide reference voltages for the positive inputs to the two comparators. The positive inputs to the comparators are available for use as the pins THRES and TRIG. The outputs of the comparators are connected to S (set) and R (reset) pins of a flip-flop (see Chap. 28). The 555 provides two types of output. The OUT pin is a buffered push-pull output capable of sourcing or sinking a very respectable 200 mA. It is intended that the load be connected between this output and either GND or the positive supply. The other output (DISCH) is (for obvious reasons) what's called an open-collector output. This much lower current output is intended to provide feedback to the comparator inputs.

Figure 26-4 shows the 555 timer configured as an oscillator. The CONT pin is connected to GND via a 10 nF capacitor to prevent the otherwise floating input for picking up interference. The THRES and TRIG pins are connected together. When first powered up, C1 is uncharged, and starts to charge through R1 and R2. Until the voltage at the THR and TRIG reaches 1/3 of the supply voltage, the flip-flop will be set and the output will be on and the transistor connected to DISCH will be off. When the voltage at C1 exceeds 2/3 of the supply voltage, the output turns on, as does the open collector transistor, discharging C1 through R2. Once the voltage at C1 falls to below 1/3 of the supply voltage, the flip-flop is reset and the transistor turns off again. This starts C1 charging again and the oscillation continues.



26-4 The ME555 as an oscillator.

The on time (T_{on}) for each cycle is:

$$0.693\ (R1 + R2)\ C$$

The off time (T_{off}) is just:

$$0.693\ R2\ C$$

So, the overall period is:

$$0.693\ (R1 + 2R2)\ C$$

The proportion of the time that the output is high (its "duty cycle") is usually expressed as a percentage and can be calculated as:

$$100\ T_{on}/(T_{on} + T_{off})$$

Exercise 26-1

What will be the frequency and duty cycle of an oscillator using the following values: R1 and R2 1 k and C1 100 nF?

Answer

$T_{on} = 0.693\ (R1 + R2)\ C = 0.693 \times 2\ k \times 100\ n = 138.6$ microseconds
$T_{off} = 0.693\ R2\ C = 0.693 \times 1\ k \times 100\ n = 69.3$ microseconds
Period $T = 136.6 + 69.3 = 207.9$ microseconds
The frequency is therefore 4.81 kHz and the duty cycle:

$$100\ T_{on}/(T_{on} + T_{off}) = 66.67\%$$

You can save yourself a lot of time with these calculations by using an online calculator such as the one available at https://ohmslawcalculator.com/555-astable-calculator.

The 555 timer is a very flexible chip and whole books have been written about applications using the chip. The Internet is a great place to find example circuits using the 555 timer and http://www.555-timer-circuits.com/ is a good place to start.

Direct Digital Synthesis

Modern radio communication hardware using the likes of Bluetooth and WiFi do not generally use any of the oscillator types described above, with the exception of perhaps PLL. Instead they use high-speed digital hardware to create the analog signals needed for radio transmission and reception. This is often included on the same chip as a microcontroller. This approach is called direct digital synthesis (DDS).

Similarly, for lower frequencies, a microcontroller with a built-in digital-to-analog (DAC) is often used to directly generate complex audio waveforms, using mostly software.

Oscillator Stability

In an oscillator, the term "stability" can have either of two distinct meanings: constancy of frequency (or minimal frequency drift), and reliability of performance.

Frequency Stability

In the design and construction of an oscillator of any kind, the components—especially the capacitors and inductors—must, to the greatest extent possible, maintain constant values under all anticipated conditions.

Some types of capacitors hold their values better than others as the temperature rises or falls. Polystyrene capacitors behave very well in this respect. Silver-mica capacitors can work when polystyrene units aren't readily available. Air-core coils exhibit the best temperature stability of all inductor configurations. They should be wound, when possible, from stiff wire with strips of plastic to keep the windings in place. Some air-core coils are wound on hollow cylindrical cores, made of ceramic or phenolic material. Ferromagnetic solenoidal or toroidal cores aren't very good for use in oscillators because these materials change permeability as the temperature varies. This variation alters the inductance, in turn affecting the oscillator frequency.

The best oscillators, in terms of frequency stability, are crystal or resonator-controlled. This category includes circuits that oscillate at the fundamental frequency of the quartz crystal, circuits that oscillate at one of the crystal harmonic frequencies, or PLL circuits that oscillate at frequencies derived from the crystal frequency by means of programmable multiplier/dividers.

Reliability

An oscillator should always start working as soon as we apply DC power. It should keep oscillating under all normal conditions. The failure of a single oscillator can cause an entire receiver, transmitter, or transceiver to stop working.

Oscillators are designed to work into relatively high-load impedances. If we connect an oscillator to a load that has a low impedance, that load will "try" to draw a lot of power from the oscillator. Under such conditions, even a well-designed oscillator might stop working or not start up, when we first switch it on. With the exception of ICs like the 555 timer with a buffered output, oscillators aren't meant to produce powerful signals; we can use amplifiers for that purpose! You need never worry that an oscillator's load impedance might get too high. In general, as we increase the load impedance for an oscillator, its overall performance improves.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

27
CHAPTER

Wireless Transmitters and Receivers

IN WIRELESS COMMUNICATIONS, A *TRANSMITTER* CONVERTS DATA INTO *ELECTROMAGNETIC* (EM) *WAVES* intended for recovery by one or more *receivers*. In this chapter, we'll learn how to convert data to an *EM field*, and then learn how we can intercept and decode that field at remote points.

Modulation

When we *modulate* a wireless signal, we "write" data onto an EM wave. We can carry out this process by varying the amplitude, the frequency, or the phase of the wave. We can also obtain a modulated signal by generating a series of multiple-wave pulses and varying their duration, amplitude, or timing. The heart of a wireless signal comprises a sine wave called the *carrier* whose frequency can range from a few kilohertz (kHz) to many gigahertz (GHz). If we expect effective data transfer, the carrier frequency must be at least 10 times the highest frequency of the modulating signal.

The data that we modulate may be analog or digital in nature. With the demise of analog TV transmission, FM radio is one of the few remaining analog radio signals in widespread use. However, the basic principals of modulation of a carrier wave remain the same.

On/Off Keying

The simplest form of modulation involves *on/off keying* of the carrier. We can *key* the oscillator of a radio transmitter to send *Morse code*, one of the simplest known *binary digital* modulation modes. The duration of a Morse-code *dot* equals the duration of one *binary digit*, more often called a *bit*. (A binary digit is the smallest or shortest possible unit of data in a system whose only two states are "on" and "off.") A *dash* measures three bits in duration. The space between dots and dashes within a *character* equals one bit; the space between characters in a *word* equals three bits; the space between words equals seven bits. Some technicians refer to the *key-down* (full-carrier) condition as *mark* and the *key-up* (no-signal) condition as *space*. Amateur radio operators who enjoy using the Morse code send and receive it at speeds ranging from about 5 words per minute (wpm) to around 60 wpm.

Frequency-Shift Keying

We can send digital data faster and with fewer errors than Morse code allows if we use *frequency-shift keying* (FSK). In some FSK systems, the carrier frequency shifts between mark and space conditions, usually by a few hundred hertz or less. In other systems, a two-tone audio-frequency (AF) sine wave modulates the carrier, a mode known as *audio-frequency-shift keying* (AFSK). The two most common codes used with FSK and AFSK are *Baudot* (pronounced "baw-DOE") and *ASCII* (pronounced "ASK-ee"). The acronym ASCII stands for *American Standard Code for Information Interchange*.

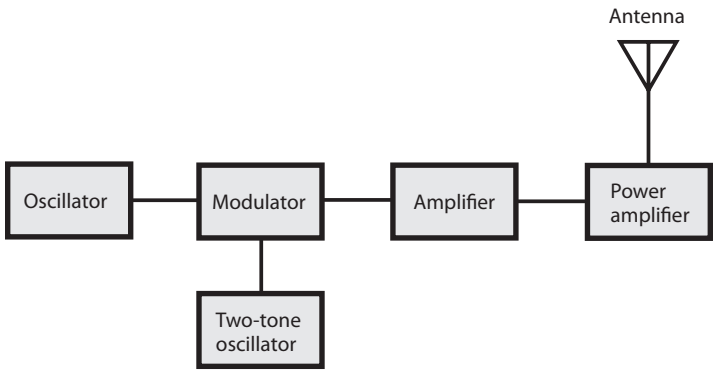
In *radioteletype* (RTTY), FSK, and AFSK systems, a *terminal unit* (TU) converts the digital signals into electrical impulses that operate a teleprinter or display characters on a computer screen. The TU also generates the signals necessary to send RTTY when an operator types on a keyboard. A device that sends and receives AFSK is sometimes called a *modem*, an acronym that stands for *modulator/demodulator*. A modem is basically the same as a TU. Figure 27-1 is a block diagram of an AFSK transmitter.

The main reason why FSK or AFSK work better than on/off keying is the fact that the space signals are identified as such, rather than existing as mere gaps in the data. A sudden noise burst in an on/off keyed signal can "confuse" a receiver into falsely reading a space as a mark, but when the space is positively represented by its own signal, this type of error happens less often at any given data speed.

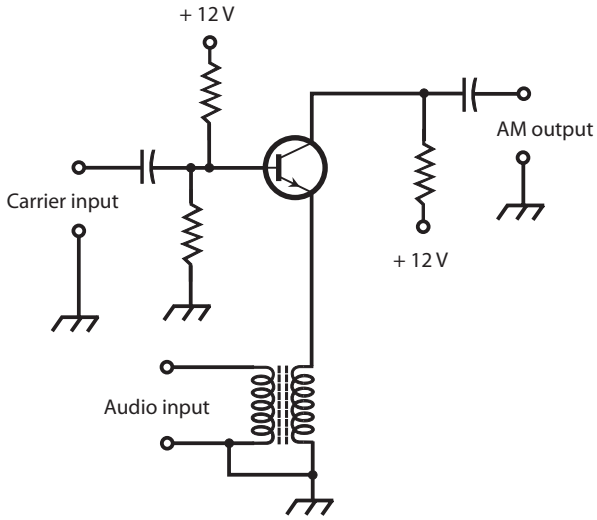
Amplitude Modulation

An AF voice signal has frequencies mostly in the range between 300 Hz and 3 kHz. We can modulate some characteristic of an RF carrier with an AF voice waveform, thereby transmitting the voice information over the airwaves. Figure 27-2 shows a simple circuit for obtaining *amplitude modulation* (AM). We can imagine this circuit as an RF amplifier for the carrier, with the instantaneous gain dependent on the instantaneous audio input amplitude. We can also think of the circuit as a mixer that combines the RF carrier and the audio signals to produce sum and difference signals above and below the carrier frequency.

The circuit shown in Fig. 27-2 performs well as long as we don't let the AF input amplitude get too great. If we inject too much audio, we get *distortion* (nonlinearity) in the transistor resulting in degraded *intelligibility* (understandability), reduced *circuit efficiency* (ratio of DC power input to useful power output), and excessive output signal *bandwidth* (the difference between the highest and



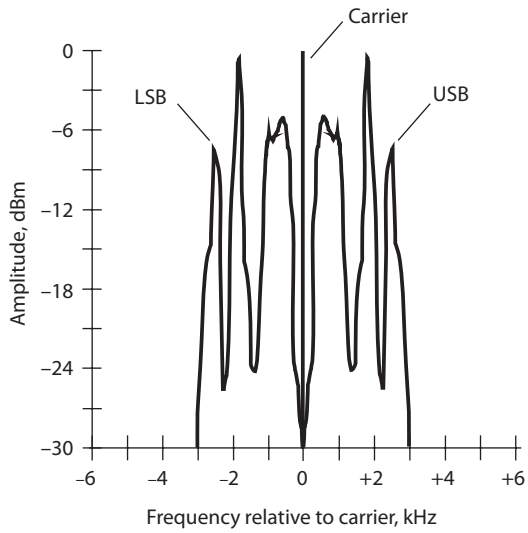
27-1 Simplified block diagram of a transmitter that uses audio-frequency-shift keying (AFSK).



27-2 An amplitude modulator using an NPN bipolar transistor.

lowest component frequency). We can express the modulation extent as a percentage ranging from 0 percent, representing an *unmodulated carrier*, to 100 percent, representing the maximum possible modulation we can get without distortion. If we increase the modulation beyond 100 percent, we observe the same problems as we do when we apply excessive AF input to a modulator circuit, such as the one shown in Fig. 27-2. In an AM signal that's modulated at 100 percent, we find that ⅓ of the signal power conveys the data, while the carrier wave consumes the other ⅔ of the power.

Figure 27-3 shows a *spectral display* of an AM voice radio signal. The horizontal scale is calibrated in increments of 1 kHz per division. Each vertical division represents 3 dB of change in signal strength. The maximum (reference) amplitude equals 0 dB relative to 1 mW (abbreviated as 0 dBm). The data exists in *sidebands* above and below the carrier frequency. These sidebands constitute sum and difference signals produced by mixing in the modulator circuit between the audio and



27-3 Spectral display of a typical amplitude-modulated (AM) voice communications signal.

the carrier. The RF energy between −3 kHz and the carrier frequency is called the *lower sideband*; the RF energy from the carrier frequency to +3 kHz is called the *upper sideband*.

The signal bandwidth equals the difference between the maximum and minimum sideband frequencies. In an AM signal, the bandwidth equals twice the highest audio modulating frequency. In the example of Fig. 27-3, all the AF voice energy exists at or below 3 kHz, so the signal bandwidth equals 6 kHz, typical of AM voice communications. In standard AM broadcasting in which music is transmitted along with voices, the AF energy is spread over a wider bandwidth, nominally 10 kHz to 20 kHz. The increased bandwidth provides for better *fidelity* (sound quality).

Single Sideband

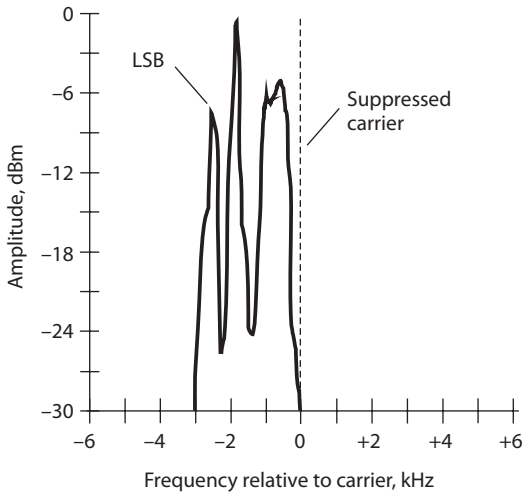
In AM at 100 percent modulation, the carrier wave consumes ⅔ of the signal power, and the sidebands exist as mirror-image duplicates that, combined, employ only ⅓ of the signal power. These properties make AM inefficient and needlessly redundant.

If we could get rid of the carrier and one of the sidebands, we'd still convey all the information we want while consuming far less power. Alternatively, we could get a stronger signal for a given amount of RF power. We could also reduce the signal bandwidth to a little less than half that of an AM signal modulated with the same data. The resulting *spectrum savings* would allow us to fit more than twice as many signals into a specific range, or *band*, of frequencies. During the early twentieth century, communications engineers perfected a way to modify AM signals in this way. They called the resulting mode *single sideband* (SSB), a term which endures to this day.

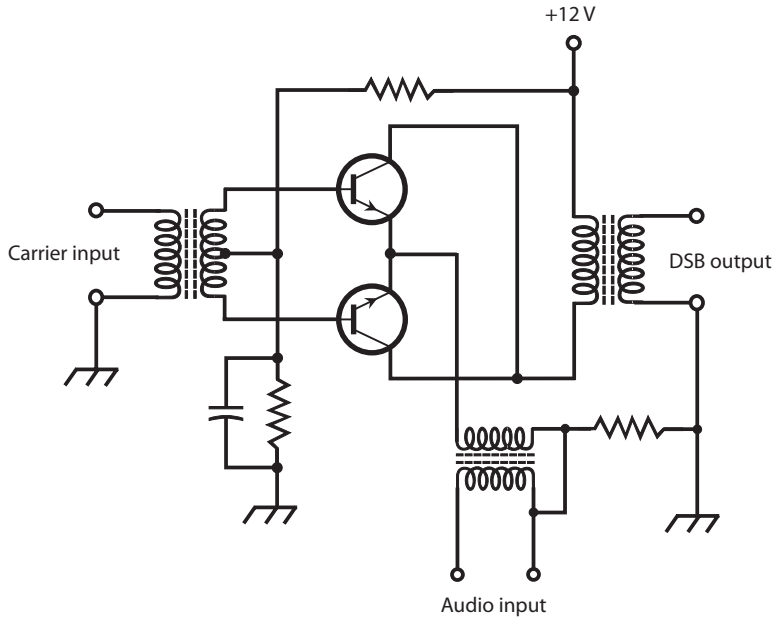
When we remove the carrier and one of the sidebands from an AM signal, the remaining energy has a spectral display resembling Fig. 27-4. In this case, we eliminate the upper sideband along with the carrier, leaving only the lower sideband. We could just as well remove the LSB along with the carrier, leaving only the upper sideband.

Balanced Modulator

We can almost completely suppress the carrier in an AM signal using a *balanced modulator*—an amplitude modulator/amplifier using two transistors with the inputs connected in push-pull and the outputs connected in parallel, as shown in Fig. 27-5. This arrangement "cancels" the carrier wave in



27-4 Spectral display of a typical single-sideband (SSB) voice communications signal, in this case lower sideband.



27-5 A balanced modulator using two NPN bipolar transistors. We connect the bases in push-pull and the collectors in parallel.

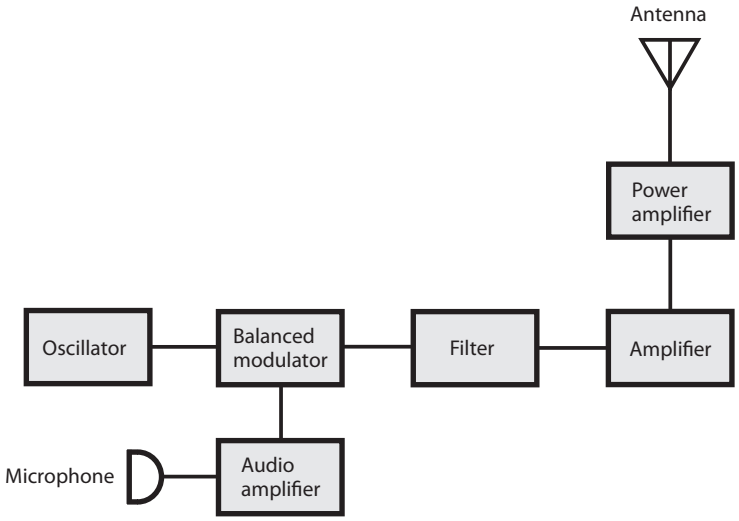
the output signals, leaving only lower sideband and upper sideband energy. The balanced modulator produces a *double-sideband suppressed-carrier* (DSBSC) signal, often called simply *double sideband* (DSB). One of the sidebands can be suppressed in a subsequent circuit by a *bandpass filter* to obtain an SSB signal.

Basic SSB Transmitter

Figure 27-6 is a block diagram of a simple SSB transmitter. The RF amplifiers that follow any type of amplitude modulator, including a balanced modulator, must all operate in a linear manner to prevent distortion and unnecessary spreading of the signal bandwidth, a condition that some engineers and radio operators call *splatter*. These amplifiers generally work in class A except for the PA, which operates in class AB or class B. We'll never see a class-C amplifier as the PA in an SSB transmitter because class-C operation distorts any signal whose amplitude varies over a continuous range.

Frequency Modulation

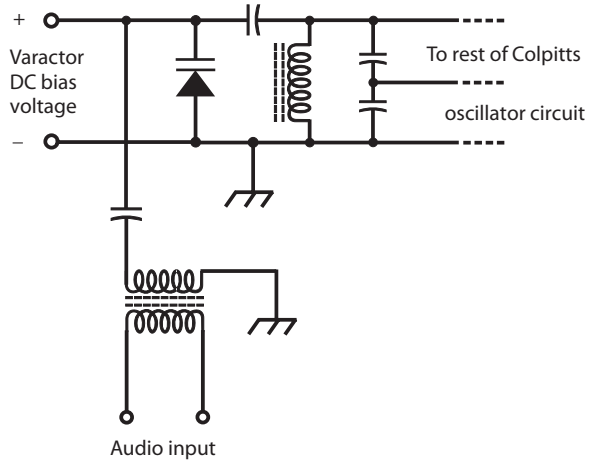
In *frequency modulation* (FM), the instantaneous signal amplitude remains constant; the instantaneous frequency varies instead. We can get FM by applying an audio signal to the varactor diode in a voltage-controlled oscillator (VCO). Figure 27-7 shows an example of this scheme, known as *reactance modulation*. This circuit employs a Colpitts oscillator, but we could use any other type of oscillator and get similar results. The varying voltage across the varactor causes its capacitance to change in accordance with the audio waveform. The fluctuating capacitance causes variations in the resonant frequency of the inductance-capacitance (*LC*) tuned circuit, causing small fluctuations in the frequency generated by the oscillator.



27-6 Block diagram of a basic SSB transmitter.

Phase Modulation

We can indirectly obtain FM if we modulate the phase of the oscillator signal. When we vary the phase from instant to instant, we inevitably provoke small variations in the frequency as well. Any instantaneous phase change shows up as an instantaneous frequency change (and vice-versa). When we employ *phase modulation* (PM), we must process the audio signal before we apply it to the modulator, adjusting the *frequency response* of the audio amplifiers. Otherwise the signal will sound muffled when we listen to it in a receiver designed for ordinary FM.



27-7 Generation of a frequency-modulated (FM) signal by employing reactance modulation in a Colpitts oscillator. We can modify other oscillator types in a similar way.

Deviation for FM and PM

In an FM or PM signal, we can quantify the maximum extent to which the instantaneous carrier frequency differs from the unmodulated-carrier frequency in terms of a parameter called *deviation*. For most FM and PM voice transmitters, the deviation is standardized at ±5 kHz. We call this mode *narrowband FM* (NBFM). The bandwidth of an NBFM signal roughly equals that of an AM signal containing the same modulating information. In FM hi-fi music broadcasting, and in some other applications, the deviation exceeds ±5 kHz, giving us a mode called *wideband FM* (WBFM).

The deviation obtainable with FM is greater, for a given oscillator frequency, than the deviation that we get with PM. However, we can increase the deviation of any FM or PM signal with the help of a *frequency multiplier*. When the signal passes through a frequency multiplier, the deviation gets multiplied along with the carrier frequency. The deviation in the final output should equal the highest modulating audio frequency if we expect optimum audio fidelity. Therefore, ±5 kHz is more than enough deviation for voice communications. For music, a deviation of ±15 kHz or ±20 kHz is required for good reproduction.

Modulation Index for FM and PM

In any FM or PM signal, the ratio of the frequency deviation to the highest modulating audio frequency is called the *modulation index*. Ideally, this figure should be somewhere between 1:1 and 2:1. If it's less than 1:1, the signal sounds muffled or distorted, and efficiency is sacrificed. Increasing the modulation index much beyond 2:1 broadens the bandwidth without providing significant improvement in intelligibility or fidelity.

Power Amplification for FM and PM

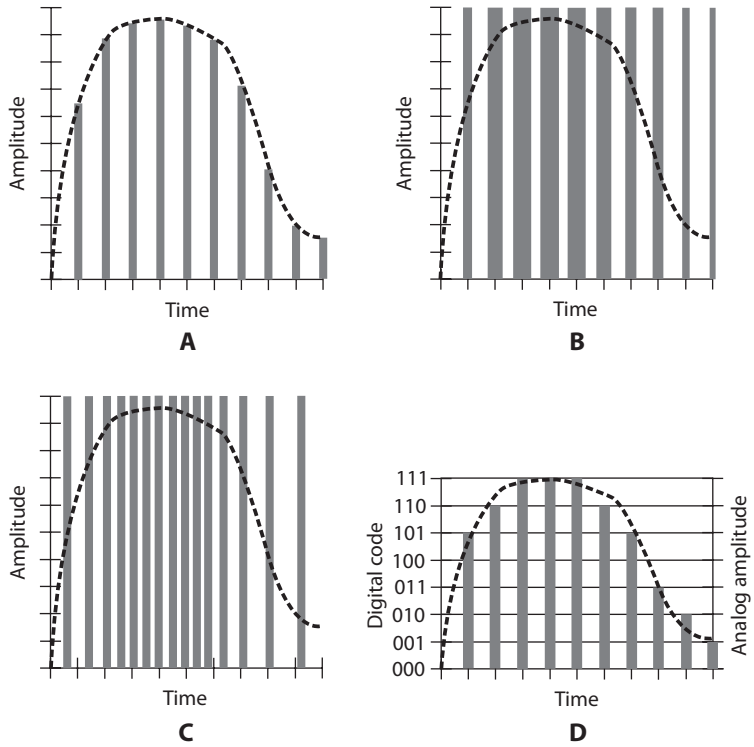
A class-C PA can function in an FM or PM transmitter without causing distortion because the signal amplitude remains constant. Nonlinearity (characteristic of class-C operation) has no meaning, let alone any adverse effects, when the signal amplitude never changes! For this reason, we'll often find class-C PAs in FM and PM transmitters, especially those that have high output power. Remember: Class C offers the best efficiency of any PA mode!

Pulse-Amplitude Modulation

We can modulate a signal by varying some aspect of a constant stream of signal pulses. In *pulse-amplitude modulation* (PAM), the strength of each individual pulse varies according to the modulating waveform. In this respect, PAM resembles AM. Figure 27-8A shows an amplitude-versus-time graph of a hypothetical PAM signal. The modulating waveform appears as a dashed curve, and the pulses appear as vertical gray bars. Normally, the pulse amplitude increases as the instantaneous modulating-signal level increases (*positive PAM*). But this situation can be reversed, so higher audio levels cause the pulse amplitude to go down (*negative PAM*). Then the signal pulses are at their strongest when there is no modulation. The transmitter works harder to produce negative PAM than it does to produce positive PAM.

Pulse-Width Modulation

We can modulate the output of an RF transmitter output by varying the width (duration) of signal pulses to obtain *pulse-width modulation* (PWM), also known as *pulse duration modulation* (PDM) as shown in Fig. 27-8B. Normally, the pulse width increases as the instantaneous modulating-signal



27-8 Time-domain graphs of various modes of pulse modulation. At A, pulse-amplitude modulation (PAM); at B, pulse-width modulation (PWM), also called pulse-duration modulation (PDM); at C, pulse-interval modulation (PIM); at D, pulse-code modulation (PCM).

level increases (*positive PWM*). But this situation can be reversed (negative *PWM*). The transmitter must work harder to accomplish negative PWM. Either way, the peak pulse amplitude remains constant.

Pulse-Interval Modulation

Even if all the pulses have the same amplitude and the same duration, we can obtain pulse modulation by varying how often the pulses occur. In PAM and PWM, we always transmit the pulses at the same time interval, known as the *sampling interval*. But in *pulse interval-modulation* (PIM), pulses can occur more or less frequently than they do under conditions of no modulation. Figure 27-8C shows a hypothetical PIM signal. Every pulse has the same amplitude and the same duration, but the time interval between them changes. When there is no modulation, the pulses emerge from the transmitter evenly spaced with respect to time. An increase in the instantaneous data amplitude might cause pulses to be sent more often, as is the case in Fig. 27-8C (*positive PIM*). Alternatively, an increase in instantaneous data level might slow down the rate at which the pulses emerge (*negative PIM*).

Pulse-Code Modulation

In *digital communications*, the modulating data attains only certain defined states, rather than continuously varying. Compared with old-fashioned *analog communications* (where the state is always continuously variable), digital modes offer improved *signal-to-noise* (S/N) *ratio*, narrower signal bandwidth, better accuracy, and superior reliability. In *pulse-code modulation* (PCM), any of the above-described aspects—amplitude, width, or interval—of a pulse sequence (or *pulse train*) can be varied. But rather than having infinitely many possible states, the number of states equals some power of 2, such as 2^2 (four states), 2^3 (eight states), 2^4 (16 states), 2^5 (32 states), 2^6 (64 states), and so on. As we increase the number of states, the fidelity and data transmission speed improve, but the signal gets more complicated. Figure 27-8D shows an example of eight-level PCM.

Analog-to-Digital Conversion

Pulse-code modulation, such as we see in Fig. 27-8D, serves as a common form of *analog-to-digital* (A/D) *conversion*. A voice signal, or any continuously variable signal, can be *digitized*, or converted into a train of pulses whose amplitudes can achieve only certain defined levels.

In A/D conversion, because the number of states always equals some power of 2, we can represent the signal as a binary-number code. Fidelity improves as the exponent increases. The number of states is called the *sampling resolution*, or simply the *resolution*. A resolution of $2^3 = 8$ (as shown in Fig. 27-8D) is good enough for basic voice communications. A resolution of $2^4 = 16$ can provide fairly decent music reproduction.

The efficiency with which we can digitize a signal depends on the frequency at which we carry out the sampling. In general, the *sampling rate* must be at least twice the highest data frequency. For an audio signal with components as high as 3 kHz, the minimum sampling rate for effective digitization is 6 kHz. For music it's higher, of course.

The Electromagnetic Field

In a radio or television transmitting antenna, electrons constantly move back and forth. Their velocity constantly changes as they speed up in one direction, slow down, reverse direction, speed up again, and so on. Any change of velocity (speed and/or direction) constitutes *acceleration*. When charged particles accelerate in a certain way, they produce an *electromagnetic* (EM) *field*.

How It Happens

When electrons move, they generate a magnetic (M) field. When electrons accelerate, they generate a *changing* M field. When electrons accelerate back and forth, they generate an *alternating* M field at the same frequency as that of the electron motion.

An alternating M field gives rise to an alternating electric (E) field, which, in turn, spawns another alternating M field. This process repeats indefinitely in the form of an EM field that *propagates* (travels) through space at the speed of light. The E and M fields expand alternately outward from the source in spherical wavefronts. At any given point in space, the lines of E flux run perpendicular to the lines of M flux. The waves propagate in a direction perpendicular to both the E and M flux lines.

Frequency versus Wavelength

All EM fields have two important properties: the *frequency* and the *wavelength*. When we quantify them, we find that they exhibit an *inverse relation:* as one increases, the other decreases. We've already learned about AC frequency. We can express EM wavelength as the physical distance between any two adjacent points at which either the E field or the M field has identical amplitude and direction.

An EM field can have any conceivable frequency, ranging from centuries per cycle to quadrillions of cycles per second (or hertz). The sun has a magnetic field that oscillates with a 22-year cycle. Radio waves oscillate at thousands, millions, or billions of hertz. Infrared (IR), visible light, ultraviolet (UV), X rays, and gamma rays comprise EM fields that alternate at many trillions (million millions) of hertz. The wavelength of an EM field can likewise vary over the widest imaginable range, from many trillions of miles to a tiny fraction of a millimeter.

Let f_{MHz} represent the frequency of an EM wave in megahertz as it travels through free space. (Technically, free space constitutes a vacuum, but we can consider the air at the earth's surface equivalent to free space for most "real-world" applications.) Let L_{ft} represent the wavelength of the same wave in feet. Then

$$L_{ft} = 984/f_{MHz}$$

If we want to express the wavelength L_m in meters, then

$$L_m = 300/f_{MHz}$$

The inverses of these formulas are

$$f_{MHz} = 984/L_{ft}$$

and

$$f_{MHz} = 300/L_m$$

Velocity Factor

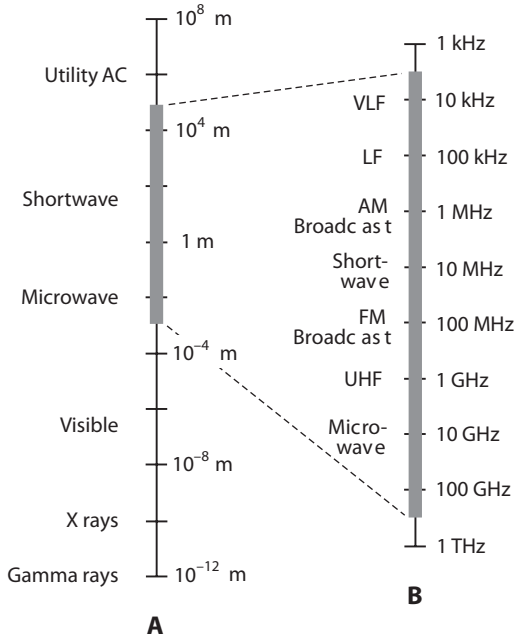
In media other than free space, EM fields propagate at less than the speed of light. As a result, the wavelength grows shorter according to a quantity called the *velocity factor*, symbolized v. The value of v can range from 0 (representing no movement at all) to 1 (representing the speed of propagation in free space, which equals approximately 186,000 mi/s or 300,000 km/s). We can also express the velocity factor as a percentage $v_{\%}$. In that case, the smallest possible value is 0 percent, and the largest is 100 percent. The velocity factor in practical situations rarely falls below 0.50 or 50 percent, and it usually exceeds 0.60 or 60 percent.

Velocity factor constitutes a crucial parameter in the design of RF transmission lines and antenna systems, when sections of cable, wire, or metal tubing must be cut to specific lengths measured in wavelengths or fractions of a wavelength. Taking the velocity factor v, expressed as a ratio, into account, we can modify the above-mentioned four formulas as follows:

$$L_{ft} = 984v/f_{MHz}$$
$$L_m = 300v/f_{MHz}$$
$$f_{MHz} = 984v/L_{ft}$$
$$f_{MHz} = 300v/L_m$$

The EM and RF Spectra

Physicists, astronomers, and engineers refer to the entire range of EM wavelengths as the *electromagnetic* (EM) *spectrum*. Scientists use logarithmic scales to depict the EM spectrum according to the wavelength in meters, as shown in Fig. 27-9A. The radio-frequency (RF) *spectrum*, which includes radio, television, and microwaves, appears expanded in Fig. 27-9B, where we label the axis



27-9 At A, the electromagnetic (EM) spectrum from 10^8 m to 10^{-12} m. Each vertical division represents two orders of magnitude (a 100-fold increase or decrease in the wavelength). At B, the radio-frequency (RF) portion of the EM spectrum, with each vertical division representing one order of magnitude (a 10-fold increase or decrease in the wavelength).

according to frequency. The RF spectrum is categorized in *bands* from *very low frequency* (VLF) through *extremely high frequency* (EHF), according to the breakdown in Table 27-1. The exact lower limit of the VLF range is a matter of disagreement in the literature. Here, we define it as 3 kHz.

Wave Propagation

Radio-wave propagation has fascinated scientists ever since Marconi and Tesla discovered, around the year 1900, that EM fields can travel over long distances without any supporting infrastructure. Let's examine some wave-propagation behaviors that affect wireless communications at radio frequencies.

Polarization

We can define the orientation of E-field *lines of flux* as the *polarization* of an EM wave. If the E-field flux lines run parallel to the earth's surface, we have *horizontal polarization*. If the E-field flux lines

Table 27-1. Bands in the RF Spectrum

Frequency Designation	Frequency Range	Wavelength Range
Very Low (VLF)	3 kHz–30 kHz	100 km–10 km
Low (LF)	30 kHz–300 kHz	10 km–1 km
Medium (MF)	300 kHz–3 MHz	1 km–100 m
High (HF)	3 MHz–30 MHz	100 m–10 m
Very High (VHF)	30 MHz–300 MHz	10 m–1 m
Ultra High (UHF)	300 MHz–3 GHz	1 m–100 mm
Super High (SHF)	3 GHz–30 GHz	100 mm–10 mm
Extremely High (EHF)	30 GHz–300 GHz	10 mm–1 mm

run perpendicular to the surface, we have *vertical polarization*. Polarization can also have a "slant," of course.

In some situations, the E-field flux lines rotate as the wave travels through space. In that case we have *circular polarization* if the E-field intensity remains constant. If the E-field intensity is more intense in some planes than in others, we have *elliptical polarization*. A circularly or elliptically polarized wave can rotate either *clockwise* or *counterclockwise* as we watch the wavefronts come toward us. The rotational direction is called the *sense* of polarization. Some engineers use the term *right-hand* instead of clockwise and the term *left-hand* instead of counterclockwise.

Line-of-Sight Wave

Electromagnetic waves follow straight lines unless something makes them bend. *Line-of-sight* propagation can often take place when the receiving antenna can't be seen visually from the transmitting antenna because radio waves penetrate nonconducting opaque objects, such as trees and frame houses, to some extent. The line-of-sight wave consists of two components called the *direct wave* and the *reflected wave*, as follows:

1. In the direct wave, the longest wavelengths are least affected by obstructions. At very low, low, and medium frequencies, direct waves can *diffract* around things. As the frequency rises, especially above about 3 MHz, obstructions have a greater and greater blocking effect on direct waves.
2. In the reflected wave, the EM energy reflects from the earth's surface and from conducting objects like wires and steel beams. The reflected wave always travels farther than the direct wave. The two waves might arrive at the receiving antenna in perfect phase coincidence, but usually they don't.

If the direct and reflected waves arrive at the receiving antenna with equal strength but 180° out of phase, we observe a *dead spot*. The same effect occurs if the two waves arrive inverted in phase with respect to each other (that is, in phase opposition). The dead-spot phenomenon is most noticeable at the highest frequencies. At VHF and UHF, an improvement in reception can sometimes result from moving the transmitting or receiving antenna only a few inches or centimeters! In mobile operation, when the transmitter and/or receiver are moving, multiple dead spots produce rapid, repeated interruptions in the received signal, a phenomenon called *picket fencing*.

Surface Wave

At frequencies below about 10 MHz, the earth's surface conducts AC quite well, so vertically polarized EM waves can follow the surface for hundreds or thousands of miles, with the earth helping to transmit the signals. As we reduce the frequency and increase the wavelength, we observe decreasing *ground loss*, and the waves can travel progressively greater distances by means of *surface-wave propagation*. Horizontally polarized waves don't travel well in this mode because the conductive surface of the earth "shorts out" horizontal E flux. At frequencies above about 10 MHz (corresponding to wavelengths shorter than roughly 30 m), the earth becomes lossy, and surface-wave propagation rarely occurs for distances greater than a few miles.

Sky-Wave EM Propagation

Ionization in the upper atmosphere, caused by solar radiation, can return EM waves to the earth at certain frequencies. The so-called *ionosphere* has several dense zones of ionization that occur at fairly constant, predictable altitudes.

The *E layer*, which lies about 50 mi (roughly 80 km) above the surface, exists mainly during the day, although nighttime ionization is sometimes observed. The E layer can provide medium-range radio communication at certain frequencies.

At higher altitudes, we find the *F_1 layer* and the *F_2 layer*. The F_1 layer, normally present only on the daylight side of the earth, forms at about 125 mi (roughly 200 km) altitude; the F_2 layer exists at about 180 mi (roughly 300 km) over most, or all, of the earth, the dark side as well as the light side. Sometimes the distinction between the F_1 and F_2 layers is ignored, and they are spoken of together as the *F layer*. Communication by means of F-layer propagation can usually be accomplished between any two points on the earth at some frequencies between 5 MHz and 30 MHz.

The lowest ionized region is called the *D layer*. It exists at an altitude of about 30 mi (roughly 50 km), and is ordinarily present only on the daylight side of the planet. This layer absorbs radio waves at some frequencies, impeding long-distance ionospheric propagation.

Tropospheric Propagation

At frequencies above about 30 MHz (wavelengths shorter than about 10 m), the lower atmosphere bends radio waves toward the surface. *Tropospheric bending* occurs because the *index of refraction* of air, with respect to EM waves, decreases with altitude. Tropospheric bending makes it possible to communicate for hundreds of miles, even when the ionosphere will not return waves to the earth.

Ducting is tropospheric propagation that occurs somewhat less often than bending, but offers more dramatic effects. Ducting takes place when EM waves get "trapped" within a layer of cool, dense air sandwiched between two layers of warmer air. Like bending, ducting occurs almost entirely at frequencies above 30 MHz.

Still another tropospheric-propagation mode is called *tropospheric scatter,* or *troposcatter*. This phenomenon takes place because air molecules, dust grains, and water droplets scatter some of the EM field. We observe troposcatter most commonly at VHF and UHF. Troposcatter always occurs to some extent, regardless of weather conditions.

Tropospheric propagation in general, without mention of the specific mode, is sometimes called *tropo*.

Auroral Propagation

In the presence of unusual solar activity, the *aurora* (northern lights or southern lights) can return radio waves to the earth, facilitating *auroral propagation*. The aurora occurs at altitudes of about 40 to 250 mi (roughly 65 to 400 km). Theoretically, auroral propagation is possible, when the aurora are active, between any two points on the earth's surface from which the same part of the aurora lie on a line of sight. Auroral propagation seldom occurs when either the transmitting station or the receiving station is located at a latitude less than 35° north or south of the equator.

Auroral propagation causes rapid and deep signal fading, which nearly always renders analog voice and video signals unintelligible. Digital modes work somewhat better, but the carrier frequency gets "spread out" or "smeared" over a band several hundred hertz wide as a result of phase modulation induced by auroral motion. This "spectral spreading" limits the maximum data transfer rate. Auroral propagation commonly takes place along with poor ionospheric propagation resulting from sudden eruptions called *solar flares* on the sun's surface.

Meteor-Scatter Propagation

Meteors produce ionized trails that persist for a fraction of a second up to several seconds. The exact duration of the trail depends on the size of the meteor, its speed, and the angle at which it

enters the atmosphere. A single meteor trail rarely lasts long enough to allow transmission of much data. However, during a *meteor shower*, multiple trails can produce almost continuous ionization for a period of hours. Ionized regions of this type can reflect radio waves at certain frequencies. Communications engineers call this effect *meteor-scatter propagation*, or sometimes simply *meteor scatter*. It can take place at frequencies far above 30 MHz and over distances ranging from just beyond the horizon up to about 1500 mi (roughly 2400 km). The maximum communications range depends on the altitude of the ionized trail, and also on the relative positions of the trail, the transmitting station, and the receiving station.

Moonbounce Propagation

Earth-moon-earth (EME) communications, also called *moonbounce*, is routinely carried on by amateur radio operators at VHF and UHF. This mode requires a sensitive receiver using a low-noise preamplifier, a large, directional antenna, and a high-power transmitter. Digital modes work far better than analog modes for moonbounce.

Signal *path loss* presents the main difficulty for anyone who contemplates EME communications. Received EME signals are always weak. High-gain directional antennas must remain constantly aimed at the moon, a requirement that dictates the use of steerable antenna arrays. The EME *path loss* increases with increasing frequency, but this effect is offset by the more manageable size of high-gain antennas as the wavelength decreases.

Solar noise can pose a problem; EME communications becomes most difficult near the time of the new moon, when the moon lies near a line between the earth and the sun. The sun constitutes a massive broadband generator of EM energy! Problems can also occur with *cosmic noise* when the moon passes near "noisy" regions in the so-called *radio sky*. The constellation *Sagittarius* lies in the direction of the center of the Milky Way galaxy, and EME performance suffers when the moon passes in front of that part of the stellar background.

The moon keeps the same face more or less toward the earth at all times, but some back-and-forth "wobbling" occurs. This motion, called *libration* (not "liberation" or "libation"!), produces rapid, deep fluctuations in signal strength, a phenomenon known as *libration fading*. The fading becomes more pronounced as the operating frequency increases. It occurs as multiple transmitted EM wavefronts reflect from various "lunagraphical" features, such as craters and mountains on the moon's surface, whose relative distances constantly change because of libration. The reflected waves recombine in constantly shifting phase at the receiving antenna, sometimes reinforcing, and at other times canceling.

Transmission Media

Data can be transmitted over various *media*, which include *cable*, *radio* (also called *wireless*), *satellite links* (a specialized form of wireless), and *fiberoptics*. Cable, radio/TV, and satellite communications use the RF spectrum. Fiberoptics uses IR or visible light energy.

Cable

The earliest cables comprised plain wires that carried DC. Nowadays, data-transmission cables more often carry RF signals that can be amplified at intervals on a long span. The use of such amplifiers, called *repeaters*, greatly increases the distances over which data can be sent by cable. Another advantage of using RF is the fact that numerous signals can travel over a single cable, with each signal on a different frequency.

Cables can consist of pairs of wires, somewhat akin to lamp cords. But more often coaxial cable, of the type described and illustrated at the end of Chap. 10, is used. This type of cable has a center conductor that carries the signals, surrounded by a cylindrical, grounded shield that keeps signals confined to the cable, and also keeps external EM fields from interfering with the signals.

Radio

All radio and TV signals consist of EM waves traveling through the earth's atmosphere or outer space. In a radio transmitting station, the RF output goes into an *antenna system* located at some distance from the transmitter. To get from the transmitter's final amplifier to the antenna, the EM energy follows a *transmission line*, also called a *feed line*.

Most radio antenna transmission lines consist of coaxial cable. Other types of cable exist for special applications. At microwave frequencies, hollow tubes called *waveguides* can transfer the energy. A waveguide works more efficiently than coaxial cable at the shortest radio wavelengths.

Radio amateurs sometimes use a *parallel-wire* transmission line in which the RF currents in the two conductors are in phase opposition so their EM fields cancel each other out. This phase cancellation keeps the transmission line from radiating, guiding the EM field along toward the antenna.

Satellite Systems

At very high frequencies (VHF) and above, some communications circuits use satellites that follow *geostationary orbits* around the earth. If a satellite orbits directly over the equator at an altitude of approximately 35,800 km (22,200 mi) and travels from west to east, it follows the earth's rotation, thereby staying in the same spot in the sky as seen from the surface. That's why we call it a *geostationary satellite*.

A single geostationary satellite lies on a line of sight with a large set of locations that covers about 40 percent of the earth's surface. Three such satellites, placed at 120° (⅓-circle) intervals around the earth, allow coverage of all human-developed regions. Only the extreme polar regions lie "out of range." We can aim a *dish antenna* at a geostationary satellite, and once we've fixed the antenna in the correct position, we can leave it alone.

Another form of satellite system uses multiple "birds" in relatively low-altitude orbits that take them over, or nearly over, the earth's poles. These satellites exhibit continuous, rapid motion with respect to the earth's surface. If enough satellites of this type exist, the entire "flock" can work together, maintaining reliable communications between any two points on the surface at all times. Directional antennas aren't necessary in these systems, which engineers call *low earth orbit* (LEO) networks.

Fiberoptics

We can modulate beams of IR or visible light, just as we can modulate RF carriers. An IR or visible light beam has a frequency far higher than that of any RF signal, allowing modulation by data at rates faster than anything possible with radio.

Fiberoptic technology offers several advantages over wire cables (which are sometimes called *copper* because the conductors usually comprise that metallic element). A fiberoptic cable doesn't cost much, doesn't weigh much, and remains immune to interference from outside EM fields. A fiberoptic cable doesn't corrode as metallic wires do. Fiberoptic cables are inexpensive to maintain and easy to repair. An optical fiber can carry far more signals than a cable because the frequency bands are far wider in terms of megahertz or gigahertz.

In theory we can "imprint" the entire RF spectrum, from VLF through EHF, onto a single beam of visible light and transmit it through an optical fiber no thicker than a strand of human hair!

Receiver Fundamentals

A wireless receiver converts EM waves into the original messages sent by a distant transmitter. Let's define a few important criteria for receiver operation, and then we'll look at two common receiver designs.

Specifications

The *specifications* of a receiver quantify how well the hardware can actually do what we design and build it to do.

Sensitivity: The most common way to express receiver sensitivity is to state the number of micro-volts that must exist at the antenna terminals to produce a certain *signal-to-noise ratio* (S/N) or *signal-plus-noise-to-noise ratio* (S+N/N) in decibels (dB). The sensitivity depends on the gain of the *front end* (the amplifier or amplifiers connected to the antenna). The amount of noise that the front end generates also matters because subsequent stages amplify its noise output as well as its signal output.

Selectivity: The *passband*, or bandwidth that the receiver can "hear," is established by a wideband *preselector* in the early RF amplification stages, and is honed to precision by narrowband filters in later amplifier stages. The preselector makes the receiver most sensitive within about plus-or-minus 10 percent (±10%) of the desired signal frequency. The narrowband filter responds only to the frequency or channel of a specific signal that we want to hear; the filter rejects signals in nearby channels.

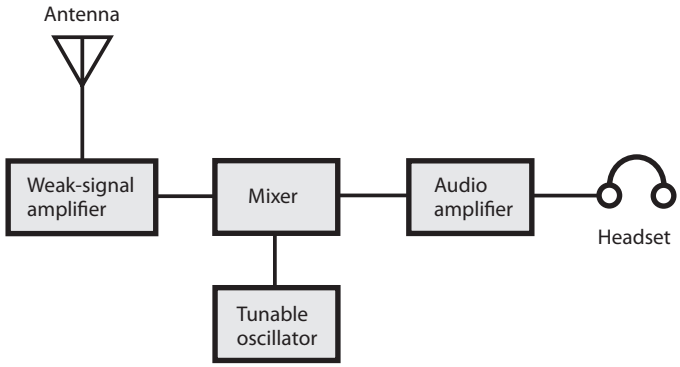
Dynamic range: The signals at a receiver input can vary over several orders of magnitude (powers of 10) in terms of absolute voltage. We define dynamic range as the ability of a receiver to maintain a fairly constant output, and yet to keep its rated sensitivity, in the presence of signals ranging from extremely weak to extremely strong. A good receiver exhibits dynamic range in excess of 100 dB. Engineers can conduct experiments to determine the dynamic range of any receiver; commercial receiver manufacturers publish this specification as a selling point.

Noise figure: The less internal noise a receiver produces, in general, the better the S/N ratio will be. We can expect an excellent S/N ratio in the presence of weak signals only when our receiver has a low noise figure, a measure of internally generated circuit noise. The noise figure matters most at VHF, UHF, and microwave frequencies. Gallium-arsenide field-effect transistors (GaAsFETs) are known for the low levels of noise they generate, even at very high frequencies. We can get away with other types of FETs at lower frequencies. Bipolar transistors, which carry higher currents than FETs, generate more circuit noise than FETs do.

Direct-Conversion Receiver

A *direct-conversion receiver* derives its output by mixing incoming signals with the output of a tunable (variable frequency) *local oscillator* (LO). The received signal goes into a mixer along with the output of the LO. Figure 27-10 is a block diagram of a direct-conversion receiver.

For reception of on/off keyed Morse code, also called *radiotelegraphy* or *continuous-wave* (CW) mode, the LO, also called a *beat-frequency oscillator* (BFO), is set a few hundred hertz above or below



27-10 Block diagram of a direct-conversion receiver.

the signal frequency. We can also use this scheme to receive FSK signals. The audio output has a frequency equal to the difference between the LO frequency and the incoming carrier frequency. For reception of AM or SSB signals, we adjust the LO to precisely the same frequency as that of the signal carrier, a condition called *zero beat* because the *beat frequency*, or difference frequency, between the LO and the signal carrier equals zero.

A direct-conversion receiver provides rather poor *selectivity*, meaning that it can't always separate incoming signals when they lie close together in frequency. In a direct-conversion receiver, we can hear signals on either side of the LO frequency at the same time. A *selective filter* can theoretically eliminate this problem. Such a filter must be designed for a fixed frequency if we expect it to work well. However, in a direct-conversion receiver, the RF amplifier must operate over a wide range of frequencies, making effective filter design an extreme challenge.

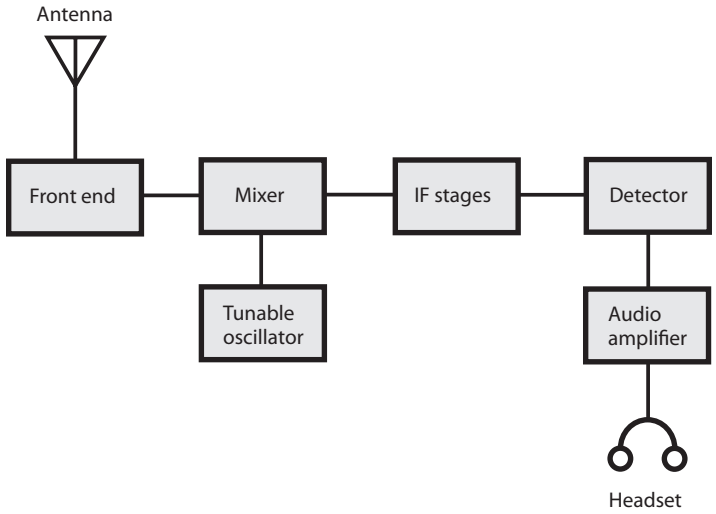
Superheterodyne Receiver

A *superheterodyne receiver*, also called a *superhet*, uses one or more LOs and mixers to obtain a constant-frequency signal. We can more easily filter a fixed-frequency signal than we can filter a signal that changes in frequency (as it does in a direct-conversion receiver).

In a superhet, the incoming signal goes from the antenna through a tunable, sensitive front end, which is a precision weak-signal amplifier. The output of the front end mixes (heterodynes) with the signal from a tunable, unmodulated LO. We can choose the sum signal or the difference signal for subsequent amplification. We call this signal the *first intermediate frequency* (IF), which can be filtered to obtain selectivity.

If the first IF signal passes straight into the detector, we call our system a *single-conversion receiver*. Some receivers use a second mixer and second LO, converting the first IF to a lower-frequency *second IF*. Then we have a *double-conversion receiver*. The IF bandpass filter can be constructed for use on a fixed frequency, allowing superior selectivity and facilitating adjustable bandwidth. The sensitivity is enhanced because fixed IF amplifiers are easy to keep in tune.

Unfortunately, even the best superheterodyne receiver can intercept or generate unwanted signals. We call external false signals *images*; we call internally generated false signals *birdies*. If we carefully choose the LO frequency (or frequencies) when we design our system, images and birdies will rarely cause problems during ordinary operation.



27-11 Block diagram of a single-conversion superheterodyne receiver.

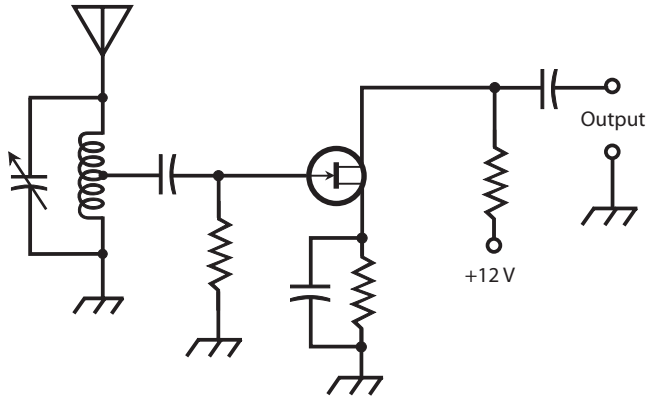
Stages of a Single-Conversion Superhet

Figure 27-11 shows a block diagram of a generic single-conversion superheterodyne receiver. Individual receiver designs vary somewhat, but we can consider this example representative. The various stages break down as follows:

- The front end consists of the first RF amplifier and often includes *LC* bandpass filters between the amplifier and the antenna. The dynamic range and sensitivity of a receiver are determined by the performance of the front end.
- The mixer stage, in conjunction with the tunable LO, converts the variable signal frequency to a constant IF. The output occurs at either the sum or the difference of the signal frequency and the tunable LO frequency.
- The IF stages produce most of the gain. We also get most of the selectivity here, filtering out unwanted signals and noise, while allowing the desired signal to pass.
- The detector extracts the information from the signal. Common circuits include the envelope detector for AM, the product detector for SSB, FSK, and CW, and the ratio detector for FM.
- One or two stages of audio amplification boost the demodulated signal to a level suitable for a speaker or headset. Alternatively, we can feed the signal to a printer, facsimile machine, or computer.

Predetector Stages

When we design and build a superheterodyne receiver, we must ensure that the stages preceding the first mixer provide reasonable gain but generate minimal internal noise. They must also be capable of handling strong signals without *desensitization* (losing gain), a phenomenon also known as *overloading*.



27-12 A tunable preamplifier for use with a radio receiver. This circuit uses an N-channel JFET.

Preamplifier

All preamplifiers operate in the class-A mode, and most employ FETs. An FET has a high input impedance ideally suited to weak-signal work. Figure 27-12 shows a simple generic RF preamplifier circuit. Input tuning reduces noise and provides some selectivity. This circuit produces a 5 dB to 10 dB gain, depending on the frequency and the choice of FET.

We must ensure that a preamplifier remains linear in the presence of strong input signals. Nonlinearity can cause unwanted mixing among multiple incoming signals. These so-called *mixing products* produce *intermodulation distortion* (IMD), or *intermod*, that can spawn numerous false signals inside the receiver. Intermod can also degrade the S/N ratio by generating *hash*, a form of wideband noise.

Front End

At low and medium frequencies, considerable atmospheric noise exists, and the design of a front-end circuit is simple because we don't have to worry much about internally generated noise. (Conditions are bad enough in the antenna!)

Atmospheric noise diminishes as we get above 30 MHz or so. Then the main sensitivity-limiting factor becomes noise generated within the receiver. For this reason, front-end design grows in importance as the frequency rises through the VHF, UHF, and microwave spectra.

The front end, like a preamplifier, must remain as linear as possible. The greater the degree of nonlinearity, the more susceptible the circuit becomes to the generation of mixing products and intermod. The front end should also have the greatest possible dynamic range.

Preselector

The preselector provides a bandpass response that improves the S/N ratio, and reduces the likelihood of overloading by a strong signal that's far removed from the operating frequency. The preselector also provides *image rejection* in a superheterodyne circuit.

We can tune a preselector by means of *tracking* with the receiver's main tuning control, but this technique requires careful design and alignment. Some older receivers incorporate preselectors that must be adjusted independently of the receiver tuning.

IF Chains

A high IF (several megahertz) works better than a low IF (less than 1 MHz) for image rejection. However, a low IF allows us to obtain superior selectivity. Double-conversion receivers have a comparatively high first IF and a low second IF to get the "best of both worlds." We can cascade multiple IF amplifiers with tuned-transformer coupling. The amplifiers follow the mixer and precede the detector. Double-conversion receivers have two series, called *chains*, of IF amplifiers. The *first IF chain* follows the first mixer and precedes the second mixer, and the *second IF chain* follows the second mixer and precedes the detector.

Engineers sometimes express IF-chain selectivity by comparing the bandwidths for two power-attenuation values, usually −3 dB and −30 dB, also called *3 dB down* and *30 dB down*. This specification offers a good description of the bandpass response. We call the ratio of the bandwidth at −30 dB to the bandwidth at −3 dB the *shape factor*. In general, small shape factors are more desirable than large ones, but small factors can prove difficult to attain in practice. When we have a small shape factor and graph the system gain as a function of frequency, we get a curve that resembles a rectangle, so we can say that the receiver has a *rectangular response*.

Detectors

Detection, also called *demodulation*, allows a wireless receiver to recover the modulating information, such as audio, images, or printed data, from an incoming signal.

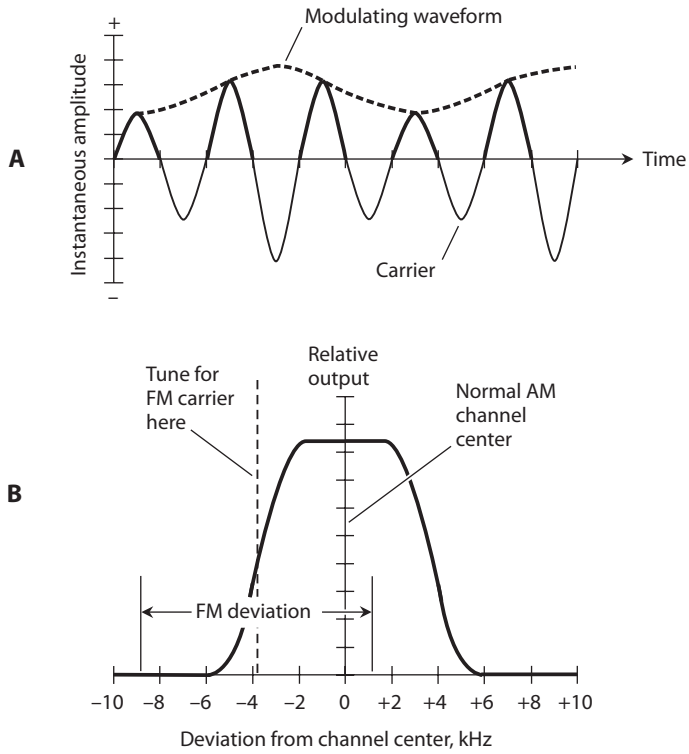
Detection of AM

A radio receiver can extract the information from an AM signal by half-wave rectifying the carrier wave and then filtering the output waveform just enough to smooth out the RF pulsations. Figure 27-13A shows a simplified time-domain view of how this process works. The rapid pulsations (solid curves) occur at the RF carrier frequency; the slower fluctuation (dashed curve) portrays the modulating data. The carrier pulsations are smoothed out by passing the output through a capacitor that's large enough to hold the charge for one carrier current cycle, but not so large that it dampens or obliterates the fluctuations in the modulating signal. We call this technique *envelope detection*.

Detection of CW and FSK

If we want a receiver to detect CW, we must inject a constant-frequency, unmodulated carrier a few hundred hertz above or below the signal frequency. The local carrier is produced by a tunable BFO. The BFO signal and the incoming CW signal heterodyne in a mixer to produce audio output at the sum or difference frequency. We can tune the BFO to obtain an audio "note" or "tone" at a comfortable listening pitch, usually 500 to 1000 Hz. This process is called *heterodyne detection*.

We can detect FSK signals using the same method as CW detection. The carrier beats against the BFO in the mixer, producing an audio tone that alternates between two different pitches. With FSK, the BFO frequency is set a few hundred hertz above or below both the *mark frequency* and the *space frequency*. The *frequency offset*, or difference between the BFO and the signal frequencies, determines the audio output frequencies. We adjust the frequency offset to get specific standard AF notes (such as 2125 Hz and 2295 Hz in the case of 170-Hz shift).



27-13 At A, envelope detection of AM, shown in the time domain. At B, slope detection of FM, shown in the frequency domain.

Slope Detection of FM and PM

We can use an AM receiver to detect FM or PM by setting the receiver frequency near, but not exactly at, the unmodulated-carrier frequency. An AM receiver has a filter with a passband of a few kilohertz and a selectivity curve such as that shown in Fig. 27-13B. If we tune the receiver so that the FM unmodulated-carrier frequency lies near either edge, or *skirt*, of the filter response, frequency variations in the incoming signal cause its carrier to "swing" in and out of the receiver passband. As a result, the instantaneous receiver output amplitude varies along with the modulating data on the FM or PM signal. In this system, known as *slope detection*, the relationship between the instantaneous deviation and the instantaneous output amplitude is nonlinear because the skirt of the passband is not a straight line (as we can see in Fig. 27-13B). Therefore, slope detection does not provide an optimum method of detecting FM or PM signals. The process can usually yield an intelligible voice, but it will ruin the quality of music.

Using a PLL to Detect FM or PM

If we inject an FM or PM signal into a PLL circuit, the loop produces an error voltage that constitutes a precise duplicate of the modulating waveform. A *limiter*, which keeps the signal amplitude from varying, can be placed ahead of the PLL so that the receiver doesn't respond to changes in the signal amplitude. Weak signals tend to abruptly appear and disappear, rather than fading, in an FM or PM receiver that employs limiting.

Discriminator for FM or PM

A *discriminator* produces an output voltage that depends on the instantaneous signal frequency. When the signal frequency lies at the center of the receiver passband, the output voltage equals zero. When the instantaneous signal frequency falls below the passband center, the output voltage becomes positive. When the instantaneous signal frequency rises above center, the output voltage becomes negative. The relationship between the instantaneous FM deviation (which, as we remember, can result indirectly from PM) and the instantaneous output amplitude is linear. Therefore, the detector output represents a faithful reproduction of the incoming signal data. A discriminator is sensitive to amplitude variations, but we can use a limiter to get rid of this problem, just as we do in a PLL detector.

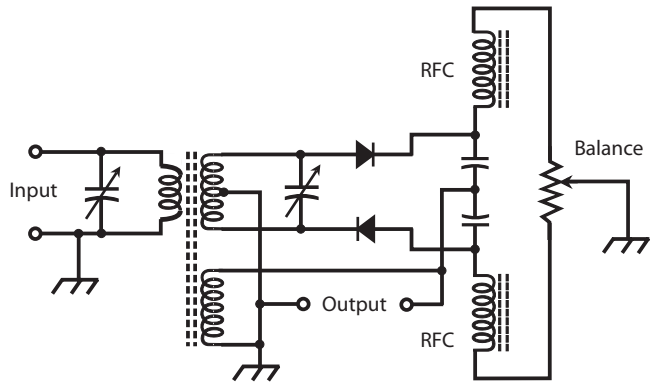
Ratio Detector for FM or PM

A *ratio detector* comprises a discriminator with a built-in limiter. The original design was developed by RCA (Radio Corporation of America), and works well in high-fidelity receivers and in the audio portions of old-fashioned analog TV receivers. Figure 27-13C illustrates a simple ratio detector circuit. The potentiometer marked "balance" should be adjusted experimentally to get optimum received-signal audio quality.

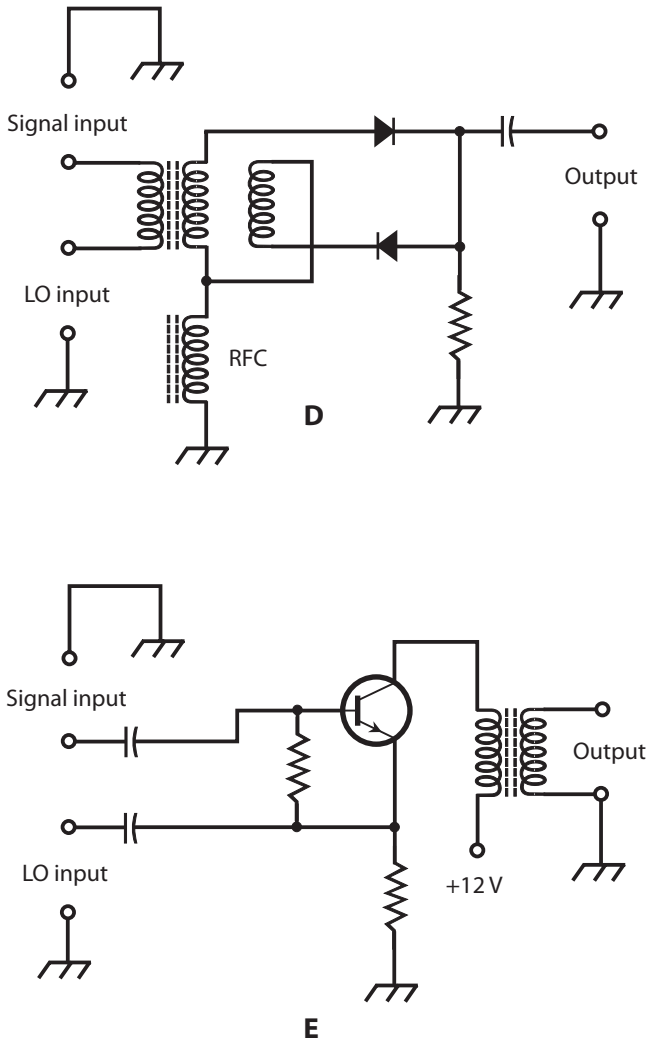
Detection of SSB

For reception of SSB signals, most communications engineers prefer to use a *product detector*, although a direct-conversion receiver can do the job. A product detector also facilitates reception of CW and FSK. The incoming signal combines with the output of an unmodulated LO, reproducing the original modulating signal data. Product detection occurs at a single frequency, rather than at a variable frequency, as in direct-conversion reception. The single, constant frequency results from mixing of the incoming signal with the output of the LO.

Figures 27-13D and 27-13E are schematic diagrams of product-detector circuits, which can also serve as mixers in superhet receivers. In the circuit shown at D, diodes are used, so we do not get any amplification. The circuit shown at E employs a bipolar transistor biased for class-B mode, providing some gain if the incoming signal has been sufficiently amplified by the front end before it arrives at the detector input. The effectiveness of the circuits shown in Figs. 27-13D or 27-13E



27-13 At C, a ratio detector circuit for demodulating FM signals.



27-13 At D, a product detector using diodes. At E, a product detector using an NPN bipolar transistor biased for class-B operation.

lies in the nonlinearity of the semiconductor devices. This nonlinearity facilitates the heterodyning necessary to obtain sum and difference frequency signals that result in data output.

Postdetector Stages

We can obtain selectivity in a receiver by tailoring the frequency response in the AF amplifier stages following the detector, in addition to optimizing the RF selectivity in the IF stages preceding the detector.

Filtering

In a communications system, a human voice signal requires a band ranging from about 300 Hz to 3000 Hz for a listener to easily understand the content. An *audio bandpass filter*, with a passband

of 300 Hz to 3000 Hz, can improve the intelligibility in some voice receivers. An ideal voice audio bandpass filter has little or no attenuation within the passband range but high attenuation outside the passband range, along with a near-rectangular response curve.

A CW or FSK signal requires only a few hundred hertz of bandwidth. Audio CW filters can narrow the response bandwidth to 100 Hz or less, but passbands narrower than about 100 Hz produce *ringing*, degrading the quality of reception at high data speeds. With FSK, the bandwidth of the filter must be at least as large as the difference (shift) between mark and space, but it need not (and shouldn't) greatly exceed the frequency shift.

An *audio notch filter* is a *band-rejection filter* with a sharp, narrow response. Band-rejection filters pass signals only below a certain lower cutoff frequency or above a certain upper cutoff frequency. Between those limits, in the so-called *bandstop range*, signals are blocked. A notch filter can "mute" an interfering unmodulated carrier or CW signal that produces a constant-frequency tone in the receiver output. Audio notch filters are tunable from at least 300 Hz to 3000 Hz. Some AF notch filters work automatically; when an interfering AF tone appears, the notch finds and "mutes" it within a few tenths of a second.

Squelching

A *squelch* silences a receiver when no incoming signals exist, allowing reception of signals when they appear. Most FM communications receivers use squelching systems. The squelch is normally *closed*, cutting off all audio output (especially receiver hiss, which annoys some communications operators) when no signal is present. The squelch *opens*, allowing everything to be heard, if the signal amplitude exceeds a *squelch threshold* that the operator can adjust.

In some systems, the squelch does not open unless an incoming signal has certain pre-determined characteristics. This feature is called *selective squelching*. The most common way to achieve selective squelching is the use of a *subaudible* (below 300 Hz) *tone generator* or an AF *tone-burst generator* in the transmitter. The squelch opens only in the presence of signals modulated by a tone, or sequence of tones, having the proper characteristics. Some radio operators use selective squelching to prevent unwanted transmissions from "coming in."

Specialized Wireless Modes

Communications engineers have a long history of innovation, developing numerous exotic wireless modes. In recent years, new modes have emerged; we can expect more to come, each of which offers specific advantages under strange or difficult conditions. Three common examples follow.

Synchronized Communications

Digital signals require less bandwidth than analog signals to convey a given amount of information per unit of time. The term *synchronized communications* refers to any of several specialized digital modes in which the transmitter and receiver operate from a common frequency-and-time standard to optimize the amount of data that can be sent in a communications channel or band.

In synchronized digital communications, also called *coherent communications*, the receiver and transmitter operate in lock-step. The receiver evaluates each transmitted data bit for a block of time lasting for the specified duration of a single bit. This process makes it possible to use a receiving filter having extremely narrow bandwidth. The synchronization requires the use of an external

frequency-and-time standard, such as that provided by the National Institute of Standards and Technology (NIST) radio station WWV in the United States. Frequency dividers generate the necessary synchronizing signals from the frequency-standard signal. A tone or pulse appears in the receiver output for a particular bit if, but only if, the average signal voltage exceeds a certain value over the duration of that bit. False signals caused by filter ringing, sferics, or ignition noise are generally ignored because they rarely produce sufficient average bit voltage.

Experiments with synchronized communications have shown that the improvement in S/N ratio, compared with nonsynchronized systems, is several decibels at low to moderate data speeds.

Multiplexing

Signals in a communications channel or band can be intertwined, or *multiplexed*, in various ways. The most common methods are *frequency-division multiplexing* (FDM) and *time-division multiplexing* (TDM). In FDM, the channel is broken down into subchannels. The carrier frequencies of the signals are spaced so that they don't overlap. Each signal remains independent of all the others. A TDM system breaks signals down into segments of specific time duration, and then the segments are transferred in a rotating sequence. The receiver stays synchronized with the transmitter by means of an external time standard, such as the data from "shortwave" station WWV. Multiplexing requires an *encoder* that combines or "intertwines" the signals in the transmitter, and a *decoder* that separates or "untangles" the signals in the receiver.

Spread-Spectrum

In *spread-spectrum communications*, the transmitter varies the main carrier frequency in a controlled manner, independently of the signal modulation. The receiver is programmed to follow the transmitter frequency from instant to instant. The whole signal, therefore, "roams" up and down in frequency within a defined range.

In spread-spectrum mode, the probability of *catastrophic interference*, in which one strong interfering signal can obliterate the desired signal, is near zero. Unauthorized people find it impossible to eavesdrop on a spread-spectrum communications link unless they gain access to the *sequencing code*, also known as the *frequency-spreading function*. Such a function can be complex, and, of course, it must be kept secret. If neither the transmitting operator nor the receiving operator divulge the sequencing code to anyone, then (ideally) no unauthorized listener will be able to intercept it.

During a spread-spectrum contact between a given transmitter and receiver, the operating frequency can fluctuate over a range of several kilohertz, megahertz, or tens of megahertz. As a band becomes occupied with an increasing number of spread-spectrum signals, the overall noise level in the band appears to increase. Therefore, a practical limit exists to the number of spread-spectrum contacts that a band can handle. This limit is roughly the same as it would be if all the signals were constant in frequency, and had their own discrete channels. The main difference between fixed-frequency communications and spread-spectrum communications, when the band gets crowded, lies in the *nature* of the mutual interference.

A common method of generating spread-spectrum signals involves so-called *frequency hopping*. The transmitter has a list of channels that it follows in a certain order. The transmitter "jumps" or "hops" from one frequency to another in the list. The receiver must be programmed with this same list, in the same order, and must be synchronized with the transmitter. The *dwell time* equals the length of time that the signal remains on any given frequency; it's the same

as the time interval at which frequency changes occur. In a well-designed frequency-hopping system, the dwell time is short enough so that a signal will not be noticed by an unauthorized listener using a receiver set to a constant frequency, and also will not cause interference on any frequency. The sequence contains numerous *dwell frequencies*, so the signal energy is diluted to the extent that, if someone tunes to any particular frequency in the sequence, they won't notice the signal.

Another way to obtain spread spectrum, called *frequency sweeping*, requires frequency-modulating the main transmitted carrier with a waveform that guides it "smoothly" up and down over the assigned band. The "sweeping FM" remains entirely independent of the actual data that the signal conveys. A receiver can intercept the signal if, but only if, its instantaneous frequency varies according to the same waveform, over the same band, at the same rate, and in the same phase as that of the transmitter. The transmitter and receiver in effect "roam all over the band," following each other from moment to moment according to a "secret map" that only they know.

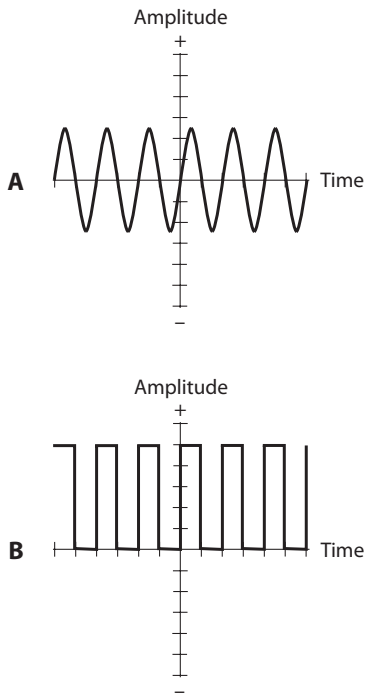
Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

28
CHAPTER

Digital Basics

ENGINEERS AND TECHNICIANS CALL AN ELECTRONIC SIGNAL *DIGITAL* WHEN IT CAN ATTAIN A LIMITED number of well-defined states. Digital signals contrast with *analog* signals, which vary over a continuous range and, therefore, can, in theory, attain infinitely many different instantaneous states. Figure 28-1 shows an example of an analog signal (at A) and a digital signal (at B).



28-1 An analog wave (A) and a digital rendition of the same wave (B).

Numeration Systems

In everyday life, most of us deal with the *decimal number system*, which makes use of digits from the set {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}. Machines, such as computers and communications devices, offer other numeration schemes.

Decimal

The familiar decimal number system is also called *base 10* or *radix 10*. When we express nonnegative integers in this system, we multiply the rightmost digit by 10^0, or 1. We multiply the next digit to the left by 10^1, or 10. The power of 10 increases as we move to the left. Once we've multiplied the digits, we add up all the resulting values. For example:

$$8 \times 10^0 + 5 \times 10^1 + 0 \times 10^2 + 2 \times 10^3 + 6 \times 10^4 + 8 \times 10^5 = 862{,}058$$

Binary

The *binary number system* denotes numbers using only the digits 0 and 1. We'll sometimes hear this system called *base 2* or *radix 2*. When we express nonnegative integers in binary notation, we multiply the rightmost digit by 2^0, or 1. The next digit to the left is multiplied by 2^1, or 2. The power of 2 increases as we continue to the left, so we get a "fours" digit, then an "eights" digit, then a "16s" digit, and so on. For example, consider the decimal number 94. In the binary system, we would write this quantity as 1011110. It breaks down into the sum

$$0 \times 2^0 + 1 \times 2^1 + 1 \times 2^2 + 1 \times 2^3 + 1 \times 2^4 + 0 \times 2^5 + 1 \times 2^6 = 94$$

Hexadecimal

Another system used in computer work is the *hexadecimal number system*. It has 16 (2^4) symbols: the usual 0 through 9 plus six more, represented by the uppercase English letters A through F, yielding the digit set {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F}. This system is sometimes called *base 16* or *radix 16*. All of the hexadecimal digits 0 through 9 represent the same values as their decimal counterparts. However, we have the following additional digits:

- Hexadecimal A equals decimal 10
- Hexadecimal B equals decimal 11
- Hexadecimal C equals decimal 12
- Hexadecimal D equals decimal 13
- Hexadecimal E equals decimal 14
- Hexadecimal F equals decimal 15

When we express nonnegative integers in hexadecimal notation, we multiply the rightmost digit by 16^0, or 1. We multiply the next digit to the left by 16^1, or 16. The power of 16 increases as we continue to the left, so we get a "256s" digit, then a "4096s" digit, and so on. For example, we write the decimal quantity 35,898 in hexadecimal form as 8C3A. Remembering that C = 12 and A = 10, we can break the hexadecimal number down into the following sum:

$$\mathrm{A} \times 16^0 + 3 \times 16^1 + \mathrm{C} \times 16^2 + 8 \times 16^3 = 35{,}898$$

Digital Logic

Digital logic, also called simply *logic*, is the form of "reasoning" used by electronic machines. Engineers also use the term in reference to the circuits that make up digital devices and systems.

Boolean Algebra

Boolean algebra constitutes a system of logic using the numbers 0 and 1 with the operations AND (multiplication), OR (addition), and NOT (negation). Combinations of these operations give us two more, called NAND (NOT AND) and NOR (NOT OR). This system, which gets its name from the nineteenth-century British mathematician *George Boole*, plays a vital role in the design of digital electronic circuits.

- The AND operation, also called *logical conjunction*, operates on two or more quantities. Let's denote it using an asterisk, for example X * Y.
- The NOT operation, also called *logical inversion* or *logical negation*, operates on a single quantity. Let's denote it using a minus sign (−), for example −X.
- The OR operation, also called *logical disjunction*, operates on two or more quantities. Let's denote it using a plus sign (+), for example X + Y.

Table 28-1A breaks down all the possible input and output values for the above-described Boolean operations, where 0 indicates "falsity" and 1 indicates "truth." In mathematics and philosophy courses involving logic, you can expect to see other symbols used for conjunction and disjunction.

Theorems

Table 28-1B shows several logic equations that hold true under all circumstances, that is, for all values of the *logical variables* X, Y, and Z. We call such facts *theorems*. Statements on either side of the equals (=) sign in each case are *logically equivalent*, meaning that one is true *if and only if (iff)* the other is true. For example, the statement X = Y means "If X then Y, and if Y then X." Boolean theorems allow us to simplify complicated *logic functions*, facilitating the construction of a circuit to perform a specific digital operation using the smallest possible number of switches.

Positive versus Negative Logic

In so-called *positive logic*, a circuit represents the binary digit 1 with an *electrical potential* of approximately +5 V DC (called the *high state* or simply *high*), while the binary digit 0 appears

Table 28-1A. Boolean Operations

X	Y	−X	X * Y	X + Y
0	0	1	0	0
0	1	1	0	1
1	0	0	0	1
1	1	0	1	1

Table 28-1B. Common Theorems in Boolean Algebra

Theorem (Logic Equation)	What It's Called
X + 0 = X	OR identity
X * 1 = X	AND identity
X + 1 = 1	
X * 0 = 0	
X + X = X	
X * X = X	
−(−X) = X	Double negation
X + (−X) = 1	
X * (−X) = 0	Contradiction
X + Y = Y + X	Commutative property of OR
X * Y = Y * X	Commutative property of AND
X + (X * Y) = X	
X * (−Y) + Y = X + Y	
(X + Y) + Z = X + (Y + Z)	Associative property of OR
(X * Y) * Z = X * (Y * Z)	Associative property of AND
X * (Y + Z) = (X * Y) + (X * Z)	Distributive property
−(X + Y) = (−X) * (−Y)	DeMorgan's Theorem
−(X * Y) = (−X) + (−Y)	DeMorgan's Theorem

as little or no DC voltage (called the *low state* or simply *low*). Some circuits employ *negative logic*, in which little or no DC voltage (low) represents logic 1, while +5 V DC (high) represents logic 0. In another form of negative logic, the digit 1 appears as a negative voltage (such as −5 V DC, constituting the low state) and the digit 0 appears as little or no DC voltage (the high state because it has the more positive voltage). To avoid confusion, let's stay with positive logic for the rest of this chapter!

Logic Gates

All digital electronic devices employ switches that perform specific logical operations. These switches, called *logic gates*, can have anywhere from one to several inputs and (usually) a single output.

- A *logical inverter*, also called a *NOT gate*, has one input and one output. It reverses, or inverts, the state of the input. If the input equals 1, then the output equals 0. If the input equals 0, then the output equals 1.
- An *OR gate* can have two or more inputs (although it usually has only two). If both, or all, of the inputs equal 0, then the output equals 0. If any of the inputs equal 1, then the output equals 1. Mathematical logicians would tell us that such a gate performs an *inclusive-OR operation* because it "includes" the case where both variables are high.
- An *AND gate* can have two or more inputs (although it usually has only two). If both, or all, of the inputs equal 1, then the output equals 1. If any of the inputs equal 0, then the output equals 0.

- An OR gate can be followed by a NOT gate. This combination gives us a *NOT-OR* gate, more often called a *NOR gate*. If both, or all, of the inputs equal 0, then the output equals 1. If any of the inputs equal 1, then the output equals 0.
- An AND gate can be followed by a NOT gate. This combination gives us a *NOT-AND* gate, more often called a *NAND gate*. If both, or all, of the inputs equal 1, then the output equals 0. If any of the inputs equal 0, then the output equals 1.
- An *exclusive OR gate*, also called an *XOR gate*, has two inputs and one output. If the two inputs have the same state (either both 1 or both 0), then the output equals 0. If the two inputs have different states, then the output equals 1. Mathematicians use the term *exclusive-OR operation* because it doesn't "include" the case where both variables are high.

Table 28-2 summarizes the functions of the above-defined logic gates, assuming a single input for the NOT gate and two inputs for the others. Figure 28-2 illustrates the schematic symbols that engineers and technicians use to represent these gates in circuit diagrams.

Clocks

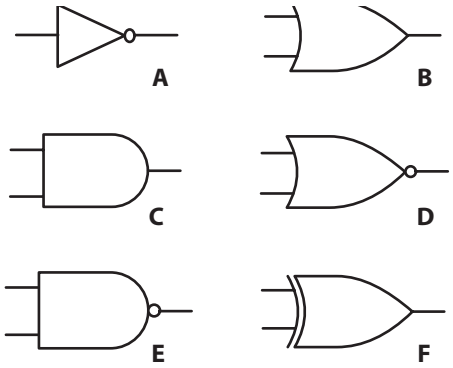
In electronics, the term *clock* refers to a circuit that generates pulses at high speed and at precise, constant time intervals. The clock sets the tempo for the operation of digital devices. In a computer, the clock acts like a metronome for the *microprocessor*. We express or measure clock speeds as frequencies in *hertz* (Hz). One hertz equals one pulse per second. Higher-frequency units work out as follows, just as they do with analog wave signals:

- A *kilohertz* (kHz) equals 1000 or 10^3 pulses per second
- A *megahertz* (MHz) equals 1,000,000 or 10^6 pulses per second
- A *gigahertz* (GHz) equals 1,000,000,000 or 10^9 pulses per second
- A *terahertz* (THz) equals 1,000,000,000,000 or 10^{12} pulses per second

In positive logic, a clock generates brief high pulses at regular intervals. The normal state is low.

Table 28-2. Logic Gates and Their Characteristics

Gate Type	Number of Inputs	Remarks
NOT	1	Changes state of input.
OR	2 or more	Output high if any inputs are high. Output low if all inputs are low.
AND	2 or more	Output low if any inputs are low. Output high if all inputs are high.
NOR	2 or more	Output low if any inputs are high. Output high if all inputs are low.
NAND	2 or more	Output high if any inputs are low. Output low if all inputs are high.
XOR	2	Output high if inputs differ. Output low if inputs are the same.



28-2 An inverter or NOT gate (A), an OR gate (B), an AND gate (C), a NOR gate (D), a NAND gate (E), and an XOR gate (F).

Flip-Flops

A *flip-flop* is a specialized circuit constructed from logic gates, known collectively as a *sequential gate*. In a sequential gate, the output state depends on both the inputs and the outputs. The term "sequential" comes from the fact that the output depends not only on the states of the circuit at any given instant in time, but also on the states immediately preceding. A flip-flop has two states, called *set* and *reset*. Usually, the set state corresponds to logic 1 (high), and the reset state corresponds to logic 0 (low). In schematic diagrams, a flip-flop is usually shown as a rectangle with two or more inputs and two outputs. If the rectangle symbol is used, the letters FF, for "flip-flop," are printed or written at the top of the rectangle, either inside or outside. Several different types of flip-flop exist, as follows.

- In an *R-S flip-flop*, the inputs are labeled R (reset) and S (set). Engineers call the outputs Q and −Q. (Often, rather than −Q, we'll see Q′, or perhaps Q with a line over it.) Table 28-3A shows the input and output states. If R = 0 and S = 0, the output states remain at the values they've attained for the moment. If R = 0 and S = 1, then Q = 1 and −Q = 0. If R = 1 and S = 0, then Q = 0 and −Q = 1. When S = 1 and R = 1, the circuit becomes unpredictable. You may remember that back in Chap. 26, we used a R-S flip-flop contained in the 555 timer IC.
- In a *synchronous flip-flop*, the states change when triggered by the signal from an external clock. In *static triggering,* the outputs change state only when the clock signal is either high or low. This type of circuit is sometimes called a *gated flip-flop*. In *positive-edge triggering*, the outputs change state at the instant the clock pulse is positive-going. In *negative-edge triggering*, the outputs change state at the instant the clock pulse is negative-going. The abrupt rise or fall of a pulse looks like the edge of a cliff (Fig. 28-3).
- In a *master/slave (M/S) flip-flop*, the inputs are stored before the outputs can change state. This device comprises essentially two R-S flip-flops in series. We call the first flip-flop the *master* and the second flip-flop the *slave*. The master functions when the clock output is high, and the slave acts during the next ensuing low portion of the clock output. The time delay prevents confusion between the input and output.
- The operation of a *J-K flip-flop* resembles the functioning of an R-S flip-flop, except that the J-K device has a predictable output when the inputs both equal 1. Table 28-3B shows the input and output states for this type of flip-flop. The output changes only when a triggering pulse is received.

Table 28-3. Flip-Flop States

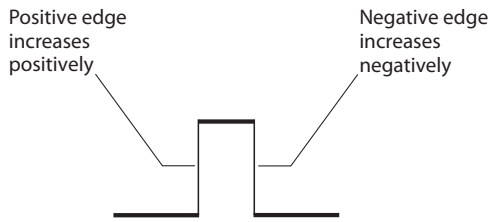
A: R-S Flip-Flop			
R	**S**	**Q**	**−Q**
0	0	Q	−Q
0	1	1	0
1	0	0	1
1	1	?	?
B: J-K Flip-Flop			
J	**K**	**Q**	**−Q**
0	0	Q	−Q
0	1	1	0
1	0	0	1
1	1	−Q	Q

- The operation of an *R-S-T flip-flop* resembles that of an R-S flip-flop, except that a high pulse at an additional T input causes the circuit to change state.
- A circuit called a *T flip-flop* has only one input. Each time a high pulse appears at the input, the output reverses its state, either from 1 to 0 or else from 0 to 1. Note the difference between this type of circuit and a simple inverter (NOT gate)!

Counters

A *counter* literally counts digital pulses one by one. Each time the counter receives a high pulse, the binary number in its *memory* increases by 1. A *frequency counter* can measure the frequency of an AC wave or signal by tallying up the cycles over a precisely known period of time. The circuit consists of a *gate*, which begins and ends each counting cycle at defined intervals. (Don't confuse this type of gate with the logic gates described a few moments ago.) The counter's accuracy depends on the *gate time*, or how long the gate remains open to accept pulses for counting. As we increase the gate time in a frequency counter, the accuracy improves. Although the counter tallies up the pulses as binary numbers, the display shows base-10 digital numerals.

28-3 Digital pulse edges are either positive-going (abruptly increasing in the positive sense) or negative-going (abruptly increasing in the negative sense).



Binary Communications

If we want to attain *multilevel signaling* (digital transmission with more than two states), we can represent each different signal level with a unique group of binary digits, representing a specific binary number. A group of three binary digits can represent up to 2^3, or eight, levels. A group of four binary digits can represent up to 2^4, or 16, levels. The term "binary digit" is commonly contracted to *bit*. A bit is represented by either logic 0 or logic 1. Some engineers call a group of eight bits an *octet*, and in many systems an octet corresponds to a unit called a *byte*.

Forms of Binary Signaling

Engineers have invented numerous forms, or *modes*, of binary communication. Three classical examples follow.

1. *Morse code* is the oldest binary mode. The logic states are called *mark* (key-closed or on) and *space* (key-open or off). Morse code is obsolete in modern systems, but amateur radio operators consider it a backup mode that can work in case digital signal-reading machines fail.
2. *Baudot*, also called the *Murray code*, is a five-unit digital code not widely used by today's digital equipment, except in a few antiquated *teleprinter* systems. There exist 2^5, or 32, possible representations.
3. The *American National Standard Code for Information Interchange* (ASCII) is a seven-unit code for the transmission of text and simple computer programs. There exist 2^7, or 128, possible representations.

Bits

We can represent large quantities of data according to powers of 2 or according to powers of 10. This duality can cause some confusion. Here's how the units build up when we talk about bits:

- A *kilobit* (kb) equals 10^3 or 1000 bits
- A *megabit* (Mb) equals 10^6 or 1,000,000 bits
- A *gigabit* (Gb) equals 10^9 or 1,000,000,000 bits
- A *terabit* (Tb) equals 10^{12} bits or 1000 Gb
- A *petabit* (Pb) equals 10^{15} bits or 1000 Tb
- An *exabit* (Eb) equals 10^{18} bits or 1000 Pb

We use power-of-10 multiples of *bits per second* (bps, kbps, Mbps, Gbps, and so on) to express data speed when we transmit digital signals from one location (called the *source*) to another location (called the *destination*).

Bytes

Data quantity in *storage* or *memory* (residing in a fixed location rather than propagating from one place to another) is specified in units that comprise power-of-2 multiples of bytes. Here's how the units build up:

- A *kilobyte* (KB) equals 2^{10} or 1024 bytes
- A *megabyte* (MB) equals 2^{20} or 1,048,576 bytes
- A *gigabyte* (GB) equals 2^{30} or 1,073,741,824 bytes

- A *terabyte* (TB) equals 2^{40} bytes or 1024 GB
- A *petabyte* (PB) equals 2^{50} bytes or 1024 TB
- An *exabyte* (EB) equals 2^{60} bytes or 1024 PB

Note also the following conventions concerning abbreviations for the units and prefix multipliers:

- The lowercase b stands for "bits."
- The uppercase B stands for "bytes."
- The lowercase k stands for 10^3 or 1000.
- The uppercase K stands for 2^{10} or 1024.
- We always denote the prefix multipliers M, G, T, P, and E in uppercase.

Baud

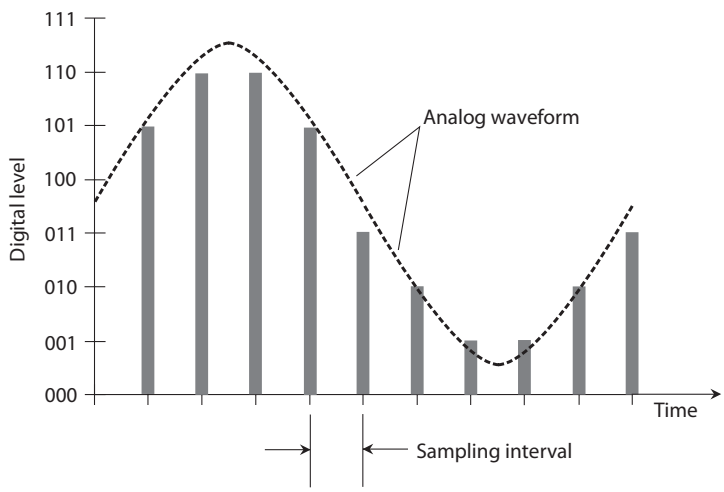
The term *baud* refers to the number of times per second that a signal changes state. We'll read about baud (sometimes called *baud rate*) only in texts and papers dated before about 1980. Bits per second (bps) and baud represent *qualitatively* different parameters, even though they might come out *quantitatively* close to each other for a particular digital signal. Some engineers used to speak and write about bps and baud as if they meant the same thing. They don't!

Data Conversion

We can convert an analog signal into a string of pulses whose amplitudes have a finite number of states, usually some power of 2. This scheme constitutes *analog-to-digital* (A/D) *conversion*; the reverse of digital-to-analog (D/A) conversion.

Figure 28-4 shows the functional difference between analog and digital signals. Imagine sampling the curve to obtain a sequence, or *train,* of pulses (A/D conversion), or smoothing out the pulses to obtain the curve (D/A conversion).

We can transmit and receive binary data one bit at a time along a single line or channel. This mode constitutes *serial data transmission*. Higher data speeds can be obtained by using multiple

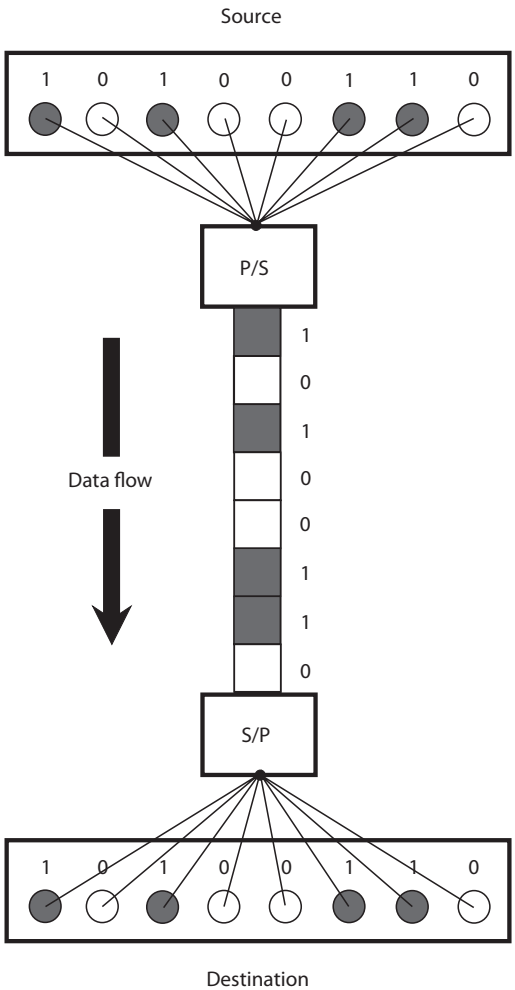


28-4 An analog waveform (dashed curve) and an 8-level digital representation of the same curve (vertical bars).

lines or a wideband channel, sending independent sequences of bits along each line or subchannel. Then we have *parallel data transmission*.

Parallel-to-serial (P/S) *conversion* involves the reception of bits from multiple lines or channels, and their retransmission one by one along a single line or channel. A *buffer* stores the bits from the parallel lines or channels while they await transmission along the serial line or channel. *Serial-to-parallel* (S/P) *conversion* involves the reception of bits one by one from a serial line or channel, and their retransmission in batches along several lines or channels. The output of an S/P converter cannot go any faster than the input, but we can find such a system useful when we want to interface between a serial-data device and a parallel-data device.

Figure 28-5 illustrates a circuit that employs a P/S converter at the source and an S/P converter at the destination. In this example, the words comprise eight-bit bytes. However, the words could have 16, 32, 64, or even 128 bits, depending on the communications scheme.



28-5 A communications circuit employing parallel-to-serial (P/S) conversion at the source and serial-to-parallel (S/P) conversion at the destination.

Data Compression

Data compression provides us with a way to maximize the amount of digital information that a machine can store in a given space, or that we can send within a certain period of time.

Digital image and audio files can be compressed in either of two ways. In *lossless image compression*, detail is not sacrificed; only the redundant bits are eliminated. In *lossy image compression*, we lose some detail, although the loss is rarely severe enough to degrade the quality of the image or sound to an objectionable extent.

Digital Signal Processing

Digital signal processing is a way of using digital technology to manipulate analog signals, typically audio or video. This can be done for a variety of purposes, examples are:

- filtering
- noise reduction
- noise cancelling headphones
- captioning of video

Signals are first changed into digital form by A/D conversion. Then the digital data is manipulated using software on the microcontroller. Finally, the digital signal is changed back to the original voice or video by a D/A converter.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

This page intentionally left blank

4
PART

Specialized Devices and Systems

This page intentionally left blank

29
CHAPTER

Microcontrollers

A *MICROCONTROLLER* IS A SINGLE IC THAT CONTAINS WHAT WOULD HAVE PASSED FOR MOST OF A home computer in the 1980s. It includes a simple 8-bit (sometimes more) microprocessor, nonvolatile flash memory that holds a program to be run, and *random access memory* (RAM) to hold temporary data and values. Many microcontrollers also have some *electrically erasable programmable read-only memory* (EEPROM) that is used for nonvolatile storage of program data. That is, unlike normal RAM, the contents are not lost when the microcontroller is not powered. A microcontroller also has numerous *general-purpose input/output* (GPIO) pins that you can use to interface with sensors, switches, LEDs, displays, etc. It is essentially a computer on a chip.

Benefits

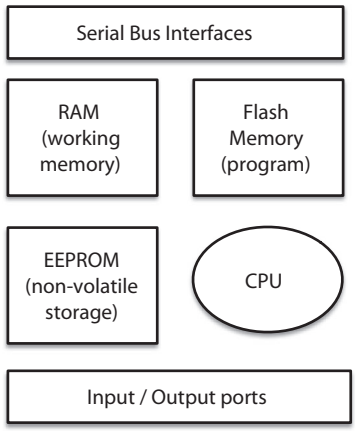
The benefit of having a single microcontroller chip instead of many separate logic chips is cost savings. You will find microcontrollers in almost any item of consumer electronics that you care to mention. You will find them in everything from an electric toothbrush to a car (which probably has tens of microcontrollers in it).

Figure 29-1 shows how a typical microcontroller chip is organized. The *central processing unit* (CPU) fetches instructions one at a time from the flash memory. These instructions together form the program running on the microcontroller and include things such as:

- Add numbers together
- Compare numbers
- Jump to another part of the program depending on a comparison
- Read a digital input
- Write a digital output

Microcontrollers perform things one step at a time and use a *clock* to trigger each step. The clock is an oscillator of generally 1 MHz to 20 MHz for a low-power microcontroller and hundreds of MHz for higher power processors. Each time the clock ticks over, another instruction is performed.

Microcontrollers interact with the world through their GPIO pins. These can be configured to be inputs or outputs, just like the inputs and outputs to logic gates described in Chap. 28



29-1 Essential components of a microcontroller.

"Digital Basics." However, because what links the pins is software rather than, say, a hardware AND gate, the microcontroller is almost infinitely flexible, limited only by its speed, memory, and the ability of you, the programmer, to write the program that controls it.

Microcontrollers generally have a method of converting between continuous analog signals and the on/off world of digital electronics. This means that their use is not restricted to just the digital, but they can also be pressed into service for some analog electronic applications.

A special subspecies of microcontroller called the *digital signal processing* (DSP) microcontroller is optimized for such analog activities. An echo effects box for a guitar or even tone controls for audio amplifiers are often implemented by digitizing an analog signal, doing some math, and then converting the digital value back into an analog signal again.

The downside of all this flexibility is that the microcontroller is not made knowing how to perform all this magic; it must be programmed, and so the person wanting to use it has to learn a little computer programming as well as electronics. The programs for a microcontroller are generally short and simple, and you will learn a lot more about this in the next chapter in which we look at the extremely popular microcontroller board, the Arduino.

Microcontroller boards, such as the Arduino, combine a microcontroller chip with supporting components like voltage regulators and a USB programming interface so that you can get programs onto your microcontroller without having to have separate programming hardware.

All Shapes and Sizes

Microcontrollers are available in all sorts of package sizes, from three pins to hundreds of pins, so you can pick a device with the features, performance, price, and number of GPIO pins that you need for your design. In fact, there are so many different devices that it is not easy to navigate the vast array of devices on offer. To simplify things a little, devices are often grouped into families of microcontrollers that have the same basic structure and programming instructions, but have different numbers of pins and different quantities of flash memory for program storage.

The ATTiny family of microcontrollers from the manufacturer Microchip are typical of low-end microcontrollers. For higher power microcontrollers, many manufacturers base their design on those of the microcontroller design company ARM. These devices access data and perform operations on it 32 bits at a time rather than the 8 bits of the ATTiny series. Table 29-1 shows the features

Table 29-1. A Selection of Microcontrollers

Microcontroller	Architecture	Max. Clock	Package Pins	GPIO Pins	Flash Memory (kB)	Guide Price ($)
ATtiny4	8-bit	20 MHz	8	6	4	0.60
ATtiny3227	8-bit	20 MHz	24	22	21	1.20
STM32F030	ARM 32-bit	48 MHz	20	18	16	1.75
STM32F303	ARM 32-bit	72 MHz	48	37	64	7.50

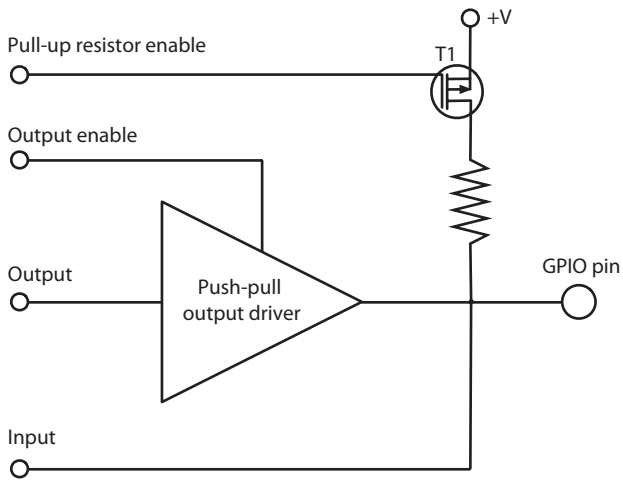
of these common microcontrollers, including some popular ARM devices from the manufacturer ST Electronics. The guide prices are somewhat arbitrary but are based on the price for one chip at the time of writing.

General-Purpose Input/Output (GPIO) Pins

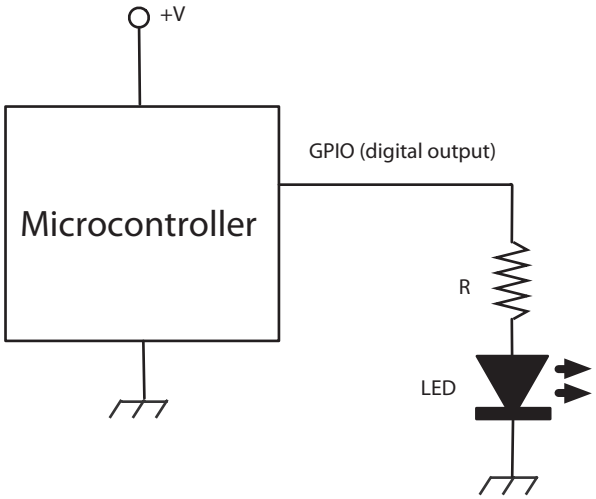
The clever part of a GPIO pin is the "general purpose" part. This means that your program running on the microcontroller can decide whether the pin should be an input or an output and also, if it is an input, whether or not to enable a built-in resistor to pull the input up to the positive supply of the microcontroller.

Figure 29-2 is a simplified schematic diagram for a typical GPIO pin. The output drive is *push-pull*. This means that when configured as an output, the pin can either *sink* (receive) or source current, although generally only between 10 mA and 40 mA in each direction. In addition, control logic allows the whole output to be disabled so that the pin can act as a digital input.

Microcontrollers often have some pins that can be used as analog inputs. This requires the pin to be connected to a comparator; by comparing the voltage at the input to a series of different voltages generated within the microcontroller, the pin can be used to measure a voltage.



29-2 Simplified schematic of a GPIO pin.



29-3 Lighting an LED with a digital output.

Digital Outputs

The most common thing to do with a GPIO pin is to have it act as a digital output. That is, to turn something on and off. This could be something connected directly to the GPIO pin, like an LED, or it could use a transistor or relay to provide more current or voltage, or both, to whatever is being controlled.

Figure 29-3 shows how you would typically connect the GPIO pin of a microcontroller to an LED. The series resistor is necessary because GPIO pins can typically supply only a few tens of milliamps.

Setting the digital output HIGH will set the pin at the supply voltage for the microcontroller (5 V or 3.3 V) and setting it LOW will set it to 0 V.

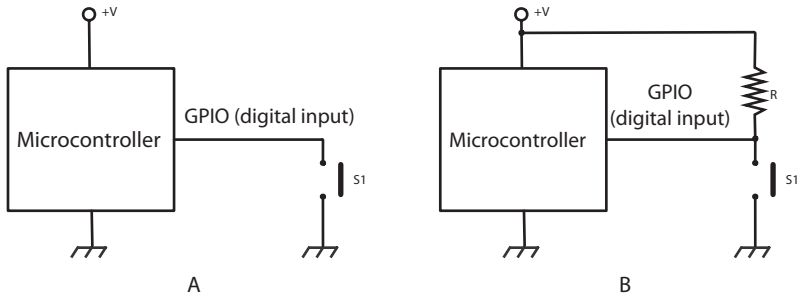
Most microcontrollers can operate at 5-V or 3.3-V logic levels (5-V micros will nearly always work with 3.3-V ones, but the reverse is not always true). In this case, if you assume that the microcontroller is operating at 5 V, the resistor has a value of 470 Ω and the LED has a forward voltage of 2 V, then $I = E/R = (5 - 2)/470 = 6.28$ mA. That's enough for the LED to light reasonably brightly.

Digital Inputs

When you configure a GPIO pin to be a digital input, it is because you want the microcontroller to do something in response to that input. So the digital input might be connected to a push switch or a motion detector sensor that has a digital output.

When a microcontroller "reads" a digital input, the input will either be HIGH or LOW. HIGH usually means over half of the microcontroller's supply voltage (2.5 V for a 5-V microcontroller and 1.65 V for a 3.3-V microcontroller) and LOW is below that threshold.

It's fine to connect a 3.3-V digital output to a 5-V input, as the 3.3-V HIGH voltage will still be above the 2.5-V input threshold, but you must not connect a 5-V digital output to a 3.3-V digital input unless the input is described as "5-V tolerant." It is not uncommon for a 3.3-V microcontroller



29-4 Connecting a switch to a digital input.

to have some 5-V-tolerant inputs just to make it easier to interface them in systems with multiple supply voltages.

Because microcontrollers can implement all sorts of complex logic, there is usually no need to connect a switch to a microcontroller that is any more complex than a simple momentary action, push-to-make switch. Such switches are arranged to switch to ground, using a pull-up resistor that keeps the input high until it is connected to ground by the switch. Figure 29-4A shows a switch connected to a digital input using the internal pull-up resistor of the microcontroller, and Fig. 29-4B shows a similar switch using an external pull-up resistor.

To avoid the need for an external pull-up resistor, the pin logic for a microcontroller normally includes a built-in pull-up resistor that can be enabled or disabled from the program running on your microcontroller. The value of this resistor is often specified as a range rather than an exact value, and this might typically be 30 k to 50 k. Under most circumstances this will work just fine, but if you have a long lead to your switch, then you may need to supply your own external and much smaller value of resistor, say, 270 Ω. In this case, the resistor should be connected as shown in Fig. 29-4B.

Switch contacts tend to bounce. That is, when you press the button, you do not get a single clean closure of the contacts, but rather the contacts close and open a number of times in quick succession. For some applications, this does not matter, but if, say, alternate presses of the switch button are used by the microcontroller to turn something else on and off, then if the digital input receives an even number of bounces in rapid succession, it could seem like nothing happened when the button was pressed.

Debouncing switches for use in digital circuitry without a microprocessor requires hardware debouncing; however, if the switch is connected to a microcontroller, then it can be debounced using software without the need for any extra components. You will see an example of how to do this with an Arduino microcontroller board in Chap. 30.

An ordinary digital input normally requires the program running on the microcontroller to repeatedly read the value of the digital input until it changes. This means that some very short pulses on a digital input might be missed because they went high then low again before the program on the microcontroller could register the change. This won't matter for pressing a button, but for such very short-pulsed inputs (microseconds), then some of the microcontroller's GPIO pins can be designated as "interrupts." These are guaranteed to be registered by the microcontroller and cause your program to stop whatever it was doing and instead run an interrupt service routine.

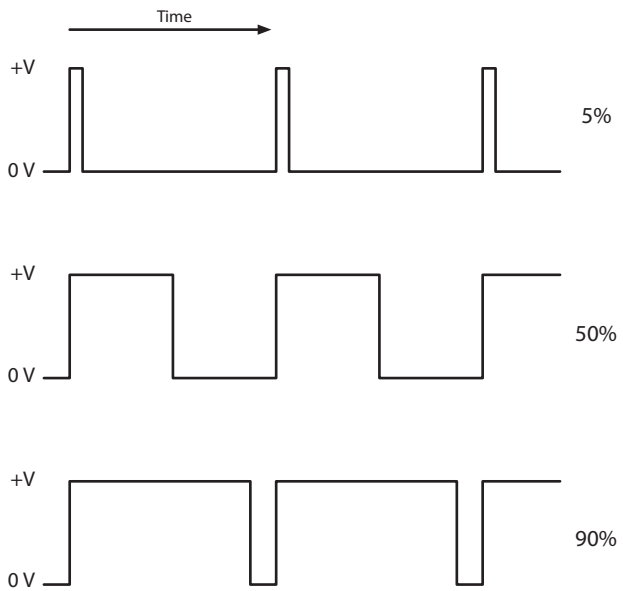
PWM Outputs

Some microcontrollers provide true analog outputs. These allow the output to be varied in a large number of steps (often 1024 or 4096) between 0 and the microcontroller supply voltage. These microcontrollers are the exception though, and in general, microcontrollers use a technique called PWM (see the section "The Class D Amplifier" in Chap. 25 and the section "Pulse-Width Modulation" in Chap. 27) to provide something that approximates an analog output.

Rather than using PWM to transmit a signal, as described in Chap. 27, or to amplify an audio signal, as described in Chap. 25, a microcontroller uses low-frequency PWM to control the power to, say, a motor or LED, controlling its speed or brightness by altering the duration of the pulses powering the output device.

Figure 29-5 shows how PWM works. If the pulses are short (say, high for just 5% of the time), then only a small amount of energy is delivered with each pulse. The longer the pulse, the more energy is supplied to the load. When powering a motor, this will control the speed at which the motor rotates. When driving an LED using PWM, the brightness appears to change. In fact, an LED can turn on and off millions of times per second, and so the PWM pulses will become pulses of light. The human eye and brain do the averaging trick for us, making the brightness of the LED vary with the pulse length.

Most microcontrollers have hardware PWM available on some, or all, of their pins. If they do not, then PWM can be implemented in software quite easily. The pulse frequency of the PWM outputs is generally around 500 Hz but is also configurable to be higher or lower, depending on the application.



29-5 How PWM works.

Analog Inputs

Converting a voltage (between 0 and the microcontroller supply voltage) to a number for use by the program running on the microcontroller requires the use of an *A/D converter* (analog-to-digital converter), perhaps more often called an ADC these days.

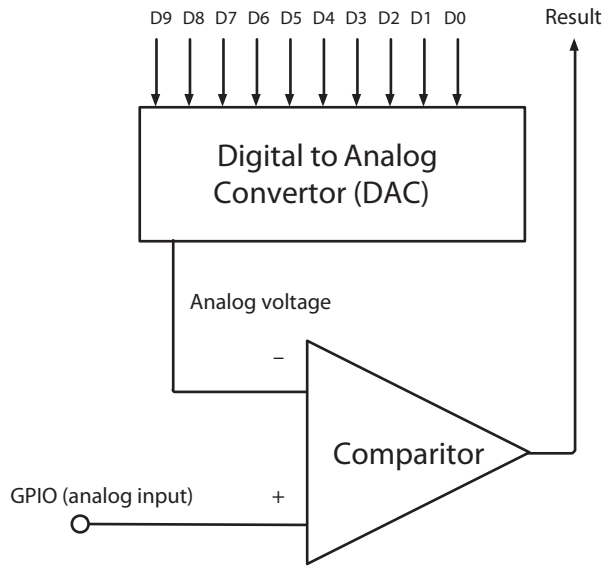
Nearly all microcontrollers use a technique called *successive approximation* to convert the analog voltage to a digital value. This technique uses a *digital-to-analog converter* (DAC), which converts a digital value to an analog voltage, and a comparator, as shown in Fig. 29-6.

The GPIO pin being used as an analog input is connected to one input of a voltage comparator. The other input of the comparator is connected to the output of a DAC. The DAC generates an analog voltage depending on the binary number on D0 to D9 supplied by the microcontroller. So, if this value is 0 (D0 to D9 all low) then the output voltage of the DAC will be 0. If, on the other hand, D0 to D9 are all high (the maximum value), then the output voltage from the DAC is the microcontroller supply voltage.

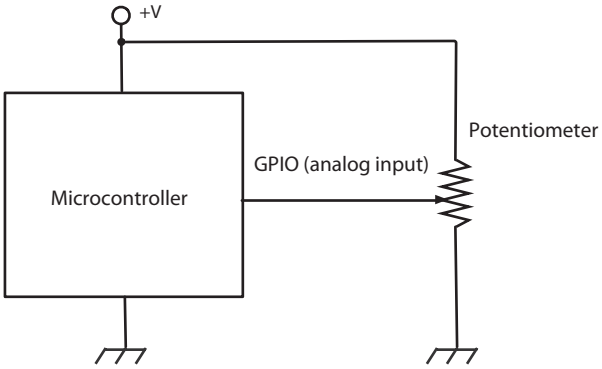
It's the comparator's job to provide feedback to the microcontroller in a game of "higher/lower" to gradually home in on a number input to the DAC that just balances the input from the GPIO pin. This takes a number of steps. In fact, the number of steps is equal to the number of bits input to the DAC (10 in Fig. 29-6).

The trick is to start with the highest value bit (D9) in Fig. 29-6. Assuming a 5-V microcontroller, then if bit D9 is high and all the other bits are low, then the output voltage of the DAC will be 2.5 V (half of 5 V). If the GPIO input voltage is above this, then the comparator output will be high. If it is less, then the output of the comparator will be low. So, if the output of the comparator (Result in Fig. 29-6) is high, the microcontroller leaves bit D9 set and moves on to the next bit D8, and applies the same test using the comparator. This is repeated for the other bits of the DAC.

Microcontrollers have different bit resolutions of ADC. The example here is 10 bits, which is typical of many microcontrollers and is what is used in the ATmega328 microcontroller used by the



29-6 Successive approximation ADC.



29-7 Attaching a potentiometer to the analog input of a microcontroller.

Arduino. Some microcontrollers have 8- or 12-bit resolutions. The more bits of resolution, the more precise the reading, but the longer it takes to make.

An 8-bit resolution ADC gives a value between 0 and 255 ($2^8 - 1$). So, a voltage of 3 V at the GPIO pin of a 5-V microcontroller would result in a reading of $3/5 \times 255 = 153$. For a microcontroller with 10-bit ADC, the range of value would be 0 to 1023 ($2^{10} - 1$).

Figure 29-7 shows how you can attach a potentiometer to the analog input of a microcontroller so that the position of the potentiometer's knob can be read by the program running on the microcontroller.

The resistance of the potentiometer needs to be much lower than the input impedance of the comparator. Most microcontroller comparators will have an input impedance in the megohm range, so a potentiometer of perhaps 10 k should be fine.

Dedicated Serial Hardware

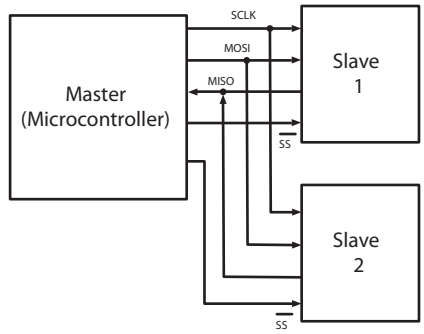
Most microcontrollers have one or more serial data interfaces. At least one such interface is needed to program the microcontroller, and generally this is the *Serial Peripheral Interface* (SPI). This can be used both for programming the microcontroller and also as an interface to other peripheral chips or microcontrollers, once the microcontroller has been programmed.

Although you can simulate (sometimes called "bit-banging") all of the serial communication mechanisms described below using software, microcontrollers will provide hardware implementations of some of the interface types that make programming the microcontroller easier and higher speed communication possible.

Serial Peripheral Interface (SPI)

SPI uses four data lines for communication. Figure 29-8 shows how a number of peripherals can be connected to the data "bus." Note that there can be only one master device that is responsible for controlling the other devices.

Each of the slave devices has a dedicated *slave select* (SS) line, and the master enables just the device it wants to communicate with. Two data lines are needed because separate lines are used for each direction of communication. *Master Out Slave In* (MOSI) carries the data from the master to



29-8 SPI allows two or more peripherals to be connected to the data "bus."

the slave device, and *Master In Slave Out* (MISO) the reverse. A separate clock signal synchronizes the data transmission.

Note: The terms "Master" and "Slave" have negative connotations, and you will find that sometimes the word "Master" is replaced by "Controller" and "Slave" by "peripheral." This also applies to the acronyms, so MOSI becomes COPI and MISO becomes CIPO.

I²C

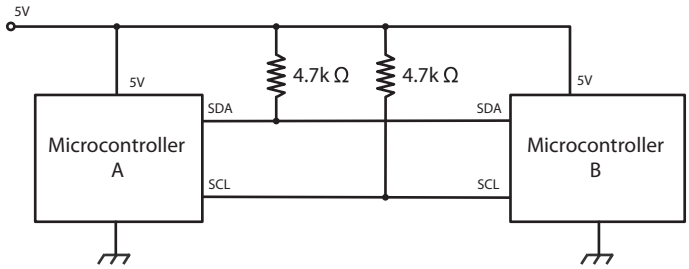
Inter-Integrated Circuit (I^2C), also sometimes known as *two-wire interface* (TWI), serves much the same purpose as SPI, although it has two wires for data rather than the four of SPI. I^2C is often used by displays and other peripheral modules that are designed to be connected to a microcontroller.

The two data lines of I^2C are open-drain connections that operate as both inputs and outputs at the microcontroller. They need pull-up resistors so that when not being driven low as an output, they are high. The protocol that controls the direction of data flow also ensures that the situation never occurs in which one end of the bus is connected to a digital output that is low and the other end to a digital output that is high.

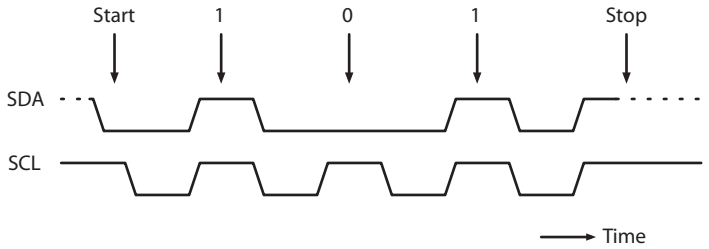
Figure 29-9 shows how two microcontrollers might communicate with each other using I^2C.

I^2C devices are either masters or slaves, and there can be more than one master device per bus. In fact, devices are allowed to change roles, although this is not usually done.

The *Serial Clock Line* (SCL) data line is a clock and the other, the *Serial Data Line* (SDA), carries the data. The timing of these pins is shown in Fig. 29-10.



29-9 Microcontroller-to-microcontroller communication using I^2C.



29-10 Timing diagram for I^2C.

The master supplies the SCL clock. When there is data to be transmitted, the sender (master or slave) takes the SDA line out of tri-state and sends data as logic highs or lows in time with the clock signal. When transmission is complete, the clock can stop, and the SDA pin is taken back to tri-state.

Serial

Many microcontrollers include hardware for yet another type of interface called *Serial*. This is an old standard with its roots dating back to the days of teletypes. Some computers can still be found that have Serial ports. In the "old days," people attached modems to them for communicating over phone lines with other computers.

The normal voltages used in the signals for Serial ports conform to the RS232 standard and use voltages that swing both positive and negative with respect to GND. This is not terribly convenient when using microcontrollers. For this reason, microcontrollers often use the same communication protocol, but at logic levels. This is called *TTL Serial*, or often just *Serial*. Let's use a capital S to distinguish this type of communication from general serial communication.

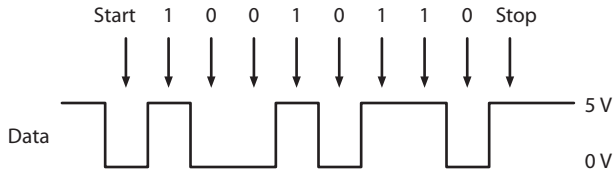
Electrically, TTL Serial uses two data pins, Transmit (Tx) and Receive (Rx). It is not a bus, and the connection is point to point, so there are no problems of addressing different devices. If you want to connect more than one Serial device to your microcontroller, then the microcontroller will need more than one Serial port.

Another remnant from early computer history is the nomenclature around the bandwidth of serial connections. A serial connection has to be set to the same baud rate at both ends of the connection. The *baud rate* (see also Chap. 28) is the number of times that the signal can change state per second. This is taken to be the same as the bits transmitted per second of the data payload, but that does include start, stop, or possible parity bits, so the actual transmission of data in bits per second is a little slower than the baud rate. To simplify matching up the baud rates at each end of the connection, a set of standard baud rates is used: 110, 300, 600, 1200, 2400, 4800, 9600, 14,400, 19,200, 38,400, 57,600, 115,200, 128,000, and 256,000.

Of these, 1200 is probably the slowest baud rate commonly in use and many TTL Serial devices will not go as high as 115,200. Perhaps the most common rate is 9600 baud. Devices often default to this rate, but can be configurable to other rates.

As well as the baud rate, other parameters that define a Serial connection are the number of bits per word, the type of parity bit, and the number of start and stop bits. Almost universally, these are defined as 8, none, and 1 respectively, which often gets abbreviated to 8N1.

Bits are simply sent as high or low logic levels (Fig. 29-11). As there is no separate clock signal, timing is critical, so after the start bit, the receiver will sample at the appropriate rate until it has read the eight data bits and the one stop bit. The least significant bit of the data is sent first.



29-11 An example of TTL Serial.

When connecting two Serial devices together, you connect the Tx pin of one device to the Rx pin of the other and vice-versa. For some devices, such as GPS receivers, only one-way communication is needed, as the device just repeatedly transmits data. In this case, the link needs to be established in only one direction.

Most microcontrollers have built-in hardware for TTL Serial called a *Universal Asynchronous Receiver Transmitter* (UART).

USB

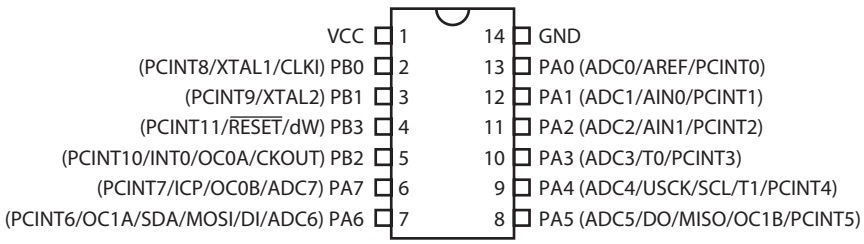
While electrically similar to RS232, *Universal Serial Bus* (USB) is one of the most complex interfaces around in terms of software and communication protocols. This complexity is because it can be used with full-sized computers with a vast array of peripherals, from keyboards and mice to printers and cameras. Fortunately, its use with microcontrollers is generally simplified.

While microcontrollers generally do not have USB hardware built in, most families of microcontrollers will have a member with USB hardware that allows the microcontroller to communicate with a PC or emulate a mouse or keyboard. In fact, even if the microcontroller does not have hardware USB support, a Serial interface and some slick software will normally allow the microcontroller to "talk" USB.

Microcontrollers are often used to implement USB peripherals, such as keyboards and mice. The USB connector has two power pins (GND and 5 V as well as two data pins D+ and D−). The USB standard dictates that a USB host (usually your computer) must supply 5 V and at least 500 mA on its supply terminals.

An Example—The ATtiny44

Figure 29-12 shows the pin connections of the ATtiny44 microcontroller taken from its datasheet. As you can see, each pin has a variety of roles. This is typical of a modern microcontroller in which flexibility is paramount.



29-12 The pinout of an ATtiny44.

Looking at Fig. 29-12, you can see that the only two pins that do not have multiple roles are pin 14 (GND) and pin 1 (VCC, the positive supply). The other pins are theoretically all available to be used as GPIO pins. The supply voltage can be anything between 2.7 V and 5.5 V.

The GPIO pins on the ATtiny44 are grouped into two ports, port A (PA) and port B (PB). Starting with the right side of the chip, pins 13 to 8 are labelled PA0 to PA5. These pins are all available for use as GPIO pins on port A. The other functions for those pins are indicated in parentheses. The pins PA0 to PA5 are also marked as ADC0 to ADC5. This indicates that all of those pins can be used as analog inputs.

Note that the SPI pins needed for programming (MOSI, MISO and USCK) can also be configured for use as GPIO pins.

On the left of the IC, you can see that pins 2 and 3 have the second function of XTAL1 and XTAL2. This allows you to connect a crystal between those pins to set the clock frequency of the microcontroller, which produces a much more accurate clock frequency than that provided by the internal resonator circuit of the microcontroller. This particular microcontroller will work with clock frequencies up to 20 MHz.

The choice between using the internal resonator circuit of the microcontroller for the clock frequency or sacrificing two GPIO pins for a more accurate clock frequency is made during the programming of the microcontroller chip.

Programming Languages

The machine code of a microcontroller is a series of instructions that tell it what to do. The instructions are not intended to be used directly by humans. Instead, the human programmer creates a text file containing a program written in a so-called *high-level language* on his/her main computer. This text file is then fed into a program called a *compiler*, which converts the high-level language into a machine code file that is installed into the flash memory of the microcontroller.

There are many programming languages that are used to program microcontrollers. The most common of these is the C programming language. C has the advantage that it is fairly easy to use, while still being relatively "close to the metal," so that programs written with it generally compile down to be compact and efficient. This is something that you need when programming a microcontroller with its limited memory.

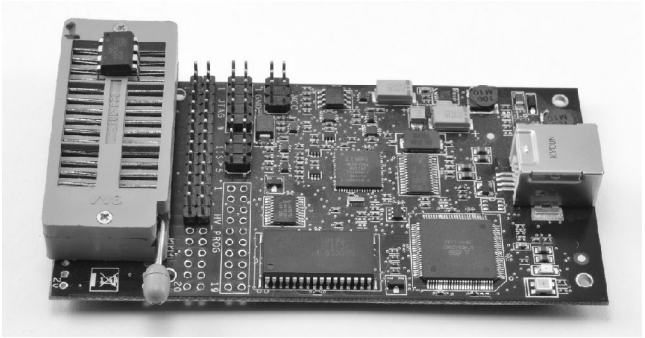
In the next chapter, you will learn a little C programming on a microcontroller.

Programming a Microcontroller

There are various ways of getting a program onto a microcontroller, and they all involve the use of a regular computer, and often specialized programming hardware, such as the AVR Dragon programmer shown in Fig. 29-13.

You can also connect the programming interface using header leads, if for example you need to program a surface mount microcontroller already soldered onto a board. This board has a *zero insertion force* (ZIF) socket on the left that holds the microcontroller IC while it is being programmed. On the right of Fig. 29-13, you can see the USB interface that is used to connect the programmer to your computer.

During programming, the microcontroller's SPI pins are used to allow access to the microcontroller's flash memory.



29-13 The AVR Dragon programmer.

Programming hardware, like the AVR Dragon, can also be used to program microcontrollers that are already soldered into place on a PCB, using a method called *In-Circuit Serial Programming* (ICSP). Here, the SPI pins of the microcontroller are connected to header pins on the PCB so that they can be connected to programming hardware using a "programming cable."

Another approach to programming microcontrollers is to program the microcontroller with a *bootloader,* using one of the methods described above. The bootloader runs each time that the microcontroller is restarted. It pauses for a fraction of a second to see if its Serial interface is being used to send it a new program. If it is, then it copies the received data into its flash memory and then restarts itself. Note that the bootloader is not destroyed in this process, so it can be used the next time a program needs to be uploaded.

Allowing the microcontroller to be programmed over Serial rather than SPI means that it can be programmed using hardware no more complex than a USB or Serial adaptor. The Arduino actually includes such a converter on board, so that all the hardware that you need to program an Arduino is a USB cable.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

30
CHAPTER

Arduino

ELECTRONICS HOBBYISTS HAVE BEEN USING MICROCONTROLLERS AS LONG AS THOSE DEVICES HAVE existed. But they have become extremely popular on the back of the runaway success of the Arduino microcontroller boards. Here are some of the reasons for the success of the Arduino platform:

- Built-in USB interface and programmer (no extra hardware required)
- Simple to use *Integrated Development Environment* (IDE) with which to write your programs
- Works with Microsoft Windows, Mac, and Linux computers
- Simple programming language
- Standard GPIO socket arrangement for the addition of plug-in "shields"
- Open-source hardware design, so you get to see the schematic diagram of the board and understand exactly how it works

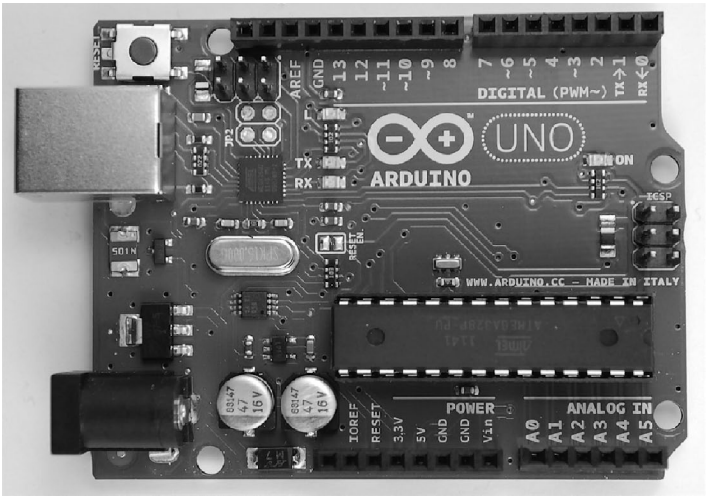
The Arduino Uno/Genuino

Although there is now a wide range of different Arduino models, the "classic" Arduino model is the Arduino Uno, more specifically the Arduino Uno revision 3. This is the device that most people think of as an Arduino. After legal difficulties, the originators of the Arduino have had to change the brand name to Genuino rather than Arduino. The boards are electrically identical, but cosmetically, look slightly different. Figure 30-1 shows an Arduino Uno with key components annotated.

The Arduino Uno's USB port is used when programming the Arduino, but it can also be used to communicate with the Arduino from your PC and to provide 5-V DC to the device.

The Arduino Uno actually has two microcontroller chips on the board. The main microcontroller is the ATmega328 chip fitted into an IC socket. The second microcontroller is just to the right of the USB socket. This second microcontroller has a built-in USB interface and serves the single purpose of acting as a USB interface to the ATmega328. The USB interface microcontroller has its own *In-Circuit Serial Programming* (ICSP) header that allows it to be programmed during manufacture.

In the top left of the board, there is a reset switch. Pressing this takes the RESET pin of the ATmega328 low, causing the microcontroller to reset.



30-1 The Arduino Uno.

The top side of the board has two sections of female header pins connected to the GPIO pins of the ATmega328 labeled 0 to 13. Pin 13 is actually wired to an LED built onto the board, known as the "L" LED. This is useful because it allows you to experiment with the Arduino GPIO pins as a digital output without having to attach any external components. The pins 0 and 1 double as a TTL Serial connection that is also used for communication and programming over USB, so it is best not to use pins 0 and 1 for anything else.

After pin 13, there is a ground pin (GND) and a pin labeled AREF. The AREF pin can be connected to a voltage lower than 5 V to compress the range of the analog inputs, although this feature is not often used. The Arduino Uno operates at 5 V, and so all its digital outputs will provide 5 V when HIGH.

The two unlabeled sockets to the left of AREF are the SDA and SCK pins of an I^2C interface.

The ICSP header pins on the right of the board are for factory programming of the ATmega328 with the bootloader that allows all subsequent programming of the ATmega328 to take place over USB rather than over ICSP, thus avoiding the need for special programming hardware.

The ATmega328 itself is mounted in an IC socket. This allows easy replacement of the IC should it be accidentally damaged by, say, shorting out an output pin. A socket also means that you can use the Arduino to program an ATmega328 and then remove the programmed IC and install it in an electronics project without having to commit the entire Arduino to the project. You can buy ATmega328 ICs already programmed with the Arduino bootloader to replace the chip.

At the bottom right of the board, you have six pins labeled A0 to A5 that are primarily intended for use as analog inputs, although they can all also be used as digital inputs or outputs.

The final section of the socket header is mostly concerned with power. The V_{in} pin can be used as an alternative to the DC Power Socket or the USB connector to power the Arduino from an unregulated 7-V to 12-V input. There are also regulated 5-V and 3.3-V power supplies from the voltage regulators built onto the board, or if being powered over USB, 5-V DC from the USB socket.

To allow for projects that require accurate timing, the Arduino uses an external 16-MHz crystal to provide the frequency for the clock to the ATmega328.

Setting Up the Arduino IDE

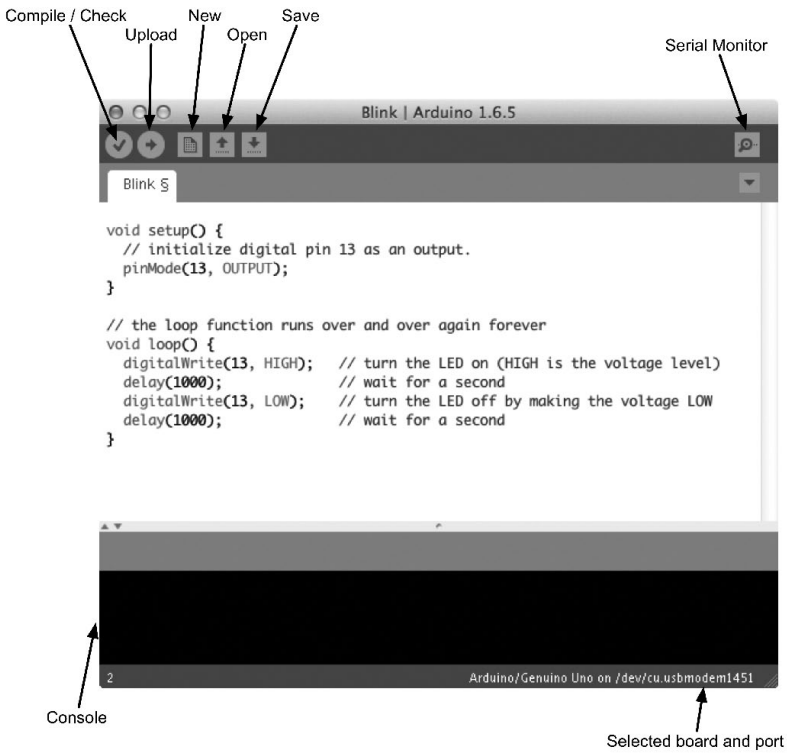
The Arduino integrated development environment (IDE) is not limited in use to just official Arduino microcontroller boards. In fact, there are probably more people using non-Arduino boards with the Arduino IDE than there are using official Arduino boards. It has become a common way of programming all manner of microcontrollers and has the advantage that you don't have to learn how to use each chip manufacturer's proprietary development environment.

The Arduino IDE is a program that you run on your Windows, Mac, or Linux PC that lets you type in your program and then upload it to your Arduino board over USB. To install the latest version of the Arduino IDE, follow the instructions on the following website:

https://www.arduino.cc/en/Main/Software

When you have the software installed, run the Arduino IDE; then from the "File" menu, select the option "Examples," "01 Basic," and then "Blink," and you should see a window that looks something like Fig. 30-2.

The "Compile/Check" button compiles the Arduino C code without actually attempting to upload it to the board itself. The "Upload" button first compiles the Arduino C code and then uploads it to the Arduino over USB.



30-2 The Arduino IDE.

In the Arduino world, programs are called "sketches." The next three buttons allow you to start a new sketch, open an existing sketch, or save the current sketch. Each sketch is saved as a file with the extension .iso. Each sketch file will automatically be placed inside a folder of the same name as the sketch file, but without the extension.

The top-right icon on the Arduino IDE's toolbar opens a separate window called the "Serial Monitor" that allows you to communicate with the Arduino from your computer.

The console area of the IDE is where any error messages will appear when you try to compile a sketch or upload it to an Arduino.

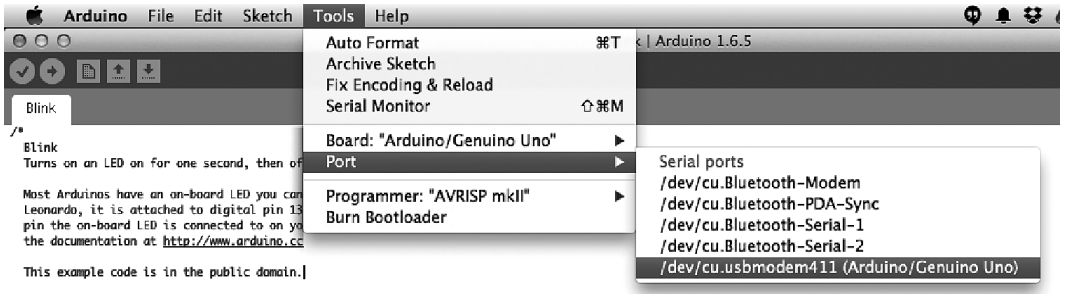
Programming "Blink"

The Arduino Uno has a tiny LED attached to pin 13, labeled "L." Your first step with an Arduino should be to make this LED blink. You may find that when you plug the LED lead into your Arduino, the "L" LED is already blinking. This happens because Arduinos are generally sold with the LED Blink sketch preinstalled. To prove that you are making the LED blink, you can make it blink faster.

When you start up the IDE and open the Blink sketch, the text of the sketch appears as shown below:

```
/*
  Blink
  Turns on an LED on for one second, then off for one second, repeatedly.

  Most Arduinos have an on-board LED you can control. On the Uno and
  Leonardo, it is attached to digital pin 13. If you're unsure which
  pin the on-board LED is connected to on your Arduino model, check
  the documentation at http://www.arduino.cc

  This example code is in the public domain.

  modified 8 May 2014
  by Scott Fitzgerald
 */

// the setup function runs once when you press reset or power the board
void setup() {
  // initialize digital pin 13 as an output.
  pinMode(13, OUTPUT);
}

// the loop function runs over and over again forever
void loop() {
  digitalWrite(13, HIGH);   // turn the LED on (HIGH is the voltage level)
  delay(200);               // wait for a second
  digitalWrite(13, LOW);    // turn the LED off by making the voltage LOW
  delay(200);               // wait for a second
}
```

The text between `/*` and `*/` is called "comment." It is not program code; it's just comments to explain the sketch. Similarly, many of the lines of actual code have a // after them followed by a description of what the code does. This is also "comment" rather than program code. When using the // style of comment, the comment begins with the // and ends at the end of the line.



30-3 Selecting the serial port.

Edit the two lines that say: `delay(1000)` to read `delay(200)` and then save the sketch. When you come to save the sketch, the IDE will ask you to choose a new name for it because it is a read-only, built-in example. Save the sketch in a different location.

If you haven't already done so, connect your Arduino to your PC with a USB cable. You then need to let the IDE know which USB port your Arduino is connected to, so from the Tools menu, select the "Port:" option, as shown in Fig. 30-3, and select the option that has (Arduino/Genuino Uno) after it. If you have a Mac or Linux computer, then the name will be something like the one shown in Fig. 30-3; if you are using Windows, then it will be called COM followed by a number.

You also need to tell the IDE what type of board is connected, so from the Tools menu, select "Board:" and then make sure that "Arduino/Genuino Uno" is selected.

Click on the upload button, and after the sketch has compiled, you should see the LEDs labeled TX and RX flicker while the sketch is uploaded into the flash memory of the ATmega328. This should take only a few seconds, after which the "L" LED should blink fairly fast.

Programming Fundamentals

Before taking apart the Blink sketch and adding some bells and whistles, let's explore the fundamental concepts of programming.

A program, or in Arduino terminology a "sketch," is a text file that contains a series of commands to be executed by the computer. In the case of the Arduino, these commands are in a programming language called C.

The commands will run (or "be executed") in order, one after the other. A microcontroller can do only one thing at a time. For example the lines below will be executed one after the other to make the LED blink.

```
digitalWrite(13, HIGH);   // turn the LED on (HIGH is the voltage level)
delay(200);               // wait for a second
digitalWrite(13, LOW);    // turn the LED off by making the voltage LOW
delay(200);
```

These commands are pretty self-explanatory, or if nothing else, explained by the comments that follow them. Note that the comment "wait for a second" is now wrong because we changed the delay. It should now say, "wait for 200 milliseconds (⅕ of a second)."

There is a little more to the structure of this sketch that we will come to in the next section. But for now, the important thing to understand is that these programming commands will be executed in turn, one after the other.

In fact, it is not the text of the above program that is copied into the flash memory of the ATmega328 microcontroller, but rather the compiled version of this code. In other words, the Arduino IDE "compiles" the text that we typed into a compact machine code form that is then copied into the flash memory of the ATmega328 by the bootloader program on the ATmega328.

Setup and Loop

If you strip away the comments from the Blink example, you are left with the following lines of actual program code.

```
void setup() {
  pinMode(13, OUTPUT);
}

void loop() {
  digitalWrite(13, HIGH);
  delay(200);
  digitalWrite(13, LOW);
  delay(200);

}
```

You can see that the code is divided into two blocks or "functions." The first block starts with the line `void setup {` and the end of `setup` is marked by a }. In this case, there is just a single line between { and } that specifies that pin 13 should be set to be an output. But in other situations, there may be multiple lines of code in `setup`. The `setup` function runs once when the Arduino resets.

The second function `loop` will be run over and over. As soon as it finishes the last line between its { and }, it will start over with the first line. Notice how each of the lines of code within `loop` ends in a semicolon.

The first line inside `loop` uses the Arduino command `digitalWrite` to set pin 13 high. The next line `delay` instructs the Arduino processor to do nothing for 200 milliseconds or ⅕ of a second. The next two lines set pin 13 low and then delay for another 200 milliseconds before the cycle starts again.

Variables and Constants

One of the most important ideas in programming is that of *variables*. A variable is a way of giving a name to a value. For example, the Blink sketch makes pin 13 toggle on and off, but if you wanted to use pin 10 instead, you would have to change every occurrence of "13" to "10" (three places) in the sketch. That's not so bad in this sketch, but in a more complicated sketch, there might be lots of places where the pin number would need to be replaced, and missing one of them when making a change to the program could result in a bug that takes awhile to track down and fix.

To avoid problems like this, it is a good idea to use a variable to give the pin a name. This has the added advantage that rather than just appearing as a number that could mean anything, the pin now has a logical name that will tell someone looking at the code what it is for. In the code below, Blink has been improved by the addition of a variable to define the LED pin. The changes are highlighted in bold. Because typing in code is tedious and error prone, all the example sketches are available as a download from the GitHub page for the book at the website

https://github.com/simonmonk/tyee6

To download them, click on the "Download ZIP" button on the GitHub page. Extract the ZIP file to some convenient location. Each listing in this chapter has a comment at the top that identifies the sketch file in the downloads.

```
// blink_variable
int ledPin = 13;

void setup() {
  pinMode(ledPin, OUTPUT);
}

void loop() {
  digitalWrite(ledPin, HIGH);
  delay(200);
  digitalWrite(ledPin, LOW);
  delay(200);
}
```

You could take this a stage further and use a variable for the delay value like this:

```
int ledPin = 13;

int blinkDelay = 200;
void setup() {
  pinMode(ledPin, OUTPUT);
}
void loop() {
  digitalWrite(ledPin, HIGH);
  delay(blinkDelay);
  digitalWrite(ledPin, LOW);
  delay(blinkDelay);
}
```

Variable names must be a single word (no spaces), and by convention, they start with a lowercase letter and use an uppercase letter to separate the different parts of the variable name so that the variable name can be usefully descriptive. In this example, the most logical name for the variable (ignoring all syntax rules and conventions for a moment) would be `LED pin`, but we are discouraged from starting the variable name with an uppercase letter, and we cannot have a space in the variable name. So we end up with a variable name of `ledPin` that satisfies the requirements of starting with a lowercase letter and being all one word. The uppercase `P` of `Pin` helps us to read the variable name.

Variables are defined right at the start of the program before `setup`. The word `int` is short for *integer* and specifies that the variable will contain a whole number. Later on we will meet other types of variables.

In the example above, once `ledPin` and `blinkDelay` are defined, there are no instructions later on in the program code that would change them. This means that they can be called *constants*. You can, optionally, tell the Arduino compiler that these variables are constants by prefixing them with the word `const`, as shown below:

```
const int ledPin = 13;
const int blinkDelay = 200;
```

The code will work whether you insert `const` or not, but if you do include `const`, it tells someone reading the code that the variable isn't going to change later in the program, and it also allows the compiler to make slightly more efficient and compact code, saving you a few bytes of program size and RAM usage while the program runs.

The Serial Monitor

It can sometimes be a little difficult to know what the Arduino is doing. Yes, it can blink its built-in LED, but apart from that, if you have a problem with the software, it can be tricky to work out just what is going wrong. The same USB interface that allows you to upload a sketch from your computer to your Arduino can also be used to allow the Arduino to communicate with your computer and provide valuable feedback on the value of variables and what the program is doing.

To try out the Serial Monitor, modify your Blink sketch so that it appears as below.

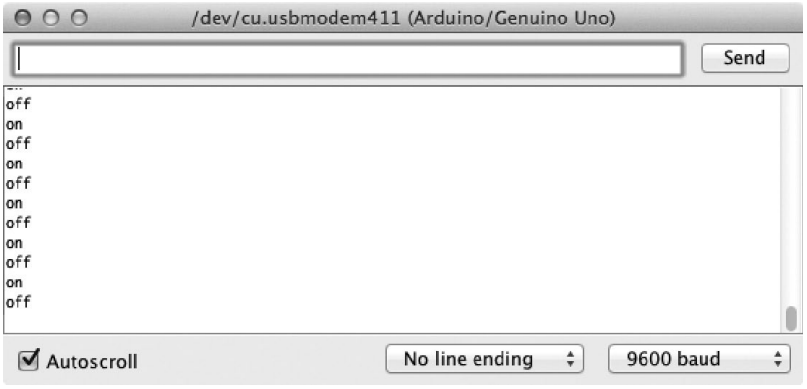
```
// blink_serial_monitor
void setup() {
  pinMode(13, OUTPUT);
  Serial.begin(9600);
}

void loop() {
  Serial.println("on");
  digitalWrite(13, HIGH);
  delay(200);
  Serial.println("off");
  digitalWrite(13, LOW);
  delay(200);
}
```

The line that has been added to `setup` starts serial communication with the Arduino IDE running on your computer at the baud rate specified. A baud rate of 9600 is the default. You can change this to a higher or lower value, but you will have to also change the drop-down list on the Serial Monitor to match the value that you use in the sketch.

Now, inside the `loop` function, two new lines have been added that send a text message of `on` and `off` over the serial connection to the Serial Monitor.

Upload the sketch to your Arduino, and you should find that it behaves just like it did before you modified it. However, if you click on the Serial Monitor icon on the Arduino IDE, then the Serial Monitor window will open, as shown in Fig. 30-4.

Each time the LED is turned on and off, you will see the message `on` or `off` appear in the Serial Monitor.



30-4 The Serial Monitor.

The values `on` and `off` inside the `Serial.println` commands are called *strings*. They are the C language's way of representing text. If you are from a conventional programming background, then you are probably familiar with the concept of strings and use them extensively in your programming. Strings are not commonly used when writing programs for an Arduino, as an Arduino is often used in applications that do not have any means of displaying text. The exceptions to this are situations in which the Arduino is communicating with a device like your PC that does have a means of displaying text, or when you have attached display hardware to the Arduino.

Ifs

As a program runs, the normal sequence of events is to run one command after the other. However, sometimes you will need to run only some of the commands if a condition is true. For example, you might want to run only certain commands when a switch, connected to a digital input, has been pressed. Another example might be that you want to run the command to turn a digital output on only if the temperature, measured using a temperature sensor, is greater than a certain value. The mechanism for doing this is to use the C `if` command. Assuming that you already have a variable called `temperature` that contains the temperature, you could write:

```
if (temperature > 90) {
  Serial.println("Its hot!");
}
```

Don't worry for now where `temperature` gets its value from. The important point is the structure of the `if` command. After the word `if`, there is an expression in parentheses called the "condition," and in this case, the condition is that the value in the variable `temperature` is greater than (`>`) 90. There then follows an { to indicate the start of a block of code. All the lines of code between this { and the corresponding } will be run only if the temperature indeed exceeds 90. The line `Serial.println("Its hot!");` is tabbed right to show that it belongs to the `if` command.

Iteration

Another common departure from simply executing a series of commands one after the other is to repeat those commands a number of times. Although the `loop` function in an Arduino sketch will repeat the commands it contains indefinitely, sometimes you need more control over how many times some lines of code are executed.

The main C language command for repeating things a number of times is the `for` command. The following sketch uses a `for` command to send the numbers 1 to 10 to the Serial Monitor.

```
// count_to_ten_once
void setup() {
  Serial.begin(9600);
  for (int i = 1; i <= 10; i++) {
    Serial.print(i);
  }
}

void loop() {
}
```

Because we want the Arduino to count to 10 only once, the `for` loop is in `setup`. If we wanted the sketch to count to 10 over and over, then we would move the three lines of the `for` loop to the currently empty `loop` function so that the sketch looked like this:

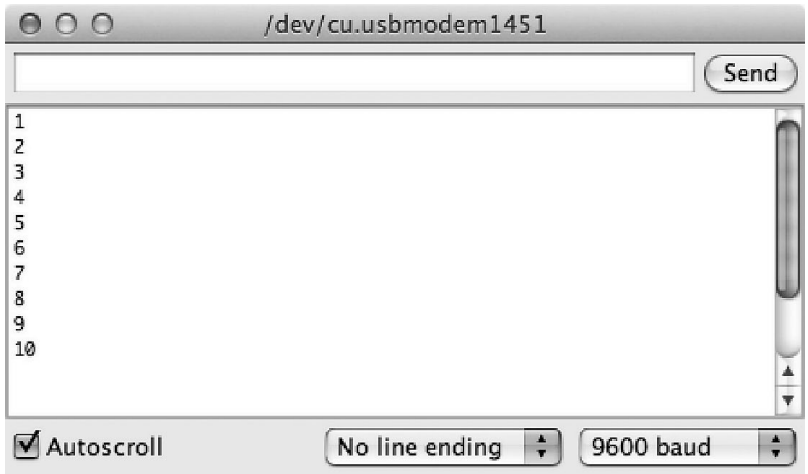
```
// count_to_ten_repeat
void setup() {
  Serial.begin(9600);
}

void loop() {
  for (int i = 1; i <= 10; i++) {
    Serial.println(i);
  }
}
```

The parenthesized expression after the word `for` contains three little snippets of code separated by semicolons. The first defines a counter variable called `i`, and the second section is the condition for staying in the loop. In this case, that means that the code will stay in the loop as long as `i` is less than or equal to (<=) 10. The final section (`i++`) means that 1 will be added to the value of `i` each time around the loop. The lines of code in between the { and } of the `for` loop will be run repeatedly until `i` is no longer less than or equal to 10.

The end result of this process is that the numbers between 1 and 10 get displayed on the Serial Monitor, as shown in Fig. 30-5.

There is a second type of `loop` command in C called a `while` loop that you will find useful from time to time.

The `while` command looks rather like an `if` command at first glance, but rather than just doing the things inside its { and } if the condition is true, it repeatedly executes the commands inside the { and } while the condition is true, stopping only when the condition is not true. At this point, the program continues to run the lines of code (if there are any) after the `while` loop.



30-5 The Serial Monitor output counting to 10.

If you rewrite the `for` example above to use `while` it would look like this:

```
// count_to_ten_once_while
void setup() {
  Serial.begin(9600);
  int i = 1;
  while (i <= 10) {
    Serial.println(i);
    i++;
  }
}

void loop() {
}
```

The variable `i` is now defined before the loop starts.

Functions

Functions are named blocks of code. Every sketch has to contain a `setup` and a `loop` function, but you can also define your own functions. This tends to happen if there are some lines of code that you want to use in several places in your sketch. Rather than repeat them in the code, you define them as a function. Writing you own functions helps to make your programs easier to understand. In the following example sketch, there is a user-defined function called `blink` that makes the "L" LED pin 13 blink 10 times.

```
// blink_function_broken
const int ledPin = 13;
void setup() {
  pinMode(ledPin, OUTPUT);
}

void loop() {
}

void blink() {
  for (int i = 1; i <= 10; i++) {
    digitalWrite(ledPin, HIGH);
    delay(1000);
    digitalWrite(ledPin, LOW);
    delay(1000);
  }
}
```

If you were to run this sketch, then the LED would not actually blink at all. That is because, although you have defined the `blink` function saying what the code must do and giving it a name, nowhere in the sketch do you actually "call" the function, telling it to run.

This separation of *defining* a function versus *running* a function is a very important distinction. In defining the function, we have created a named piece of code that knows how to "blink," but nowhere have we actually told it to go ahead and do some blinking. It's a bit like giving someone written instructions for making a cup of coffee but not actually telling them to go ahead and make a cup.

To fix this so that you actually get some blinking of the LED, you can put the line highlighted in bold into `setup`.

```
// blink_function
const int ledPin = 13;
void setup() {
  pinMode(ledPin, OUTPUT);
  blink();
}
void loop() {
}

void blink() {
  for (int i = 1; i <= 10; i++) {
    digitalWrite(ledPin, HIGH);
    delay(1000);
    digitalWrite(ledPin, LOW);
    delay(1000);
  }
}
```

It does not matter where in the sketch you define the function, although it is most common to put your own functions after `setup` and `loop`. To execute or "call" the `blink` function, use the name of the function followed by `()`.

The current `blink` function is pretty inflexible: it can blink only 10 times, it always blinks `ledPin`, and the fixed delay value means it can only blink at one speed. You can make this function much more flexible and general purpose by "parameterizing" the function so it reads

```
// blink_function_params
const int ledPin = 13;

void setup() {
  pinMode(ledPin, OUTPUT);
  blink(ledPin, 20, 200);
}

void loop() {
}

void blink(int pin, int times, int period) {
  for (int i = 1; i <= times; i++) {
    digitalWrite(pin, HIGH);
    delay(period);
    digitalWrite(pin, LOW);
    delay(period);
  }
}
```

The changes are highlighted in bold. The first change is that the call to `blink` inside `setup` now has three things in the parentheses separated by commas. These are called *parameters* and will be "passed to" the function when it is called. The first is the pin to blink (in this case `ledPin`). The second is the number of times to flash, and the last is the delay period between turning the pin on and off.

The parameters inside the function are called *local variables* because they apply only inside the function. So when the function is called and the first parameter supplied is `ledPin`, the value of `ledPin` will be transferred to the local variable `pin` inside the function. Such local variables are accessible only within the function itself, whereas the other variables that we have met so far, such as `ledPin`, are called *global variables* because they are accessible throughout the sketch.

Data Types

A variable of type `int` in Arduino C uses two bytes of data. The type `int` is used for most variables. An exception is when the `int` range of −32768 to 32767 is not enough because you want to represent a number greater than 32767 or less than −32768, in which case a `long` using four bytes of data will give you big numbers.

Another situation in which an `int` won't work arises when you want to represent real numbers that have digits after the decimal place. The `float` data type uses the binary equivalent of scientific notation. That is, the number is split into a mantissa and exponent. This gives an enormous range of values but limited precision.

Using 0.0 rather than just 0 helps to emphasize that the number is a real number and not just an integer. Table 30-1 breaks down the data types available.

So far our variables have all been `int`s and declared like

```
int x = 0;
```

Setting the initial value for the variable by following the declaration with = and then a value is optional, but considered good practice, as it removes any ambiguity about the value of the variable.

To declare a `float` you write something like

```
float x = 0.0;
```

Table 30-1. Data Types in Arduino C

Type	Memory (bytes)	Range	Notes
`boolean`	1	true or false (0 or 1)	
`char`	1	−128 to +128	Used to represent an ASCII character code, e.g., A is represented as 65. Its negative numbers are not normally used.
`byte`	1	0 to 255	Often used for communicating serial data, as a single unit of data.
`int`	2	−32768 to +32767	
unsigned `int`	2	0 to 65536	Can be used for extra precision where negative numbers are not needed. Use with caution as arithmetic with `int`s may cause unexpected results.
`long`	4	−2,147,483,648 to 2,147,483,647	Needed only for representing very big numbers.
unsigned `long`	4	0 to 4,294,967,295	See unsigned `int`
`float`	4	−3.4028235E+38 to + 3.4028235E+38	
`double`	4	same as `float`	Normally this would be 8 bytes and higher precision than `float` with a greater range. However on Arduino it is the same as `float`.

Generally speaking, when performing calculations that involve a mixture of different types (say `int`s and `float`s), the compiler does a pretty good job of automatically converting types, as you would expect. For example, the following code would produce the expected result of 25,000,000.00:

```
// calc_1
void setup() {
  Serial.begin(9600);
  float x = 5000.0;
  int y = 5000;
  float result = x * y;
  Serial.println(result);
}

void loop() {
}
```

The dangerous situations in which you may not get the result you expect arise when the combined number gets too big for the `int` data type, for example:

```
// calc_2
void setup() {
  Serial.begin(9600);
  int x = 500;
  int y = 500;
  int result = (y * x) / 1000;
  Serial.println(result);
}

void loop() {
}
```

In this case, you would expect `result` to be 250, or 250,000/1000. However, if you run this sketch, the result will actually be −12. That happens because the first step of the calculation is to multiply 500 by 500, which gives a result of 250,000, which is above the limit for an `int`. In C, once the number exceeds the limit, the value "wraps-around" to negative numbers, giving meaningless results like the one here. If you try changing `x` and `y` to be `long`s the result will be as expected.

In summary, whenever you are doing arithmetic, always think through the maximum values that might arise, even for intermediate values in the calculations, and use a data type that has a big enough range for them.

Interfacing with GPIO Pins

When it comes to using the Arduino's GPIO pins, there are a number of built-in functions that you use first to define whether the pin is to act as an input or an output, and then, either to read the value of an input into the sketch or write a value to an output.

Setting the Pin Mode

Unless otherwise specified, an Arduino pin is an input without the pin's pull-up resistor being active. The `pinMode` command allows you to choose whether the pin is to act as an input or output and whether the pull-up resistor should be enabled or not.

Normally the mode of a pin is set in the `setup` function, but you can change the mode of a pin at any time as the sketch runs.

The `pinMode` built-in function takes two parameters. The first is the pin whose mode is to be set, and the second is the *mode*. The mode must be INPUT, INPUT_PULLUP, or OUTPUT. These are constants defined in Arduino C.

Digital Read

To read the digital value of an input pin, the built-in function `digitalRead` is used. This takes the pin to be read as its parameter and *returns a value* of 0 or 1. Returning a value means that the result of reading the digital input can be assigned to a variable, as shown in the following example sketch:

```
// digital_read
const int inputPin = 7;

void setup() {
  Serial.begin(9600);
  pinMode(inputPin, INPUT);
}

void loop() {
  int x = digitalRead(inputPin);
  Serial.println(x);
  delay(1000);
}
```

The variable `x` is defined inside the `loop` function, making it a local variable accessible only within the `loop` function. The result of the `digitalRead` function call will be assigned to the variable and will be 1 if the pin is high and 0 if the pin is low. Two special constants called `HIGH` and `LOW` are defined in the Arduino code so that you can use them instead of 1 and 0 respectively.

If you wanted to make a somewhat over-engineered light switch, you could attach a switch to pin 7 of the Arduino, and then install the following sketch to turn the built-in "L" LED on when the switch is pressed.

```
 // digital_read_switch
const int switchPin = 7;
const int ledPin = 13;

void setup() {
  pinMode(switchPin, INPUT_PULLUP);
  pinMode(ledPin, OUTPUT);
}

void loop() {
  if (digitalRead(switchPin) == LOW) {
    digitalWrite(ledPin, HIGH);
  }
  else {
    digitalWrite(ledPin, LOW);
  }
}
```

Rather than use a variable to store the result of the `digitalRead` (as we did in the previous example), the `digitalRead` function is used directly in the condition part of the `if` command. To compare two things to see if they are equal, you use a double equals sign (==) rather than the

single sign (=) that you use to assign a value to a variable. The condition is that the result of calling `digitalRead` be `LOW` because the input is normally `HIGH` due to the pull-up resistor and only goes `LOW` when the switch is closed.

In this case, the `if` statement has an `else` counterpart that is run if the condition is not true.

The above example could, of course, easily be made without an Arduino at all, simply by putting a switch, an LED, and a current limiting resistor in series and connected to a voltage source.

If you wanted to change things so that pressing a push switch toggled the LED between on and off, then you would need the sketch to use a variable to keep track of whether the LED was last on or off (called its *state*). This is what such a sketch would look like:

```
// digital_read_toggle
const int switchPin = 7;
const int ledPin = 13;

int ledState = LOW;

void setup() {
  pinMode(switchPin, INPUT_PULLUP);
  pinMode(ledPin, OUTPUT);
}

void loop() {
  if (digitalRead(switchPin) == LOW) {
    ledState = ! ledState;
    digitalWrite(ledPin, ledState);
    delay(100);
    while (digitalRead(switchPin) == LOW) {}
  }
}
```

Now, when `digitalRead` detects that the switch button has been pressed, the variable `ledState` is toggled using the not (`!`) command. This sets `ledState` to be the inverse of its current setting. That is, if `ledState` is `HIGH`, it sets it `LOW` and vice-versa.

The `digitalWrite` function is then used to set the output to the new state. After that, there is a delay of 100 milliseconds that allows time for the switch contacts to settle, as they often "bounce" between high and low during the pressing of the switch.

You don't want `ledState` to be immediately toggled again, so the `while` loop effectively waits until the switch has been released.

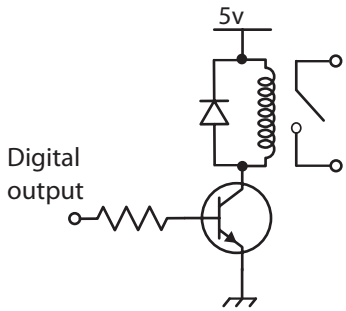
Digital Write

You have already used the `digitalWrite` command to turn the "L" LED on the Arduino board on and off. The command takes two parameters, the first being the pin to control and the second being 1 or 0 for `HIGH` and `LOW`, respectively.

The command will control the pin only if the pin has already been set to be a digital output using the `pinMode` command.

An Arduino pin can source or sink 40mA without any risk of damaging the ATmega328 microcontroller, which is fine for controlling an LED directly, but other devices, such as a relay or DC motor, will require some current amplification.

Because we are concerned only with turning things on and off, the amplification does not need to be linear, so a simple control with a single transistor is all that is required in most circumstances.



30-6 Controlling a relay from a digital output.

Figure 30-6 shows the use of an NPN bipolar transistor to provide the 50 to 100mA (or so) needed to drive the coil of a typical relay. The resistor should limit the current to less than 40mA, so 150 Ω would be ideal. A simple low-cost transistor, like the 2N3904, is suitable. The diode across the relay coil is necessary to snub (that is, suppress) any pulses of voltage resulting from driving the inductive load of the relay coil.

Analog Input

The Arduino has six pins labeled A0 to A5 that can provide 10-bit analog inputs. The built-in function `analogRead` returns a number between 0 and 1023 corresponding to the voltage at the analog input. The following sketch illustrates this, along with the math necessary to convert the reading into a voltage and print it out at the Serial Monitor.

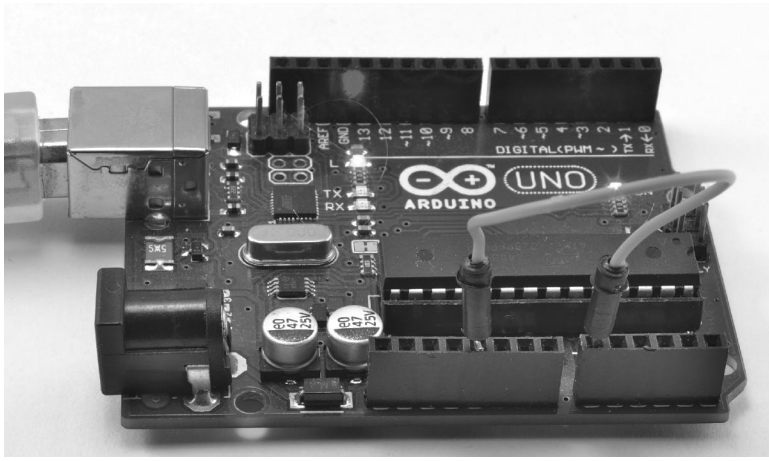
```
// analog_read
const int analogPin = A0;

void setup() {
  Serial.begin(9600);
}

void loop() {
  int reading = analogRead(analogPin);
  float volts = reading * 5.0 / 1023.0;
  Serial.println(volts);
  delay(1000);
}
```

The `int reading` is multiplied by 5.0 and divided by 1023.0. That is, both 5.0 and 1023.0 have a decimal point so that C knows they are `floats` rather than `ints`.

The Serial Monitor will display the voltage at pin A0 once a second, and you can actually use this scheme to measure some of the voltages on the Arduino board itself. Figure 30-7 shows a jumper wire linking A0 to GND.

While this is in place, the readings should be 0 V. Change the jumper wire so that it now links A0 and the "3-V" pin. The Serial Monitor should report a voltage of around 3.3 V. Finally connect A0 to the Arduino's "5-V" pin. Figure 30-8 shows the output of the Serial Monitor.

Although you need to be careful not to exceed the maximum input voltage to an analog input of 5 V, you can, of course, use a pair of resistors as a voltage divider, if you need to measure higher voltages.



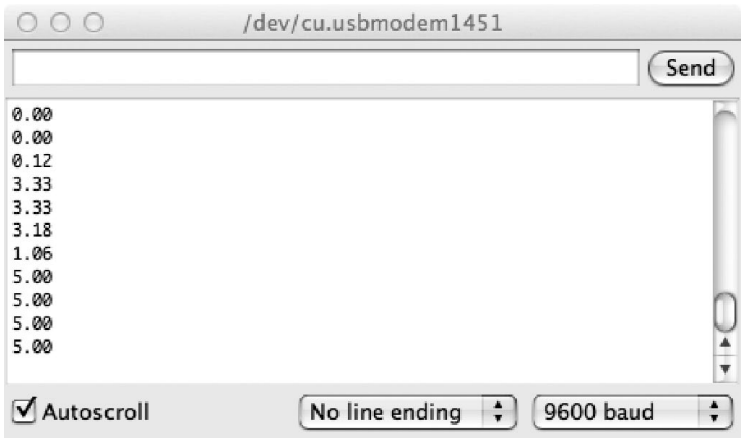
30-7 An Arduino measuring its own voltages.

Analog Write

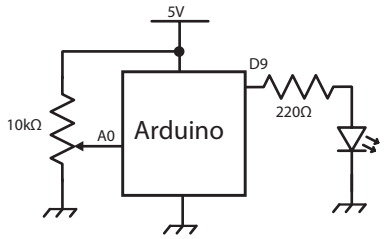
An Arduino Uno does not have true analog outputs, but rather uses *pulse width modulation* (PWM) as described in Chap. 29. Only the pins on an Arduino Uno marked with a ~ (3, 5, 6, 9, 10, and 11) can provide hardware supported PWM.

The command to set the duty cycle of such an output is the `analogWrite` command. This takes two parameters: the pin to control and the duty cycle as an `int` between 0 and 255. A value of 0 means the duty cycle will be 0 and the output will be fully off. A value of 255 and it will be fully on. In fact it's just like `digitalWrite` except that instead of the value being 0 or 1 (LOW or HIGH) the value is between 0 and 255.

To illustrate, we can use an Arduino with a potentiometer attached to control the brightness of an LED, as shown in Fig. 30-9.



30-8 The Serial Monitor reporting analog readings.



30-9 Controlling LED brightness with a potentiometer.

The sketch for this project simply takes the analog reading from the potentiometer (0 to 1023) and divides it by 4 to put it in the range 0 to 255 needed by analogWrite.

```
// analog_write
const int potPin = A0;
const int ledPin = 9;

void setup() {
  pinMode(ledPin, OUTPUT);
}

void loop() {
  int reading = analogRead(potPin);
  analogWrite(ledPin, reading / 4);
}
```

As you turn the potentiometer knob, the apparent brightness of the LED will increase as the duration of the pulses arriving at it increases. This way of controlling the apparent brightness of an LED actually works much better than controlling the voltage to the LED because, until the voltage reaches the working forward voltage of the LED (typically at least 1.6 V), the LED will not emit any light at all.

The Arduino C Library

There are a large number of commands available in the Arduino library, some of which you have already met. The most commonly used commands are listed in Table 30-2. For a full reference to all the Arduino commands, see the official Arduino documentation at the website

http://www.arduino.cc

Libraries

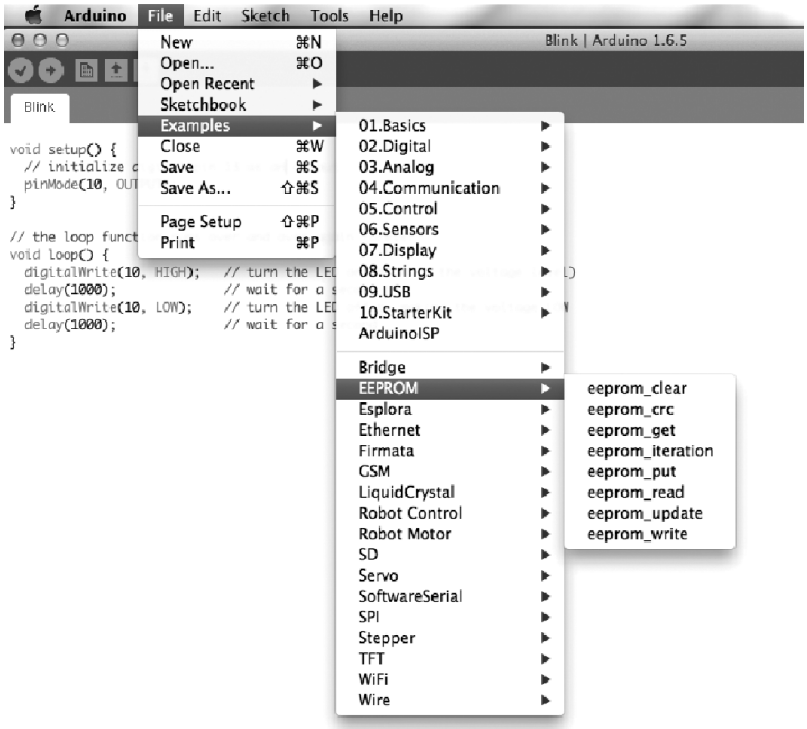
The Arduino IDE uses the concept of libraries to organize code that you might want to use in your sketches. These libraries contain program code that others have written and shared. This makes it easier to interface with certain types of hardware.

For example, the Arduino IDE comes supplied with a load of libraries pre-installed that you can make use of. It is a convention that libraries should include some example sketches that will get you started quickly when using the libraries. You can get an idea of the libraries included with the Arduino IDE from the "File" and then the "Examples" menu (see Fig. 30-10).

The top half of the example sketches list is not related to libraries, but below the line, all the sketches are contributed from libraries. For example, to store values persistently so that they are not

Table 30-2. Arduino Library Functions

Command	Example	Description
Digital IO		
`pinMode`	`pinMode(8, OUTPUT);`	Sets pin 8 to be an output. The alternative is to set it to be `INPUT` or `INPUT_PULLUP`.
`digitalWrite`	`digitalWrite(8, HIGH);`	Sets pin 8 high. To set it low, use the constant `LOW` instead of `HIGH`
`digitalRead`	`int i;` `i = digitalRead(8);`	This will set the value of `i` to `HIGH` or `LOW` depending on the voltage at the pin specified (in this case pin 8).
`pulseIn`	`i = pulseIn(8, HIGH)`	Returns the duration in microseconds of the next `HIGH` pulse on pin 8
`tone`	`tone(8, 440);`	Make pin 8 oscillate at 440 Hz.
`noTone`	`noTone(8);`	Cut short the playing of any tone that was in progress on pin 8.
Analog IO		
`analogRead`	`int r;` `r = analogRead(A0);`	Assigns a value to `r` of between 0 and 1023: 0 for 0 V, 1023 if the pin A0 is 5 V.
`analogWrite`	`analogWrite(9, 127);`	This command outputs a PWM signal. The duty cycle is a number between 0 and 255, 255 being 100 %. This must be used by one of the pins marked as PWM on the Arduino board (3, 5, 6, 9, 10, and 11).
Time Commands		
`millis`	`unsigned long l;` `l = millis();`	The variable type `long` in Arduino is represented in 32 bits. The value returned by `millis()` will be the number of milliseconds since the last reset. The number will wrap around after approximately 50 days.
`micros`	`long l;` `l = micros();`	See `millis`, except this is microseconds since the last reset. It will wrap after approximately 70 minutes.
`delay`	`delay(1000);`	Delay for 1000 milliseconds or 1 second.
`delayMicroseconds`	`delayMicroseconds(100000);`	Delay for 100,000 microseconds. Note the minimum delay is 3 microseconds, the maximum is around 16 milliseconds.
Interrupts		
`attachInterrupt`	`attachInterrupt(1,` `myFunction, RISING);`	Associates the function `myFunction` with a rising transition on interrupt 1 (D3 on an Uno).
`detachInterrupt`	`detachInterrupt(1);`	Disables any interrupt on interrupt 1.



30-10 Library example sketches.

lost when the Arduino is reset, you can use the EEPROM library. The "eeprom_clear" sketch is shown below with some of the comments removed for brevity.

```
#include <EEPROM.h>
void setup()
{
  for ( int i = 0 ; i < EEPROM.length() ; i++ ) {
    EEPROM.write(i, 0);
  }

  // turn the LED on when we're done
  digitalWrite(13, HIGH);
}
void loop(){ /** Empty loop. **/ }
```

To show the Arduino IDE that a library is required the `#include` command is used, followed by the name of the library's header file. The best way to get the command right for your own sketch is just to copy it from one of the example sketches.

Other libraries that are included with the IDE are for network programming with network hardware (Ethernet), liquid-crystal displays (LCDs), reading and writing SD cards (SD) controlling servomotors (Servo), and stepper motors (Stepper).

If you have a piece of hardware that you want to control from your Arduino, then chances are there will already be a library for it included in the Arduino IDE, or there will be a library written by someone else that you can install in your Arduino IDE.

The Arduino community is very good at creating and sharing libraries. In the spirit of open-source cooperation, libraries are almost always provided free of charge and without any kind of licensing restrictions on their use. When you find a library that you want to use, it will be in the form of a ZIP file. Download the ZIP file, and then from the Arduino IDE, select the menu option "Sketch," then "Include Library," and then "Add ZIP Library," and navigate to the ZIP file you just downloaded. This will install the library, and if you now look at the menu option "File" and then "Examples," you should see a new set of examples for the library you just installed.

Special Purpose Arduinos

The Arduino Uno is by far the most commonly used official Arduino however, there are other models of Arduino that are better suited to some situations. Some of these are official Arduino boards, and others are made by third parties and use the Arduino IDE or a different IDE, but the same Arduino C language.

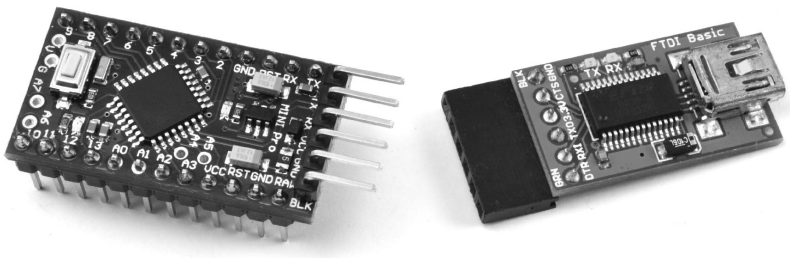
There are many Arduino and Arduino-compatible boards. New boards are being developed all the time, so rather than attempt a comprehensive survey, let's focus on three representative Arduino boards.

The Arduino Pro Mini

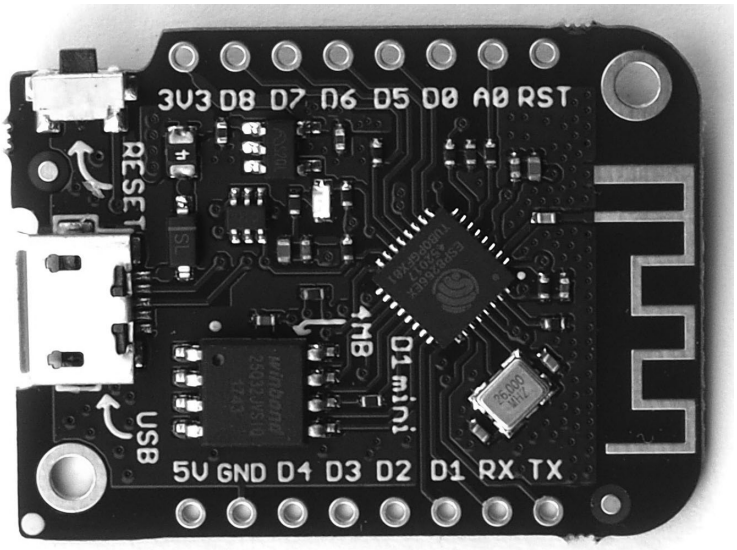
The Arduino Uno has a lot of components and features. This can make it a little wasteful to embed an Arduino Uno into a project. For example, the Arduino has a built-in USB interface, and if you need this only while programming the board, it is unnecessary to have that USB interface permanently attached to your project.

The Arduino Pro Mini (Fig. 30-11) separates the USB interface from the rest of the Arduino so that the Arduino itself is smaller and lower-cost. This makes it more reasonable to embed an Arduino permanently into a project once you have perfected it. If you need to update the Arduino's sketch, you can just reattach the USB interface and program it again.

Programming an Arduino Pro Mini is just like programming an Arduino Uno. You just have to select a board type of "Arduino Pro or Pro Mini" in the "Board" section of the "Tools" menu.



30-11 The Arduino Pro Mini (left) and USB interface (right).



30-12 The Wemos D1.

Wemos D1

The Wemos D1 (Fig. 30-12) uses the ESP8266 processor, can be programmed from the Arduino IDE and offers tremendous value for money, as it includes Wi-Fi capabilities making it well-suited to Internet of Things projects.

If you need more GPIO pins, and/or more processing power, then it's causing the LOLIN32 and other boards based on the ESP32 processor also offer great value for money.

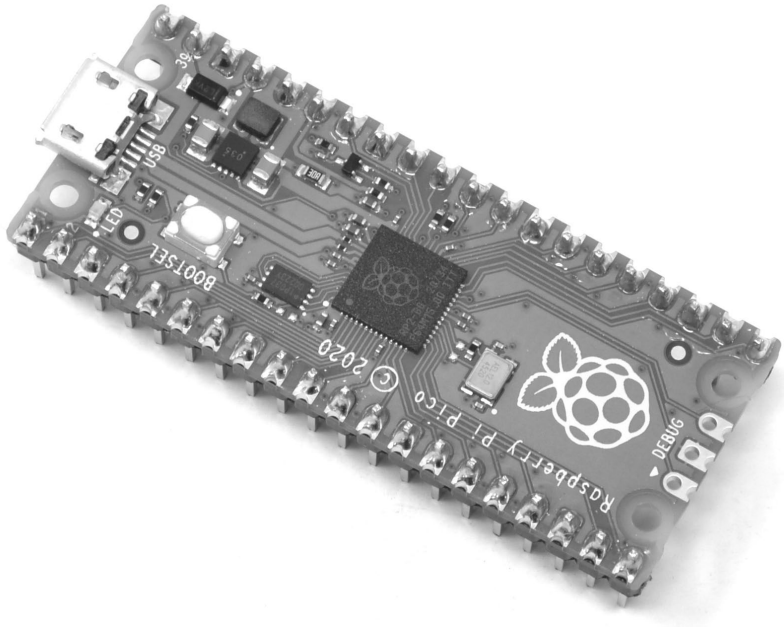
Raspberry Pi Pico

The Raspberry Pi Pico (Fig. 30-13) can also be programmed from the Arduino IDE. It uses the RP2040 processor designed by the Raspberry Pi foundation, and offers an interesting alternative to the Arduino Pro Mini in terms of size. However, it has a built-in USB interface and a powerful ARM m0 chip that has two processors.
New section

Adafruit Feather Boards

Adafruit have pioneered a wide ranging system of microcontroller and shield boards in their Feather range. Boards are available for a huge range of microcontroller ICs including the RP2040 and ESP boards mentioned previously.

These boards are often programmed in the Python programming language, but most can also be programmed using the Arduino IDE. You can find out more about this range of boards here: https://www.adafruit.com/category/943.



30-13 The Raspberry Pi Pico.

Shields

The Arduino header sockets can be used to attach so-called *shields*, which stack on top of the Arduino and are available for all sorts of purposes. An Internet search will find you no end of different shields for all sorts of purposes, including:

- Motor control
- Relays
- Ethernet and Wi-Fi
- Various types of display
- Sensors

Shields often have an accompanying Arduino library that makes them easy to use.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

31
CHAPTER

Transducers and Sensors

IN THIS CHAPTER, YOU'LL LEARN ABOUT ELECTRONIC DEVICES THAT CONVERT ENERGY FROM ONE form to another, devices that can detect phenomena and measure their intensity.

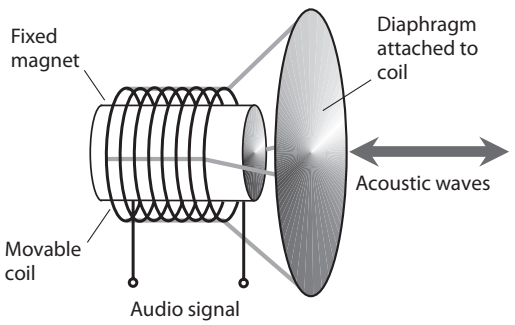
Wave Transducers

In electronics, *wave transducers* convert AC or DC into acoustic or electromagnetic (EM) waves. They can also convert these waves into AC or DC signals.

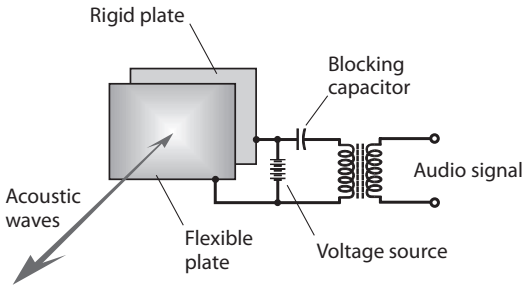
Dynamic Transducer for Sound

A *dynamic transducer* comprises a coil and magnet that translates mechanical vibration into varying electrical current or vice-versa. The most common examples are the *dynamic microphone* and the *dynamic speaker*.

Figure 31-1 is a functional diagram of a dynamic transducer. A diaphragm is attached to a coil that can move back and forth rapidly along its axis. A permanent magnet rests inside the coil. Sound waves cause the diaphragm and coil to move together, producing fluctuations in the magnetic field within the coil. As a result, audio AC flows in the coil, having the same waveform as the sound that strikes the diaphragm.



31-1 Functional diagram of a dynamic sound transducer.



31-2 Functional diagram of an electrostatic sound transducer.

If we apply an audio signal to the coil, the AC in the wire generates a magnetic field that produces forces on the coil. These forces cause the coil to move, pushing the diaphragm back and forth to create acoustic waves in the surrounding air.

Electrostatic Transducer for Sound

An *electrostatic transducer* takes advantage of the forces produced by electric fields. Two metal plates, one flexible and the other rigid, are placed parallel to each other and close together, as shown in Fig. 31-2.

In an *electrostatic pickup*, incoming sound waves vibrate the flexible plate, producing small, rapid changes in the spacing, and therefore, the capacitance, between the plates. We apply a constant DC voltage between the plates. As the plate-to-plate capacitance varies, the electric field intensity between them fluctuates, causing variations in the current through the transformer primary winding. Audio signals appear across the secondary winding.

In an *electrostatic emitter*, fluctuating currents in the transformer produce changes in the voltage between the plates. This AC voltage results in electrostatic field variations, producing forces that push and pull the flexible plate in and out. The motion of the flexible plate produces sound waves in the air.

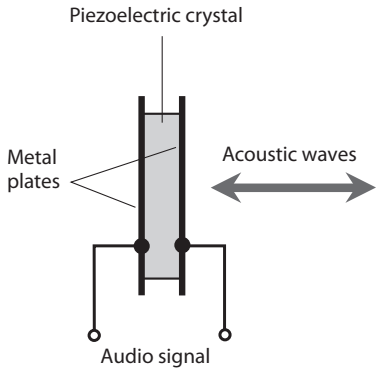
We can use electrostatic transducers in most applications in which dynamic transducers will work. Advantages of electrostatic transducers include light weight and good sensitivity. The relative absence of magnetic fields can also constitute an asset in certain situations.

Piezoelectric Transducer for Sound and Ultrasound

Figure 31-3 shows a *piezoelectric transducer* that consists of a slab-like *crystal* of quartz or ceramic material sandwiched between two metal plates. Piezoelectric transducers can function at higher frequencies than dynamic or electrostatic transducers can, so piezoelectric transducers are favored in ultrasonic applications such as intrusion detectors.

When acoustic waves strike one or both of the plates, the metal vibrates. This vibration transfers to the crystal by mechanical contact. The crystal generates weak electric currents when subjected to the mechanical stress. Therefore, an AC voltage develops between the two metal plates, with a waveform identical to that of the acoustic disturbance.

If we apply an electrical audio signal to the plates, the fluctuating current causes the crystal to vibrate in sync with the current. The metal plates vibrate also, producing an acoustic disturbance in the surrounding medium.



31-3 Functional diagram of a piezoelectric transducer for sound and ultrasound.

Transducers for RF Energy

The term *radio-frequency* (*RF*) *transducer* is a fancy expression for an *antenna*. Two basic types exist: the *receiving antenna* and the *transmitting antenna*.

Transducers for IR and Visible Light

Many wireless devices transmit and receive energy in the infrared (IR) spectrum, which spans frequencies higher than those of radio waves but lower than those of visible light. Some wireless devices transmit and receive EM signals in the visible range, although we will encounter them less often than we find IR devices.

The most common IR transmitting transducer is the infrared-emitting diode (IRED). When you apply fluctuating DC to the device, it emits IR rays. The fluctuations in the current constitute modulation, causing variations in the intensity of the rays emitted by the semiconductor P-N junction. The modulated IR carries information such as which channel you want your television (TV) set to "see" or whether you want to raise or lower the volume. You can focus an IR beam using optical lenses or mirrors to *collimate* the rays (make them parallel) for line-of-sight transmission through clear air over distances of up to several hundred meters.

Infrared receiving transducers resemble photodiodes or photovoltaic cells. The fluctuating IR energy from the transmitter strikes the P-N junction of the receiving diode. If the receiving device is a photodiode, you apply a current to it. This current varies rapidly in accordance with the signal waveform on the IR beam from the transmitter. If the receiving device is a photovoltaic cell, it produces the fluctuating current all by itself, without the need for an external power supply. In either case, the current fluctuations are weak, and you must amplify them before sending them to the equipment (TV set, garage-door opener, oven, security system, or whatever) controlled by the wireless system.

Displacement Transducers

A *displacement transducer* measures a distance or angle traversed, or the distance or angle separating two points. Conversely, a displacement transducer can convert an electrical signal into mechanical movement over a certain distance or angle. A device that measures or produces movement in a straight line constitutes a *linear displacement transducer*. If it measures or produces movement through an angle, we call it an *angular displacement transducer*.

Pointing and Control Devices

A *joystick* can produce movement or control variable quantities in two dimensions. The device has a movable lever attached to a ball bearing within a control box. You can manipulate the lever by hand up and down, or to the right and left. Some joysticks allow you to rotate the lever for control in a third dimension. Joysticks are used in computer games, for entering coordinates into a computer, and for the remote-control of robots.

A *mouse* is a peripheral device commonly used with personal computers. By sliding the mouse around on a flat surface, you can position a cursor or arrow on the computer display. Push-button switches on the top of the unit actuate the computer to perform whatever function the cursor or arrow shows. These actions are called *clicks*.

A *trackball* resembles an inverted mouse, or a two-dimensional joystick without the lever. Instead of pushing the device around on a flat surface, you manipulate a ball bearing with the index finger of one hand, causing the display cursor to move vertically and horizontally. Push-button switches on a computer keyboard, or on the trackball box itself, actuate the functions.

An *eraser-head pointer* is a rubber button approximately five millimeters (5 mm) in diameter, usually placed in the center of a computer keyboard. You move the cursor on the display by pushing against the button. Clicking is done with button switches on the keyboard.

A *touch pad* is a sensitive plate approximately the size and shape of a credit card. You place your index finger on the plate and move your finger around, producing movement of the display cursor or arrow. You can do clicks just as you do with a trackball or eraser-head pointer.

Electric Motor

An *electric motor* converts electrical energy into angular (and in some cases linear) mechanical energy. Motors can operate from either AC or DC, and range in size from tiny devices used in microscopic robots to huge machines that pull passenger trains. You learned the basic principle of the DC motor in Chap. 8. In a motor designed to work with AC, no commutator exists. Instead, the alternations in the current keep the polarity correct at all times, so the shaft does not "lock up." The rotational speed of an AC motor depends on the frequency of the applied AC. With 60-Hz AC, for example, the rotational speed equals 60 revolutions per second (60 r/sec) or 3600 revolutions per minute (3600 r/min). When you connect a motor to a load, the rotational force required to turn the shaft increases and the motor draws increasing power from the source.

Stepper Motor

A *stepper motor* turns in small increments, rather than continuously. The *step angle*, or extent of each turn, varies depending on the particular motor. It can range from less than 1° of arc to a quarter of a circle (90°). A stepper motor turns through its designated step angle and then stops, even if the coil current continues. When the shaft of a stepper motor has come to rest with current going through its coils, the shaft resists external rotational force; it "tries to stay in place."

Conventional motors run at hundreds or thousands of revolutions per minute. A stepper motor usually runs at far lower speeds, almost always less than 180 r/min. A stepper motor has the most turning power when operated at its slowest speeds, and the least turning power when operated at its highest speeds.

When we supply current pulses to a stepper motor at a constant frequency, the shaft rotates in increments, one step for each pulse. In this way, the device can maintain a precise speed. Because of the braking effect, this speed holds constant over a wide range of mechanical turning resistances.

Stepper motors work well in applications requiring point-to-point motion. Specialized robots can perform intricate tasks with the help of microcomputer-controlled stepper motors.

Electric Generator

An *electric generator* is constructed in much the same way as an AC motor, although it functions in the opposite sense. Some generators can also function as motors; we call dual-purpose devices of this type *motor/generators*.

A typical generator produces AC from the mechanical rotation of a coil in a strong magnetic field. Alternatively, a permanent magnet can be rotated within a coil of wire. We can drive the rotating shaft with a gasoline engine, a steam turbine, a water turbine, a wind turbine, or any other source of mechanical power. A commutator can work with a generator to produce pulsating DC output, which we can filter, if desired, to obtain pure DC to operate electronic equipment.

Small portable gasoline-powered generators, capable of delivering a few kilowatts, can be purchased in department stores or home-and-garden stores. Larger generators, which usually burn propane or methane ("natural gas"), allow homes or businesses to retain a continuous supply of electrical power during a utility disruption. The largest generators are found in power plants and can produce many kilowatts.

Small generators can function in synchro systems. These specialized generators allow remote control of robotic devices. A generator can be used to measure the speed at which a vehicle or rolling robot moves. The shaft of the generator is connected to one of the wheels, and the generator output voltage and frequency vary directly with the angular speed of the wheel. In this case we have a *tachometer*.

Optical Encoder

Potentiometers are often used as a volume control, but they have end-stops so that their angular range is limited. Optical encoders are also used to sense angular position.

An optical encoder comprises two LEDs, two photodetectors, and a device called a *chopping wheel*. The LEDs shine on the photodetectors through the wheel. The wheel, which has alternately transparent and opaque radial bands (Fig. 31-4), is attached to a rotatable shaft and a control knob. As we rotate the knob, the light beams are interrupted. Each interruption causes the frequency to change by a specified increment.

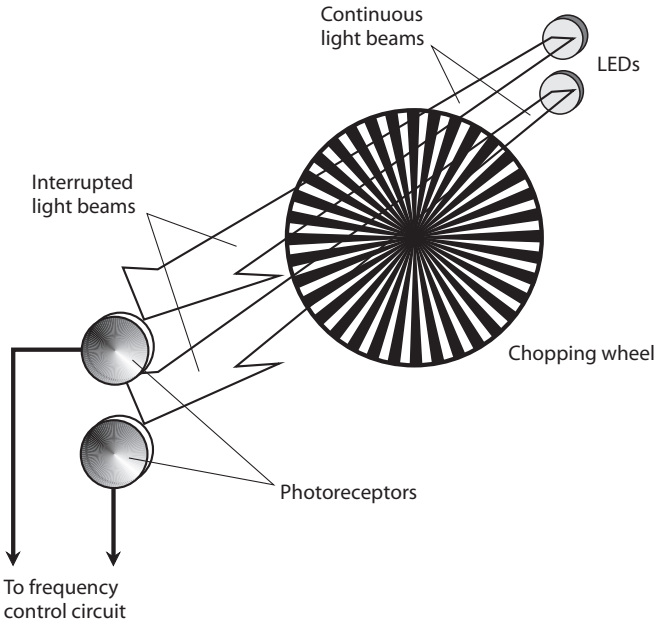
When used to detect rotation, an optical shaft encoder can sense the difference between the "up" command (clockwise shaft rotation) and the "down" command (counterclockwise shaft rotation) according to which photodetector senses each sequential beam interruption first. Optical sensors like this are expensive and a similar approach using mechanical contact called a quadrature encoder is often used in appliances.

Detection and Measurement

A *sensor* employs one or more transducers to detect or measure parameters, such as temperature, humidity, barometric pressure, pressure, texture, proximity, and the presence of certain substances.

Capacitive Pressure Sensor

Figure 31-5 portrays the functional details of a *capacitive pressure sensor*. Two metal plates, separated by a layer of compressible dielectric foam, create a variable capacitor that we connect in parallel with an inductor. The resulting inductance/capacitance (*LC*) circuit determines the frequency of

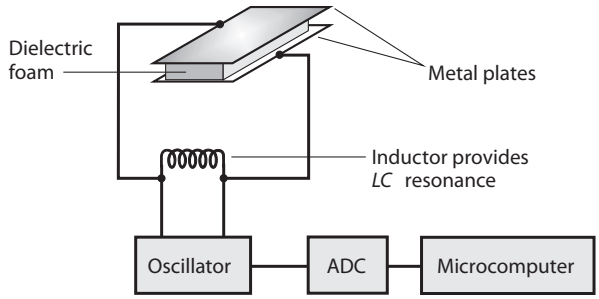


31-4 An optical encoder uses LEDs and photodetectors to sense the direction and extent of shaft rotation.

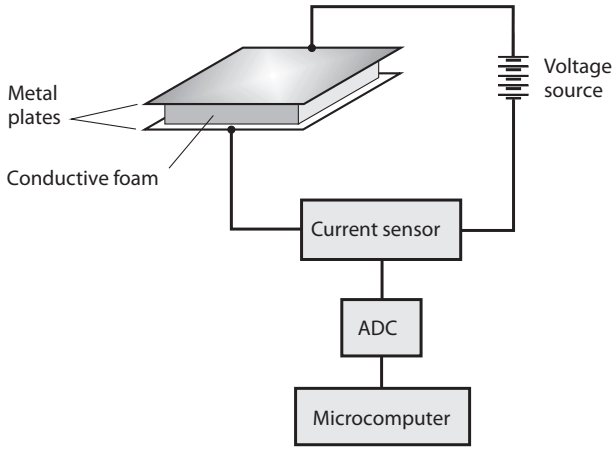
an oscillator. If an object strikes or pushes against the sensor, the plate spacing momentarily decreases, causing an increase in the capacitance and, therefore, a decrease in the oscillator frequency. When the object moves away from the transducer, the foam layer springs back to its original thickness, the plates return to their original spacing, and the oscillator frequency returns to normal.

We can convert the output of a capacitive pressure sensor to digital data using an *analog-to-digital converter* (ADC). This signal can go to a microcomputer, such as a robot controller. We can mount pressure sensors in various places on a mobile robot, such as the front, back, and sides. Then, for example, physical pressure on the sensor in the front of the robot can send a signal to the controller, which tells the machine to move backward.

A capacitive pressure sensor can be fooled by massive conducting or semiconducting objects in its vicinity. If such a mass comes near the transducer, the capacitance can change even if direct mechanical contact does not occur. We call this phenomenon *body capacitance*. In some applications, we can tolerate body capacitance; in other situations, we can't.



31-5 A capacitive pressure sensor. When force is applied, the spacing between the plates decreases, causing the capacitance to increase and the oscillator frequency to go down.



31-6 An elastomer pressure sensor detects applied force without unwanted capacitive effects.

Elastomer

When we want to avoid the effects of body capacitance, we can use an *elastomer* device instead of a capacitive device for pressure sensing. An elastomer is a flexible substance resembling rubber or plastic that can be used to detect the presence or absence of mechanical pressure.

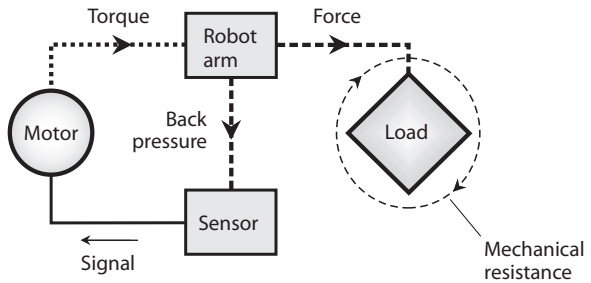
Figure 31-6 illustrates how we can employ an elastomer to detect and locate a pressure point. The elastomer conducts electricity fairly well, but not perfectly. It has a foam-like consistency, so that it can be compressed. Conductive plates are attached to opposite faces of the elastomer pad.

When pressure appears at some point in the pad, the material compresses, and its electrical resistance goes down. The drop in resistance produces an increase in the current between the plates. As the applied pressure increases, the elastomer grows thinner, and the current goes up more. The current-change data can be sent to a microcomputer, such as a robot controller.

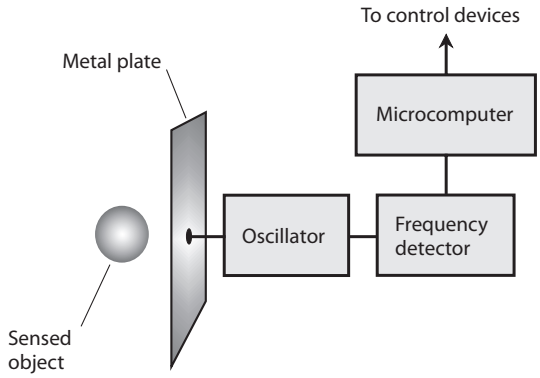
Back-Pressure Sensor

A motor produces a measurable pressure that depends on the *torque* ("turning force") that we apply. A *back-pressure sensor* detects and measures the torque that the motor exerts at any given instant in time. The sensor produces a signal, usually a variable voltage, that increases as the torque increases. Figure 31-7 is a functional diagram of the system.

Robotics engineers use back-pressure sensors to limit the forces applied by robot grippers, arms, drills, hammers, or other so-called *end effectors*. The *back voltage*, or signal produced by the sensor,



31-7 A back-pressure sensor governs the force applied by a robot arm or other mechanical device.



31-8 A capacitive proximity sensor can detect nearby conducting or semiconducting objects.

reduces the torque applied by the motor, preventing damage to objects that the robot handles, and also ensuring the safety of people working around the robot.

Capacitive Proximity Sensor

A *capacitive proximity sensor* uses an RF oscillator, a frequency detector, and a metal plate connected into the oscillator, as shown in Fig. 31-8. The resulting device *takes advantage* of the body capacitance effects that can confound capacitive pressure sensors. We design the oscillator so that any variation in the capacitance of the plate, with respect to the environment, causes the oscillator frequency to change. The frequency detector senses this change and transmits a signal to a microcomputer or robot controller.

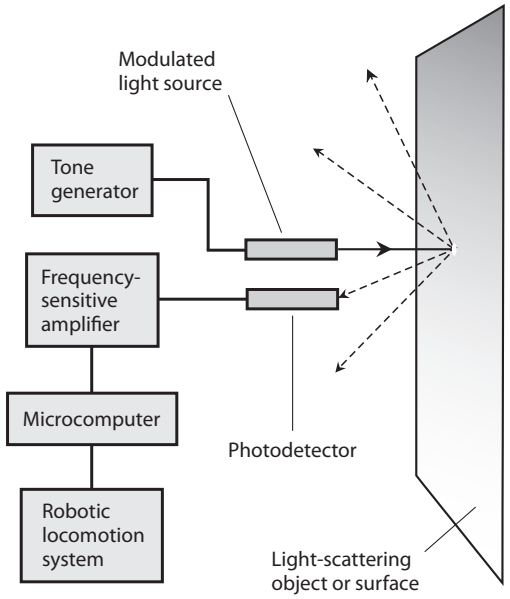
Substances that conduct electricity to some extent (such as metal, salt water, and living tissue) are sensed more easily by capacitive transducers than are materials that do not conduct, such as dry wood, plastic, glass, or dry fabric. For this reason, capacitive proximity sensors work poorly, if at all, in environments that lack conductive objects. A machine shop would present a better venue for a robot with capacitive proximity sensing than, say, a child's bedroom.

Photoelectric Proximity Sensor

Reflected light can help a robot "know" when it's approaching a physical barrier. A *photoelectric proximity sensor* contains a light-beam generator, a photodetector, a frequency-sensitive amplifier, and a microcomputer, interconnected and operated as shown in Fig. 31-9.

The light beam reflects from the object, and the photodetector picks up some of the reflected light. The tone generator modulates the light beam at a certain frequency, say, 1000 Hz. The photodetector's amplifier responds only to light modulated at that frequency. This modulation scheme prevents false imaging that could otherwise result from stray illumination from flashlights or sunlight. (Such light sources are unmodulated, and won't actuate a sensor designed to respond only to modulated light.) As the robot approaches an object, the robot controller senses the increasing intensity of the reflected, modulated beam. The robot can then steer clear of the obstruction.

Photoelectric proximity sensing doesn't work for objects that don't reflect light, or for shiny objects, such as glass windows or mirrors, approached at a sharp angle. In these scenarios, none of the light beam reflects back toward the photodetector, so the object remains "invisible" to the robot.



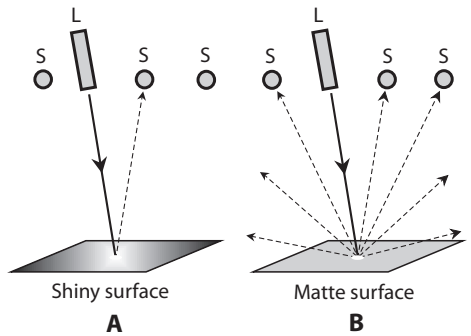
31-9 A photoelectric proximity sensor. Modulation of the light beam allows the device to distinguish between sensor-generated light and background illumination.

Texture Sensor

Texture sensing is the ability of a machine to determine whether an object has a shiny surface or a rough (matte) surface. A simple texture sensor contains a laser and several light-sensitive receptors.

Figure 31-10 shows how a combination of laser (L) and sensors (S) can tell the difference between a flat shiny surface (drawing A) and a flat matte (roughened) surface (drawing B). The shiny surface, such as a flat glass mirror or flat pane of glass, reflects light at the incidence angle only. But the matte surface, such as a sheet of paper, scatters light in all directions. The shiny surface reflects the beam back entirely to the sensor in the path of the beam whose reflection angle equals its incidence angle. The matte surface reflects the beam back to all the sensors. We can program a microcomputer to tell the difference.

Certain types of surfaces can confuse a texture sensor of the type portrayed in Fig. 31-10. For example, we would define a pile of small ice cubes as shiny on a microscopic scale but rough on a large scale. Depending on the diameter of the laser beam, the texture sensor might interpret such a



31-10 In texture sensing, lasers (L) and sensors (S) analyze a shiny surface (at A) and a matte surface (at B). Solid lines represent incident light; dashed lines represent reflected light.

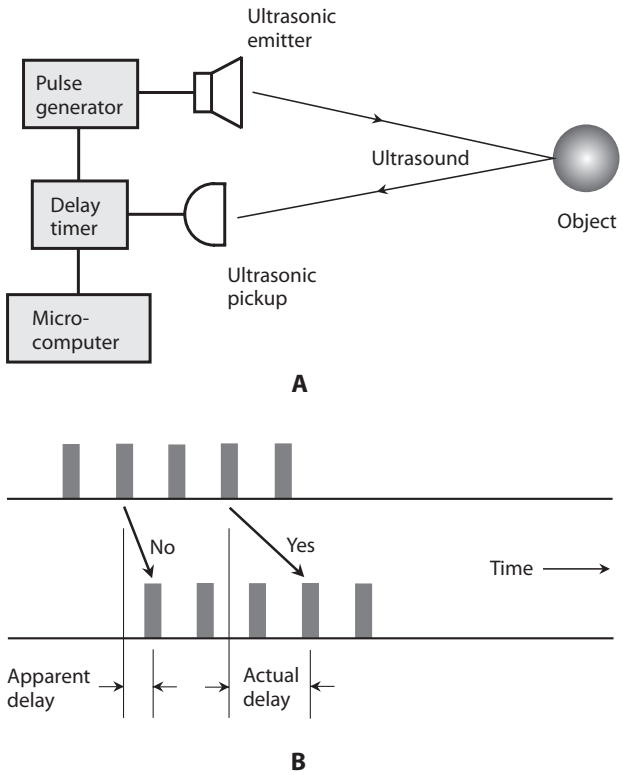
surface as either shiny or matte. The determination can also be affected by the motion of the sensor relative to the surface. A surface interpreted as shiny when standing still relative to the sensor might be interpreted as matte when moving relative to the sensor.

Sonar

Sonar is a medium-range method of proximity sensing. The acronym derives from the words *sonic navigation and ranging*. The principle is simple: Bounce acoustic waves off of objects, and measure the time it takes for the echoes to return.

Figure 31-11A shows a simple sonar system. The microcomputer can generate a *computer map* on the basis of sounds returned from various directions in two or three dimensions. This map can help a mobile robot or vessel navigate in its environment. However, the system can be "fooled" if the echo delay equals or exceeds the time interval between pulses, as shown in Fig. 31-11B. To overcome this conundrum, the microcomputer can instruct the pulse generator to send pulses of various frequencies in a defined, rotating sequence. The microcomputer keeps track of which echo corresponds to which pulse.

Acoustic waves travel faster in water than they do in the air. The amount of salt in water makes a difference in the propagation speed when sonar is used on boats (in depth finding, for example). The density of water can vary because of temperature differences as well. If the true speed of the acoustic waves is not accurately known, false readings will result. In fresh water, acoustic waves travel at about



31-11 At A, a block diagram of a medium-range sonar system. At B, unexpectedly long delays can confuse a sonar system.

1400 meters per second (m/s), or 4600 feet per second (ft/s). In salt water, acoustic waves travel at about 1500 m/s (4900 ft/s). In air, acoustic waves travel at approximately 335 m/s (1100 ft/s).

In the atmosphere, sonar can operate with audible sound waves, but ultrasound is often used instead. Ultrasound has a frequency too high to hear, ranging from about 20 kHz to more than 100 kHz. The most significant advantage of ultrasound is the fact that people who work around the sonar devices can't hear the signals, and will, therefore, not experience distraction, headaches, or other ill effects from them. In addition, ultrasonic sonar is less likely than audible sonar to be confused by people talking, heavy equipment, loud music, and other common noise sources. At frequencies higher than the range of human hearing, acoustical disturbances don't occur as often, or with as much intensity, as they do within the hearing range.

Low-cost sonar modules are available in a form that makes them very easy to use with microcontrollers. Searching for HC-SR04 will find examples of these.

Quiz

To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

32 CHAPTER

Antennas for RF Communications

WE CAN CATEGORIZE RF ANTENNAS INTO RECEIVING TYPES AND TRANSMITTING TYPES. NEARLY ALL transmitting antennas can receive signals quite well within the design frequency range. Some, but not all, receiving antennas can transmit RF signals with reasonable efficiency.

Radiation Resistance

When RF current flows in an electrical conductor, such as a wire or a length of metal tubing, some EM energy radiates into space. Imagine that we connect a transmitter to an antenna and test the whole system. Then we replace the antenna with a resistance/capacitance (*RC*) or resistance/inductance (*RL*) circuit and adjust the component values until the transmitter behaves exactly as it did when connected to the real antenna. For any antenna operating at a specific frequency, there exists a unique resistance R_R, in ohms, for which we can make a transmitter "think" that an *RC* or *RL* circuit is, in fact, that antenna. We call R_R the *radiation resistance* of the antenna.

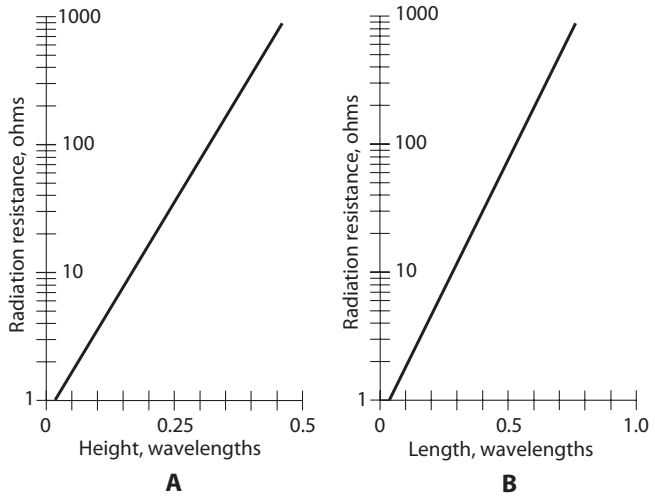
Determining Factors

Suppose that we place a thin, straight, lossless vertical wire over flat, horizontal, perfectly conducting ground with no other objects in the vicinity, and feed the wire with RF energy at the bottom. In this situation, the radiation resistance R_R of the wire is a function of its height in wavelengths. If we graph the function, we get Fig. 32-1A.

Now imagine that we string up a thin, straight, lossless wire in free space (such as a vacuum with no other objects anywhere nearby) and feed it with RF energy at the center. In this case, R_R is a function of the overall conductor length in wavelengths. If we graph the function, we obtain Fig. 32-1B.

Antenna Efficiency

We rarely have to worry about *antenna efficiency* in a receiving system, but in a transmitting antenna system, efficiency always matters. The efficiency expresses the extent to which a transmitting antenna converts the applied RF power to actual radiated EM power. In an antenna, radiation resistance R_R



32-1 Approximate values of radiation resistance for vertical antennas over perfectly conducting ground (A) and for center-fed antennas in free space (B).

always appears in series with a certain *loss resistance* (R_L). We can calculate the antenna efficiency, *Eff*, as a ratio with the formula

$$Eff = R_R / (R_R + R_L)$$

As a percentage, we have

$$Eff_{\%} = 100\ R_R / (R_R + R_L)$$

We can obtain high efficiency in a transmitting antenna only when the radiation resistance *greatly exceeds* the loss resistance. In that case, most of the applied RF power "goes into" useful EM radiation, and relatively little power gets wasted as heat in the earth and in objects surrounding the antenna. When the radiation resistance is comparable to or smaller than the loss resistance, a transmitting antenna behaves in an inefficient manner. This situation often exists for extremely short antenna radiators because they exhibit low radiation resistance. If we want reasonable efficiency in an antenna with a low R_R value, we must do everything we can to minimize R_L. Even the most concerted efforts rarely reduce R_L to less than a few ohms.

If an antenna system has a high loss resistance, it can work efficiently if we design it to exhibit an extremely high radiation resistance. When an antenna radiator measures a certain height or length at a given frequency, and if we construct it of low-loss wire or metal tubing, we can get its radiation resistance to exceed 1000 Ω. Then we can construct an efficient antenna even in the presence of substantial loss resistance.

Half-Wave Antennas

We can calculate the physical span of an EM-field half wavelength in free space using the formula

$$L_{ft} = 492 / f_o$$

where L_{ft} represents the straight-line distance in feet, and f_o represents the frequency in megahertz. We can calculate the physical span of a half wavelength in meters, L_m, using the formula

$$L_m = 150 / f_o$$

Velocity Factor

The foregoing formulas represent theoretical ideals, assuming infinitely thin conductors that have no resistance at any EM frequency. Obviously, such antenna conductors do not exist in physical reality. The atomic characteristics of wire or metal tubing cause EM fields to travel along a real-world conductor a little more slowly than light propagates through free space. For ordinary wire, we must incorporate a *velocity factor v* of 0.95 (95 percent) to account for this effect. For tubing or large-diameter wire, v can range down to about 0.90 (90 percent).

Open Dipole

An *open dipole* or *doublet* comprises a half-wavelength radiator fed at the center, as shown in Fig. 32-2A. Each side or "leg" of the antenna measures a quarter wavelength long from the *feed point* (where the transmission line joins the antenna) to the end of the conductor. For a straight wire radiator, the length L_{ft}, in feet, at a design frequency f_o, in megahertz, for a center-fed, half-wavelength dipole is approximately

$$L_{ft} = 467/f_o$$

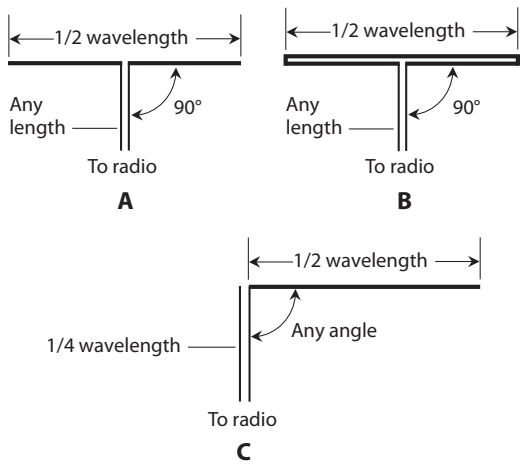
The length is meters is approximately

$$L_m = 143/f_o$$

These values assume that $v = 0.95$, as we would normally expect with copper wire of reasonable diameter. In free space, the impedance at the feed point of a center-fed, half-wave, open dipole constitutes a pure resistance of approximately 73 Ω. This resistance represents R_R alone. No reactance exists because a half-wavelength open dipole exhibits resonance, just as a tuned *RLC* circuit would if made with a discrete resistor, inductor, and capacitor. In fact, a dipole antenna (as well as many other types) constitutes an *RLC* circuit of a specialized, unique sort.

Folded Dipole

A *folded dipole antenna* consists of a half-wavelength, center-fed antenna constructed of two parallel wires with their ends connected together as shown in Fig. 32-2B. The feed-point impedance of the folded dipole is a pure resistance of approximately 290 Ω, or four times the feed-point resistance of a



32-2 Basic half-wave antennas. At A, the dipole. At B, the folded dipole. At C, the zepp.

half-wave open dipole made from a single wire. This "resistance-multiplication" property makes the folded dipole ideal for use in parallel-wire transmission lines, which usually have high characteristic impedance (Z_o) values.

Half-Wave Vertical

Imagine that we stand a half-wave radiator "on its end" and feed it at the *base* (the bottom end) against an earth ground, coupling the transmission line to the antenna through an *LC* circuit called an *antenna tuner* or *transmatch* designed to cancel out reactances over a wide range of values. Then we connect the other end of the feed line to a radio transmitter. This type of antenna works as an efficient radiator even in the presence of considerable loss resistance R_L in the conductors, the surrounding earth, and nearby objects, because the radiation resistance R_R is high.

Zepp

A *zeppelin antenna*, also called a *zepp*, comprises a half-wave radiator fed at one end with a quarter-wave section of parallel-wire transmission line, as shown in Fig. 32-2C. The impedance at the feed point is an extremely high, pure resistance. Because of the specific length of the transmission line, the transmitter "sees" a low, pure resistance at the operating frequency. A zeppelin antenna can function well at all harmonics of the design frequency. If we use a transmatch to "tune out" reactance, we can use any convenient length of transmission line.

Because of its non-symmetrical geometry, the zepp antenna allows some EM radiation from the feed line as well as from the antenna. That phenomenon can sometimes present a problem in radio transmitting applications. Amateur radio operators have an expression for it: "RF in the shack." We can minimize problems of this sort by carefully cutting the antenna radiator to a half wavelength at the fundamental frequency, and by using the antenna only at (or extremely near) the fundamental frequency or one of its harmonics.

J Pole

We can orient a zepp antenna vertically, and position the feed line so that it lies in the same line as the radiating element. The resulting antenna, called a *J pole*, radiates equally well in all horizontal directions. The J pole offers a low-cost alternative to metal-tubing vertical antennas at frequencies from approximately 10 MHz up through 300 MHz. In effect, the J-pole is a half-wavelength vertical antenna fed with an *impedance matching section* comprising a quarter-wavelength section of transmission line. The J pole does not require any electrical ground system, a feature that makes it convenient in locations with limited real estate.

Some radio amateurs "hang" long J poles, cut for 3.5 MHz or 1.8 MHz, from kites or helium-filled balloons. Such an antenna offers amazing performance, but it presents a danger if not tethered to prevent it from breaking off and flying away with the kite or balloon. Such an antenna must never be "flown" where it might fall on a utility or power lines. Long, kite- or balloon-supported wire antennas can acquire massive electrostatic charges, even in clear weather—and they *literally* attract lightning. "Flying" such an antenna in unstable weather invites deadly disaster.

Quarter-Wave Verticals

The physical span of a quarter wavelength antenna is related to frequency according to the formula

$$L_{ft} = 246v/f_o$$

where L_{ft} represents a quarter wavelength in feet, f_o represents the frequency in megahertz, and v represents the velocity factor. If we express the length in meters as L_m, then the formula becomes

$$L_m = 75v/f_o$$

For a typical wire conductor, $v = 0.95$ (95 percent); for metal tubing, v can range down to approximately 0.90 (90 percent).

A quarter-wavelength vertical antenna must be operated against a low-loss RF ground if we want reasonable efficiency. The feed-point value of R_R over perfectly conducting ground is approximately 37 Ω, half the radiation resistance of a center-fed half-wave open dipole in free space. This figure represents radiation resistance in the absence of reactance, and provides a reasonable impedance match to most coaxial-cable type transmission lines.

Ground-Mounted Vertical

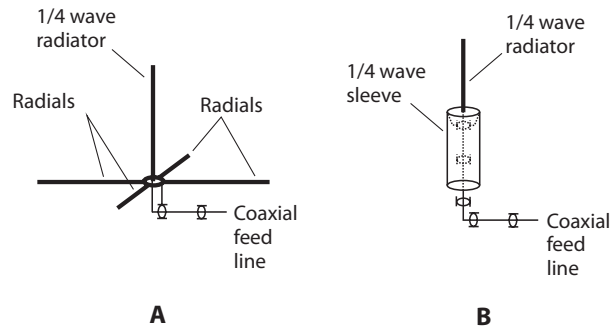
The simplest vertical antenna comprises a quarter-wavelength radiator mounted at ground level. The radiator is fed at the base with coaxial cable. The cable's center conductor is connected to the base of the radiator, and the cable's shield is connected to ground.

Unless we install an extensive *ground radial* system with a quarter-wave vertical antenna, it will exhibit poor efficiency unless the earth's surface in the vicinity is an excellent electrical conductor (salt water or a salt marsh, for example). In receiving applications, vertically oriented antennas "pick up" more human-made noise than horizontal antennas do. The EM fields from ground-mounted transmitting antennas are more likely to interfere with nearby electronic devices than are the EM fields from antennas installed high above the ground.

Ground Plane

A *ground-plane antenna* is a vertical radiator, usually ¼ wavelength tall, operated against a system of ¼-wavelength conductors called *radials*. The feed point, where the transmission line joins the radiator and the hub of the radial system, is elevated. When we place the feed point at least ¼ wavelength above the earth, we need only three or four radials to obtain low loss resistance for high efficiency. We extend the radials straight out from the feed point at an angle between 0° (horizontal) and 45° below the horizon. Figure 32-3A illustrates a typical ground-plane antenna.

A ground-plane antenna works best when fed with coaxial cable. The feed-point impedance of a ground-plane antenna having a quarter-wavelength radiator is about 37 Ω if the radials are horizontal; the impedance increases as the radials *droop*, reaching about 50 Ω at a *droop angle* of 45°.



32-3 Basic quarter-wave vertical antennas. At A, the ground-plane design. At B, the coaxial design.

You've seen ground-plane antennas if you've spent much time around *Citizens Band* (CB) fixed radio installations that operate near 27 MHz, or if you've done much amateur-radio activity in the very-high-frequency (VHF) bands at 50 or 144 MHz.

Coaxial Antenna

We can extend the radials in a ground-plane antenna straight downward, and then merge them into a quarter-wavelength-long cylinder or sleeve concentric with the coaxial-cable transmission line. We run the feed line inside the radial sleeve, feeding the antenna "through the end" as shown in Fig. 32-3B. The feed-point radiation resistance equals approximately 73 Ω, the same as that of a half-wave open dipole. This type of antenna is sometimes called a *coaxial antenna*, a term that arises because of its feed-system geometry, not merely because it's fed with coaxial cable.

Loops

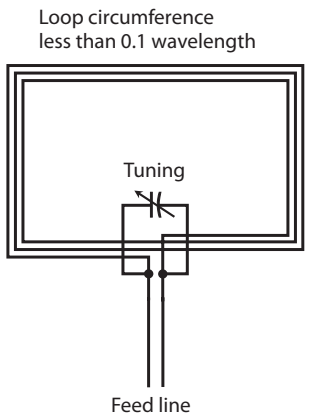
Any receiving or transmitting antenna made up of one or more turns of wire or metal tubing constitutes a *loop antenna*.

Small loop

A *small loop antenna* has a circumference of less than 0.1 wavelength (for each turn) and can function effectively for receiving RF signals. However, because of its small physical size, this type of loop exhibits low radiation resistance, a fact that makes RF transmission inefficient unless the conductors have minimal loss. While small transmitting loops exist, you won't encounter them often.

Even if a small loop has many turns of wire and contains an overall conductor length equal to a large fraction of a wavelength or more, the radiation resistance of a practical antenna is a function of its *actual circumference* in space, not the total length of its conductors. Therefore, for example, if we wind 100 turns of wire around a hoop that's only 0.1 wavelength in circumference, we have a low radiation resistance even though the total length of wire equals 10 full wavelengths!

A small loop exhibits the poorest response to signals coming from along its axis, and the best response to signals arriving in the plane perpendicular to its axis. A variable capacitor can be connected in series or parallel with the loop and adjusted until its capacitance, along with the inherent inductance of the loop, produces resonance at the desired receiving frequency. Figure 32-4 shows an example.



32-4 A small loop antenna with a capacitor for adjusting the resonant frequency.

Communications engineers and radio amateurs sometimes use small loops for *radio direction finding* (RDF) at frequencies up to about 20 MHz, and also for reducing interference caused by human-made noise or strong local signals. A small loop exhibits a sharp, deep *null* along its axis (perpendicular to the plane in which the conductor lies). When the loop is oriented so that the null points in the direction of an offending signal or noise source, the unwanted energy can be attenuated by upwards of 20 dB.

Loopstick

For receiving applications at frequencies up to approximately 20 MHz, a *loopstick antenna* can function in place of a small loop. This device consists of a coil wound on a solenoidal (rod-shaped), powdered-iron core. A series or parallel capacitor, in conjunction with the coil, forms a tuned circuit. A loopstick displays directional characteristics similar to those of the small loop antenna shown in Fig. 32-4. The sensitivity is maximum off the sides of the coil (in the plane perpendicular to the coil axis), and a sharp null occurs off the ends (along the coil axis).

Large Loop

A *large loop antenna* usually has a circumference of either a half wavelength or a full wavelength, forms a circle, hexagon, or square in space, and lies entirely in a single plane. It can work well for transmitting or receiving.

A half-wavelength loop presents a high radiation resistance at the feed point. Maximum radiation/response occurs in the plane of the loop, and a shallow, rather broad null exists along the axis. A full-wavelength loop presents a radiation resistance (and zero reactance, forming a purely resistive impedance) of about 100 Ω at the feed point. The maximum radiation/response occurs along the axis, and minimum radiation/response (though not a true null) exists in the plane containing the loop.

The half-wavelength loop exhibits a slight power loss relative to a half-wave open or folded dipole in its *favored directions* (the physical directions in which it offers the best performance). The full-wavelength loop shows a slight gain over a dipole in its favored directions. These properties hold for loops up to several percent larger or smaller than exact half-wavelength or full-wavelength circumferences. Resonance can be obtained by means of a transmatch at the feed point, even if the loop itself does not exhibit resonance at the frequency of interest.

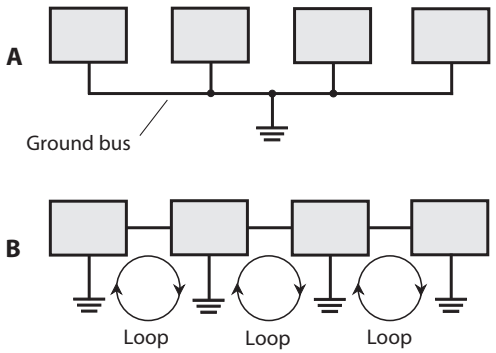
An extremely large loop antenna, measuring several wavelengths in circumference, can be installed horizontally among multiple supports such as communications towers, trees, or wooden poles. We might call this type of antenna a *giant loop*. The gain and directional characteristics of giant loops are hard to predict. If fed with open wire line using a transmatch at the transmitter end of the line, and if placed at least a quarter wavelength above the earth's surface, such an antenna can offer exceptional performance for transmitting and receiving.

Ground Systems

End-fed quarter-wavelength antennas require low-loss RF ground systems to perform efficiently. Center-fed half-wavelength antennas do not. However, good grounding is advisable for any antenna system to minimize interference and electrical hazards.

Electrical versus RF ground

Electrical grounding constitutes an important consideration for personal safety. A good electrical (that is, DC and utility AC) ground can help protect communications equipment from damage if



32-5 At A, the correct method for grounding multiple hardware equipment. At B, an incorrect method for grounding multiple equipment, creating RF ground loops.

lightning strikes in the vicinity. A good electrical ground also minimizes the risk of *electromagnetic interference* (EMI) to and from radio equipment. In a three-wire electrical utility system, the ground prong on the plug should never be defeated because such modification can result in dangerous voltages appearing on exposed metal surfaces.

A good RF ground system can help minimize EMI, even if it's not necessary for efficient antenna operation. Figure 32-5 shows a proper RF ground scheme (at A) and an improper one (at B). In a good RF ground system, each device is connected to a common *ground bus*, which, in turn, runs to the earth ground through a single conductor. This conductor should have the smallest possible physical length. A poor ground system contains *ground loops* that can act like loop antennas and increase the risk of EMI.

Radials and the Counterpoise

A surface-mounted vertical antenna should employ as many grounded radial conductors as possible, and they should be as long as possible. The radials can lie on the earth's surface or be buried a few inches underground. In general, as the number of radials of a given length increases, the overall efficiency of any vertical antenna improves if all other factors remain constant. As the radial length increases and all other factors remain constant, vertical-antenna efficiency improves. The radials should all converge toward, and connect directly to, a ground rod at the feed point.

A specialized conductor network called a *counterpoise* can provide an RF ground without a direct earth-ground connection. A grid of wires, screen, or metal sheet is placed above the earth's surface and oriented horizontally to obtain *capacitive coupling* to the earth's conductive mass. A vertical antenna is located at the center of the counterpoise. This arrangement minimizes RF ground loss, although the counterpoise won't provide a good electrical ground unless connected to a *ground rod* driven into the earth, or to the utility system ground. Ideally, a counterpoise should have a radius of at least a quarter wavelength at the lowest anticipated operating frequency.

Gain and Directivity

The *power gain* of a transmitting antenna equals the ratio of the maximum *effective radiated power* (ERP) to the actual RF power applied at the feed point. Power gain is expressed in decibels (dB). It's usually expressed in an antenna's favored direction or directions.

Suppose that the ERP, in watts, for a given antenna equals P_{ERP}, and the applied power, also in watts, equals P. We can calculate the antenna power gain using the formula

$$\text{Power Gain (dB)} = 10 \log_{10} (P_{ERP}/P)$$

In order to define power gain, we must use a *reference antenna* with a gain that we define as 0 dB in its favored direction(s). A half-wavelength open or folded dipole in free space provides a useful reference antenna. Power gain figures taken with respect to a dipole (in its favored directions) are expressed in units called dBd. Some engineers make power-gain measurements relative to a specialized system known as an *isotropic antenna*, which theoretically radiates and receives equally well in all directions in three dimensions (so it has no favored direction). In this case, units of power gain are called dBi.

For any given antenna, the power gains in dBd and dBi differ by approximately 2.15 dB, with the dBi figure turning out larger. That is,

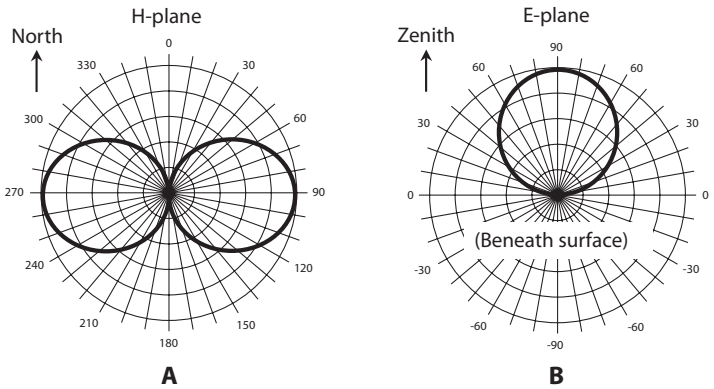
$$\text{Power Gain (dBi)} = \text{Power Gain (dBd)} + 2.15$$

An isotropic antenna exhibits a *loss* of 2.15 dB with respect to a half-wave dipole in its favored directions. This fact becomes apparent if we rewrite the above formula as

$$\text{Power Gain (dBd)} = \text{Power Gain (dBi)} - 2.15$$

Directivity Plots

We can portray antenna *radiation patterns* (for signal transmission) and response patterns (for reception) using graphical plots, such as those in Fig. 32-6. We assume, in all such plots, that the antenna occupies the center (or *origin*) of a *polar coordinate system*. The greater the radiation or



32-6 Directivity plots for a dipole. At A, the H-plane (horizontal plane) plot as viewed from high above the antenna. Coordinate numbers indicate compass-bearing (azimuth) angles in degrees. At B, the E-plane (elevation plane) plot as viewed from a point on the earth's surface far from the antenna. Coordinate numbers indicate elevation angles in degrees above or below the plane of the earth's surface.

reception capability of the antenna in a certain direction, the farther from the center we plot the corresponding point.

A dipole antenna, oriented horizontally so that its conductor runs in a north-south direction, has a *horizontal plane* (or *H-plane*) pattern similar to Fig. 32-6A. The *elevation plane* (or *E-plane*) pattern depends on the height of the antenna above *effective ground* at the viewing angle. In most locations, the effective ground is an imaginary plane or contoured surface slightly below the actual surface of the earth. With the dipole oriented so that its conductor runs perpendicular to the page, and the antenna ¼ wavelength above effective ground, the E-plane pattern for a half-wave dipole resembles Fig. 32-6B.

Forward Gain

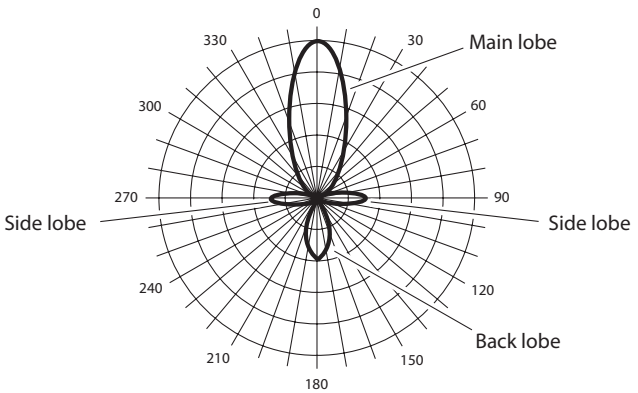
Forward gain is expressed in terms of the ERP in the *main lobe* (favored direction) of a *unidirectional* (one-directional) antenna compared with the ERP from a reference antenna, usually a half-wave dipole, in its favored directions. This gain is calculated and defined in dBd. In general, as the wavelength decreases (the frequency gets higher), we find it easier to obtain high forward gain figures.

Front-to-Back Ratio

The *front-to-back* (f/b) *ratio* of a unidirectional antenna quantifies the concentration of radiation/response *in the center* of the main lobe, relative to the direction *opposite the center* of the main lobe. Figure 32-7 shows a hypothetical directivity plot for a unidirectional antenna pointed north. The outer circle depicts the RF *field strength* in the direction of the center of the main lobe, and represents 0 dB (relative to the main lobe, not a dipole). The next smaller circle represents a field strength 5 dB down (a radiation/response level of −5 dB) with respect to the main lobe. Continuing inward, circles represent 10 dB down (−10 dB), 15 dB down (−15 dB), and 20 dB down (−20 dB). The origin represents 25 dB down (−25 dB) with respect to the main lobe, and also shows the location of the antenna. In this particular example, we can determine the f/b ratio by comparing the signal levels between north (azimuth 0°) and south (azimuth 180°). It appears to be 15 dB in Fig. 32-7.

Front-to-Side Ratio

The *front-to-side* (f/s) *ratio* provides us with another useful expression for the directivity of an antenna system. The specification applies to unidirectional antennas, and also to *bidirectional*



32-7 Directivity plot for a hypothetical antenna in the H (horizontal) plane. We can determine the front-to-back and front-to-side ratios from such a graph. Coordinate numbers indicate compass-bearing angles in degrees.

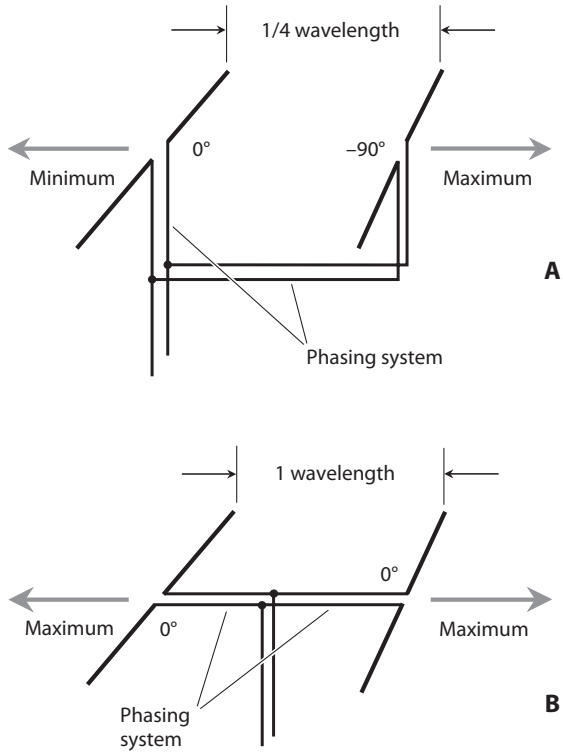
antennas that have two favored directions, one opposite the other in space. We express f/s ratios in decibels (dB), just as we do with f/b ratios. We compare the EM field strength *in the favored direction* with the field strength *at right angles* to the favored direction. Figure 32-7 shows an example. In this situation, we can define two separate f/s ratios: one by comparing the signal level between north and east (the right-hand f/s ratio), and the other by comparing the signal level between north and west (the left-hand f/s ratio). In most directional antenna systems, the right-hand and left-hand f/s ratios theoretically equal each other. However, they sometimes differ in practice because of physical imperfections in the antenna structure, and also because of the effects of conducting objects or an irregular earth surface near the antenna. In the situation of Fig. 32-7, both the left-hand and right-hand f/s ratios appear to be roughly 17 dB.

Phased Arrays

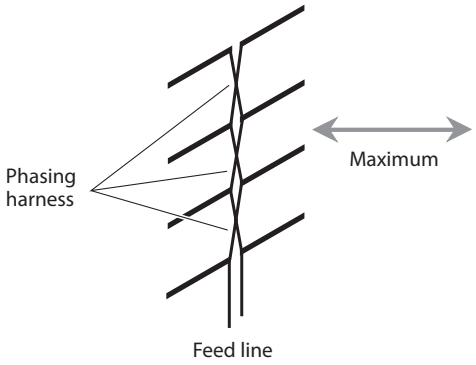
A *phased antenna array* uses two or more *driven elements* (radiators connected directly to the feed line) to produce power gain in some directions at the expense of other directions.

End-Fire Array

A typical *end-fire array* consists of two parallel half-wave open dipoles fed 90° out of phase and spaced ¼ wavelength apart, as shown in Fig. 32-8A. This geometry produces a unidirectional radiation pattern. Alternatively, the two elements can be driven in phase and spaced at a separation of one full wavelength, as shown in Fig. 32-8B, producing a bidirectional radiation pattern. When we



32-8 At A, a unidirectional end-fire antenna array. At B, a bidirectional end-fire array.



32-9 A broadside array. The elements all receive their portions of the outgoing signal in phase, and at equal amplitudes.

design the *phasing system* (also called the *phasing harness*), we must cut the branches of the transmission line to precisely the correct lengths, taking the velocity factor of the line into account.

Longwire

A wire antenna measuring a full wavelength or more, and fed at a high-current point or at one end, constitutes a *longwire antenna*. A longwire antenna offers gain over a half-wave dipole. As we increase the length of the wire, making sure that it runs along a straight line for its entire length, the main lobes get more and more nearly in line with the antenna, and their magnitudes increase. The power gain in the *major lobes* (that is, the strongest lobes, representing the favored directions) in a straight longwire depends on the overall length of the antenna; the longer the wire, the greater the gain. We can consider a longwire as a phased array because it contains multiple points along its span where the RF current attains maximum values. Each of these *current loops* acts like the center of a half-wave dipole or zepp antenna, so the whole longwire, in effect, constitutes a set of two or more half-wave antennas placed end-to-end, with each section phased in opposition relative to the adjacent section or sections.

Broadside Array

Figure 32-9 shows the geometric arrangement of a *broadside array*. The driven elements can each consist of a single radiator, as shown in this illustration, or they can consist of more complex antennas with directive properties. In the design shown here, all the driven elements are identical, and are spaced at ½-wavelength intervals along the phasing harness, which comprises a parallel-wire transmission line with a half-twist between each element to ensure that all the elements operate in phase coincidence. If we place a flat reflecting screen behind the array of dipoles in Fig. 32-9, we obtain a system known as a *billboard antenna*. The directional properties of any broadside array depend on the number of elements, on whether or not the elements have gain themselves, on the spacing among the elements, and on whether or not a reflecting screen is employed. In general, as we increase the number of elements, the forward gain, the f/b ratio, and the f/s ratios all increase.

Parasitic Arrays

Communications engineers use so-called *parasitic arrays* at frequencies ranging from approximately 5 MHz into the microwave range for obtaining directivity and forward gain. Examples include the *Yagi antenna* and the *quad antenna*. In the context of an antenna array, the term *parasitic* describes

the characteristics of certain antenna elements. (It has nothing to do with parasitic oscillation, a phenomenon that can take place in malfunctioning RF power amplifiers.)

Concept

A *parasitic element* is an electrical conductor that forms an important part of an antenna system, but that we don't connect directly to the feed line. Parasitic elements operate by means of *EM coupling* to the driven element or elements. When power gain occurs in the direction of the parasitic element, we call that element a *director*. When power gain occurs in the direction opposite the parasitic element, we call that element a *reflector*. Directors normally measure a few percent shorter than the driven element(s). Reflectors are normally a few percent longer than the driven element(s).

Yagi

The *Yagi antenna*, which radio amateurs sometimes call a *beam antenna*, or simply a *beam*, comprises an array of parallel, straight antenna elements, with at least one element acting in a parasitic capacity. (The term *Yagi* comes from the name of one of the original design engineers.) We can construct a two-element Yagi by placing a director or reflector parallel to, and a specific distance away from, a single half-wave driven element. The optimum spacing between the elements of a *driven-element/director Yagi* is 0.1 to 0.2 wavelength, with the director tuned 5 to 10 percent higher than the resonant frequency of the driven element. The optimum spacing between the elements of a *driven-element/reflector Yagi* is 0.15 to 0.2 wavelength, with the reflector tuned 5 to 10 percent lower than the resonant frequency of the driven element. Either of these designs give us a *two-element Yagi*. The power gain of a well-designed two-element Yagi in its single-favored direction is approximately 5 dBd.

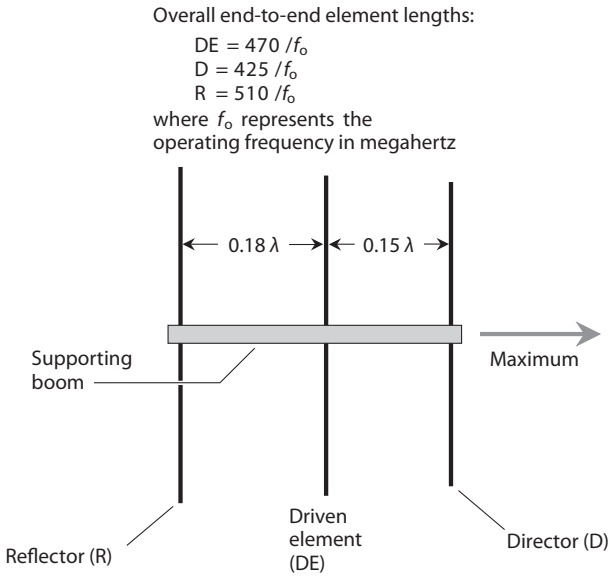
A Yagi with one director and one reflector, along with the driven element, forms a *three-element Yagi*. This design scheme increases the gain and f/b ratio as compared with a two-element Yagi. An optimally designed three-element Yagi exhibits approximately 7 dBd gain in its favored direction. Figure 32-10 is a generic example of the relative dimensions of a three-element Yagi. Although this illustration can serve as a crude "drawing-board" engineering blueprint, the optimum dimensions of a three-element Yagi in practice will vary slightly from the figures shown here because of imperfections in real-world hardware, and also because of conducting objects or terrain irregularities near the system.

The gain, f/b ratio, and f/s ratios of a properly designed Yagi antenna all increase as we add elements to the array. We can obtain four-element, five-element, or larger Yagis by placing extra directors in front of a three-element Yagi. When multiple directors exist, each one should be cut slightly shorter than its predecessor. Some commercially manufactured Yagis have upwards of a dozen elements. As you can imagine, engineers must spend a lot of time "tweaking" the dimensions of such antennas to optimize their performance.

Quad

A *quad antenna* operates according to the same principles as the Yagi, except that full-wavelength loops replace the half-wavelength elements.

A *two-element quad* can consist of a driven element and a reflector, or it can have a driven element and a director. A *three-element quad* has one driven element, one director, and one reflector. The director has a perimeter of 0.95 to 0.97 wavelength, the driven element has a perimeter of exactly one wavelength, and the reflector has a perimeter of 1.03 to 1.05 wavelength. These figures



32-10 A three-element Yagi antenna. See text for discussion of specific dimensions. The lowercase, italic Greek lambda (λ) means "wavelength."

represent electrical dimensions (taking the velocity factor of wire or tubing into account), and not free-space dimensions.

Additional directors can be added to the basic three-element quad design to form quads having any desired numbers of elements. The gain increases as the number of elements increases. Each succeeding director is slightly shorter than its predecessor. Long quad antennas are practical at frequencies above 100 MHz. At frequencies below approximately 10 MHz, quad antennas become physically large and unwieldy, although some radio amateurs have constructed quads at frequencies down to 3.5 MHz.

Antennas for Ultra-High and Microwave Frequencies

At ultra-high frequencies (UHF) and microwave frequencies, high-gain, many-element antennas have reasonable physical dimensions and mass because the wavelengths are short.

Waveguides

A *waveguide* is a specialized RF transmission line comprising a hollow metal "pipe" or "duct" with a circular or rectangular cross section. The EM field travels down the waveguide quite efficiently, provided that the wavelength is short enough (or the cross-sectional dimensions of the pipe are large enough). In order to efficiently propagate an EM field, a *rectangular waveguide* must have height and width that both measure at least 0.5 wavelength, and preferably more than 0.7 wavelength. A *circular waveguide* should measure at least 0.6 wavelength in diameter, and preferably 0.7 wavelength or more.

The characteristic impedance (Z_o) of a waveguide varies with the frequency. In this sense, a waveguide behaves differently than a coaxial or parallel-wire RF transmission line, whose Z_o value remains independent of the frequency over the entire range of wavelengths for which the line is designed.

A properly installed and maintained waveguide acts as an exceptional RF transmission line because dry air has essentially zero loss, even at UHF and microwave frequencies. However, if we

expect a waveguide to work properly, we must keep its interior free from dirt, dust, insects, spider webs, and condensation. Even a small obstruction can seriously degrade the performance and cause significant power loss.

The main limitation of a waveguide, from a practical standpoint, is its relative inflexibility, both figuratively and literally. We can't run a waveguide from one point to another in a haphazard fashion, as we can do with coaxial cable. Bends or turns in a waveguide present a particular problem, because we must make them gradually. We can't simply "turn a corner" with a waveguide! Another limitation involves the usable frequency range. Waveguides are impractical for use at frequencies below approximately 300 MHz because the required cross-sectional dimensions become prohibitively large.

Horn

The *horn antenna* has a characteristic shape like a squared-off trumpet horn. It provides a unidirectional radiation and response pattern, with the favored direction coincident with the opening of the horn. The feed line is a waveguide that joins the antenna at the narrowest point (throat) of the horn. Horns are sometimes used all by themselves, but they can also feed large *dish antennas* at UHF and microwave frequencies. The horn design optimizes the f/s ratio by minimizing extraneous radiation and response that occurs if a dipole is used as the driven element for the dish.

Dish

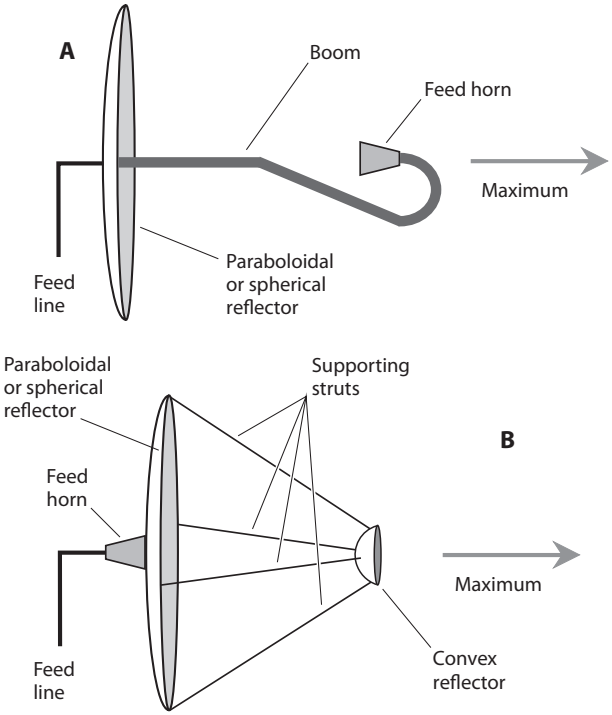
Most people are familiar with dish antennas because of their widespread use in consumer satellite TV and Internet services. Although the geometry looks simple to the casual observer, a dish antenna must be precisely shaped and aligned if we want it to function as intended. The most efficient dish, especially at the shortest wavelengths, comprises a *paraboloidal reflector*, so named because it's a section of a *paraboloid* (the three-dimensional figure that we get when we rotate a *parabola* around its axis). However, a *spherical reflector*, having the shape of a section of a sphere, can also work in most dish-antenna system designs.

A dish-antenna feed system consists of a coaxial line or waveguide from the receiver and/or transmitter along with a horn or helical driven element at the focal point of the reflector. Figure 32-11A shows an example of *conventional dish feed*. Figure 32-11B shows an alternative scheme known as *Cassegrain dish feed*. The term *Cassegrain* comes from the resemblance of this antenna design to that of a *Schmidt-Cassegrain reflector telescope*. All dish antennas work well at UHF and microwave frequencies for transmitting and receiving. For hobby use, they're physically impractical at frequencies much below 300 MHz.

As we increase the diameter of the dish reflector in wavelengths, the gain, the f/b ratio, and the f/s ratios all increase, and the width of the main lobe decreases, making the antenna more sharply unidirectional. A dish antenna must measure at least several wavelengths in diameter for proper operation. The reflecting element can consist of sheet metal, a screen, or a wire mesh. If a screen or mesh is used, the spacing between the wires must be a small fraction of a wavelength. At microwave frequencies, large dish antennas can have forward gain figures that exceed 35 dBd.

Helical

A *helical antenna* is a high-gain, unidirectional antenna that transmits and receives EM waves with circular polarization. Figure 32-12 illustrates the construction of a typical helical antenna. The reflector diameter should be at least 0.8 wavelength at the lowest operating frequency. The radius of the helix should be approximately 0.17 wavelength at the center of the intended operating frequency range. The longitudinal spacing between helix turns should be approximately 0.25 wavelength in the

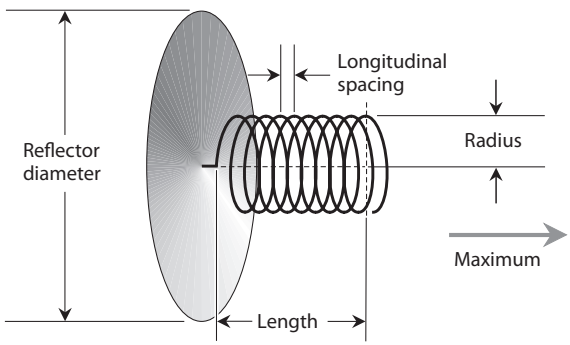


32-11 Dish antennas with conventional feed (A) and Cassegrain feed (B).

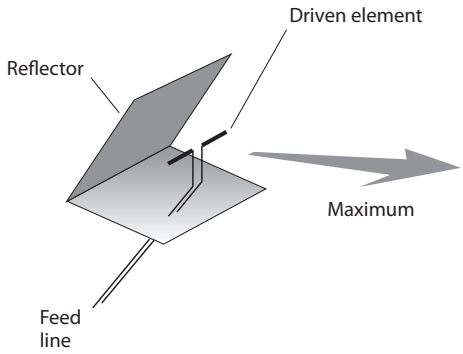
center of the operating frequency range. The entire helix should measure at least a full wavelength from end to end at the lowest operating frequency. When properly designed, this type of antenna can provide about 15 dBd forward gain. Helical antennas, like dish antennas, are used primarily at UHF and microwave frequencies.

Corner Reflector

Figure 32-13 illustrates a *corner reflector* with a half-wave open-dipole driven element. This design provides some power gain over a half-wave dipole. The reflector is made of wire mesh, screen, or sheet metal. The *flare angle* of the reflecting element equals approximately 90°. Corner reflectors are widely used in terrestrial communications at UHF and microwave frequencies. For additional



32-12 A helical antenna with a flat reflector.



32-13 A corner reflector that employs a dipole antenna as the driven element.

gain, several half-wave dipoles can be fed in phase and placed along a common axis with a single, elongated reflector, forming a *collinear corner-reflector array*.

Safety

All engineers who work with antenna systems, particularly large arrays or antennas involving the use of long lengths of wire, place themselves in physical peril. By observing some simple safety guidelines, the risk can be minimized, but there's never a complete guarantee of safety. Some basic guidelines follow.

Antennas must never be maneuvered or installed in such a way that they can fall or blow down on power lines. Also, it should not be possible for power lines to fall or blow down on an antenna, even in the event of a violent storm.

Wireless equipment having outdoor antennas should not be used during thundershowers, or when lightning exists anywhere in the vicinity. Antenna construction and maintenance should never be undertaken when lightning is visible or thunder can be heard, even if a storm appears far away. Ideally, antennas should be disconnected from electronic equipment, and connected to a substantial earth ground, at all times when the equipment is not in use.

Tower and antenna climbing constitutes a job for professionals only. No inexperienced person should ever attempt to climb any antenna support structure.

Indoor transmitting antennas can expose operating personnel to EM field energy. The extent of the hazard, if any, posed by such exposure has not been firmly established. However, sufficient concern exists among some experts to warrant checking the latest publications on the topic.

Warning!

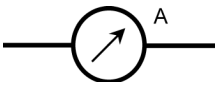
For complete information on antenna safety matters, you should consult a professional antenna engineer or a comprehensive reference devoted to antenna design and construction. You should also consult and heed all electrical and building codes for your city, state, or province.

Quiz

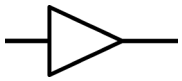
To test your knowledge of this chapter, you might like to try the online quiz for it. See the Preface of this book for details.

Schematic Symbols

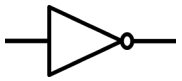
ammeter



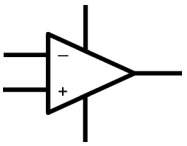
amplifier, general



amplifier, inverting



amplifier, operational



AND gate

antenna, balanced

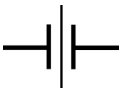
antenna, general

antenna, loop

antenna, loop, multiturn

battery, electrochemical

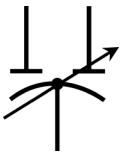
capacitor, feedthrough



capacitor, fixed

capacitor, variable

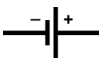
capacitor, variable, split-rotor



capacitor, variable, split-stator

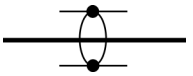
cavity resonator

cell, electrochemical



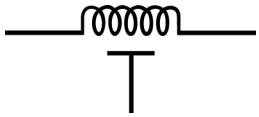
circuit breaker

coaxial cable



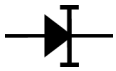
crystal, piezoelectric

delay line



diac

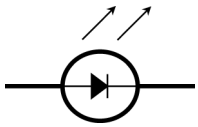
diode, field-effect



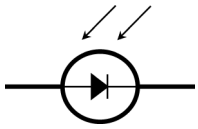
diode, general

diode, Gunn

diode, light-emitting

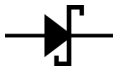


diode, photosensitive

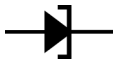


diode, PIN

diode, Schottky



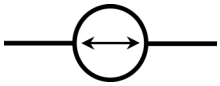
diode, tunnel



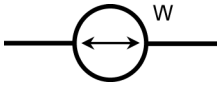
diode, varactor

diode, Zener

directional coupler



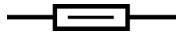
directional wattmeter



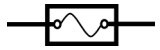
exclusive-OR gate

female contact, general

Ferrite bead



fuse



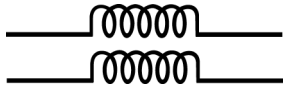
galvanometer

ground, chassis

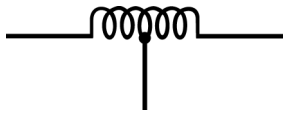
ground, earth

inductor, air core

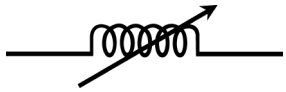
inductor, air core, bifilar



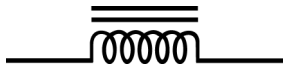
inductor, air core, tapped



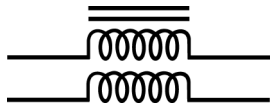
inductor, air core, variable



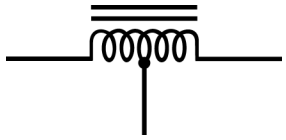
inductor, iron core



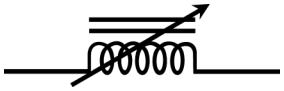
inductor, iron core, bifilar



inductor, iron core, tapped

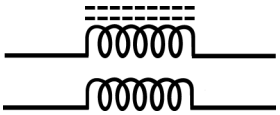


inductor, iron core, variable

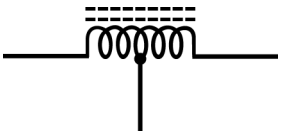


inductor, powdered-iron core

inductor, powdered-iron core, bifilar

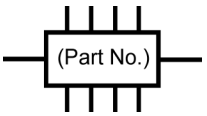


inductor, powdered-iron core, tapped

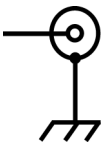


inductor, powdered-iron core, variable

integrated circuit, general



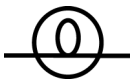
jack, coaxial or phono



jack, phone, 2-conductor

jack, phone, 3-conductor

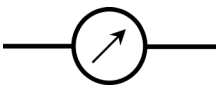
lamp, incandescent



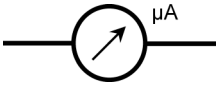
lamp, neon

male contact, general

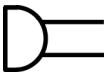
meter, general



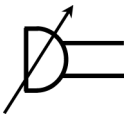
microammeter



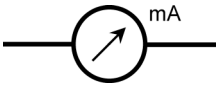
microphone



microphone, directional



milliammeter

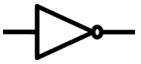


NAND gate

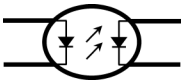
negative voltage connection

NOR gate

NOT gate



optoisolator



OR gate

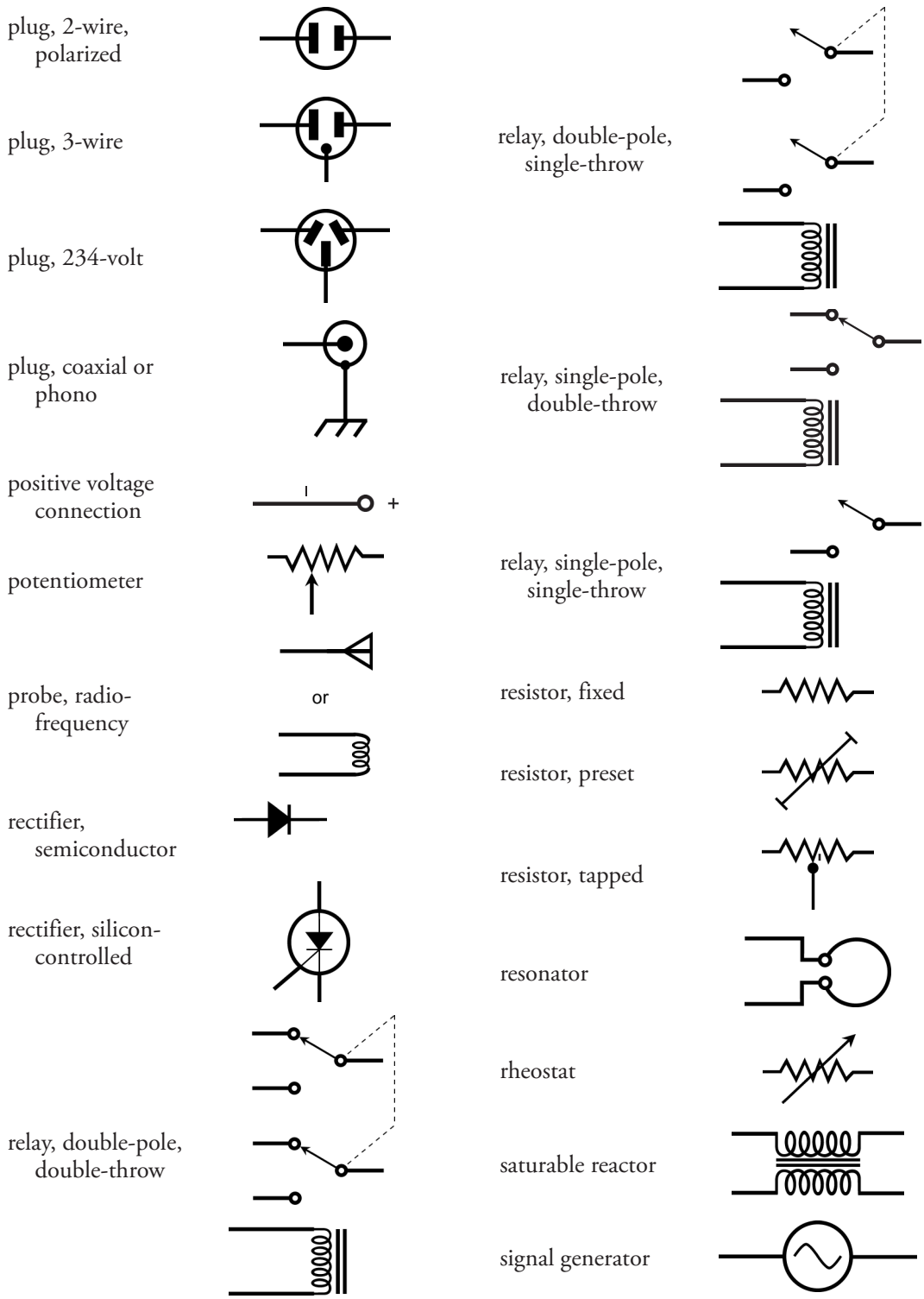
outlet, 2-wire, nonpolarized

outlet, 2-wire, polarized

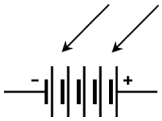
outlet, 3-wire

outlet, 234-volt

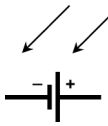
plug, 2-wire, nonpolarized


plug, 2-wire, polarized
plug, 3-wire
plug, 234-volt
plug, coaxial or phono
positive voltage connection
+
potentiometer
probe, radio-frequency
or
rectifier, semiconductor
rectifier, silicon-controlled
relay, double-pole, double-throw
relay, double-pole, single-throw
relay, single-pole, double-throw
relay, single-pole, single-throw
resistor, fixed
resistor, preset
resistor, tapped
resonator
rheostat
saturable reactor
signal generator

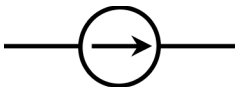
solar battery



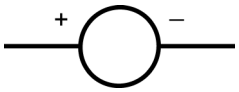
solar cell



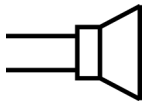
source, constant-current



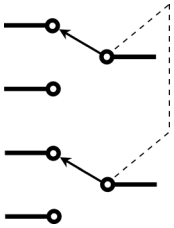
source, constant-voltage



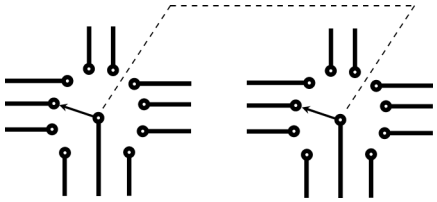
speaker



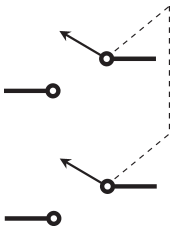
switch, double-pole, double-throw



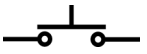
switch, double-pole, rotary



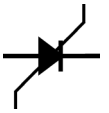
switch, double-pole, single-throw



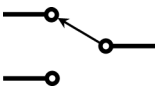
switch, momentary-contact



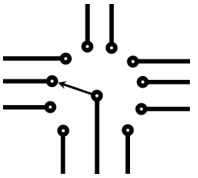
switch, silicon-controlled



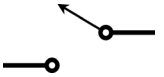
switch, single-pole, double-throw



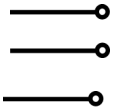
switch, single-pole, rotary



switch, single-pole, single-throw



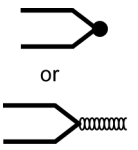
terminals, general, balanced



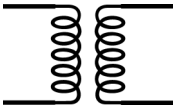
terminals, general, unbalanced

test point

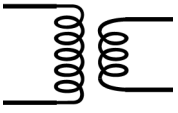
thermocouple



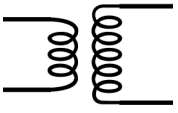
transformer, air core



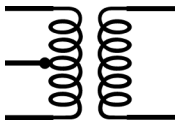
transformer, air core, step-down



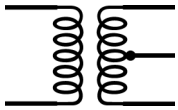
transformer, air core, step-up



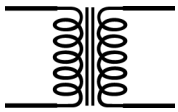
transformer, air core, tapped primary



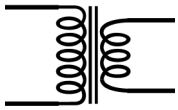
transformer, air core, tapped secondary



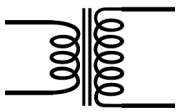
transformer, iron core



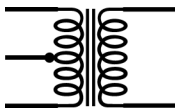
transformer, iron core, step-down



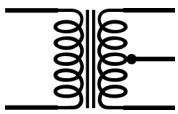
transformer, iron core, step-up



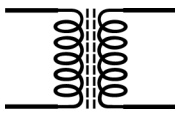
transformer, iron core, tapped primary



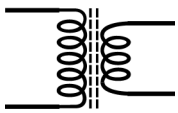
transformer, iron core, tapped secondary



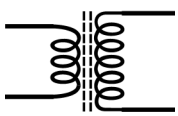
transformer, powdered-iron core



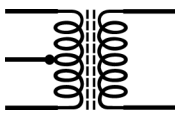
transformer, powdered-iron core, step-down



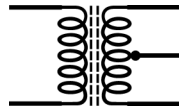
transformer, powdered-iron core, step-up



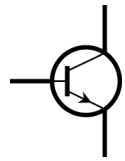
transformer, powdered-iron core, tapped primary



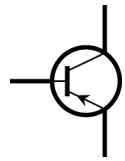
transformer, powdered-iron core, tapped secondary



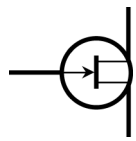
transistor, bipolar, NPN



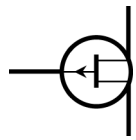
transistor, bipolar, PNP



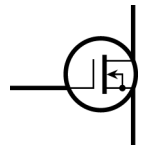
transistor, field-effect, N-channel



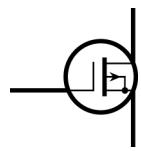
transistor, field-effect, P-channel



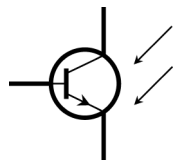
transistor, MOS field-effect, N-channel



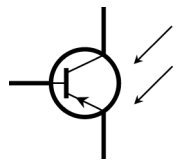
transistor, MOS field-effect, P-channel



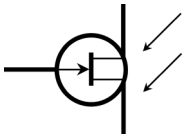
transistor, photosensitive, NPN



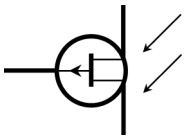
transistor, photosensitive, PNP



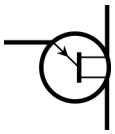
transistor, photosensitive, field-effect, N-channel



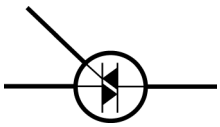
transistor, photosensitive, field-effect, P-channel



transistor, unijunction

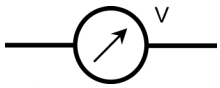


triac

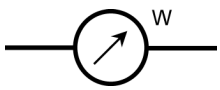


unspecified unit or component

voltmeter



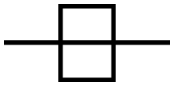
wattmeter



waveguide, circular

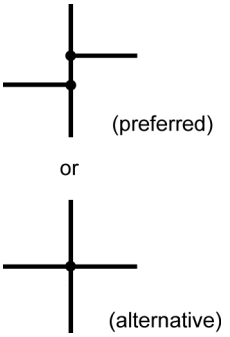
waveguide, flexible

waveguide, rectangular

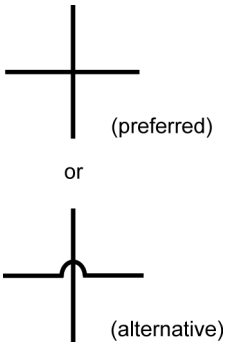


waveguide, twisted

wires, crossing, connected



wires, crossing, not connected



Suggested Additional Readings

- American Radio Relay League, Inc. *The ARRL Handbook for Radio Communications*. Newington, CT: ARRL, revised annually.
- Gibilisco, Stan, *Ham and Shortwave Radio for the Electronics Hobbyist*. New York: McGraw-Hill, 2014.
- Gibilisco, Stan, *Technical Math Demystified*. New York: McGraw-Hill, 2006.
- Gussow, Milton, *Schaum's Easy Outline of Basic Electricity Revised*. New York: McGraw-Hill, 2011.
- Kybett, Harry, *All New Electronics Self-Teaching Guide*, 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc., 2008.
- Kybett, Harry and Boysen, Earl, *Complete Electronics Self-Teaching Guide with Projects*, 4th ed. Hoboken, NJ: John Wiley & Sons, Inc., 2012.
- Mims, Forrest, *Getting Started in Electronics*. Niles, IL: Master Publishing, 2003.
- Monk, Simon, *Hacking Electronics*. New York: McGraw-Hill, 2013.
- Monk, Simon, *Programming Arduino: Getting Started with Sketches*. New York: McGraw-Hill, 2011.
- Monk, Simon, *Programming Arduino Next Steps: Going Further with Sketches*. New York: McGraw-Hill, 2013.
- Morrison, Ralph, *Electricity: A Self-Teaching Guide*, 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc., 2003.
- Platt, Charles, Make, Electronics: Learning by Discovery: A hands-on primer for the new electronics enthusiast, 2021.

- Santiago, John, *Circuit Analysis for Dummies*. Hoboken, NJ: John Wiley & Sons, Inc., 2013.
- Shamieh, Cathleen, *Electronics for Dummies*, 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc., 2015.
- Slone, Randy, *TAB Electronics Guide to Understanding Electricity and Electronics*, 2nd ed. New York: McGraw-Hill, 2000.

Index

Note: Page numbers followed by *f* indicate figures; and page numbers followed by *t* indicate tables.

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z